Alistair D. Fitt · John Norbury
Hilary Ockendon · Eddie Wilson Editors

Progress in Industrial Mathematics at ECMI 2008

Springer

MATHEMATICS IN INDUSTRY 15

THE EUROPEAN CONSORTIUM
FOR MATHEMATICS IN INDUSTRY

For further volumes:
http://www.springer.com/series/4650

Alistair D. Fitt
John Norbury
Hilary Ockendon
R. Eddie Wilson
Editors

Progress in Industrial Mathematics at ECMI 2008

With 360 Figures, 109 in color and 46 Tables

Springer

Editors
Prof. Alistair D. Fitt
University of Southampton
School of Mathematics
Southampton SO17 1BJ
United Kingdom
A.D.Fitt@maths.soton.ac.uk

Dr. Hilary Ockendon
Dr. John Norbury
University of Oxford
Mathematical Institute
St. Giles 24-29
Oxford OX1 3LB
United Kingdom
ockendon@maths.ox.ac.uk
john.norbury@lincoln.ox.ac.uk

Dr. R. Eddie Wilson
University of Bristol
Bristol Centre for Applied Nonlinear
Mathematics
Bristol BS8 1TR
United Kingdom
Re.Wilson@bristol.ac.uk

ISBN 978-3-642-12109-8 e-ISBN 978-3-642-12110-4
DOI 10.1007/978-3-642-12110-4
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2010930279

Mathematics Subject Classification (2000): 35, 60, 62, 65, 70, 74, 76, 80, 82, 91, 92, 97

Cover design: deblik, Berlin

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

The 15th European Conference on Mathematics for Industry was held in the agreeable surroundings of University College London, just 5 minutes walk from the British Museum in the heart of London, over the five warm, sunny days from 30 June to 4 July 2008. Participants from all over the world met with the common aim of reinforcing the role of mathematics as an overarching resource for industry and business.

The conference attracted over 300 participants from 30 countries, most of them participating with either a contributed talk, a minisymposium presentation or a plenary lecture. 'Mathematics in Industry' was interpreted in its widest sense as can be seen from the range of applications and techniques described in this volume. We mention just two examples. The Alan Tayler Lecture was given by Mario Primicerio on a problem arising from moving oil through pipelines when temperature variations affect the shearing properties of wax and thus modify the flow. The Wacker Prize winner, Master's student Lauri Harhanen from the Helsinki University of Technology, showed how a novel piece of mathematics allowed new software to capture real-time images of teeth from the data supplied by present day dental machinery (see ECMI Newsletter 44).

The meeting was attended by leading figures from government, business and science who all shared the same aim – to promote the application of innovative mathematics to industry, and identify industrial sectors that offer the most exciting opportunities for mathematicians to provide new insight and new ideas. The finance day in the Lloyd's building provided an alternative venue and different talk themes. The panel discussions and the conference dinner generated formal and informal interaction and wide ranging discussions.

The organizing committee is grateful to all those who helped to make the meeting so successful. We thank Professor Frank Smith and University College London for the provision of the venue, Lord Hunt and Arren Ariel of the Lighthill Institute of Mathematical Sciences, and David Youdan and Amy Marsh of the Institute of Mathematics and its Applications for all their

effort in organizing the smooth running of the conference. We are very grateful to all our sponsors for their financial support (see: `www.ecmi2008.org`), and, in particular, Dr Robert Leese and the KTN for Industrial Mathematics for their help with the design work. The editors Alistair Fitt, John Norbury, Hilary Ockendon and Eddie Wilson thank Anthony Lock for his invaluable help with the publishing of these Proceedings.

Finally, a big thank you to all our participants who share the ECMI vision of using mathematics to make a better future in Europe – we hope this publication will help in the process of achieving this goal.

Oxford,
April 2009

John Norbury (Chair)
On behalf of the Organizing Committee

Contents

Part I

Plenary Lectures

Modelling Living Tissues: Mechanical and Mechanobiological Aspects

M. Doblaré[1,2] and J.M. García-Aznar[1,2]

[1] Aragón Institute of Engineering Research (I3A), University of Zaragoza, Zaragoza, Spain, mdoblare@unizar.es
[2] Centro de Investigacíon Biomedica en Red en Bioingeniería, Biomateriales y Nanomedicina (CIBER-BBN), Zaragoza, Spain, jmgaraz@unizar.es

Summary. Mechanical modeling of living tissues is currently one of the most crucial challenges in research for mechanical engineers and mathematicians. Mechanics is a key factor to understanding the mechanisms that regulate many biological processes, such as mitosis, migration, and differentiation. This work aims to present the most crucial aspects, in the authors' opinion, to approach this challenge.

1 Introduction

"Classical science is a conversation between theory and experiment" [1]. However, nowadays, computer simulation has been recognized as a key tool for scientific research. Some of the most useful applications of computational modelling belong to Biology [2], and specifically to modelling living tissues with a structural function, supporting and transferring loads and moving other organs [3].

In fact, Mechanics has a strong influence on many biological processes characteristics of living tissues, such as, regulation of different biological processes (homeostasis), morphological and structural adaptation or tissue damage and repair, and it is responsible directly or indirectly of many diseases such as scoliosis, osteoporosis, malaria, etc. This fact has motivated that a wide number of research works have been recently developed with the purpose of modelling the active and passive behaviour of living tissues. Modelling the functional mechanical behaviour of living tissues has historically followed two approaches: (1) considering living tissues as inert structural materials, only dealing with Mechanics and (2) considering the biological reaction of living tissues to mechanical strains/stresses and the associated changes in microstructure and therefore in the mechanical behaviour itself.

The first field corresponds to classical *Biomechanics* and applies the principles of Mechanics to predict the mechanical behaviour (movement, strains and stresses) of a tissue or an organ, taking into account the acting loads, its microstructure and the external boundary conditions. The second one,

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*,
Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_1,

known as *Mechanobiology*, tries to predict the evolution of the microstructure and biological constitution of a tissue or an organ as consequence of the mechanical environment.

In both cases, however, computational modelling presents strong difficulties that are necessary to keep in mind:

- We have to deal with very complex geometries that are sometimes evolutive. Therefore, computational geometry, medical imaging and data visualization are complementary tools.
- Most tissues involve large displacements and strains and internal material constraints which require sophisticated computational and mechanical models.
- Loads, boundary conditions and interactions are usually not known and very complex, which imply the need of accurate and complex experimental protocols to estimate them.
- Living tissues are regulated by multiple biophysical stimuli, thus, coupled fields (Multiphasic Mechanics, Biology, Chemistry) with very different time scales have to be modeled.
- Living tissues are hierarchically structurally composed materials, with their macroscopic properties depending on the different spatial scales involved. Therefore, a multiscale analysis is usually required.
- In contrast to usual engineering materials, living tissues have been optimally designed by the blind force of natural selection and show the remarkable ability to adapt not only their material properties and geometry, but also their functionality to environmental changes. Consequently, living tissues are evolving materials.
- Available experimental data present a strong variability that complicates the estimation of the parameters of the model, sometimes requiring stochastic approaches.

2 Biomechanical Tissue Modelling

Traditionally, Biomechanics in tissue modelling has been divided into two main fields of application due to the main characteristics of each tissue: hard and soft tissues.

On one hand, hard tissues typically undergo small deformations and behave nearly elastically in the range of interest. The first rigorous mathematical models for biological tissues that were introduced in the mid-1970s mainly addressed hard tissues such as bone [4].

The first modelling works of bone were elastic. For example, several authors try to model its mechanical behaviour through a mixture rule: Voigt's model [5] or Reuss's model [6]. Wagner and Weiner [7] modelled bone considering a composed material defined by its microstructure. Several authors [8, 9]

proposed experimental correlations that define the mechanical properties assuming isotropic behaviour as a function of the apparent density. However, bone is a porous and anisotropic material. Therefore, additional, correlations have been proposed including the directional influence of the microstructure through the so-called "fabric tensor" [10–14]. More recently, poroelastic models have been proposed to model the complex behaviour of bone and the interaction with the fluid that flows within its pores, lacunae and canaliculi [15–19].

On the other hand, biomechanics models for soft tissues needs a more sophisticated theory involving geometrically non-linear approaches [20, 21]. Soft tissues have a non-linear stress-strain behaviour, and many of them are viscoelastic and highly incompressible. Most models consider hyperelastic anisotropic theories with different types of strain energy density functions (polynomial, exponential, stochastically-based). Polyconvexity considerations; internal constraints (incompressibility); linear and strain-dependent viscoelasticity; residual stresses; damage; and in some case (muscular tissue) coupled electro-mechanical active behaviour are only a few of the topics addressed when dealing with the structural constitutive behaviour of soft tissues [20–25].

3 Mechanobiological Tissue Modelling

The main aim of mechanobiological models is to evaluate how a mechanical stimulus can regulate biological mechanisms, such as, remodelling, healing, etc. Therefore, these models allow improving our understanding of how tissues react to changes in the mechanical environment. In this sense, there are two main approaches: phenomenological and mechanistic.

Phenomenological models are able to predict the long-term behaviour of a biological tissue under physiological and pathological loads by establishing direct relations between external causes (mechanical stimuli) and external effects (internal microstructure or morphology) without considering the intermediate actors as they are the cells.

Mechanistic models, on the other hand, try to unravel the mechanotransduction mechanisms that regulate tissue reactions, such as: how tissues interact with cells; how cells sense strain (mechanosensing); how cells express biochemical substances after sensing strain (mechanotransduction); and how individual cells communicate with each other (signalling).

3.1 Phenomenological-Based Approaches

Phenomenological models are particularly useful to predict the adaptive tissue changes regulated by mechanical factors without information of how cells actually do it. In this sense, these models have been used to solve some important engineering problems like improving implant design [26, 27], clinical therapies evaluation [28] or tissue engineering applications [29].

3.2 Mechanistic-Based Approaches

Mechanistic models try to incorporate the effect that cells exert on the evolution of the microstructure, accounting for processes like cell proliferation, differentiation, extracellular matrix production, etc. Although very difficult to validate, they are much more bio-physical and allow checking different hypotheses and design new experiments useful for a better understanding of the specific problem analyzed. Multiphasic formulations are usually used, including complex interactions between Mechanics, cells and volume growth in the framework of open systems [30, 31].

As for as the authors know, the most general mechanistic model that considers coupled equations between multiphasic and multicellular tissue mixtures in a continuum setting has been proposed by Doblare and García-Aznar [32]. This model incorporates different and crucial factors to achieve this goal: multiple species and different types of cells; sources, sinks and diffusion of both mass and cells; possible energy transfer between the different species and cells; tissue growth, differentiation, remodeling and damage; cell proliferation, migration, differentiation and necrosis.

This formulation has been particularized to different biological processes, such as, bone tissue adaptation [33] (see Fig. 1), bone healing [34, 35] (see Fig. 2) or cell migration [36].

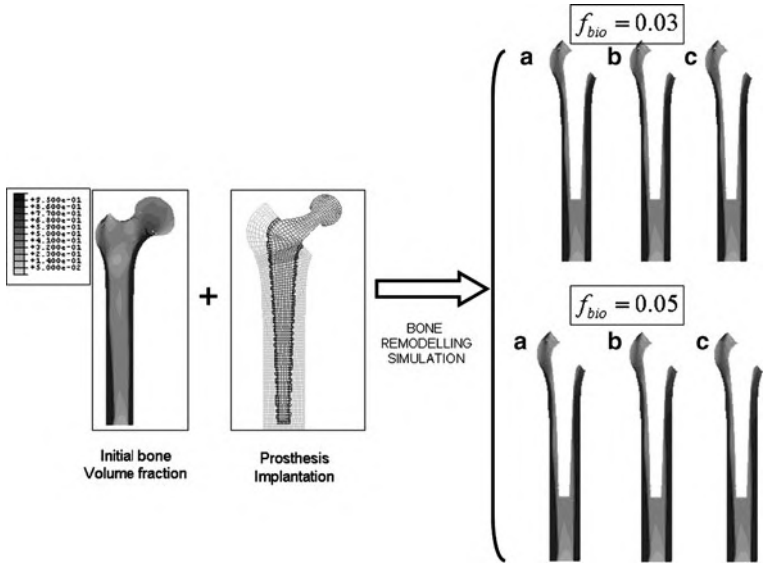


Fig. 1. Numerical simulation of the long-term adaptation process of a 2D model of a femur after implantation. Evolution of the bone volume fraction distribution for different biological factors and for different periods of time: (**a**) 330, (**b**) 660 and (**c**) 990 days [32]

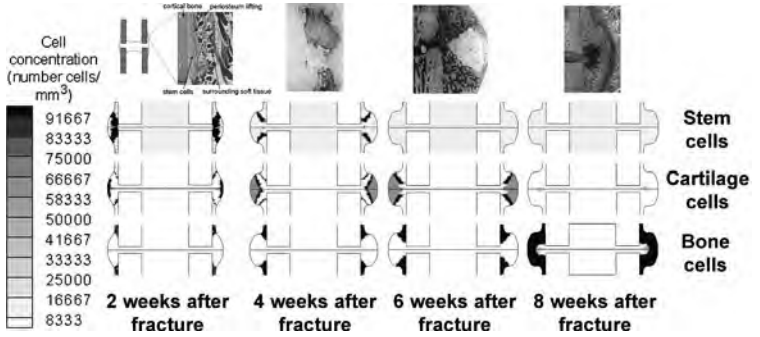


Fig. 2. Cellular distributions at different times of the healing process [32]: numerical results and histological sections (histologies taken from van der Meulen, Cornell University, NY; Sarmiento and Russell `http://www.hwbf.org/ota/bfc/index.htm`, 2002)

4 Conclusions and Further Work

Computational models including multi-scale and multi-physics approaches are a promising tool to better understand complex biophysical processes and are also essential in the growing field of quantitative and "evidence-based" Medicine. In fact, this kind of models allows exploring mechanotransduction at the cellular level and carry the information all the way up to the organ scale [4]. While the cellular scale can provide new insight into the fundamental mechanisms and help to explain signalling pathways (closer to biologists), large scale are essential to successfully address clinical and engineering problems.

Although these numerical techniques do already exists, their computational cost is still very high and the underlying biophysics is still not fully understood, so we are not able yet to fully analyze with sufficient confidence and accuracy a tissue at all the different scales incorporating all the relevant biophysical stimuli.

References

1. Kelly, K.: Science **279**(5353), 992–992 (1998)
2. Krieger, K.: Science **312**(5771), 189–190 (2006)
3. van der Meulen, M.C., Huiskes, R.: J. Biomech. **35**(4), 401–414 (2002)
4. Kuhl, E.: Comput. Methods Biomech. Biomed. Engin. **11**(5), 433–434 (2008)
5. Bonfield, W., Li, C.H.: J. Appl. Phys. **38**, 2450–2455 (1967)
6. Piekarski, K.: J. Eng. Sci. **11**, 557–565 (1973)
7. Wagner, H.D., Weiner, S.: J. Biomech. **25**(11), 1311–1320 (1992)

8. Beaupre, G.S., Orr, T.E., Carter, D.R.: J. Orthop. Res. **8**(5), 662–70 (1990)
9. Hernandez, C.J., Beaupre, G.S., Keller, T.S., Carter, D.R.: Bone **29**(1), 74–78 (2001)
10. Zysset, P.K.: J. Biomech. **36**(10), 1469–1485 (2003)
11. Zysset, P.K., Curnier, A.: J. Biomech. **29**(12), 1549–1558 (1996)
12. Doblare, M., Garcia, J.M.: J. Biomech. **35**(1), 1–17 (2002)
13. Turner, C.H., Cowin, S.C., Rho, J.Y., Rice, J.C.: J. Biomech. **23**(6), 549–561 (1990)
14. Zysset, G., Kabel, J., van Rietbergen, B., Odgaard, A., Huiskes, R., Curnier, A.: J. Elast. **53**(2), 125–146 (1998–1999)
15. Manfredini, P., Cocchetti, G., Maier, G., Redaelli, A., Montevecchi, F.M.: J. Biomech. **32**(2), 135–144 (1999)
16. Wang, L., Fritton, S.P., Cowin, S.C., Weinbaum, S.: J. Biomech. **32**(7), 663–672 (1999)
17. Smit, T.H., Burger, E.H., Huyghe, J.M.: J. Bone. Miner. Res. **17**(11), 2021–2029 (2002)
18. Burger, E.H., Klein-Nulend, J., Smit, T.H.: J. Biomech. **36**(10), 1453–1459 (2003)
19. Fornells, P., García-Aznar, J.M., Doblaré, M.: Ann. Biomed. Eng. **35**(10), 1687–1698 (2007)
20. Alastrue, V., Pea, E., Martinez, M.A., Doblaré, M.: Ann. Biomed. Eng. **35**(10), 1821–1837 (2007)
21. Peña, E., Perez del Palomar, A., Calvo, B., Martinez, M.A., Doblaré, M.: Arch. Comput. Methods Eng. **14**(1), 47–91 (2007)
22. Zamir, E.A., Taber, L.A.:J. Biomech. Eng. **126**, 276–283 (2004)
23. Taber, L.A., Humphrey, J.D.: J. Biomech. Eng. **123**, 528–535 (2001)
24. Holzapfel, G.A., Gasser, T.C.: Comput. Meth. App. Mech. Eng. **190**, 4379–4430 (2001)
25. Huyghe, J.M.R.J., Molenaar, M.M., Baaijens, F.P.T.: J. Biomech. Eng. **129**, 776–785 (2007)
26. Huiskes, R., Weinans, H., Grootenboer, H.J., Dalstra, M., Fudala, B., Slooff, T.J.: J. Biomech. **20**(11–12), 1135–1150 (1987)
27. García, J.M., Doblare, M., Cegoñino, J.: Comput. Mater. Sci. **25**, 100–114 (2002)
28. Prendergast, P.J.: Clin. Biomech. **12**(6), 343–366 (1997)
29. Sanz-Herrera, J.A., García-Aznar, J.M., Doblaré, M.: Biomech. Model. Mechanobiol. **7**(5), 355–366 (2008)
30. Lubarda, V.A., Hoger, A.: Int. J. Solids Struct. **39**, 4627–4664 (2002)
31. Garikipati, K., Arruda, E.M., Grosh, K., Narayanan, H., Calve, S.: J. Mech. Phys. Solids. **52**(7), 1595–1625 (2004)
32. Doblaré, M., García-Aznar, J.M.: Arch. Comput. Methods Eng. **13**(4), 471–513 (2006)
33. García-Aznar, J.M., Rueberg, T., Doblaré, M.: Biomech. Model. Mechanobiol. **4**(2–3), 147–167 (2005)
34. Gomez-Benito, M.J., García-Aznar, J.M., Kuiper, J.H., Doblaré, M.: J. Theor. Biol. **235**(1), 105–119 (2005)
35. García-Aznar, J.M., Kuiper, J.H., Gomez-Benito, M.J., Doblaré, M., Richardson, J.B.: J. Biomech. **40**(7), 1467–1476 (2006)
36. Moreo, P., García-Aznar, J.M., Doblaré, M.: Acta Biomater. **4**(3), 613–621 (2007)

New Mathematical Approaches for Image Reconstruction in the Oil and Medical Industries

M. Moscoso

Gregorio Millán Institute, Universidad Carlos III de Madrid, Leganés 28911, Spain, moscoso@math.uc3m.es

Summary. The problem of reconstructing images from measurements at the boundary of a domain belongs to the class of inverse problems. Although in different applications the techniques used to create the images work under different physical principles and map different physical parameters, they all share similar mathematical foundations. I will present here two mathematical approaches for image reconstruction. The first one is used to solve the so called history matching problem in the oil industry, and the second one is specially designed for the application of optical molecular imaging in biomedicine.

1 Introduction

Imaging is a broad field which covers all aspects of the analysis, modification, compression, visualization, and generation of images. There are at least two major areas in imaging science in which applied mathematics has a strong impact: image processing, and image reconstruction. In image processing the input is a (digital) image such as a photograph, while in image reconstruction the input is the data gathered on the boundary of an object. In the latter case, the data is limited, and its poor information content is not enough to generate an image to start with.

Image processing techniques apply numerical algorithms to either improve a given image or to extract information about the image [1]. Image segmentation is typically used for the latter purpose. It refers to the process of partitioning an image into multiple regions (locating objects and boundaries) in order to simplify its representation for its further analysis. Each region shares the same properties or characteristics such as color, intensity or texture. Different techniques have been applied for image segmentation. We mention here, graph partitioning methods in which the image is modelled as a graph; level-set methods in which an initial shape is evolved towards the object boundary; and statistical methods in which we view a region of an image as one realization of a random process (probability distribution functions and

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*, Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_2,

histograms are used to estimate the characteristics of the regions). We will not discuss the mathematics of image processing here.

On the other hand, image reconstruction refers to the techniques used to create an image of the interior of a body from data collected on its boundary [2]. Mathematically, an image reconstruction can be seen as the solution of an inverse problem in which the cause is inferred from the effect. We will show here two different applications in the oil and medical industries. The first application is the so called history matching problem where we want to estimate the unknown properties of a reservoir, such as its porosity and permeability, from the production data. We will apply to this problem an adjoint technique. The second application is the inverse fluorescent source problem in optical molecular imaging. We obtain here explicit solutions for a point source and a voxel source from which we estimate the location, size and total strength of a general source.

The goal of this paper is to illustrate the role of the imaging techniques in these applications. In the oil industry, for example, they improve our ability to design a good management strategy to increase the productivity and life of a reservoir. They help to better understand the reservoir behavior so that its performance can be predicted and controlled with higher reliability. On the other hand, in the biomedical application of optical molecular imaging they are used to monitor cellular and structural changes associated with predisease states such as dysplastic progression.

2 Reservoir Characterization

Oil fields typically extend over large areas, possibly several hundred kilometers across and full exploitation entails multiple wells scattered across the area. Initially, the natural differential pressure displaces hydrocarbons from the reservoir, into the wellbore and up to the surface. This is the primary recovery stage. As oil production takes place, the reservoir pressure declines, and eventually, the primary recovery stage reaches its limit. Typically, only a small fraction, around the 15% of the initial oil in place is produced during the primary recovery stage. During the second stage, water is injected into the production zone to sweep the oil from the reservoir. The secondary recovery stage reaches its limit when the injected fluid is produced in considerable amounts at the production wells and the production is no longer economical. Around 40% of the field's oil is produced during this stage. The third stage of oil production uses sophisticated techniques that alter the original properties of the oil. Its purpose is to improve oil displacement or fluid flow in the reservoir. It allows another 10% of the filed's oil to be recovered.

We consider here the case of 'secondary recovery' where water is injected through several injection wells conveniently located in order to enhance oil production. Potential problems associated with waterflood techniques include

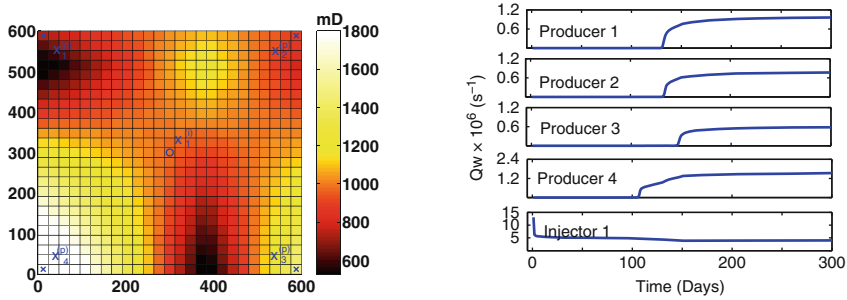


Fig. 1. (**a**) Permeability distribution: 5-spot example; (**b**) extracted water flows at the producers (shown in figure (**a**) with x's) and injected water flow at the injector (shown in figure (**a**) with a *circle*)

inefficient recovery due to the unknown variable permeability, that is due to the impact of the unknown geological heterogeneity on flow during oil recovery. Therefore, there is a need to estimate the permeability distribution inside the reservoir in order to optimize oil production. Proper characterization of the reservoir heterogeneity helps to better understand the reservoir behavior so its performance can be predicted and controlled with higher reliability.

2.1 The Direct Problem: Governing Equations

Secondary oil recovery techniques involve the simultaneous flow of up to three fluid phases. It requires the solution of the equations of a multiphase flow in a porous medium. We will consider here only water and oil, and we will neglect gas. We will also neglect the effects of gravity and capillary pressure. For describing the flow dynamics in the reservoir $\Omega \subset \mathbb{R}^n$ (n $= 2, 3$), we use a simplified Black-Oil model [3]:

$$-\nabla \cdot \Big[T \nabla p\Big] = Q, \qquad \text{in} \quad \Omega \times [0, t_f], \tag{1}$$

$$\phi \frac{\partial S_w}{\partial t} - \nabla \cdot \big[T_w \nabla p\big] = Q_w, \qquad \text{in} \quad \Omega \times [0, t_f], \tag{2}$$

where $p(\mathbf{x}, t)$ and $S_w(\mathbf{x}, t)$ are the unknowns of the problem which represent the pressure and the water saturation at position $\mathbf{x}$ and time t, respectively. The water saturation S_w measures the volume fraction of water. $\phi(\mathbf{x})$ is the porosity, and T and T_w are the transmissibilities, which are known functions which depend linearly on the permeability K, the parameter to be reconstructed, and nonlinearly on S_w,

$$T_w = K(\mathbf{x}) \frac{K_{rw}(S_w)}{\mu_w}, \quad T_o = K(\mathbf{x}) \frac{K_{ro}(S_w)}{\mu_o}, \quad T = T_w + T_o. \tag{3}$$

$K_{rw}(S_w)$, $K_{ro}(S_w)$, μ_w and μ_o denote the relative permeabilities and the viscosities of each phase, respectively. Hereafter, the subscript 'w' stands for 'water', while the subscript 'o' stands for 'oil'. $Q(\mathbf{x},t)$ and $Q_w(\mathbf{x},t)$ define the total flow and the water flow at the wells, respectively. They are given by

$$Q = cT\sum_{j=1}^{N_i}(p_{wb_j}^{(i)} - p)\delta(\mathbf{x}-\mathbf{x}_j^{(i)}) + cT\sum_{j=1}^{N_p}(p_{wb_j}^{(p)} - p)\delta(\mathbf{x}-\mathbf{x}_j^{(p)}), \tag{4}$$

$$Q_w = cT\sum_{j=1}^{N_i}(p_{wb_j}^{(i)} - p)\delta(\mathbf{x}-\mathbf{x}_j^{(i)}) + cT_w\sum_{j=1}^{N_p}(p_{wb_j}^{(p)} - p)\delta(\mathbf{x}-\mathbf{x}_j^{(p)}), \tag{5}$$

where $\mathbf{x}_j^{(i)}$, $j=1,\ldots,N_i$, denote the locations of the N_i injector wells, $\mathbf{x}_j^{(p)}$, $j=1,\ldots,N_p$, denote the locations of the N_p production wells, and $p_{wb_j}^{(i)}$, $p_{wb_j}^{(p)}$ are the imposed well bore pressures at the N_i injector wells and at the N_p production wells, respectively. Here, c is a constant that depends on the well model. Since $p_{wb_j}^{(i)}$ ($p_{wb_j}^{(p)}$) are larger (smaller) than the reservoir pressure at the injector (production) wells, Q and Q_w are positive (negative) at the injector (production) wells.

Equation (2) is the conservation law for water in a porous medium and (1) is obtained by combining the conservation laws for water and oil in order to eliminate the time derivative term. It is assumed that the flow obeys Darcy's law ($\mathbf{u}_l(\mathbf{x},t) = -\frac{K(\mathbf{x})K_{rl}(S_w)}{\mu_l}\nabla p(\mathbf{x},t), l=w,o$) which defines the velocity of each phase in the medium. Equations (1) and (2) are solved with the following initial and boundary conditions:

$$S_w(\mathbf{x},0) = S_w^0(\mathbf{x}) \qquad \text{in} \quad \Omega, \tag{6}$$

$$p(\mathbf{x},0) = p^0(\mathbf{x}) \qquad \text{in} \quad \Omega, \tag{7}$$

$$\nabla p\cdot\boldsymbol{\nu} = 0 \qquad \text{on} \quad \partial\Omega, \tag{8}$$

where $\boldsymbol{\nu}$ is the outward unit normal to $\partial\Omega$. The boundary condition (8) implies no flux across the boundary.

Equations (1)–(8) define the *direct problem* for the dynamic production history at the extraction wells. It refers to the resolution of the equations describing the flow within the reservoir assuming that the properties of the porous media, defined by $K(\mathbf{x})$ and $\phi(\mathbf{x})$, are known. The properties of the fluids are defined by μ_w, μ_o, $K_{rw}(S_w)$, and $K_{ro}(S_w)$. The well bore pressures $p_{wb_j}^{(i,p)}$ are known functions of time at the well's positions.

The left side of Fig. 1 shows a five-spot layout with an injector well (o) in the center (location $\mathbf{x}_1^{(i)}$) and four production wells (x) at the corners of a two-dimensional reservoir (locations $\mathbf{x}_j^{(p)}$, $j=1\ldots4$, being $j=1$ the well in the upper left corner and numbered in the clockwise direction). Also shown is the real permeability distribution in milli-Darcys (mD). The water injected at the injection well displaces the oil in the reservoir towards the production

wells. Time resolution of the flow equations provides the time evolution of pressure and flow at each point of the reservoir. Of particular interest is the oil and water flow rate at each production well. The right hand side of Fig. 1 shows the time history of water flow rate (Q_w) at each well obtained by solving the direct problem. Notice that water arrival occurs first at well four since it is surrounded by a region of high permeability.

2.2 The Inverse Problem: The Adjoint Method

In the *inverse problem* we assume that the water flow rate at each well is known but the permeability distribution is unknown. Hence, the unknown of interest in the inverse problem is the permeability K that we want to estimate from the water production rates. We will start with an initial permeability guess (typically some constant distribution) and will iteratively modify it until the actual water production rate at each well is matched by the simulator.

Adjoint techniques are particularly useful in large scale inverse problems where relatively few independent experiments can be performed for gathering data but many parameters need to be reconstructed. Since typically only one experiment is performed in history matching due to the simultaneous production process, the adjoint technique is therefore much faster in this application. Adjoint techniques have also been applied with great success in other applications of geophysical and medical imaging (see [2], and references therein). Other techniques have also been applied to the history matching problem. Among them, we mention shape-based reconstructions that use level-set techniques [4,5].

The forward operator described in the previous section can be written in abstract form as

$$M : P \longrightarrow D, \qquad M[K] = Q_w[K]|_{\Omega_+ \times [0,t_f]}, \tag{9}$$

where Q_w is obtained by solving the direct problem for a given permeability distribution K (1)–(8). Here, we denote the space of permeability distributions K by P, the data space by D, and the set of measurement locations ('well-locations') by $\Omega_+ := \{\mathbf{x}_1^{(p)}, \mathbf{x}_2^{(p)}, \ldots, \mathbf{x}_{N_p}^{(p)}\}$. At each of these positions, the water flow is measured during a time $0 \leq t \leq t_f$, such that the data space D is given by $D = (L_2([0, t_f]))^{N_p}$.

For some guess K of the permeability, and given the measured data $\tilde{G}$ (water flow rate) at these production wells, we also define the residual operator

$$R[K] = M[K] - \tilde{G}. \tag{10}$$

Equation (10) describes the mismatch between the physically measured data and the data corresponding to a guess K. In the inverse problem, we ideally want to find a permeability distribution $\hat{K}$ in P such that

$$R[\hat{K}] = 0. \tag{11}$$

This equation has a solution in the situation where the data $\tilde{G}$ are in the range of M. Most likely this is not the case if we use real data, so we generalize our criterion for a solution defining the least squares cost functional

$$\mathcal{J}(K) = \frac{1}{2}\|R(K)\|^2, \tag{12}$$

and searching for a minimizer of this cost functional. This cost functional defines the differences between the predicted model (as described by K) and the actual observed measurements in an L_2 norm sense.

In order to find an 'update' δK for our permeability K we linearize (in a Newton-type fashion) the nonlinear operator R (assuming that this linearized operator $R'[K]$ exists and is well-defined) and write

$$R[K + \delta K] = R[K] + R'[K]\delta K + 0(||\delta K||^2), \tag{13}$$

where the linearized operator $R'[K]$ represents the Frechet derivative of R at K, which is closely related to the 'sensitivity functions' of the parameter profile with respect to the data. Using (13) we want to look for a correction δK such that $R[K + \delta K] = 0$. Neglecting terms of order $0(||\delta K||^2)$ in (13), this amounts to solving

$$R'[K]\delta K = -R[K]. \tag{14}$$

A classical solution to the ill-posed linear inverse problem (14) is the minimum-norm solution

$$\delta K_{MN} = -R'[K]^* \left(R'[K]R'[K]^*\right)^{-1} R[K], \tag{15}$$

where $R'[K]^*$ represents the adjoint operator of $R'[K]$. In our application, the operator $C = (R'[K]R'[K]^*)^{-1}$ is ill-conditioned and expensive to calculate, so it will be replaced by the identity operator I (note that C 'just' maps vectors from the data space back into the data space, so it can be considered as a 'filtering operator'). Therefore, we end up with simply applying the adjoint operator $R'[K]^*$ to the residuals R for calculating the update direction

$$\delta K = -R'[K]^* R[K]. \tag{16}$$

Note that the operator $R'[K]^*$ maps the residuals back into the parameter space for obtaining the update. Therefore, in order to determine δK in each step from (16), we will need an efficient method for applying $R'[K]^*$ to a given vector ρ of the data space. Next, we show how to compute it (the details can be found in [6]).

Let us consider a small perturbation δK in the permeability distribution K that leads to small perturbations W and q in the saturation and the pressure, respectively. Here we assume that the pressure remains nearly unchanged so that ∇q is negligible. This is so because the pressure is a smooth function compared to the saturation. Using a heuristic approach to derive an expression

for R', we introduce $K + \delta K$ and $S_w + W$ in (2) and we neglect second order terms. Then, W solves the initial value problem

$$\phi \frac{\partial W}{\partial t} - \nabla \cdot \big[\frac{\partial T_w}{\partial S_w} W \nabla p\big] - \frac{\partial Q_w}{\partial S_w} W = \frac{\delta K}{K} Q_w + \nabla \cdot \big[\frac{\delta K}{K} T_w \nabla p\big] \quad \text{in } \Omega \tag{17}$$

$$W(\mathbf{x}, 0) = 0 \qquad \text{in } \Omega \tag{18}$$

where S_w and p are the solutions of (1)–(8). From the value of W we derive the linearized response of the data to a perturbation δK in the permeability distribution, which is given by

$$R'[K]\delta K = \frac{\partial Q_w}{\partial S_w} W \Big|_{\Omega_+ \times [0, t_f]} . \tag{19}$$

The adjoint operator $R'[K]^*$ is defined by

$$\langle R'[K]\delta K, \rho \rangle_D = \langle \delta K, R'[K]^* \rho \rangle_P , \tag{20}$$

where $\langle\ ,\ \rangle_D$ and $\langle\ ,\ \rangle_P$ denote the inner products in the data and parameter spaces, respectively. We assume that the inner products in the parameter space P and in the data space D are given by

$$\langle f, g \rangle_D = \sum_{j=1}^{N_p} \int_0^{t_f} f_j\ g_j\ dt\ ; \quad \langle A, B \rangle_P = \int_\Omega A\ B\ d\mathbf{x}\ , \tag{21}$$

where $f_j = f(\mathbf{x}_{p_j}, t)$ and $g_j = g(\mathbf{x}_{p_j}, t)$, $j = 1, \ldots, N_p$, are time functions defined at the production well positions $\mathbf{x}_{p_j}$. The following adjoint form of the linearized residual operator has been derived in [6].

Let $\rho \in D$ be an arbitrary function in the data space. Then $R'[K]^*\rho$ is given by

$$R'[K]^* \rho = \int_0^{t_f} \frac{T_w}{K} \nabla p \nabla z\ dt \tag{22}$$

where z is the solution of the adjoint equation

$$-\phi \frac{\partial z}{\partial t} + \frac{\partial T_w}{\partial S_w} \nabla p \nabla z - (z - \sum_{j=1}^{N_p} \rho\ \delta(\mathbf{x} - \mathbf{x}_j^{(p)})) \frac{\partial Q_w}{\partial S_w} = 0 \quad \textit{in } \Omega \tag{23}$$

$$z(\mathbf{x}, t_f) = 0 \quad \textit{in } \Omega, \tag{24}$$

and S_w and p are the solutions of (1)–(8).

Note that, as typical for the adjoint scheme, the system (23)–(24) physically models some kind of *backpropagation* with respect to the linearized forward model. The residuals are applied at the production wells as artificial injectors, and propagated backward in time (notice the minus sign in front of the time derivative in (23) and the prescribed final value conditions in (24), compared to a plus sign in (17) and initial values in (18)) and in space by the

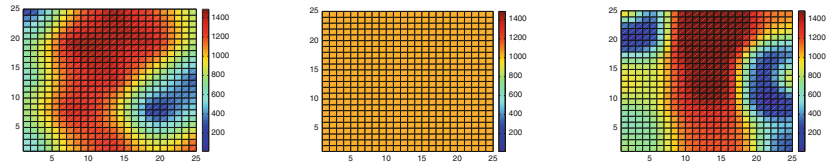


Fig. 2. Example of a reconstructed permeability distribution. *Left image*: reference profile. *Center image*: initial profile. *Right image*: reconstructed profile

system (23)–(24). Equation (22) uses these backpropagated fields to extract an update direction by combining forward and adjoint fields at each location.

In Fig. 2 we show an example of a reconstruction applying the adjoint method. The reference permeability distribution is shown in the left image. The well configuration used in this example is the same shown in Fig. 1. There is one injector well at the center of $600 \times 600\,\mathrm{m}^2$ reservoir and four producer wells at the corners. The reservoir is discretized by a 25×25 uniform spatial grid. Our initial model, shown in the center image, consists of a uniform permeability distribution of 1,040 mD. The estimated permeability distribution at the end of the 20th iteration is shown in the right image. We can observe a very good agreement between the reference and estimated permeability distributions.

3 Optical Molecular Imaging

Optical molecular imaging is showing great promise for monitoring several cellular and structural changes associated with predisease states such as dysplastic progression [7–10]. In this application, near-infrared fluorescent probes are used to mark specific cellular targets within the tissue that re-emit light upon excitation by an external light source. These markers act as internal sources that can be imaged from measurements of the light intensity at the tissue surface. The goal of determining the internal fluorescent source distribution from boundary measurements can be stated as an inverse source problem.

Several challenges arise in optical molecular imaging due to the multiple scattering of light in tissues. Physically, multiple scattering causes severe image blurring and, therefore, one cannot make use of direct images. Rather, one must develop methods to reconstruct images from scattered light measurements.

To model diagnostic measurements we use the radiative transport equation. It describes accurately light propagation in tissues [11–13]. To study the inverse fluorescent source problem we use the integral formulation of

this equation in three dimensions. Using the point source and voxel source solutions, we estimate the location, size and total strength of a general source [14].

3.1 The Direct Problem: The Radiative Transfer Theory

Modeling fluorescence in tissues must account for the following stages: (1) propagation of excitation light from the tissue's surface into its interior, (2) absorption by fluorophores, (3) conversion to fluorescence, (4) emission of the fluorescent light from the fluorophores, and (5) propagation of that light back up to the tissue surface. We assume here continuous illumination and that the absorption and emission spectra of the fluorescent molecules do not overlap. Hence, the excitation and emission processes take place at different wavelengths denoted by λ_x and $\lambda_m > \lambda_x$, respectively. Accordingly, the forward model describing the transport of excitation and emission light can be written as:

$$\mathbf{\Omega} \cdot \nabla I_x + (\mu_a^x + \mu_a^{x \to m}) I_x - \mu_s^x L I_x = 0, \tag{25}$$

$$\mathbf{\Omega} \cdot \nabla I_m + \mu_a^m I_m - \mu_s^m L I_m = S_{x \to m}. \tag{26}$$

In these equations, I_x (I_m) is the specific intensity for the exciting (emission) light at wavelength λ_x (λ_m). They depend on direction $\mathbf{\Omega} \in \mathbb{S}^2$ ($\mathbb{S}^2$ denotes the unit sphere) and position $\mathbf{r} \in \mathbb{R}^3$. At the excited (emission) wavelength λ_x (λ_m), the absorption and scattering coefficients are denoted by μ_a^x and μ_s^x (μ_a^m and μ_s^m), respectively. The absorption by fluorophores in (25) is given by the fluorophore absorption coefficient, $\mu_a^{x \to m}$. The isotropic source term

$$S_{x \to m}(\mathbf{r}) = \eta U_x(\mathbf{r}) \mu_a^{x \to m}(\mathbf{r}) \tag{27}$$

is the product of the quantum efficiency η of the fluorophore, the average excited intensity

$$U_x(\mathbf{r}) = \frac{1}{4\pi} \int_{\mathbb{S}^2} I_x(\mathbf{\Omega}, \mathbf{r}) \, \mathrm{d}\, \mathbf{\Omega}, \tag{28}$$

and the fluorophore absorption coefficient $\mu_a^{x \to m}$. This average excited intensity excites the fluorophore molecules from their ground state to an excited state. The quantum efficiency η quantifies the conversion to fluorescence.

The scattering operations $LI_{x,m}$ in (25) and (26) are defined as

$$LI_{x,m}(\mathbf{\Omega}, \mathbf{r}) = -I_{x,m}(\mathbf{\Omega}, \mathbf{r}) + \int_{\mathbb{S}^2} f_{x,m}(\mathbf{\Omega} \cdot \mathbf{\Omega}') I_{x,m}(\mathbf{\Omega}', \mathbf{r}) \, \mathrm{d}\, \mathbf{\Omega}'. \tag{29}$$

The scattering phase functions $f_{x,m}$ in (29) give the fraction of light scattered in direction $\mathbf{\Omega}$ due to light incident in direction $\mathbf{\Omega}'$ at wavelengths $\lambda_{x,m}$, respectively.

3.2 The Inverse Problem: A Semi-Analytical Method

The objective in optical molecular imaging is to reconstruct the fluorescent source $S_{x\to m}(\mathbf{r})$ in the domain D using measured data taken from the boundary surface ∂D. Since the coupling between (25) and (26) is only through the source term in (26), and the goal of our inverse problem is precisely to reconstruct it, we can consider only the second equation (26).

Because the only source of light in this problem is the fluorescent source, we prescribe boundary conditions of the form:

$$I_m(\mathbf{\Omega}, \boldsymbol{\rho}) = 0, \quad \mathbf{\Omega} \cdot \mathbf{n}(\boldsymbol{\rho}) > 0, \quad \boldsymbol{\rho} \in \partial D, \tag{30}$$

with $\mathbf{n}(\boldsymbol{\rho})$ denoting the *inward* normal at $\boldsymbol{\rho} \in \partial D$. Moreover, we impose that I_m is bounded everywhere in the halfspace. Our measured data R is the specific intensity at the boundary for all directions pointing out of D:

$$R(\mathbf{\Omega}, \boldsymbol{\rho}) = I_m(\mathbf{\Omega}, \boldsymbol{\rho}), \quad \mathbf{\Omega} \cdot \mathbf{n}(\boldsymbol{\rho}) < 0, \quad \boldsymbol{\rho} \in \partial D. \tag{31}$$

For the inverse fluorescent source problem, we wish to reconstruct the fluorescent source $S(\mathbf{r})$ with the measured data given by (31) and the direct problem given by (26) subject to (30). Here, we focus on the case of planar fluorescent reflectance imaging. The domain is modeled as a halfspace $D = \{z > 0\}$ bounded by the plane $\partial D = \{z = 0\}$. The halfspace is composed of a uniform absorbing and scattering medium. The constant absorption and scattering coefficients, denoted by μ_a and μ_s, respectively, are assumed to be known. The scattering phase function f is also assumed to be known.

In the method introduced in [14] we use the Green's function for the radiative transport equation and the general representation formula to find key properties of a fluorescent source such as its location an size. The Green's function is computed analytically as an expansion of plane wave solutions. The plane wave solutions are computed numerically. Using the Green's function for the radiative transport equation, we represent the measured data R as the superposition of interior sources and surface sources. With this representation, we can subtract off contributions from surface sources explicitly from the measured angular data yielding a quantity that depends only on the interior source of interest. Finally, we derive closed-form analytical solutions to recover a point source and a voxel source. For more details we refer the reader to [14].

We point out that the analysis in [14] relies on full angular measurements at the tissue boundary. However, one does not always have access to this data in general, but only to that given by the limited angular aperture of the detector. An extension to this theory to treat limited angular data can be found in [15].

In Fig. 3 we show the performance of our approach. We show results in which we estimate the location and size of general a fluorescent source

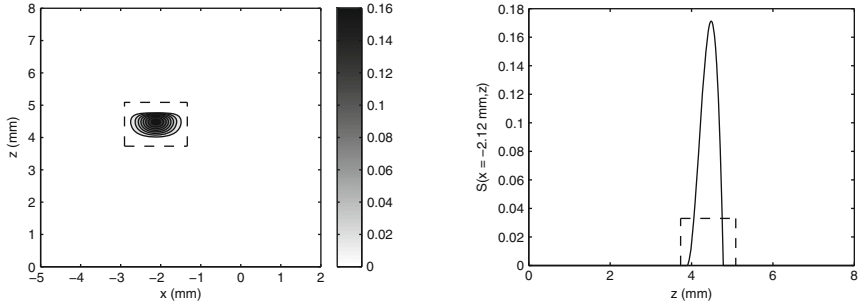


Fig. 3. Example of the estimation of the location and size of a general source. *Left plot*: contour of the true source and outline of the recovered pixel source. *Right plot*: slice of the source at $x = -2.1$ mm and the recovered pixel source. (From Kim and Moscoso [14])

$$S(x,z) = S_0\, e^{-(x-x_0)^2/w^2} \times \begin{cases} -(z-a)^2(z-b) & z \in [a,b], \\ 0 & \text{otherwise}, \end{cases} \tag{32}$$

by recovering the parameters of a pixel source. We consider a halfspace $z > 0$ composed of a uniform absorbing and scattering medium. The absorption and scattering coefficients are $\mu_a = 0.01\,\text{mm}^{-1}$, and $\mu_s = 1.0\,\text{mm}^{-1}$, respectively. We used the Henyey–Greenstein scattering phase function with asymmetry parameter $g = 0.8$. Hence, $\mu_s(1-g)/\mu_a = 20$ which is in the range of optical properties of tissues and tissue phantoms. The numerical simulations were computed in two dimensions: x and z.

The parameters used in (32) for the fluorescent source are: $S_0 = 1.64$, $x_0 = -2.1$, $w = 0.43$, $a = 3.89$ and $b = 4.78$. With these parameter values, the total strength, defined as

$$S_{total} = \int_0^\infty \int_{-\infty}^\infty S(x,z)\, dx\, dz, \tag{33}$$

is $S_{total} \approx 0.0654$.

The left plot in Fig. 3 shows the contours of the true source and the outline of the recovered pixel source. The middle plot shows the slice of the source at $x = -2.1$ mm and the recovered pixel source. The right plot shows the source at $z = 4.335$ mm and the recovered pixel source. The pixel that we recovered has parameter values: $S_0 = 0.033$, $x_1 = -2.9044$, $x_2 = -1.3356$, $z_1 = 3.7319$ and $z_2 = 5.0862$, so that the total strength recovered is $S_{total} = 0.0701$. We observe that the pixel source captures the correct location, the size, and the total strength of the source very well. Figure 3 validates our theory.

4 Conclusions and Further Work

In this paper we wanted to stress that imaging is more than showing that an inverse problem may have a unique solution under circumstances that are rarely satisfied in practice. Modern imaging approaches deal with understanding the trade off between data size, the quality of the image, the computational complexity of the forward model used to generate the measurements, and the complexity and stability of the numerical algorithm employed to obtain the images. One neither has all the data he wants, nor can solve a very general forward model to invert the data. Finally, progress can hardly be carried out without a deep understanding of the mathematical model with which we interpret the data and without efficient and well designed numerical algorithms to solve the mathematical model.

Acknowledgments

The author thanks his coauthors Oliver Dorn, Pedro González-Rodríguez, Manuel Kindelan and Arnold Kim for our extended collaboration on imaging and other interesting problems. The author also acknowledges support from the Spanish Ministry of Education and Science (grant no FIS2007-62673) and by the Autonomous Region of Madrid (grant no S-0505/ENE/0229, COMLIMAMS).

References

1. Suri, J.S., Farag, A.: Deformable Models: Theory and Biomaterial Applications, Springer, New York (2007)
2. Natterer, F., Wübbeling, F.: Mathematical Methods in Image Reconstruction. SIAM Monographs on Mathematical Modeling and Computation. SIAM, Philadelphia (2001)
3. Thomas, G. W.: Principles of Hydrocarbon Reservoir Simulation, Prentice-Hall, New Jersey (1982)
4. Villegas, R., Dorn, O., Moscoso, M., Kindelan, M., Mustieles, F.J.: Proceedings of the Paper C015, Society of Petroleum Engineers SPE-paper 1002911 (2006)
5. Villegas, R., Dorn, O., Moscoso, M., Kindelan, M.: Progress in Industrial Mathematics at ECMI 2006, vol. 12, pp. 597–602. (2006)
6. González-Rodríguez, P., Kindelan, M., Moscoso, M., Dorn, O.: Inverse Probl. **21**, 565–590 (2005)
7. O'Leary, M.A., Boas, D.A., Li, X.D., Chance, B., Yodh, A.G.: Opt. Lett. **15**, 158–160 (1996)
8. Hawrysz, D.J., Sevick-Muraca, E.M.: Neoplasia. **2**, 388–417 (2000)
9. Ntziachristos, V., Weissleder, R.: Opt. Lett. **26**, 893–895 (2001)
10. Graves, E.E., Ripoll, J., Weissleder, R., Ntziachristos, V.: Med. Phys. **30**, 901–911 (2003)
11. Moscoso, M., Keller, J.B., Papanicolaou, G.: J. Opt. Soc. Am. A. **18**(4), 948–960 (2001)

12. Kim, A.D., Keller, J.B.: J. Opt. Soc. Am. A. **20**, 92–98 (2003)
13. Kim, A.D., Moscoso, M.: J. Biomed. Opt. **10**, 034015 (2005)
14. Kim, A.D., Moscoso, M.: Inverse Probl. **22**, 23–42 (2006)
15. González-Rodríguez, P., Kim, A.D., Moscoso, M.: J. Opt. Soc. Am. A. **24**, 3456–3466 (2007)

Continuum Models: Helping to Guide Industry

Colin Please

School of Mathematics, University of Southampton, Southampton, SO17 1BJ, UK, cpp@maths.soton.ac.uk

Summary. This is a summary of a plenary talk given at ECMI 2008 emphasizing the importance of applied mathematics to our economies, in particular the use of continuum models to give insight and hence guide industrial developments. This is illustrated using examples from industrial study groups, where collaboration between mathematicians and industrial partners has yielded great insight.

1 Industrial Mathematics

Industrial mathematics is fundamental to the knowledge base of the economy of every country, and by its very nature it is interdisciplinary. The number of different types of applications is huge, and mathematicians have a vast array of tools to apply to help understand these problems. However, as has been discussed throughout this conference, currently it is not obvious where all these industrial mathematicians are. Andreas Schuppert suggested in his plenary talk, entitled "Mathematics in Industry – cost factor or key for profits", that industry structures are now project-based rather than subject-based. To address this shift in strategy, industrial mathematicians must develop a wider skill base, so that they can identify appropriate mathematics for different situations, be it continuum models, statistics, operations research, computational models, etc. Since in-house research/development departments have all but disappeared, opportunities are greater for mathematicians to provide external consultancy services. One of the best ways to train and educate industrial mathematicians is to exploit the study group workshops [16]. This format has been exported all over the world, (ECMI, UK, Denmark, the Netherlands, Ireland, China and elsewhere), and has expanded to cover many different areas, (industry, plant science, medicine, geoscience). Within the study group format, mathematicians learn how to approach completely new problems, and how to collaborate with researchers from other disciplines.

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*, Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_3,

2 Continuum Models in Industry

Continuum models are a rather specialised field, but they have numerous applications. However many continuum models have already been captured in computer software, and we can give many examples, e.g. "finite element analysis" of deforming solids, "CFD" of airflow over wings, heat flow in furnaces, chemical reactions in flames, and pollution dispersion in rivers. As a result, the expertise of industrial mathematicians is not so important in these areas. It is however, essential for "non-standard" problems. To identify these problems is not trivial, as we cannot expect the person with the problem to understand that mathematics would be useful. We need to be proactive and search out the opportunities.

Below we present three study group problems where continuum models have proved to be very useful. The emphasis is on simple mathematics and we shall see that even simple models are able to give good insight.

2.1 Cosmetics (Collaborators: G. Pettet, R. Colasanti, J. Malda, Z. Upton)

The manufacture of cosmetics is a multi-million dollar industry and there are countless problems that need solving. It is an interesting fact that the skin is the largest organ in the body, typically $1.5\,\mathrm{m}^2$ surface area, and it is a very special sort of material. It provides many functions for the body such as protection from external sources and cooling [17]. Typical problems that the cosmetic companies are interested in include wrinkle reduction, moisturising and drug delivery (through patches or injections). They are also interested in the impact of products on the skin, such as irritation or damage caused by washing up powders.

Recent legislation has made it illegal to test detergents or cosmetics on live animals, so research is now focused on developing an artificial skin for testing. A group in Brisbane is growing artificial skin and is interacting with applied mathematicians as part of their efforts to understand how to improve their procedures. A snapshot of the skin growing in vitro is displayed in Fig. 1.

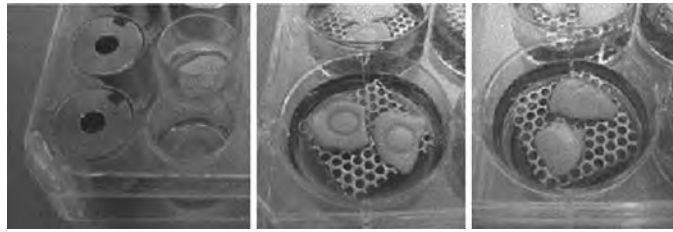


Fig. 1. Artificial skin growing in vitro. Reproduced from [1]

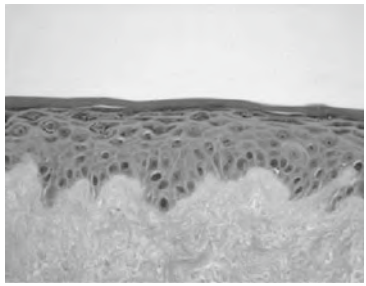


Fig. 2. Close-up photograph of artificial skin with the different layers labelled [2]

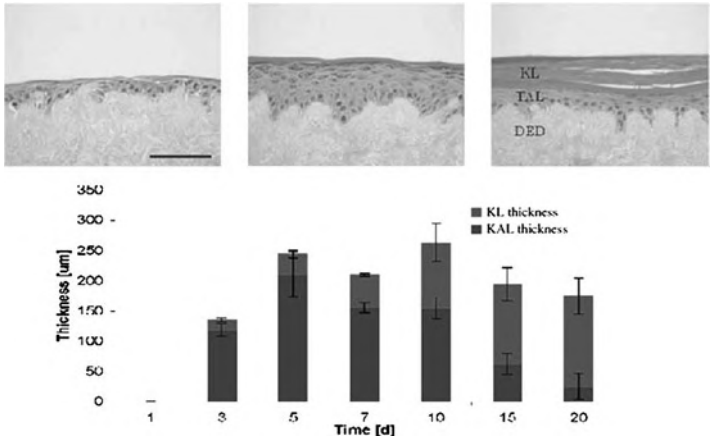


Fig. 3. Growth behaviour of human skin equivalent (HSE) [2]

Artificial skin, commonly referred to a human skin equivalent (HSE), is grown from skin taken from people. The structure of artificial skin is displayed in Fig. 2. Surplus skin is removed from the body and the top layer (the epidermis including the cornified layer (CL)) is stripped off from the dermis. The cells in the dermis are then all removed to create de-epithelised dermis (DED). A single layer of epidermal cells is then placed onto the exposed dermis and allowed to grow. The problem is to understand the resulting growth. The cells grow in the transit-amplifying layer (TAL) until they reach a certain height, where they differentiate to make the CL. In Fig. 3, we display snapshots of the different stages of growth, along with a graph showing the thicknesses of the TAL and the CL as time progresses. The first snapshot shows the TAL growing with a very thin CL on the top. As we progress to the second snapshot,

the TAL and the CL have both increased in thickness, although the CL is still much thinner than the TAL. However in the third snapshot, the TAL has thinned a lot, while the CL has thickened. The measurements displayed in the graph show how the thickness of the layers develops. This behaviour is unusual, as we would expect everything to continue thickening. The three main questions that the experiments raised were:

- Why does the TAL grow and then shrink?
- What controls the thickness of the final layers?
- Why is the interface between the TAL and the CL so flat?

To answer these questions, it is essential to understand the process of growth. Enhanced levels of nutrients cause the cells to proliferate, while high levels of calcium increase the rate of differentiation of cells from the TAL to the CL. The hypothesis suggested by the experimentalists was that the growth process depended on fluid flow. They argued that while there is no CL, fluid flow (weeping) driven by evaporation keeps calcium levels low and keeps nutrient levels high. Then as the CL recovers, the fluid flow reduces. However calculation of the Peclet number suggested that fluid flow must be irrelevant, and diffusion is more important.

An alternative theoretical approach was taken that exploited well-known observations on calcium within the epidermis. Measurements (see Fig. 4) show that there is more calcium near the CL than lower in the epidermis. It is well known that cells differentiate faster when the calcium levels are high and that differentiated cells appear to contain no calcium. Hence a model with a self-sustaining calcium gradient was developed. Calcium levels remain low initially so differentiation is slow. As the layer grows calcium at the upper surface increases in concentration as calcium is dumped into the extracellular region due to differentiation. Finally this creates the self-sustained calcium gradient.

To model the growth process the following assumptions were made:

- Cells are created/proliferate only at the lower interface between the TAL and the basement layer.
- Cells move vertically upward and are pushed from below.
- Calcium diffuses freely between cells.
- Cells take up calcium from their surroundings.
- Differentiation occurs when the concentration of calcium in the cells is sufficiently large.

We programmed these assumptions in 2D using a cellular automata program and ran simulations to visualise what happens [3]. We considered two scenarios, letting the cells start to proliferate on a flat or a bumpy layer. In both cases the cells grow upwards, the calcium concentration increases at the top and then there is a wave of retreat as the cells differentiate and throw out calcium. The process is self-sustaining.

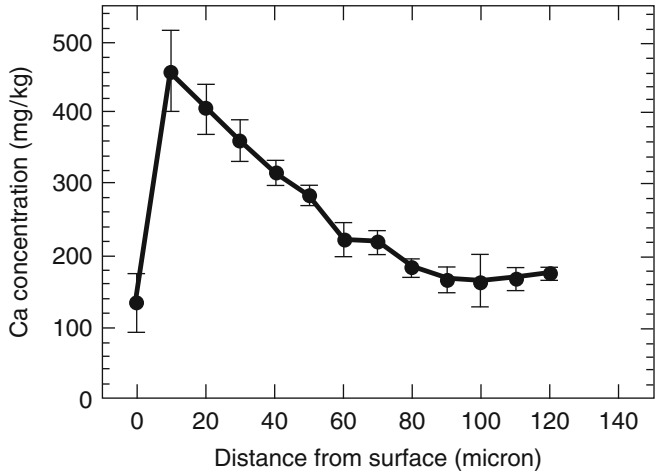


Fig. 4. Graph of calcium concentration versus distance from surface. Reproduced from [4]

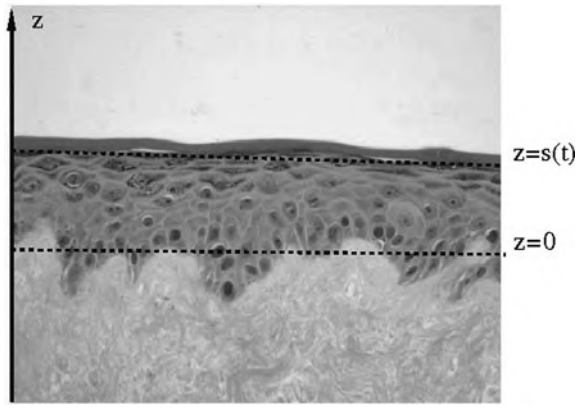


Fig. 5. A simple 1D model of the skin. The interface between the TAL and the basement layer is at $z = 0$, while the interface between the TAL and the CL is a free boundary at $z = s(t)$. Based on [2]

The cellular automata simulations produced reasonable behaviour, so we decided to write a continuum model using the same assumptions. To keep the model simple, we considered 1D, as depicted in Fig. 5. The interface between the TAL and the basement layer is at $z = 0$, while the interface between the TAL and the CL is a free boundary at $z = s(t)$. Letting $B(z, t)$ be the calcium bound in the cells and $C(z, t)$ be the freely-diffusing calcium, then the non-dimensional governing equations are

$$\left.\begin{aligned} \mu\frac{\partial^2 C}{\partial z^2} &= \mathcal{T}(C,\,B) \\ \frac{\partial B}{\partial t}+\frac{\partial B}{\partial z} &= \mathcal{T}(C,\,B) \end{aligned}\right\}, \quad 0 \le z \le s(t), \tag{1}$$

with boundary conditions

$$C = C_0, \quad B = B_0, \qquad \text{on } z = 0, \tag{2}$$

$$\left.\begin{aligned} \frac{\partial C}{\partial z} &= \alpha B\,\mathcal{D}(B) \\ 1-\frac{ds}{dt} &= \mathcal{D}(B) \end{aligned}\right\}, \quad \text{on } z = s(t), \tag{3}$$

and initial conditions

$$C = C_0, \quad B = B_0, \quad s = 0, \qquad \text{at } t = 0. \tag{4}$$

Here $\mathcal{T}(C,\,B)$ is the transfer rate between intra-cellular and extra-cellular calcium, $\mathcal{D}(B)$ is the rate of differentiation to cornified cells, and α is a constant. To progress we chose the following forms for $\mathcal{T}(C,\,B)$ and $\mathcal{D}(B)$,

$$\mathcal{T}(C,\,B) = C - B, \tag{5}$$

$$\mathcal{D}(B) = \epsilon + \lambda H(B-1), \tag{6}$$

where $H(\cdot)$ is the Heaviside function and λ and ϵ are constants. Equation (5) assumes that the exchange of calcium between cells is linearly dependent on concentrations. In formula (6) the ϵ term allows the cells to differentiate slowly all the time, while the Heaviside function means that when $B > 1$ the cells differentiate rapidly when the calcium level exceeds a critical value of $B = 1$. The parameter ϵ is extremely important. If ϵ is taken to be zero, then we obtain a trivial solution where the TAL simply grows and cornification never occurs. Taking $\epsilon \ll 1$ corresponds to growing a very thick TAL before $B = 1$ and cornification occurs. Therefore ϵ is the trigger mechanism which switches on the calcium gradient.

Examples of the behaviour of the model are shown in Figs. 6 and 7. In both cases the model reproduces the gradient of the calcium concentration, which increases towards the top of the TAL ($z = s(t)$). The graphs of layer thickness versus time show that $s(t)$ increases and then decreases, replicating the observed growth and retreat of the TAL. The model also reproduces the thickening in the CL which is shown between the curves of $s(t)$ and the outer surface $\eta(t)$. Depending on the values chosen for ϵ and λ, the thickness of the TAL may oscillate over time, as can be seen in Fig. 7. This oscillatory behaviour has yet to be seen experimentally.

To summarize, the continuum model fits the observed experimental data, although there is some data fitting due to lack of experimental evidence for

the different parameters. Future work includes extending the model to 2D or 3D to see if the free boundary between the TAL and the CL remains flat. We also need to consider more details of the flow of cells in the TAL, and we should incorporate other mechanisms to understand the trigger (cell death, potassium etc.).

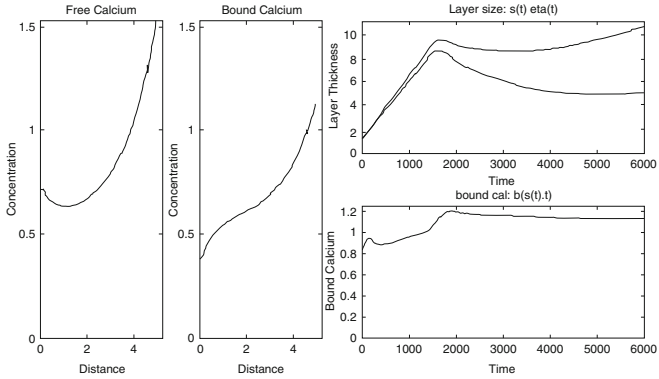


Fig. 6. Typical conditions. The graphs of free and bound calcium versus z are snapshots at time $t = 6{,}000$, where the interface $s(t) \approx 4.8$. The graph of layer thickness versus time shows the top of the TAL ($s(t)$, *lower line*) and the CL ($\eta(t)$, *upper line*)

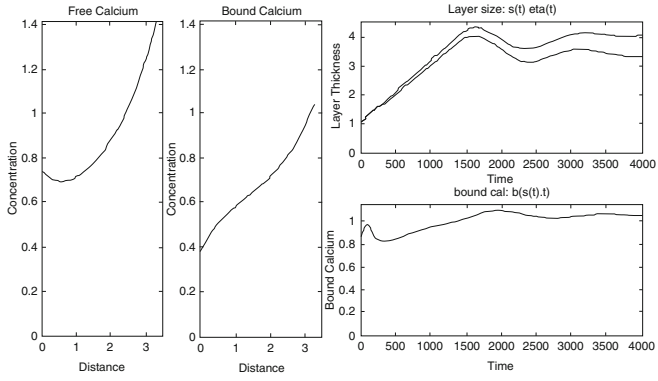


Fig. 7. Differentiation has high sensitivity to calcium. The graphs of free and bound calcium versus z are snapshots at time $t = 4{,}000$, where the interface $s(t) \approx 3.3$. The graph of layer thickness versus time shows the top of the TAL ($s(t)$, *lower line*) and the top of the CL ($\eta(t)$, *upper line*)

2.2 Optical Fibres (Collaborators: D. Abbott, P. Howell, A. Fitt, C. Voyce, B. Tilley, D. Schwendeman, T. Monro and Others)

The second problem that we shall consider concerns optical fibres. Optical fibres are used extensively in communications systems or optical detectors. These fibres have spawned many areas of research including investigation of their electromagnetic and mechanical properties. In this section we are interested in the actual manufacturing process. This problem was brought to a study group by Corning Inc [5].

Optical fibres work by using changes in the refractive index to guide light down the glass core of the fibre. Usually the glass is doped to provide the change in refractive index, but an alternative is to make the fibre with an array of holes down the centre, with cross-sections such as those as displayed in Fig. 8. The transition from the air to the glass provides the required changes in refractive index. One way to make the holey fibre is to take hollow straws of glass $\approx$1 m long and pack them together into a bundle of diameter $\approx$3 cm, with a solid straw in the middle, thereby creating a 'blank'. The blank is put in a furnace and drawn out into a fibre $\approx$100 km long with diameter measured in μm. The aspect ratio of the blank therefore changes by order 10^5. Corning are interested in what happens to the fibre in the drawing process, especially as surface tension will try to close the holes off. They are also interested in the behaviour of any bubbles within fibres, as these will cause problems. The evolution of bubbles and holes may be described by identical equations. There are a number of papers that study this problem including [7, 8] and [9] and the work here follows these ideas closely. The precise problem is discussed in more detail in the presentations available at [5].

To gain a basic understanding of the behaviour we take a simple case and consider the behaviour of one hollow straw of glass. We assume that the straw

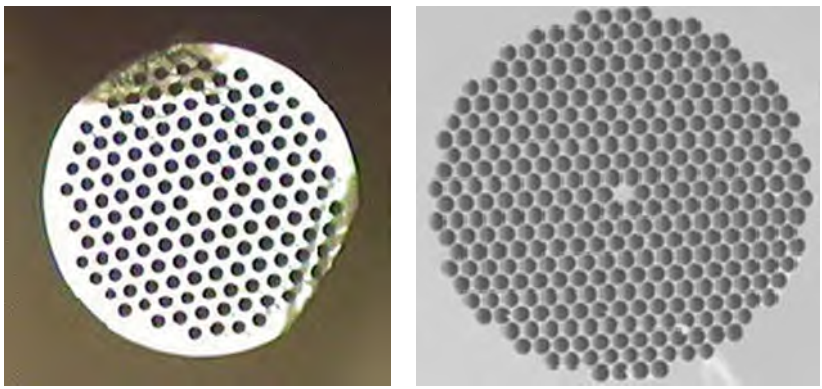


Fig. 8. Example cross-sections of 'holey' optical fibres [6]

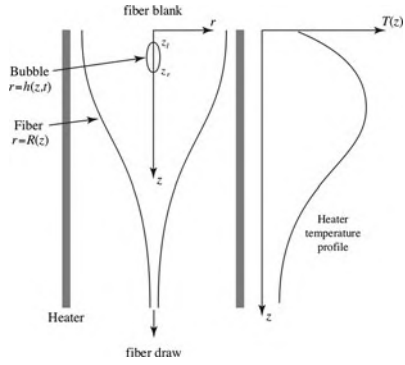


Fig. 9. Schematic of a fibre with a single concentric bubble or hole. Reproduced from [5]

is axis-symmetric and that the glass is a Newtonian viscous fluid. A schematic of the model is shown in Fig. 9. The radius of the hole is given by $r = h_1(z,t)$, where z is the vertical coordinate measured from the exit of the furnace. The radius of the glass is given by $r = h_2(z,t)$. Taking advantage of the long thin aspect ratio to neglect appropriate terms, the governing equations for the evolution of the fibre in the z-direction are

$$\rho(h_2^2 - h_1^2)\left(\frac{\partial w}{\partial t} + w\frac{\partial w}{\partial z} + g\right) = \frac{\partial}{\partial z}\left(3\mu(h_2^2 - h_1^2)\frac{\partial w}{\partial z} + \gamma(h_1 + h_2)\right), \quad (7)$$

$$\frac{\partial(h_2^2 - h_1^2)}{\partial t} + \frac{\partial\big(w(h_2^2 - h_1^2)\big)}{\partial z} = 0, \quad (8)$$

$$\frac{\partial h_1^2}{\partial t} + \frac{\partial w h_1^2}{\partial z} = \frac{p h_1^2 h_2^2 - \gamma h_1 h_2(h_1 + h_2)}{\mu(h_2^2 - h_1^2)}. \quad (9)$$

Here ρ and μ are the density and viscosity of the glass and γ is the surface tension of the glass in air. Gravity is represented, as usual, by g. The vertical velocity of the glass is denoted by $w(z,\, t)$ and $p(z,\, t)$ represents the pressure, above atmospheric, of gas in the hole. Boundary conditions are prescribed on the free surfaces $r = h_1(z,t)$ and $r = h_2(z,t)$, and also at the top $z = 0$ and the bottom $z = L$. Equation (7) represents conservation of momentum in the z-direction, (8) represents conservation of mass, and (9) represents conservation of momentum in the radial direction. The coefficient of $\partial w/\partial z$ in the first equation is known as the Trouton viscosity.

Suppose that we now consider the small hole limit, so that $h_1(z,t) \ll h_2(z,t)$. Then to leading order (7)–(9) reduce to

$$\rho h_2^2 \left(\frac{\partial w}{\partial t} + w \frac{\partial w}{\partial z} + g \right) = \frac{\partial}{\partial z} \left(3\mu h_2^2 \frac{\partial w}{\partial z} + \gamma h_2 \right), \tag{10}$$

$$\frac{\partial (h_2^2)}{\partial t} + \frac{\partial (w h_2^2)}{\partial z} = 0, \tag{11}$$

$$\frac{\partial h_1^2}{\partial t} + \frac{\partial (w h_1^2)}{\partial z} = p h_1^2 - \gamma h_1 \, . \tag{12}$$

Note that in this case the last equation for the radius of the hole has decoupled. This means that as we draw the blank, the glass acts as if there were no hole and the hole is forced to change in response to the glass flow. We may rewrite the last equation as

$$\frac{\partial a}{\partial t} + \frac{\partial (wa)}{\partial z} = pa - \gamma \sqrt{a}, \tag{13}$$

where $a(z, t)$ is the cross-sectional area of the hole. The evolution of a depends on the competition between the pressure p in the hole, which tries to keep the hole open, and the surface tension γ, which tries to close the hole. Corning want to keep the hole open, and tried changing the pressure in the hole to achieve this. However this was not very successful, and this may be explained by considering (13). This is a hyperbolic equation for a, and we can see that if the term pa dominates, then a will grow exponentially, while if $\gamma\sqrt{a}$ dominates, then the hole will pinch off. It is therefore extremely difficult to use the pressure to control the difference in these two terms and make the hole stay at the unstable equilibrium point $a = (\gamma/p)^2$.

Equation (13) also enables us to analyse how the shape and size of a bubble changes as a fibre is drawn. Suppose that at $t = 0$, the top and bottom of the bubble lie at $z_{\min}$ and $z_{\max}$ respectively. Then the bounding characteristics generated by (13) will tell us within what range the bubble lies. The bounding characteristics will have a profile similar to that depicted in Fig. 10, so that the bubble elongates as drawing progresses, (as long as pinch-off does not occur).

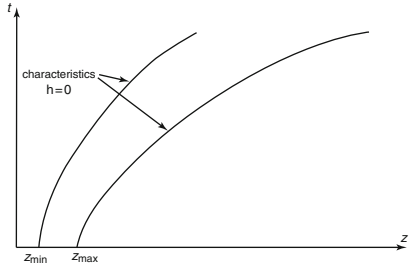
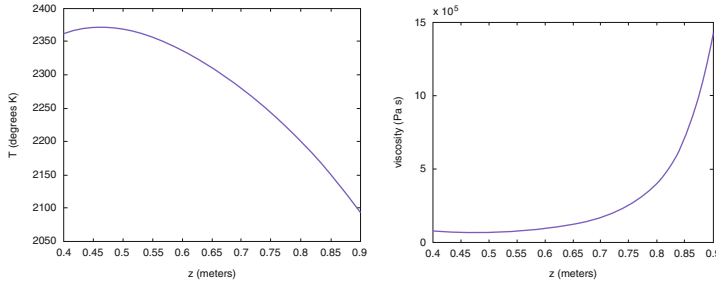


Fig. 10. Typical bounding characteristics which determine bubble evolution during the drawing process. Reproduced from [5]



(a) Temperature variations. (b) Viscosity variations.

Fig. 11. Typical temperature and viscosity variations in the z-direction as the fibre is drawn. Reproduced from [5]

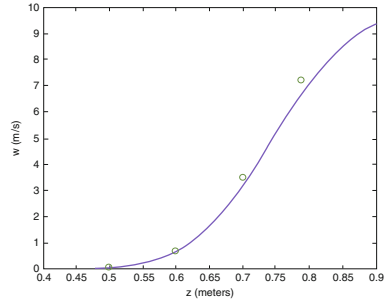


Fig. 12. Vertical velocity of the glass in the steady state, $w(z)$. Reproduced from [5]

To allow us to solve (10)–(12) we may reasonably assume that the density and surface tension of the glass are constant. However it is not so clear that we can assume a constant viscosity as the temperature variations down the fibre are large. We show typical temperature and viscosity variations in Fig. 11. Perhaps surprisingly the variations in viscosity are not extreme and as an approximation we may take it to be constant at around 10^6 Pa s. Imposing appropriate boundary conditions and assuming a steady glass flow, we solved (10) and (11) numerically for the velocity w, and the results are shown in Fig. 12.

Using this solution for $w(z)$, we may solve (12) for the evolution of a bubble. In doing this we will assume that the pressure in the bubble is spatially uniform, this assumes the gas can move easily inside it, that the mass of gas is constant and that the gas is governed by the ideal gas law. In this way the pressure varies both due to the changes in bubble shape and due to temperature variations along the fibre. The numerical solution for the evolution of

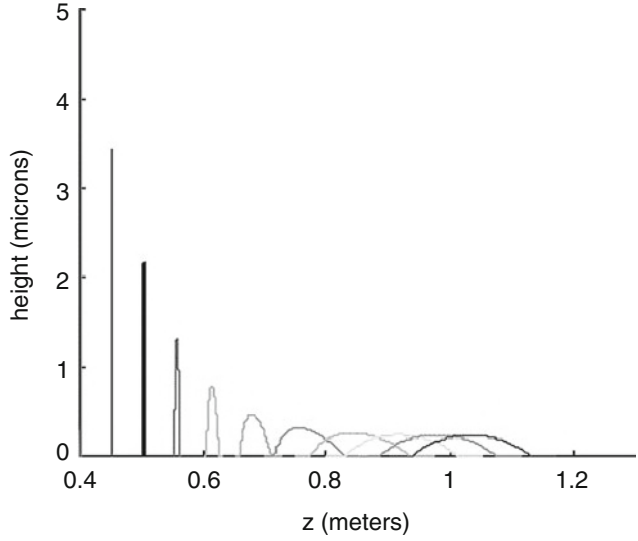


Fig. 13. Numerical solution for the steady state evolution of a bubble down the fibre. Reproduced from [5]

a bubble is shown in Fig. 13. The sharp peak at $z \approx 0.4$ represents a bubble near the top of the fibre with maximum radius 5 μm, with the top and bottom of the bubble very close together (5 μm). The subsequent shapes represent the evolution of the bubble as it progresses down the fibre, and we can see that it elongates dramatically and the radius decreases. For example, the peak at $z \approx 1$ m shows that the maximum radius of the bubble has decreased to approximately 0.25 μm, while the distance between the top and bottom of the bubble has increased to approximately 0.2 m. Corning calls these elongated bubbles threads, and would like to eliminate them.

The pressure within the bubble is shown in Fig. 14 and indicates that it increases in the hot region and then decreases in the solidifying region. Note also that, although the bubbles are elongated by the stretching of the glass they can be shown to be shorter than the length that would be predicted by the simple argument about bounding characteristics because the ends pinch-off.

Although many assumptions have been made, the continuum model proposed above has been able to provide great insight into the control of holes and the size of bubbles. A further item of interest would be to consider the effect of gas leaving the bubble by diffusion into the surrounding glass, which would affect the pressure. In addition, realistic fibres enclose thousands of bubbles,

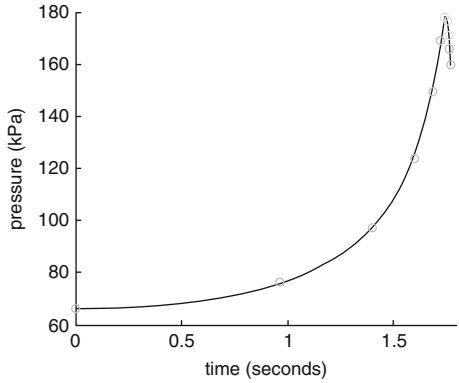


Fig. 14. Pressure in the bubble versus time (*circles* correspond to the times of the *bubble shapes* in Fig. 13). Reproduced from [5]

not just one, so it would be interesting to apply homogenisation techniques to extend the analysis for the single bubble.

2.3 Semiconductors (D. Schwendeman, P. Kramer, T. Witelski, L. Borucki)

Years ago the field of semiconductor modelling was perfect for the use of asymptotic analysis, as the important non-dimensional parameter, the ratio of the Debye length to the device size, was $O(10^{-10})$. However as technology has advanced and the size of the devices has decreased, this number is now approximately 10^{-1} or larger, and as a result the original asymptotic results are less applicable! There are two main aspects to semiconductor modelling:

1. Manufacturing processes
   - Etching
   - Lithography
   - Deposition
   - Implantation
2. Electrical behaviour
   - Quantum effects
   - Solar cell efficiency

In particular, there are many opportunities for the use of mathematics in modelling the quantum effects. Currently the engineers include these effects in a very much empirical manner (for some interesting quantum analysis see [10]).

In this section we concentrate on manufacturing aspects of the devices. The semiconductors are constructed from multiple layers, which are added

in separate processes (deposition, lithography, etching). It is very important to have a flat surface for each lithographic stage. As the multiple layers are added, each layer becomes more bumpy, and this affects the focusing required for accurate lithography. To avoid this difficulty, the surface is frequently polished flat using Chemical Mechanical Polishing. This process is not only abrasive, but also uses dissolving chemicals.

The Chemical Mechanical Polishing process uses three spinning discs. A cartoon of the polisher is shown in Fig. 15. The semiconductor wafer to be polished is stuck onto the bottom of the wafer carrier, which spins the wafer round and presses it onto the polishing pad. The polishing pad revolves in the opposite direction at the same angular speed (by having the same angular speed the relative velocity of the wafer and the polishing pad is independent of position so polishing will be quite uniform). A cross-section of this pad, as shown in Fig. 16, is abrasive and provides the mechanical element of the polishing while its surface is sprayed with a slurry to provide the chemical element of the polishing. There is a third spinning disc, which is the conditioning disc. As the polishing pad spins round it wears down and the conditioner renews the surface.

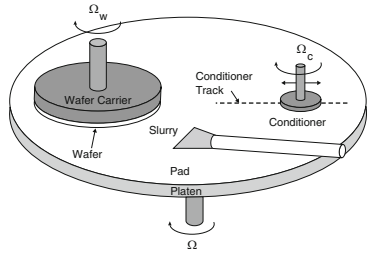


Fig. 15. Cartoon of Chemical Mechanical Polisher. Reproduced from [11]

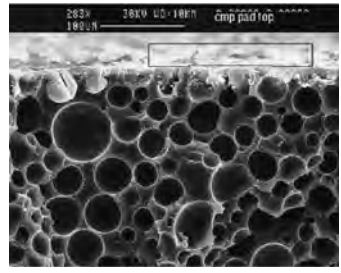
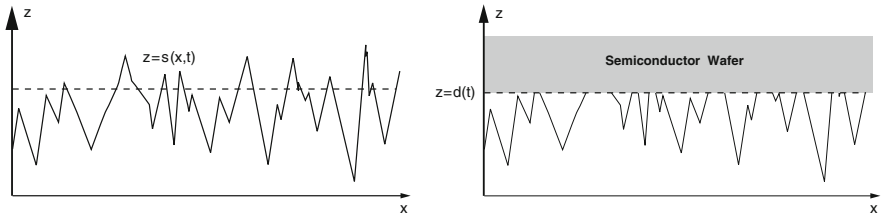
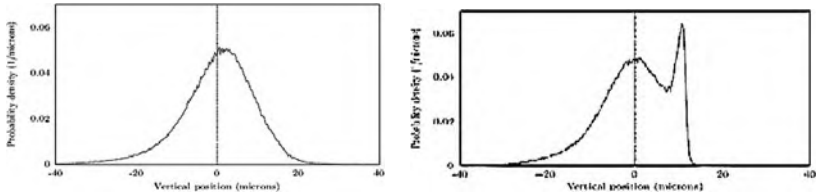


Fig. 16. Cross-section through a polishing pad. Reproduced from [11]



(a) Surface of polishing pad before polishing

(b) Surface of polishing pad after polishing



(c) Probability density function of height of surface before polishing

(d) Probability density function of height of surface after polishing

Fig. 17. Model of the surface of the polishing pad. Reproduced from [11]

The problem posed by Motorola at the study group was to investigate how the rough surface of the polishing pad changes as it is subjected to both wear and to conditioning. We were able to pose this as a continuum problem and a more detailed description much of the following is given in [11] and the study group report [12].

Let us first of all consider the wearing of the polishing pad. Before polishing, the surface of the pad will have a jagged surface, which we may represent by the function $z = s(x, t)$ as displayed in Fig. 17(a). If we let $\phi(z, t)$ be the probability density function of the height $s(x, t)$, then a graph of $\phi(z, t)$ will have a profile of the form depicted in Fig. 17(c). When the pad is used to polish the semiconductor wafer, because the wafer is hard and the pad quite compliant, the jagged surface will be flattened at a certain height $z = d(t)$, as shown in Fig. 17(b) and the flatten regions will then wear. Because of the wear the amount of pad surface at height $z = d(t)$ will increase and a spike develops in the probability density function at this height, as we can see in Fig. 17(d). This general behaviour will now be described mathematically.

The displacement of any point on the pad surface by the semiconductor wafer pressing down on it is given by $(z - d(t))H\big(z - d(t)\big)$, where $H(\cdot)$ is the Heaviside function. If we assume that the wear-rate of the surface is proportional to the square root of the displacement (Hertzian indentor), then by conservation of probability, we may write down

$$\frac{\partial \phi}{\partial t} + \frac{\partial}{\partial z}\left(\beta\sqrt{(z-d(t))\ H\big(z-d(t)\big)}\ \phi\right) = 0, \tag{14}$$

where β is the constant of proportionality.

Letting $q(z, t)$ be the fraction of solid pad in any plane z =constant, then $q(z, t) = \text{Prob}\big(z < s(x, t)\big)$, which is the cumulative density function. We therefore have the following relationship between q and ϕ:

$$\phi(z, t) = -\frac{\partial q(z, t)}{\partial z}, \tag{15}$$

and (14) may be rewritten, after using conditions when $z \to -\infty$, as

$$\frac{\partial q}{\partial t} + \beta\sqrt{(z-d(t))\ H\big(z-d(t)\big)}\ \frac{\partial q}{\partial z} = 0. \tag{16}$$

Now let us consider the conditioning process. The conditioner consists of a circular plate on which a regular array of small sharp diamonds are adhered. As the conditioner is pressed against the polishing pad the diamonds cut grooves into the surface of the polishing pad. The cuttings from the pad are removed in the slurry and the interwoven grooves create the new surface of the polishing pad. As a simple model we shall assume that we need consider only one diamond and that this moves randomly cutting grooves in a prescribed spatial interval. A cartoon of this process is shown in Fig. 18. Where the diamond is represented by the thin shaded triangle, and its endpoint is assumed to be at a height $z = h(t)$. The pad has a groove cut in it in the region $z - h(t) > 0$, and if the cutter has straight sides, it may be shown that

$$\frac{\partial q}{\partial t} = -\gamma\big(z-h(t)\big)H\big(z-h(t)\big)q, \tag{17}$$

where $q(z, t)$ is the cumulative density function mentioned above and γ is a constant of proportionality which measures the sharpness, or more specifically the steepness of the sides, of the cutter. Putting (16) and (17) together, for simultaneous wearing and conditioning, $q(z, t)$ satisfies the equation

$$\frac{\partial q}{\partial t} = \beta\sqrt{(z-d(t))\ H\big(z-d(t)\big)}\ \frac{\partial q}{\partial z} - \gamma\big(z-h(t)\big)H\big(z-h(t)\big)q, \tag{18}$$

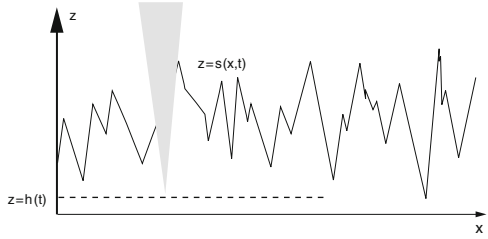


Fig. 18. Conditioning process using a *diamond* (represented by the *shaded triangle*) to cut a groove in the polishing pad below

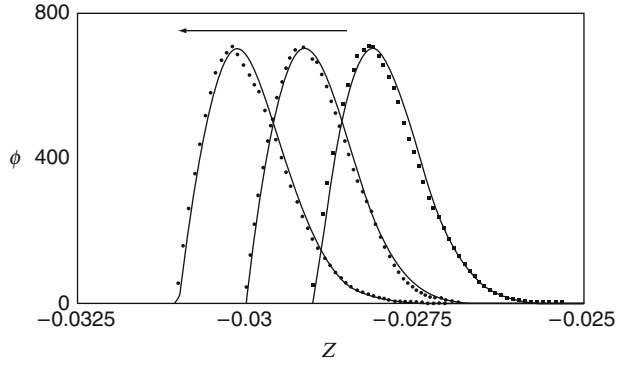


Fig. 19. Probability density function for wear of the polishing pad. The model results are represented by the *continuous line*, and the experimental results by *dots*. Reproduced from [11]

which is a linear hyperbolic partial differential equation. Usually the conditioning plate and the wafer are pressed down at a given rate, c, so we consider a "steady" problem where

$$h(t) = h_0 - ct, \quad d(t) = D + h(t) = D + h_0 - ct. \tag{19}$$

Looking for a travelling wave solution, $q(\eta)$, where η is a moving variable defined by $\eta = z - h(t)$, (18) reduces to a linear ordinary differential equation:

$$c\frac{\partial q}{\partial \eta} = \beta\sqrt{(\eta - D)\ H(\eta - D)}\ \frac{\partial q}{\partial \eta} - \gamma\eta H(\eta)q\ . \tag{20}$$

Solving (20), with appropriate conditions at infinity, in the case when no conditioning is applied and $\gamma = 0$ allows us to determine a probability density function ϕ as shown in Fig. 19. Solving (20) to include both the wear and the conditioning of the pad produces a probability density function as shown in Fig. 20. In both graphs the experimental results are marked with dots, and the model gives an excellent fit.

To conclude, the continuum model gives an excellent fit to the experimental data with very little parameter fitting. There are of course improvements that can be made to the model. We currently assume that the material of the polishing pad is solid. In fact it actually contains many cavities, as can be seen in Fig. 16. The initial data is not a delta function, representing a flat surface, but has a distribution, representing the cavities in the pad, and each new cut of the conditioner now exposes material with this distribution rather than cutting solid material.

2.4 Wine Making

The last problem is to do with wine making, and, rather than demonstrating the effectiveness of continuum modelling, it serves as cautionary tale to

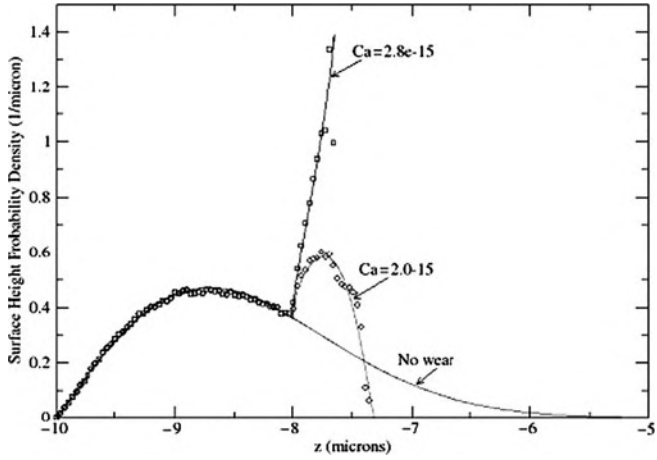


Fig. 20. Probability density function for simultaneous wear and conditioning of the polishing pad. The model results are represented by the *continuous line*, and the experimental results by *dots*. Reproduced from [11]

modellers working with industry. There are many problems in the wine making industry, for example how to spray the grapes to ensure total coverage [13], but one particular problem was brought by an Australian company that had problems with the labels starting to peel off their bottles [14, 15]. The modellers at the study group duly went away and considered many interesting problems, such as bubbles under the labels, the dynamics of the labelling machine and the modelling of the glue on the paper. However, it was only when someone did a simple experiment of actually trying to put a label onto a bottle that the answer became apparent. Paper is not orthotropic and tries to curve in one direction. To label bottles securely, the curvature of the label must be at right angles to the curvature of the bottle. Then the curvatures counteract each other and the label stays on. This fact has been well-known for years, was assumed to have been accounted for, but has somehow been overlooked in the quality assurance process. So the warning to modellers is that, before leaping ahead with all sorts of complicated mathematics, ensure that you have understood some of the basics first.

3 Overview

This article has reviewed several areas where continuum mathematics can be applied to solve industrial problems. For many physical problems some continuum models are relatively accessible to non-mathematicians as they are already coded into usable software. To be effective industrial mathematicians must therefore interact strongly with the industrialists to identify

"non-standard" problems because these provide opportunities for mathematics to have a significant impact in understanding the problem. In seeking these new problems it is important to note that although it is very motivating to involve complicated mathematics, the complexity of the mathematics may not always be correlated to the insight that it gives. The examples quoted show that simple mathematics, suitably applied, can produce understanding that allows problems to be solved. Finally the scope of continuum mathematics is vast and hence it can guide industrialists in an extremely wide range of areas.

Acknowledgements

Many thanks to Chris Bell for his assistance in the creation of this manuscript.

References

1. Topping, G., Malda, J., Dawson, R.A., Upton, Z.: Prim. Intention. **14**, 14–21 (2006)
2. Malda, J.: personal communication.
3. http://ric.colasanti.googlepages.com/camodels2
4. Mauro, T., Bench, G., Sidderas-Haddad, E., Feingold, K., Elias, P., Cullander, C.: J. Invest. Dermatol. **111**(6), 1198–1201 (1998)
5. http://www.math.wpi.edu/MPI2008/corning/corningMPI2008.pdf
6. Monroe, T.M.: personal communication.
7. Fitt, A.D., Furusawa, K., Monro, T.M., Please, C.P., Richardson, D.J.: J. Eng. Math. **43**, 201–227 (2002)
8. Voyce, C.J., Fitt, A.D., Monro, T.M.: Opt. Express. **12**, 5810–5820 (2004)
9. Voyce, C.J., Fitt, A.D., Monro, T.M.: J. Lightwave Technol. **26**(7), 791–798 (2008)
10. Cumberbatch, E., Uno, S., Abebe, H.: Euro. J. Appl. Math. **17**, 465–489 (2006)
11. Borucki, L.J., Witelski, T., Please, C.P., Kramer, P.R., Schwendeman, D.W.: J. Eng. Math. **50**, 1–24 (2004)
12. http://www.wpi.edu/Academics/Depts/Math/News/MPI2005/FinalReports/Araca.pdf
13. Barry, S.I., Weber, R.O.: In Hewitt J. (ed.) Proceedings of the 2001 Mathematics-in-Industry Study Group, pp. 28–40. ISBN 0-9578623-1-8, MISG. (2001)
14. Broadbridge, P., Fulford, G.R., Fowkes, N.D., Chan, D.Y.C., Lassig, C.: SIAM Rev. **41**(2), 363–372 (1999)
15. Hewitt, J. (ed.): Proceedings of the 1996 Mathematics-in-Industry Study Group, pp. 103–113. ISBN 0-646-28979-9, MISG. (1996)
16. http://miis.maths.ox.ac.uk/
17. http://www.infovisual.info/03/036_en.html

Wax Segregation in Oils: A Multiscale Problem

Mario Primicerio

Department of Mathematics "Ulisse Dini", University of Florence, v.le Morgagni 67/a, 50134, Florence, Italy, primicer@math.unifi.it

Preface

It is for me a great honour to be invited to deliver the Alan Tayler memorial lecture during this conference.

I had the fortune of sharing a long friendship with Alan, started more than forty years ago. But beyond the sincere friendship, we shared a common way of looking at applied mathematics and its relations with industry.

Well before the nowadays popular slogans on "knowledge-based economy", Alan was deeply conscious that mathematics could (and should) be a fundamental driving force in promoting innovation in industry and more generally in the society.

Alan put into this goal all of his enthusiasm and all of his effective action as a leader and as an organizer.

His contribution to the foundation of ECMI, to its first activities, in obtaining the first ECMI-contract from the EU it is well known to the ECMI "old guard". But it is up to us to act so that also our younger colleagues could thank Alan Tayler for the momentum he gave to the development of industrial mathematics in Europe.

This lecture is conceived in his spirit and not just in his memory.

1 Introduction

In the last few years our group was involved in a long-term research program partially supported by the societies of the ENI group (Enitecnologie and Agip), that is the main Italian holding in oil industry.

The program aims at understanding the behaviour of waxy crude oils subject to temperatures gradients. Indeed, this class of oils is characterized by the fact that they contain a relatively large amount of heavy hydrocarbons

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*,
Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_4,

(paraffins, asphaltenes etc.) that – as we shall discuss in detail in the following – may crystallize and eventually form gel-like structures thus influencing the motion of the oil in the pipeline.

The experimental evidence is that when these oils are pumped in pipelines crossing zones at relatively low temperature (as e.g. in the submarine pipelines) a deposit is formed at the walls that grows and hinders the flow, so that periodic "cleaning" operations are to be scheduled to keep a high efficiency of the transportation and to avoid a possible total clogging of the line. The research on the possible mechanisms responsible for the phenomenon and on their mathematical modelling is very active (see e.g. some general papers and reviews like [1–5]). The research program includes both an experimental part and a section aimed at modelling and simulation. The former is implemented in three laboratories: Eni Milano, the Istituto Donegani in Novara and the Department of Chemistry of the University of Florence. The latter is mainly done by our group: Antonio Fasano, Lorenzo Fusi and myself together with Loredana Faienza and Alessandro Monti and some others co-workers (Alberto Mancini, Fabio Rosso) who joined the team from time to time. A helpful contribution also came from John Ockendon.

It has to be noted that the cooperation among the teams is very intense, and this is witnessed by a number of papers in which experimental results are discussed in the framework of the mathematical models presented (see e.g. [6–8]).

To deal with a relatively simple situation, we will refer to an "ideal mixture" that mimics the behaviour of a real oil. It is a mixture of a given standard "wax" and a "solvent" (decane). The wax we chose has been characterized by its spectrum obtained by gas chromatography.

2 Segregation/Dissolution of Wax

For any waxy crude oil, and in particular for our "ideal mixture" with a given wax concentration c, a temperature T_{CL} can be defined such that, for $T > T_{CL}$ all wax is dissolved in the solvent while for $T < T_{CL}$ part of the wax segregates. Temperature T_{CL} is called **cloud temperature** or W.A.T. (wax appearance temperature). We are supposing that the system is always at thermodynamic equilibrium, a fact that is by no means granted.

Cloud temperature is usually determined by differential scanned calorimetry (DSC), and the measure can be made by raising or lowering the temperature of the sample: in the first case the temperature at which the peak in the heat exchanged occurs is sometimes called wax disappearance temperature (WDT) while the term WAT is used for the result obtained when the measure is performed with decreasing temperature.

Moreover, the determination of WAT and of WDT is influenced by the rate at which the temperature is varied. Here, we report the data of a typical experiment [9–11]:

Table 1.

WDT 10°C/min	$T_1 = 29.3$°C
WDT 1°C/min	$T_2 = 26.5$°C
WAT 1°C/min	$T_3 = 21.3$°C
WAT 10°C/min	$T_4 = 20.2$°C

A few comments on these results are in order (for general theoretical discussion see e.g. [12–22]):

1. The differences between T_1 and T_2 and between T_3 and T_4 show that the process of dissolution (crystallization) is not instantaneous but that the system takes some time to reach the thermodynamical equilibrium.
2. The difference between WAT and WDT shows that undercooling occurs practically always, as it is rather usual in phase-change processes.
3. A difference of about 5°C between T_2 and T_3 is commonly found in different situations of concentration.

We note that in the literature the term "cloud temperature" or "cloud point" is sometimes related to an optical determination. Of course the accuracy of this measurement is strongly dependent on the method used [23, 24], since it is difficult to measure the variation of optical properties when only micro-crystals are present (a possible colloidal transition state has been also supposed to exist). Moreover, the method is applicable to our "ideal mixture" that is optically transparent but practically useless when commercial oils are concerned. Let us come back to the definition of T_{CL} and assume we can associate a value T_{CL} to each value of the concentration c of wax in the mixture. As it can easily be expected, it turns out that T_{CL} is a monotonically increasing function of c. For our purposes, it will be useful to consider the inverse function of $T_{CL}(c)$ and to define $c_{SAT}(T)$ as the maximum amount of wax that can be added to a unit volume the solvent kept at temperature T without producing any crystallization. It can be seen as the *solubility* of wax in the solvent as a function of the temperature.

To model the phenomenon, we will use the following functions:

1. $c(\mathbf{x}, t)$: total wax concentration at point $\mathbf{x}$ at time t.
2. $C(\mathbf{x}, t)$: concentration of dissolved wax.
3. $G(\mathbf{x}, t)$: concentration of segregated (crystallized) wax.

Of course it is:

$$c(\mathbf{x}, t) = C(\mathbf{x}, t) + G(\mathbf{x}, t). \tag{1}$$

2.1 Case of Thermodynamical Equilibrium

As we will see, the phenomenon we are studying is a typical multiscale phenomenon, so that it is quite possible that in the time scale of the experiment the process of dissolution/segregation can be considered to be instantaneous. In this case we have

$$C(\mathbf{x},t) = \min(c(\mathbf{x},t), c_{SAT}(T(\mathbf{x},t))) \tag{2}$$

$$G(\mathbf{x},t) = \max(0, c(\mathbf{x},t) - c_{SAT}(T(\mathbf{x},t))) \tag{3}$$

so that (1) is automatically satisfied.

2.2 A Case of Macroscopic Kinetics

Consider the case in which the thermodynamical equilibrium is reached in finite time with a characteristic time constant. If we still remain in the framework of a macroscopic description, we should postulate the existence of a sort of chemical potential acting as the driving force of the phenomenon.

The simplest assumption we can postulate is that the rate of segregation/dissolution is proportional to the deviation from the thermodynamical equilibrium i.e.

$$\frac{\partial G}{\partial t} = \theta\beta(C(\mathbf{x},t) - c_{SAT}(\mathbf{x},t)) \tag{4}$$

where $\beta > 0$ is the inverse of the characteristic time and θ is a factor that ensures that G_t vanishes if both $(C - c_{SAT})^+$ and G are zero. Thus

$$\theta = H(G + (C - c_{SAT})^+), \tag{5}$$

where H is the Heaviside jump function

$$H(z) = \begin{cases} 0, & \text{if } z \le 0 \\ 1, & \text{if } z > 0. \end{cases} \tag{6}$$

A simple generalization consists in assuming different values of β for $(C - c_{SAT})$ positive and negative and/or to include the possible dependence of β on the temperature.

2.3 A Microscopic Description

A possible microscopic description of the process of segregation (crystallization) is based upon two mechanisms: **nucleation** and **growth**. One defines $\dot{\nu}$ to be the rate of birth of new crystals per unit volume of the solution and $\dot{\rho}$ as the radial growth of the crystals that are assumed approximately spherical.

We will neglect the radius of the newborn crystals and we will assume that $\dot{\nu}$ and $\dot{\rho}$ (both depending on C and T) are such that their ratio is constant. This is the so-called **isokinetic assumption** that can be written as:

$$\begin{cases} \dot{\rho} = \dot{\rho}_0 F(C,T), \\ \dot{\nu} = \dot{\nu}_0 F(C,T). \end{cases} \tag{7}$$

Under these assumptions (and normalizing the quantities so that the density is equal to one) we have:

$$\frac{\partial G}{\partial t} = 4\pi\dot{\rho}(t)\int_0^t \frac{\dot{\nu}_0}{\dot{\rho}_0}\dot{\rho}(\tau)\left[\int_\tau^t \dot{\rho}(s)ds\right]^2 d\tau, \tag{8}$$

and after some simple manipulations we get

$$\frac{\partial G}{\partial t} = 4\left(\frac{\pi\dot{\nu}_0\dot{\rho}_0^3}{3}\right)^{1/4} G^{3/4}F(C,T). \tag{9}$$

Consequently, we can obtain the number of crystallites per unit volume

$$N(t) = \left(\frac{3}{\pi}\right)^{1/4}\left(\frac{\dot{\nu}_0}{\dot{\rho}_0}\right)^{3/4} G^{1/4} \tag{10}$$

and the average radius

$$\bar{R}(t) = 4^{-1/3}\left(\frac{3}{\pi}\right)^{1/4}\left(\frac{\dot{\rho}_0}{\dot{\nu}_0}\right)^{1/4} G^{1/4}. \tag{11}$$

Of course, to complete the description of the process we have to specify the form of the function F in (7). We note that, in any case, this picture can only refer to the crystallization (of course it does not apply to dissolution) and hence F has to vanish if and only if thermodynamical equilibrium has been reached and thus if C reaches the value c_{SAT}. The simplest choice leads us to

$$\frac{\partial G}{\partial t} = K(\dot{\nu}_0\dot{\rho}_0^3)^{1/4}[C(x,t) - c_{SAT}(T(x,t))]G^{3/4}, \tag{12}$$

or, more generally to

$$\frac{\partial G}{\partial t} = K(\dot{\nu}_0\dot{\rho}_0^3)^{1/4}[C(x,t) - c_{SAT}(T(x,t))]^q G^{3/4} \tag{13}$$

for some q, in accordance with typical models for crystallization of polymer melts (see [25, 26]).

We note that this model is based on concepts similar to the ones used in [25–30], with the difference that the phenomenon of "impingement" among growing crystals is much less relevant in the present case because concentration of wax in solvent is very low.

In the literature of waxy oils an approach similar to the one illustrated above has been adopted in [31–33], but just in spatially homogeneous cases and when the cooling rate is constant. Under these assumptions the so-called "Avrami thumb rule" [34] is applied.

We conclude this section showing a comparison between the approaches illustrated under (2.1), (2.2) and (2.3) (see Fig. 1).

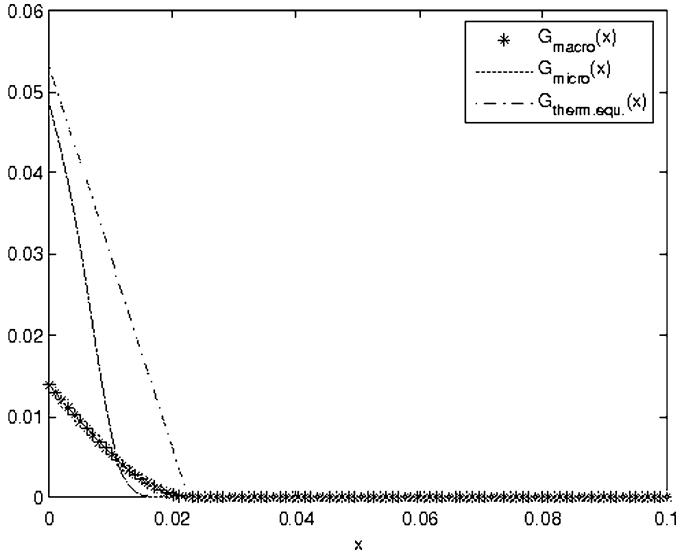


Fig. 1. Comparison of function G for the three approaches of macroscopic kinetics $G_{macro}(x)$, microscopic kinetics $G_{micro}(x)$ and thermal equilibrium $G_{therm.equ.}(x)$. The picture refers to a simulation of a stratum $0 < x < L$ where the boundary $x = L$ is at a constant temperature $T >$WAT while the boundary $x = 0$ is being cooled at a temperature $T <$WAT; we assume that saturation is linear in x and that $G = 0$ initially

3 Diffusion/Convection of Heat

Heat transfer turns out to be the driving force for the deposition of wax since temperature is the key quantity in the process of change of state of wax.[1] Conversely, one can ask how much the process of segregation/dissolution of wax influences the thermal field.

Within the experimental uncertainty, one can claim that the state of aggregation of wax (and even its concentration in the mixture) does not have an important effect on the thermal diffusivity of oil (see [35, 36]).

Concerning the latent heat associated to the change of phase, it is around $10\,\mathrm{J\,g^{-1}}$. Since concentration of wax is below 10% and the heat capacity of oil is about $5\,\mathrm{J\,g^{-1}\,K^{-1}}$ and the change of phase takes place across a few degrees, we can claim that the effect of latent on the determination of thermal field

[1] Indeed, sedimentation by gravity is negligible since wax and oil have almost the same density; indeed, the deposit on the walls of pipelines has the same thickness at every point of a given cross section.

can be neglected.[2] This means that the latter can be found without knowing C and G, as least as far as just conduction is considered. Of course this is no longer true in general when convection has to be taken into account, since the rheological properties of the mixture can be strongly influenced by the state of aggregation of wax and since the presence of deposit determines the motion.

4 Diffusion/Convection of Wax: Gelification

Let us turn our attention to the diffusion of segregated and dissolved wax in the solvent. We assume the validity of Fick's law and denote the diffusivity of segregated and dissolved wax by D_G and D respectively.

Of course, one expects that

$$D_G \ll D. \tag{14}$$

Moreover, the mobility of dissolved wax within the mixture is hindered by the presence of wax crystals, at least if G is "large enough". More precisely, one can see that when G exceeds a threshold value G^*, crystallites tend to aggregate and to entrap liquid (i.e. oil + dissolved wax) and form a gel. Process of gelification is not instantaneous but follows a kinetics that we can model by introducing a quantity $g(x,t)$ characterizing the **degree of gelification** whose evolution is governed by the following law

$$\frac{\partial g}{\partial t} = \Phi(G - G^*) \tag{15}$$

where Φ is a nondecreasing function of its argument, $\Phi(0) = 0$. In general, Φ will also depend on temperature (and will be monotonically decreasing). Of course (15) only refers to mixtures at rest since motion can strongly influence in contrasting gelification.

Consistently with the description above, we will assume that D (and also D_G whenever it will be taken into consideration) is a given decreasing function of $K = G(1+g)$, vanishing when K exceeds some critical value K^*.

The degree of gelification is also relevant to the phenomenon of adhesion of wax aggregates to the pipe walls (deposition). In principle, in order to model this phenomenon, one should also include the nature of the wall, its rugosity and so on. For simplicity, we can assume that when G reaches the value G^* at the wall, then the deposit begins to form and grow; in other words we identify the deposit with the region where $g > 0$.

Another phenomenon that is observed is the so-called **ageing** of the gel. In our model, this fact is explained both by the diffusion of liquid wax within

[2] On the other hand, the mathematical problem to be solved if one takes this coupling into account is far from being trivial (see [37]).

the gel (although with lower diffusivity) and by the consequent additional segregation that takes place whenever

$$G(x,t) < C(x,t) - C_{SAT}(T(x,t)). \tag{16}$$

When a mixture is brought at $T_1 \gg T_{CL}$ and then put in a vessel whose walls are kept at $T_2 < T_{CL}$ the following facts are observed (see [36, 38] and also [39, 40]):

1. A deposit is formed at the walls and its final thickness is reached in a short time.
2. The concentration of wax in the deposit continues to increase.
3. In the deposited layer concentration of wax decreases when approaching the walls.

An additional information we got from experiment and from literature is that the mechanism of gelification – and hence its influence on diffusivity – is strongly dependent on the nature of wax since the geometry of crystallites plays an important role [11].

A Multiscale Problem

Summing up we have briefly discussed five processes that are relevant to the phenomenon to be studied, and each of them has a corresponding time scale:

1. Thermal diffusion (t_1).
2. Segregation of wax (t_2).
3. Diffusion of dissolved wax (t_3).
4. Diffusion of segregated wax (t_4).
5. Gelification (t_5).

Moreover, when the motion of the mixture is to be taken into account, we have also

6. Motion of the fluid (t_6).

Of course the model that can be used should take into account the practical cases to be studied. Since they span over a large variety of situations (depending on the type of oil and of thermal conditions) in the following we will consider different scenarios separately.

5 Thermodynamical and Thermal Equilibrium

A first scenario is studied in [41] where it is assumed that

$$t_1,\ t_2 \ll t_3,\ t_4 \ll t_5,\ t_6, \tag{17}$$

corresponding to a situation in which the fluid is at rest, the thermal field attains its asymptotic (stationary) profile in a very short time interval and phase equilibrium is instantaneously reached.

Under these assumptions, we will consider a one-dimensional geometry having in mind the interpretation of experiments done on commercial oils with a laboratory device called "cold finger" where a steady thermal gradient is applied between two co-axial cylinders kept at constant temperatures T_1 and T_2 and the gap between the two cylinders is filled by oil with given wax concentration c^*. Of course at least one of the two thermostats is maintained at a temperature below $T_{CL}(c^*)$ [7,8].

Just to simplify notation, we refer here to plane (rather than cylindrical) symmetry, and for the same reason we will assume that T_{CL} depends linearly on concentration in the range of interest, so that

$$\frac{dc_{SAT}}{dT} = \gamma,\ \gamma > 0 \text{ constant.} \tag{18}$$

Since we have assumed that temperature reaches its stationary (linear) profile, we have

$$c_{SAT}(x) = A + Bx,\ x \in [0, l] \tag{19}$$

where $A = c_{SAT}(T_1)$ and B are positive constants assuming that T_2 (i.e. the temperature at the boundary $x = l$) is higher than the temperature T_1 of the wall $x = 0$ ($B = \gamma(T_2 - T_1)/l$).

But assuming that

$$c^* > c_{SAT}(l) = A + Bl,$$

the assumption of instantaneous thermodynamical equilibrium implies that

$$C(x, 0) = c_{SAT}(x),\ G(x, 0) = c^* - c_{SAT}(x). \tag{20}$$

As long as deposition is not taken into account, the boundary conditions are of course

$$DC_x(0, t) + D_G G_x(0, t) = DC_x(l, t) + D_G G_x(l, t) = 0,\ t > 0. \tag{21}$$

At this point we have to consider two different cases: first we will see what happens if diffusion of segregated wax plays a role, then we will discuss the case in which the crystallites can be thought to be immobile. We will start assuming that diffusivities are given and constant.

5.1 The Case $t_3 \sim t_4$ (Non-Negligible Crystal Diffusivity)

Starting from the initial situation (20) we can define $\hat{t}$ as

$$\hat{t} = \sup\{t : G(x, t) > 0, \quad x \in [0, l]\}. \tag{22}$$

This means that in the time interval $[0, \hat{t})$ the mixture is always **saturated** (that means $c > c_{SAT}(x)$). Therefore

$$C(x,t) = A + Bx, \quad 0 < x < l, \quad 0 < t < \hat{t} \tag{23}$$

$$\begin{cases} G_t - D_G G_{xx} = 0, & 0 < x < l,\ 0 < t < \hat{t} \\ G(x,0) = c^* - A - Bx, & 0 < x < l, \\ G_x(0,t) = G_x(l,t) = -\frac{D}{D_G} B, & 0 < t < \hat{t}. \end{cases} \tag{24}$$

Of course, $\hat{t} < +\infty$, if we exclude the unrealistic (and trivial) case $\dfrac{D_G}{D} > \dfrac{c_2 - c_1}{2c^* - c_1 - c_2}$ where we have written $c_i = c_{SAT}(T_i), \quad i = 1, 2.$

Since G_x is negative, by maximum principle, the definition of $\hat{t}$ implies $G(l, \hat{t}) = 0$ and, for any $t > \hat{t}$ a free boundary $x = s(t)$, $s(\hat{t}) = l$ will exist separating the **saturated** region $(0, s(t))$ where $G > 0$, from the **unsaturated** region $(s(t), l)$ where $c(x,t) = C(x,t) < c_{SAT}(x)$.

More specifically,

$$C(x,t) = A + Bx, \quad 0 < x < s(t), \quad t > \hat{t}, \tag{25}$$

$$\begin{cases} G_t - D_G G_{xx} = 0, & 0 < x < s(t), \quad t > \hat{t} \\ G(x,\hat{t}) = \hat{G}(x), & 0 < x < s(\hat{t}) = l, \\ G_x(0,t) = -\frac{D}{D_G} B, & t > \hat{t}, \\ G(s(t),t) = 0, & t > \hat{t}, \end{cases} \tag{26}$$

where $\hat{G}(x)$ is found as $G(x, \hat{t})$ from the solution of (24).[3]

On the other hand

$$\begin{cases} C_t - DC_x x = 0, & s(t) < x < l,\ t > \hat{t} \\ C_x(l,t) = 0, & t > \hat{t} \\ C(s(t),t) = A + Bs(t), & t > \hat{t}, \end{cases} \tag{27}$$

and

$$G(x,t) = 0, \quad s(t) < x < l, \quad t > \hat{t}. \tag{28}$$

Mass conservation, i.e. flux continuity across $x = s(t)$ provides the free boundary condition that completes the problem

$$DB + D_G G_x(s(t)_-, t) = DC_x(s(t)_+, t). \tag{29}$$

Problem (25)–(29) is an implicit two-phase free boundary problem. In [42] it is proved that it can be immediately reduced to a form for which the results

[3]The latter exists and is unique within the class of bounded functions.

of [42] and [43] can be applied and thus prove that a classical solution exists globally.

The asymptotic profile $(C_\infty, G_\infty, s_\infty)$ of the solution is the following:

$$G_\infty(x) = \begin{cases} \frac{D}{D_G} B(s_\infty - x), & x \in [0, s_\infty], \\ 0, & x \in [s_\infty, l] \end{cases} \tag{30}$$

$$C_\infty(x) = \begin{cases} A + Bx, & x \in [0, s_\infty], \\ A + Bs_\infty, & x \in [s_\infty, l] \end{cases} \tag{31}$$

and s_∞ is found from the global mass balance as the unique positive solution of the algebraic equation

$$\frac{B}{2}\left(\frac{D - D_G}{D_G}\right) s_\infty^2 + Bs_\infty - (c^* - A)l = 0. \tag{32}$$

As we anticipated, the model above does not include a specific mechanism for deposition (i.e. for adhesion to the cold wall $x = 0$) and assumes that D and D_G are constant. In the spirit of Sect. 4 we can say that this implies that G is always below the critical value G^*, i.e. when $G_\infty(0) = \dfrac{D}{D_G} Bs_\infty < G^*$.

The Deposit

A possible way of incorporating deposition in the model above is to assume that all the wax[4] arriving at the cold wall sticks to its surface and does not take part in the diffusion process. This fact can be modelled introducing a second free boundary $x = \sigma(t)$ where $\sigma(t)$ represent the thickness of the deposit or assuming that such thickness is negligible and that the wax reaching $x = 0$ simply leaves the system; this corresponds to replacing the third condition in (26) by $G_x(0, t) = 0$. This approach (with or without the free boundary $\sigma(t)$) has been used to interpret the data of the cold finger experiment (see [7]); in [41] the difference of heat between the mixture and the deposit has been also taken into account.

A basic difficulty of this approach is to evaluate the wax concentration in the deposit i.e. the amount of oil (or, rather, of mixture) that is "entrapped" and, if not, to estimate how much the displacement of the liquid caused by the deposit is relevant to the process [44, 45].

A possible way of answering this question is to perform the experiment until the asymptotic situation is reached and to weigh the total mass M_D^∞ per unit surface of the deposit. Knowing the mass of wax initially present and the quantity that is still in the solution (at a concentration equal to $c_{SAT}(0)$), the mass of the deposited wax M_w^∞ can be calculated. Hence the mass of entrapped oil is given by $M_D^\infty - M_w^\infty$. Nevertheless, the experiment is delicate since it lasts for several hours and its results are still not conclusive [7]. We add

[4] Or a given fraction of it.

that, under the assumption of linear dependence of c_{SAT} on T, the gradient of solubility γ can be evaluated by means of two asymptotic measures M_D^∞ of the deposited mass corresponding to two values of T_2 (say $\bar{T}_2$ and $\bar{\bar{T}}_2$). Indeed

$$\gamma = \frac{|\bar{M}_D^\infty - \bar{\bar{M}}_D^\infty|}{l|\bar{T}_2 - \bar{\bar{T}}_2|}.$$

Hindered Diffusion

An alternative approach consists in prescribing the dependence of D (and of D_G) on G, as was discussed in Sect. 4, or even on $G(1+g)$.[5]

Some preliminary simulations have been done (not taking into account g) and assuming that D is constant for $G < G^*$ and jumps to zero at G^*. Similar results were obtained imposing the threshold G^* not to G but to $G + C$.

In all these simulations the deposit was defined as the region where G (or $G + C$) exceeds G^*.

5.2 The Case $t_3 \ll t_4$

If we assume that, in the time scale of the experiment, the segregated wax is practically immobile the mathematical aspects of the model change totally, since letting D_G tend to zero is a singular perturbation of the problem.

Indeed, if we start from the same initial situation (20) with the natural boundary condition

$$C_x(0,t) = C_x(l,t) = 0, \tag{33}$$

The unsaturated region appears from the very beginning (i.e. $\hat{t} = 0$). Moreover, in order to make the model consistent it is necessary either to introduce a boundary layer close to $x = 0$ or to postulate a mechanism of deposition as we did above.

Let us confine ourselves to the approach used in Sect. 5. More specifically let us consider its simplest case in which the dissolved wax reaching $x = 0$ is assumed to be simply leaving the system.

Thus, we have the following problem

$$G(x,t) = \begin{cases} G_0(x) = c^* - c_{SAT}(x),\ 0 < x < s(t),\ t > 0 \\ 0, \qquad\qquad\qquad\qquad s(t) < x < l \quad t > 0 \end{cases} \tag{34}$$

$$C(x,t) = A + Bx,\ 0 < x < s(t),\ t > 0 \tag{35}$$

[5] In the latter case, we have to assume $t_6 \sim t_3, t_4$.

while, in the unsaturated region we have

$$\begin{cases} C_t - DC_{xx} = 0, & s(t) < x < l,\ t > 0 \\ C_x(l,t) = 0, & t > 0, \\ C(s(t),t) = A + Bs(t), & t > 0, \\ s(0)) = l \end{cases} \tag{36}$$

with the free boundary condition

$$DB - DC_x(s(t),t) = -G_0(s(t))\dot{s}(t),\ t > 0. \tag{37}$$

This is a free boundary problem formally similar to a Stefan-type problem and its well-posedness in a classical sense is proved in [46].

6 Phase Equilibrium in a Transient Thermal Field: No Gelification

In this section we will assume that

$$t_2 \ll t_1,\ t_3,\ t_4, \ll t_5,\ t_6$$

so that the dissolution/segregation of wax can be considered as instantaneous while heat conduction and wax diffusion occur over the same time scale.

6.1 A General Problem: Weak Solution

Let $Q_{\tilde{t}} \equiv \Omega \times (0,\tilde{t})$ be a general smooth cylinder in $\mathbb{R}^3 \times \mathbb{R}$ and assume that initial and boundary conditions are given for temperature on $\Omega \times 0$ and $\partial\Omega \times (0,\tilde{t})$. In the assumptions of Sect. 3 the function $T(\mathbf{x},t)$ can be found and we can define Q^+ as the (so far unknown) subset of $Q_{\tilde{t}}$ where $c(\mathbf{x},t) > c_{SAT}(\mathbf{x},t)$, i.e. where $G(\mathbf{x},t) > 0$ and $C(\mathbf{x},t) = c_{SAT}(T(\mathbf{x},t))$.

Assume that D and D_G are constant and define

$$\mathcal{L}_1 C = C_t - D\Delta C,\ \mathcal{L}_2 G = G_t - D_G \Delta C.$$

Mass conservation implies that

$$\mathcal{L}_1 C + \mathcal{L}_2 G = 0,\ \text{in } Q^+. \tag{38}$$

But $C(\mathbf{x},t) = c_{SAT}(T(\mathbf{x},t))$ in Q^+ and hence $\mathcal{L}_1 C$ is a known quantity $q(\mathbf{x},t)$

$$q(\mathbf{x},t) = \frac{\partial}{\partial t} c_{SAT}(T(\mathbf{x},t)) - D\Delta c_{SAT}(T(\mathbf{x},t)). \tag{39}$$

Thus we have

$$\mathcal{L}_2 G = -q(\mathbf{x},t),\ \text{in } Q^+ \tag{40}$$

Now set $Q^- = Q_{\tilde{t}} \setminus Q^+$ so that

$$G(\mathbf{x},t) = 0,\ \mathcal{L}_1 C = 0,\ \text{ in } Q^-. \tag{41}$$

If both Q^+ and Q^- are non-void and are separated by a smooth surface S, then we have

$$G = 0,\ c = C = c_{SAT},\ \text{ on } S \tag{42}$$

$$\left[D_G \frac{\partial G}{\partial \mathbf{n}} + D \frac{\partial C}{\partial \mathbf{n}} \right]_{S^+} = \left[D \frac{\partial C}{\partial \mathbf{n}} \right]_{S^-} \tag{43}$$

where $\mathbf{n}$ is the normal vector to $S \times \{t\}$ and $[\,]_{S^+}$ (resp. $[\,]_{S^-}$) denote the limit of the quantity in brackets when $(\mathbf{x},t) \in Q^+$ (resp. $\in Q^-$).

Defining

$$U(\mathbf{x},t) = c(\mathbf{x},t) - c_{SAT}(T(\mathbf{x},t)) \tag{44}$$

the problem can be written formally as

$$U_t - \nabla \cdot \left(D \left[1 + \left(\frac{D_G}{D} - 1 \right) H(U) \right] \nabla U \right) \in -q(\mathbf{x},t) \tag{45}$$

where H is the Heaviside graph.

Weak solutions $U \in H^{j,j/2}(Q_{\tilde{t}}) \cap W^{1,0}(Q_{\tilde{t}})$ for some $j \in (0,1)$ and for any $\tilde{t} > 0$ has been proved to exist in [45].

Remark 1. Note that U is positive in Q^+ and negative in Q^- and that the (45) could be interpreted as the model for the diffusion of two immiscible chemical substances (of concentration U in Q^+ and $-U$ in Q^-) that diffuse in a host medium and undergo, on the contact surface, a fast chemical reaction whose products precipitate. In this picture the term $-q(\mathbf{x},t)$ would represent a volumetric source/sink.

6.2 One-Dimensional Classical Solutions

In one-dimensional cases, more information can be obtained. Once again we refer for simplicity to planar symmetry $x \in (0,l)$.

Let us fix the temperature at $x = l$ at a value

$$T(l,t) = T_2, \quad t > 0, \tag{46}$$

and let

$$T(0,t) = T_2 - \phi(t), \quad t > 0 \tag{47}$$

with $\phi(t)$ monotonically increasing, $\phi(0) = 0$. To be specific we take $\phi(t) = \lambda t$.

Furthermore, we assume that

$$T(x,0) = T_2,\ 0 < x < l, \tag{48}$$

$$c(x,0) = c^* < c_{SAT}(T_2),\ 0 < x < l, \tag{49}$$

so that in the initial situation all wax is dissolved (i.e. the slab is completely unsaturated).

Of course, no segregation will take place until the time $\bar{t}$ such that

$$c_{SAT}(T_2 - \lambda \bar{t}) = c^*. \tag{50}$$

Recall that, in our assumption, the thermal field can be found independently of the knowledge of $C(x,t)$ and $G(x,t)$.

For $t > \bar{t}$ a region $Q^+ \equiv \{(x,t) : 0 < x < s(t),\ t > \bar{t}\}$ will appear where $G > 0$ and $C(x,t) = c_{SAT}(T(x,t))$.

Within Q^+ we have

$$\frac{\partial G}{\partial t} - D_G \frac{\partial^2 G}{\partial x^2} = -\frac{\partial}{\partial t} c_{SAT}(T(x,t)) + D \frac{\partial^2}{\partial x^2} c_{SAT}(T(x,t)) \tag{51}$$

and on $x = 0$ the following condition has to be fulfilled for $t > \bar{t}$

$$\left[D_G \frac{\partial G}{\partial x} \right]_{x=0} + \left[D \frac{\partial}{\partial x} c_{SAT}(T(x,t)) \right]_{x=0} = 0. \tag{52}$$

On the other hand, the region $Q^- \equiv \{(x,t) : s(t) < x < l,\ t > \bar{t}\}$ is such that $G = 0$ and hence

$$\frac{\partial}{\partial t} C(x,t) - D \frac{\partial^2}{\partial x^2} C(x,t) = 0,\ (x,t) \in Q^- \tag{53}$$

and

$$\left[\frac{\partial C}{\partial x} \right]_{x=l} = 0,\ t > \bar{t}. \tag{54}$$

Finally, the free boundary is characterized by the conditions

$$C(s(t),t) = c_{SAT}(T(s(t),t)),\ \text{ i.e. } G(s(t),t) = 0, t > \bar{t}, \tag{55}$$

$$\left[D \frac{\partial}{\partial x} C(T(x,t)) \right]_{x=s(t)^+} = \left[D \frac{\partial}{\partial x} c_{SAT}(T(x,t)) + D_G \frac{\partial G}{\partial x} \right]_{x=s(t)^+}. \tag{56}$$

This free boundary problem is considered in [46] and its well-posedness in the classical sense is proved.

7 Phase Equilibrium in a Transient Thermal Field with Gelification

Here, we consider cases in which gelification takes place, but its characteristic time t_5 (as well as t_2) is negligible with respect to t_1, t_3 and t_4 so that the process can be thought of as a change of phase, occurring at a given temperature depending on temperature.

In this case a phase diagram for oil-wax mixtures can be drawn in the (c, T) plane in which (see Fig. 2) zone A corresponds to dissolved wax ($T > T_{CL}(c)$, i.e. $c < c_{SAT}(T)$), in zone B we have coexistence of dissolved and segregated phases, while zones C and D correspond to gel and a $T_{GEL}(G)$ (or, equivalently, $G = G_{GEL}(T)$) can be defined. Zone D is separated from C by a line where $T = T_{DEP}(G)$ (or $G = G_{DEP}(T)$) and corresponds to a situation in which diffusivity of wax vanishes. Thus, zone C is called the gel zone and zone D is called the deposit.

In [47] a model problem in one space dimension is studied with initial and boundary conditions as in (46)–(49). Thermal field is computed such that $T_t < 0$, $T_x > 0$, $T_{xt} > 0$.

Since the case $D_G = 0$ is considered, the problem turns out to be a hyperbolic-parabolic free boundary problem. Indeed the evolution of G in the saturated zone and in the gel is governed by

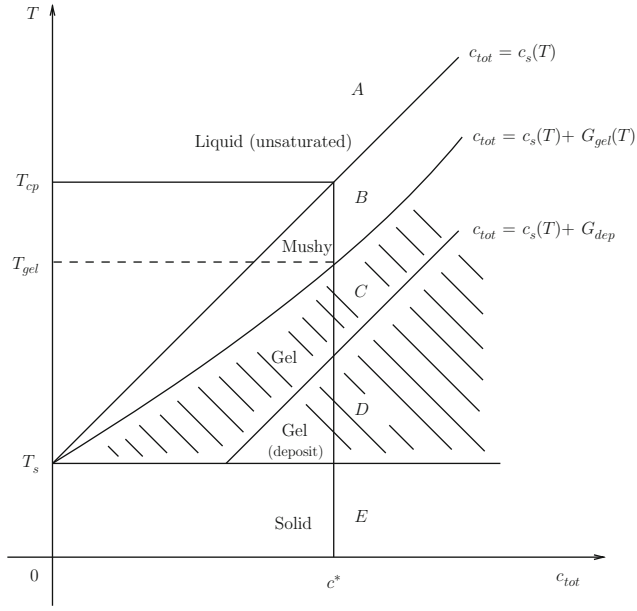


Fig. 2. Phase diagram for oil-wax mixtures

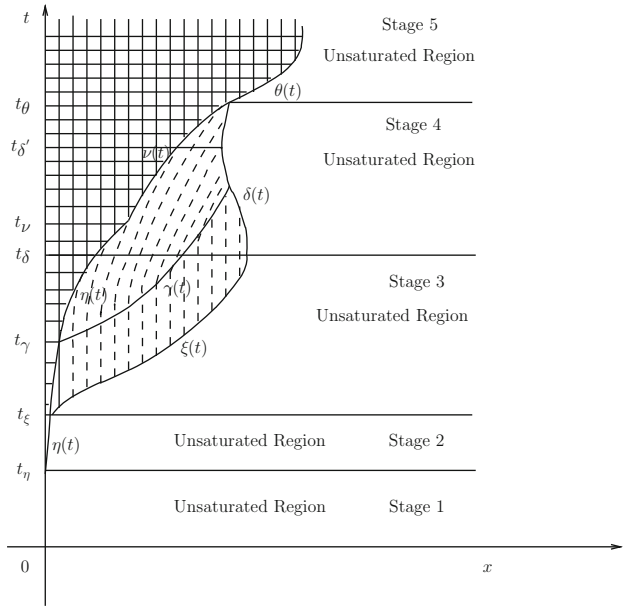


Fig. 3. Qualitative behaviour of the solution

$$\frac{\partial G}{\partial t} = \lambda \frac{\partial T}{\partial t} \tag{57}$$

if one assumes, for simplicity that c_{SAT} is a linear function of T.

The problem exhibits several free boundaries and its analysis is rather delicate; results on well-posedness in classical sense and on characterization of the free boundaries can be found in [47]. As expected (see Fig. 3 where the qualitative behaviour of the solution is illustrated), the saturated region disappears in the long run and only an unsaturated region and the deposit are eventually present.

8 Thermal Equilibrium with Crystallization Kinetics

In this section we will consider situations in which phase transition is not assumed to occur instantaneously, i.e. we allow t_2 to be of the same order of t_3. Hence we assume

$$t_1 \ll t_2,\ t_3,\ t_4,\ t_5 \ll t_6 \tag{58}$$

and we consider both the macroscopic and the microscopic description of the crystallization kinetics.

8.1 A Problem with Uniform Temperature

This situation is elementary from the point of view of its mathematical description, but provides a useful insight to interpret the experimental results.

Assume the mixture with wax concentration c^* is initially at a uniform temperature $T_0 > T_{CL}$. Assume that the boundary of the domain Ω occupied by the mixture has a prescribed time dependent temperature, e.g.

$$T(\mathbf{x}, t) = T_0 - \lambda t, \quad \mathbf{x} \in \partial\Omega,\ t > 0, \tag{59}$$

for some $\lambda > 0$. Since we are assuming that t_1 is negligible w.r.t. the time scale of the experiment and we disregard the effect of latent heat of crystallization (see Sect. 2), we have

$$T(\mathbf{x}, t) = T_0 - \lambda t, \quad \mathbf{x} \in \Omega,\ t > 0. \tag{60}$$

Therefore, starting from the time t^* such that

$$T_0 - \lambda t^* = T_{CL}(c^*)$$

we will have that the segregation (and, eventually, gelification) starts and $G(\mathbf{x}, t)$ will become positive and increasing (and independent on t) for $t > t^*$.

Its time evolution can be described by (4) or (12) according to the point of view we want to assume. In both cases we have one parameter to fit the experimental data (or two if we take (13)).

The results of numerical simulations [48] and the comparison with experimental data, show that it is quite difficult to discriminate between the two models. Investigation in this sense is still going on.

8.2 A Problem with Constant Thermal Gradient

This problem has been studied in [49] and in [46]. In both papers the kinetics of crystallization is described by a macroscopic equation. We also note that in [46] gelification is not taken into account (and thus, in (58), we would have $t_5 \gg t_2,\ t_3,\ t_4$).

The temperature is stationary and, referring once more to planar symmetry we will write

$$T(x, t) = a + bx, \quad 0 < x < l,\ t > 0. \tag{61}$$

We also assume that both $c_{SAT}(T)$ and $c_{GEL}(T)$ are linear. Therefore, they can be written as two linear function of x that we denote by $c_1(x)$ and $c_2(x)$ respectively:

$$c_1(x) = a_1 + b_1 x, \quad 0 < x < l, \tag{62}$$
$$c_2(x) = a_2 + b_2 x, \quad 0 < x < l, \tag{63}$$

and the positive constants a_i, b_i $(i = 1, 2)$ are such that

$$c_1(x) < c_2(x), \quad 0 < x < l. \tag{64}$$

Recalling (4)–(6) and neglecting diffusion of the segregated phase, we have

$$\frac{\partial G}{\partial t} = H(G + (C - c_{SAT})^+)\beta[C(x,t) - c_1(x)], \quad 0 < x < l,\ t > 0. \tag{65}$$

As we noted in Sect. 2, dependence of β on $T(x)$ could also be taken into account.

Now, let us turn our attention to $C(x,t)$. In [49] it is postulated that diffusivity of dissolved wax jumps to zero in the gelified part. Hence C will have to satisfy, in a suitable weak sense, the following equation

$$C_t - [DH(c_2 - c)C_{xx}] = -G_t, \quad 0 < x < l,\ t > 0, \tag{66}$$

where

$$c(x,t) = C(x,t) + G(x,t), \quad 0 < x < l,\ t > 0. \tag{67}$$

Moreover

$$C_x(0,t) = 0, \quad t > 0, \tag{68}$$

$$C_x(l,t) = 0 \quad t > 0, \tag{69}$$

$$C(x,0) = c_1(x), \quad 0 < x < l, \tag{70}$$

$$G(x,0) = c^* - c_1(x), \quad 0 < x < l, \tag{71}$$

assuming that, for $t = 0$, the mixture is everywhere saturated ($c^* > c_{SAT}(T(x,0))$).

It can be proved, [49], that

A. There exist two Lipschitz continuous functions $s(t)$, $\sigma(t)$ such that the half strip $K = (0,l) \times (0,+\infty)$ in the (x,t) plane is partitioned in three regions:
   (1) The gel region $\mathcal{G} = \{(x,t) : 0 < x < s(t),\ t > t_g,\ s(t_g) = 0\}$.
   (2) The undersaturated region $\mathcal{U} = \{(x,t) : \sigma(t) < x < l,\ t > t_u,\ \sigma(t_u) = l\}$.
   (3) The saturated region $\mathcal{S} \equiv K \setminus (\mathcal{U} \cup \mathcal{G})$.

B. In region $\mathcal{G}$ (no mass transfer).

$$c = c_2(x) \tag{72}$$

C. In region $\mathcal{S}$, c and $G \in C2,1(\mathcal{S})$, and C, C_x, $G \in C(\bar{\mathcal{S}})$ and satisfy the differential equations (65) and (66) with initial conditions (70) and (71), while boundary conditions (68) and (69) are fulfilled for $t < t_g$ and t_u respectively.

D. In region $\mathcal{U}$, $G = 0$, $C \in C^{2,1}(\mathcal{U})$ and $u,\ u_x \in C(\bar{\mathcal{U}})$. Moreover

$$C_t - DC_{xx} = 0 \quad (x,t) \in \mathcal{U}, \tag{73}$$

E. On $x = \sigma(t)$ it is

$$G(\sigma(t),t) = 0, \quad t > t_u, \tag{74}$$

$$[C]^+_- = [C_x]^+_- = 0, \quad t > t_u, \tag{75}$$

where by $[\,]^+_-$ we denote the jump of the quantity in bracket across the curve $x = \sigma(t)$.

F. On $x = s(t)$ it is

$$C(s(t),t) + G(s(t),t) = c_2(s(t)), \quad t > t_g, \tag{76}$$

$$C_x(s(t),t) = 0, \quad t > t_g. \tag{77}$$

Assuming the microscopic description of the crystallization process simply consists in substituting (12) or (13)–(65). The results are quite similar as in the case of uniform temperature.

9 Variable Thermal Fields and Crystallization Kinetics: Is Diffusion Relevant?

When the domain occupied by the mixture is not "thin" and the waxes that are contained in the oil are "heavy" enough, neither t_1 nor t_2 can be thought to be negligible with respect to the time scale of the experiment. On the other hand, some authors tend to disregard the influence of diffusion in the interpretation of experimental results.

In any case, we still retain the assumption that the thermal field can be determined independently of the knowledge of $C(x,t)$, $G(x,t)$.

Referring to the experimental situation, we consider a cylinder of radius R, containing a mass M of oil (per unit axial length) at concentration c^*. The initial temperature T_0 is larger than $T_{CL}(c^*)$ while the surface of the cylinder is maintained, for any $t > 0$ at a temperature T_{EXT} that is less than $T_{CL}(c^*)$.

The thermal field is the solution of the following parabolic problem

$$\begin{cases} \frac{\partial T}{\partial t} = D\left\{\frac{\partial^2 T}{\partial r^2} + \frac{1}{r}\frac{\partial T}{\partial r}\right\}, & 0 < r < R,\ t > 0 \\ T(r,0) = T_0, & 0 < r < R \\ \frac{\partial T}{\partial r}(0,t) = 0, & t > 0 \\ T(R,t) = T_{EXT}, & t > 0, \end{cases} \tag{78}$$

that can be expressed in series of Bessel functions (see e.g. [50, Chap. VII]).

Neglecting diffusion, we can compute $G(x,t)$ according to the macroscopic and microscopic crystallization kinetics (as well as to the assumption of phase equilibrium (3)). The corresponding simulations seem to show that a model

that does not take diffusion (at least of the dissolved wax) into account cannot interpret the experimental results. Once again, it is difficult to discriminate between macroscopic and microscopic crystallization kinetics, unless we confine to the initial stage of the phenomenon.

10 Deposition in Moving Mixtures

As we said in Sect. 1, the final aim of the research is a descriptive and predictive model for wax deposition on the walls of pipelines. There exists a huge literature on this subject (see e.g. [51–59]) aimed at interpreting the results on experimental loops and field experiments.

At the present stage of our research program, we can claim that our model can actually be used in a quite large class of "field" conditions, i.e. in cases in which quantities like temperature, mean axial velocity, wax concentration can be thought to be independent of the radial coordinate r within a bulk core $0 < r < R - \delta$, δ being the thickness of a boundary layer.

This problem has been studied in [60]. The basic assumption is that the flow in the pipeline is turbulent and that molecular diffusion is a thin boundary layer is the only mechanism responsible for deposition.

The model includes the mechanism of ablation that has the effect of limiting the thickness of the deposit.

Just to give a rough idea of the model, the starting point is to write down the quasi-steady profile of the temperature in the boundary layer

$$T(r,z,t) = -a(z,t)(R-\delta)\ln\left(\frac{r}{R-\delta}\right)+T_c(z,t),\quad R-\delta < r < R,\ z > 0,\ t > 0 \tag{79}$$

In (79) $a(z,t)$ is a coefficient that has to be determined and T_c is the temperature of the bulk of the fluids; in general the analysis may take into account the variation of δ with z and t that will be ignored here to simplify the discussion.

Imposing thermal balance allows us to find $a(z,t)$ that represents the thermal gradient in the boundary layer as a function of z and t. Since the model applies to oils for which the characteristic time of crystallization is negligible w.r.t. the time scale of the experiment, the thickness of the deposit can be found applying the techniques we used in Sect. 5. As long as the "core" remains saturated, an explicit approximated formula for the thickness of the deposit can be found

$$\delta = \frac{Rt}{T_{CL}t_0}\frac{T_0 - T_l}{\mu}\exp\left[-\frac{2\pi D_T}{\mu Q}z\right]H(T_{CL} - T_W)^6 \tag{80}$$

[6] In (80), we did not write the ablation term.

where Q is the volumetric flow rate, T_l is the temperature of the surroundings and

$$\mu = \frac{k}{hR} \tag{81}$$

where k is the thermal conductivity of the mixture and h the heat transfer coefficient between the pipe and the surroundings.

Since $T_W(z)$ is the temperature of the wall and H is the Heaviside function, it is clear that the line will have a "deposit free zone" z_F that can be easily calculated.

For the case in which the core may desaturate, the analysis has some additional difficulties and we refer the reader to the original paper.

As a conclusion, we report that the model is consistent with field experiments made on two different pipelines, where the discrepancy between measured and calculated quantities is below 10%.

References

1. Azevedo, L.F.A., Teixeira, A.M.: A critical review of the modeling of wax deposition mechanism. Petrol. Sci. Technol. **21**(3&4), 393–408 (2003)
2. Faienza, L., Fusi, L.: The literature on waxy crude oils. Tech. Rep. Dept. U. Dini (2008)
3. Escobar-Remolina, J.C.M.: Characteristics of wax precipitation in synthetic mixtures and fluids of petroleum. Fluid Phase Equilib. **240**, 197–203 (2006)
4. Wu, C.H., Creek, J.L., Wang, K.S., Carlson, R.M., Cheung, S., Shuler, P.J., Tang, Y.: Measurements of wax deposition in paraffin solutions. AIChE Proceedings of the 4th International Symposium on wax Thermodynamics and Deposition (2002)
5. Zougary, M.I., Sopkow, T.: Introduction to Crude Oil Wax Crystallization Kinetics: Process Modeling. Ind. Eng. Chem. Res. **46**, 1360–1368 (2007)
6. Correra, S., Fasano, A., Fusi, L., Merino-Garcia, D.: Calculating deposit formation in the pipelining of waxy crude oils. Meccanica. **42**, 149–165 (2007)
7. Correra, S., Fasano, A., Fusi, L., Primicerio, M.: Modeling wax diffusion in crude oils: the cold finger device. Appl. Math. Model.**31**, 2286–2298 (2006)
8. Correra, S., Fasano, A., Fusi, L., Primicerio, M., Rosso, F.: Wax Diffusivity under given thermal gradient: a mathematical model. ZAMM. **87**, 24–36 (2007)
9. Garcia, D.M.: Progress in cold flow I. Enitecnologie Techincal Report (2004)
10. Hammami, A., Ratulowski, J., Coutinho, J.A.P.: Cloud Points and can we measure or model them. Petrol. Sci. Technol. **21**(3&4), 345–358 (2003)
11. Vignati, E., Piazza, R., Visintin, R.F.G., Lapasin, R., D'Antona, P., Lockhart, T.P.: Wax crystallization and aggregation in a model crude oil. Eni Technical Report (2006)
12. Coutinho, J.A.P., Knudsen, K., Andersen, S.I.: A local composition model for paraffinic solis solution. Chem. Eng. Sci. **51**(12), 3273–3282 (1996)
13. Coutinho, J.A.P.: Predictive UNIQUAC: a new model for the description of multiphase Solid-Liquid Equilibria in Complex Hydrocarbon Mixtures. Ind. Eng. Chem. Res. **37**, 4870–4875 (1998)

14. Hansen, J.H., Fredenshund, A., Pedersen, K.S., Ronnigsen, H.P.: A Thermodynamic model for predicting wax formation in crude oils. AIChE J. **34**(12), 1937–1942 (1988)
15. Ji, H.-Y., Tohidi, B., Told, A.: Wax phase equilibria: developing a thermodynamic model using a systematic approach. Fluid Phase Equilib. **216** (2004)
16. Lira-Galeana, C., Firoozabadi, C., Prausnitz, J.M.: Thermodynamics of wax precipitation in petroleum mixtures. AIChE J. **42**(1) (1996)
17. Pan, H., Firoozabadi, A.: Pressure and composition effect on wax precipitation: experimental data and model results. SPE Prod. Facil. 250–259 (1997).
18. Prausnitz, J.M., Lichtenthaler, R.N., Azevedo, E.G.: Molecular Thermodynamics of Fluid Phase Equilibria. Prentice-Hall, New Jersey (1986)
19. Weintgarten, J.S., Euncher, J.A.: Methods for predicting wax precipitation and deposition. SPE Prod. Eng. 121–126 (1988)
20. Won, K.W.: Thermodynamics for solid solution-liquid-vapor-equilibria: wax phase formation from heavy hydrocarbon mixtures. Fluid Phase Equilib. **30** (1986)
21. Won, K.W.: Thermodynamic calculation of cloud point temperatures and wax phase composition of refined hydrocarbon mixture. Fluid Phase Equilib., **53** (1989)
22. Zuo, J.Y., Zhang, D.D., Ng, H.J.: An improved thermodynamic model for precipitation from petroleum fluids. Chem. Eng. J. **56** (2001)
23. Coutinho, J.A.P., Daridon, J.-L.: The limitations of the cloud point measurement techniques and the influence of the oil composition and its detection. Petrol. Sci. Technol. **23**, 1113–1128 (2005)
24. Coutinho, J.A.P., Edmonds, B., Moorwood, T., Szczepanski, R., Zhang, X.: Reliable wax predictions for flow assurance. Energy Fuels. **20**, 1081–1088 (2006)
25. Andreucci, D., Fasano, A., Primicerio, M.: On a mathematical model for the crystallization of polymers. In: Wacker, H.J. et al. (eds.) Proceedings of ECMI 4, pp. 3–16. Kluwer, Dordrecht (1991)
26. Berger, J., Schneider, W.: A zone model of rate controlled solidification. Plast. Rubber Process. Appl. **6**, 127–133 (1986)
27. Andreucci, D., Primicerio, M., Borrelli, L., Capasso, V., Li, P.: Polymer crystallization kinetics: the effect of impingement. Math. Eng. Ind. **4**, 249–263 (1993)
28. Malkin, A.Ya., Beghisev, V.P., Keapin, I.A., Bolgov, S.A.: General treatment of polymer crystallization kinetics I. Polym. Eng. Sci. **24**, 1396–1401 (1984)
29. Rabesiaka, J., Kovacs, A.I.: Isothermal crystallization kinetics of polyethylene III. J. Appl. Phys. **32**, 2314–2320 (1961)
30. Tobin, M.: Theory of phase transition kinetics with growth site impingement I. J. Polym. Sci. **14**, 399–406 (1974).
31. Hammami, A., Mehrotra, A.K.: Non-isothermal crystallization kinetics of n-paraffins with chain lenght between thirty and fifty. Thermochim Acta.**211**, 137–153 (1992)
32. Hammami, A., Mehrotra, A.K.: Non-isothermal crystallization kinetics of even-numbered and odd-numbered normal alkanes. Thermochim Acta.**215**, 197–209 (1993)
33. Ozawa, T.: Kinetics of non-isothermal crystallization. Polymer. **12**, 150–158 (1971)

34. Avrami, M.: Kinetics of phase change I: general theory. J. Chem. Phys. **7** (1939)
35. Correra, S., Andrei, M., Carniani, C.: Wax diffusivity: is it a physical property or a pivotal parameter? Petrol. Sci. Technol. **21**(9), 1539–1554 (2003)
36. Garcia, D.M., Correra, S., Margarone, M.: Kinetics of waxy gel formation from batch experiments. 7th Conference on Petroleum Phase Behaviour and Fouling, Asheville (2006)
37. Gianni, R., Petrova, A.G.: One dimensional problem for heat and mass transport in oil-wax solution. Rend. Mat. Accad. Lincei. **9**(16), 181–196 (2005)
38. Garcia, D.M.: Wax cold flow: report II. Enitecnologie Technical Report (2005)
39. Paso, K.G., Fogler, H.S.: Influence of n-paraffin composition on the aging of wax-oil gel deposits. AIChE J. **49** (2003)
40. Paso, K.G., Fogler, H.S.: Bulk stabilization in wax deposition systems. Energy Fuels. **18**(4), 1005–1013 (2004)
41. Fasano, A., Primicerio, M.: Temperature driven mass transport in concentrated saturated solutions. Prog. Nonlinear Differ. Equ. Appl. **61**, 91–108 (2005)
42. Cannon, J.R., Fasano, A.: Boundary value multidimensional problems in fast chemical reactions. Arch. Rat. Mech. Anal. **53**, 1–13 (1973)
43. Ladyszenskya, O.A., Solonnikov, V.A., Ural'ceva, N.N.: Linear and Quasi-linear Equations of Parabolic Type. America Mathematical Society, Providence (1968)
44. Comparini, E., Talamucci, F.: A general model for wax diffusion in crude oils under thermal gradient. In: Cutello, V. et al. (eds.) Applied and Industrial Mathematics in Italy, pp. 259–270. World Scientific, Singapore (2007)
45. Fasano, A., Primicerio, M.: Heat and mass transfer in non-isothermal partially saturated solutions. In: Fergola, P. et al. (eds.) New Trends in Mathematical Physics, pp. 34–44. World Scientific, Singapore (2003)
46. Fasano, A., Primicerio, M.: Wax deposition in crude oils: a new approach. Rend. Mat. Accad. Lincei. **9**, 251–263 (2005)
47. Fasano, A., Fusi, L., Primicerio, M., Ockendon, J.: Gelification and mass transport in static non-isothermal waxy solution, Eur. J. Appl. Math. **20**, 93–122 (2009)
48. Faienza, L.: Numerical simulations for deposition in paraffine-decane mixtures. Tech. Rep. Dept. U. Dini (2008)
49. Faienza, L., Fasano, A., Primicerio, M.: Gelification of hydrocarbons, a model problem. In Jeltsch, R. et al. (eds.) Some Topics in Industrial Mathematics, pp. 120–133. World Scientific, Singapore (2007)
50. Carslaw, H.S., Jaeger, J.C.: Conduction of heat in solids. Clarendon, Oxford (1959).
51. Bath, N., Mehrotra, A.K.: Modeling of deposition from "waxy" mixtures in a pipeline via moving boundary formulation. Ind. Eng. Chem. Res. **44** (2005)
52. Bern, P.A., Winthers, V.R., Cairns, J.: Wax deposition in pipelines. European Offshore Petroleum Conference & Exhibition, pp. 21–24. London (1980)
53. Creek, J.L., Lund, H.J., Brill, J.P., Volk, M.: Wax deposition in a single phase flow. Fluid Phase Equilib. **158**(1), 801–811 (1999)
54. Fong, N., Mehrotra, A.K.: Deposition under turbulent flow of wax-solvent mixtures in a bench-scale flow-loop apparatus with heat transfer. Energy Fuels. **21**(3), 1263–1276 (2007)
55. Hsu, J.J.C., Santamaria, M.M., Brubaker, J.P.: Wax deposition of waxy live crudes under turbulent flow conditions. SPE 28480
56. Kok, M.V., Saracoglu, O.: Mathematical modelling of wax deposition in crude oil pipeline system. SPE. 1–7 (2007)

57. Nazar, S.A.R., Dabir, B., Vaziri, H., Islam, M.R.: Experimental and mathematical modeling of wax deposition and propagation in pipes transporting crude oil. SPE. 67328, 239–248 (2001)
58. Singh, P., Fogler, H.S., Nagorajan, N.: Prediction of wax content in a pipeline: an application of the controlled-stress rheometer. J. Rheol. **43**(6), (1999)
59. Svendsen, J.A.: Mathematical modeling of wax deposition in oil pipeline system. AIChE J. **39**(8), 1377–1388 (1993)
60. Fasano, A., Fusi, L., Correra, S.: Mathematical models for waxy crude oils. Meccanica. **39**, 441–483 (2004)

Chebfun: A New Kind of Numerical Computing

R.B. Platte[1] and L.N. Trefethen[2]

[1] School of Mathematical and Statistical Sciences, Arizona State University, Tempe, AZ 85287-1804, USA, platte@math.asu.edu
[2] Oxford University Mathematical Institute, Oxford OX1 3LB, UK trefethen@maths.ox.ac.uk

Summary. The functionalities of the chebfun and chebop systems are surveyed. The chebfun system is a collection of MATLAB codes to manipulate functions in a manner that resembles symbolic computing. The operations, however, are performed numerically using polynomial representations. Chebops are built with the aid of chebfuns to represent linear operators and allow chebfun solutions of differential equations. In this article we present examples to illustrate the simplicity and effectiveness of the software. Among other problems, we consider edge detection in logistic map functions and the solution of linear and nonlinear differential equations.

1 Introduction

For a long time there have been two kinds of mathematical computation: symbolic and numerical. Symbolic computing manipulates algebraic expressions exactly, but it is unworkable for many applications since the space and time requirements tend to grow combinatorially. Numerical computing avoids the combinatorial explosion by rounding to 16 digits at each step, but it works just with individual numbers, not algebraic expressions.

The chebfun system introduced in 2004 by Battles and Trefethen [1] aims to combine the feel of symbolics with the speed of numerics. The idea is to represent functions by Chebyshev expansions whose length is determined adaptively to maintain an accuracy of close to machine precision. The system has been significantly extended since its introduction. Among other developments, it now handles piecewise smooth functions on arbitrary intervals [2] and linear operators [3]. The latter extension was made possible by T. A. Driscoll who implemented the *chebop* class. In this article, we review the main features of the software and demonstrate its effectiveness through many examples, including solution of differential equations.

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*, Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_5,

The chebfun system is implemented in object-oriented MATLAB. One of the guiding principles in its design is the analogy of commands available for vectors and those implemented in the chebfun package for functions. Once a chebfun object has been created, commands like `diff`, `sum` and `norm` can be used to compute its derivative, definite integral, and norm, respectively.

The simplicity of its use is illustrated in the following example, where the number of roots, maximum and L_1-norm of the function $f(x) = J_0(x)\sin x$ are computed in the interval $[0, 100]$.

```
>> f = chebfun(@(x) besselj(0,x).*sin(x), [0 100]);
>> length(roots(f))
ans = 64
>> max(f)
ans = 0.644562281456927
>> norm(f,1)
ans = 6.295294435933753
```

Similarly, the chebop extension to linear operators relies on underlying polynomial-based spectral methods. The analogy here, to some extent, is between linear operators and matrices. With chebops, commands such as `diff` and `sum` are used to define differential and integral operators, while "`*`" and "`\`" are used to apply operators in forward and inverse modes. The following commands, for example, can be used to differentiate $f(x) = \sin(x)$ in $[-\pi, \pi]$ using chebop notation.

```
[d,x] = domain([-pi,pi]);
D = diff(d);
df = D*sin(x);
```

One of the main strengths of chebops is how simple the syntax is for solving differential equations. To solve the boundary value problem

$$u''(x) + u'(x) + u(x) = \sin(x), \quad x \in (-\pi, \pi), \quad u(\pm\pi) = 0,$$

for instance, one only needs to define the operator and appropriate boundary conditions and type `\`,

```
L = D^2 + D + eye(d) & 'dirichlet';
sol = L\sin(x);
```

This article is organized in two main sections. In Sect. 2 we review basic aspects of the chebfun system, including piecewise polynomial representations and apply the chebfun edge detector to locate break points of piecewise constant functions that are limits of logistic map sequences. In Sect. 3 we briefly describe the syntax of the chebop system and give examples to illustrate how simple and effective it is.

2 Chebfuns

In this section we give some insight into the underlying theory and implementation of the system. More detailed information can be found in [1] and [4].

2.1 Funs: Smooth Representations on $[-1, 1]$

The original chebfun class implemented by Battles in 2004 for smooth functions on $[-1, 1]$ is now called *fun*. A chebfun object consists of one or more funs. Each smooth piece is mapped to the interval $[-1, 1]$ and represented by an expansion in Chebyshev polynomials of the form

$$f_N(x) = \sum_{j=0}^{N} \lambda_j T_j(x), \qquad x \in [-1, 1], \tag{1}$$

where $T_j(x) = \cos(j \arccos(x))$. When constructing a fun object, the system computes the coefficients λ_j by interpolating the target function f at $N + 1$ Chebyshev extreme points,

$$x_j = \cos \frac{\pi j}{N}, \qquad j = 0, \ldots, N.$$

The polynomial degree N is automatically determined so that the representation is as accurate as possible in double precision arithmetic.

Polynomial interpolation in Chebyshev nodes is known to be near-optimal for approximating functions that are smooth, converging geometrically for analytic functions [1]. Fast Fourier transforms (FFTs) can be used to map function values $f(x_j)$ to coefficients λ_j, and vice versa, in $O(N \log N)$ operations. Figure 1 presents the polynomial representation of the Bessel function J_0 and its corresponding Chebyshev coefficients. The construction process begins by sampling the target function at $2^n + 1$ points, with $n = 3, 4, \ldots$. The optimal degree N is then determined such that $|\lambda_j|$ is close to zero, relative to the coefficient of largest magnitude, for all $j > N$.

The left graph of Fig. 1 was obtained with the following commands:

```
f = chebfun(@(x) besselj(0,x),[0 100]); plot(f,'.-')
```

and the coefficients were plotted using

```
c = chebpoly(f); semilogy(flipud(abs(c)),'.')
```

The execution of the first command constructs the chebfun object from an anonymous function evaluated in the specified interval. Once a chebfun object has been created, there are a number of methods that can be used to operate on it. The list of methods can be obtained by typing `methods chebfun`. The syntax is, in most cases, the same as the usual MATLAB calls for vectors. The integral of f from 0 to 100, for instance, is obtained with the command `sum`.

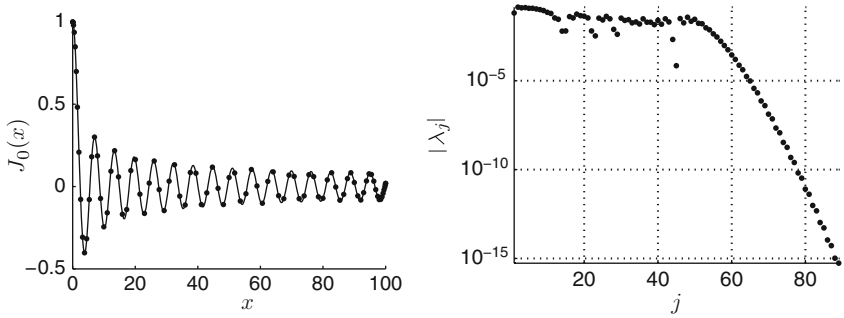


Fig. 1. *Left*: Chebfun representation by a polynomial of degree 88 of the Bessel function J_0 on the interval $[0, 100]$. The dots mark the 89 Chebyshev interpolation points. *Right*: semilog plot of the magnitude of the corresponding Chebyshev coefficients

```
>> sum(f)
ans = 0.922662556960163
```

All digits in this answer are correct except the last one. Integrals are computed efficiently by Clenshaw–Curtis quadrature in $O(N)$ operations once the coefficients are obtained with the aid of the FFT. Similarly, `cumsum(f)` returns the indefinite integral of the chebfun `f`.

Rootfinding plays a key role in the chebfun system. The method we use makes use of a recursion proposed by Boyd [5]. The main idea behind this approach is that the roots of a polynomial of the form (1) are the eigenvalues of an $N \times N$ *colleague matrix* [6]. To avoid the cubic growth of the number of operations required by eigenvalue computations, the algorithm uses recursive subdivision of intervals to bring the degree of the polynomial representation to at most 100, improving the overall operation count to $O(N^2)$.

Here is an example where rootfinding is used to obtain all local maxima of `f`.

```
df = diff(f);
xcrit = roots(df);
ddf = diff(df);
xmax = xcrit(ddf(xcrit)<0);
plot(f), hold on, plot(xmax,f(xmax),'o')
```

The result is displayed in Fig. 2. Also shown in this figure is the global minimum, which is computed in a similar way with just one function call: `[ymin,xmin]=min(f); plot(xmin,ymin,'*')`.

The evaluation of a chebfun at arbitrary points is carried out using the barycentric formula introduced by Salzer [7, 8]. The formula has been proved to be stable by Higham in [9] and requires $O(MN)$ operations to evaluate a chebfun at M points. The `plot` command, for instance, relies on evaluations at thousands of points.

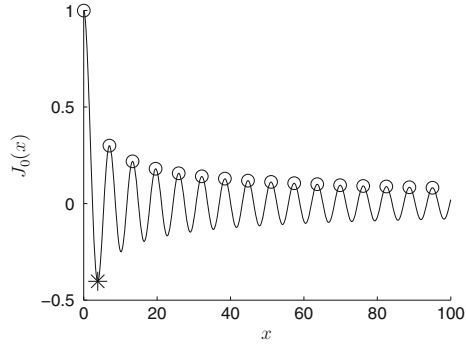


Fig. 2. Local maxima (*opencircle*) and global minimum (*asterisk*) of J_0 in [0,100]

2.2 Piecewise Representations

The chebfun system also handles piecewise smooth functions [2]. Piecewise representations can result from certain operations on smooth functions such as

```
abs, sign, floor, ceil, round, fix, min, max
```

among others. They may also be defined using the chebfun constructor. In the construction precess, each smooth piece may be explicitly defined or obtained through an edge detection procedure.

The main components of a chebfun with several pieces are the endpoints of the interval, the breakpoints, and the corresponding *funs*, which are objects representing each smooth piece. When breakpoints are introduced by operations on chebfuns, they are, in most cases, obtained by rootfinding. In the following code segment, for instance,

```
>> x = chebfun(@(x) x);
>> f = sin(4*x.^2).*floor(1.5*sin(5*x));
>> norm(f,1)
ans = 0.936713707137759
```

zerofinding is used twice. To find the breakpoints of the piecewise constant chebfun `floor(1.5*sin(5*x))`, the system finds all values of x that satisfy $1.5\sin(5x) - n = 0$ for $n = -1, 0, 1$. To compute the L_1 norm of f, it first obtains a piecewise representation of $|f|$, which also requires rootfinding.

Chebfun also comes with an efficient edge detector, since in many situations, one may want to construct a representation from samples of a function. To this end, the constructor works in two splitting modes that may be selected by the user: `splitting on` and `splitting off` – the current default is `on`. The following example illustrates the edge detector in action:

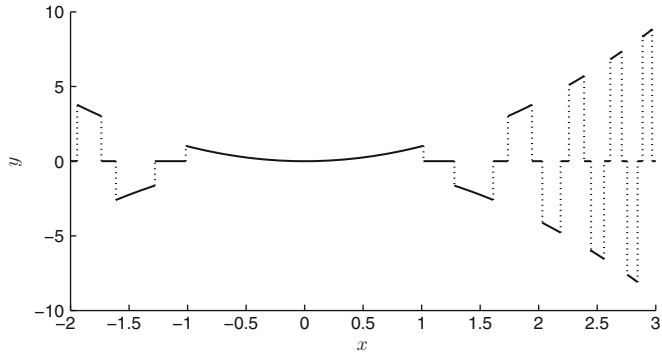


Fig. 3. Plot of the chebfun corresponding to $f(x) = x^2 \text{ round}(\cos x^3)$

```
splitting on
f = chebfun(@(x) x.^2.*round(cos(x.^3)),[-2 3]);
plot(f)
```

The result is shown in Fig. 3. The breakpoints are stored in the field `f.ends`. The edge detection algorithm uses bisection and finite differences to locate jumps in function values accurately to machine precision, as well as jumps in first, second and third derivatives [2].

In `splitting off` mode, the system disables the splitting algorithm. This mode is recommended when the target functions are smooth since in such cases manipulating global approximations is often more efficient. Most operations in the chebop system are restricted to this mode.

The Logistic Map

Simple examples of piecewise smooth functions arise throughout applied mathematics and are easily manipulated in the chebfun system. For one set of examples, see the online chebfun guide [10]. Here, we shall push the system harder with a more challenging example. Many chebfun computations finish in a fraction of a second; the results we shall show have taken minutes.

We use the logistic map to illustrate some strengths and limitations of piecewise polynomial representations. The map is given by the recurrence formula

$$x_{k+1} = r x_k (1 - x_k), \tag{2}$$

with $x_k \in [0, 1]$ and $r \in [0, 4]$, and is often used to model simple population dynamics and to illustrate key properties of dynamical systems such as chaos. The bifurcation diagram for the logistic map is shown in Fig. 4.

Suppose we are interested in representing the map functions, $g_r^k : x_0 \mapsto x_k$, and studying their convergence. For $r = 4$, it is possible to derive a simple polynomial representation [11],

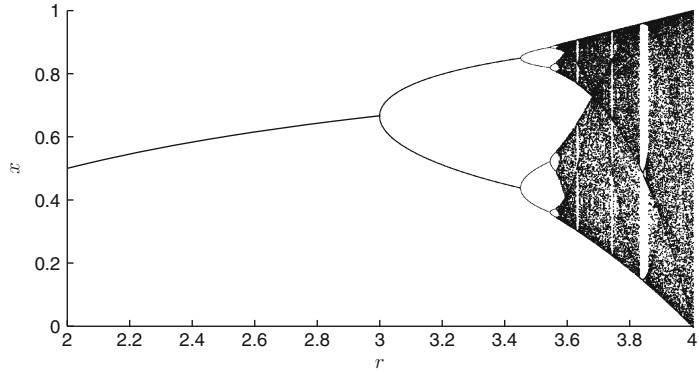


Fig. 4. The famous bifurcation diagram for the logistic map, showing period doubling as a route to chaos

$$g_4^k(x_0) = \frac{1 - \cos(2^k \arccos(1 - 2x_0))}{2},$$

but in general, nonrecursive expressions are not available. The maps g_r^k are polynomials of order 2^k, but their chebfun representations often have smaller degrees. For $1 < r < 3$, the functions g_r^k converge to a constant, $1 - 1/r$, if we exclude the singular endpoints: $x = 0$ and $x = 1$. Here are the degrees of chebfuns for $g_{2.5}^k$:

```
>> xk = chebfun(@(x) x, [0.001 .999]);
>> for k = 1:51
       xk = 2.5*xk.*(1-xk); deg(k) = length(xk)-1;
   end
>> deg
deg =
  Columns 1 through 9
     2     4     8    16    32    64   112   178   284
  Columns 10 through 18
   434   574   544   554   522   522   522   496   488
  Columns 19 through 27
   488   474   470   390   390   388   388   354   352
  Columns 28 through 36
   352   352   338   330   330   258   258   256   256
  Columns 37 through 45
   158   158   158   158   158   106   106   106   106
  Columns 46 through 51
   106    72    72     0     0     0
```

Despite the initial exponential growth in degree, the length of the chebfuns reaches a maximum of 574, and for $k \geq 49$, the chebfun representations of $g_{2.5}^k$

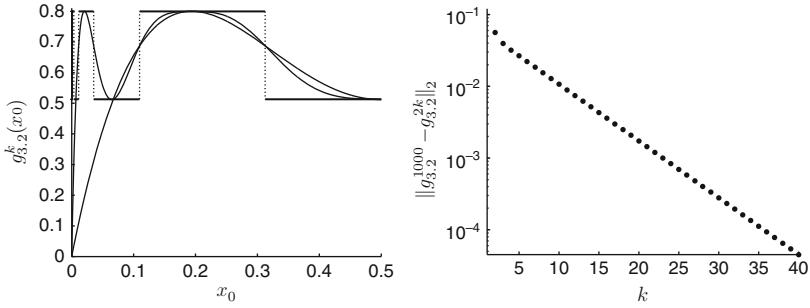


Fig. 5. *Left*: piecewise chebfun representation of $g^k_{3.2}$, $k = 2, 4$ and 1,000. *Right*: convergence plot of $g^{2k}_{3.2}$ in the L_2-norm

are constant functions with the correct value 0.6. In the exact arithmetic of symbolic computing, for $k = 49$ the degree would be 562,949,953,421,312.

It is also interesting to look at the convergence of these functions in the two-cycle region, $3 < r < 3.44\ldots$, where the subsequences $\{g_r^{2k-1}\}$ and $\{g_r^{2k}\}$ converge to piecewise constant functions. With the aid of the chebfun automatic edge detection algorithm, we can represent these limiting functions and compute the rates for convergence as follows for $r = 3.2$:

```
g1000 = chebfun(@(x) logistic(3.2,1000,x), [0 0.5]);
xk = chebfun(@(x) x, domain(g1000));
delta = zeros(40,1);
for k = 1:80
   xk = 3.2*xk.*(1-xk);
   if mod(k,2)==0, delta(k/2) = norm(xk-g1000); end
end
plot(g1000), figure, semilogy(delta,'.')
```

The result is shown in Fig. 5 together with the graphs of $g^2_{3.2}$ and $g^4_{3.2}$. The first line of the execution above requires the function `logistic.m`:

```
function x = logistic(r,n,x)
     for k=1:n, x = r*x.*(1-x); end
```

Notice that the functions g_r^k are symmetric about $x = 0.5$, so in the example above we only considered the interval $[0, .5]$. The subsequences $\{g_r^{2k-1}\}$ and $\{g_r^{2k}\}$ cannot converge uniformly because of the jumps in the limit. In the (default) L_2-norm, on the other hand, they converge very fast. The right plot in Fig. 5 indicates exponential convergence. We point out that the chebfun representation of $g^{1000}_{3.2}$ has 31 break points, most of them near $x = 0$, with the spaces between them decaying exponentially.

A similar cascade of break points can be observed in the 4-cycle region. In fact, as the parameter r is increased, the number of jump locations also

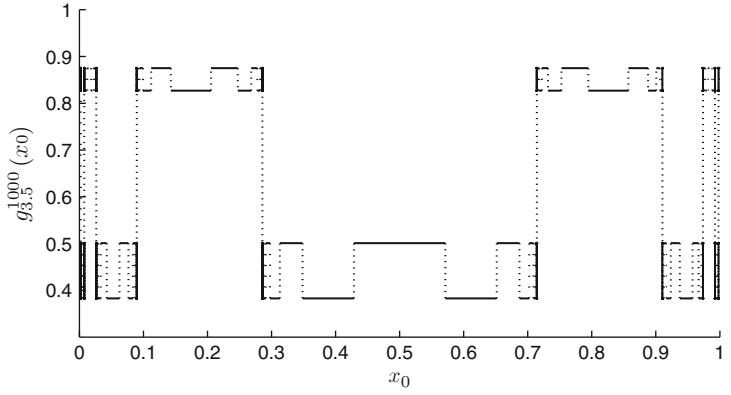


Fig. 6. Piecewise chebfun representation of $g_{3.5}^{1000}$. Now there are four constant values instead of two

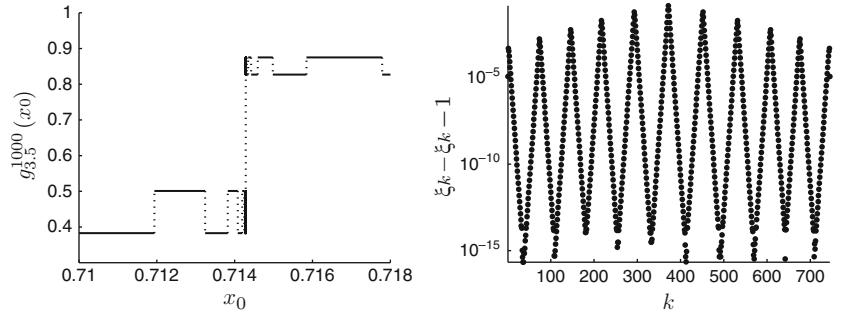


Fig. 7. *Left*: plot of $g_{3.5}^{1000}$ near the unstable fixed point $x = 1 - 1/3.5$. *Right*: semilog plot of the distance between breakpoints of $g_{3.5}^{1000}$

increases. Figure 6 shows the chebfun representation for $r = 3.5$ and $k = 1{,}000$, `g1000 = chebfun(@(x) logistic(3.5,1000,x), [0.001 0.999])`. Notice that now there are several clusters of break points. A detailed plot of $g_{3.5}^{1000}$ around the unstable fixed point $x = 1 - 1/3.5 = 0.714285\ldots$ is presented in Fig. 7. The semilog plot on the right of this figure shows that the distance between neighboring break points decreases exponentially near some critical values. In this plot, ξ_k denotes jump locations which were recovered from the field `g1000.ends`. This graph was generated with `semilogy(diff(g1000.ends))`.

Because these subsequences of polynomials are converging to piecewise constant functions, the pointwise convergence is slower near the location of a

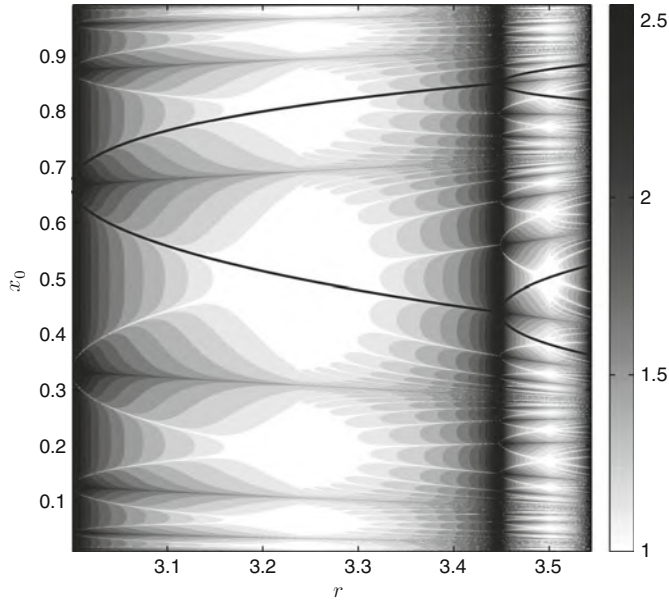


Fig. 8. Number of iterations needed to converge (to a tolerance of 10^{-5}) to the 2-cycle and 4-cycle limits as a function of x_0 and r. The grayscale map shows the $\log_{10}$ of the number of iterations. The bifurcation diagram is superimposed (*solid lines*). Figures 5 and 6 correspond to vertical sections through this plot at $r = 3.2$ and $r = 3.5$, respectively

jump in the limit function. Figure 8 shows a grayscale map of the logarithm of the number of iterations required for a subsequence $\{g_r^k(x_0)\}$ to converge to its limit, to a tolerance of 10^{-5}. This figure is not the result of a chebfun computation; it is provided to give insight into the convergence of chebfun computations. Notice that near bifurcation points, convergence is very slow, regardless of the starting value. Away from these regions, convergence is fast almost everywhere. The locations of slow convergence in this case seem to coincide with the jump locations in the limiting function. Similar convergence maps have been presented in [12].

Finally, the logistic map can also be used to illustrate some limitations of piecewise polynomial representations. Near bifurcation points, for instance, chebfun representations of the maps g_r^k can only be obtained for very small values of k, since the degree of the representations grows exponentially with k and the limit is not achieved in thousands of iterations. Similarly, near or

at the chaotic regimes, the maps are impossible to represent for large k due to the complexity of these functions.

3 Chebops

The chebop system developed by Driscoll et al. [3] is an extension of the chebfun system to handle linear operators. Here, the analogy is between matrices and continuous operators rather than vectors and functions.

A chebop object is defined by a domain, a chebfun, or another chebop. Identity, differentiation and integration operators, for instance, are defined using the domain class:

```
[d,x] = domain(0,1);
D = diff(d)     % differentiation
I = eye(d)      % identity
S = cumsum(d)   % integration
```

We point out that `domain` returns a domain object and a chebfun, in this case `x`. The multiplication operator, on the other hand, is defined by a chebfun and the exponential operator by a chebop. These operators can then be combined to generate other chebops. For example, `L = D^2 + 5*I` defines the operator $L : u \mapsto \partial^2 u/\partial x^2 + 5u$.

In chebops, multiplication has been overloaded to apply operators to chebfuns and other chebops. This can be illustrated as follows:

```
u = sin(3*pi*x)
f = L*u
```

Now, suppose that we would like to solve the differential equation $Lu = f$ for u. Of course, the backward operation requires boundary conditions for uniqueness. For example, if the desired boundary conditions are homogeneous Dirichlet at $x = 0$ and Neumann at $x = 1$, we augment L with

```
L.lbc = 'dirichlet'  % left boundary condition
L.rbc = 'neumann'    % right boundary condition
```

and the solution of the differential equation can then be obtained using the backslash command, which has been overloaded to invert chebops:

```
sol = L\f
```

The algorithms used in the chebop system are described in [3]. When inverting these operators, as in the solution of differential equations, chebops rely on adaptive spectral collocation methods that are also based on Chebyshev polynomials [13,14]. Lazy evaluations of the associated spectral discretization matrices are performed to compute the solution. As in the chebfun system, the polynomial degree of the solution of a differential equation is determined

by the relative magnitudes of Chebyshev coefficients. In the present implementation, most chebops operations are restricted to global representations, i.e., to `splitting off` mode.

We give a number of examples that illustrate the use of chebops in the solution of linear and nonlinear ODEs, PDEs, and eigenvalue problems. The codes used to solve each problem are provided here, and more examples can be downloaded from the chebfun website [10].

3.1 Linear Differential Equations with Variable Coefficients

Consider the hypergeometric equation

$$xy'' + (5 - x)y' + y = \sin(5x), \quad x \in (1, 6), \tag{3}$$

subject to homogeneous Neumann boundary conditions. The chebop syntax to obtain a solution is

```
[d, x] = domain([1 6]);
D = diff(d);
L = diag(x)*D^2 + diag(5-x)*D + eye(d) & 'neumann';
u = L\sin(5*x);
plot(u)
```

Here `diag` is used to define the multiplication operator and `&` to define the boundary conditions. When this code is executed, the system adaptively determines that the desired solution can be represented to approximately machine precision by a polynomial of degree 47. The plot is shown in the left of Fig. 9. The maximum value of the residual in this calculation is

```
>> norm(L*u-sin(5*x),inf)
ans = 3.925115787950517e-11
```

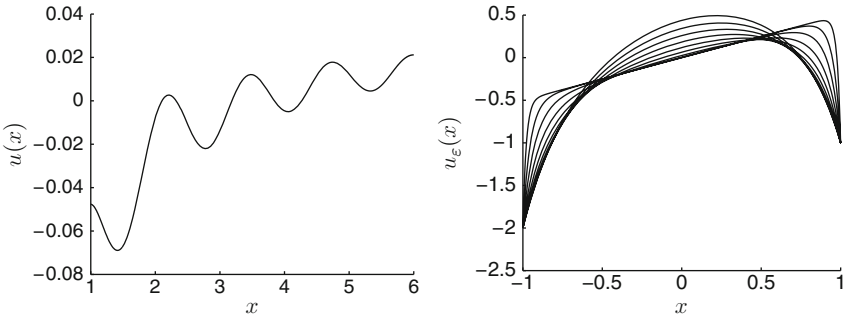


Fig. 9. Chebop solution of two boundary value problems. *Left*: the hypergeometric equation (3). *Right*: the boundary layer problem (4) with $\varepsilon = 0.02, 0.04, \ldots, 0.2$

Our next example is the singularly perturbed problem [15],

$$\varepsilon y'' - xy' + y = 0, \quad x \in (-1, 1), \quad y(-1) = -2, \quad y(1) = -1. \tag{4}$$

Chebops handle boundary layers well, as the clustering of Chebyshev nodes provide good resolution near the endpoints of the interval. The following commands generate plots for several values of ε:

```
figure, hold on
[d,x] = domain(-1,1);
D = diff(d);
for ep = 0.02:0.02:0.2
    L = ep*D^2-diag(x)*D+eye(d);
    L.lbc = -2; L.rbc = -1;
    plot(L\0)
end
```

The solutions correspond to polynomials of degrees 64, 50, 42, 38, 36, 34, 34, 28, 28, 28, and are presented on the right of Fig. 9.

3.2 The Orr–Sommerfeld Eigenvalue Problem

The chebop system also overloads the command `eigs` to solve eigenvalue problems. The eigenvalues of the 1D Laplacian operator on $[0, \pi]$, for instance, can be easily computed with

```
>> d = domain(0,pi);
>> L = -diff(d,2) & 'dirichlet';
>> eigs(L,6)
ans =
   0.999999999999991
   3.999999999999823
   8.999999999999659
  15.999999999999831
  25.000000000000089
  35.999999999999893
```

The command `eigs` has been overloaded instead of `eig` because, in MATLAB, the latter is used to return all eigenvalues of a matrix, which is not possible for differential operators. The details of which eigenvalues are returned by `eigs` can be found in [3].

Our next example is an Orr–Sommerfeld generalized eigenvalue problem arising in the eigenvalue stability analysis of plane Poiseuille fluid flow. The Orr–Sommerfeld equation is given by

$$\frac{d^4u}{dx^4} - 2\alpha^2\frac{d^2u}{dx^2} + \alpha^4 u - i\alpha R\left[(1-x^2)\left(\frac{d^2u}{dx^2} - \alpha^2 u\right) - 2u\right] = \lambda\left(\frac{d^2u}{dx^2} - \alpha^2 u\right)$$

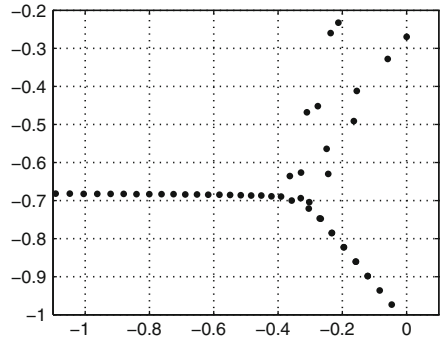


Fig. 10. Rightmost eigenvalues of the Orr–Sommerfeld operator in the complex plane for $R = 5772.22$ and $\alpha = 1.02056$

where R is the Reynolds number and α a wave number. Orszag showed in [16] that $R = 5772.22$, $\alpha = 1.02056$ are critical values, with one of the eigenvalues crossing to the right half of the complex plane. We repeat his eigenvalue computation using chebops.

```
[d,x] = domain(-1,1);
I = eye(d); D = diff(d);
R = 5772.22; alpha = 1.02056;
B = D^2 - alpha^2;
A = B^2/R - 1i*alpha*(2+diag(1-x.^2)*B);
A.lbc(1) = I; A.lbc(2) = D;
A.rbc(1) = I; A.rbc(2) = D;
e = eigs(A,B,50,'LR');
```

We confirm Orszag's result by showing these eigenvalues in Fig. 10 and computing their largest real part:

```
>> max(real(e))
ans = 6.129513257887425e-09
```

3.3 Linear Partial Differential Equations

Certain linear partial differential equations can also be handled by chebops. In the following example, we use the exponential operator `expm` to advance in time. Writing a linear partial differential equation in the form $u_t = Lu$, we have $u(t + \Delta t, x) = \exp(\Delta t L)u(t, x)$, assuming that $\exp(\Delta t L)$ is well defined. The following code solves the convection-diffusion equation,

$$u_t = 0.05u_{xx} - xu_x, \qquad x \in (-2, 2), \tag{5}$$

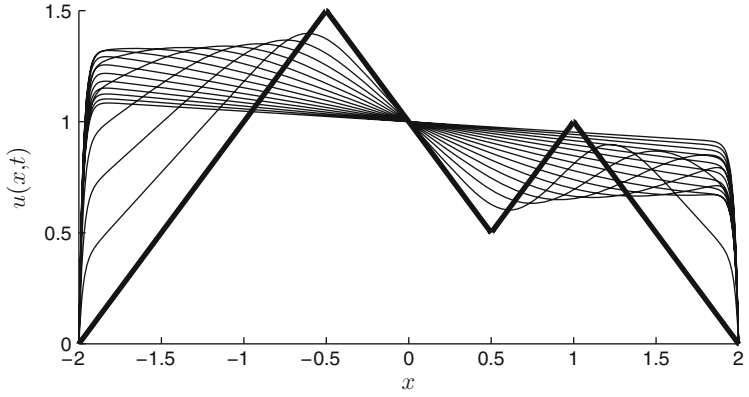


Fig. 11. Solution to the PDE (5) at several times t. The initial condition is shown by a *thick line*

with homogeneous Dirichlet boundary conditions and initial condition

$u(0, x) = -|x + 0.5| + |x - 0.5| - |x - 1| + 2.$

```
[d,x] = domain(-2,2);
splitting on
u = chebfun(@(x) -abs(x+0.5)+abs(x-0.5)-abs(x-1)+2, d);
splitting off
L = 0.05*diff(d,2)-diag(x)*diff(d);
dt = 0.2; expmL = expm(dt*L & 'dirichlet');
plot(u,'k', 'linewidth',4), hold on
for t = 0:dt:3
    u = expmL*u;
    plot(u,'k')
end
```

The result of this execution is presented in Fig. 11. Notice that despite the lack of smoothness in the initial condition, chebops can be used in the solution of this problem as u is smooth for all $t > 0$.

Chebops can also be used to solve nonlinear PDEs with implicit or semi-implicit time-stepping schemes. One example, involving the nonlinear cubic Schrödinger equation, is presented in [3].

3.4 Nonlinear Boundary-Value Problems

While linear equations can be solved with "\", nonlinear problems require iterative algorithms.[1] In our next example we use Newton's method together with chebop technology to solve the boundary-value problem

[1]Solutions to nonlinear boundary value problems have been automated in Chebfun Version 3 via automatic differentiation. The example in this section can now be solved with "\".

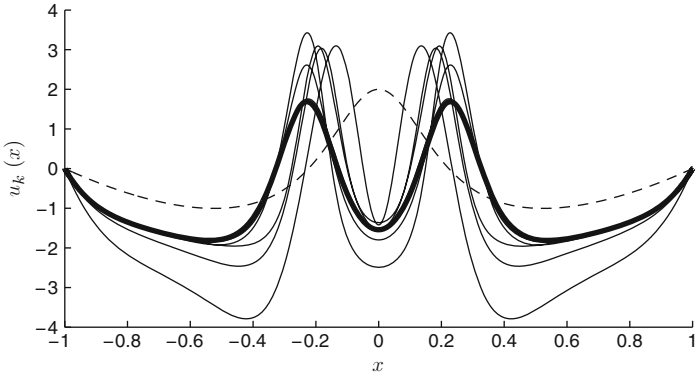


Fig. 12. Newton's method solution of (6). Intermediate iterates u_k are shown together with the initial guess (*dashed*) and the final solution (*thick line*) – cf. Fig. 9.26 in [17]

$$\varepsilon u'' + 2(1-x^2)u + u^2 = 1, \quad x \in (-1,1), \tag{6}$$

with homogeneous Dirichlet boundary conditions. This equation, due to Carrier, is discussed at length by Bender and Orszag [17]. The problem has many solutions, some of which can be approximated by boundary-layer theory. The following code was used to generate the solution plotted in Fig. 12. The figure also shows the intermediate Newton method iterates.

```
[d,x] = domain(-1,1);
D2 = diff(d,2); F = diag(2*(1-x.^2));
u = 2*(x.^2-1).*(1-2./(1+20*x.^2));
eps = 0.01; nrmdu = Inf;
plot(u,'--k'), hold on
while nrmdu > 1e-10
    r = eps*D2*u + F*u + u.^2 - 1;
    A = eps*D2 + F + diag(2*u) & 'dirichlet';
    A.scale = norm(u); delta = -(A\r);
    u = u+delta;  nrmdu = norm(delta)
    plot(u,'k')
end
plot(u,'k', 'linewidth',4)
```

3.5 Ground State Solution of the 3D Cubic Schrödinger Equation

Our final example, which comes to us from Roudenko and Holmer [20], is related to radial solutions of the cubic Schrödinger equation in $\mathbb{R}^3$,

$$iu_t + \Delta u + |u|^2 u = 0.$$

Using the separation of variables $u(x,t) = e^{it}v(x)$, we obtain a nonlinear equation for v,

$$-v + \Delta v + |v|^2 v = 0. \tag{7}$$

This equation has an infinite number of solutions in $H^1(\mathbb{R}^3)$. The solution of minimal mass is positive, radial, and exponentially decaying and is called *the ground state* [18].

We shall seek a positive radial solution to (7) with exponential decay. Because the current implementation of the chebfun and chebop systems is restricted to bounded domains, we perform the change of variables $r = \tilde{r}/(1-\tilde{r})$, $\tilde{r} \in [0,1]$, and $Q(\tilde{r}) = v(\tilde{r}/(1-\tilde{r}))$. An equation for Q can then be written as

$$\tilde{r}\left[-Q + (1-\tilde{r})^4 Q_{\tilde{r}\tilde{r}} + Q^3\right] + 2(1-\tilde{r})^4 Q_{\tilde{r}} = 0, \qquad \tilde{r} \in (0,1), \tag{8}$$

with boundary conditions $Q_{\tilde{r}}(0) = Q(1) = 0$. As in the previous example, we use Newton's method to find a solution.

```
[d,r] = domain(0,1); D = diff(d); D2 = D^2;
Q = chebfun(@(r) 4*sech(2*r./(1-r+eps)), d);
nrmdu = Inf;
while nrmdu > 1e-13
   res = r.*(Q.^3-Q+(1-r).^4.*(D2*Q)) + 2*(1-r).^4.*(D*Q);
   A = diag(r)*(diag(3*Q.^2)-eye(d)+diag((1-r).^4)*D2)+ ...
       2*diag((1-r).^4)*D;
   A.rbc = 'dirichlet' ; A.lbc = 'neumann';
   A.scale = norm(Q);  delta = -(A\res);
   Q = Q+delta;  nrmdu = norm(delta)
end
plot(Q)
```

The resulting plot is shown in Fig. 13.

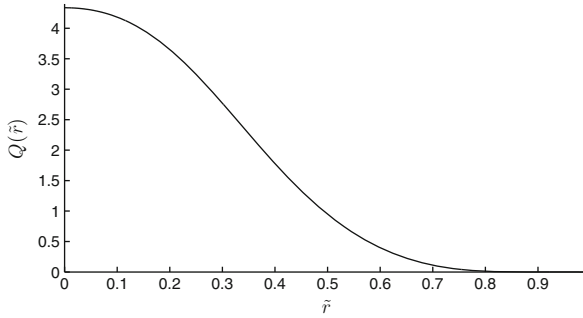


Fig. 13. The solution of (8) computed with Newton's method

4 Concluding Remarks

A brief review of the chebfun and chebop systems has been presented and several examples provided to demonstrate how simple and effective the system is. Some capabilities of the software have not been mentioned here, such as quasimatrices [19]. The system is evolving and efforts are currently being made to extend it to handle unbounded domains via mapped polynomial representations. We hope that the change of variables performed in the final example may be handled automatically in future releases.

The computations presented in this paper were carried out with the October 2008 release of chebfun Version 2. The code is freely available under a BSD-type software license, and can be found together with a user's guide and other information at http://www.maths.ox.ac.uk/chebfun/.

Acknowledgments

The chebfun system is currently a joint project with Ricardo Pachón and Toby Driscoll. Toby Driscoll is the principal author of the chebop system, to which another key contributor was Folkmar Bornemann.

References

1. Battles, Z., Trefethen, L.N.: An extension of MATLAB to continuous functions and operators. SIAM J. Sci. Comput. **25**(5), 1743–1770 (2004)
2. Pachón, R., Platte, R.B., Trefethen, L.N.: Piecewise-smooth chebfuns. IMA J. Numer. Anal. doi:10.1093/imanum/drp008 (2009)
3. Driscoll, T.A., Bornemann, F., Trefethen, L.N.: The chebop system for automatic solution of differential equations. BIT Numer. Math. **48**(4), 701–723 (2008)
4. Trefethen, L.N.: Computing numerically with functions instead of numbers. Math. Comput. Sci. **1**(1), 9–19 (2007)
5. Boyd, J.P.: Computing zeros on a real interval through Chebyshev expansion and polynomial rootfinding. SIAM J. Numer. Anal. **40**(5), 1666–1682 (2002)
6. Good, I.J.: The colleague matrix, a Chebyshev analogue of the companion matrix. Quart. J. Math. **12**, 61–68 (1961)
7. Berrut, J.-P., Trefethen, L.N.: Barycentric Lagrange interpolation. SIAM Rev. **46**(3), 501–517 (2004)
8. Salzer, H.E.: Lagrangian interpolation at the Chebyshev points $X_{n,\nu} \equiv \cos(\nu\pi/n), \nu = 0(1)n$; some unnoted advantages. Comput. J. **15**, 156–159 (1972)
9. Higham, N.J.: The numerical stability of barycentric Lagrange interpolation. IMA J. Numer. Anal. **24**(4), 547–556 (2004)
10. Trefethen, L.N., Pachón, R., Platte, R.B., Driscoll, T.A.: Chebfun version 2. http://www.maths.ox.ac.uk/chebfun/ (2008)
11. Sprott, J.C.: Chaos and Time-Series Analysis. Oxford University Press, New York (2003)

12. Bresten, C.L., Jung, J.-H.: A study on the numerical convergence of the discrete logistic map. Commun. Nonlinear Sci. **14**(7), 3076–3088 (2009)
13. Fornberg, B.: A Practical Guide to Pseudospectral Methods. Cambridge University Press, Cambridge (1996)
14. Trefethen, L.N.: Spectral Methods in MATLAB. SIAM, Philadelphia, PA (2000)
15. O'Malley, R. Jr.: Singularly perturbed linear two-point boundary value problems. SIAM Rev. **50**(3), 459–482 (2008)
16. Orszag, S. A.: Accurate solution of the Orr-Sommerfeld stability equation. J. Fluid Mech. **50**, 689–703 (1971)
17. Bender, C.M., Orszag, S.A.: Advanced Mathematical Methods for Scientists and Engineers. I. Springer, New York (1999). Reprint of the 1978 original
18. Weinstein, M.I.: Nonlinear Schrödinger equations and sharp interpolation estimates. Comm. Math. Phys. **87**(4), 567–576 (1982/1983).
19. Trefethen, L.N.: Householder triangularization of a quasimatrix. IMA J. Numer. Anal. doi:10.1093/imanum/drp018 (2009)
20. Holmer, J., Roudenko, S.: A sharp condition for scattering of the radial 3D cubic nonlinear Schrödinger equation. Commun. Math. Phys. **282**(2), 435–467 (2008)

Part II

Minisymposia

Minisymposium *Asymptotic Analysis*

D. Dominici[1] and R.B. Paris[2]

[1] Department of Mathematics, State University of New York at New Paltz, 1 Hawk Dr., New Paltz, NY 12561-2443, USA, dominicd@newpaltz.edu
[2] University of Abertay Dundee, Dundee, UK, r.paris@abertay.ac.uk

Many of the problems facing mathematicians and scientists involve such difficulties as non-linear governing equations and complex boundary conditions that preclude their exact solution. Consequently, solutions are approximated using numerical techniques, analytic techniques or combinations of both. Foremost among the analytic techniques are the systematic methods of perturbations (asymptotic expansions) in terms of a small or large parameter or coordinate.

The advantage of allowing parameters to become small or large is that in surprisingly many cases, even when there do exist explicit expressions for the functions we are interested in, this procedure does yield simple asymptotic approximations, when the influence of less important factors falls off.

In recent years, asymptotic methods have been used extensively in several fields of pure and applied mathematics including algebra, geometry, analysis, differential and difference equations, probability theory, number theory, special functions and combinatorics.

This section contains works by speakers in the Minisymposium MS10 Asymptotic Analysis. The articles cover a wide range of topics, including singular perturbations, asymptotic inversion, special functions and entropic measures.

Jesús Sánchez-Dehesa, from the Universidad de Granada, studies very highly excited (Rydberg) states of hydrogenic atoms with energy levels

$$E = -\frac{Z^2}{2\eta^2}, \quad \text{with} \quad \eta = n + \frac{D-3}{2}, \quad n = 1, 2, \ldots$$

where Z is the nuclear charge and D is the dimension. He calculates their Shannon information entropy $S(\rho)$, defined by

$$S(\rho) = -\int \rho(\mathbf{r}) \log \rho(\mathbf{r}) d\mathbf{r},$$

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*, Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_6,

where $\rho(\mathbf{r})$ denotes the quantum-mechanical probability of finding an electron in the volume element $(\mathbf{r}, \mathbf{r}+d\mathbf{r})$, asymptotically as $n \to \infty$. To accomplish this, he uses asymptotic properties of Laguerre and Gegenbauer polynomials.

Diego Dominici from the State University of New York at New Paltz, analyzes the zeros of the Hermite polynomials $H_n(x)$ asymptotically as $n \to \infty$. Denoting by $\zeta_1^n > \zeta_2^n > \cdots > \zeta_n^n$ the zeros of $H_n(x)$, enumerated in decreasing order, he derives the asymptotic approximation $\zeta_j^n \sim \sqrt{2n}\sin(\tau_j^n)$, where τ_j^n is given by the Kapteyn series

$$\tau_j^n = \frac{\pi}{2} - \frac{\pi}{2}(4j-1)N^{-1} - \sum_{k=1}^{\infty} \frac{1}{k} \mathrm{J}_k[(1-N^{-1})k] \sin\left(\frac{4j-1}{N} k\pi\right),$$

where $N = 2n+1$.

Ester Pérez Sinusía from the Universidad Pública de Navarra, studies the importance of the error function in the approximation of the solution of the singularly perturbed convection-diffusion equation

$$-\varepsilon \Delta U + \overrightarrow{v} \cdot \overrightarrow{\nabla} U = 0$$

with discontinuous boundary conditions, where $\epsilon > 0$ is a small perturbation parameter and $\overrightarrow{v}$ is a constant vector. She presents examples in two and three dimensions, including a quarter plane, an infinite strip, a rectangle and an octant.

Renato Spigler from the Università "Roma Tre", discusses the singular perturbation of parabolic partial differential equations with or without boundary layers. This type of problem is characterised by the presence of a parameter ϵ that tends to zero to produce either a reduction in the order of the equation or a change in its type. The usual treatment of such problems involves the introduction of a boundary layer. However, there are cases where no boundary layer is required during the passage to the limit. Simple model examples are given in which conditions on the data are obtained for there to be no boundary layer as $\epsilon \to 0^+$.

Nico Temme from the Centrum voor Wiskunde en Informatica, considers the asymptotic inversion of cumulative distribution functions of the form

$$F_a(\eta) = \sqrt{\frac{a}{2\pi}} \int_{-\infty}^{\eta} e^{-\frac{1}{2}a\zeta^2} f(\zeta)\, d\zeta,$$

where $a > 0$, $\eta \in \mathbb{R}$, and f is analytic and real on $\mathbb{R}$ with $f(0) = 1$. In particular, the normal distribution, the incomplete gamma function and the incomplete beta function can be written in this form. As a particular example, he analyzes the hyperbolic distribution, given by

$$F(y) = C \int_{-\infty}^{y} e^{-\alpha\sqrt{\delta^2+(x-\mu)^2}+\beta(x-\mu)}\, dx, \quad y \in \mathbb{R},$$

where $\alpha > 0$, $|\beta| < \alpha$, δ and μ are arbitrary real constants and C is the normalizing constant which gives $F(\infty) = 1$.

Asymptotics of Orthogonal-Polynomial Functionals and Shannon Information Entropy of Rydberg Atoms

J.S. Dehesa[1,4], S. López-Rosa[1,4], A. Martínez-Finkelshtein[3,4], and R.J. Yáñez[2,4]

[1] Departamento de Física Atómica, Molecular y Nuclear, Universidad de Granada, 18071-Granada, Spain, dehesa@ugr.es
[2] Departamento de Matemática Aplicada, Universidad de Granada, 18071-Granada, Spain, slopez@ugr.es
[3] Departamento de Estadística y Matemática Aplicada, Universidad de Almería, La Cañada, 04120 Almería, Spain, andrei@ual.es
[4] Instituto Carlos I de Física Teórica y Computacional, Universidad de Granada, 18071-Granada, Spain, ryanez@ugr.es

Summary. The asymptotics of entropic integrals of Laguerre and Gegenbauer polynomials is used to calculate the Shannon information entropy of Rydberg atoms (i.e. giant atoms of hydrogenic type), which provides a bulky spreading measure of their charge density much more appropriate than the standard deviation or Heisenberg measure. These systems are stepping stones from the quantum to classical worlds. Indeed because of its exaggerated properties, a Rydberg atom is a good laboratory for investigating how quantum and classical physics correspond when the latter involves irregular (chaotic) orbits.

1 Introduction

Rydberg atoms [1,2] are swollen atoms with energy, i.e. giant atoms of hydrogenic type. They were theoretically predicted in the early days of Quantum Mechanics and first detected in 1965 in the interstellar space, but they were only produced in 1970 at Argonne National Laboratory. Nowadays, atoms in Rydberg states mimic circular and elliptic classical orbits of specified eccentricity by means of laser excitation in presence of perpendicular electric and magnetic fields.

They present exaggerated properties such as long radiative lifetimes and strong long-range interactions, which allow them to be good laboratories for investigating how quantum and classical physics correspond when the latter involves irregular (chaotic) orbits. Indeed, they probe the shadowy realm where the quantum world of the atom gives way to the familiar classical world.

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*, Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_7,

Moreover, they show evidence for chaos in which the motion of the Rydberg electrons become hard, even impossible to predict. For these reasons, Rydberg atoms are considered stepping stones from the quantum to the classical worlds.

Here we investigate the spatial extension of the quantum-mechanical density of a D-dimensional Rydberg system in position space by means of the Shannon information entropy given by

$$S(\rho) = -\int \rho(\mathbf{r}) \log \rho(\mathbf{r}) d\mathbf{r} \tag{1}$$

This quantity not only controls the bulky spreading of the atomic charge but also it is an uncertainty measure much more appropriate than the renowned standard deviation or Heisenberg measure, mainly because the latter gives a large weight to the tails of the density. Moreover, the Shannon entropy is a basic variable of the emerging information theory of quantum-mechanical systems (see e.g. [3]) which is the foundational pillar of the modern quantum information and computation [4].

Here we first express the Shannon entropy of general hydrogenic systems in terms of entropic functionals of Laguerre and Gegenbauer polynomials, and then we use their asymptotics to accurately determine the Shannon entropy of the Rydberg states (i.e. highly excited states) of hydrogenic atoms.

This work is structured as follows. First, in Sect. 2, the Schrödinger wave equations of a particle moving in a D-dimensional central potential is given and its physical solutions (the wavefunctions of the allowed quantum-mechanical states) are presented. In this section, the associated probability density is shown to be separated out in two radial and angular parts, which are basically controlled by the Rakhmanov densities of Laguerre and Gegenbauer polynomials respectively. Then, in Sect. 3, the Shannon entropy of general hydrogenic systems is expressed in terms of the entropic functionals of these orthogonal polynomials. Finally, in Sect. 4, the asymptotics of these functionals is used to obtain the Shannon entropy of the highly excited (Rydberg) atomic states in terms of the quantum numbers, the nuclear charge and the dimensionality.

2 The Schrödinger Equation of a D-Dimensional Central Potential

The quantum-mechanical motion of a particle (say, an electron) in the D-dimensional Coulomb potential $V(r) = -Z/r$ is governed by the Schrödinger equation

$$\left(-\frac{1}{2}\boldsymbol{\nabla}_D^2 - \frac{Z}{r}\right) \Psi(\mathbf{r}) = E \ \Psi(\mathbf{r})$$

in the appropriate atomic units, where $\boldsymbol{\nabla}_D$ denotes the gradient operator associated to the D-dimensional position vector $\mathbf{r} = (r, \theta_1, \theta_2, \ldots, \theta_{D-1})$. The

physical solutions of this equation, which correspond to the wavefunctions of our system, are characterized (see e.g. [3,5]) by the energy eigenvalues

$$E = -\frac{Z^2}{2\eta^2}, \qquad \text{with} \qquad \eta = n + \frac{D-3}{2}; \qquad n = 1, 2, 3, \ldots, \tag{2}$$

and the eigenfunctions

$$\Psi_{n,l,\{\mu\}}(\mathbf{r}) = R_{n,l}(r)\mathcal{Y}_{l,\{\mu\}}(\Omega_{D-1}), \tag{3}$$

where η and $(l, \{\mu\}) \equiv (l \equiv \mu_1, \mu_2, \ldots, \mu_{D-1})$ denote the radial hyperquantum number and the angular hyperquantum numbers associated to the variables $(\theta_1, \theta_2, \ldots, \theta_{D-1}) \equiv \Omega$, which may have all values consistent with the inequalities $l \equiv \mu_1 \geq \mu_2 \geq \cdots \geq |\mu_{D-1}| \equiv |m| \geq 0$. The radial part $R_{n,l}(r)$ is given by

$$R_{n,l}(r) = \left(\frac{\lambda^{-D}}{2\eta}\right)^{1/2} \left[\frac{\omega_{2L+1}(\hat{r})}{\hat{r}^{D-2}}\right]^{1/2} \tilde{\mathcal{L}}^{2L+1}_{\eta-L-1}(\hat{r}), \tag{4}$$

where $\tilde{\mathcal{L}}^{\alpha}_{k}(x)$ denotes the orthonormal Laguerre polynomials of degree k and parameter α, and the ground orbital angular momentum hyperquantum number L and the adimensional parameter $\hat{r}$ are

$$L = l + \frac{D-3}{2}, \qquad l = 0, 1, 2, \ldots \qquad \text{and} \qquad \hat{r} = \frac{r}{\lambda}, \qquad \text{with} \qquad \lambda = \frac{\eta}{2Z}. \tag{5}$$

The angular part $\mathcal{Y}_{l,\{\mu\}}(\Omega_{D-1})$ is given by the hyperspherical harmonics [6]

$$\mathcal{Y}_{l,\{\mu\}}(\Omega_{D-1}) = \frac{1}{\sqrt{2\pi}} e^{im\varphi} \prod_{j=1}^{D-2} \tilde{C}^{\alpha_j+\mu_{j+1}}_{\mu_j-\mu_{j+1}}(\cos\theta_j)(\sin\theta_j)^{\mu_{j+1}}, \tag{6}$$

with $\alpha_j = \frac{1}{2}(D-j-1)$ and $\tilde{C}^{\lambda}_{k}(x)$ denotes the orthonormal Gegenbauer polynomials of degree k and parameter λ.

Then, the quantum-mechanical probability to find the electron in the volume element $(\mathbf{r}, \mathbf{r} + d\mathbf{r})$ is

$$\begin{aligned} \rho_{n,l,\{\mu\}}(\mathbf{r})d\mathbf{r} &= \left|\Psi_{n,l,\{\mu\}}(\mathbf{r})\right|^2 d\mathbf{r} = R^2_{n,l}(r)r^2 dr \left|\mathcal{Y}_{l,\{\mu\}}(\Omega_{D-1})\right|^2 d\Omega \\ &\equiv D_{n,l}(r)r^2 dr \times \Pi(\Omega)d\Omega, \end{aligned} \tag{7}$$

where $D_{n,l}(r) \equiv R^2_{n,l}$ denotes the radial probability density which gives the probability per radial interval to find the particle between r and $r + dr$, and $\Pi(\Omega) \equiv |\mathcal{Y}_{l,\{\mu\}}(\Omega_{D-1})|^2$ describes the spatial profile of our system.

3 The Shannon Entropy of Hydrogenic Systems

The intrinsic randomness or uncertainty of our system (i.e. a hydrogenic atom with nuclear charge Z) is best defined by (1) and (7) which characterize the Shannon entropy of the density $\rho_{n,l,\{\mu\}}(\mathbf{r})$.

It turns out that

$$S(\rho_{n,l,\{\mu\}}) = S(R_{n,l}) + S(\mathcal{Y}_{l,\{\mu\}}) \tag{8}$$

where the radial part is given by

$$\begin{aligned} S(R_{n,l}) &= -\int_0^\infty r^{D-1} R_{n,l}^2(r) \log R_{n,l}^2 dr \\ &= A(n,l,D) - \frac{1}{2\eta} E_1\left(\tilde{\mathcal{L}}_{\eta-L-1}^{2L+1}\right) - D\ln Z \end{aligned} \tag{9}$$

and the angular part is

$$\begin{aligned} S(\mathcal{Y}_{l,\{\mu\}}) &= -\int_{S_{D-1}} |\mathcal{Y}_{l,\{\mu\}}(\Omega_{D-1})|^2 \ln \left|\mathcal{Y}_{l,\{\mu\}}(\Omega_{D-1})\right|^2 d\Omega_{D-1} \\ &= B(l,\{\mu\},D) + \sum_{j=1}^{D-2} E_2\left(\tilde{C}_{\mu_j-\mu_{j+1}}^{\alpha_j+\mu_{j+1}}\right). \end{aligned} \tag{10}$$

Relations (4) and (6) were used in the second equality of (9) and (10), respectively. We have obtained the values

$$\begin{aligned} A(n,l,D) = -2l\left[\frac{2\eta-2L-1}{2\eta} + \psi(\eta+L+1)\right] + \frac{(\Omega_{D-1})3\eta^2 - L(L+1)}{\eta} \\ + \ln\left[\frac{(\eta-L-1)!}{(\eta+L)!}\right] - \ln\left[\frac{2^{D-1}(\eta-L-1)!}{\eta^{D+1}(\eta+L)!}\right] \end{aligned}$$

$$\begin{aligned} B(l,\{\mu\},D) = \ln 2\pi - 2\sum_{j=1}^{D-2}\mu_{j+1} \\ \times\left[\psi(2\alpha_j+\mu_j+\mu_{j+1}) - \psi(\alpha_j+\mu_j) - \ln 2 - \frac{1}{2(\alpha_j+\mu_j)}\right], \end{aligned}$$

for the terms A and B. The entropic functionals $E_1(\tilde{y}_n)$ and $E_2(\tilde{y}_n)$ of the polynomials $\{\tilde{y}_n(x)\}$, orthonormal with respect to the weight function $\omega(x)$, are defined by

$$E_1(\tilde{y}_n) = -\int_0^\infty x\omega(x)\tilde{y}_n^2(x)\ln\tilde{y}_n^2(x)dx, \tag{11}$$

and

$$E_2(\tilde{y}_n) = -\int_{-1}^{+1} \omega(x)\tilde{y}_n^2(x)\ln\tilde{y}_n^2(x)dx, \tag{12}$$

respectively. It is well known that the weight functions of Laguerre and Gegenbauer polynomials, $\tilde{L}_k^\alpha(x)$ and $\tilde{C}_k^\lambda(x)$, are given by

$$\omega_\alpha(x) = x^\alpha e^{-x}; \qquad\qquad \omega_\lambda^*(x) = (1-x^2)^{\lambda-\frac{1}{2}},$$

respectively. The Gegenbauer entropic functional $E_2(\tilde{C}_k^{\lambda})$ involved in the evaluation of the angular entropy given by (10) can be numerically computed quite accurately by means of the recent algorithm of Buyarov et al. [7]. The Laguerre entropic functional $E_1(\tilde{L}_k^{\alpha})$ involved in the radial entropy given by (9) can be analytically calculated only for the very few lowest degrees of the polynomial. In the general case it is a formidable open task. Here, in the next section, we solve this problem in the asymptotic case, i.e. for large value of the degree.

4 Shannon Entropy of Rydberg Atoms and Asymptotics of Laguerre and Gegenbauer Polynomials

Here we calculate the Shannon entropy of highly and very highly excited (Rydberg) states of hydrogenic atoms with nuclear charge Z and dimensionality D, which is the main result of this work. To do that we have to determine the value of $S(\rho_{n,l,\{\mu\}})$ given by (8) for large n. Since the angular part does not depend on n, everything reduces to the evaluation of the radial entropy $S(R_{n,l})$ for large values of n. To solve this problem we need to use the following asymptotical results [8] for the entropic integral $E_1(\tilde{\mathcal{L}}_k^{\alpha})$ of the orthonormal Laguerre polynomials $\tilde{\mathcal{L}}_k^{\alpha}$ for fixed $\alpha > -1$ and $k \to \infty$:

$$\begin{aligned} E_1(\tilde{\mathcal{L}}_k^{\alpha}) &= -\int_0^{\infty} x\omega_{\alpha}(x) \left[\tilde{\mathcal{L}}_k^{\alpha}(x)\right]^2 \ln\left[\tilde{\mathcal{L}}_k^{\alpha}(x)\right]^2 dx \\ &= -6k^2 + (2\alpha+1)k\ln k + \ln(2\pi) - 2\alpha - 4 + o(1). \end{aligned} \tag{13}$$

The combination of (8), (9) and (13) have allowed us to obtain the value

$$S(\rho) = 2D\ln n + (2-D)\ln 2 + \ln\pi + D - 3 - D\ln Z + S(\mathcal{Y}_{l,\{\mu\}}) + o(1), \tag{14}$$

for the position Shannon entropy of the Rydberg D-dimensional hydrogenic state characterized by the angular hyperquantum numbers $(l, \{\mu\})$ and a very large radial quantum number n. The value of $S(\mathcal{Y}_{l,\{\mu\}})$ is a fixed number, which does not depend on n and can be numerically computed in an accurate way as indicated previously.

In momentum space we can work in a similar manner with the corresponding wavefunctions, which are basically controlled by Gegenbauer polynomials. The use of the asymptotics for the entropic integral $E_2(\tilde{C}_k^{\alpha})$ of the orthonormal Gegenbauer polynomials $\tilde{C}_k^{\alpha}$ found [9–11] as

$$\begin{aligned} E_2(\tilde{C}_k^{\alpha}) &= -\int_{-1}^{+1} \omega_{\alpha}^*(x) \left[\tilde{C}_k^{\alpha}(x)\right]^2 \ln\left[\tilde{C}_k^{\alpha}(x)\right]^2 dx \\ &= \ln\pi + (1-2\alpha)\ln 2 - 1 + o(1), \end{aligned} \tag{15}$$

has allowed us to obtain the value

$$S(\gamma) = -D\ln n + (D+2)\ln 2 + \ln \pi - D - 2 + D\ln Z + S(\mathcal{Y}_{l,\{\mu\}}) + o(1), \quad (16)$$

for the momentum Shannon entropy of the Rydberg D-dimensional hydrogenic state characterized by the angular hyperquantum numbers $(l, \{\mu\})$ and a very large principal quantum number n.

Finally, let us highlight that the net Shannon entropy sum $S(\rho) + S(\gamma)$ has the value

$$S(\rho) + S(\gamma) = D\ln n + 4\ln 2 + 2\ln \pi - 5 + 2S(\mathcal{Y}_{l,\{\mu\}}) + o(1),$$

which can be shown to fulfill the entropic uncertainty relation of Bialynicki-Birula and Mycielski [12].

Acknowledgement

The authors gratefully acknowledge the Spanish MEC grant FIS2005-00973 and FIS2008-02380 (J.S.D, S.L.R and R.J.Y) and MTM2005-08648-C02-01 and MTM2008-06689-C02-01 (A.M.F), and the excellence grants FQM-481, 1735 (J.S.D, S.L.R, A.M.F. and R.J.Y) and 2445 (J.S.D, S.L.R and R.J.Y) of the Junta de Andalucía. They belong to the Andalusian research group FQM-207 (J.S.D, S.L.R and R.J.Y) and FQM-229 (A.M.F.). One of us (S.L.R.) acknwoledges the FPU scholarship of the Spanish Ministerio de Educación y Ciencia.

References

1. Gallagher, T.F.: Rydberg Atoms. Cambridge University Press, Cambridge (1994)
2. Lundeen, S.R.: Adv. At. Mol. Opt. Phys. **52**, 161–208 (2005)
3. Dehesa, J.S., López-Rosa, S., Martínez-Finkelshtein, A., Yáñez, R.J.: Information theory of D-dimensional hydrogenic systems: Applications to circular and Rydberg states. Int. J. Quant. Chem. **110**, 1529–1548 (2009)
4. Nielsen, M.A., Chuang, I.L.: Quantum Computation and Quantum Information. Cambridge University Press, Cambridge (2000) 2nd. Printing
5. Aquilanti, V., Cavalli, S., Coletti, C.: Chem. Phys. **214**, 1–13 (1997)
6. Avery, J.: Hyperspherical Harmonics and Generalized Sturmians. Kluwer, Dordrecht (2000)
7. Buyarov, V., Dehesa, J.S., Martínez-Finkelshtein, A., Sánchez-Lara, J.: SIAM J. Sci. Comput. **26**, 488–509 (2004)
8. Dehesa, J.S., Yáñez, R.J., Aptekarev, A.I., Buyarov, V.: J. Math. Phys. **39**(6), 3050–3060 (1998)
9. Aptekarev, A.I., Dehesa, J.S., Yáñez, R.J.: J. Math. Phys. **35**, 4423–4428 (1994)
10. Aptekarev, A.I., Buyarov, V.S., Dehesa, J.S.: Russian Acad. Sci. Sb. Math. **82**, 373–395 (1995)
11. Aptekarev, A.I., Buyarov, V.S., van Assche, W., Dehesa, J.S., Dokl. Math. **53**, 47–49 (1996)
12. Bialynicki-Birula, I., Mycielski, J.: Common. Math. Phys. **44**, 129 (1975)

Asymptotic Analysis of the Zeros of Hermite Polynomials

Diego Dominici

Department of Mathematics, State University of New York at New Paltz, 1 Hawk Dr., New Paltz, NY 12561-2443, USA, dominicd@newpaltz.edu

Summary. We analyze the zeros of the Hermite polynomials $H_n(\xi)$ asymptotically as $n \to \infty$. Our formulas involve some special functions and they yield very accurate approximations.

1 Introduction

The Hermite polynomials $H_n(x)$ are defined by [1]

$$H_n(x) = n! \sum_{k=0}^{\lfloor \frac{n}{2} \rfloor} \frac{(-1)^k}{k!(n-2k)!} (2x)^{n-2k} \tag{1}$$

for $n = 0, 1, \ldots$. They satisfy the orthogonality condition [2]

$$\int_{-\infty}^{\infty} e^{-x^2} H_m(x) H_n(x) dx = \sqrt{\pi} 2^n n! \delta_{nm}$$

and the reflection formula

$$H_n(-x) = (-1)^n H_n(x). \tag{2}$$

The Hermite polynomials have been extensively studied since the pioneer article of C. Hermite [3] in 1864 (they were previously considered by Fourier and Chebyshev). They have many applications in the physical sciences and are particularly important in the quantum mechanical treatment of the harmonic oscillator [4]. We refer the interested reader to [5] for further properties and references.

The zeros of the Hermite polynomials are very important in applied mathematics and related fields. Several authors have investigated their properties and connections with the zeros of other special functions, see [6–10] and [11].

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*,
Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_8,

In this article we analyze the asymptotic behavior of the zeros of $H_n(x)$ as $n \to \infty$, using and the asymptotic results derived in [12]. We obtain asymptotic approximations that can be expressed in terms of Kapteyn series and present some numerical computations showing the accuracy of our results.

2 Previous Results

In [12, Theorem 5], we studied the differential-difference equation satisfied by the Hermite polynomials

$$H_{n+1} + H_n' = 2xH_n,$$

and obtained, among other results, the asymptotic formula

$$H_n\left[\sqrt{2n}\sin(\theta)\right] \sim \sqrt{\frac{2}{\cos(\theta)}}\exp\left\{\frac{n}{2}\left[\ln(2n) - \cos(2\theta)\right]\right\} \tag{3}$$
$$\times \cos\left\{n\left[\frac{1}{2}\sin(2\theta) + \theta - \frac{\pi}{2}\right] + \frac{\theta}{2}\right\}, \quad n \to \infty,$$

with $-\frac{\pi}{2} < \theta < \frac{\pi}{2}$. This formula is valid in the interval $\left(-\sqrt{2n}, \sqrt{2n}\right)$ where the zeros of $H_n(x)$ are located and therefore can be used to study the asymptotic behavior of the zeros.

In Fig. 1 we graph

$$H_n\left[\sqrt{2n}\sin(\theta)\right]\exp\left\{-\frac{n}{2}[\ln(2n) - \cos(2\theta)]\right\}$$

and

$$\sqrt{\frac{2}{\cos(\theta)}}\cos\left\{n\left[\frac{1}{2}\sin(2\theta) + \theta - \frac{\pi}{2}\right] + \frac{\theta}{2}\right\},$$

with $n = 20$.

The approximation is very good throughout the interval $\left(0, \sqrt{2n}\right)$ and it breaks down when x approaches the value $\sqrt{2n}$, in the neighborhood of which a new formula in terms of the Airy function needs to be considered (see [12, Theorem 3]).

We note that writing

$$\xi = \sqrt{2n}\sin(\theta)$$

and considering the leading term of (3) as $n \to \infty$, we obtain

$$H_n(\xi) \sim \sqrt{2}\exp\left\{\frac{n}{2}[\ln(2n) - 1] + \frac{\xi^2}{2}\right\}\cos\left(n\frac{\pi}{2} - \xi\sqrt{2n}\right),$$

in agreement with formula (4.14.9) in [13].

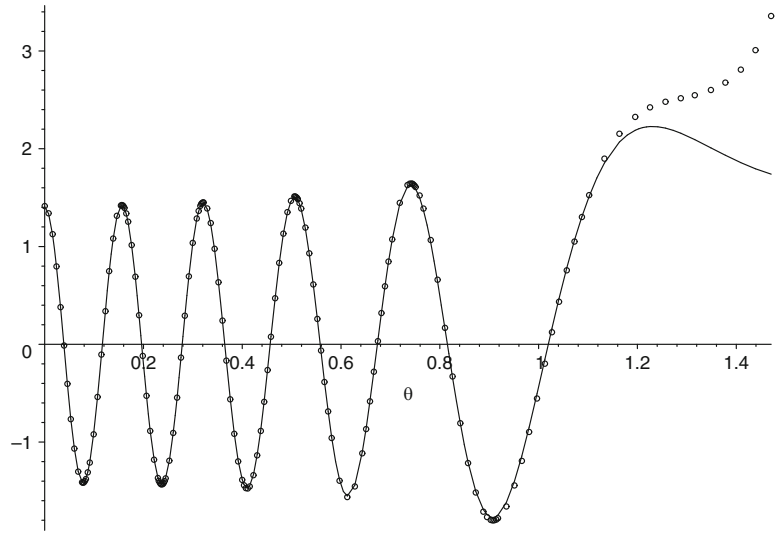


Fig. 1. A comparison of the exact (*solid curve*) and asymptotic (*circle*) values of $H_{20}(x)$ in the interval $\left(0, \sqrt{2n}\right)$

3 Zeros

Let us denote by $\zeta_1^n > \zeta_2^n > \cdots > \zeta_n^n$ the zeros of $H_n(\xi)$, enumerated in decreasing order. It then follows from (3) that $\zeta_j^n \sim \sqrt{2n}\sin(\tau_j^n)$, where τ_j^n satisfies

$$n\left[\frac{1}{2}\sin(2\tau_j^n) + \tau_j^n - \frac{\pi}{2}\right] + \frac{\tau_j^n}{2} = \frac{\pi}{2} - j\pi, \quad 1 \le j \le n.$$

In general, we can rewrite an equation of the form

$$n\left[\frac{1}{2}\sin(2t) + t - \frac{\pi}{2}\right] + \frac{t}{2} = A$$

as Kepler's equation

$$E - \varepsilon\sin(E) = M, \tag{4}$$

with

$$E = 2t, \quad M = 2\frac{2A + n\pi}{2n+1}, \quad \varepsilon = -\frac{2n}{2n+1}. \tag{5}$$

It is well known [14] that the solution of (4) can be expressed as a Kapteyn series

$$E = M + 2\sum_{k=1}^{\infty}\frac{1}{k}\mathrm{J}_k(k\varepsilon)\sin(kM), \tag{6}$$

where $\mathrm{J}_k(\cdot)$ is a Bessel function of the first kind.

Table 1. A comparison of the exact and asymptotic values of the positive zeros of $H_{20}(\xi)$

ζ_j^{20}	$\sqrt{40}\sin(\tau_j^{20})$
5.3875	5.3939
4.6037	4.6056
3.9448	3.9450
3.3479	3.3482
2.7888	2.7891
2.2550	2.2550
1.7385	1.7382
1.2341	1.2337
.73747	.73827
.24534	.24532

Thus, using (5) in (6) with $A = \frac{\pi}{2} - j\pi$, we obtain

$$\tau_j^n = \pi\frac{1+n-2j}{2n+1} + \sum_{k=1}^{\infty}\frac{1}{k}\mathrm{J}_k\left(-\frac{2n}{2n+1}k\right)\sin\left(2\pi\frac{1+n-2j}{2n+1}k\right), \qquad (7)$$

for $1 \le j \le n$. Using the reflection formula [15] $\mathrm{J}_k(-x) = (-1)^k\mathrm{J}_k(x)$, we can write (7) as

$$\tau_j^n = \frac{\pi}{2} - \frac{\pi}{2}(4j-1)N^{-1} - \sum_{k=1}^{\infty}\frac{1}{k}\mathrm{J}_k\left[(1-N^{-1})k\right]\sin\left(\frac{4j-1}{N}k\pi\right), \qquad (8)$$

where $N = 2n+1$.

Formula (8) is, to our knowledge, new and has not been considered before. In Table 1 we compare the exact positive zeros of $H_{20}(\xi)$ with the asymptotic approximation (8).

We note that the approximation is worse for $j = 1$ (largest zero) since (3) breaks down as $\theta \to \frac{\pi}{2}$. It would be interesting to see if there exists a Kapteyn series which is *exactly* equal to $\arcsin\left(\frac{\zeta_j^n}{\sqrt{2n}}\right)$ for all values of j and n.

References

1. Ismail, M.E.H.: Classical and quantum orthogonal polynomials in one variable. Encyclopedia of Mathematics and its Applications, vol. 98. Cambridge University Press, Cambridge (2005)
2. Szegő, G.: Orthogonal Polynomials, 4th edn. American Mathematical Society, Providence, R.I. (1975)
3. Hermite, C.: Sur un nouveau développement en série de functions. Compt. Rend. Acad. Sci. Paris. **58**, 93–100 (1864)
4. Van Assche, W.: Entropy of Hermite polynomials with application to the harmonic oscillator. Bull. Belg. Math. Soc. Simon Stevin. (suppl.), 85–96 (1996) Numerical analysis (Louvain-la-Neuve, 1995)

5. Erdélyi, A., Magnus, W., Oberhettinger, F., Tricomi, F.G.: Higher Transcendental Functions, vols. I, II. McGraw-Hill Book Company, Inc., New York-Toronto-London (1953)
6. Calogero, F., Perelomov, A. M.: Asymptotic density of the zeros of Hermite polynomials of diverging order, and related properties of certain singular integral operators. Lett. Nuovo Cimento (2), **23**(18), 650–652 (1978)
7. Dette, H., Studden, W.J.: Some new asymptotic properties for the zeros of Jacobi, Laguerre, and Hermite polynomials. Constr. Approx. **11**(2), 227–238 (1995)
8. Elbert, Á., Muldoon, M.E.: Inequalities and monotonicity properties for zeros of Hermite functions. Proc. Roy. Soc. Edinburgh Sect. A. **129**(1), 57–75 (1999)
9. Gawronski, W.: Strong asymptotics and the asymptotic zero distributions of Laguerre polynomials $L_n^{(an+\alpha)}$ and Hermite polynomials $H_n^{(an+\alpha)}$. Analysis. **13**(1–2), 29–67 (1993)
10. Pittaluga, G., Sacripante, L.: Bounds for the zeros of Hermite polynomials. Ann. Numer. Math. **2**(1–4), 371–379 (1995) Special functions (Torino, 1993)
11. Ricci, P.E.: Improving the asymptotics for the greatest zeros of Hermite polynomials. Comput. Math. Appl. **30**(3–6), 409–416 (1995)
12. Dominici, D.: Asymptotic analysis of the Hermite polynomials from their differential-difference equation. J. Difference Equ. Appl. **13**(12), 1115–1128 (2007)
13. Lebedev, N.N.: Special Functions and their Applications. Dover Publications Inc., New York (1972)
14. Watson, G.N.: A Treatise on the Theory of Bessel Functions. Cambridge Mathematical Library. Cambridge University Press, Cambridge (1995)
15. Spanier, J., Oldham, K.B.: An Atlas of Functions. Hemisphere, New York (1987)

The Error Function in the Study of Singularly Perturbed Convection-Diffusion Problems with Discontinuous Boundary Data

J.L. López[1], E. Pérez Sinusía[1], and N.M. Temme[2]

[1] Dpto. de Ingeniería Matemática e Informática, Universidad Pública de Navarra, Pamplona 31006, Spain, jl.lopez@unavarra.es, ester.perez@unavarra.es
[2] CWI, PO Box 94079, 1090 GB Amsterdam, The Netherlands Nico.Temme@cwi.nl

Summary. We show the importance of the error function in the approximation of the solution of singularly perturbed convection-diffusion problems with discontinuous boundary conditions. It is observed that the error function (or a combination of them) provides an excellent approximation and reproduces accurately the effect of the discontinuities on the behaviour of the solution at the boundary and interior layers.

1 Introduction

We consider the model convection-diffusion problem $-\varepsilon\Delta U+\overrightarrow{v}\cdot\overrightarrow{\nabla}U=0$ in Ω where Ω is an open set in $\mathbb{R}^2$ or $\mathbb{R}^3$, $\varepsilon>0$ and $\overrightarrow{v}$ is a constant vector. Besides the small perturbation parameter ε, other sources of singular behaviour for the solution of singular perturbation problems are the discontinuities of the boundary data. We consider for this problem Dirichlet boundary data piecewise constant: $U|_{\partial\Omega}=0$ or 1 with jump discontinuities (of height 1) at some points in $\partial\Omega$.

In [1–7], we have analyzed this problem in a number of two and three-dimensional unbounded and bounded domains Ω with discontinuous boundary data at $\partial\Omega$. For all these problems, we have found that the solution in the singular limit $\varepsilon\to 0^+$ and away from the discontinuity points of the boundary data can be approximated in the form

$$U=U_0(1+\mathcal{O}(\sqrt{\varepsilon})), \tag{1}$$

where U_0 is an error function or a combination of error functions. In the next section we describe the asymptotic approximation and layer structure of the solutions found in examples considered in our earlier papers. In the conclusion section we discuss the (in our opinion) universality of the complementary error function as basic approximant of the solution of this kind of problems.

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*,
Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_9,

In what follows, we will consider the polar coordinates $x = r\sin\phi$, $y = r\cos\phi$, $\overrightarrow{r} := (x, y)$, $w := 1/(2\varepsilon)$ and $\zeta(x, y) := \sqrt{r - x\sin\beta - y\cos\beta}$. In all the problems analyzed we consider $U \in \mathcal{C}(\tilde{\Omega}) \cap \mathcal{D}^2(\Omega)$ and U bounded on bounded subsets of $\tilde{\Omega}$, where $\tilde{\Omega}$ is the closed set $\bar{\Omega}$ indented at the discontinuity points of the boundary conditions.

2 Examples in Two-Dimensional Domains

In this section we consider $\overrightarrow{v} = (\sin\beta, \cos\beta)$, with $\beta \in [0, \pi/2)$.

2.1 A Quarter Plane

For $(x, y) \in \tilde{\Omega}_1 := \bar{\Omega}_1 \setminus \{(0, 0)\}$ and $0 \le \beta < \pi/2$ the solution of the problem

$$\begin{cases} -\varepsilon\Delta U + \overrightarrow{v} \cdot \overrightarrow{\nabla} U = 0, & (x, y) \in \Omega_1 := (0, \infty) \times (0, \infty), \\ U(x, 0) = 0, \quad U(0, y) = 1, \end{cases} \tag{P_1}$$

can be approximated in $\tilde{\Omega}_1$ by (1) with $U_0(x, y) = \operatorname{erfc}[\sqrt{w\zeta(x, y)}]$ for $\beta = 0$ and

$$U_0(x, y) = \frac{1}{2}\operatorname{erfc}\left[\sqrt{w\zeta(x, y)}\right], \qquad \text{for } 0 < \beta < \pi/2.$$

Then, the first order approximation is a complementary error function that exhibits an interior layer of width $\mathcal{O}(\sqrt{\varepsilon})$ and parabolic level lines of equation $r - \overrightarrow{v} \cdot \overrightarrow{r} = C \cdot \varepsilon$ near the half-line $t\overrightarrow{v}$, $t > 0$ (see Fig. 1b). For further details we refer to [1].

2.2 An Infinite Strip

For $(x, y) \in \tilde{\Omega}_2 := \bar{\Omega}_2 \setminus \{(a, 0), (b, 0)\}$ and $0 \le \beta \le \pi/2$, the solution of

$$\begin{cases} -\varepsilon\Delta U + \overrightarrow{v} \cdot \overrightarrow{\nabla} U = 0, & (x, y) \in \Omega_2 := (-\infty, \infty) \times (0, 1), \\ U(x, 0) = \chi_{[a,b]}(x), U(x, 1) = 0, & a < b, \end{cases} \tag{P_2}$$

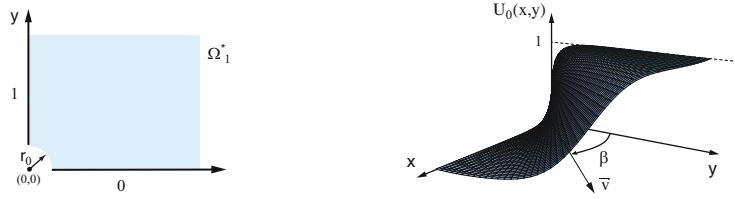


Fig. 1. (**a**) Indented region $\tilde{\Omega}_1$ (**b**) First order approximation $U^0_{\pi/4}(x, y)$ to the solution of (P_1) for $\varepsilon = 0.1$ and $\beta = \pi/4$. Near the half-line $t\overrightarrow{v}$, $t > 0$ an internal parabolic layer occurs

in $\tilde{\Omega}_2$ is of the form (1) with

$$
\begin{aligned}
U_0(x,y) = \frac{1+\delta_{\beta,\pi/2}}{2} \Bigg\{ & \operatorname{sign}\left[\beta - \arctan\left(\frac{x-a}{y}\right)\right] \operatorname{erfc}\left(\sqrt{w\zeta(x-a,y)}\right) \\
& - e^{2(y-1)w\cos\beta} \operatorname{sign}\left[\beta - \arctan\left(\frac{x-a}{2-y}\right)\right] \operatorname{erfc}\left(\sqrt{w\zeta(x-a,2-y)}\right) \\
& - \operatorname{sign}\left[\beta - \arctan\left(\frac{x-b}{y}\right)\right] \operatorname{erfc}\left(\sqrt{w\zeta(x-b,y)}\right) \\
& + e^{2(y-1)w\cos\beta} \operatorname{sign}\left[\beta - \arctan\left(\frac{x-b}{2-y}\right)\right] \operatorname{erfc}\left(\sqrt{w\zeta(x-b,2-y)}\right) \Bigg\} \\
& + \frac{1}{2}\left[\chi_A(x,y) + \chi_{A_0}(x,y) - e^{2(y-1)w\cos\beta}\left(\chi_B(x,y) + \chi_{B_0}(x,y)\right)\right].
\end{aligned}
$$

The region A is limited by the lines $y = 0$, $y = 1$, $x = a + y\tan\beta$ and $x = b + y\tan\beta$. Region B is limited by the lines $y = 0$, $y = 1$, $x = a + (2-y)\tan\beta$ and $x = b + (2-y)\tan\beta$.

In this case, U_β^0 is a combination of four error functions plus step functions multiplied by exponential functions of y, and it exhibits two interior layers of width $\mathcal{O}(\sqrt{\varepsilon})$ and level lines of equation $\zeta(x-c,y) = \text{constant}$ with $c = a, b$. It presents a regular boundary layer of width $\mathcal{O}(\varepsilon)$ near the piece of the outflow boundary situated between the points $(a+\tan\beta, 1)$ and $(b+\tan\beta, 1)$ and it also exhibits two corner layers of area $\mathcal{O}(\sqrt{\varepsilon}) \times \mathcal{O}(\varepsilon)$ near the points $(a+\tan\beta, 1)$ and $(b+\tan\beta, 1)$ (see Fig. 2b). The reader is referred to [1] for further information.

2.3 A Rectangle

For $(x,y) \in \tilde{\Omega}_3 := \bar{\Omega}_3 \setminus \{(0,0), (\pi a, 0)\}$ and $\beta \in (0, \pi/2]$, the solution of

$$
\begin{cases}
-\varepsilon \Delta U + \overrightarrow{v} \cdot \overrightarrow{\nabla} U = 0, & (x,y) \in \Omega_3 := (0, \pi a) \times (0, \pi), \\
U(x,0) = 1, & \\
U(x,\pi) = U(0,y) = U(\pi a, y) = 0, &
\end{cases} \tag{P_3}
$$

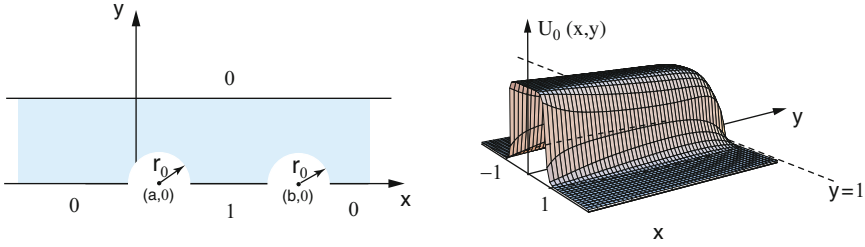


Fig. 2. (**a**) Indented region $\tilde{\Omega}_2$ (**b**) Graph of the first order approximation, $U_\beta^0(x,y)$, to the solution of the problem (P_2) for $\varepsilon = 0.1$ and $\beta = 0$

where $0 \le \beta < 2\pi$ and $a > 0$, can be approximated in $\tilde{\Omega}_3$ by (1) with

$$U_0(x,y) = e^{wy\cos\beta}\frac{\sinh[(\pi-y)w\cos\beta]}{\sinh[\pi w\cos\beta]} \times \Big\{\chi_A(x,y) - e^{2w(x-\pi a)\sin\beta}\chi_B(x,y) + \frac{(1+\delta_{\beta,\pi/2})}{2}\Big[\operatorname{sign}\Big(\beta - \arctan\Big(\frac{x}{y}\Big)\Big)\operatorname{erfc}\sqrt{w\zeta(x,y)} - e^{2(x-\pi a)w\sin\beta}\operatorname{sign}\Big(\beta - \arctan\Big(\frac{2\pi a - x}{y}\Big)\Big)\operatorname{erfc}\sqrt{w\zeta(2\pi a - x,y)} + e^{2(x-\pi a)w\sin\beta}\operatorname{sign}\Big(\beta - \arctan\Big(\frac{\pi a - x}{y}\Big)\Big)\operatorname{erfc}\sqrt{w\zeta(\pi a - x,y)}\Big]\Big\}.$$

The regions A and B are defined by $A := \{(x,y) \in \Omega_3, y < x\cot\beta\}$ and $B := \{(x,y) \in \Omega_3, (\pi a - x)\cot\beta < y < (2\pi a - x)\cot\beta\}$.

Then, the first order approximation of the solution of (P_3) is a linear combination of error functions and elementary functions. The error functions present interior/boundary layers of width $\mathcal{O}(\sqrt{\varepsilon})$. The exponential factors are giving boundary layers of width $\mathcal{O}(\varepsilon)$ (see Fig. 3b). For more details we refer to [5].

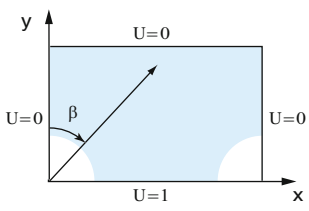
3 Examples in Three-Dimensional Domains

3.1 An Octant

For $(x,y,z) \in \tilde{\Omega}_4 := \Omega_4 \cup \{(x,y,0); x,y>0\} \cup \{(0,y,z); y \ge 0, z>0\} \cup \{(x,0,z); x \ge 0, z>0\}$ (see Fig. 4a), the solution of the problem

$$\begin{cases} -\varepsilon\Delta U + U_z = 0, & \text{in } \Omega_4 := (0,\infty)\times(0,\infty)\times(0,\infty), \\ U(x,y,0) = 1, U(0,y,z) = U(x,0,z) = 0, & \text{for } (x,y,z) \in \tilde{\Omega}_4, \end{cases} \tag{P_4}$$

can be approximated in $\tilde{\Omega}_4$ by (1) with



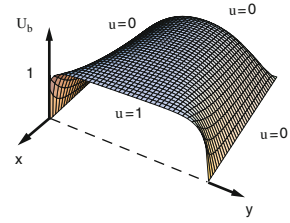


Fig. 3. (**a**) Indented region $\tilde{\Omega}_3$ (**b**) First order approximation $U^0_\beta(x,y)$ to problem (P_3) for $\varepsilon = 0.1$ and $\beta = 0$

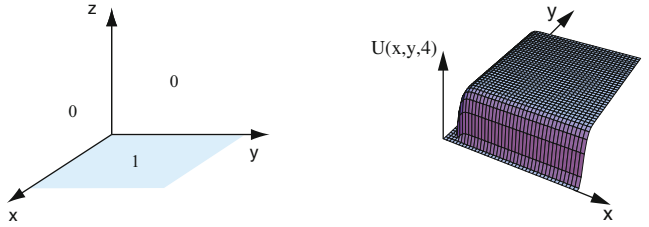


Fig. 4. (**a**) Domain Ω_4 and Dirichlet conditions of problem (P_4) (**b**) Graph of the first order approximation for the solution of problem (P_4) for $\varepsilon = 0.1$

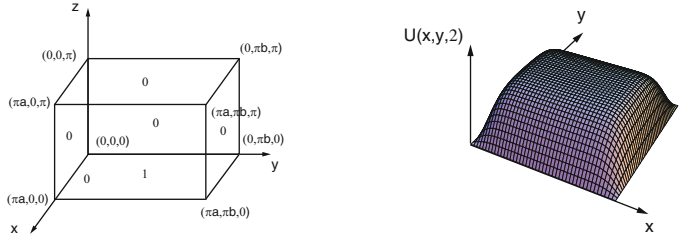


Fig. 5. (**a**) Domain Ω_5 and Dirichlet conditions of problem (P_5) (**b**) Graph of the first order approximation for the solution of problem (P_5) for $\varepsilon = 0.1$

$$U_0(x,y,z) = \operatorname{erf}\sqrt{w\zeta(x,z)}\operatorname{erf}\sqrt{w\zeta(y,z)}.$$

The solution of this problem has boundary layers along the planes $x = 0$ and $y = 0$ of size $\mathcal{O}(\sqrt{\varepsilon})$ (see Fig. 4b). For further information consult [4,6].

3.2 A Cuboid

For $(x,y,z) \in \widetilde{\Omega}_5 := \bar{\Omega}_5 \setminus \big\{\{(0,y,0),(\pi a,y,0); 0 \le y \le \pi b\} \cup \{(x,0,0), (x,\pi b,0); 0 \le x \le \pi a\}\big\}$ (see Fig. 5a), the solution of the problem

$$\begin{cases} -\varepsilon\Delta U + U_z = 0 & \text{in } \Omega_5 := (0,\pi a)\times(0,\pi b)\times(0,\pi), \\ U(0,y,z) = U(\pi a,y,z) = U(x,0,z) = 0, & \\ U(x,\pi b,z) = U(x,y,\pi) = 0, U(x,y,0) = 1, & \text{for } (x,y,z) \in \widetilde{\Omega}_5. \end{cases} \tag{P_5}$$

can be approximated in $\tilde{\Omega}_5$ by (1) with

$$\begin{aligned} U_0(x,y,z) = &\Big[\operatorname{erfc}\sqrt{\omega\zeta(x,z)} - \operatorname{erf}\sqrt{\omega\zeta(x-\pi a,z)}\Big] \\ &\times \Big[\operatorname{erfc}\sqrt{\omega\zeta(y,z)} - \operatorname{erf}\sqrt{\omega\zeta(y-\pi b,z)}\Big] \\ &\times \frac{e^{\omega z}\sinh[\omega(\pi - z)]}{\sinh[\omega\pi]}. \end{aligned}$$

Then, the first order approximation of the solution of problem (P_5) is a combination of products of error functions. See [7] for further information.

4 Conclusions

It is clear from the above examples that the (complementary) error function plays a fundamental role in the approximation of these problems in $\tilde{\Omega}$ (away from the discontinuities of the boundary conditions) as $\varepsilon \to 0^+$. It seems that the error function shows up as a universal approximant. But this fact is not surprising: the complementary error function

$$u(x, y; \tilde{x}, \tilde{y}) := \frac{1}{2} \operatorname{erfc}\left[\frac{1}{\sqrt{2\varepsilon}} \zeta(x - \widetilde{x}, y - \widetilde{y})\right], \qquad (\tilde{x}, \tilde{y}) \quad \text{fixed}, \qquad (2)$$

is an exact solution of the 2D convection-diffusion partial differential equation with constant convection vector $\overrightarrow{v}$ and satisfies approximately the Dirichlet data: consider the line defined by the convection vector $\overrightarrow{v}$ emanating from a discontinuity point $(\widetilde{x}, \widetilde{y}) \in \partial\Omega$ defined by $\{(x, y) \,|\, \zeta(x - \widetilde{x}, y - \widetilde{y}) = 0\}$, then $u(x, y; \tilde{x}, \tilde{y}) \simeq 0$ at one side of this line and $u(x, y; \tilde{x}, \tilde{y}) \simeq 1$ at the other side, approximating in this way the boundary condition that only takes the values 0 or 1. Moreover, this function always lies between the values 0 and 1 and those limiting values are approached rapidly. It exhibits a rapid transition from one value to another when $\overrightarrow{r}$ crosses the above mentioned line. A "maximum principle" states that the solution of this problem must have its values between 0 and 1, so function (2) reproduces this property of the exact solution. Furthermore, the arguments of the complementary error function describe approximately the shape and size of the singular layers as well as their location. The singular parameter ε controls the incline of the singular layers: the smaller ε is, the steepest the shape of U is on the singular layer. The size of the transition region (singular layer) is $\mathcal{O}(\sqrt{\varepsilon})$.

The layer structure of the solution of these problems is described by the (complementary) error function or combinations of error functions. From a numerical point of view, these functions can be very useful to design stable numerical methods. For the construction of local grids their arguments give an idea about the mesh size and location of refined meshes. Moreover, in the analysis of any numerical method it is important to derive sharp bounds for the derivatives of the solutions in terms of ε. The derivation of the approximations obtained in [1–7] may be used to obtain those bounds.

References

1. López, J.L., Pérez Sinusía, E.: Stud. Appl. Math. **113**(1), 57–89 (2004)
2. López, J.L., Pérez Sinusía, E.: Acta Appl. Math. **82**(1), 101–117 (2004)
3. López, J.L., Pérez Sinusía, E.: J. Comp. Appl. Math. **181**, 1–23 (2005)
4. López, J.L., Pérez Sinusía, E., Temme, N.M.: Stud. Appl. Math. **116**, 303–319 (2006)
5. López, J.L., Pérez Sinusía, E.: P. Roy. Soc. Edinb. A. **137A**, 93–109 (2007)
6. López, J.L., Pérez Sinusía, E., Temme, N.M.: J. Math. Anal. Appl. **328**(2), 931–945 (2007)
7. López, J.L., Pérez Sinusía, E.: IMA J. Appl. Math. **74**, 35–45 (2009)

Singular Perturbations of Parabolic Equations With or Without Boundary Layers

Denis R. Akhmetov[1], Mikhail M. Lavrentiev, Jr.[1], and Renato Spigler[2]

[1] Sobolev Institute of Mathematics SD RAS and Novosibirsk State University, Novosibirsk, Russia denis_r_akhmetov@yahoo.com, mmlavr@nsu.ru
[2] University "Roma Tre", Rome, Italy spigler@mat.uniroma3.it

1 Introduction

Singular perturbations of partial differential equations (PDEs) are encountered due to the nature of certain physical models (e.g., small viscosity in Navier–Stokes equations), or to analyze some asymptotic limiting behavior (long time, long distances). In such cases, sometimes, certain usually nondimensional groups of terms are first identified, e.g., electron to ion mass ratio. Besides, singular perturbations are encountered for regularization purposes, e.g., in the numerical treatment of hyperbolic or ultraparabolic PDEs, like the Fokker-Planck equation, through parabolic regularization.

We should remind that by "singular perturbation" we refer to cases when the order of a given PDE formally drops when, e.g., a certain parameter is set to zero, hence the order of the PDEs or its type changes.

Applications are numerous. Just recall an industrial application, that to industrial plasma – wall interaction in semiconductor etching [1]. This is only to stress the importance of the boundary conditions (BCs) at the wall for kinetic equations.

In both, regular and singular perturbation problems, only in few instances, mostly with ordinary differential equations (ODEs), full asymptotic series expansions in the (e.g.) smallness parameter can be obtained. In the ODEs theory, we can identify two steps: a formal part and analytic validity part.

Most often, such expansions do not exist, and one should be satisfied with "lowest order" information, i.e., just obtaining the limiting behavior of solutions.

2 Boundary Layer or Not?

In singular perturbation problems, as a rule, boundary layers (and/or internal layers) arise as the small parameter goes to zero. The solution behaves differently in some subdomains of the space domain, and this requires a different

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*,
Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_10,

asymptotic treatment. This is due to a nonuniform behavior on the entire space-domain. There are however cases when no boundary layer arises.

In [2], suitable approach was proposed for the search of conditions on data of certain given singularly perturbed problems, under which *no* boundary layer exists (i.e., is not required). In such cases, there exists a sequence of regularized solutions, u^{ε_n}, which converges *uniformly* to a solution, v, of the corresponding limiting problem, as $\varepsilon_n \to 0^+$. Such conditions are suitable "higher-order compatibility conditions" on the boundary data.

3 Boundary Layer

An example of this kind (indeed, the most frequent) is the linear Fokker-Planck equation on a half-space, in the Kramers–Smoluchoswski limit [3],

$$\varepsilon^2 \frac{\partial f}{\partial t} + \varepsilon v \frac{\partial f}{\partial x} = \frac{\partial^2 f}{\partial v^2} + \frac{\partial (vf)}{\partial v}$$

for $t > 0,\ x > 0,\ -\infty < v < +\infty$, $\varepsilon > 0$ small. Of course, here $f = f^\varepsilon$. Another is given by the nonlinear analogue of the previous case [4],

$$\varepsilon^2 \frac{\partial f}{\partial t} + \varepsilon v \frac{\partial f}{\partial x} = \delta(f)\, \frac{\partial^2 f}{\partial v^2} + \theta(f)\, \frac{\partial (vf)}{\partial v},$$

with f replaced by the number density

$$n = n(x,t) := \int_{-\infty}^{+\infty} f(x,v,t)\, dv$$

in δ and θ. In fact, here $f = f^\varepsilon$ and $n = n^\varepsilon$. Again, we have also the nonlinear problem [5–7]

$$\varepsilon^2 \frac{\partial f}{\partial t} + \varepsilon v \frac{\partial f}{\partial x} = \frac{\partial (vf)}{\partial v} + \frac{\partial^2 f}{\partial v^2} + \varepsilon^\alpha \left[F(n) f + S(x,v,t)\right],$$

where is the number density, and $\alpha = 0, 1, 2, \ldots$, e.g. Here $f = f^\varepsilon$ and $n = n^\varepsilon$., ε is related to the mean free path, and the kinetics of certain chemical reactions can be described. According to the relative strength of the nonlinearity $[\alpha]$, transport or chemical reaction effects dominate: the lowest-order density is then governed by pure diffusion, reaction-diffusion, or chemical equilibrium.

4 No Boundary Layer

It was shown that no boundary layer is needed, deriving ultraparabolic equations of the Fokker–Planck type from their parabolic regularizations. The explanation rests on the high *regularity* enjoyed by solutions in such problems.

In [2], the conjecture was that, if a given singularly perturbed problem, subject to certain initial condition (IC) and BCs, (1) has a unique solution, $u^\varepsilon(x,y,t)$, for every $\varepsilon > 0$, and (2) there exists a unique solution, $v(x,y,t)$, to the reduced equation, subject to *the same* IC and BCs, (that is to the equation obtained setting formally $\varepsilon = 0$), then the passage to the limit as $\varepsilon \to 0^+$ does *not* require any boundary layer.

Simple model examples are:

1. A hyperbolic limiting equation,

$$u_t = \varepsilon u_{yy} + u_y, \quad \text{on } \Pi := \{(y,t) \in [0,1] \times [0,+\infty)\},$$

with

$$(u, u_y)|_{y=0} = (u, u_y)|_{y=1}, \quad u(y,0) = \varphi(y),$$

being $\varphi(y) \in C^\infty(\mathbf{R})$ and periodic with period 1.

Then, $u^\varepsilon \in C^\infty(\Pi)$ and

$$\|u^\varepsilon(y,t)\|_{C^k(\Pi)} \le M_k,$$

for all $k = 1, 2, \ldots$, and all $\varepsilon \in (0,1)$, M_k being independent of ε.

2. On $Q_\infty := \{(x,y,t) \in [0,1]^2 \times [0,+\infty)$, an ultraparabolic limiting equation,

$$u_t = u_{xx} + \varepsilon u_{yy} + u_y,$$

with $u|_{x=0} = 0$, $u|_{x=1} = 0$, $(u, u_y)|_{y=0} = (u, u_y)|_{y=1}$, $u(x,y,0) = \varphi(x,y)$.

Then, if the compatibility conditions

$$D_x^{2n}\varphi|_{x=0} = 0, \ D_x^{2n}\varphi|_{x=1} = 0$$

hold for $y \in \mathbf{R}$ and all $n \in \mathbf{N}_0$, then there exists a unique classical solution $u^\varepsilon(x,y,t) \in C^\infty(Q_\infty)$ with

$$\|u^\varepsilon(x,y,t)\|_{C^k(Q_\infty)} \le M_k.$$

for all $k = 1, 2, \ldots$, and all $\varepsilon \in (0,1)$, M_k being independent of ε. Note that here we have Dirichlet data in x and periodicity in y. The compatibility conditions above are the necessary conditions in order to obtain a classical smooth solution.

3. Three BV problems were considered, on $Q_T := \{(x,y,t) \in [0,1]^2 \times [0,T]\}$, for limiting ultraparabolic equations, like

$$v_t + k(x,y,t)v_y = a(x,y,t)v_{xx} + b(x,y,t)v_x + c(x,y,t)v + f(x,y,t),$$

using the notation: $\Gamma := \Gamma_1 \cup \Gamma_2$, being Γ_1 and Γ_2 that part of the boundary of Q_T where $y = 0$ and $k(x,y,t) > 0$, and where $y = 1$ and $k(x,1,t) < 0$, respectively (which can be empty).

3.1. Dirichlet BCs in x:

$$v|_{x=0} = 0, \quad v|_{x=1} = 0, \quad v|_\Gamma = \psi(x,y,t), \quad v(x,y,0) = \varphi(x,y).$$

3.2. Neumann BCs in x:

$$v_x|_{x=0} = 0, \quad v_x|_{x=1} = 0, \quad v|_\Gamma = \psi(x,y,t), \quad v(x,y,0) = \varphi(x,y).$$

3.3. Mixed-type BCs in x:

$$[v_x + \beta(x,y,t)v]_{x=0} = 0, \quad [v_x + \gamma(x,y,t)v]_{x=1} = 0,$$

$$v|_\Gamma = \psi(x,y,t), \quad v(x,y,0) = \varphi(x,y).$$

There exists a unique (weak) solution to each of these problems.

Three singularly perturbed problems for the parabolic equation

$$u_t + k(x,y,t)u_y = \varepsilon u_{yy} + a(x,y,t)u_{xx}$$

$$+ b(x,y,t)u_x + c(x,y,t)u + f(x,y,t), \tag{1}$$

with $a(x,y,t) \geq a_0 > 0$ (of course, here $u = u^\varepsilon$) were considered:
A) on the unbounded domain $Q := \{(x,y,t) \in \mathbf{R} \times [0,1] \times [0,T]\}$, with periodic BCs in y, $(u,u_y)|_{y=0} = (u,u_y)|_{y=1}$, and IV $u(x,y,0) = \varphi(x,y)$;
B) on the bounded domain $Q_T := \{(x,y,t) \in [0,1]^2 \times [0,T]\}$, with homogeneous BCs in x and periodic BCs in y, $u|_{x=0} = 0$, $u|_{x=1} = 0$, $(u,u_y)|_{y=0} = (u,u_y)|_{y=1}$, and IV $u(x,y,0) = \varphi(x,y)$;
C) a Dirichlet problem on the bounded domain Q_T, with homogeneous BCs,

$$u|_{x=0} = 0, \ u|_{x=1} = 0, \quad u|_{y=0} = 0, \ u|_{y=1} = 0,$$

and IV $u(x,y,0) = \varphi(x,y)$.

The purpose is to find conditions on the data under which no boundary layer is needed, when $\varepsilon \to 0^+$.
Hypothesis 1. – If all coefficients and source term in (1) are periodic in y, and the IV satisfies the corresponding compatibility conditions, then the singular perturbation problems (A) and (B) have *no* boundary layers. The same is true for problem (C), if

$$k|_{y=0} = f|_{y=0} = 0, \quad k|_{y=1} = f|_{y=1} = 0, \tag{2}$$

which are the Dirichlet BCs (w.r.t. y) for k and f.
Hypothesis 2. – If all coefficients and source term in (1) satisfy the conditions in (2), and if the IC, φ, satisfies the corresponding compatibility condition, then the singular perturbation problem has *no* boundary layer. These hypotheses have been proved to be true in precise theorems.

In [2], it was suggested that no boundary layers are required whenever singularly perturbed and limiting problems both possess a unique solution, subject to all the same boundary conditions. In that statement, absence of boundary layers should be understood as having solutions uniformly bounded

along with its first derivatives. However, "weak" boundary layers – so to say – i.e., the occurrence of possible unboundedness of the second derivative is not forbidden. This case was kindly pointed by Martin Stynes.

It may be considered remarkable that the previous phenomenon occurs even with some *nonlinear* equations, namely the nonlinear integro-differential equation

$$\frac{\partial f}{\partial t} = \frac{\partial^2 f}{\partial \omega^2} + \frac{\partial}{\partial \omega}\left[(\omega - \Omega - \mathcal{K}(\theta, t))f\right] - \omega \frac{\partial f}{\partial \theta},$$

on the unbounded slab $Q_T := \{(\theta, \omega, t, \Omega) \in [0, 2\pi] \times \mathbf{R} \times [0, T] \times [-G, G]\}$, where

$$\mathcal{K}(\theta, t) := K \int_{-G}^{G} \int_{-\infty}^{+\infty} \int_{0}^{2\pi} g(\Omega') \sin(\theta' - \theta) f(\theta', \omega', t, \Omega')\, d\theta' d\omega' d\Omega',$$

subject to the IC and BCs $f|_{t=0} = f_0(\theta, \omega, \Omega)$, $f|_{\theta=0} = f|_{\theta=2\pi}$, [2,8,11]. This equation governs the time evolution of populations of nonlinearly coupled random oscillators (a generalization of the Kuramoto equation). Indeed, we observed the "phenomenon of no boundary layer" first when studying such problem, and what is remarkable is, rather, that it occurs in linear problems.

Therefore we have the Cauchy problem

$$(\mathcal{P}^\varepsilon) \qquad L^\varepsilon u^\varepsilon = f(x, y, t) \quad \text{in } H_T \cap \{t > 0\},$$
$$u^\varepsilon(x, y, 0) = \varphi(x, y) \quad \text{for } (x, y) \in \mathbf{R}^2,$$

where $H_T := \mathbf{R} \times \mathbf{R} \times [0, T]$,

$$L^\varepsilon u^\varepsilon := u^\varepsilon_t + k(x, y) u^\varepsilon_y - \varepsilon u^\varepsilon_{yy} - a(x, y) u^\varepsilon_{xx} - b(x, y) u^\varepsilon_x - c(x, y) u^\varepsilon,$$

and the "reduced problem"

$$(\mathcal{P}^0) \qquad L^0 u^0 = f(x, y, t),$$

etc., *formally* obtained setting $\varepsilon = 0$ in $\mathcal{P}^\varepsilon$ above.

First, ε-*uniform* estimates are obtained for bounded classical solutions of the *parabolic* problem $\mathcal{P}^\varepsilon$ in the anisotropic Sobolev space $W_2^{3,2,1}(\mathcal{Q}_T)$, where $\mathcal{Q}_T := \{(x, y, t) \in \mathbf{R} \times [0, 1] \times [0, T]\}$. Hence, *no* BL is required. Then, using this fact, existence of global in time strong solutions to the *ultraparabolic* problem $\mathcal{P}^0$ has been established. Here, coefficients and r.h.s. are allowed to be unbounded [9]. Results apply even to certain nonlinear integro-differential PDEs, such as that above, generalizing the Kuramoto equation. *Optimal* decay estimates for global in time strong solutions to such equations were also established [10].

Acknowledgements

The authors would like to thank Professor Martin Stynes of the National University of Ireland, Cork, Ireland, for his interest and for suggesting an simple but enlightening example. This concerns a singularly perturbed parabolic problem whose solution is uniformly bounded in the C^1 norm, while its second derivative grows to infinity as the small parameter approaches zero.

References

1. Cardinali, A., Matte, J.P., Shoucri, M., Spigler, R.: Numerical study of plasma-wall transition using an Eulerian Vlasov code. Eur. Phys. J. D. **30**, 81–92 (2004)
2. Akhmetov, D.R., Lavrentiev, M.M. Jr., Spigler, R.: Singular perturbations for certain partial differential equations without boundary-layer. Asymptot. Anal. **35**(1), 65–89 (2003)
3. Spigler, R.: Boundary-layer theory in the Kramers-Smoluchowski limit for the Fokker-Planck equation on a half-space. Boll. Un. Mat. Ital. (7) **1-B**, 918–938 (1987)
4. Spigler, R.: A boundary-layer theory for the nonlinear Fokker-Planck equation on the half-pace. Boundary and interior layers – computational and asymptotic methods, Proceedings of the 5th International Conference, BAIL-V (Shanghai/China, 1988), 326–331, Boole Press Conference Series 12, Boole, Dún Laoghaire (1988)
5. Spigler, R., Zanette, D.H.: Reaction-diffusion models from the Fokker-Planck formulation of chemical processes. IMA J. Appl. Math. **49**, 217–229 (1992)
6. Spigler, R., Zanette, D.H.: Asymptotic analysis and reaction-diffusion approximation for BGK kinetic models of chemical processes in multispecies gas mixtures. Z. angew. Math. Phys. (ZAMP) **44**, 812–827 (1993)
7. Spigler, R., Zanette, D.H.: A BGK model for chemical processes: The reaction-diffusion approximation. M^3AS: Math. Models Methods Appl. Sci. **4**, 35–47 (1994)
8. Acebrón, J.A., Bonilla, L.L., Pérez Vicente, C.J., Ritort, F., Spigler, R.: The Kuramoto model: a simple paradigm for synchronization phenomena. Rev. Modern Phys. **77**, 137–185 (2005)
9. Akhmetov, D.R., Lavrentiev, M.M. Jr., Spigler, R.: Singular perturbations for parabolic equations with unbounded coefficients leading to ultraparabolic equations. Differ. Integr. Equ. **17**(1–2), 99–118 (2004)
10. Akhmetov, D.R., Spigler, R.: Uniform and optimal estimates for solutions to singularly perturbed parabolic equations. J. Evol. Equ. **7**, 347–372, (2007)
11. Akhmetov, D.R., Lavrentiev, M.M. Jr., Spigler,R.: Existence and uniqueness of classical solutions to certain nonlinear integrodifferential Fokker-Planck-type equations. Electron. J. Differ. Equ., **2002**(24), 1–17 (2002)

The Asymptotic Inversion of Certain Cumulative Distribution Functions

Amparo Gil[1], Javier Segura[2] and Nico Temme[3]

[1] Departamento de Matemática Aplicada y Ciencias de la Computación, ETSI Caminos, Canales y Puertos, Universidad de Cantabria, 39005-Santander, Spain, amparo.gil@unican.es
[2] Departamento de Matemáticas, Estadística y Computación, Universidad de Cantabria, 39005 Santander, Spain, javier.segura@unican.es
[3] CWI, PO Box 94079, 1090 GB Amsterdam, The Netherlands, nico.temme@cwi.nl

Summary. The inversion of cumulative distribution functions is an important topic in statistics, probability theory and econometrics, in particular for computing percentage points of the distribution functions. The numerical inversion of these distributions needs accurate starting values, and for the standard distributions powerful asymptotic formulas can be used to obtain these values. It is explained how a uniform asymptotic expansions of a standard form representing several well-known distribution functions can be used for the asymptotic inversion of these functions. As an example we consider the inversion of the hyperbolic cumulative distribution function.

1 Introduction

We consider functions of the form

$$F_a(\eta) = \sqrt{\frac{a}{2\pi}} \int_{-\infty}^{\eta} e^{-\frac{1}{2}a\zeta^2} f(\zeta)\, d\zeta, \tag{1}$$

where $a > 0$, $\eta \in \mathbb{R}$, and f is analytic and real on $\mathbb{R}$ with $f(0) = 1$.

The special case $f = 1$ gives the normal distribution

$$P(\eta\sqrt{a}) = \sqrt{\frac{a}{2\pi}} \int_{-\infty}^{\eta} e^{-\frac{1}{2}a\zeta^2}\, d\zeta = \tfrac{1}{2}\mathrm{erfc}\left(-\eta\sqrt{a/2}\right), \tag{2}$$

where $\mathrm{erfc}\, z$ is the complementary error function

$$\mathrm{erfc}\, z = \frac{2}{\sqrt{\pi}} \int_{z}^{\infty} e^{-t^2}\, dt. \tag{3}$$

As shown in [1, 2] and [3, Chap. 10] the incomplete gamma functions and the incomplete beta function – which are the basic functions for several

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*, Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_11,

distribution functions – can be written in this form. In these references we have used uniform asymptotic expansions for inverting these distribution functions for large values of one or two parameters.

We explain how the incomplete gamma function

$$P(a,x)=\frac{1}{\Gamma(a)}\int_0^x t^{a-1}e^{-t}\,dt \tag{4}$$

can be written in the standard form (1). Let $\lambda=\frac{x}{a}$ and $t=a\tau$. Then

$$\Gamma^*(a)P(a,x)=\sqrt{\frac{a}{2\pi}}\int_0^\lambda e^{-a(\tau-\ln\tau-1)}\,\frac{d\tau}{\tau}, \tag{5}$$

where

$$\Gamma^*(a)=\Gamma(a)a^{-a}e^a\sqrt{\frac{a}{2\pi}}=1+\mathcal{O}(1/a). \tag{6}$$

The transformation

$$\tau-\ln\tau-1=\tfrac{1}{2}\zeta^2,\quad \operatorname{sign}(\tau-1)=\operatorname{sign}(\zeta) \tag{7}$$

gives the standard form

$$\Gamma^*(a)P(a,x)=\sqrt{\frac{a}{2\pi}}\int_{-\infty}^\eta e^{-\frac{1}{2}a\zeta^2}f(\zeta)\,d\zeta,\quad f(\zeta)=\frac{1}{\tau}\frac{d\tau}{d\zeta}, \tag{8}$$

with

$$\lambda-\ln\lambda-1=\tfrac{1}{2}\eta^2,\quad \operatorname{sign}(\lambda-1)=\operatorname{sign}(\eta). \tag{9}$$

2 Asymptotic Representation of $F_a(\eta)$

By using Laplace's asymptotic method (see [4, Chap. 2]) it is not difficult to find the asymptotic estimates for large positive a and fixed values of η:

$$F_a(\eta)=\begin{cases}-f(\eta)/(\eta\sqrt{2a\pi})e^{-a\eta^2}[1+\mathcal{O}(1/a)], & \text{if } \eta<0;\\ \frac{1}{2}+\mathcal{O}(1/\sqrt{a}), & \text{if } \eta=0;\\ 1+\mathcal{O}(1/a), & \text{if } \eta>0.\end{cases} \tag{10}$$

We see that the asymptotic behaviour of $F_a(\eta)$ is completely different in the three cases distinguished. Moreover, the asymptotic forms do not pass into each other when η changes sign. By using an integration by parts procedure we can obtain a single asymptotic representation of $F_a(\eta)$ which is valid for all $\eta\in\mathbb{R}$. We write in (1) $f(\eta)=[f(\eta)-f(0)]+f(0)$, where $f(0)=1$, and use (2). Then we obtain by repeating integration by parts steps:

$$F_a(\eta)=\tfrac{1}{2}\operatorname{erfc}(-\eta\sqrt{a/2})F_a(\infty)\ +\frac{e^{-\frac{1}{2}a\eta^2}}{\sqrt{2\pi a}}S_a(\eta), \tag{11}$$

where, as $a \to \infty$,

$$F_a(\infty) \sim \sum_{n=0}^{\infty} \frac{A_n}{a^n}, \quad A_0 = 1, \quad S_a(\eta) \sim \sum_{n=0}^{\infty} \frac{C_n(\eta)}{a^n}, \tag{12}$$

uniformly with respect to $\eta \in \mathbb{R}$. The coefficients follow from the following recursive scheme. Let $f_0(\eta) = f(\eta)$. Then, for $n = 0, 1, 2, \ldots$, define

$$f_{n+1}(\eta) = \frac{d}{d\eta} \frac{f_n(\eta) - f_n(0)}{\eta}, \tag{13}$$

and we have

$$A_n = f_n(0), \quad C_n(\eta) = \frac{f_n(0) - f_n(\eta)}{\eta}. \tag{14}$$

3 The Asymptotic Inversion Method

Let $p \in (0, 1)$ and a a large positive parameter. Then we are interested in the value η that solves the equation

$$F_a(\eta) = F_a(\infty)\, p. \tag{15}$$

We use the representation in (11) and define a number η_0 that solves the reduced equation

$$\tfrac{1}{2}\mathrm{erfc}(-\eta_0 \sqrt{a/2}) = p. \tag{16}$$

Then for the requested value η we assume the expansion

$$\eta \sim \eta_0 + \frac{\eta_1}{a} + \frac{\eta_2}{a^2} + \frac{\eta_3}{a^3} + \ldots, \quad a \to \infty, \tag{17}$$

and try to find the coefficients $\eta_1, \eta_2, \eta_3, \ldots$. To obtain the η_j we can substitute the expansion for η into (11) and use formal power series manipulations. For the asymptotic inversion of the incomplete gamma and beta functions we have used techniques based on differential equations; see [1,2] and [3, Chap. 10]. In the next section we consider a different example.

The method based on differential equations runs as follows. From (1), (15) and (16) we obtain

$$\frac{dp}{d\eta_0} = \sqrt{\frac{a}{2\pi}} e^{-\frac{1}{2}a\eta_0^2}, \quad \frac{dp}{d\eta} = \sqrt{\frac{a}{2\pi}} \frac{f(\eta)}{F_a(\infty)} e^{-\frac{1}{2}a\eta^2}, \tag{18}$$

from which we obtain, upon dividing,

$$f(\eta) \frac{d\eta}{d\eta_0} = F_a(\infty) e^{\frac{1}{2}a(\eta^2 - \eta_0^2)}. \tag{19}$$

Substituting (17) we obtain for η_1 after perturbation analysis in first order for large a

$$f(\eta_0) = e^{\eta_0 \eta_1} \quad \Longrightarrow \quad \eta_1 = \frac{1}{\eta_0} \ln f(\eta_0). \tag{20}$$

For higher order terms $\eta_j, j \geq 2$, we need in (19) more coefficients in the asymptotic expansion of $F_a(\infty)$ (see (12), (30) and (31)) and we have to expand

$$f(\eta) = f(\eta_0) + (\eta - \eta_0) f'(\eta_0) + \tfrac{1}{2}(\eta - \eta_0)^2 f''(\eta_0) + \ldots. \tag{21}$$

4 The Hyperbolic Cumulative Distribution

The hyperbolic distribution was introduced in [5] and is given by

$$F(y) = C \int_{-\infty}^{y} e^{-\alpha\sqrt{\delta^2 + (x-\mu)^2} + \beta(x-\mu)}\, dx, \quad y \in \mathbb{R}, \tag{22}$$

where $\alpha > 0$, $|\beta| < \alpha$, δ and μ are arbitrarily real constants, and C is the normalizing constant which gives $F(\infty) = 1$. The value of C is given by

$$C = \frac{\omega}{2\alpha\delta^2 K_1(\omega)}, \quad \omega = \delta\sqrt{\alpha^2 - \beta^2}, \tag{23}$$

where $K_1(\omega)$ denotes the modified Bessel function of the third kind of order 1 (see [6, Chap. 9] or [7, Chap. 9]).

4.1 A Few Transformations

We transform the function $F(y)$ into the standard form. Because $|\beta| < \alpha$, we can write $\beta = \alpha \tanh\theta$. We substitute in (22) $x = \mu + \delta \sinh(\theta + t)$, and obtain

$$F(y) = \frac{1}{2K_1(\omega)} \int_{-\infty}^{\tau} e^{-\omega \cosh t} \frac{\cosh(t+\theta)}{\cosh\theta}\, dt, \tag{24}$$

where ω is given in (23) and

$$\tau = \operatorname{arcsinh} \frac{y-\mu}{\delta} - \theta, \quad \cosh\theta = \frac{\alpha}{\sqrt{\alpha^2 - \beta^2}}. \tag{25}$$

Next we use the transformation

$$\cosh t = 1 + 2\zeta^2, \quad \Longrightarrow \quad t = 2\operatorname{arcsinh}\zeta, \tag{26}$$

which gives

$$F(y) = \frac{e^{-\omega}}{K_1(\omega)} \int_{-\infty}^{\eta} e^{-\frac{1}{2}a\zeta^2} f(\zeta)\, d\zeta, \tag{27}$$

where

$$a = 4\omega, \quad \eta = \sinh \tfrac{1}{2}\tau, \tag{28}$$

and

$$f(\zeta) = \frac{1 + 2\zeta^2 + 2\tanh\theta\zeta\sqrt{\zeta^2+1}}{\sqrt{\zeta^2+1}}. \tag{29}$$

We see that $f(0) = 1$ and it follows that we can write $F(y)$ in the form

$$F(y) = \frac{F_a(\eta)}{F_a(\infty)}, \quad F_a(\infty) = \sqrt{\frac{2\omega}{\pi}}e^{\omega}K_1(\omega), \tag{30}$$

where $F_a(\eta)$ has the standard form (1). We have (see [6, Equation (9.7.2)])

$$F_a(\infty) = 1 + \frac{3}{8\omega} + \mathcal{O}(1/\omega^2), \quad \omega \to \infty. \tag{31}$$

It follows also that the inversion problem $F(y) = p$ when a is large can be written in the form (15). When we have found η from the expansion (17), we compute $\tau = 2\text{arcsinh}\,\eta$ and finally (see (25))

$$y = \mu + \delta\sinh(\theta + \tau), \quad \theta = \text{arctanh}\frac{\beta}{\alpha}. \tag{32}$$

4.2 A Numerical Example

When a is large the function $F_a(\eta)$ approaches the unit step function and the numerical inversion needs accurate starting values for, say, Newton's method, in particular when in (15) p is very small or very close to unity.

In [8] analytic approximations for these $p-$values are constructed of the inverse function F^{-1} of $F(y)$ given in (22). With these approximations a numerical algorithm from Mathematica is used to compute the inverse F^{-1} from the differential equation satisfied by this function.

We demonstrate our approach by taking $\alpha = 5$, $\beta = 3$, $\mu = 0$ and $\delta = 1, 10, 100$. These values give $\omega = 4, 40, 400$ and $a = 16, 160, 1600$, respectively.

First we compute η_0 from (16) and next η_1 from (20), with $f(\eta)$ given in (29). The computed value η then follows from (17) (with two terms). Next we compute $\tau = 2\text{arcsinh}\,\eta$ (see (28)), and with τ we can compute y by inverting the second equation in (25) with $\theta = \text{arctanh}(\beta/\alpha)$.

In Table 1 we give for several values of p and δ the computed value y, and the relative error $|F(y) - p|/p$. We observe that the approximations of y become indeed better when the large parameter $a = 4\delta\sqrt{\alpha^2 - \beta^2}$ increases. Also, the approximations are better when $p \sim 1$.

Table 1. Values y and relative errors $\Delta = |F(y) - p|/p$ of the inversion $F(y) = p$, where $F(y)$ is given in (22) for $\alpha = 5$, $\beta = 3$, $\mu = 0$, and several values of δ and p

δ	1		10		100	
p	y	Δ	y	Δ	y	Δ
0.0001	−1.1087	$0.43\,10^{-1}$	1.2413	$0.82\,10^{-3}$	53.110	$0.14\,10^{-4}$
0.1	0.1646	$0.10\,10^{-1}$	5.2635	$0.17\,10^{-3}$	67.317	$0.12\,10^{-4}$
0.2	0.4071	$0.22\,10^{-2}$	6.0654	$0.22\,10^{-3}$	69.985	$0.11\,10^{-4}$
0.3	0.5963	$0.25\,10^{-2}$	6.6627	$0.24\,10^{-3}$	71.931	$0.95\,10^{-5}$
0.4	0.7708	$0.54\,10^{-2}$	7.1866	$0.24\,10^{-3}$	73.608	$0.83\,10^{-5}$
0.5	0.9465	$0.70\,10^{-2}$	7.6884	$0.23\,10^{-3}$	75.188	$0.72\,10^{-5}$
0.6	1.1361	$0.76\,10^{-2}$	8.2023	$0.21\,10^{-3}$	76.779	$0.60\,10^{-5}$
0.7	1.3565	$0.74\,10^{-2}$	8.7664	$0.18\,10^{-3}$	78.496	$0.49\,10^{-5}$
0.8	1.6397	$0.62\,10^{-2}$	9.4462	$0.14\,10^{-3}$	80.523	$0.36\,10^{-5}$
0.9	2.0826	$0.40\,10^{-2}$	10.426	$0.90\,10^{-4}$	83.367	$0.21\,10^{-5}$
0.9999	5.8365	$0.79\,10^{-5}$	16.767	$0.28\,10^{-6}$	99.863	$0.57\,10^{-8}$

References

1. Temme, N.M.: Math. Comp. **58**, 755–764 (1992)
2. Temme, N.M.: J. Comput. Appl. Math. **41**(1–2), 145–157 (1992)
3. Gil, A., Segura, J., Temme, N.M.: Numerical Methods for Special Functions. SIAM, Philadelphia, PA (2007)
4. Wong, R.: Asymptotic approximations of integrals. Classics in Applied Mathematics, vol. 34. SIAM, Philadelphia, PA (2001)
5. Barndorff-Nielsen, O.E.: Proc. Roy. Soc. London Ser. A **353**, 401–419 (1977)
6. Abramowitz, M., Stegun, I.A.: Handbook of mathematical functions. Nat. Bur. Standards AMS, vol. 55. U.S. Govt. Printing Office, Washington, D.C. (1964)
7. Temme, N.M.: Special Functions: An Introduction to the Classical Functions of Mathematical Physics. Wiley, New York (1996)
8. Leobacher, G., Pillichshammer, F.: Computing **69**, 291–303 (2002)

Minisymposium *Asymptotic Properties of Complex Random Systems and Applications*

Malwina J. Luczak

Department of Mathematics, London School of Economics, Houghton Street, London WC2A 2AE, United Kingdom, m.j.luczak@lse.ac.uk

There has been recent intense activity in the study of the asymptotic character of sequences of random processes arising e.g. in computer science, statistical physics and mathematical biology. These may model the emergence of certain graph properties; load-sharing among links or servers; the survival and extinction of species; co-operation and competition in a social context; spread of epidemics; DNA, RNA and amino-acid sequences. Under appropriate conditions, a sequence of processes converges to the solution of a differential equation, which may be interpreted as a functional law of large numbers. Such approximations are of great significance as a way to interpret the qualitative behaviour of a complicated, multi-faceted structure in terms of a considerably simpler one. Unfortunately, it is often difficult to prove their validity, especially when the random process has an unbounded number of components in the limit. We would hope that over the coming years, the intense interest in the field will produce a coherent and widely applicable theory. At present, it often appears that each new problem defies the existing theory in an interesting way. In organising the ECMI Minisymposium 'Asymptotic properties of complex random systems and applications', our motivation was to bring these problems into focus and highlight their importance in modelling of real-world situations. Our aim was thus to generate interest among the applied mathematics community, in the hope that interesting new insights and ideas may result. The following sections summarise the contents of the four talks given.

Andrew Barbour's talk. Andrew Barbour, from the University of Zürich, talked on 'Laws of large numbers for epidemic models with countably many types' (work joint with Malwina Luczak). We establish a quantitative law of large numbers for a large class of stochastic epidemic models. It was previously known that certain host-parasite systems can be approximated by systems of differential equations, but rates of convergence were not available. With such diseases, it is natural to distinguish hosts according to the number of parasites they carry. Since it is not usually possible to prescribe a fixed upper limit for the parasite load, this leads to models with countably infinitely

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*, Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_12,

many types, one for each possible number of parasites. This causes difficulty with many arguments which for finitely many types would be quite standard; in particular, proving limit results is a much more delicate issue. A further difficulty is that the operator driving the deterministic limit is non-Lipschitz.

Carl Graham's talk. Carl Graham, from École Polytechnique, lectured on 'A multiclass mean-field model with graph structure for TCP flows', based on his joint work with Philippe Robert. TCP is one of the core protocols used on the Web and other communication networks. Unlike previous studies of TCP window evolutions, the authors consider interaction between diverse kinds of TCP flows through the congestion they create along flow routes. Resources may consist of switches, buffers, links or processors. Flow characteristics include the route, utilisation of specific resources, and the round trip time (influenced by congestion). A Markovian multi-class mean-field interacting model for the window size evolution of a large number of TCP flows is analysed. In the limit as the numbers of flows in different classes become large while keeping the relative weight of each class fixed, the process converges to a deterministic function solving a non-linear differential equation. Also, the system is chaotic, i.e. different flows become approximately independent.

Petra Berenbrink's talk. Petra Berenbrink, from Simon Fraser University, gave a talk entitled 'Distributed selfish load-balancing', based on joint research with Tom Friedetzky, Iman Hajirasouliha, and Zengjian Hu. A congestion game model is considered with n identical resources and m players with weighted tasks. The system goal is to allocate every task to exactly one resource, and the goal of each selfish player is to be allocated to a resource with minimum total load. Agents migrate from overloaded to underloaded resources in a distributed setting, until the allocation becomes balanced. An allocation is a Nash Equilibrium if no player can benefit from changing their strategy. The authors analyse a simple, decentralised protocol converging to a Nash equilibrium, proving bounds on the rate in terms of n, m and Δ (maximum task weight). Proofs involve analysing a suitable potential function.

Ilkka Norros's talk. Ilkka Norros, from VTT, lectured on 'Features of power-law random graphs' (joint work with Hannu Reittu). A power-law random graph model is considered in a regime where the vertex degree has finite mean and infinite variance. Power-law graphs are commonly used to model inhomogeneous random networks, such as the Internet. These graphs have some remarkable features: e.g. with high probability there are subgraphs with arbitrary edge densities, and the typical distance between a pair of vertices in the giant component of a graphs of size N is $O(\log \log N)$. Also, the random graph has a robust structure in that the deletion of highest-degree vertices does not decrease the relative size of the giant, even though it does cause a moderate increase in distances between vertices.

A Multi-Class Mean-Field Model with Graph Structure for TCP Flows

C. Graham[1] and Ph. Robert[2]

[1] CNRS – École Polytechnique, Route de Saclay, 91128 Palaiseau, France
carl@cmap.polytechnique.fr
[2] INRIA Paris – Rocquencourt, Domaine de Voluceau, 78153 Le Chesnay, France
philippe.robert@inria.fr

Summary. A Markovian mean-field multi-class model for the interaction of several classes of permanent connections in a network is analyzed. Connections create congestion at the nodes they utilize, and adapt their throughput to the congestion they encounter in a way similar to the Transmission Control Protocol (TCP).

1 Introduction

The Internet can be described as a very large distributed system for data transmission, with self-adaptive capabilities to the different congestion events that regularly occur. In this paper, a packet level model of the self-adaptive behavior of data flows submitted to Additive Increase Multiplicative Decrease (AIMD) algorithms, similar to Transmission Control Protocol (TCP), is established and studied. Throughput grows linearly in the number of known successful packet transmissions. When a loss is detected, the throughput is sharply reduced by multiplication by some factor $r < 1$ (usually $1/2$).

Studies up to now usually consider a *single* node carrying *similar* connections, see e.g. Ott et al. [1], Adjih et al. [2], Baccelli et al. [3], Dumas et al. [4], and Guillemin et al. [5].

This proceeding announces *without proof* results in Graham and Robert [6], still work in progress at the time of ECMI 2008, in which the interaction due to the simultaneous transmission of *several* classes of permanent connections is rigorously analyzed. A class of connections is characterized, in particular, by the set of nodes it uses, and how, at those nodes, the connections create some congestion and adapt to the total congestion encountered.

For mean-field limit proofs for systems of statistically *indistinguishable* objects, assuming mean-field limit convergence of initial conditions, Sznitman [7] has developed compactness-uniqueness methods, as well as coupling methods between the system and an i.i.d. system. Mean-field studies of stochastic communication networks have been performed notably by Dobrushin and his co-authors, see Karpelevitch et al. [8]. See also Graham [9].

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*,
Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_13,

The model of interest here features *dissimilar* connections classified in a finite number of *classes* according to their characteristics. Few convergence proofs for such *multi-class* systems exist, and those in Graham and Méléard [10] require a structure lacking here. So, Graham and Robert [6] develop a coupling method which extends the methods in Sznitman [7], yielding more tractable non-linear limit equations.

The scope is then to study the equilibrium behavior of the *limit* system, and hopefully to establish that the equilibrium behavior for a finite number of connections converges to it. This can be seen as the inversion of limits

$$\lim_{N\to\infty}\lim_{t\to\infty} = \lim_{t\to\infty}\lim_{N\to\infty}$$

where t and N are time and size parameters.

We refer to Graham and Robert [6] for a more complete introduction with a survey of the literature in the domain, and rigorous proofs.

2 The Markovian Network Model

The network has $J \geq 1$ nodes and accommodates $K \geq 1$ classes of sizes $N_k \geq 1$ for $1 \leq k \leq K$ of permanent connections (or flows, streams, etc.). Let

$$N = (N_1, \ldots, N_K)\,, \qquad |N| = N_1 + \cdots + N_K\,.$$

We study the connection *transmission rate*, governed by the *window size* restricting the quantity of data allowed to be in transit at one time.

An *allocation matrix* $A = (A_{jk}, 1 \leq j \leq J, 1 \leq k \leq K)$ describes the utilization of nodes by the connections. We have $A_{jk} \geq 0$, and if $w_{n,k} \geq 0$ is the state of the n-th class k connection, its utilization of node j is given by $A_{jk}w_{n,k}$. The total utilization u_j of node j by the various connections is then

$$u_j = \sum_{k=1}^{K}\sum_{n=1}^{N_k} A_{jk}w_{n,k}\,, \qquad 1 \leq j \leq J\,.$$

An example is $A_{jk} = 1$ if a class k connection uses node j and else $A_{jk} = 0$.

The quantity u_j represents the level of congestion at node j, in particular the loss rate of a connection going through it will depend on it. There are functions $a_k : \mathbb{R}_+ \times \mathbb{R}_+^J \to \mathbb{R}_+$ and $b_k : \mathbb{R}_+ \times \mathbb{R}_+^J \to \mathbb{R}_+$ for $1 \leq k \leq K$, such that, when the resource utilization vector of the network is $u = (u_j, 1 \leq j \leq J)$ and the state of a class k connection is w_k then,

- This state increases linearly at rate $a_k(w_k, u)$.
- A loss occurs at rate $b_k(w_k, u)$ and causes a jump from w_k to $r_k w_k$.

A natural form for such functions (with slight abuse of notation) is

$$a_k(w_k, u) = a_k(u), \qquad b_k(w_k, u) = w_k \beta_k(u), \tag{1}$$

$$a_k(u) = \left(\tau_k + \sum_{j=1}^{J} t_{jk}(u_j) \right)^{-1}, \qquad \beta_k(u) = \delta_k + \sum_{j=1}^{J} d_{jk}(u_j), \tag{2}$$

where $\tau_k > 0$ can be interpreted as the round trip time (RTT) between source and destination, and $\delta_k \geq 0$ as the loss rate, of class k connections in a non-congested network, and $t_{jk}(u_j) \geq 0$ as the additional RTT delay and $d_{jk}(u_j) \geq 0$ as the additional loss rate at node j when its utilization is u_j.

The SDE Representation and the Mean-Field Scaling

The Markov process describing the state of the connections is given by

$$W^N(t) = (W^N_{n,k}(t), 1 \leq n \leq N_k,\ 1 \leq k \leq K)\,, \qquad t \in \mathbb{R}_+\,,$$

where $W^N_{n,k}(t)$ is the state of the n-th connection of class k at time t. It can be represented by the solution of a stochastic differential equation (SDE): for $1 \leq k \leq K$ and $1 \leq n \leq N_k$,

$$\begin{aligned} dW^N_{n,k}(t) &= a_k(W^N_{n,k}(t-), U^N(t-))\, dt \\ &\quad -(1 - r_k) W^N_{n,k}(t-) \int \mathbf{1}_{\{0 \leq z \leq b_k(W^N_{n,k}(t-), U^N(t-))\}} \mathcal{N}_{n,k}(dz, dt) \end{aligned} \tag{3}$$

with $U^N(t) = (U^N_j(t), 1 \leq j \leq J)$ and

$$U^N_j(t) = \sum_{k=1}^{K} A_{jk} \sum_{n=1}^{N_k} W^N_{n,k}(t)\,,$$

where $(\mathcal{N}_{n,k}, 1 \leq k \leq K, 1 \leq n \leq N_k)$ are independent Poisson processes with Lebesgue intensity measure on $\mathbb{R}^2_+$. Existence and uniqueness of solutions is classical if a_k is Lipschitz and b_k bounded, $1 \leq k \leq K$.

A scaling is used to reduce the high dimensionality of (3) in order to investigate its qualitative and quantitative properties. It is assumed that

$$N_k \to \infty\,, \quad \frac{N_k}{|N|} = \frac{N_k}{N_1 + \cdots + N_K} \to p_k\,, \qquad 1 \leq k \leq K\,, \tag{4}$$

where $p_k \geq 0$ and $p_1 + \cdots + p_K = 1$. The capacity is accordingly scaled by setting $\overline{U}^N = U^N/|N|$ in the functions a_k and b_k. We obtain the mean-field scaled SDE: for $1 \leq k \leq K$ and $1 \leq n \leq N_k$,

$$\begin{aligned} dW^N_{n,k}(t) &= a_k(W^N_{n,k}(t-), \overline{U}^N(t-))\, dt \\ &\quad -(1 - r_k) W^N_{n,k}(t-) \int \mathbf{1}_{\{0 \leq z \leq b_k(W^N_{n,k}(t-), \overline{U}^N(t-))\}} \mathcal{N}_{n,k}(dz, dt) \end{aligned} \tag{5}$$

with $\overline{U}^N(t) = (\overline{U}^N_j(t), 1 \le j \le J)$ and

$$\overline{U}^N_j(t) = \sum_{k=1}^{K} \frac{N_k}{|N|} A_{jk} \overline{W}^N_k(t) \quad \text{with} \quad \overline{W}^N_k(t) = \frac{1}{N_k} \sum_{n=1}^{N_k} W^N_{n,k}(t)\,.$$

This multi-class mean-field system interacts through the *scaled utilization* vector $\overline{U}^N(t)$, or the *scaled state* vector $\overline{W}^N(t) = (\overline{W}^N_k(t), 1 \le k \le K)$.

3 The Non-Linear Limit Process

When N goes to infinity, in view of (5), mean-field behavior is expected: the connection evolutions should become independent, and for class k connections should converge in law to that of $(W_k(t), t \ge 0)$, where the stochastic process $(W(t), t \ge 0) = \big((W_k(t), 1 \le k \le K), t \ge 0\big)$ solves the non-linear SDE

$$dW_k(t) = a_k\left(W_k(t-), u_W(t)\right)\, dt - (1 - r_k) W_k(t-) \int \mathbf{1}_{\{0 \le z \le b_k(W_k(t-), u_W(t))\}}\, \mathcal{N}_k(dz, dt) \quad (6)$$

for $1 \le k \le K$, with $u_W(t) = (u_{W,j}(t), 1 \le j \le J)$ and

$$u_{W,j}(t) = \sum_{k=1}^{K} A_{jk} p_k \mathbf{E}(W_k(t)),$$

where $(\mathcal{N}_k, 1 \le k \le K)$ are i.i.d. Lebesgue intensity Poisson point processes.

In this *non-linear* SDE, the evolution of the process $(W(t), t \ge 0)$ depends not only on its instantaneous value but also on the mean utilization vector $u(t)$, or on the mean value $\mathbf{E}(W(t))$. Its infinitesimal generator depends at time t *on the law* of $W(t)$ itself, which thus solves non-linear equations.

We seek results valid for a_k and b_k of the form (1)–(2), where b_k has a quadratic behavior. To control the long-time evolution or the stationary behavior of $W(t)$, initial conditions cannot be assumed to be uniformly bounded, so that exponential and Gaussian moment assumptions are introduced.

Condition (C) *Holds for a family of random variables* $\{X^\alpha_0, \alpha \in \mathcal{S}\}$ *in* $\mathbb{R}^K_+$, *for* (b_k), *and for* $\varepsilon > 0$ *when at least one of the two conditions is satisfied:*

1. *for* $1 \le k \le K$, *the function* $b_k : \mathbb{R}_+ \times \mathbb{R}^J_+ \to \mathbb{R}_+$ *is Lipschitz, and*
$$\sup_{\alpha \in \mathcal{S}} \mathbf{E}\big(\exp(\varepsilon \|X^\alpha_0\|)\big) < \infty\,,$$
2. *for* $1 \le k \le K$, $b_k(w, u) = w\beta_k(u)$ *and* $\beta_k : \mathbb{R}^J_+ \to \mathbb{R}_+$ *is Lipschitz, and*
$$\sup_{\alpha \in \mathcal{S}} \mathbf{E}\big(\exp\left(\varepsilon \|X^\alpha_0\|^2\right)\big) < \infty\,.$$

Theorem 1. *If the functions* $a_k : \mathbb{R}_+ \times \mathbb{R}_+^J \to \mathbb{R}_+$, $1 \le k \le K$ *are bounded and Lipschitz and if Condition (C) holds for* W_0, (b_k) *and* $\varepsilon > 0$, *then there is pathwise existence and uniqueness of a solution* $(W(t), t \ge 0)$ *of the non-linear SDE* (6) *starting at* W_0, *with continuous dependence on the initial condition.*

4 The Mean-Field Limit for Converging Initial Data

The fundamental notions of exchangeability and chaoticity, see Aldous [11] and Sznitman [7], must be extended to such multi-class models. We use the notation $\lim_{N\to\infty}$ for the limit along an arbitrary subsequence of $N = (N_k)_{1\le k\le K} \in \mathbb{N}^K$ such that $\min_{1\le k\le K} N_k$ goes to infinity.

Definition 1. *The family of r.v.* $(X_{n,k}, 1 \le n \le N_k,\ 1 \le k \le K)$ *is* multi-exchangeable *if its law is invariant under permutation of the indexes* within *the classes: for* $1 \le k \le K$ *and all permutations* σ_k *of* $\{1, \dots, N_k\}$, *we have*

$$\mathcal{L}\big(X_{\sigma_k(n),k}, 1 \le n \le N_k,\ 1 \le k \le K\big) = \mathcal{L}\big(X_{n,k}, 1 \le n \le N_k,\ 1 \le k \le K\big)\,.$$

A sequence $(X_{n,k}^N, 1 \le n \le N_k,\ 1 \le k \le K)$ *of multi-class random variables indexed by* $N = (N_k)_{1\le k\le K} \in \mathbb{N}^K$ *is* $P_1 \otimes \cdots \otimes P_K$-multi-chaotic *if*

$$\lim_{N\to\infty} \mathcal{L}\big(X_{n,k}^N, 1 \le n \le m,\ 1 \le k \le K\big) = P_1^{\otimes m} \otimes \cdots \otimes P_K^{\otimes m}\,, \qquad \forall m \ge 1\,,$$

where P_k *for* $1 \le k \le K$ *is a probability measure on* $\mathbb{R}_+$.

The following theorem is the main result. It uses the topology of uniform convergence on compact sets for the sample path spaces.

Theorem 2. *In the mean-field scaling* (4), *if*

1. *The initial values* $(W_{n,k}^N(0), 1 \le n \le N_k,\ 1 \le k \le K)$ *are multi-exchangeable and* $P_{1,0} \otimes \cdots \otimes P_{K,0}$*-multi-chaotic, and*
2. *The functions* $a_k : \mathbb{R}_+ \times \mathbb{R}_+^J \to \mathbb{R}_+$, $1 \le k \le K$, *are bounded and Lipschitz, and Condition (C) holds for* $\{W_1^N(0),\ N \in \mathbb{N}^K\}$, (b_k) *and* $\varepsilon > 0$,

then, as N *goes to infinity, the processes*

$$\big((W_{n,k}^N(t), t \ge 0), 1 \le n \le N_k,\ 1 \le k \le K\big)$$

solving the SDE (5) *with initial values* $(W_{n,k}^N(0))$ *are multi-exchangeable and* P_W*-multi-chaotic, where* $P_W = P_{W_1} \otimes \cdots \otimes P_{W_K}$ *is the law of the process* $(W(t), t \ge 0) = ((W_k(t), t \ge 0), 1 \le k \le K)$, *the solution of the non-linear SDE* (6) *with initial law* $P_{1,0} \otimes \cdots \otimes P_{K,0}$.

5 Invariant Laws and a Fixed Point Equation

We consider the probability densities given for $0 < r < 1$ and $\rho > 0$ by

$$H_{r,\rho}(x) = \frac{\sqrt{2\rho/\pi}}{\prod_{n=0}^{+\infty}(1-r^{2n+1})} \sum_{n=0}^{+\infty} \frac{r^{-2n}}{\prod_{k=1}^{n}(1-r^{-2k})} e^{-\rho r^{-2n} x^2/2}, \quad x \in \mathbb{R}_+,$$

which have first moment (expected value)

$$\int_{x\geq 0} x H_{r,\rho}(x)\, dx = \sqrt{\rho}\psi(r), \qquad \psi(r) = \sqrt{\frac{2}{\pi}} \prod_{n=1}^{+\infty} \frac{1-r^{2n}}{1-r^{2n-1}}.$$

Theorem 3. *If the functions a_k and b_k, $1 \leq k \leq K$, are of the form (1) with $\beta_k > 0$, and a_k and β_k are Lipschitz functions and a_k is bounded, then the invariant laws for solutions $(W(t), t \geq 0)$ of (6) are in one-to-one correspondence with the solutions $u = (u_j)_{1\leq j\leq J} \in \mathbb{R}_+^J$ of the fixed point equation*

$$u_j = \sum_{k=1}^{K} A_{jk} p_k \psi(r_k) \sqrt{\frac{a_k(u)}{\beta_k(u)}}, \quad 1 \leq j \leq J,$$

and the invariant law corresponding to such a solution u^* has density $w = (w_k)_{1\leq k\leq K} \mapsto \prod_{k=1}^{K} H_{r_k,\rho_k}(w_k)$ with $\rho_k = a_k(u^*)/\beta_k(u^*)$, see above.

References

1. Ott, T.J., Kemperman, J.H.B., Mathis, M.: The stationary behavior of Ideal TCP Congestion Avoidance. Unpublished manuscript, August (1996)
2. Adjih, C., Jacquet, Ph., Vvedenskaya, N.: Performance evaluation of a single queue under multi-user TCP/IP connections. Tech. Report RR-4141, INRIA, March 2001, http://hal.archives-ouvertes.fr/docs/00/07/24/84/PDF/RR-4141.pdf (2001)
3. Baccelli, F., McDonald, D.R., Reynier, J.: A mean-field model for multiple TCP connections through a buffer implementing RED. Perform. Eval. **49**, 77–97 (2002)
4. Dumas, V., Guillemin, F., Robert, Ph.: A markovian analysis of additive-increase multiplicative-decrease (AIMD) algorithms. Adv. Appl. Probab. **34**(1), 85–111 (2002)
5. Guillemin, F., Robert, Ph., Zwart, B.: AIMD algorithms and exponential functionals. Ann. Appl. Probab. **14**(1), 90–117 (2004)
6. Graham, C., Robert, Ph.: Interacting Multi-class Transmissions in Large Stochastic Networks. Ann. Appl. Probab. **19**(6), 2334–2361 (2009)
7. Sznitman, A.S.: Topics in propagation of chaos, École d'été de Saint-Flour. Lecture Notes in Maths, vol. 1464, pp. 167–243. Springer, Berlin (1989)
8. Karpelevich, F.I., Pechersky, E.A., Suhov, Yu.M.: Dobrushin's approach to queueing network theory. J. Appl. Math. Stochastic Anal. **9**(4), 373–397 (1996) MR1429262 (98d:60182)

9. Graham, C.: Kinetic limits for large communication networks. In: Bellomo, N., Pulvirenti, M. (eds.) Modelling in Applied Sciences: A Kinetic Theory Approach, pp. 317–370. Birkhauser, Boston (2000)
10. Graham, C., Méléard, S.: Chaos hypothesis for a system interacting through shared resources. Probab. Theor. Relat. Field. **100**, 157–173 (1994)
11. Aldous, D.J.: Exchangeability and related topics, École d'été de Probabilités de Saint-Flour XIII. Lecture Notes in Mathematics, vol. 1117, pp. 1–198. Springer, New York (1985)

Minisymposium *Charge and Spin Transport in Nanostructures*

L.L. Bonilla and M. Carretero

G. Millán Institute, Fluid Dynamics, Nanoscience & Industrial Mathematics, Universidad Carlos III, 28911 Leganés, Spain, bonilla@ing.uc3m.es, manuel.carretero@uc3m.es

Electronic transport is the basis of many nanotechnology applications. Spin transport and spintronics are used to create better computer memories, whereas basic science and many interesting device applications are being actively pursued. In nanoelectronic devices, the interplay between charge, spin and vibrational degrees of freedom determines their main electronic and transport features. Moreover, the dimensionality and the number of atoms determines the more suitable theoretical framework and numerical techniques for each particular system. In this minisymposium, different models of quantum charge and spin transport in low-dimensional nanostructures were discussed.

Prof. V. Romano (U. of Catania, Italy) considers the problem of describing electrons in a single band subject to an external electrostatic potential and in equilibrium with a phonon bath. He analyzes the semiclassical limit in which the electron wavelength is small compared to the scale of the potential ($\hbar \to 0$). The method consists of writing an equation for the equilibrium density matrix, transforming this equation via the Wigner transform in an equation for the Wigner function, and expanding nonlocal terms thereof in powers of $\hbar$. The solution of the resulting equation is then approximately solved by regular perturbation methods. Results are given for a band with a nonparabolic Kane dispersion relation.

Dr. L. Barletti (U. of Florence, Italy) and collaborators discuss superlattices (SL) with Rashba spin-orbit effects. These structures are artificial one-dimensional crystals (with finitely many periods). In materials with spin-orbit effects, electrons with different spin have different energies and can transport spin. The paper presents a simple quantum kinetic equation for the SL. Using singular perturbations, Barletti et al. derive spatially nonlocal equations for the electric field and the spin-up and spin-down electron populations, and solve them numerically to show that this SL may behave as a spin oscillator.

In another example of semiconductor-based spintronics, a different spin oscillator can be achieved by applying a static magnetic field to a weakly

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*, Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_14,

coupled SL if at least one period contains magnetic impurities. Dr. M. Carretero (Carlos III University, Spain) and collaborators analyze and solve numerically a spatially discrete model of this system, demonstrating its behavior as an injector of spin polarized time-periodic current.

Prof. G. Platero (CSIC, Spain) discusses the use of double quantum dots as spin-current rectifiers. Quantum dots (QD) are artificial atoms, two QD separated by a barrier (double quantum dots, DQD) are artificial molecules. Attaching contacts to a DQD, electrons with a precise value of spin can tunnel through the barrier from one QD if there is an available state in the other dot and appropriate voltage bias is held between the contacts. Otherwise the Pauli principle precludes tunneling (spin-Coulomb blockade). Thus the DQD acts as a nanoscale spin rectifier, blocking current in one bias direction and allowing it in the other. Platero analyzes a simple transport model for this system and compares it to available experiments.

The Equilibrium Wigner Function in the Case of Nonparabolic Energy Bands

V. Romano

Department of Mathematics and Computer Science, University of Catania,
Viale A. Doria 6, 95125, Catania, Italy, romano@dmi.unict.it

Summary. By solving the Bloch equation the expression of the equilibrium Wigner function is obtained up to first order in the scaled Planck constant for arbitrary energy bands.

1 Introduction

Due to the extreme miniaturization of the electron devices, the simulation requires advanced transport models that take into account also quantum effects.

In [1] a model based on the maximum entropy principle has been proposed by including the quantum corrections with a Chapmann–Enskog expansion starting from the Wigner equation. In the drift-collision dominated regime an explicit form of the Wigner function has been obtained up to first order in the square of the scaled Planck constant in the effective mass approximation. A key point is constituted by the equilibrium Wigner function. It has been determined for the first time in [3] in the effective mass approximation while in [2] a procedure based on the Bloch equation has been devised.

In the present paper we write the Bloch equation for an arbitrary energy band assuming that it is defined in all the space, as appropriate for some analytical approximations like Kane's dispersion relation. The free streaming pseudo-differential operator is defined as a multiplication operator in the space of Fourier transforms. The general form of the solution up to second order terms in the scaled Plank is determined. In the case of the Kane dispersion relation an explicit formula is given and it shown that, at variance with the parabolic band, a quantum correction is present even in the bulk case.

2 The Bloch Equation

The physical situation is represented by an electron gas which is in equilibrium with a thermal bath of phonons at a constant temperature T_L. We suppose

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*,
Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_15,

that the energy bands are represented by a function $\mathcal{E}(p)$ defined in $\Re^3$, which depends only on the modulus of the crystal momentum p and it is even. Several analytical approximation as the parabolic band and the Kane dispersion relation satisfy the previous conditions. Moreover we will work in the single electron approximation.

Under the previous assumptions the system is described by the density matrix $\rho(r,s)$ with r,s position vectors belonging to $\Re^3$. If we denote by H the Hamiltonian, the equilibrium is parametrized by the inverse of the temperature $\beta = \frac{1}{k_B T_L}$ and is defined, in the Boltzmann limit of the Fermi-Dirac statistics, by $\rho^{(eq)}(r,s,\beta) = \exp(-\beta H)$, where of course the exponential must be intended in the operatorial sense. Expanding the exponential gives an approximation of $\rho^{(eq)}(r,s,\beta)$ but the procedure is rather cumbersome.

An alternative way has been devised in [2]: starting from the Schrödinger equation one derives the following equation for the equilibrium density matrix $\rho^{(eq)}(r,s,\beta)$

$$\frac{\partial}{\partial\beta}\rho^{(eq)}(r,s,\beta) = -\frac{1}{2}\left(H\rho^{(eq)} + \rho^{(eq)}H\right), \tag{1}$$

called the Bloch equation, augmented by the condition $\rho^{(eq)}(r,s,0) = \delta(r-s)$

The Hamiltonian H is given for a general energy band $\mathcal{E}$ by

$$H(x,p) = \mathcal{E}(p) - qV(x) - \Phi \tag{2}$$

with q absolute electron charge and V the electrostatic potential. Φ is the quasi Fermi potential which is constant at equilibrium. In an operatorial sense $\mathcal{E}(p)$ acts as $\mathcal{E}(-i\hbar\nabla_x)$, e.g. in the parabolic case $\mathcal{E}(p) = p^2/(2m^*)$ the corresponding operator is $-(\hbar^2/2m^*)\Delta_x$, m^* being the effective electron mass. In the sequel in order to simplify the notation the same symbol will be used both for the operator and its symbol.

After the change of variables

$$\begin{cases} r = x + \frac{\hbar}{2}\eta, \\ s = x - \frac{\hbar}{2}\eta \end{cases}$$

we introduce the Wigner function

$$w(x,p,t) = \mathcal{F}^{-1}[\rho(r,s,t)](x,p,t) = \frac{1}{(2\pi)^3}\int_{\Re^3_\eta} \rho\left(x + \frac{\hbar}{2}\eta, x - \frac{\hbar}{2}\eta\right) e^{ip\cdot\eta}\, d\,\eta,$$

with $\mathcal{F}$ the Fourier transform and $\mathcal{F}^{-1}$ its inverse. In particular $w^{(eq)}(x,p,\beta)$ is the equilibrium Wigner function given by $\mathcal{F}^{-1}[\rho^{(eq)}](x,p,\beta)$.

By substituting the expression of H into (1) and applying $\mathcal{F}^{-1}$ one has

$$\begin{aligned} \frac{\partial}{\partial\beta}w^{(eq)}(x,p,\beta) = &-\frac{1}{2}\mathcal{F}^{-1}\left[\mathcal{E}(-i\hbar\nabla_r)\rho^{(eq)} + \mathcal{E}(-i\hbar\nabla_s)\rho^{(eq)}\right] \\ &+\frac{q}{2}\mathcal{F}^{-1}\left[\left(V(r) + V(s)\right)\rho^{(eq)}\right] + \Phi w^{(eq)}. \end{aligned} \tag{3}$$

By introducing the convolution operator $f * g = \int f(x-t)g(t)dt$, we have

$$\begin{aligned}
&\mathcal{F}^{-1}\left[\left(V(r)+V(s)\right)\rho^{(eq)}\right](x,p,\beta)\\
&\quad=\mathcal{F}^{-1}\left[\left(V(r)+V(s)\right)\right]*w^{(eq)}(x,p,\beta)\\
&\quad=\int_{\Re^3_q}\mathcal{F}^{-1}\left[\left(V(r)+V(s)\right)\right](x,p-q,\beta)w^{(eq)}(x,q,\beta)\,dq\\
&\quad=\frac{1}{(2\pi)^3}\int_{\Re^3_q\times\Re^3_\eta}[V(x+\frac{\hbar}{2}\eta)+V(x-\frac{\hbar}{2}\eta)]w^{(eq)}(x,q,\beta)e^{i(p-q)\cdot\eta}dqd\eta.
\end{aligned}\tag{4}$$

Similarly, since $i\hbar\nabla_r = i\frac{\hbar}{2}\nabla_x + i\nabla_\eta$ and $i\hbar\nabla_s = i\frac{\hbar}{2}\nabla_x - i\nabla_\eta$, by taking into account that p and η are conjugate variables, we define

$$\begin{aligned}
&\mathcal{F}^{-1}\left[\mathcal{E}(-i\hbar\nabla_r)\rho^{(eq)}+\mathcal{E}(-i\hbar\nabla_s)\rho^{(eq)}\right](x,p,\beta)\\
&\quad=(2\pi)^{-3}\int_{\Re^3_\nu\times\Re^3_{x'}}\left[\mathcal{E}\left(p+\frac{\hbar\nu}{2}\right)+\mathcal{E}\left(p-\frac{\hbar\nu}{2}\right)\right]\\
&\qquad w^{(eq)}(x',p,\beta)e^{i(x-x')\cdot\nu}\,d\nu dx'.
\end{aligned}\tag{5}$$

By expanding up to first order in $\hbar^2$, one has

$$\begin{aligned}
\mathcal{E}\left(p+\frac{\hbar\nu}{2}\right)+\mathcal{E}\left(p-\frac{\hbar\nu}{2}\right)&=2\mathcal{E}(p)+\frac{1}{4}\frac{\partial^2\mathcal{E}}{\partial p_i\partial p_j}\nu_i\nu_j\hbar^2+o(\hbar^2),\\
V(x+\frac{\hbar}{2}\eta)+V(x-\frac{\hbar}{2}\eta)&=2V(x)+\frac{1}{4}\frac{\partial^2 V}{\partial x_i\partial x_j}\eta_i\eta_j\hbar^2+o(\hbar^2),
\end{aligned}$$

where summation over repeated indexes is understood, and the Bloch equation up to first order in $\hbar^2$ reads

$$\begin{aligned}
&\frac{\partial}{\partial\beta}w^{(eq)}(x,p,\beta)=-\mathcal{E}(p)w^{(eq)}(x,p,\beta)+\frac{\hbar^2}{8}\frac{\partial^2\mathcal{E}}{\partial p_i\partial p_j}\frac{\partial^2 w^{(eq)}(x,p,\beta)}{\partial x_i\partial x_j}\\
&+qV(x)w^{(eq)}(x,p,\beta)-\frac{q\hbar^2}{8}\frac{\partial^2 V}{\partial x_i\partial x_j}\frac{\partial^2 w^{(eq)}(x,p,\beta)}{\partial p_i\partial p_j}+\Phi w^{(eq)}(x,p,\beta),
\end{aligned}\tag{6}$$

with initial condition $w^{(eq)}(x,p,0)=1$.

3 The Equilibrium Wigner Function

We look for a solution of (6) of the form

$$w^{(eq)}(x,p,\beta)=w^{(0)}(x,p,\beta)+\hbar^2 w^{(1)}(x,p,\beta)+o(\hbar^2).$$

At zero order (6) gives

$$\frac{\partial}{\partial\beta}w^{(0)}(x,p,\beta)=-\mathcal{E}(p)w^{(0)}(x,p,\beta)+qV(x)w^{(0)}(x,p,\beta)+\Phi w^{(0)}(x,p,\beta).$$

where from $w^{(0)}(x,p,\beta)=\exp[-\mathcal{E}(p)\beta+\beta(\Phi+qV(x))]$.

At first order in $\hbar^2$ (6) gives

$$\frac{\partial}{\partial\beta}w^{(1)}(x,p,\beta) = -\mathcal{E}(p)w^{(1)}(x,p,\beta) + \frac{1}{8}\frac{\partial^2\mathcal{E}}{\partial p_i\partial p_j}\frac{\partial^2 w^{(0)}(x,p,\beta)}{\partial x_i\partial x_j} + qV(x)w^{(1)}(x,p,\beta) - \frac{q}{8}\frac{\partial^2 V}{\partial x_i\partial x_j}\frac{\partial^2 w^{(0)}(x,p,\beta)}{\partial p_i\partial p_j} + \Phi w^{(1)}(x,p,\beta). \tag{7}$$

We solve the last equation via separation of constants looking for solution of the form

$$w^{(1)}(x,p,\beta) = g(x,p,\beta)w^{(0)}(x,p,\beta)$$

with the function g satisfying the equation

$$\frac{\partial g}{\partial\beta} = \frac{1}{8w^{(0)}}\frac{\partial^2\mathcal{E}}{\partial p_i\partial p_j}\frac{\partial^2 w^{(0)}(x,p,\beta)}{\partial x_i\partial x_j} - \frac{q}{8w^{(0)}}\frac{\partial^2 V}{\partial x_i\partial x_j}\frac{\partial^2 w^{(0)}(x,p,\beta)}{\partial p_i\partial p_j} \tag{8}$$

and the initial condition

$$g(x,p,0) = 0.$$

One finds

$$g(x,p,\beta) = \frac{q\beta^2}{8}\frac{\partial^2\mathcal{E}}{\partial p_i\partial p_j}\frac{\partial^2 V}{\partial x_i\partial x_j} + \frac{q^2\beta^3}{24}\left[\frac{\partial^2\mathcal{E}}{\partial p_i\partial p_j}\frac{\partial V}{\partial x_i}\frac{\partial V}{\partial x_j} - \frac{\partial^2 V}{\partial x_i\partial x_j}v_iv_j\right], \tag{9}$$

where $v = \nabla_p\mathcal{E}(p)$ is the electron velocity. The equilibrium Wigner equation is therefore given by

$$w^{(eq)}(x,p,\beta) = \exp\left[-\mathcal{E}(p)\beta + \beta\left(\Phi + qV(x)\right)\right]\left\{1 + \frac{q\beta^2\hbar^2}{8}\frac{\partial^2\mathcal{E}}{\partial p_i\partial p_j}\frac{\partial^2 V}{\partial x_i\partial x_j} + \frac{q^2\beta^3\hbar^2}{24}\left[\frac{\partial^2\mathcal{E}}{\partial p_i\partial p_j}\frac{\partial V}{\partial x_i}\frac{\partial V}{\partial x_j} - \frac{\partial^2 V}{\partial x_i\partial x_j}v_iv_j\right]\right\} + o(\hbar^2). \tag{10}$$

In the particular case of a parabolic band

$$\mathcal{E}(p) = \frac{p^2}{2m^*}, \quad v = \frac{p}{m^*}$$

with m^* electron effective mass, and one obtains the same results as in [3]

$$w^{(eq)}(x,p,\beta) = \exp\left[-\frac{\beta p^2}{2m^*} + \beta\left(\Phi + qV(x)\right)\right]\left\{1 + \frac{q\beta^2\hbar^2}{8m^*}\Delta V + \frac{q^2\beta^3\hbar^2}{24m^*}\left[|\nabla V|^2 - m^*\frac{\partial^2 V}{\partial x_i\partial x_j}v_iv_j\right]\right\} + o(\hbar^2).$$

It is convenient (see for example [1]) to parametrize $w^{(eq)}(x,p,\beta)$ in term of the local density instead of the quasi Fermi potential Φ.

By defining the density as

$$n(x,t) = \int_{\Re_p^3} w^{(eq)}(x,p,\beta)dp$$

and eliminating $\exp[\beta(qV+\phi)]$, one has

$$w^{(eq)}(x,p,\beta) = \frac{n(x,t)e^{-\beta\mathcal{E}}\exp}{A_0(\beta,m^*)}\left\{1+\hbar^2\left[\left(\frac{q\beta^2}{8}\frac{\partial^2 V}{\partial x_i\partial x_j}+\frac{q^2\beta^3\hbar^2}{24}\frac{\partial V}{\partial x_i}\frac{\partial V}{\partial x_j}\right)\right.\right.$$
$$\left.\left.\left(\frac{\partial^2\mathcal{E}}{\partial p_i\partial p_j}-\frac{A_{ij}(\beta,m^*)}{A_0(\beta,m^*)}\right)-\frac{q\beta^3\hbar^2}{24}\frac{\partial^2 V}{\partial x_i\partial x_j}\left(v_iv_j-\frac{B_{ij}(\beta,m^*)}{A_0(\beta,m^*)}\right)\right]\right\}+o(\hbar^2) \tag{11}$$

where

$$A_0(\beta,m^*)=\int_{\Re^3}e^{-\beta\mathcal{E}}\,dp,\quad A_{ij}(\beta,m^*)=\int_{\Re^3}e^{-\beta\mathcal{E}}\frac{\partial^2\mathcal{E}}{\partial p_i\partial p_j}\,dp,$$
$$B_{ij}(\beta,m^*)=\int_{\Re^3}e^{-\beta\mathcal{E}}v_iv_j\,dp.$$

4 The Case of the Kane Dispersion Relation

In the case of the Kane dispersion relation

$$\frac{p^2}{2m^*}=\mathcal{E}\,(1+\alpha\mathcal{E})$$

with α nonparabolicity factor while

$$v=\frac{p}{m^*(1+2\alpha\mathcal{E})}$$

and

$$\frac{\partial^2\mathcal{E}}{\partial p_i\partial p_j}=\frac{1}{m^*(1+2\alpha\mathcal{E})}\left[\delta_{ij}-\frac{2\alpha}{m^*(1+2\alpha\mathcal{E})^2}p_ip_j\right].$$

By expressing the elementary volume $d\,p$ as

$$dp=m^*\sqrt{2m^*\mathcal{E}(1+\alpha\mathcal{E})}(1+2\alpha\epsilon)\,d\mathcal{E}d\Omega,$$

$d\,\Omega$ being the elementary solid angle, the coefficients appearing in the Wigner function can be written as

$$A_0(\beta,m^*)=4\pi m^*\sqrt{2m^*}\int_0^\infty e^{-\beta\mathcal{E}}\sqrt{\mathcal{E}(1+\alpha\mathcal{E})}\mathcal{E}(1+2\alpha\mathcal{E})\,d\mathcal{E}$$
$$=4\pi m^*\sqrt{2m^*}\;d_0(\beta),$$
$$A_{ij}(\beta,m^*)=4\pi\sqrt{2m^*}\;\delta_{ij}\int_0^\infty e^{-\beta\mathcal{E}}\left[\sqrt{\mathcal{E}(1+\alpha\mathcal{E})}-\frac{4\alpha\left[\mathcal{E}(1+\alpha\mathcal{E})\right]^{3/2}}{3(1+2\alpha\mathcal{E})^2}\right]d\mathcal{E},$$
$$B_{ij}(\beta,m^*)=\frac{8\pi}{3}\sqrt{2m^*}\;\delta_{ij}\int_0^\infty e^{-\beta\mathcal{E}}\frac{\left[\mathcal{E}(1+\alpha\mathcal{E})\right]^{3/2}}{(1+2\alpha\mathcal{E})}\,d\mathcal{E},$$

obtaining the equilibrium Wigner function

$$w^{(eq)}(x,p,\beta) = \frac{n(x,t)e^{-\beta\mathcal{E}}}{4\pi m^*\sqrt{m^*}d_0(\beta)}\left\{1+\hbar^2\left[\left(\frac{q\beta^2}{8}\frac{\partial^2 V}{\partial x_i\partial x_j}+\frac{q^2\beta^3}{24}\frac{\partial V}{\partial x_i}\frac{\partial V}{\partial x_j}\right)\right.\right.$$
$$\left[\frac{\delta_{ij}}{m^*(1+2\alpha\mathcal{E})}-\frac{2\alpha p_i p_j}{(m^*)^2(1+2\alpha\mathcal{E})^3}-\frac{\delta_{ij}}{m^*d_0(\beta)}\right.$$
$$\left.\int_0^{+\infty} e^{-\beta\mathcal{E}}\left(\sqrt{\mathcal{E}(1+\alpha\mathcal{E})}-\frac{4\alpha}{3}\frac{[\mathcal{E}(1+\alpha\mathcal{E})]^{3/2}}{(1+2\alpha\mathcal{E})^2}\right)d\mathcal{E}\right]$$
$$-\frac{q\beta^3}{24}\frac{\partial^2 V}{\partial x_i\partial x_j}\left(v_i v_j-\frac{2\delta_{ij}}{3m^*d_0(\beta)}\right.$$
$$\left.\left.\left.\times\int_0^{\infty} e^{-\beta\mathcal{E}}\frac{[\mathcal{E}(1+\alpha\mathcal{E})]^{3/2}}{1+2\alpha\mathcal{E}}\,d\mathcal{E}\right)\right]\right\}.$$

Remark 1. In the bulk case the $\hbar^2$ correction vanishes in the parabolic band approximation and $w^{(eq)}(x,p,\beta)$ reduces to the semiclassical Maxwellian. Instead when the energy bands are described by the Kane dispersion relation, $w^{(eq)}(x,p,\beta)$ in the bulk case is given by

$$w^{(eq)}(x,p,\beta) = \frac{n(x,t)e^{-\beta\mathcal{E}}}{4\pi m^*\sqrt{m^*}d_0(\beta)}\left\{1+\hbar^2\frac{q^2\beta^3}{24}E_iE_j\right.$$
$$\left[\frac{\delta_{ij}}{m^*(1+2\alpha\mathcal{E})}-\frac{2\alpha p_i p_j}{(m^*)^2(1+2\alpha\mathcal{E})^3}-\frac{\delta_{ij}}{m^*d_0(\beta)}\right.$$
$$\left.\left.\int_0^{+\infty} e^{-\beta\mathcal{E}}\left(\sqrt{\mathcal{E}(1+\alpha\mathcal{E})}-\frac{4\alpha}{3}\frac{[\mathcal{E}(1+\alpha\mathcal{E})]^{3/2}}{(1+2\alpha\mathcal{E})^2}\right)d\mathcal{E}\right]\right\},$$

with $E_i = -\partial V/\partial x_i$ the components of the electric field. This implies that the quantum correction affects all the transport parameters even in the bulk case when more realistic approximations of the energy bands are used.

Acknowledgments

The author acknowledges the financial support by M.I.U.R. (PRIN 2007 *Equazioni cinetiche e idrodinamiche di sistemi collisionali complessi*) and the EU Marie Curie RTN project **COMSON** grant n. MRTN-CT-2005-019417.

References

1. Romano, V.: J. Math. Phys. **48**, 123504-1–123504-24 (2007)
2. Gardner, C., Ringhofer, C.: SIAM J. Appl. Math. **54**, 409–427 (1994)
3. Wigner, E.: Phys. Rev. **40**, 749–759 (1932)

Nonlinear Electron and Spin Transport in Semiconductor Superlattices

L.L. Bonilla[1], L. Barletti[2] and M. Alvaro[1]

[1] G. Millán Institute of Fluid Dynamics, Nanoscience & Industrial Mathematics, Universidad Carlos III, Leganés, Spain, bonilla@ing.uc3m.es, barletti@math.unifi.it

[2] Dipartimento di Matematica "Ulisse Dini", Università di Firenze, Italy, mariano.alvaro@uc3m.es

Summary. Nonlinear charge transport in strongly coupled semiconductor superlattices is described by two-miniband Wigner–Poisson kinetic equations with BGK collision terms. The hyperbolic limit, in which the collision frequencies are of the same order as the Bloch frequencies due to the electric field, is investigated by means of the Chapman–Enskog perturbation technique, leading to nonlinear drift-diffusion equations for the two miniband populations. In the case of a lateral superlattice with spin-orbit interaction, the corresponding drift-diffusion equations are used to calculate spin-polarized currents and electron spin polarization.

1 Introduction

Semiconductor superlattices are essential ingredients in fast nanoscale oscillators, quantum cascade lasers and infrared detectors. A superlattice (SL) is a quasi-one-dimensional crystal originally proposed by Esaki and Tsu to observe Bloch oscillations, i.e., the periodic coherent motion of electrons in a miniband when an electric field is applied. Once the materials were grown, many interesting nonlinear phenomena were observed, such as self-oscillations of the current through the SL due to charge dipole motion, multistability of stationary charge and field profiles, etc. See the review [1].

Nonlinear charge transport in SLs has been widely studied in the last decade using balance equations for electron densities and electric field. These equations are either proposed using phenomenological arguments or derived ad hoc from kinetic theories [1]. Systematic derivations are scarce. For a single-miniband SL, the Chapman–Enskog (CE) method applied to a semiclassical Boltzmann–Poisson system whose collision term is of Bhatnagar–Gross–Krook (BGK) type yields a generalized drift-diffusion equation (GDDE) [2], and a quantum drift-diffusion equation (QDDE) when applied to a Wigner–Poisson–BGK (WPBGK) system [3]. The quantum WPBGK system contains two pseudo-differential operators, involving the band dispersion relation and the electric potential. The leading order approximation in the hyperbolic limit

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*, Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_16,

balances collisions and electric potential, and its solution is not obvious because the potential is an a priori unknown solution of the Poisson equation. SLs are simpler because their Wigner functions are periodic in the reciprocal lattice, the potential terms become multiplication operators in Fourier space, and the leading order approximation is straightforward to solve [3].

For sufficiently high applied electric fields, electrons may populate higher minibands, then be scattered to the lowest, etc. Moreover, SLs with diluted magnetic impurities subject to a magnetic field may present spin polarization effects whose understanding is crucial to develop spintronic devices [4]. Even without magnetic impurities, spin polarization could appear due to Rashba spin-orbit interaction [5]. Once we consider electron spin, each miniband is split in two and single-miniband SLs become two-miniband SLs. We shall systematically derive quantum balance equations by the CE method.

2 Wigner Description of a Two-Miniband Superlattice

We shall consider a 2×2 Hamiltonian $\mathbf{H}(x, -i\partial/\partial x)$, in which

$$\begin{aligned}\mathbf{H}(x,k) &= [h_0(k) - eW(x)]\boldsymbol{\sigma}_0 + \mathbf{h}(k)\cdot\boldsymbol{\sigma}] \qquad (1)\\ &\equiv \begin{pmatrix} (\alpha+\gamma)(1-\cos kl) - eW(x) + g & -i\beta\sin kl \\ i\beta\sin kl & (\alpha-\gamma)(1-\cos kl) - eW(x) - g \end{pmatrix}.\end{aligned}$$

Then $h_0 = \alpha(1-\cos kl)$, $h_1 = 0$, $h_2 = \beta\sin kl$, $h_3 = \gamma(1-\cos kl) + g$, and

$$\boldsymbol{\sigma}_0 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \boldsymbol{\sigma}_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \boldsymbol{\sigma}_2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \boldsymbol{\sigma}_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}. \qquad (2)$$

The Hamiltonian (1) corresponds to the simplest 2×2 Kane model in which the quadratic and linear terms $(kl)^2/2$ and kl are replaced by $(1-\cos kl)$ and $\sin kl$, respectively. For a SL with two minibands, $2g$ is the miniband gap and $\alpha = (\Delta_1 + \Delta_2)/4$ and $\gamma = (\Delta_1 - \Delta_2)/4$, provided Δ_1 and Δ_2 are the miniband widths. In the case of a lateral SL, $g = \gamma = 0$, and $h_2\boldsymbol{\sigma}_2$ corresponds to the precession term in the Rashba spin-orbit interaction [5]. The other term, the intersubband coupling, depends on the momentum in the y direction and we have not included it here. Small modifications of (1) represent a single miniband SL with dilute magnetic impurities in the presence of a magnetic field B: $g = \gamma = h_2 = 0$, and $h_1 = \beta(B)$ [4]. As in the case of a single miniband SL, $W(x)$ is the electric potential.

The energy minibands $\mathcal{E}^{\pm}(k)$ are the eigenvalues of the free Hamiltonian $\mathbf{H}_0(k) = h_0(k)\boldsymbol{\sigma}_0 + \mathbf{h}(k)\cdot\boldsymbol{\sigma}$ and are given by

$$\mathcal{E}^{\pm}(k) = h_0(k) \pm |\mathbf{h}(k)|. \qquad (3)$$

The corresponding spectral projections are $\mathbf{P}^{\pm}(k) = (\boldsymbol{\sigma}_0 \pm \boldsymbol{\nu}(k)\cdot\boldsymbol{\sigma})/2$, with $\boldsymbol{\nu} = \mathbf{h}/|\mathbf{h}(k)|$, so that we can write $\mathbf{H}_0(k) = \mathcal{E}^{+}(k)\mathbf{P}^{+}(k) + \mathcal{E}^{-}(k)\mathbf{P}^{-}(k)$.

We shall now write the WPBGK equations for the Wigner matrix written in terms of the Pauli matrices $\boldsymbol{\sigma}_s$:

$$\mathbf{f}(x,k,t) = \sum_{s=0}^{3} f^s(x,k,t)\boldsymbol{\sigma}_s = f^0(x,k,t)\boldsymbol{\sigma}_0 + \mathbf{f}(x,k,t)\cdot\boldsymbol{\sigma}. \tag{4}$$

The Wigner components are real and can be related to the coefficients of the Hermitian Wigner matrix by $f_{11} = f^0 + f^3$, $f_{12} = f^1 - if^2$, $f_{21} = f^1 + if^2$, $f_{22} = f^0 - f^3$. The populations of the minibands with energies $\mathcal{E}^{\pm}$ are the moments:

$$n^{\pm}(x,t) = \frac{l}{2\pi}\int_{-\pi/l}^{\pi/l} \left[f^0(x,k,t) \pm \boldsymbol{\nu}\cdot\mathbf{f}(x,k,t)\right]\, dk, \tag{5}$$

and the total electron density is $n^+ + n^-$.

We shall restrict ourselves to the Rashba case, $g = \gamma = h_3 = 0$, from now on. Then $\boldsymbol{\nu} = (0,1,0)$ and $n^{\pm}$ are the densities of electrons having spin $\pm$. After some algebra, we can obtain the following WPBGK equations for the Wigner components

$$\frac{\partial f^0}{\partial t} + \frac{\alpha}{\hbar}\sin kl\, \Delta^- f^0 + \frac{\beta\cos kl}{\hbar}\,\Delta^- f^2 - \Theta f^0 = Q^0[f], \tag{6}$$

$$\frac{\partial \mathbf{f}}{\partial t} + \frac{\alpha\sin kl}{\hbar}\Delta^-\mathbf{f} + \frac{\beta}{\hbar}[\boldsymbol{\nu}\Delta^- f^0\cos kl + \Delta^+(\boldsymbol{\nu}\times\mathbf{f})\sin kl] - \Theta\mathbf{f} = \mathbf{Q}[f], \tag{7}$$

$$\varepsilon\,\frac{\partial^2 W}{\partial x^2} = \frac{e}{l}\,(n^+ + n^- - N_D), \tag{8}$$

$$\Theta f^s(x,k,t) = \sum_{j=-\infty}^{\infty} \frac{ejl}{i\hbar}\langle F(x,t)\rangle_j e^{ijkl} f_j^s(x,t), \tag{9}$$

where we have put $f^s(x,k,t) = \sum_{j=-\infty}^{\infty} f_j^s(x,t)\, e^{ijkl}$ and

$$\langle u\rangle_j(x,t) = \frac{1}{jl}\int_{-jl/2}^{jl/2} u(x+s,t)\, ds.$$

Our collision model contains two terms: a BGK term which tries to send $f^0 \pm \boldsymbol{\nu}\cdot\mathbf{f}$ to its (collision-broadened) Fermi–Dirac local equilibrium, and a scattering term which tries to equalize n^+ and n^- [4]:

$$Q^0[f] = -\frac{f^0 - \Omega^0}{\tau}, \quad \mathbf{Q}[f] = -\frac{\mathbf{f} - \boldsymbol{\Omega}}{\tau} - \frac{\mathbf{f}}{\tau_{\mathrm{sc}}}, \tag{10}$$

$$\Omega^0 = \frac{\phi^+ + \phi^-}{2}, \quad \boldsymbol{\Omega} = \frac{\phi^+ - \phi^-}{2}\,\boldsymbol{\nu}, \tag{11}$$

where (see [6] for details and [7] for the numerical method employed)

$$\phi^{\pm}(k;\mu^{\pm}) = \int_{-\infty}^{+\infty} \frac{D_\Gamma\left(E-\mathcal{E}^{\pm}(k)\right)}{1+\exp\left(\frac{E-\mu^{\pm}}{k_BT}\right)}\,dE \tag{12}$$

$$D_\Gamma(E) = \frac{\sqrt{2m^*}}{2\pi\hbar L_z}\int_0^{\infty}\frac{\delta_\Gamma(E_y+E_1-E)}{\sqrt{E_y}}\,dE_y, \quad \delta_\Gamma(E) = \frac{\sqrt{2}\,\Gamma^3/\pi}{\Gamma^4+E^4} \tag{13}$$

$$\frac{l}{2\pi}\int_{-\pi/l}^{\pi/l}\phi^{\pm}(k;n^{\pm})\,dk = n^{\pm}. \tag{14}$$

In (12), $\mu^{\pm} = \mu^{\pm}(n^{\pm})$ solve (14). Our collision model satisfies charge continuity. In fact, from (6) to (8) we obtain:

$$\frac{\partial}{\partial t}(n^+ + n^-) + \Delta^-\left[\frac{l}{\pi\hbar}\int_{-\pi/l}^{\pi/l}(\alpha\,\sin kl\,f^0 + \beta\,\cos kl\,f^2)\,dk\right] = 0. \tag{15}$$

Since $\Delta^- u(x) = l\,\partial\langle u(x)\rangle_1/\partial x$, (15) provides charge continuity. From (8) and (15), we get Ampère's law ($J(t)$ is the total current density):

$$\varepsilon\frac{\partial F}{\partial t} + \left\langle \frac{el}{\pi\hbar}\int_{-\pi/l}^{\pi/l}(\alpha\,\sin kl\,f^0 + \beta\,\cos kl\,f^2)\,dk\right\rangle_1 = J(t). \tag{16}$$

3 Quantum Drift-Diffusion Equations with Spin-Orbit Interaction

In the simpler case of a lateral SL with the precession term of Rashba spin-orbit interaction (but no intersubband coupling), we can obtain explicit rate equations for $n^{\pm}$ by means of the CE method. First of all, we should decide the order of magnitude of the terms in the WPBGK equations (6) and (7) in the hyperbolic limit. In this limit, the collision frequency $1/\tau$ and the Bloch frequency $eF_Ml/\hbar$ are of the same order, and the scattering time $\tau_{\rm sc}$ is much longer than the collision time τ. Then, a suitable small parameter λ can be introduced [2] such that the scaled Wigner equations read as follows:

$$\lambda\frac{\partial f^0}{\partial t} + \lambda\frac{\alpha}{\hbar}\sin kl\,\Delta^- f^0 + \lambda\frac{\beta\cos kl}{\hbar}\,\Delta^- f^2 - \Theta f^0 = Q^0[f], \tag{17}$$

$$\lambda\frac{\partial \mathbf{f}}{\partial t} + \lambda\frac{\alpha\sin kl}{\hbar}\Delta^-\mathbf{f} + \lambda\frac{\beta}{\hbar}[\boldsymbol{\nu}\Delta^- f^0\cos kl + \Delta^+(\boldsymbol{\nu}\times\mathbf{f})\sin kl] \tag{18}$$

$$-\Theta\mathbf{f} = -\frac{\mathbf{f}-\boldsymbol{\Omega}}{\tau} - \lambda\frac{\mathbf{f}}{\tau_{\rm sc}}, \tag{19}$$

To derive the reduced balance equations, we use the following CE ansatz:

$$f(x,k,t;\lambda) = f^{(0)}(k;n^+,n^-,F) + \sum_{m=1}^{\infty} f^{(m)}(k;n^+,n^-,F)\,\lambda^m, \tag{20}$$

$$\varepsilon\frac{\partial F}{\partial t} + \sum_{m=0}^{\infty} J_m(n^+,n^-,F)\,\lambda^m = J(t), \quad \frac{\partial n^\pm}{\partial t} = \sum_{m=0}^{\infty} A_m^\pm(n^+,n^-,F)\,\lambda^m. \tag{21}$$

$A_m^\pm$ and J_m are related through the Poisson equation (8), so that

$$A_m^+ + A_m^- = -\frac{l}{e}\frac{\partial J_m}{\partial x}. \tag{22}$$

Following the CE procedure up to order 2 (see [6] for details) we obtain

$$\frac{\partial n^\pm}{\partial t} + \Delta^- D_\pm(n^+,n^-,F) = \mp R(n^+,n^-,F), \tag{23}$$

$$\varepsilon\,\frac{\partial F}{\partial t} + e\,\langle D_+ + D_-\rangle_1 = J, \tag{24}$$

$$D_\pm = -\frac{\alpha}{\hbar}\,\mathrm{Im}(\varphi_1^0 \pm \varphi_1^2 + \psi_1^0 \pm \psi_1^2) \pm \frac{\beta}{\hbar}\,\mathrm{Re}(\varphi_1^0 \pm \varphi_1^2 + \psi_1^0 \pm \psi_1^2), \tag{25}$$

$$R = \frac{n^+ - n^-\theta(\mu^-\mathcal{E}_{\min}^+)}{\tau_{sc}}. \tag{26}$$

Here, $\varphi \equiv f^{(0)}$ and $\psi \equiv f^{(1)}$ can be explicitly calculated and yield

$$\begin{aligned} D_\pm = {} & \frac{(\alpha\vartheta_1 \pm \beta)\phi_1^\pm}{\hbar\,(1+\vartheta_1^2)} \mp \frac{\tau\,(\phi_1^+ - \phi_1^-)\,[2\alpha\vartheta_1 \pm \beta(1-\vartheta_1^2)]}{2\hbar\tau_{\rm sc}(1+\vartheta_1^2)^2} \\ & + \frac{[2\alpha\vartheta_1 \pm \beta(1-\vartheta_1^2)]\,\alpha\tau}{\hbar^2\,(1+\vartheta_1^2)^2}\,\frac{\partial\phi_1^\pm}{\partial n^\pm}\left[\Delta^-\left(\frac{\alpha\vartheta_1 \pm \beta}{\hbar\,(1+\vartheta_1^2)}\phi_1^\pm\right) \pm \frac{\hbar}{\alpha\tau_{sc}}(n^+ - n^-)\right] \\ & + \frac{\alpha\,(3\vartheta_1^2 - 1) \pm \beta\vartheta_1(3-\vartheta_1^2)}{\hbar(1+\vartheta_1^2)^3}\,\frac{l\tau^2\phi_1^\pm}{\hbar\varepsilon}\left(\frac{J}{e} - \left\langle\left\langle\frac{\alpha\,(\phi_1^+ + \phi_1^-)\vartheta_1}{\hbar(1+\vartheta_1^2)}\right\rangle_1\right\rangle_1\right. \\ & \left. - \left\langle\left\langle\frac{\beta\,(\phi_1^+ - \phi_1^-)}{\hbar(1+\vartheta_1^2)}\right\rangle_1\right\rangle_1\right) - \frac{(\alpha^2+\beta^2)\tau}{2\hbar^2(1+\vartheta_1^2)}\,\Delta^- n^\pm \\ & + \frac{\tau}{2\hbar^2(1+\vartheta_1^2)}\left[(\alpha^2 - \beta^2 \mp 2\alpha\beta\vartheta_1)\,\Delta^-\left(\frac{\phi_2^\pm}{1+\vartheta_2^2}\right)\right. \\ & \left. + [(\beta^2 - \alpha^2)\vartheta_1 \mp 2\alpha\beta]\,\Delta^-\left(\frac{\vartheta_2\phi_2^\pm}{1+\vartheta_2^2}\right)\right], \end{aligned} \tag{27}$$

where $\phi_j^\pm$ are the Fourier components of the local Fermi–Dirac equilibrium functions (12) and $\vartheta_j \equiv \frac{\tau e j l}{\hbar}\,\langle F\rangle_j$.

The simulations shown below are based on the quantum drift-diffusion equations (23)–(27). Figure 1a shows electric current vs. field in a spatially uniform stationary state, for a Fermi–Dirac statistics, for different values of the level broadening parameter Γ, and for the Botzmann statistics without

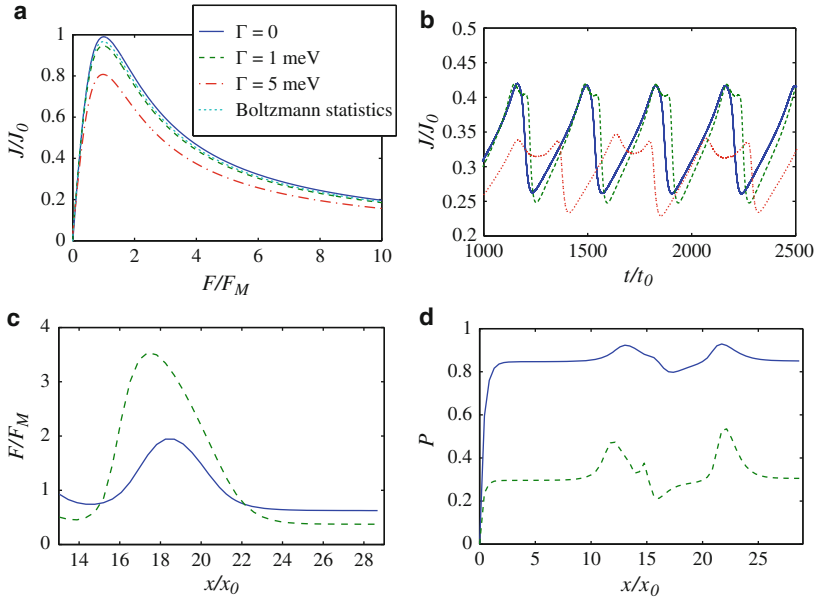


Fig. 1. Electric current vs. field in the stationary case (**a**). Total current vs. time (**b**), electric field profile (**c**) and polarization profile (**d**) during current self-oscillations

broadening. Figures 1b–1d illustrate (for different values of Γ and also for the Botzmann case), a non-stationary behavior showing stable, self-sustained current and spin oscillations. They are due to the periodic formation of a pulse of the electric field at the cathode $x = 0$ and its motion through the superlattice. Plot (b) is total current density vs. time while (c) and (d) show the electric field (c) and spin polarization (d) profiles during current self-oscillations. We have used the following values of the parameters: $\alpha = 8\,\text{meV}$, $\beta = 2.63\,\text{meV}$, $L_z = 3.1\,\text{nm}$, $T = 5\,\text{K}$, $\tau = 5.56 \times 10^{-14}\,\text{s}$, $\tau_{\text{sc}} = 5.56 \times 10^{-13}\,\text{s}$, $N_D = 4.048 \times 10^{10}\,\text{cm}^{-2}$, $m^* = 0.0992$, $V = 3\,\text{V}$. The plot units are the following: $F_M = 23.42\,\text{kV/cm}$, $x_0 = 19.4\,\text{nm}$, $t_0 = 0.082\,\text{ps}$, $J_0 = 3.94 \times 10^4\text{A/cm}^2$.

References

1. Bonilla, L.L., Grahn, H.T.: Rep. Prog. Phys. **68**, 577 (2005)
2. Bonilla, L.L., Escobedo, R., Perales, A.: Phys. Rev. B **68**, 241304(R) (2003)
3. Bonilla, L.L., Escobedo, R.: Math. Mod. Meth. Appl. Sci. **15**(8), 1253 (2005)
4. Sánchez, D., MacDonald, A.H., Platero, G.: Phys. Rev. B **65**, 035301 (2002)
5. Kleinert, P., Bryksin, V.V., Bleibaum, O.: Phys. Rev. B **72**, 195311 (2005)
6. Bonilla, L.L., Barletti, L., Alvaro, M.: SIAM J. Appl. Math. **69**(2), 494–513 (2008)

Self-Sustained Spin-Polarized Current Oscillations in Multiquantum Well Structures

M. Carretero[1,2], L.L. Bonilla[1,2], R. Escobedo[3] and G. Platero[4]

[1] G. Millán Institute, Fluid Dynamics, Nanoscience & Industrial Mathematics, Universidad Carlos III, 28911 Leganés, Spain, manuel.carretero@uc3m.es
[2] Unidad Asociada al Instituto de Ciencia de Materiales de Madrid, CSIC, 28049 Cantoblanco, Spain, bonilla@ing.uc3m.es
[3] Departamento de Matemática Aplicada y Ciencias de la Computación, Universidad de Cantabria, 39005 Santander, Spain, escobedo@unican.es
[4] Instituto de Ciencia de Materiales de Madrid, CSIC, 28049 Cantoblanco, Spain gplatero@csic.es

Summary. A semiconductor multiquantum well structure exhibits self-sustained spin-polarized current oscillations if one or more of its wells are doped with Mn. Analysis and numerical solution of a nonlinear spin transport model yield the minimal number of wells and the range of doping density needed to find oscillations.

1 Introduction

Spintronics is a multidisciplinary field whose central theme is the active manipulation of spin degrees of freedom in solid states systems. Among the fields that are involved in spintronics, magnetoelectronics has achieved important results regarding magnetoresistive effects which are important since they can be used for magnetic read heads in computer hard drives and non-volatile random access memory [1]. Semiconducting materials offer the possibility of new device functionalities not realizable in metallic systems. In particular, Diluted Magnetic Semiconductors (DMS) with nonlinear current-voltage characteristics can be associated with non-magnetic semiconductors to produce efficient spin injectors [2, 3] or be used as spin oscillators [4, 5]. The present work models a dc voltage biased II-VI semiconductor Multiquantum Well Structure (MQWS) attached to normal contacts with at least one quantum well (QW) doped with Mn, thereby constituting a DMS. An external magnetic field causes splitting of energy levels in the DMS and this induces spin polarization in the MQWS. We find that MQWS with at least four QWs exhibit Self Sustained Current Oscillations (SSCOs) that can be used to design spin oscillators. There are interesting spatio-temporal patterns for long MQWS.

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*, Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_17,

2 Governing Equations

2.1 Theoretical Model

The sample under consideration consists of an n-doped ZnSe/(Zn,Cd,Mn)Se weakly coupled MQWS. Under an external magnetic field B, the DMS in QWs doped with Mn^{++} (with spin $S = 5/2$) have spin-dependent energy levels: $E_j^{\pm} = E_j \mp \Delta/2$ for electron spin $s = \pm 1/2$. The level splitting Δ is a function of B and the Mn density [3], and we can consider it as a tunable parameter.

The governing equations describing our model [5], for a MQWS of N wells, are:

$$F_i - F_{i-1} = \frac{e}{\varepsilon}\,(n_i^+ + n_i^- - N_D), \tag{1}$$

$$e\,\frac{\mathrm{d}n_i^{\pm}}{\mathrm{d}t} = J_{i-1\to i}^{\pm} - J_{i\to i+1}^{\pm} \pm \frac{n_i^- - n_i^+/\Theta_i}{\tau_{\mathrm{sf}}}, \tag{2}$$

$$\Theta_i = 1 + e^{\frac{E_{1,i}^- - \mu_i^+}{\gamma_\mu}}, \tag{3}$$

where $i = 1, \ldots, N$. Here n_i^+, n_i^- and $-F_i$ are the two-dimensional (2D) spin-up and spin-down electron densities, and the average electric field at the ith MQWS period, respectively. The voltage bias condition is $\sum_{i=0}^{N} F_i l = V$ for an applied voltage V. We have denoted the spin-dependent subband energies (E) by $E_{j,i}^{\pm} = E_j \mp \Delta_i/2$, with $\Delta_i = \Delta$ or 0, depending on whether the ith well contains magnetic impurities. N_D, ε, $-e$, l, τ_{sf} and $-J_{i\to i+1}^{\pm}$ are the 2D doping density at the QWs, the average permittivity, the electron charge, the width of a MQWS period, the spin-flip scattering time, and the tunneling current density across the ith barrier, respectively. For numerical convenience, the right hand side of (2) contains a smoothed form of the scattering term used in [3]. Time-differencing (1) and inserting (2) in the result, we obtain the following form of Ampère's law,

$$\varepsilon\,\frac{\mathrm{d}F_i}{\mathrm{d}t} + J_{i\to i+1} = J(t) = \frac{1}{N+1}\sum_{i=0}^{N} J_{i\to i+1}, \tag{4}$$

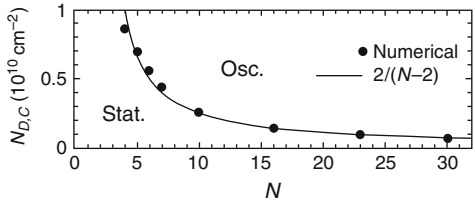


Fig. 1. Minimal doping density N_D for SSCOs vs the number of wells N for $\Delta = 15$ MeV

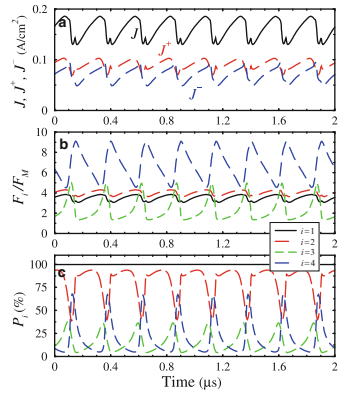


Fig. 2. (**a**) Tunneling current, (**b**) electric field and (**c**) polarization, as a function of time at the i QW during SSCOs for $N = 4$. *Solid line* ($i = 1$); *dot-dashed* ($i = 2$); *dashed* ($i = 3$); *long-dashed* ($i = 4$)

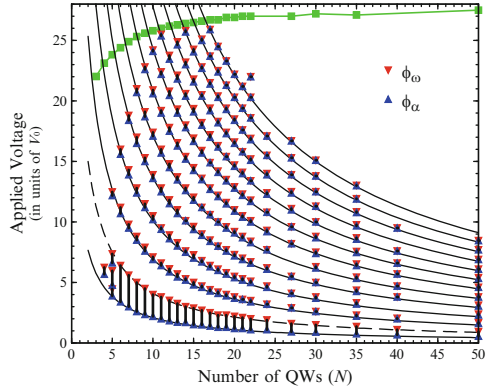


Fig. 3. Phase diagram of average electric field ϕ vs N for a MQWS containing Mn in its first QW. The SSCOs begin at *triangles* ϕ_α and end at *inverted triangles* ϕ_ω

where $J_{i\to i+1} = J^+_{i\to i+1} + J^-_{i\to i+1}$. In (4), $J(t)$ is the total current density. Tunneling currents are calculated taking into account that spin up and down electrons have different energies:

$$J^{\pm}_{i\to i+1} = \frac{e\, v^{(f)\pm}(F_i)}{l}\left\{ n_i^{\pm} - a \ln\left[1 + e^{-\frac{eF_i l}{k_B T}}\left(e^{\frac{n^{\pm}_{i+1}}{a}} - 1\right)\right]\right\}, \qquad (5)$$

for $i = 1, \ldots, N-1$, with $a = \frac{m^* k_B T}{2\pi\hbar^2}$ [6]. As boundary tunnelling currents for $i = 0$ and N, we use (5) with $n_0^{\pm} = n_{N+1}^{\pm} = \kappa N_D/2$ [3]. Initially, we set $F_i = V/[l(N+1)]$, $n_i^{\pm} = N_D/2$, $v^{(f)\pm}(F_i)$ is the "forward tunneling velocity", see details in [6]. The currents $J^{\pm}_{i\to i+1}$ are functions of F_i, $n_i^{\pm}$ and $n_{i+1}^{\pm}$. For constant values $n_i^{\pm} = N_D/2$ and $F_i = F$, the tunneling current density at a nonmagnetic QW has a maximum J_M at a value F_M of the field. In terms of

F_M, the voltage bias condition can be written as a condition for the average field ϕ:

$$\frac{1}{(N+1)F_M}\sum_{i=0}^{N} F_i = \phi \equiv \frac{V}{V_0} = \frac{V}{(N+1)F_M l}. \tag{6}$$

3 Results

3.1 Short Devices: Spin Injector

Nonmagnetic MQWSs do not exhibit SSCOs: at least one QW has to contain magnetic impurities (Mn). Let that QW be the first one, next to the injecting contact. Figure 1 shows that the MQWS should have at least four QWs and sufficient doping density ($N > N_{DC}$) for SSCOs to exist. The critical doping density is approximately given by $N_{DC} = \frac{2}{N-2} \times 10^{10}$ cm^{-2}.

Figure 2 shows the time evolution of the spin-polarized current densities, the electric field and the spin polarization (defined as $P_i = (n_i^+ - n_i^-)/(n_i^+ + n_i^-)$) at the different periods of a 4-well MQWS with normal contacts. SSCOs are caused by repeated nucleation and motion of electric field pulses which are charge dipole waves. During one oscillation period, the first QW is fully polarized, the second QW is highly polarized and the third and the fourth QWs are strongly polarized when the dipole wave is traversing them. These results should be useful to build an oscillatory spin polarized current injector.

3.2 Long Devices: Spatio-Temporal Patterns

For typical values of the parameters (see [5]) our model exhibits a variety of stationary states with electric file domains (EFD) and SSCOs.

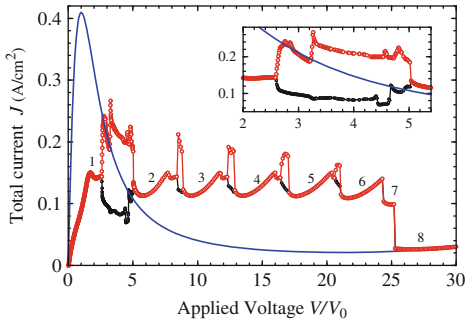


Fig. 4. Current-voltage characteristics for a 8-period MQWS. The maximum and minimum of the SSCOs has been represented with *circles* in each voltage interval where they exist

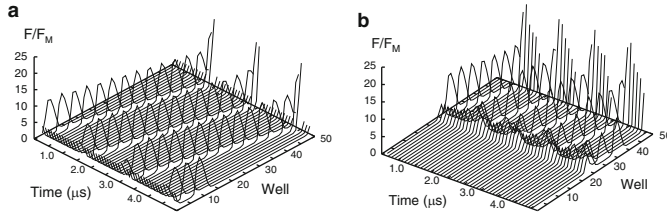


Fig. 5. Electric field profile vs QW index and time for $N = 50$, during SSCOs, if the magnetic QW is (**a**) $i = 1$, (**b**) $i = 25$

Phase Diagram

For $N_D = 10^{10}\text{cm}^{-2}$, SSCOs appear in several intervals of the average field (6), $\phi^N_{\alpha,k} < \phi < \phi^N_{\omega,k}$, $k \in [1, 2, \ldots]$. The number and width of these intervals of oscillatory solutions depend on N, as shown in Fig. 3, where $\phi^N_{\alpha,k}$ and $\phi^N_{\omega,k}$ are marked with triangles and inverted triangles, respectively. We observe that the sequence of ϕ at which oscillations appear can be approximated by the formula $\phi^N_{\alpha,k} = 38k/(N+1)$, which provides the solid lines in Fig. 3.

Current-Voltage Characteristics

As can be seen in Fig. 3, the first bias interval for which there are SSCOs is the widest. For instance, for $N = 8$ this interval is $2.59 < \phi < 5.04$. The current-voltage characteristics (I-V) for a 8-period MQWS is depicted in Fig. 4, in which we show the maximum and minimum values of the current during SSCOs for biases in the voltage intervals $(\phi^8_{\alpha,k}, \phi^8_{\omega,k})$, $k = 1, \ldots, 5$.

Stationary States

As a general rule, the stationary states for $\phi < \phi^N_{\alpha,1}$ and for $\phi^N_{\omega,kmax} < \phi$ are spatially almost uniform. In the other stationary intervals of the I-V diagram, the profiles of F_i, P_i and $n_i^{\pm}$ are not uniform and there are two EFDs, a low field domain adjacent to the cathode and a high field domain that extends to the anode, separated by a domain wall in which the field increases.

SSCOs

Examples of SSCOs for long MQWS are shown in Fig. 5 when only one well is magnetic. We have found that if the only magnetic QW is the ith (with $1 \leq i < N-3$), the charge dipoles are emitted at this well, and dipole motion is limited to the last $N - i$ QWs.

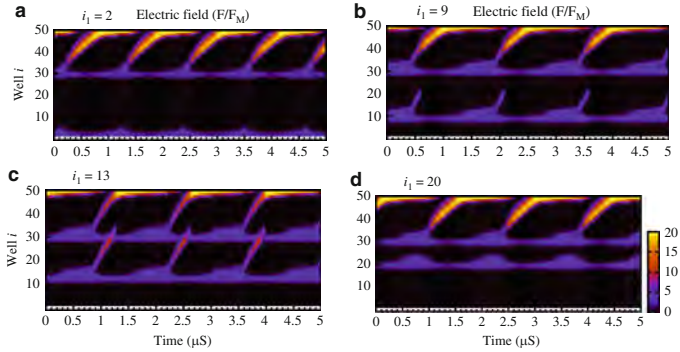


Fig. 6. Density plot of the electric field profile during SSCOs for $N = 50$ and DMS QW at $i_2 = 30$ and: (**a**) $i_1 = 2$, (**b**) $i_1 = 9$, (**c**) $i_1 = 13$, (**a**) $i_1 = 20$

MQWS with Two Magnetic Quantum Wells

MQWS with two QWs containing magnetic impurities exhibit a very rich dynamical behavior. Figure 6 shows the density plot of the electric field profile during SSCOs for a 50-well structure having two magnetic QWs. We have: (a) and (d) inhibition of dipole triggering at the first magnetic QW; (b) short and (c) long movements of the dipole waves.

4 Conclusions and Further Work

We have studied the dynamical behavior of MQWS with DMS in at least one QW. Spin oscillators may be designed using the results for short devices while long devices exhibit interesting patterns. Future work will explore switching the magnetic field through Δ.

References

1. Bandyopadhyay, S., Cahay, M.: Introduction to Spintronics, 2nd edn. CRC, Boca Raton (2008)
2. Zutic, I., Fabian, J., Das Sarma, S.: Rev. Mod. Phys. **76**, 323 (2004)
3. Sánchez, D., MacDonald, A.H., Platero, G.: Phys. Rev. B **65**, 035301 (2002)
4. Béjar, M., Sánchez, D., Platero, G., MacDonald, A.H.: Phys. Rev. B **67**, 045324 (2003)
5. Bonilla, L.L., Escobedo, R., Carretero, M., Platero, G.: Appl. Phys. Lett. **91**, 092102 (2007)
6. Bonilla, L.L.: J. Phys. Cond. Matter **14**, R341 (2002)

Spin Dynamics in Quantum Dots

Gloria Platero[1], Jesús Iñarrea[1,2], and Carlos López-Monís[1]

[1] Instituto de Ciencia de Materiales, CSIC, Cantoblanco, Madrid, 28049, Spain, gplatero@csic.es
[2] Escuela Politécnica Superior, Universidad Carlos III, Leganes, Madrid, Spain

Summary. Leakage current of double quantum dot systems in the spin blockade regime has been attributed to hyperfine interactions. In this work electron transport through double quantum dots is analyzed in the spin blockade regime, in the presence of hyperfine interaction by means of rate equations. In agreement with experiment, current hysteresis as a function of magnetic field is found. This behavior comes from the interplay between dynamic nuclear spin polarization and the electronic energy states renormalization due to the Overhauser shifts induced by the nuclei.

1 Introduction

The Pauli exclusion principle can play an important role in current rectification [1, 2] in both molecular and semiconductor nano-structures transport. Spin blockade (SB), which occurs in double quantum dots (DQDs) over certain ranges of gate voltage, external magnetic field, and bias voltage is one important example. The interplay between Coulomb and spin blockade can be used to block current in one direction of bias voltage while allowing it to flow in the opposite one. Because of this property DQDs can function as externally controllable spin-Coulomb rectifiers that have potential application in spintronics, as spin memories and transistors. Spin relaxation processes [3–5], induced by spin-orbit (SO) scattering [6] or hyperfine (HF) interactions [7–12], produce a leakage current which limits the SB resistance. Spin-flip (sf) relaxation times in QDs are rather long however and the SB resistance is large.

In this paper we report on a model for transport through two weakly coupled vertical QD's in the spin blockade configuration [13, 14]. Recent experiments [1] show current leakage in the Spin Blockade regime which is attributed to hyperfine interaction between the nuclei spins and the electronic spins. On a spin blockade plateau, current flow between dots is possible only when each dot has one electron and their spins are opposite. A finite bias voltage allows an electron in the left dot to tunnel sequentially to the double

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*,
Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_18,

occupied singlet state in the right dot and then to the collector. In this circumstance there is an approximately even chance that electrons in left and right dots will have the same spin when the left dot electron is refreshed from the source. When that happens, the Pauli exclusion principle prevents tunneling. Current flow stops until a spin flip takes place. The time averaged current is consequently strongly suppressed. The SB blockade regime is conveniently tuned by an external magnetic field (B). Fields applied in the plane of the quantum dots introduce a Zeeman energy splitting of the levels. Increasing the field allows to tune the relative energy between states with antiparallel spins and with parallel spins and bring them close to degeneracy. In this situation, it has been shown, both theoretically and experimentally that the current presents instabilities and hysteretic behavior [15]. We explain this behavior by accounting for the interplay between dynamic nuclear polarization and Overhauser shifts suffered by the electronic levels which are induced by the nuclei spins. Recent experiments by Koppens et al. show similar instabilities and bistable regions in the current as a function of magnetic field [16] through a lateral DQD, which likely have a similar explanation. Also, current hysteresis has been observed in InAs quantum dots by the group of Ensslin [17].

2 Electronic Transport Through Double Quantum Dots: Role of Hyperfine Interaction

2.1 Theoretical Model

We consider the Hamiltonian: $H = H_L + H_R + H_T^{LR} + H_{leads} + H_T^{l,D}$, where $H_L, (H_R)$ is the Hamiltonian for the isolated left (right) QD modeled as one-orbital Anderson impurity. H_T^{LR} and $H_T^{l,D}$ describe tunneling between QDs and between leads and QDs respectively, and H_{leads} is the Hamiltonian for the leads. In the presence of an external magnetic field and hyperfine interaction there is an additional contribution to the hamiltonian:

$$\hat{H} = g_e \mu_B \mathbf{S}\cdot\mathbf{B} + \frac{A}{N}\sum_{i=1}^{N}\left[S_z I_z^i + \frac{1}{2}(S_+ I_-^i + S_- I_+^i)\right] \tag{1}$$

where the average hyperfine coupling constant is $A \simeq 90\,\mu\text{eV}$ for GaAs and B is the external magnetic field. The basis considered consists on the eigenstates for the isolated quantum dots.

Rate equations for the occupation probabilities ρ_{ss} corresponding to the electronic states become:

$$\dot{\rho}(t)_{ss} = \sum_{m\neq s} W_{sm}\rho_{mm} - \sum_{k\neq s} W_{ks}\rho_{ss} \tag{2}$$

where W_{ij} are transition rates for the tunneling through the contact barriers and for the tunneling through the interdot barrier. We consider incoherent interdot tunnel and we account for both elastic and inelastic inter-dot

tunneling [13,14]. We consider as well transition rates which involve spin-flip coming from hyperfine interaction, as we will describe below. We calculate the electronic spin-flip scattering rate $W_{i,j}^{sf}$ using a microscopic model that accounts for HF interactions: The HF interaction can then be separated into mean-field and flip-flop contributions: $\hat{H} = \hat{H}_z + \hat{H}_{sf}$ where $\hat{H}_z = A\langle I_z\rangle S_z$ has an effective nuclear field $B_N = A\langle I_z\rangle/g_e\mu_B$ contribution which is added to the external magnetic field contribution to produce an effective Zeeman splitting of the levels and

$$\langle I_z\rangle = \frac{1}{N}\sum_{i=1}^{N}(I_z^i) = \left[\frac{N^{\uparrow} - N^{\downarrow}}{N^{\uparrow} + N^{\downarrow}}\right]|I_z| = P|I_z| \tag{3}$$

where $P = \left[\frac{N^{\uparrow}-N^{\downarrow}}{N^{\uparrow}+N^{\downarrow}}\right]$ is the nuclear spin polarization where $N^{\uparrow(\downarrow)}$ is the number of nuclei with spin up(down), in a QD. $\hat{H}_{sf} = (A/2N)\sum_i \left[S_+ I_-^i + S_- I_+^i\right]$ is the flip-flop interaction responsible for mutual electronic and nuclear spin flips. Because of the mismatch between nuclear and electronic Zeeman energies transitions must be accompanied at low temperature by phonon emission. We approximate the spin-flip transition rate from parallel-spin to opposite-spin configurations by:

$$\frac{1}{\tau_{sf}} \simeq \frac{2\pi}{\hbar}|<\hat{H}_{sf}>|^2 \frac{\gamma}{\Delta E^2 + \gamma^2} \tag{4}$$

where γ is the electronic state life-time broadening which is of the order of μeV, i.e., of the order of the phonon scattering rate [3,4]. ΔE is the difference between the energy of a state with one electron in each dot with aligned spins ($|\downarrow,\downarrow\rangle/|\uparrow,\uparrow\rangle$)and the energy of a state with one electron in each dot with opposite spin orientation ($|\uparrow,\downarrow\rangle/|\downarrow,\uparrow\rangle$) (see Fig. 1). The latter are *mixed* due to interdot tunneling with the intradot singlet state in the right QD ($|0,\downarrow,\uparrow\rangle$). The energy of the *mixed* state with antiparallel spins is calculated perturbatively and depends mainly on the interdot tunneling (t) and the right and left dots level detuning.

In resonance, at $B \neq 0$, ΔE depends on the Zeeman energy due to the external field B and on the additional Zeeman splitting due to the magnetic field induced by the nuclei:

$$\Delta E = E_{(|\downarrow,\downarrow\rangle/|\uparrow,\uparrow\rangle)} - E_{(|\uparrow,\downarrow\rangle/|\downarrow,\uparrow\rangle)} = g_e\mu_B B + \frac{A}{2}P \tag{5}$$

The equations that describe the time evolution of the nuclear spin polarization for both dots include the flip-flop interaction and the nuclear spin relaxation time τ_{relax} (that we include phenomenologically and that is much longer than the electron-nuclei spin scattering time) become:

$$\dot{P}_L = W_{6,3}^{sf}\rho_3 - W_{5,4}^{sf}\rho_4 - \frac{P_L}{\tau_{relax}} \tag{6}$$

$$\dot{P}_R = W_{5,3}\rho_3 - W_{6,4}^{sf}\rho_4 - \frac{P_R}{\tau_{relax}} \tag{7}$$

Here, for instance:

$$|4\rangle \equiv |\downarrow,\downarrow\rangle \rightarrow |5\rangle \equiv |\uparrow,\downarrow\rangle \Rightarrow W^{sf}_{5,4} = \left[\frac{1}{\tau_{sf}}\right]_L \left[\frac{1+P_L}{2}\right], \tag{8}$$

where L and R mean left and right dot respectively. The system of time evolution equations for the electronic states occupations ρ_i and nuclear polarization of the left and right dot is self-consistently solved. From that we calculate the total current through the system which is the physical observable of interest.

For $B = 0$, experiments [1] show a weak peak in the current at low V_{DC} followed by a wide plateau, and then finally a very strong peak at $V_{DC} \geq 6$ MeV, in good agreement with the results plotted in Fig. 1. The leakage current observed in the plateau is due to the finite probability for electrons in the QD's to flip their spin by interaction with nuclei.

In Fig. 2 we show I/B (B in-plane) for V_{DC} near the center of the SB region. When sweeping up and down the magnetic field, we find current hysteretic behavior in agreement with experiment [15]. The source of this behavior is the interplay between the induced nuclei polarization due to HF interaction and the energy shift induced in the electronic states by the nuclear magnetic field which modifies the spin-flip rate. At small B, for the DC voltage that we have considered the $|\downarrow,\downarrow\rangle$ state has lower energy than the state with antiparallel spins, and then, at low temperatures spin-flip has low probability. Increasing B, the state $|\downarrow,\downarrow\rangle$ becomes higher in energy than the state with

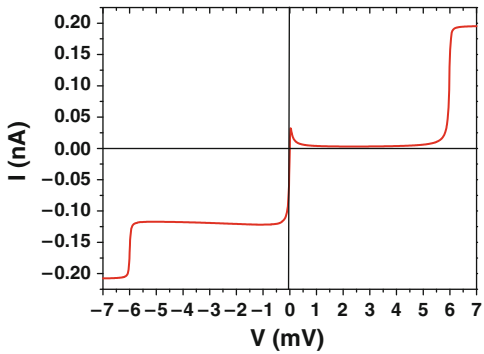


Fig. 1. Stationary I/V_{DC} (B = 0). At low V_{DC}, I takes place when one electron from the (1, 1) state with two electrons, one in each dot, with opposite spins tunnels to the singlet double occupied state in the right QD (0, 2). Once one electron tunnels from the emitter contact to the left dot with the same spin polarization as the electron in the right dot, the current drops abruptly due to spin blockade. A finite current leakage is observed due to spin flip induced by HF interaction. At $V_{DC} \geq 6$ MeV the chemical potential of the right lead crosses the (1, 1) state with parallel spins and the right QD becomes suddenly discharged producing a large peak in I

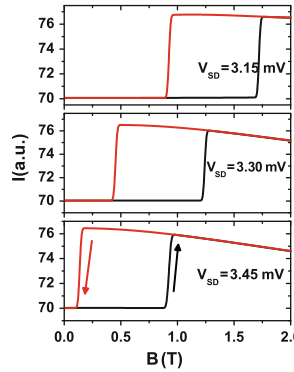


Fig. 2. I/B in the SB region for a DQD under in-plane B for different DC voltages. The current shows hysteresis reflecting strong non-linearities induced by the interplay of electron and nuclear spin dynamics. Electronic levels energies depend on the level detuning which depends on the source-drain voltage. This is the reason why the hysteresis region shifts with voltage

antiparallel spins and then, electrons have a finite spin-flip rate and relax to states with antiparallel spins, producing a small leakage current. In this case, as the electronic spin flips from down to up the nuclei spin flips from up to down and the effective field produced by the nuclei is aligned with the external field. This feed-back mechanism between the nuclei polarization and the electron state renormalization implies hysteresis in the electronic current as a function of B as observed experimentally [15].

In conclusion we have proposed a model which describes charge transport through double quantum dots in the spin-blockade regime including HF interactions. The interplay between electronic charge occupation and spin polarization of the nuclei is accounted for by solving coupled rate equations self-consistently. We interpret current features seen experimentally [15] at the SB regime as evidence for hyperfine interaction between electronic and nuclei spins in the double quantum dot structure. At the SB plateau, electronic spin flip from states with parallel spins to states with antiparallel spins states produces a nuclear field which shift the electronic levels. This shift modifies the spin-flip rate and therefore the nuclei polarization. This feed-back is responsible of the strongly non linear current behaviour.

Acknowledgements

This work has been supported by the MCYT (Spain) under grant MAT2008-02626/NAN.

References

1. Ono, K., et al.: Science **297**, 1313 (2002)
2. Johnson, A.C., et al.: Phys. Rev. B **72**, 165308 (2005)
3. Fujisawa, T., et al.: Nature (London) **419**, 278 (2002)
4. Elzerman, J.M., et al.: Nature **430**, 431 (2004)
5. Gywat, O., et al.: Phys. Rev. B **69**, 205303 (2004)
6. Golovach, V.N., et al.: Phys. Rev. Lett. **93**, 016601 (2004)
7. Erlingsson, S.I., et al.: Phys. Rev. B **64**, 195306 (2001)
8. Erlingsson, S.I., et al.: Phys. Rev. B **66**, 155327 (2002)
9. Erlingsson, S.I., et al.: Phys. Rev. B **72**, 033301 (2005)
10. Khaetskii, A.V., et al.: Phys. Rev. B **61**, 12639 (2000)
11. Merkulov, I.A., et al.: Phys. Rev. B **65**, 205309 (2002)
12. Coish, W.A., et al.: Phys. Rev. B **70**, 195340 (2004)
13. Iñarrea, J., Platero, G., MacDonald, A.H.: Phys. Rev. B **76**, 085329 (2007)
14. Iñarrea, J., Lopez-Monís, C., MacDonald, A.H., Platero, G.: Appl. Phys. Lett. **91**, 252112 (2007)
15. Ono, K., Tarucha, S.: Phys. Rev. Lett. **92**, 256803 (2004)
16. Koppens, F.H., et al.: Science **309**, 1346 (2005)
17. Pfund, A., et al.: Phys. Rev. Lett. **99**, 036801 (2007)

Relocation Dynamics During Voltage Switching in Spin-Polarized Superlattices

R. Escobedo[1], M. Carretero[2,3], L.L. Bonilla[2,3], and G. Platero[4]

[1] Departamento de Matemática Aplicada y Ciencias de la Computación, Universidad de Cantabria, 39005 Santander, Spain, escobedo@unican.es
[2] G. Millán Institute, Fluid Dynamics, Nanoscience and Industrial Mathematics, Universidad Carlos III de Madrid, 28911 Leganés, Spain, manuel.carretero@uc3m.es
[3] Unidad Asociada al Instituto de Ciencia de Materiales de Madrid, CSIC, 28049 Cantoblanco, Madrid, Spain bonilla@ing.uc3m.es
[4] Instituto de Ciencia de Materiales de Madrid, CSIC, 28049 Cantoblanco, Madrid, Spain, gplatero@csic.es

Summary. A numerical study of domain wall relocation during voltage switching in a semiconductor device whose current-voltage characteristic exhibits the alternation of intervals of stationary states with intervals of oscillatory states is presented. Voltage switching between voltage intervals always induces the emission of a relocation wave, but relocation can be permanent or not depending on the voltage increase.

1 Introduction

A II-VI semiconductor superlattice (SL) having its first quantum well (QW) doped with magnetic ions and attached to normal contacts is studied numerically when a sudden increase is applied to the external applied voltage. A numerical study has been recently presented showing that, for a large range of values of the physical parameters, the current-voltage characteristic curve J–V representing the variation of the total current density J with respect to the applied voltage V exhibits voltage intervals of stationary states alternating with voltage intervals of self-sustained spin-polarized current oscillations due to the repeated triggering of charge dipole waves at the magnetic well and their motion towards the collector. See Fig. 1.

The alternation of intervals of stationary and oscillatory solutions makes this structure suitable to study the electric field domain wall relocation phenomenon due to a sudden increase of the voltage across the SL, reported from experiments by Luo et al. [1–5] and studied numerically in Bonilla et al. [6,7].

The J–V curve in Fig. 1 shows that four possible scenarios exist for applying the voltage jump, depending on if the initial and final values are located in an interval of stationary or oscillatory states. These relocation scenarios are

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*, Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_19,

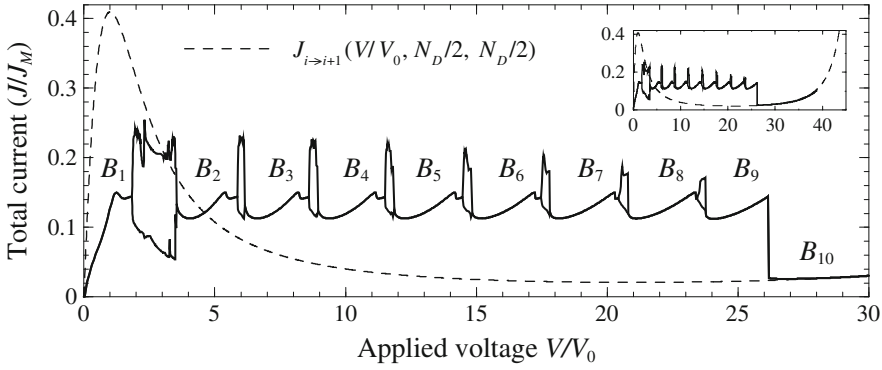


Fig. 1. J–V characteristic for a structure with $N = 12$ QWs (*solid line*), showing the stationary states branches B_k, $k = 1, \ldots, N-2 = 10$, and the intervals of oscillatory states, where maximum and minimum current values are depicted. *Dashed line* is $J_{i\to i+1}(V/V_0, N_D/2, N_D/2)$. *Inset*: same for a larger range of voltage

described numerically in terms of two voltage absorption mechanisms. Also, a minimum voltage increase for a relocation to be permanent is detected.

2 Governing Equations

The main variables are the two-dimensional (2D) spin-up and spin-down electron densities $n_i^+(t)$ and $n_i^-(t)$, the average electric field at the ith SL period $-F_i(t)$ and the applied voltage $V(t)$. The ith SL period starts at the right end of the $(i-1)$th barrier and finishes at the right end of the ith barrier.

The model equations are, for $i = 1, \ldots, N$,

$$\varepsilon \frac{\mathrm{d}F_i(t)}{\mathrm{d}t} = J(t) - J_{i\to i+1}(t), \tag{1}$$

$$e \frac{\mathrm{d}n_i^{\pm}(t)}{\mathrm{d}t} = J_{i-1\to i}^{\pm}(t) - J_{i\to i+1}^{\pm}(t) \pm \frac{1}{\tau_{\mathrm{sf}}}\Big(n_i^-(t) - A(\mu_i^+)\, n_i^+(t)\Big), \tag{2}$$

$$\sum_{i=0}^{N} F_i(t) l = V(t). \tag{3}$$

Here $A(\mu_i^+) = \left(1 + e^{\frac{E_{1,i}^- - \mu_i^+(t)}{\gamma_\mu}}\right)^{-1}$, (1) is the Ampère's law obtained by time-differencing the Poisson's equation $\varepsilon(F_i - F_{i-1}) = e(n_i^+ + n_i^- - N_D)$ and inserting (2) in the result, and (3) is the voltage bias condition.

There $J_{i\to i+1} = J_{i\to i+1}^+ + J_{i\to i+1}^-$ are the tunneling currents densities across the ith barrier, calculated by the Transfer Hamiltonian method taking into account that spin up and down electrons have different energies, and provided that scattering-induced broadening of energy levels is much smaller than subband energies and chemical potentials [8]: $J_{i\to i+1}^{\pm} = B(F_i, n_i^{\pm}, n_{i+1}^{\pm})\, v^{(f)\pm}(F_i)$,

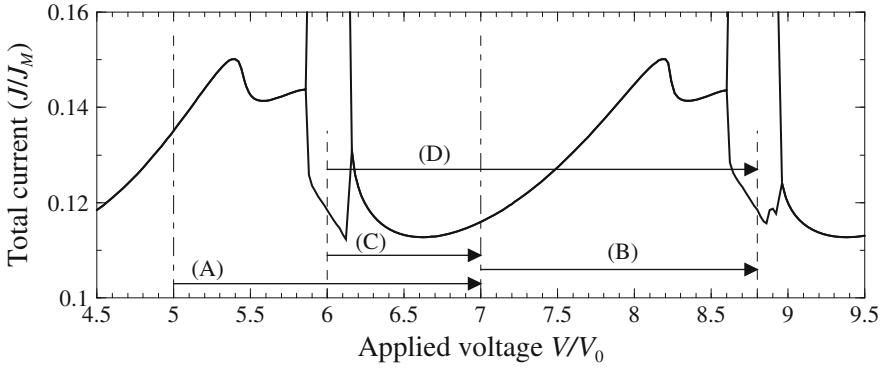


Fig. 2. Branches B_2 and B_3 (*solid line*) and second and third intervals of oscillatory solutions corresponding to the J–V depicted in Fig. 1. *Vertical lines* denote onset and end of voltage switchings: stationary states (*dot-dashed*) and oscillatory states (*dashed lines*). *Horizontal arrows* denote the four scenarios: (*A*) from SS to SS, (*B*) from SS to OS, (*C*) from OS to SS and (*D*) from OS to OS

where $v^{(f)\pm}$ is the spin-dependent "forward tunneling velocity" (which is a sum of Lorentzians). The expressions of $v^{(f)\pm}$ and B can be found in [8].

The total current density $J(t)$ is independent of i, as it can be written as

$$J(t) = \frac{1}{N+1}\left[\sum_{i=0}^{N} J_{i\to i+1}(t) + \frac{\varepsilon}{l}\frac{dV(t)}{dt}\right] \tag{4}$$

by adding (1) for all i and time-differencing (3).

The rest of the parameters are the spin-dependent subband energies (measured from the bottom of the ith well) $E^{\pm}_{j,i} = E_j \mp \Delta_i/2$, with $\Delta_i = \Delta$ or 0, depending on whether the ith well contains magnetic impurities, and N_D, ε, $-e$, $l = d+w$, and $\mu^{\pm}_i(t)$, which are, respectively, the 2D doping density at the QWs, the average permittivity, the electron charge, the width of a SL period (d and w are barrier and well widths), and the chemical potentials at the ith SL period for electrons with spin $\pm 1/2$, related to the electron densities by $n^{\pm}_i(t) = \rho \ln[C(\mu^{\pm}_i(t))]$, whose detailed expression can be found in [8].

The tunneling current density $J_{i\to i+1}$ is a function of F_i, $n^{\pm}_i$ and $n^{\pm}_{i+1}$, which has, for constant values $n^{\pm}_i = N_D/2$ and $F_i = F$ at a nonmagnetic QW, a maximum J_M at a value F_M of the field. In terms of F_M, the voltage bias condition (3) can be written as a condition for the average field $\phi(t)$:

$$\phi(t) \equiv \frac{V(t)}{(N+1)F_M l} = \frac{1}{(N+1)F_M}\sum_{i=0}^{N} F_i(t). \tag{5}$$

As boundary tunnelling currents ($i = 0, N$), B is evaluated at $n^{\pm}_0 = n^{\pm}_{N+1} = \kappa N_D/2$ (identical normal contacts with $\kappa \geq 1$) [9]. Initially, we set $F_i(0) = V(0)/[l(N+1)]$ (i.e. $\phi(0) = F_i(0)/F_M$), and $n^{\pm}_i = N_D/2$ (normal QWs).

3 Results

A sample with $d = 10\,\text{nm}$, $w = 5\,\text{nm}$, $m^* = 0.16m_0$, $\tau_{\text{sf}} = 10^{-9}$ s (normal QWs) and 10^{-11}s (magnetic QWs), $\varepsilon = 7.1\varepsilon_0$, $T = 5\,\text{K}$, $E_1 = 15.76\,\text{MeV}$, $E_2 = 61.99\,\text{MeV}$, $\gamma = 1\,\text{MeV}$, $\gamma_\mu = 0.1\,\text{MeV}$ and $\kappa = 1$ is considered [9]. Only the first QW has magnetic impurities, yielding a spin splitting $\Delta = 12\,\text{MeV}$.

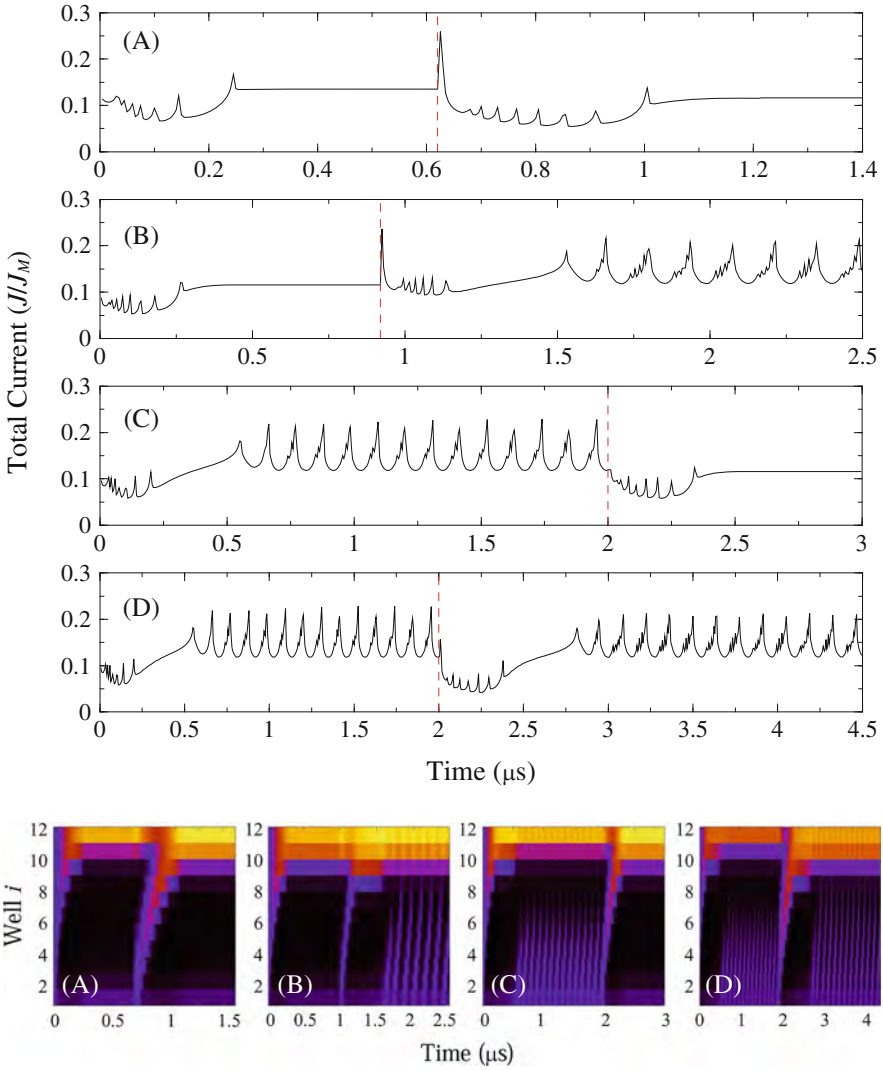


Fig. 3. Total current density $J(t)$ and electric field density plots for the four relocation scenarios: (A) SS → SS, (B) SS → OS, (C) OS → SS and (D) OS → OS

Figure 2 shows the four possible scenarios, depending on if the initial and final voltages correspond to a stationary state (SS) or an oscillatory state (OS).

Figure 3 shows the total current density $J(t)$ (solid 2D curves) and the corresponding electric field distribution $\{F_i(t)\}_{i=0}^{N}$ (density plots) for the four scenarios described in Fig. 2. Vertical dashed lines denote the time at which the voltage switching is applied.

The current oscillations due to the repeated triggering of charge dipole waves at the magnetic well $i = 1$ and their motion towards the collector $i = N$, together with the motion of these waves, can be observed in the three scenarios involving voltage values in intervals of oscillatory solutions; see scenarios B, C and D in Fig. 3.

The electric field profile outside these recycling waves consists of a domain wall located at a quantum well i_a, which separates a low field domain going from $i = 1$ to i_a and a high field domain going from i_a to $i = N$. The sudden increase of the applied voltage always induces the nucleation of a large wave which travels towards the domain wall, where it is absorbed; this is the mechanism by which the electric field profile absorbs the increase of area imposed by (5). The size of this relocation wave is always larger than the size of the typical waves corresponding to current oscillations. When the relocation wave arrives to the domain wall located at i_a (here located near the collector), the electric field accommodates itself to the stable solution corresponding to the final voltage value. This accommodation induces the relocation of the domain wall, which moves to the quantum well $i_a - 1$, except in scenario B, where the domain wall remains pinned at QW i_a. Why is this?

There are two mechanisms to absorb the voltage increase: one is by triggering a recycling wave whose area accounts for the voltage increase, and the other is by relocating the domain wall to a previous quantum well. In the first case, the domain wall relocation takes place only during the recycling of the electric field wave, whereas in the second case the relocation is permanent.

When the final state is an oscillatory state, both mechanisms can be used. Scenario D corresponds to a permanent relocation, in which the voltage increase (of size $\Delta\phi = 2.8$) is absorbed both by triggering a recycling wave and by moving back the domain wall. In turn, scenario B shows that the voltage increase (of size $\Delta\phi = 1.8$) can be absorbed by triggering a recycling wave without moving the domain wall except when the wave has to recycle.

When the final voltage corresponds to a stationary state, only the second mechanism is available: the voltage increase must be absorbed by the electric field by moving back the domain wall to the previous quantum well $i_a - 1$; this is what happens in scenario A (in which the voltage increase is of size $\Delta\phi = 2$), and also in scenario C (where $\Delta\phi = 1$), in which the resulting stationary electric field must take into account both the voltage increase and the voltage carried by the recycling waves. When the final voltage corresponds to a stationary state, the domain wall relocation is permanent.

It is not necessary to cross an interval of oscillatory solutions to induce the triggering of a relocation electric field wave yielding a permanent relocation. A relocation similar to the one obtained in scenario A can take place without leaving a stationary branch, provided the voltage increase is large enough.

In the stationary branch B_3, for a final voltage $\phi_{\text{end}} = 8.5$, a critical voltage $\phi_c = 6.65$ exists such that if the initial voltage ϕ_{ini} is lower than ϕ_c (i.e. a voltage increase greater than $\Delta\phi = 1.85$), a relocation wave is triggered and a permanent relocation takes place, whereas if $\phi_{\text{ini}} > \phi_c$, the voltage increase is absorbed by the stationary field profile without relocation nor wave nucleation.

A similar behaviour has been found for other values of $\phi_{\text{end}} \in B_3$ around 8.5, for which a minimum voltage increase $(\Delta\phi)_{\min} = 1.8$ is required to induce the relocation scenario A without leaving the stationary branch B_3.

4 Conclusions and Further Work

A numerical study of electric field domain wall relocation scenarios under a sudden voltage increase in spin-polarized semiconductor structures has been carried out. The current-voltage characteristic of these devices displays voltage intervals with stable stationary states alternating with intervals of self-sustained spin-polarized current oscillations which are due to the repeated triggering of charge dipole waves at the magnetic well and their motion towards the collector. The four different scenarios present the induction of a large electric field wave which accounts for the voltage increase when the voltage switching is applied. The scenarios have been described in terms of the electric field distribution, and two mechanisms of voltage absorption have been detected: the relocation of the domain wall of the electric field, which moves back from quantum well i_a to $i_a - 1$, and the triggering of a recycling electric field wave, which absorbs (part of) the voltage increase during its travel inside the structure, but which is accompanied by the domain relocation during its recycling. When the final voltage is located in an interval of oscillatory states in the current-voltage characteristic, both mechanisms may appear at the same time. Finally, we have observed that a relocation can take place without leaving a stationary branch during the voltage switching, provided the voltage increase is greater than a critical value.

Future work include progressive (not sudden) voltage switchings along an interval of time. This *ramping time* has proved to have an important influence in relocation scenarios [1–7].

References

1. Luo et al.: Phys. Rev. B **57** (1998)
2. Amann et al.: Phys. Rev. E **63** (2001)
3. Rogozia et al.: Phys. Rev. E **63** (2001)
4. Rogozia et al.: Phys. Rev. B **65** (2002)

5. Rogozia et al.: Physica B **314** (2002)
6. Bonilla, L.L., Escobedo, R., Dell Acqua, G.: Phys. Rev. B **73**, 115341/13 (2006)
7. Bonilla, L.L., Escobedo, R., Dell Acqua, G.: J. Comput. Appl. Math. **204** (2007)
8. Bonilla, L.L.: J. Phys. Condens. Matter **14**, R341–R381 (2002)
9. Bonilla, L.L., Escobedo, R., Carretero, M., Platero, G.: Appl. Phys. Lett. **91**, 092102/3 (2007)

Minisymposium *Dynamical Systems Methods in Aerospace Engineering*

B. Krauskopf[1] and M.H. Lowenberg[2]

[1] Department of Engineering Mathematics, University of Bristol, UK
B.Krauskopf@bristol.ac.uk
[2] Department of Aerospace Engineering, University of Bristol, UK
M.Lowenberg@bristol.ac.uk

The 1990s and early 2000s saw substantial research into the application of nonlinear dynamical systems theory in the field of aerospace engineering. This focussed principally on the flight mechanics behaviour of aircraft operating at extremes of their flight envelopes, where aerodynamic and other phenomena are significantly nonlinear. Useful results were obtained, and some practical tools were developed, for example, for the analysis of underslung loads below a helicopter, and the analysis of flight control law robustness. In spite of these initial efforts, methods from nonlinear dynamics have not as yet entered the industrial mainstream. On the other hand, there is today a growing realisation that, for the industry to develop and improve its products, it must face the fact that many of its problems are indeed nonlinear in nature. Hence, the necessary advances require that this type of behaviour is properly accounted for.

The contributions for this minisymposium showcase examples of recent nonlinear studies of aeronautical applications which are indeed being integrated into industry. Specifically, these studies show how continuation methods and bifurcation analysis are incorporated into the investigation and evaluation of a variety of aircraft systems – both from a vibration and a rigid-body motion perspective. In each case, the approach brings a new extended capability in the understanding of nonlinear engineering systems and the papers show the promise offered by these techniques to the aerospace sector in both analysis and design.

Rezgui et al. focus on the stability of a rotor in autorotation, which is investigated via the bifurcation analysis of a periodically forced system. Despite the model being relatively simple and of low order, it yields multiple autorotation solutions. At the same time an experimental rig of the system is utilised to generate experimental bifurcation diagrams. The paper shows how a coordinated implementation of both the numerical and experimental systems, the latter used to tune the model and the former to help select test conditions that are meaningful and achievable, provides a low-order model that is able to generate a rich variety of verifiable results not previously achieved.

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*,
Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_20,

The subject discussed by Rankin et al. is the nonlinear analysis of an aircraft turning on the ground. A mathematical model of low order, developed and verified with direct input from industry, reveals regions of instability of turning that may be encountered in practice as certain parameters are varied. The paper shows the results in a graphical manner intended to be accessible and easy to interpret for engineers not well versed in nonlinear systems theory. Hence, nonlinear modelling and analysis techniques become engineering tools that can be used, for example, to evaluate new design concepts at an early stage.

Shimmy in aircraft landing gear systems is a common problem encountered in design and operations, yet the methods for understanding and hence eliminating shimmy are often inadequate. In the contribution by Krauskopf et al. bifurcation analysis is applied to a mathematical model of an aircraft nose landing gear with geometric nonlinearity, which takes the form of five coupled first-order ordinary differential equations. The results clearly show the parameter regions in which one or more shimmy modes exist. The practical use of such a stability map is illustrated by simulated take-off runs of a light and a heavy aircraft.

The final contribution by Coetzee discusses the potential for nonlinear dynamical systems theory from an aerospace industry perspective. Several recent examples of their industrial use are cited, but the focus is on opportunities for the application of nonlinear methods in landing gear and related systems. Apart from specific technical challenges, the paper also discusses the need for providing the necessary framework and management support for nonlinear modelling in an industrial context.

Taken together, the contributions to the minisymposium provide useful lessons to aid in the adoption of nonlinear methods within the aerospace industry. Engineers need to understand the value not only of the traditional complex linear models, but also the power of reduced-order nonlinear models to capture important behaviour. They need to learn how to generate such models, how to analyse them with advanced tools, and how to interpret the results properly. Certainly, software for continuation and bifurcation analysis needs to be made more user-friendly for this community, and graphical methods need to be developed for both inputting information and presenting results.

In conclusion, we hope that the discussions presented here will contribute to a growing recognition of the practical benefits of nonlinear modelling and analysis in the aerospace industry and beyond.

A Combined Numerical/Experimental Continuation Approach Applied to Nonlinear Rotor Dynamics

D. Rezgui, M. Lowenberg, and P. Bunniss

Department of Aerospace Engineering, University of Bristol, UK, BS8 1TR,
Djamel.Rezgui@bristol.ac.uk, m.lowenburg@bristol.ac.uk,
peter.bunniss@bristol.ac.uk

Summary. Presented with complex systems exhibiting nonlinear behaviour, engineers in industry may face difficulties in understanding the system, both from a mathematical modeling perspective and also when trying to set up representative experiments. Here, a systematic approach combining numerical and experimental parameter continuation is applied to the investigation of complex nonlinear rotor behaviour. The aim is to show the benefits of co-ordinating numerical and physical tests in order to build a mathematical model that adequately captures the system dynamics. In this study the problem involves a dynamical system operating in a nonlinear periodic manner, with constraints on its states and parameters. The system is an autogyro rotor for which the approach generates a simple mathematical model yielding multiple possible autorotative conditions not previously identified in a systematic way; it also provides an explanation for unsafe operating scenarios.

1 Introduction

The aeromechanical stability of a rotor is a complex nonlinear problem, which involves interactions between different sources of nonlinearity. We consider the rotor of an autogyro which, unlike in a helicopter, is not driven by a power source but is kept rotating by the air flow through it; an engine with propeller provides forward thrust. It is known that when operating in autorotation at high forward speeds, the rotor can undergo unstable flapping behaviour (which has resulted in accidents) but the mechanisms at play are not well understood.

A number of methods have previously been applied to the stability of helicopter blades, ranging from pure time history simulation (time integration techniques) to parametric resonance analysis, Floquet stability theory and perturbation methods. However, these depend on assumptions that may be questionable for autorotating rotors (where rotation rate is variable, whereas for a helicopter it can be considered fixed) and inadequate to cover their entire stability picture.

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*,
Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_21,

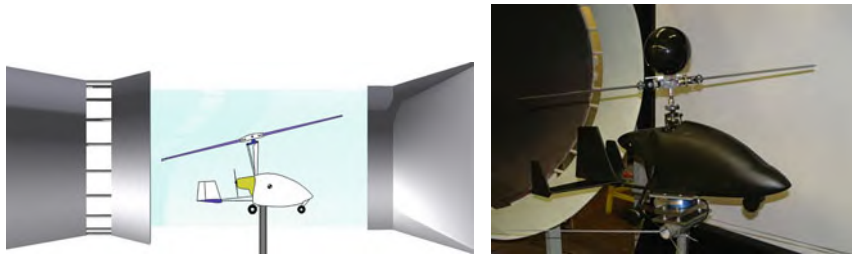


Fig. 1. Schematic view and photograph of autogyro rig in wind tunnel

Continuation and bifurcation methods have been successfully deployed to study the stability and control of a helicopter model, represented as a periodically forced nonlinear system with constant rotor speed [1]. The approach has now been extended to autorotating rotors using a relatively low order nonlinear mathematical model. These investigations using continuation and bifurcation methods [2, 3] have confirmed that an autorotating rotor can undergo unstable behaviour. This includes the scenario observed in practice when the rotor is lightly loaded (i.e. at high speed).

To complement the numerical studies, an instrumented physical model of an autogyro was constructed for testing in a wind tunnel. The experiments were performed in the University of Bristol low-speed closed-return open-jet wind tunnel, with a 1.1 m jet diameter and maximum attainable velocity is about 33 m/s. Figure 1 depicts the rig in the tunnel.

The experimental rig [4] comprises a two-bladed teeter rotor of 1 m diameter, free to flap about a hinge located at the rotor shaft axis. An airframe similar to a production autogyro with an enclosed cockpit was constructed in order to cover the rotor support frame with an aerodynamically faired shape. The following measurements were taken from the rig: rotor blade flapping angle, pitch of each blade, rotor speed and azimuthal position and forces and moments acting on the rig. The signals measured on the rotating parts were transmitted to a computer outside the tunnel by wireless telemetry.

The rotor is modeled mathematically as a dynamical system in the form of a set of nonlinear ordinary differential equations. The rotor has two rigid blades and a 2-D individual blade element approach is used to model the aerodynamic loads on each. The blades are assumed to be rigidly connected and hence have one flapping degree of freedom. The lead-lag motion is captured in the rotational degree of freedom around the shaft axis; the flapping coordinate for the blades, β, is dependent on azimuth angle, ψ. The equations of motion for the rotor in both the flapping and the rotation senses are second order, giving a total of four rotor states (ψ, $\dot{\psi}$, β and $\dot{\beta}$); see for example [5].

The blades have no geometric twist or taper and the aerofoil profile is considered to be a NACA0015 section. Aerodynamic loads for each individual blade element are calculated numerically utilising nonlinear look-up tables for lift and drag coefficients of this aerofoil, defined from experimental data

for a 360° range of angle of attack [6]. The rotor inflow is captured via a 3-state Pitt–Peters dynamic wake model [7, 8], modified to account for the rotor being in autorotation. The total number of states (rotor and inflow dynamics) is seven. The parameters of interest here are the forward speed, V, and the longitudinal shaft angle, θ_{shaft}. Further details can be found in [4].

2 Experimental and Numerical Bifurcation Studies

The steady state periodic solutions and their stability are determined numerically from the 7-state mathematical model as parameters are varied, using the continuation and bifurcation software `Auto` [9]. Bifurcation diagrams generated from this model [3] have shown that unloaded rotors in autorotation are prone to instability. This is due to the branches of stable and unstable periodic orbits moving into closer proximity as the rotor shaft angle is reduced – which would be the case at higher speeds.

The first experimental runs in the present study entailed taking measurements at stable autorotative conditions over a range of wind speeds for different shaft angles, with other parameters fixed. Since the data is periodic with rotor azimuth position, average peak values for each cycle were computed over several runs. These peak values for both the rotational velocity and the flapping angle are plotted as filled circles in Fig. 2 for $\theta_{shaft} = 7°$; the solid curve fitted through these points is a stable limit cycle branch in this experimental bifurcation diagram. The multiple circles at each wind speed, especially for flapping angle, reveal the scatter in the readings from different runs.

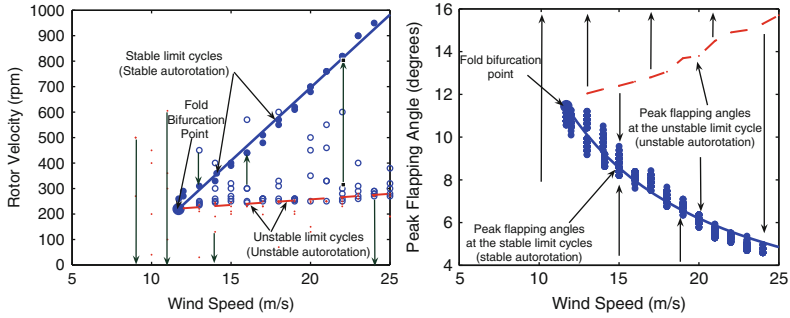


Fig. 2. Experimental bifurcation diagram of the rotational velocity (*left*) and the flapping angle (*right*) for $\theta_{shaft} = 7°$. *Solid circles* denote stable steady autorotation; the *solid curve* fitted to these points is the stable branch. *Hollow circles* represent initial points from which a stable steady autorotation state was achieved; *dots* show attempts where autorotation was not possible. The *dashed curve* represents the unstable autorotation steady state

Figure 2 illustrates that the rotational velocity increases almost linearly with the forward wind speed, while the flapping angle has an inverse relationship with the wind speed. The same trends are found for the other shaft angles tested. This shows that more flow goes through the rotor the faster it rotates, increasing the centrifugal stiffening and hence lowering the flapping angle. It is also clear that autorotation is not possible below a certain wind speed value (indicated as a fold bifurcation); to the left of this point, the rotor speed decays and the rotor flapping oscillation diverges until a safety mechanism is activated.

Next, the experiment was run to test the ability of the rotor to autorotate at wind speeds higher than the fold point but starting from low initial rotor velocities. For every shaft angle setting, rotor speed thresholds were found above which the rotor can achieve steady autorotation. The hollow circles in Fig. 2 illustrate attempts where a steady autorotation state was achieved from the initial condition (i.e. the trajectory evolved towards the solid circles). The dots are attempts leading to unstable rotation, characterised by diverging flapping and decaying rotational speed. By running a number of tests, it was possible to define a rotor speed boundary that separates the two scenarios: the dashed curve in Fig. 2. At this boundary, the behaviour of the rotor appears to be steady but due to flow disturbances it will diverge – either to the stable autorotation state or an unstable condition. This is interpreted as an unstable autorotation branch in this experimental bifurcation diagram.

To develop a deeper understanding of the nonlinear mechanisms underlying the rotor behaviour, we return to numerical continuation and bifurcation analysis of the mathematical model; this allows parameter dependence to be investigated beyond the limitations of the experiment (e.g. higher wind speed or flapping angles). Initially this did not yield a bifurcation diagram of the same shape as the experimental one, although stable and unstable branches were located. The experimental bifurcation diagram was therefore used to modify, or 'tune', the mathematical model with the aim of producing at least qualitatively the same behaviour. The numerical model uses a simple aerodynamic representation, neglecting characteristics such as blade tip and root losses, blade-to-blade interaction, unsteady aerodynamics, airframe and tunnel interaction effects, rotor downwash, etc. Also, friction acting on the rotor shaft is ignored. For simplicity, the model was adapted by incorporating a friction term in the rotation sense of the rotor to attempt to capture the effects of all unmodelled phenomena. The frictional term is formulated as a resisting torque assumed proportional to the rotational velocity; see [4]. The tuning of the coefficients in this term was performed to match rotor speed to experimental values.

Figure 3 depicts the bifurcation diagram obtained from the modified numerical model at various shaft angles. The shape of the stable autorotation branches are very similar to those obtained from experiment, although the fold bifurcation points are located at slightly higher wind speed values. These results illustrate how the essence of the stability of a rotor in autorotation can

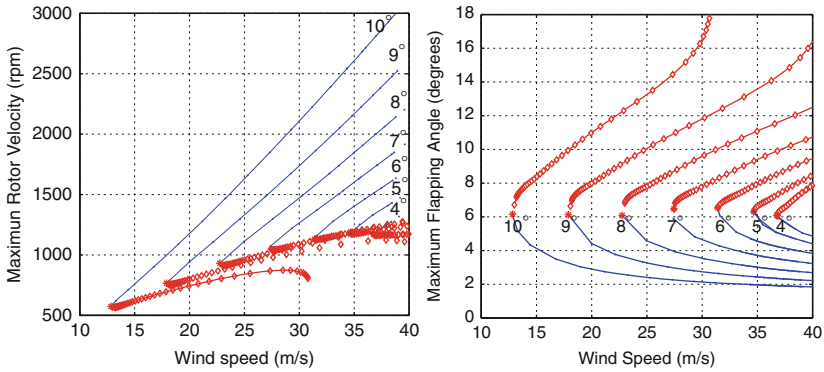


Fig. 3. Bifurcation diagram for the numerical rotor model at shaft angles of 4° to 10°. *Solid dotted curve* is stable; curve with *hollow diamonds* is unstable

be predicted to a high level, when continuation and bifurcation techniques are adapted, even though the numerical model is relatively simple. Furthermore, the figure shows that the wind speed value at which the fold bifurcation points exist increases as the shaft angle is reduced. Therefore, the boundary of rotor stability can be constructed by 2-parameter continuation (not shown), yielding the safe operating limits of the autogyro.

If the rotor flapping angles on both the stable and unstable branches are compared to those obtained from experiment, it is seen that their overall curve shapes are very similar. However, the amplitudes of the flapping angles computed are smaller than those of the physical rotor, particularly close to the bifurcation point. This is not unexpected since the model tuning was performed only for the rotor velocity state. The quantitative aspect of the analysis can be improved by incorporating a higher fidelity rotor model.

3 Conclusions

An example of a combined numerical-experimental approach to generating bifurcation diagrams has been shown to yield powerful information on the dynamics of a complex physical nonlinear system. The numerical results provide a qualitative framework to explain the experimental outcomes, which can in turn be used to validate the predictions and tune the model. In this case, an autorotating rotor was studied: a simple mathematical model was defined and then tuned using results from an experimental bifurcation diagram. Numerical continuation of this model showed, for the first time, the presence of both stable and unstable autorotation branches and their parameter dependence.

Results for other shaft angles (not shown) reveal that the smaller this angle is, the higher the wind speed below which autorotation cannot be sustained, i.e. the fold bifurcation occurs at higher forward speed. Thus a locus of fold

bifurcation points denotes the minimum permissible shaft angle required for autorotation at the corresponding wind speed: if the pilot reduces the shaft angle to decrease the lift coefficient for high speed flight, the rotor may enter an unstable condition where stable autorotation gives way to divergent flapping behaviour. In this way the bifurcation diagrams provide an explanation for high-speed autorotative instability as experienced in actual flight.

Future work will investigate following of unstable solutions directly in the experiment, as has already been achieved in a simpler nonlinear system [10].

Acknowledgments

This research was funded by *the Algerian Ministry for Higher Education and Scientific Research.*

References

1. Bedford, R., Lowenberg, M.: Flight dynamics analysis of periodically forced rotorcraft model. In: AIAA AFM Conference, number AIAA-2006-6634 (2006)
2. Rezgui, D., Bunniss, P.C., Lowenberg, M.H.: The stability of rotor blade flapping motion in autorotation using bifurcation and continuation analysis. Proceedings of 32nd European Rotorcraft Forum, 2006
3. Rezgui, D., Lowenberg, M.H., Bunniss, P.C.: Experimental and numerical analysis of the stability of an autogyro teetering rotor. Proceedings of the American Helicopter Society 64th Annual Forum, April 2008
4. Rezgui, D., Lowenberg, M., Bunniss, P.: Integrated experimental and numerical techniques in studying nonlinear rotor blade dynamics. In: ICNPAA Conference, 2008
5. Bramwell, A.R.S., Done, G., Balmford, D.: Bramwell's Helicopter Dynamics. Butterworth-Heinemann, Oxford (2001)
6. Sheldahl, R.E., Klimas, P.C.: Aerodynamic characteristics of seven airfoil sections through 180 degrees angle of attack for use in aerodynamic analysis of vertical axis wind turbines. SAND80-2114, Sandia National Laboratories, Albuquerque, New Mexico, March 1981
7. Chen, R.: A survey of nonuniform inflow models for rotorcraft flight dynamics and control applications. Vertica **14**(2), 147–184 (1990)
8. Peters, D., Morillo, J., Nelson, A.: New developments in dynamic wake modelling for dynamics applications. Proceedings of the American Helicopter Society 57th Annual Forum, May 2001
9. Doedel, E., Champneys, A., Fairgrieve, T., Kuznetsov, Y., Sandstede, B., Wang, X.: Auto97: Continuation and bifurcation software for ordinary differential equations. A.R.C. Technical Report C.P. No. 101 (14,757), http://indy.cs.concordia.ca/auto/, September 2007
10. Sieber, J., Gonzalez-Buelga, A., Neild, S., Wagg, D., Krauskopf, B.: Experimental continuation of periodic orbits through a fold. Phys. Rev. Lett. **100**(24), 244101 (2008)

Operational Parameter Study of an Aircraft Turning on the Ground

J. Rankin[1], B. Krauskopf[1], M. Lowenberg[2], and E. Coetzee[3]

[1] Engineering Mathematics, University of Bristol, Bristol, BS8 1TR, UK, j.rankin@bris.ac.uk, b.krauskopf@bris.ac.uk
[2] Aerospace Engineering, University of Bristol, Bristol, BS8 1TR, UK, m.lowenberg@bris.ac.uk
[3] Landing Gear Systems, Airbus, Bristol, BS99 7AR, UK, etienne.coetzee@airbus.com

Summary. Safety and economy are primary concerns in the study of ground manoeuvres for commercial aircraft, the ultimate goal being automation and optimisation of taxi operations. The application of mathematical and computer modeling to this problem is beneficial due to the relative costs compared with actual tests. As an example of utilising mathematical tools in the investigation of industrial problems we make use of a computer model of a passenger aircraft to perform a bifurcation analysis of turning solutions. In particular, we study how altering the longitudinal centre of gravity position of an aircraft affects its ground dynamics.

1 Introduction

During the daily service of passenger aircraft there are operational parameters that may vary considerably. Many of these parameters can have a significant effect on the ground handling properties of the aircraft. Important parameters include the loading of the aircraft in terms of passengers and fuel, runway and taxiway conditions, and wear on important components such as the tyres. In order to inform operational procedure it is important to understand how variation of these parameters affects the ground dynamics. Large costs associated with performing ground (flight) tests motivates the use of mathematical and computer modeling. In previous work a combination of flight test data and low-order computer bicycle models were used to study the ground handling properties of aircraft [1, 2], including the effect of tyre pressure on ground handling [3]. A previous study by the authors utilised a SimMechanics model to study the dynamics of aircraft on the ground under variation of thrust [4]. In this paper we use continuation analysis to perform a parameter study of a mathematical model of a passenger aircraft. Specifically, we investigate the effect that the aircraft's longitudinal centre of gravity position has on its ground handling.

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*,
Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_22,

During taxiing to and from the airport terminal a passenger aircraft will undertake various turning manoeuvres. Turns are made by applying a steering angle to the wheel and tyres of the nose gear while thrust from the engines remains constant. Our approach is to study the ensuing dynamics in terms of turning circle solutions of the system; their stability dictates whether a particular manoeuvre can be made safely. Depending on the loading of passengers, luggage and fuel levels, the centre of gravity position of the aircraft can vary considerably in day-to-day use. It is therefore of interest to treat the centre of gravity position along the longitudinal axis of the aircraft as a system parameter and to investigate changes in the turning dynamics under its variation.

We use a fully parametrised mathematical model of a typical medium sized single aisle passenger aircraft implemented in Matlab. The aircraft is modeled as a tricycle with the airframe having three translational and three rotational degrees of freedom. The equations of motion were obtained via balancing forces and moments in each degree of freedom. Nonlinear effects are included in the tyre model, depending on tyre load and slip angle, and in the aerodynamic model, depending on velocity, angle of attack and slip angle of the airframe. The steering angle δ and the centre of gravity position CG are the free parameters in our analysis. The centre of gravity position is measured as the percentage along the mean aerodynamic chord (MAC), taken from the leading edge; negative values represent a position in front of the leading edge.

The tool used here is numerical continuation; specifically, we perform a bifurcation analysis with the software package AUTO [5]. Continuation analysis is a powerful tool used to study steady-state solutions of dynamical systems [6], which are tracked under the variation of system parameters; during computations solutions are monitored to detect bifurcations, which are qualitative changes in the dynamics [7,8]. Identifying where bifurcations occur is important because they may form boundaries of safe behaviour. The use of continuation and bifurcation analysis to study ground manoeuvres is a computationally inexpensive way of analyzing the dynamics under variation of several parameters.

2 Bifurcation Analysis of Turning Solutions

We present a bifurcation analysis of aircraft turning solutions; the results are represented as one-parameter and two-parameter bifurcation diagrams. In our model fixed-radius turning circles correspond to steady-states of the system. The analysis focuses on how (steady-state) turning circle solutions change under variation of parameters. In the one-parameter study the CG position is kept fixed and the steering angle δ is varied; solutions are plotted against a state variable. In the two-parameter study we also vary CG and the results are represented as a surface of solutions that describe the dynamics over the entire range of δ and CG.

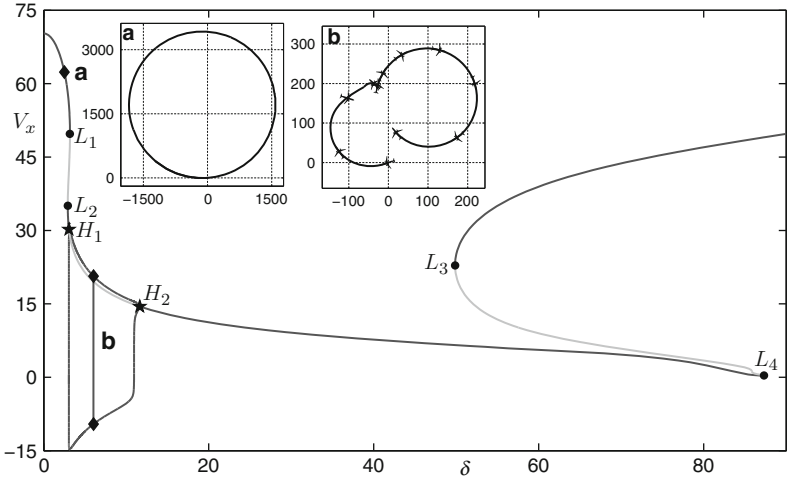


Fig. 1. One-parameter bifurcation diagram in δ for $CG = 35\%$ with a single branch of solutions; stable parts are *black* and unstable parts are *grey*. Changes in stability occur at the bifurcation points L_{1-4} and H_{1-2}. The maximum and minimum forward velocity of a branch of periodic solutions between H_1 and H_2 are plotted in *black*. *Insets* (**a**) and (**b**) show examples of the aircraft's motion

2.1 One-Parameter Study

Figure 1 shows a one-parameter bifurcation diagram in δ for $CG = 35\%$, where the forward velocity V_x of the aircraft is used as a measure of the solution. A single branch of solutions initiates in the top left of the figure and terminates in the top right; changes of stability occur at the limit point bifurcations L_{1-4} and Hopf bifurcations H_{1-2}. On the branch of solutions, stable parts are black and unstable parts are grey. Periodic solutions exist between H_1 and H_2 and their maximum and minimum velocities are plotted as black curves. Qualitatively different types of behaviour can be observed at the labeled points (a) and (b). The respective insets in Fig. 1 show a top down view of a CG-trace of the aircraft in the horizontal ground plane; in (b) markers are drawn to scale and show the aircraft's attitude along the CG-trace.

At the initial point where $\delta = 0$, the aircraft travels in a straight line with a constant velocity of $V_x = 70$ m/s due to constant thrust from the engines. As steering is applied ($\delta > 0$), the solutions represent fixed large radius turns. For example, at (a) the aircraft follows a stable turning circle of radius $r \approx 1.7$ km with a forward velocity of 63 m/s. This type of solution with a small steering angle, large radius turn persists from the initial point up to the bifurcation L_1; the radius of the turn decreases as L_1 is approached. At the bifurcation the turning moment generated by the nose gear tyres overcomes the stabilising aerodynamic force generated by the tail fin of the aircraft [4].

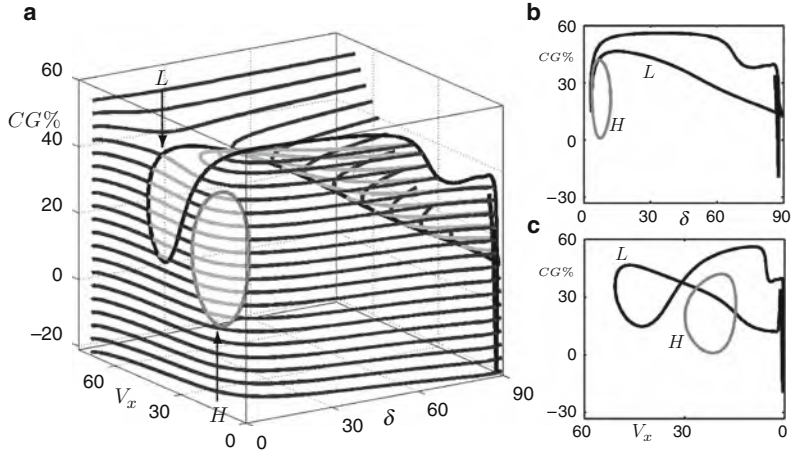


Fig. 2. Panel (**a**) shows a surface plot of solutions in (δ, V_x, CG)-space; stable solutions are *black* and unstable solutions are *grey*. The loci of limit point bifurcations L is the *thick black curve* and the locus of Hopf bifurcations H is the *thick grey curve*. Panels (**b**) and (**c**) show two-dimensional projections of the bifurcation curves onto the (δ, CG)-plane and (V_x, CG)-plane, respectively

When the steering angle is increased beyond L_1, the aircraft loses velocity rapidly over a transient period and starts to follow a solution in the region between the Hopf bifurcations H_1 and H_2. Hopf bifurcations are associated with the onset of periodic motion [8]. In this case, passing a Hopf bifurcation represents a change in which the aircraft attempts to follow a turning circle that is too tight and, therefore, there is a loss of lateral stability associated with the main landing gear tyres saturating. For example, at (b) the aircraft attempts to follow an unstable turning solution with radius $r \approx 125$ m but loses lateral stability, enters a spin and briefly travels backwards before coming to a halt. The aircraft then moves off under constant thrust, repeating the motion periodically relative to the unstable turning solution with a maximum and minimum velocity of 20 and -10 m/s, respectively. A detailed description of this undesirable behaviour that persists between H_1 and H_2 is given in [4]. Between H_2 and L_4 high steering angle, small radius turns can be observed, and between L_3 and the end point at the top right high steering angle, large radius turns can be observed for which the nose gear is almost perpendicular to the direction of motion and, hence, is effectively dragged along the ground.

2.2 Two-Parameter Bifurcation Study

One-parameter continuation runs, as in Sect. 2.1, were computed over a range of CG at discrete points. When plotted together in (δ, V_x, CG)-space the individual bifurcation curves form a surface of solutions. Two-parameter continuation was used to compute the loci of bifurcations continuously under the

variation of both δ and CG. Combining the results from these two computations into a single plot is an effective way of representing the behaviour over the complete range of δ and CG in a single figure. Two-dimensional projections of bifurcation curves show certain features more clearly.

Figure 2a shows the resulting surface plot of solutions in (δ, V_x, CG)-space; again stable solutions are black and unstable solutions are grey. Changes in stability occur at bifurcation curves on the surface. The curve L of limit point bifurcations is represented by the thick black closed curves and the curve H of Hopf bifurcations by the thick grey closed curve. The one-parameter case discussed above represents a horizontal slice of Fig. 2a at $CG = 35\%$. The bifurcations in Fig. 1 lie on the locus curves in Fig. 2a, L_1, L_2, L_3 and L_4 on L, and H_1 and H_2 on H.

Figures 2b and 2c show two-dimensional projections of the bifurcation curves onto the (δ, CG)-parameter plane and the (V_x, CG)-plane, respectively. In the (δ, CG)-parameter plane bifurcation curves bound regions with different numbers of solutions, each with a specific stability. In the largest region, not bounded by any of the bifurcation curves, a single stable turning circle solution exists. In the region bounded by the Hopf bifurcation curve H a single unstable turning circle solution exists and the attracting solution is a periodic motion relative to this unstable turning circle, as was discussed in Sect. 2.1. In the region bounded by the limit point bifurcation curve L two stable and one unstable turning circle solutions exist. Figure 1 provides an example of traversing each region in the parameter δ. A hysteresis loop results when traversing the regions bounded by limit point curves in different directions. The same data plotted in the (V_x, CG)-plane reveals the relative positions of the bifurcation curves in terms of the forward velocity V_x.

Within the operational range of $CG \in (10\%, 40\%)$, the laterally unstable behaviour inside the region bounded by H in Fig. 2 persists. However, for $CG < 15\%$ (a forward position) no limit point bifurcations will be observed at low steering angles as seen clearly in Fig. 2b. This means that the region of laterally unstable dynamics could be approached more suddenly and at lower velocities. Taking values of CG outside of the operational range (an extreme forward or aft position) results in uniformly stable behaviour at low steering angles, where intersections with L and H are not possible. In Fig. 2b there is a region for small $\delta < 3^\circ$ to the left of L and H for which no bifurcations occur. This bound does not change under variation of CG and could provide a limit for steering angles used in high-velocity turns.

3 Conclusions

A comprehensive bifurcation analysis of a mathematical model of a typical single aisle passenger aircraft was performed in terms of the steering angle and the aircraft's longitudinal centre of gravity (CG) position. A one-parameter study in the steering angle illustrated different types of solutions and their

bifurcations. These results were extended to a two-parameter study by computing solution branches over a range of CG positions and tracking the loci of the bifurcations continuously in the parameter plane. Combining the results gives a complete account of the possible turning dynamics of the aircraft under variation of both parameters.

The results presented here reveal how changing an aircraft's CG position can affect its ground dynamics. Over the operational range of the CG position there is a region of laterally unstable dynamics existing between two Hopf bifurcations. Depending on the CG position, this unsafe region of dynamics can be approached in different ways at small steering angles. With an aft position the region can be approached at high velocity by passing a limit point bifurcation, but with a forward position the solutions can be approached more suddenly at a lower velocity by passing one of the Hopf bifurcations. Additionally, a steering angle of 3° was identified as an upper bound independent of CG position for making stable high-velocity turns.

Ongoing work focuses on the sensitivity of the results presented here to variation of the additional parameters, for example, the mass and thrust of the aircraft. However, there are many other parameters that are of interest, including the track-width of the main landing gears, runway conditions and tyre properties. Physical phenomena associated with changes in qualitative dynamics are also the subject of ongoing studies.

Acknowledgments

This research was supported by an Engineering and Physical Sciences Research Council (EPSRC) Case Award grant in collaboration with Airbus in the UK.

References

1. Klyde, D., Myers, T., Magdaleno, R., Reinsberg, J.: J Guid. Control Dyn. **25**(3), 546–552 (2002)
2. Klyde, D., Sanders, E., Reinsberg, J., Kokolios, A.: J. Guid. Control Dyn. **27**(1), 41–51 (2004)
3. Klyde, D., Magdaleno, R., Reinsberg, J.: J. Guid. Control Dyn. **26**(4), 558–564 (2003)
4. Rankin, J., Coetzee, E., Krauskopf, B., Lowenberg, M.: J. Guid. Control Dyn. **32**(2), 500–511 (2009)
5. Doedel, E., Champneys, A., Fairgrieve, T., Kuznetsov, Y., Sandstede, B., Wang, X.: Auto 97. http://indy.cs.concordia.ca/auto/, May 2001
6. Krauskopf, B., Osinga, H.M., Galán-Vioque, J.E. (Eds.): Numerical Continuation Methods for Dynamical Systems. Springer, Dordrecht (2007)
7. Guckenheimer, J., Holmes, P.: Nonlinear Oscillations, Dynamical Systems and Bifurcations of Vector Fields, Applied Mathematical Sciences, vol. 42. Springer, New York (1983)
8. Strogatz, S.: Nonlinear Dynamics and Chaos. Springer, New York (2000)

Geometric Nonlinearities of Aircraft Systems

B. Krauskopf[1], P. Thota[1], and M. Lowenberg[2]

[1] Engineering Mathematics, University of Bristol, Bristol, BS8 1TR, UK
B.Krauskopf@bristol.ac.uk, Phani.Thota@bristol.ac.uk
[2] Aerospace Engineering, University of Bristol, Bristol, BS8 1TR, UK
M.Lowenberg@bristol.ac.uk

Summary. Nonlinearities due to geometric effects, in particular, via angular variables that are not small, are important for aircraft operation. Geometric nonlinearities have a strong effect on the dynamics of the aircraft system under consideration, and they are especially pronounced in aircraft ground operations. As a concrete example we consider here the effect of a non-zero rake angle on the dynamics of a nose landing gear. More specifically, we use tools from bifurcation theory to investigate the stability of the straight-rolling motion during a take-off run.

1 Introduction

Many systems of an aircraft operate in such a way that nonlinearities need to be taken into account to describe their dynamics correctly. Sources of nonlinearities include nonlinear properties of individual components (for example, tyres and dampers), range limits of control surfaces and, in particular, geometric nonlinearities due to the fact that angular variables are not small. As a specific example, we consider the role that geometric nonlinearities play in the phenomenon of shimmy oscillations in aircraft landing gears during high-speed straight-line rolling. Due to their implications for passenger comfort, safety and maintenance costs, shimmy oscillations are an unwanted type of dynamics. They may occur in any wheeled vehicle, including cars, pulled trailers, motorcycles and indeed aircraft; see the overview papers [1–3].

We consider here shimmy oscillations of the nose landing gear of a mid-size passenger aircraft, as sketched in Fig. 1. A nose landing gear consists of a strut, attached to the aircraft body, to which a wheel is mounted with an offset from the strut axis, called the caster length. The system's dynamics are dominated by the interplay between the two basic modes [4]: the torsional mode of rotation around the strut axis, and the lateral mode of deflection of the entire gear from side to side. These two modes are coupled via the nonlinear interaction of the elastic tyre with the ground. The overall landing gear system is characterised by geometric nonlinearities, because the torsional

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*,
Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_23,

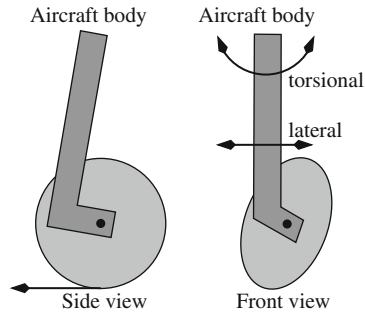


Fig. 1. A non-zero rake angle (of the strut with the vertical) of an aircraft nose landing gear results in a tilt of the tyre plane; the two main modes are the torsional mode of rotation around the steering axis, and the lateral mode of sideways motion of the gear around its attachment point

and the lateral mode may show dynamics of considerable amplitude during shimmy oscillations. An important feature of an aircraft nose landing gear is the presence of a non-zero rake angle of the steering axis with the vertical, typically in the range of 0°–10°. A positive rake angle introduces additional geometric nonlinearities into the problem. First of all, it contributes to an overall effective caster length, which in turn enters the coupling between the two modes. Furthermore, steering results in a tilt of the wheel, meaning that the wheel plane is not perpendicular to the ground; see the front view in Fig. 1.

We model the nose landing gear by equations for the torsional mode ψ, the lateral mode δ and the lateral deformation λ of the tyre (for which we use the well-established stretched string model [5]). Overall we obtain a mathematical model in the form of five coupled nonlinear ordinary first-order differential equations. The model depends on a number of parameters, including the dimensions of the landing gear, stiffnesses and dampings of the two modes and parameters specifying the tyre forces. The values of these parameters were chosen to represent a midsize passenger aircraft (with a rake angle of 9°); see [6] for details of the model and the specific values of the modelling parameters.

2 Bifurcation Analysis of Shimmy Oscillations

The landing gear moves at horizontal velocity V, subject to a vertical force F_z that is exerted by the aircraft body (which is modelled as a block of mass). It is therefore natural to study the dynamics of the nose landing gear in dependence on the operational parameters V and F_z. Figure 2 shows how the operational range of the (V, F_z)-plane is divided into regions of qualitatively different

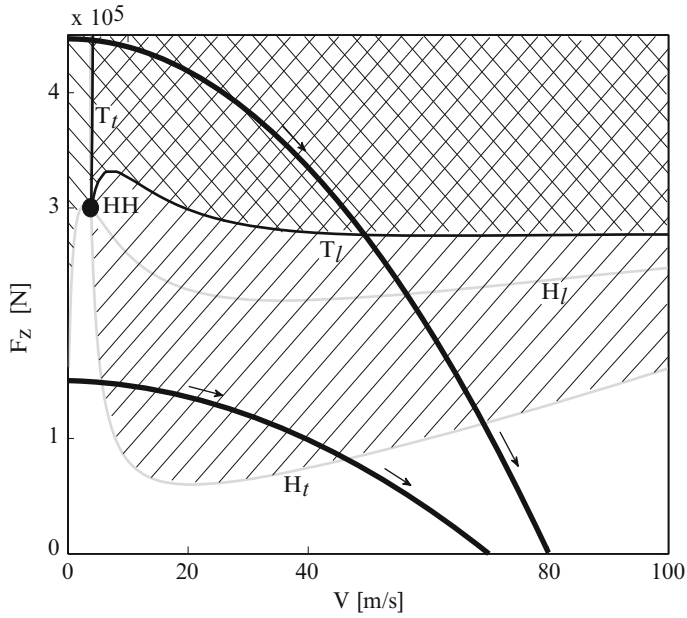


Fig. 2. Two-parameter bifurcation diagram in the (V, F_z)-plane, consisting of curves H_t and H_l of Hopf bifurcations (*grey*) and curves T_t and T_l of torus bifurcation (*black*). The straight-rolling solution is stable in the white region; torsional and lateral shimmy oscillations occur in the regions of right-slanted and left-slanted shading, respectively. The two *thick black curves* are two simulated take-off runs, of a light and a heavy aircraft, respectively

dynamics. The boundaries between regions are given by curves of bifurcations, which have been computed with the continuation software AUTO [7].

In the white region in Fig. 2 the straight-rolling motion is stable, that is, the nose landing gear does not show shimmy oscillations. Stability is lost when one of two Hopf bifurcation curves, H_t or H_l, is crossed. Specifically, crossing H_t corresponds to an undamping of the torsional mode. The ensuing torsional shimmy oscillations are stable in the region of right-slanted shading and they are characterised by oscillations of the landing gear around the strut axis. By contrast, crossing H_l corresponds to an undamping of the lateral mode, meaning that the gear shows lateral shimmy oscillations in the plane perpendicular to the direction of travel. This type of shimmy is stable in the region of left-slanted shading. The curves H_t and H_l intersect at a double-Hopf point HH, which gives rise to two curves, T_t and T_l, of torus (or Neimark-Sacker) bifurcations [8]. Crossing these two curves corresponds to the undamping of the second mode, which gives rise to the creation of an invariant torus. We find that the bifurcating torus is unstable throughout. As a result the curves T_t and T_l bound a large region where torsional and lateral shimmy oscillations

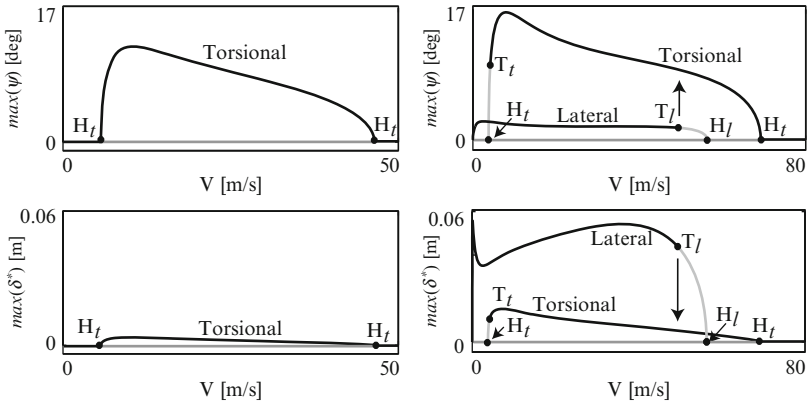


Fig. 3. One-parameter continuations along the two simulated take-off runs of a light (*left column*) and a heavy (*right column*) aircraft. The top panels show the maximum of the torsion angle ψ and the bottom panels the maximum of the lateral bending stroke δ^*; stable parts of branches are *black* and unstable parts *grey*

are both stable. In this region of bistability, it depends on the initial condition which type of shimmy the landing gear performs.

3 Shimmy Dynamics During Take-Off

Figure 2 gives a comprehensive picture of the behaviour of the aircraft over the relevant ranges of forward velocity V and downward force F_z. Each point in the (V, F_z)-plane corresponds to a type of dynamics and it typically lies in one of the regions that were identified. Hence, the bifurcation diagram in Fig. 2 illustrates the robustness of a typical choice of V and F_z with respect to small changes of their values. On the other hand, larger changes that result in a crossing of bifurcation curves lead to qualitative changes of the behaviour of the system.

To demonstrate how the information in Fig. 2 can be used in practice we consider the dynamics of the nose landing gear during take-off. During a take-off run the aircraft accelerates from zero velocity to its take-off speed, during which the vertical force F_z on the nose landing gear decreases from its maximal (static) value to zero. Hence, a take-off run corresponds to a one-dimensional curve in the (V, F_z)-plane. Two examples of take-off runs (chosen to feature shimmy oscillations), one for a light and one for a heavy aircraft, are shown as bold black curves in Fig. 2. Owing to the quadratic dependence of lift on velocity, they have been modelled as parabolas. One immediately notices that the two take-off runs are qualitatively different, because they intersect different regions of the (V, F_z)-plane. Notice further that the exact shape of these curves is not crucial, as long as the same regions are encountered in the same order.

Figure 3 shows the results of two one-parameter continuations with AUTO along the two take-off runs. Shown are the amplitudes of the torsion angle ψ and of the lateral bending stroke δ^* (the lateral stroke of the strut at ground level). The take-off run for the light aircraft case, shown in the left column of Fig. 3, starts at $F_z = 150\,\mathrm{kN}$ and ends at a take-off speed of 70 m/s. The straight-rolling motion is stable, but then loses stability when the curve H_t is crossed in Fig. 2. The amplitude of the ensuing torsional shimmy oscillations increases rapidly up to a maximum of about 14°. It then decreases as the aircraft accelerates. Finally, at about 45 m/s the straight-rolling motion regains stability and the torsional shimmy oscillations disappear. Notice that the lateral bending stroke δ^* shows small amplitude oscillations during torsional shimmy; namely, it follows the torsional mode passively due to the coupling via the tyre [6].

The take-off run for the heavy aircraft case is shown in the right column of Fig. 3; it starts as a vertical force of $F_z = 450\,\mathrm{kN}$ and ends at a take-off speed of 80 m/s. This take-off run is such that the straight-rolling motion is unstable from the very beginning. Instead at low speeds the nose landing gear performs lateral shimmy oscillations with a lateral stroke amplitude of around 5 cm; again due to the coupling via the tyre, the torsional mode follows this motion with small amplitude. The lateral shimmy oscillations are stable until the curves T_l is encountered in Fig. 2 at a velocity of about 50 m/s. This curve marks the boundary of the bistable region and the system switches to the branch of torsional shimmy oscillations, as is indicated by the arrows in Fig. 3 (right column). The torsional shimmy oscillations gradually decrease and finally disappear at around 70 m/s just prior to take-off. We remark that during the switching from lateral to torsional shimmy oscillations one may encounter quasiperiodic (two-frequency) shimmy oscillations as long transients; see [6] for more details.

4 Conclusions

We presented a study of aircraft nose gear shimmy as an example of how geometric nonlinearities influence the dynamics of aircraft systems. Specifically, we performed a bifurcation analysis of a mathematical model that describes the interaction of the torsional and lateral modes via the elastic tyre. Geometric nonlinearities arise from the fact that the amplitudes of the torsion angle and the bending stroke may be substantial – an effect that is further enhanced by the geometric nature of the coupling between the two modes via a non-zero rake angle. Torsional and lateral shimmy oscillations occur in large regions in the plane of velocity versus vertical force, including in a region of bistability. One-parameter continuations along take-off runs for a light and a heavy aircraft demonstrated how shimmy oscillations are encountered in practice when the different regions are crossed.

An obvious question is how the bifurcation diagram presented here depends on the parameters that specify the landing gear, especially on those that have a bearing on geometric nonlinearities. The influence of the rake angle has been considered in [6], where it was found that the region of torsional shimmy oscillations shrinks with an increase of the rake angle. Our present work focuses on the dependence of the bifurcation structure on other parameters, those that determine the geometry of the nose landing gear as well as those that specify tyre properties. The study of additional effects, for example, dynamics of vertical shock absorbers in the presence of a rough runway, can be addressed via an expansion of the model of the nose landing gear. Furthermore, we also intend to model and study the dynamics of main landing gears of different geometries (with different numbers of wheels). In the longer term, our goal is to couple the dynamics of individual gears via a flexible fuselage to obtain a realistic, yet tractable model to describe aircraft ground dynamics.

Acknowledgments

This research has been supported by Airbus in the UK.

References

1. Dengler, M., Goland, M., Herrman, G.: A bibliographic survey of automobile and aircraft wheel shimmy. Technical report, Midwest Research Institute, Kansas city, MO, USA, (1951)
2. Pritchard, I.J.: An overview of landing gear dynamics. NASA Technical Reports, NASA/TM-1999-209143, (1999)
3. Smiley, R. F.: Correlation, evaluation, and extension of linearized theories for tyre motion and wheel shimmy. Report submitted to the National Advisory Committee for Aeronautics, Report 1299, (1957)
4. Thota, P., Krauskopf, B., Lowenberg, M.: Modeling of nose landing gear shimmy with lateral and longitudinal bending and a non-zero rake angle. Proceedings of AIAA 2008. (2008)
5. B. von Schlippe and Dietrich, R.: Shimmying of a pneumatic wheel. Report submitted to the National Advisory Committee for Aeronautics, NACA TM 1365, (1947)
6. Thota, P., Krauskopf, B., Lowenberg, M.: Interaction of torsion and lateral bending in aircraft nose landing gear shimmy. Nonlinear Dyn. **57**(3), 455–467 (2009)
7. Doedel, E., Champneys, A., Fairgrieve, T., Kuznetsov, Y., Sandstede, B., Wang, X.: Auto 97. http://indy.cs.concordia.ca/auto/, May 2001
8. Guckenheimer, J., Holmes, P.: Nonlinear Oscillations, Dynamical Systems and Bifurcations of Vector Fields. Springer, New York (1983)

Application of Nonlinear Dynamics in Civil Aerospace

E. Coetzee

Airbus, Landing Gear Systems, Bristol, BS99 7AR, UK,
etienne.coetzee@bristol.ac.uk

Summary. Nonlinear analysis techniques, especially methods from bifurcation theory, have emerged as valuable tools over the last 20 years, particularly due to the advent of the modern computer. Originally developed as part of dynamical systems and chaos theory, they gradually are finding their way into applications areas from all walks of life. As far as the aerospace industry is concerned, methods from nonlinear dynamics were used initially for the prediction of aircraft flight dynamics at high angle of attack flight regimes, where traditional methods have failed. They are now being used within Airbus to analyse aspects of the dynamics of aircraft on the ground. Specific aerospace applications, where nonlinear dynamics techniques are expected to make an impact, include the design of flexible structures and mechanisms, and the dynamics of a braking wheel. Challenges related to the industrialisation of such methods are also discussed.

1 Introduction

Landing gear engineers observe nonlinear phenomena such as hysteresis, backlash and stiction on a daily basis, without necessarily appreciating the full meaning behind these observations. A wheel that locks up during braking is a good example. Many conflicting requirements need to be considered during the design, where the weight and pavement loading needs to be minimised, and the shock absorption maximised. The lateral stability on the ground is determined by the position of the gears, along with the tyre and oleo (shock damper) characteristics. Experience has shown that the use of different tyres can mean the difference between a stable and an unstable aircraft. Landing gears contain highly nonlinear components, including tyres, brakes and oleos, and therefore traditional analysis is usually done at some very specific design conditions. There is a perceived need to characterise the behaviour of the system over a wide variety of parameters, and this is the industrial domain where methods from nonlinear dynamics can and should be brought to bear. We discuss here some of the open avenues for this approach within the specific context of ground dynamics of passenger aircraft.

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*,
Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_24,

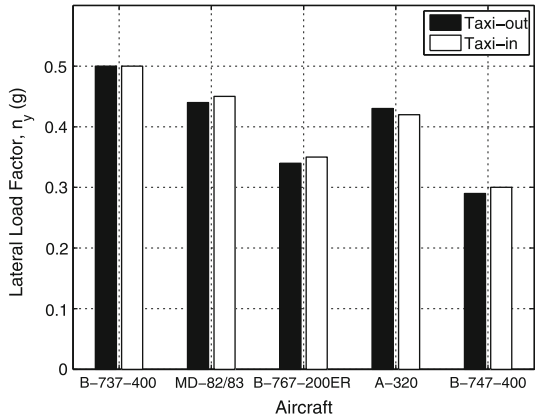


Fig. 1. Comparison of equal probability lateral load factors during ground turning for five aircraft after [1]

2 Aircraft Ground Manoeuvres

Ground operations tend to be performed at constant thrust settings, because the thrust is adjusted only occasionally by the pilot with the aim of altering the velocity. One issue is to find points (in terms of operational input) where the aircraft becomes uncontrollable during a turn. A loss of stability is dependant on several parameters, such as the steering angle, entry velocity of the turn, tyre properties, and the runway condition. Mathematically, stability loss corresponds to a limit point (or fold) bifurcation or a Hopf bifurcation, which makes it possible to classify the dynamics of a turning aircraft on the ground with the use of continuation methods. In this way, physical causes for the loss of stability have been identified [2–4]. Specifically, limit points and Hopf bifurcations bound regions in parameter space where the tyres are saturated, so that they cannot provide enough side force to maintain a specific manoeuvre.

An ongoing study by the FAA has been aiming to identify what type of lateral loading conditions can be experienced by in-service aircraft. The goal is to validate the conservative design factors that are currently required during the design phase. Current regulations require an 0.5 *g-level* at the centre of gravity, even though it is known from experience that such high *g-levels* are not possible in larger aircraft. The results from the study indicates that the actual *g-levels* experienced by airline operators are approximately 0.3 g for wide-body aircraft, such as the Boeing 747, and 0.43 g for narrow-body aircraft, such as the Airbus A320; Fig. 1 shows a summary of the expected peak *g-levels* as extracted from the report [1]. It would be of great benefit if the influence of the main parameters could be studied during the preliminary design phases of a project, where some analysis is indeed already done by means of detailed nonlinear simulations. Our experience with the bifurcation study of ground

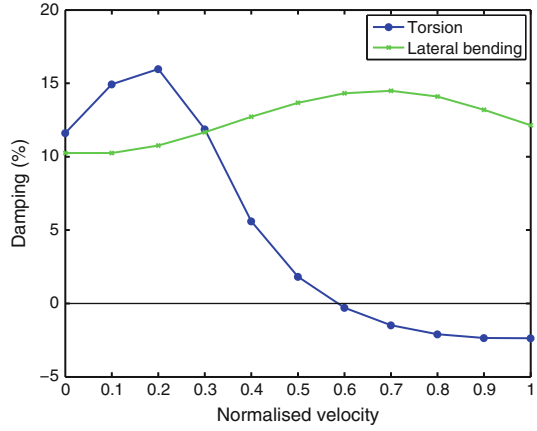


Fig. 2. An example of linear shimmy analysis

manoeuvres indicates that continuation methods may be used as a new tool to provide a reasonable estimate of the maximum *g-levels*, as well as where such operating conditions will occur.

3 Landing Gear Shimmy

Shimmy oscillations of a landing gear are undesirable due to the safety and maintenance aspects involved with the occurrence of this phenomena. Linear shimmy analysis is typically done at specific operating points for design purposes, while detailed nonlinear simulations are usually performed only after an incident occurred. Torsional and/or lateral motion can be observed during shimmy oscillations, and the contribution of each mode may be dependant on the initial conditions of the system. Linear shimmy methods calculate the damping in the system while the velocity is varied to identify the onset of shimmy as a point where either the torsional mode or the lateral mode has zero damping. Figure 2 shows an example of such an analysis.

Pilots often report the onset and disappearance of shimmy oscillations between certain velocities, indicating a trajectory across a boundary of Hopf-bifurcations. There are still many differing opinions with regards to the main parameters that influence shimmy, and they result in differing maintenance actions that are recommended when shimmy occurs. Hydraulic shimmy dampers are installed on some aircraft to prevent oscillations in the steering system, but this adds weight.

The development of a nonlinear model of a nose landing gear, and its subsequent bifurcation analysis, has demonstrated the coupled nature of the torsional and the lateral modes via nonlinear tyre forces in the presence of geometric nonlinearities [5, 6]. Future research will focus on the construction

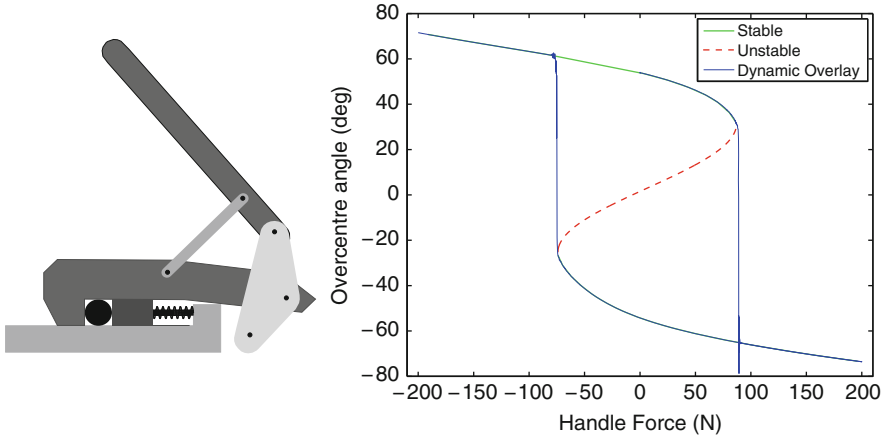


Fig. 3. A mechanism with a hysteresis loop of its force-reaction diagram

of a comprehensive map of all the types of shimmy under different operational conditions, as well as the development of preliminary rules to avoid shimmy already at the design phase of an aircraft.

4 Dynamics of a Braking Wheel

The longitudinal traction force of a braked wheel is a consequence of the relative difference between the vehicle velocity and the velocity of the wheel at the contact patch, which is also known as wheel slip [7]. It depends on the normal force on the wheel, as well as the friction coefficient between the wheel and the road surface. A free-rolling wheel is defined to have a slip-value of 0, while a locked wheel has a slip-value of 1 [8]. It is known that a hysteresis loop exists when a brake torque is applied [9]. This means that the brake torque where lockup occurs and where control is regained could be very different. Recent research on aircraft has also shown that the unstable point after which lockup occurs, does not necessarily occur at the peak value on the slip curve. In fact, braking is one of the most nonlinear processes in aircraft, and understanding it fully will require the use of advanced methods from dynamical systems theory.

5 Landing Gear Mechanisms

A mechanism is defined as a combination of parts, that are joined in a specific way, to perform a certain function. Figure 3 shows an example of a latch that contains several pinned arms and a spring. A relatively small force can be applied to the handle of the latch, yet the clamping force on a component could

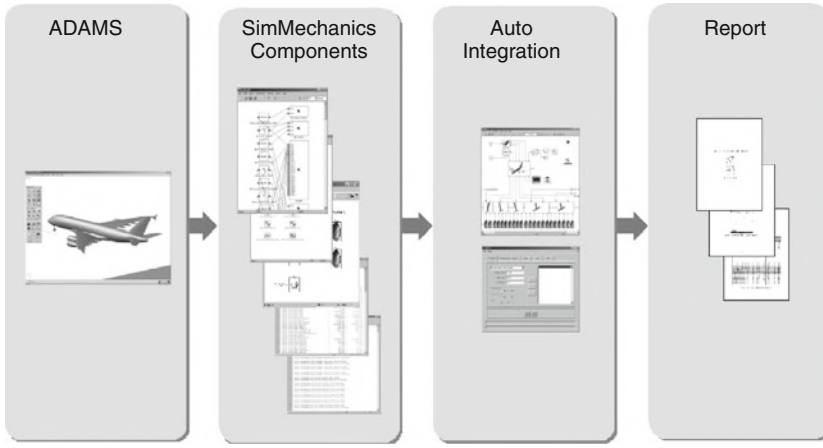


Fig. 4. Suggested integrated environment for the nonlinear analysis of linked models of aircraft components

become significant. A point could also be reached when the handle "jumps" to a new position where no additional force is needed to hold the part in place. This jump indicates the presence of a fold bifurcation as shown in Fig. 3. The envelope of where this fold occurs can be calculated by varying the spring stiffness and applied force. A landing gear effectively is a mechanism quite similar to a latch. Importantly, the landing gear needs to reach a downlock solution at a certain applied force. Nonlinear dynamics methods are being used in ongoing research to map out the envelope of downlock solutions of different types of landing gears as a function of gear spring stiffness and applied force values.

6 Conclusions and Outlook

Several case studies have clearly demonstrated that methods from nonlinear dynamics allow engineers to discover, and explain, the rich dynamical behaviour that is observed during aircraft operations on a daily basis. Traditional linear methods are adequate for many engineering systems, but nonlinear effects need to be considered if a system is to be used to its full potential.

In spite of their huge potential, bifurcation theory methods are presently being used only by small pockets of engineers in the aviation industry. In fact, when one wants to introduce nonlinear dynamics into the engineers' normal toolsets one encounters both societal and technological challenges. Primarily, the societal ones relate to management support and education. The technology needs to be supported by all tiers of management, and a strong business case needs to be made to gain this support. The technological challenge is

one of education and development of the right tools. Training is needed to familiarise engineers with the vocabulary and tools of dynamical systems theory, which are still largely unknown to the average engineer. Indeed, there is a need to learn how to formulate a problem in a way conducive to nonlinear analyses, and how to interpret the results. A level of intuition similar to that concerning, say, Bode diagrams, needs to be developed for the interpretation of bifurcation diagrams. At the same time more emphasis should be placed on the development of well-documented, industrial, integrated toolsets for nonlinear analyses. Whilst several software tools are freely available, they were developed primarily for research purposes. The overall goal is to develop an integrated and user-friendly environment where validated models can be studied with bifurcation software. Figure 4 shows an example of what such an environment may look like at a high level.

References

1. Tipps, D., Rustenburg, J., Skinn, D., DeFiore, T.: Side load factor statistics from commercial aircraft ground operations. Technical Report UDR-TR 2002-00119, Federal Aviation Administration, U.S. Department of Transportation, Federal Aviation Administration, Office of Aviation Research, Washington, DC 20591, January 2003
2. Coetzee, E., Krauskopf, B., Lowenberg, M.: Nonlinear aircraft ground dynamics. In ICNPAA, editor, International Conference on Nonlinear Problems in Aviation and Aerospace, June 2006
3. Rankin, J., Coetzee, E., Krauskopf, B., Lowenberg, M.: AIAA Modeling and Simulation Technologies Conference, (AIAA-2008-6529) (2008)
4. Rankin, J., Coetzee, E., Krauskopf, B., Lowenberg, M.: Operational parameter study of an aircraft turning on the ground. Prog. Ind. math. (2010). doi:10.1007/978-3-642-12110-4
5. Thota, P., Krauskopf, B., Lowenberg, M.: Proceedings of AIAA 2008, (2008)
6. Thota, P., Krauskopf, B., Lowenberg, M.: Geometric nonlinearities of aircraft systems. Prog. Ind. math. (2010). doi:10.1007/978-3-642-12110-4
7. Wong, J.: Theory of Ground Vehicles, 3rd edn. Wiley-Interscience, New York (2001)
8. Blundell, M., Harty, D.: The Multibody Systems Approach to Vehicle Dynamics. SAE, Warrendale, PA (2004)
9. Olson, B.: Nonlinear dynamics of longitudinal ground vehicle traction. Master's thesis, Michigan State University, 2001

Minisymposium *Global System Dynamics and Policies*

Steven Bishop

University College London, Gower St, London WC1E 6BT, UK

This Mini-Symposium highlighted some of the best ways in which global system dynamics can assist policy makers in industry and government through powerful applications combining many disciplines taken from physical, natural and social sciences. This need for a multi-disciplinary approach has recently been recognised by the European Commission by the funding of a Coordinated Action award called GSD (see www.globalsystemdynamics.eu).

The event was opened by Ralph Dum (GSD's EU Scientific Officer). He explained that there was a considerable interest in seeing how a complex systems approach could be used to improve our understanding when it comes to setting policy.

Under the title Visualising Europe's Future, Jacquie McGlade (Executive Director of the European Environment Agency, EEA) gave an overview of the EEA's findings over recent years. She stated that science needs to provide clear evidence-based hypotheses on how we can tackle some of the local, and increasingly global, challenges. Actions have only just started, and better data and improved methods for data collection are required to monitor effectiveness, which will also help us to account for the respective costs of any such actions. One area where the EEA is at the forefront is monitoring urban development. Better data means that decision makers have more information to inform policy. Visualisation must be used to aid our understanding of the spatial planning throughout Europe.

Julian Hunt (UCL/UK's House of Lords) stated that a systems approach is extremely useful when modelling problems that involve networks of groups which may be operating at different scales but interact at certain points. In particular such an approach can be applied when there is a sudden transition in the network corresponding to a breakdown. Policy makers need to have simulations of models at their fingertips in order to be able to make crucial decisions, often in a very short time frame. The relationship between the speed of operation and the speed in which they respond to external influences is critical to system behaviour. These ideas work well on a conceptual level but clearly need more refinement for specific problems.

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*,
Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_25,

The focus of Klaus Hasselmann's (founding Director, the Max Planck Institute of Meteorology) talk was specifically related to policy for climate change taking into account the key socio-economic aspects. Aspects of globalisation of our businesses and economy must be taken into account when trying to develop truly effective policies. Agent-based models allow the effects of choices made by different actors (e.g. governments) to be explored. A method was presented for constructing computer-efficient coupled climate-socio-economic models. This type of model may not yet be able to be used in a predictive manner but rather as a tool for understanding how the various aspects are inter-related. This has the additional advantage of being simple enough so as to improve the interactions between the policy makers and the scientists.

Bert de Vries (Netherlands Environmental Assessment Agency and Professor of Global Change and Energy at the Copernicus Institute for Sustainable Development and Innovation of Utrecht University) explained that scenarios are a useful way of exploring our increasing complex world, particularly the climate-energy issue. It is clear that opinions and values must be taken into account. Science should offer novel, integrated ways to deal with the sustainable management in social-ecological systems or human-environment systems. Simulation and visualisation methods, such as gaming experiments, must be used to explore situations which, in turn, will improve the interface between scientific insights and uncertainties, on the one hand, and the policy makers and public on the other.

Henri Berestycki (Ecole des Hautes Etudes en Sciences Sociales, Paris and Director of the Centre d'analyse et de Mathématique Sociales of the French CNRS) heads a multidisciplinary team that uses complex systems modelling applied to problems from the social sciences. They apply methods from mathematics, including techniques from nonlinear PDEs and reaction-diffusion equations, and incorporate concepts and methods borrowed from the statistical physics of disordered systems to provide a framework for their studies. He explained that their modelling goals are two-fold. Firstly they seek models that exhibit generic properties, but then they also model specific problems, and confirm results by comparison with empirical data. In the past, efforts have been directed at biodiversity, sustainable development and on how people make a choice under social influence. Here models consider a large number of agents which have to make a binary choice (to buy or not buy) and link/compare this to the usual Nash equilibria when individual choice depends on others choice. However, as is typical of nonlinear systems we now have multiple equilibria. One particular problem discussed was the modelling of crime patterns. This work considers the diffusion of illegal behaviour, the analysis of crime time series, attempting to separate the global trend from local fluctuations.

Carlo Jaeger (PIK and Chair European Climate Forum) gave a stimulating talk based on a single figure of economic and social trends. It was a masterclass in eclectic teaching since several of the points he wished to raise had already been aired in discussions. His approach generally is to try and

promote the development of a model that prevents confusion between the various existing techniques which range from traditional economic equilibrium models to those which consider complex adaptive systems. He has been invited to demonstrate these ideas to German decision makers. However on the day, rather than discussing how any models, no matter which you choose, can be used to model rapid changes in our society or economic growth, we should first use these models to discover why our system remained stable for apparently large portions of time. Only when we can understand this will be able to consider the catalogue of inter-linked actions that lead to major shifts in human socio-economic systems.

Systems Approaches for Critical Decisions

Julian Hunt[1,2], Steven Bishop[3] and Yulia Timoshkina[4]

[1] CERFACS, IMFT Toulouse, 31400, France
[2] Arizona State University, Tempe, AZ 85212, USA
[3] Department of Mathematics, University College London, WC1H 0AY, UK
[4] Cambridge Centre for Energy Studies, University of Cambridge, CB2 1QA, UK

Summary. In this paper three types of system analysis are considered at a conceptual level which are relevant for decision making, namely: (a) breakdown in connected transport or other networks, when a change in modelling may be needed during critical transitions; (b) systems with dynamical boundary processes in smooth and sudden transitions; (c) critical transitions and sensitivities of the throughput and behaviour of systems depending on the relation between their 'speeds' of operation and response to external influences.

1 Description of General Systems Dynamics (GSD)

Natural and artificial entities, or systems, from molecules all the way through to whole societies, consist of many disparate elements operating simultaneously but with some level of connection between them [1]. In models of many environmental, engineering, social and economic/financial systems [2] a choice is made between statistical and quasi-deterministic methods. But an exclusive choice between these two approaches may not be necessary [3]. For example, in seasonal weather forecasts (www.metoffice.gov.uk) the two methods are currently used simultaneously. Some applications are described below for the conceptual application of a systems approach in making critical decisions particularly when systems are undergoing significant transitions. They can, perhaps, guide us how to operate systems so as to minimise the adverse effects of external or unexpected internal influences.

2 Applications of GSD for Critical Decisions

2.1 Breakdowns in Connected Networks

Studying patterns of restricted paths in idealised mathematical networks is a powerful method of studying the operation of real and virtual networks.

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*,
Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_26,

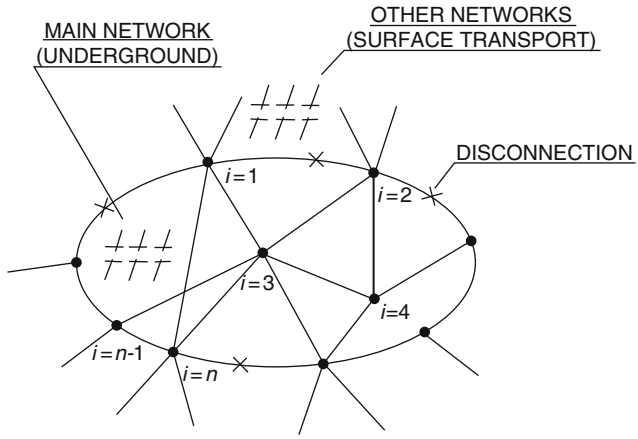


Fig. 1. Breakdown in connected networks-showing the effect of a few disconnections near nodes in a main network (e.g. underground) diffusing into other networks (e.g. surface transport)

They are particularly revealing when elements of the networks are changed, for example, by disruptions or improvements.

Following Euler, the key quantity describing paths is the connectedness matrix A_{ij} between n nodes, the magnitude of whose elements define the quality of the connection (or lines), e.g. probability between 0 and 1, between nodes i and j. For example, this defines the total number of significant connections N_i at node i (the sum of all the values of j for which $A_{ij} \neq 0$ and $i \neq j$). Consider the effects of N_b breakages in the connections of the network (see Fig. 1). An assumption has to be made about whether the nodes at either end of the broken connections also fail. If each of these has an average of $< N_i >$ connections (e.g.=5 for central London tube nodes), it means that the total number of connections affected is about $2N_b < N_i >$. So a certain number of deliberate or accidental breakages (disconnections) can affect a high proportion of the central part of a network [4].

However, the operation of the underground network with a finite number of high capacity lines is closely connected to a much larger more diffuse network, consisting of surface transportation and walkers etc. There are parallels with movement of oil and water through porous rock and through connected cracks in the rock, or urban networks of fractured water mains. For planning changes, responding to breakdowns, one form of simplification is to reduce the complexity the networks to fewer edges by averaging over many elements.

In a city with dense transport networks we may represent the movement of people as a diffusive flux F_p equal to the spatial variation (or 'gradient') of the number of people per unit area, and the diffusivities of the coupled networks D_1 and D_2 for flow. These diffusivities vary greatly across the city especially with breakdowns. The variations of the fluxes depend on local sources and sinks in the network (i.e. the numbers of people entering and leaving unit

area S, e.g. people entering or leaving activity areas (S_A) and overwhelmingly in emergencies by the movement of people away from areas of danger (S_D) as communicated and/or perceived. In dynamic situations the decision takers can vary all these parameters through physical controls (e.g. road blocks reducing the value of D_2) and communication. Simulations can solve the diffusion equation and rapidly display results as different scenarios are tried.

2.2 Systems with Dynamical Boundaries

Many systems are defined in relation to a finite physical, or non-physical, space which has boundaries (B) (e.g. a static organisation such as a city, or moving human/animal groups, or abstract boundaries, such as defined by 'areas' of activity and their scales in businesses or academia). Just as with networks, analysis can provide guidance about these systems when the boundaries and boundary processes undergo significant changes – drawing on the recent general theories of complex evolving and disrupting surfaces in turbulent fluid flow [5], and new concepts about how flooding patterns can change [6].

Richardson first showed the power of applying these concepts to social systems in his analysis of the frequencies of conflicts between nations, which he correlated with the lengths L of the boundaries B that separated them [7]. This led him to the famous conclusion that the smaller the scale l of the wiggles of the boundary shape the greater the length, according to the fractal relation $L \propto [l]^{-d}$, where $0 < d < 1$).

Consider a space within a continuous closed boundary B (Fig. 2). Outside B the key variable A, say, is A_0. Inside B, $A = A_0 + \Delta$. This changes when the surface undergoes severe disruption. With an evolving boundary, B moves outwards at an average boundary (or entrainment) 'speed' V_b. In many cases the boundary is porous, so that there is a flux of external 'activity' that crosses B in proportion to the flux (entrainment) 'speed' V_f.

One class of confined system with evolving boundaries is where the activity within the boundary is changing as the boundary spreads and exchange processes occur across the boundary. In other types of system Δ protects the system within B against an external activity A_0, e.g. the reduced flood hazard or lower wind damage in an urban area produced by deflection of water/wind by the buildings, or the reduced threat or competition to people, animals or organisations produced by joint defence against adversaries. In both cases $\Delta < 0$.

Within these boundaries, as the magnitude of the flux V_f of external activity crossing B (e.g. of fluid flow or of external bodies) grows, the protection within B might decrease (e.g. greater competition) or increase (e.g. economic advantage of immigration) in proportion to V_f and inversely with the length L_B of the boundary. The number of exchanges between insiders and outsiders would increase with L_B and this might trigger conflict [7].

Above a critical threshold, typically defined by the external action A_{crit}, the external and internal processes inter-mingle. Typically the mean

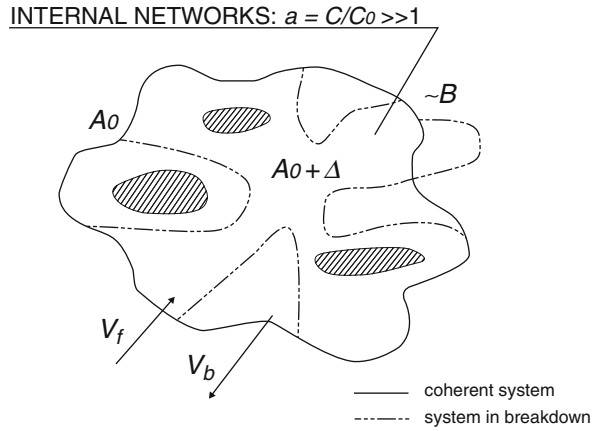


Fig. 2. A system with dynamical boundary processes undergoing transitions. A distortion and break up of the boundary produces large fluxes in and out, and large fluctuation A'

differential activity Δ decreases while the fluctuations A' and flux speed V_f increase. This might be associated with change in the activity within a fixed boundary (e.g. water/wind flow rising to high enough levels within an urban area that it becomes more hazardous inside B with infrastructure collapse than outside B). Or changes occur associated with the shape of the surface B becoming highly distorted as it breaks up into smaller areas each with surfaces denoted by dashed line, e.g. as a diseased population spreads or as spatial systems (as clouds and organisations break up).

These are also generic features of systems defined within multiple, interacting boundaries, such as when they merge or split, which applied to flow systems and adjoining nations [7].

2.3 Critical Dynamics of System-Processes Affected by Non-Local Influences

In many physical and non-physical systems there are various kinds of throughput, Q say, which are made up of 'movements' or transfers of quantities A (objects, activity, ideas etc). In changing conditions, the rate of accumulation always has to be considered at the same time as throughput. The systems involve large numbers of moving and evolving elements, which may include A and also extend beyond A, such as frameworks, external controls etc. Typically the throughput is controlled by local interactions between elements (as in transport/flow systems and in social organisations) and other 'non-local' influences or signals (Σ) coming from elsewhere in the system. Σ can be considered to be distinct from the quantity A. But Σ may be affected by large changes in A, such as when sudden changes and 'shocks' occur. A system dynamics approach also has to take into account its response process in order to estimate the speed (c) at which A is affected by the 'signals'.

The processes of accumulation and throughput with varying external influences have characteristic patterns of gradual and sharp variations that are common to many systems. Fluid flows provide a good example which show how similar behaviour occurs in different liquid and gaseous systems. These concepts have already been used to analyse and control non-fluid systems. In fluids non-local signals are waves moving with a speed c that in general differs from the speed of the flow V, though it may be affected by the flow at distant points. The equation for the change of A affected by the wave moving in one dimension shows how any arbitrary 'activity' moves at speed c. In river flows or on water surfaces, A could be the flow speed or the heights of waves, and c is 'wave' speed at which the current changes or the wave height moves. Its magnitude depends partly on the form of A, as well as on the particular system. In gases, which are compressible, c is proportional to the density – for air this is the familiar sound speed of 300 m/s, very fast compared to long waves of 3 m/s in a typical shallow river.

Where the flow has a speed $V < c$, it responds immediately to the any variations elsewhere in the system (e.g. along a river). However when the critical ratio V/c (the Froude number for liquids or Mach number for gases) exceeds 1, the flow is faster than the speed of the waves or signals from elsewhere and are less dependent of non-local influences (e.g. what happens downstream). The responses to influences are quite different to those in subcritical systems. Typically the throughput is locally obstructed (e.g. a fast flow of traffic being blocked) followed by a sudden change in the local and the overall flow occur, such as a hydraulic jump (a frothy wave on a stream) or shock wave (in front of an aircraft) in which there are intense local agitations [8]. Downstream of the 'shock', the river level rises, and in gases the density rises as in traffic density ('waves') on highways. As is well known there is a bumper-to-tail slow flow where $V/c < 1$ and free flowing supercritical traffic where $V/c > 1$. As V/c increases, the throughput of traffic increases gradually, as the traffic responds to non-local influences e.g. controls or obstructed flow. This understanding has led to traffic controls that maximise Q and reduce the chance of large waves or shocks, by ensuring that V/c is below its critical value. The patterns of mass movements of people in streets and buildings have many of the same smooth/shock transitions, often with deadly results.

There are also social and intellectual systems with non-local influences where the variation of throughput Q have similar characteristic variations depending on the relation between the speed at which the system operates (V) and the speed (c) with which information is considered or at which changes to the system propagate through it. For example, organisations in a subcritical mode ($V/c < 1$) operate smoothly, but probably not very sensitively, in response to external and non-local influences. In a super-critical mode ($V/c > 1$), they have to respond quickly to external influences, but they are at greater risk of the whole organisation experiencing sudden changes in its activity A, that are similar to shocks in flow and traffic systems. These ideas might also guide research into how individuals operate in the modern world where a certain imposed 'speed' V is required to deal with their activities (which they

can choose to some extent). Their effectiveness is affected by how this imposed speed V relates to each individual's innate speed c of processing information and responding to external influences. Probably greatest contentment comes from operating close to the critical ratio; they might also be one component of happiness [9].

3 Conclusions

Wilson [10] has commented that science is rich in concepts that have wide potential application through the methodology of complex systems analysis. But detailed modelling and measurement can greatly increase the value of system studies for decision making, because component models differ considerably between different systems. However, there are some generic issues of complex modelling that need to be discussed and teased out before non-technical policy makers will begin to use systems thinking and techniques more widely, and use the results intelligently.

Acknowledgments

We acknowledge the financial support of the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European Commission, under the FET-Open grant agreement GSD, number 221955. JH acknowledges support of the 5th European Conference on Complex Systems in Jerusalem 2008 and funding from Midi-Pyrenees Innovation for the preparation of this paper.

References

1. Smuts, J.: Holism and Evolution. Macmillan and Co., London (1926)
2. Hasselmann, K.: Application of System Dynamics to Climate Policy Assessment. Proceedings of the ECMI2009, London (2008)
3. Hunt, J.C.R.: Communicating big themes in applied mathematics. In Mathematical Modelling – Education, Engineering and Economics. Proceedings of the ICTMA12, London (2006a)
4. Angeloudis, P., Fisk, D.: Large subway systems as complex networks. Physica A. **367**, 553–558 (2006)
5. Hunt, J.C.R., Eames, I., Westerweel, J.: Mechanics of inhomogeneous turbulence and interfacial layers. J. Fluid Mech. **554**, 499–519 (2006)
6. Stelling, J.: Perspectives. Understandable Complexity. Science **376**, 9 (2007)
7. Hunt, J.C.R.: Life and Work of L.F. Richardson. In: Collected works of L.F. Richardson. Cambridge University Press, UK (1993)
8. Lighthill, M.J., Whitham, G.B.: On kinematic waves. II: A theory of traffic flow on long crowded roads. Proceedings of the Royal Society, London (1955)
9. Layard, P.: Happiness: Lessons from a New Science. Penguin Books, UK (2005)
10. Wilson A.: Complex Spatial Systems: The Modelling Foundations of Urban and Regional Analysis. Prentice Hall, Harlow, UK (2000)

Application of System Dynamics to Climate Policy Assessment

Klaus Hasselmann

Max Planck Institute of Meteorology, Hamburg; European Climate Forum,
klaus.hasselmann@zmaw.de

Summary. To provide useful scientific advice to climate policymakers, a paradigm shift in mathematical economics is needed from general equilibrium concepts to agent-based system dynamics. For effective communication, the simulation models should be simple. This can best be achieved by developing models as a hierarchy, progressing from simple to more complex versions. Examples are given of work currently being carried out in the EU project "Global System Dynamics and Policies".

1 Integrated Assessment of Climate Change

Through the persistent efforts of the UN Intergovermental Panel on Climate Change (IPCC), [5], and the mounting observational evidence, the reality of human induced climate change is today no longer seriously disputed. Governments worldwide are committed to implementing effective climate mitigation policies. However, in contrast to the central role of IPCC in bringing the climate problem to the attention of the public and policymakers, the impact of IPCC in developing effective policies to combat climate change has been marginal [3]. This can be largely attributed to the reliance on general equilibrium macroeconomic models in the assessment of climate policies [1]. The general equilibrium approach is unable to capture the basic dynamic processes that must be invoked to transform our present fossil-based socio-economic system to a sustainable carbon-free system. It ignores also other important aspects of globalization that cannot be separated from the problem of global climate change, such as widespread poverty and growing rich-poor inequalities, with associated migration pressures and increases in conflict potential. Similarly excluded are shorter-term processes such as business cycles, recessions and financial instabilities, which although traditionally disregarded in economic growth models, represent important considerations in the unavoidable short-term/long-term trade-off decisions of policymakers.

A central goal of the EU networking project "Global Systems Dynamics and Policies" is to overcome these shortcomings by creating a network of

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*,
Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_27,

researchers cooperating in the development of a new generation of integrated assessment models based on dynamical agent-based models. The standard general equilibrium paradigm of main-stream neo-liberal economics is based on Adam Smith's famous "invisible hand": although the economy is governed by the diverse actions of innumerable competing players, the net outcome is nevertheless an optimal equilibrium state in which the integrated welfare of all players is maximized. In contrast, the multi-agent paradigm views the economy as a nonlinear system with many degrees of freedom that is inherently chaotic, exhibiting random fluctuations and major instabilities (dramatically exemplified by the most recent global financial crisis). An approximately stable growth path can be maintained only if the instabilities are understood and counteracted by appropriate government policies. The inherent dynamics of the socio-economic system and the important role of governments becomes particularly relevant in the context of climate change.

2 The Model Hierarchy MADIAMS

The attainable complexity of a multi-actor model is limited by two natural constraints: the available data, and the difficulty of distinguishing between competing hypotheses if the model contains too many free parameters. To ensure that one remains within these limitations, it is useful to develop models as a model hierarchy, beginning with the simplest model at the lowest level, and successively introducing more processes at higher model levels, until one reaches a limiting level of complexity determined by the data and parameter constraints.

As illustration, we consider a model hierarchy MADIAMS (Multi-Actor Dynamic Integrated Assessment Model System) developed from an earlier single-level model MADIAM [8](see also [2]). The hierarchy is divided into three model levels M1, M2 and M3, each of which can be further sub-divided into sub-levels M1a, M1b, ..., M2a, M2b, ..., M3a, M3b, ... depending on the number and type of sectors, regions, actors, etc. The lowest-level model M1 describes a macroeconomic system governed by the actions of three representative actors: firms, households and banks. Governments are included in the next model level M2, while the highest model level M3 contains also a climate module.

The lowest model level M1 is similar to the core macroeconomic model of the original MADIAM model, but with an important difference: instead of filtering out faster variations in the supply and demand of consumer goods by regarding these as equilibrated with respect to the slower time scales of the mean growth of physical and human capital, all three production outputs, including consumer goods, are treated in M1 as dynamic, non-equilibrated stock variables. This enables a combined investigation of both fast and slow dynamic processes. The model is thereby able to simulate business cycles, recessions and the impact of the counteracting stabilization policies of a

central bank. This provides the necessary background for the investigation of the combined impact of long-term climate mitigation measures and short-term monetary and fiscal stabilization policies in the higher model levels M2 and M3.

The inclusion of governments in model level M2 enables the consideration of fiscal policy in addition to monetary policy in stabilizing economic growth, thereby illuminating the different assumptions on actor behavior underlying the long-standing debate between post-Keynsians and monetarists. On longer time scales, various model sub-versions simulate the effects of government climate policies in the form of a carbon price, subsidies, or direct emission regulations. Also included as an option at model level M2 is the role of the media in influencing consumer preferences and public support for climate policies.

The third model level M3, finally, is completed to a fully coupled climate-socio-economic integrated assessment model by incorporating the climate sub-module NICCS (Nonlinear Impulse response coupled Climate-Carbon-cycle System) [4] of the original MADIAM model. NICCS computes the greenhouse gas forcing by CO_2 emissions and the resultant climate change in the form of regionally dependent changes in near-surface temperature and sea-level (represented in both cases by the dominant first empirical orthogonal functions). The back-interaction of the computed climate change on the macroeconomic system is expressed in terms of simple aggregate impact functions. Not considered in the original MADIAM version of the model is the interaction between different economic regions via trade, an important extension that still needs to be implemented.

3 Simulation Examples

The following simulation examples illustrate two basic points: (1) long-standing debates over the role of actor behavior in governing macroeconomic dynamics can be readily quantified and illuminated by translation into simple system-dynamics models, and (2) even for very simple models it is nevertheless often difficult to predict a priori the outcome of assumed actor behavior (although this can normally be readily reconstructed a posteriori). Thus system dynamics should be seen primarily as a learning and expository tool.

Figure 1, from a model M1 simulation, shows two different growth paths resulting from two equally plausible supply strategies of firms in response to changing demands for consumer goods. In simulation *S*, firms strive to maintain a chosen target level of the goods *stock* by adjusting the investments in the consumer goods production sector at a rate proportional to the deviation of the goods level from the target level. In simulation *F*, in contrast, the adjustment rate was set proportional to the difference between the *flows* into and out of the goods stock. Simulation *S* favors short term consumption over profits and long-term growth, while the reverse holds for simulation *F*. The point here is not which of the two hypotheses is closer to reality (a question that

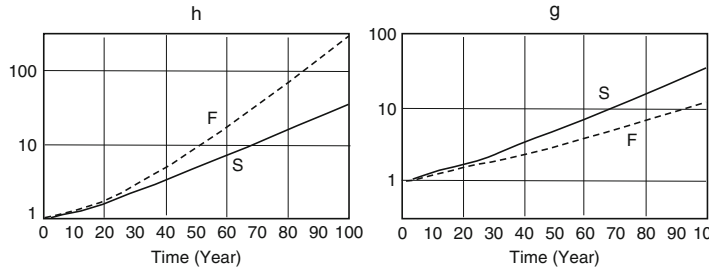


Fig. 1. Growth paths of human capital h (*left panel*) and physical capital k (*right panel*) for two different firm supply strategies S, F, in response to variable consumer goods demand; S (*full curves*): maintainence of target consumer goods stock; F (*dashed curves*): flow balance between consumer goods production and demand

can be decided only by comparisons with data and/or stakeholder interviews) but the signicant differences in growth paths resulting from elementary differences in actor behavior – features that cannot be captured in a traditional actor-independent growth model.

A simple modification of the assumed behaviours of consumers and firms in model M1 gives rise to business cycles (Fig. 2, right panel). The relevant feedback interactions are indicated in the left panel[1]. A decrease in consumption *delcons* (triggered, for example, by a decrease in consumer confidence) induces a slow-down in production *dely*, with an associated reduction in employment by firms, further reducing consumer confidence, and so on. This positive feedback loop alone would result in exponential decay or growth (a recession or boom, depending on the initial conditions). However, the exponential instabilities are converted into a periodic cycle through a stabilizing negative feedback loop (bottom two boxes), representing the willingness of firms to employ more labor once wages *delw* have been sufficiently depressed by the reduced employment level.

There exist, or course, many alternative explanations of business cycles, with numerous associated proposals for their control through appropriate monetary or fiscal policies [6]. The present example underlines the earlier comment that macroeconomic hypotheses can be readily expressed in appropriate system dynamics terms, but the outcome of the model simulations, even for the simple model shown in Fig. 2, is normally strongly dependent on the details of the hypothesized actor behaviour and difficult to forsee. Thus, in the present example, the cycles can have very different amplitudes and periods, or can revert to exponential growth or decay, depending on the values of the feedback coefficients ($fac1, fac2, fac2a, fac3$) characterizing the inter-actor coupling.

[1]The diagram represent a stocks-and-flows sketch generated by the system-dynamics graphic-modeling tool Vensim. Stocks are represented as boxed variables, rates of change by closed-cross symbols, integrations by double arrows, sources and sinks as clouds, and interdependencies by single-arrow connections.

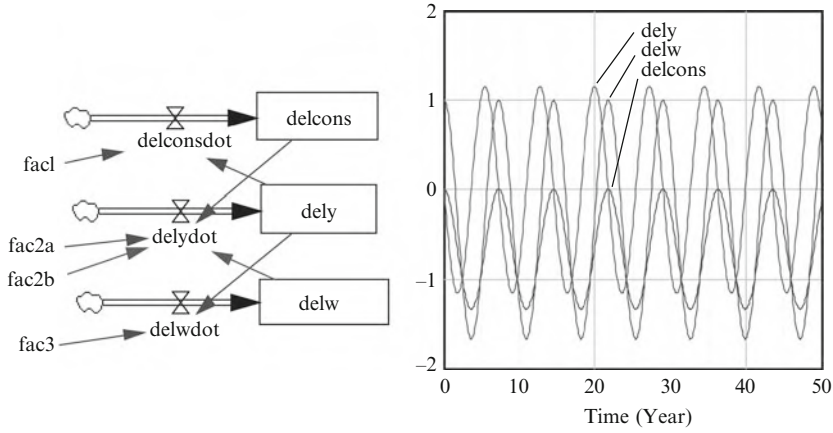


Fig. 2. *Left panel*: Business cycle model of feedbacks between modifications of consumption (delcons), production (dely) and wage levels (delw). *Dashed lines* represent positive feedbacks driving exponential instabilities, dotted lines negative feedbacks leading to oscillations. The variables fac1,. . . , fac3 denote feedback coefficients which control whether the instabilities lead to oscillations or exponential decay or growth. *Right panel*: a resulting oscillation, in normalized units

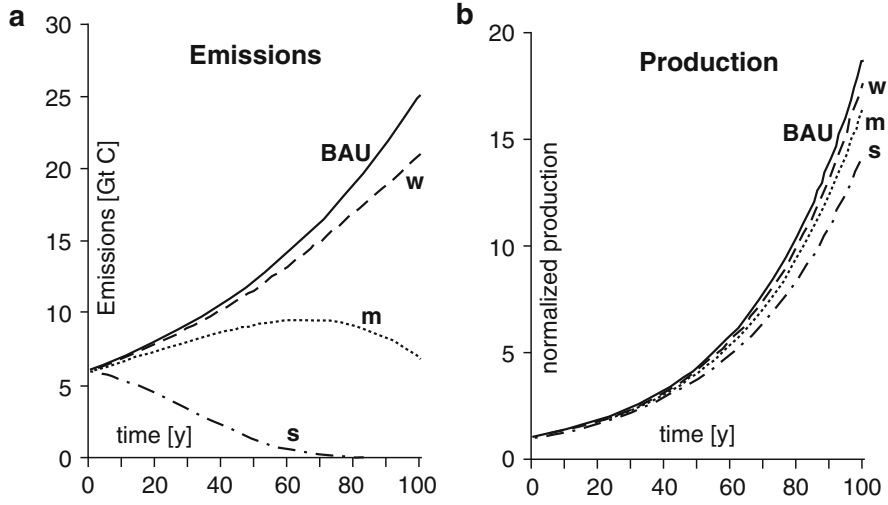


Fig. 3. Impact of various climate mitigation policies (*left*) on economic growth paths (*right*). Significant differences in emissions for weak, medium and strong mitigation policies are seen to have only a minor impact on long term economic growth

The last simulation example, from the third-level model M3 [8] (Fig. 3), illustrates the impact of government climate policies on CO_2 emissions and economic growth. The simulations support the conclusions of the Stern report [7] and other authors that the emissions responsible for global warming can

be reduced to acceptable levels at only a minor long-term economic cost of the order of 1% GDP. However, this result is again strongly actor dependent, for example with respect to the assumed response of firms and consumers to government policies.

4 Conclusions

An understanding of the interrelations between climate change and climate change policies requires the application of dynamic models that simulate the behavior of the key socio-economic actors. For an effective communication between scientists and policymakers the models should be as simple as possible. This is dictated also by the inherent uncertainties of human behavior and the unpredictability of future technological developments. Although necessarily simple, simulation models nevertheless represent the only reliable tool for deducing the implications of the assumptions regarding human behavior and future technological developments that are unavoidable in making climate policy decisions.

Acknowledgement

Support for this work through the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European Commission, under the FET-Open grant agreement GSD, number 221955, is gratefully acknowledged.

References

1. Barker, T.: The Economics of Avoiding Dangerous Climate Change, An editorial essay. Clim. Change **89** (2008)
2. de Vries, B.: SUSCLIME: a simulation/game on population and development in a climate-constrained world, Simulation and Development, vol. 29, pp. 216–237. (1998)
3. Hasselmann, K., Barker, T.: The Stern Review and the IPCC fourth assessment report: implications for the interaction between policymakers and climate experts. An editorial essay. Clim. Change **8**9, 219–229 (2008)
4. Hooss, G., Voss, R., Hasselmann, K., Maier-Reimer, E., Joos, F.: A nonlinear impulse response model of the coupled carbon cycle-climate (NICCS). Clim. Dyn. **1**8, 189–202 (2001)
5. Intergovernmental Panel on Climate Change: Fourth Assessment Report, Working Groups 1,2 and 3, Cambridge University Presse, 2007
6. Lucas, R.E. Jr.: Models of Business Cycles. Oxford; Blackwell (1987)
7. Stern, N.: The economics of climate change. The Stern Review (2007)
8. Weber, M., Barth, V., Hasselmann, K.: A Multi-Actor Dynamic Integrated Assesment Model (MADIAM) of induced technological change and sustainable economic growth. Ecol. Econ. **5**4, 306–327 (2005)

Minisymposium *Multivariate and/or Multidimensional Image Processing in Biomedical Applications*

J. Angulo and D. Jeulin

CMM-Centre de Morphologie Mathématique, Mathématiques et Systèmes, MINES Paristech; 35, rue Saint Honoré - 77305 Fontainebleau cedex, FRANCE
jesus.angulo@ensmp.fr, dominique.jeulin@ensmp.fr

Nowadays many different modalities are available in medical imaging, including computed tomography (CT) scans, functional or dynamic contrast-enhanced magnetic resonance imaging (fMRI) or (DCE-MRI), positron emission tomography (PET). The 2D/3D + time images produced by these advanced devices are useful for cancer diagnosis, radiotherapy or surgery planning, active study of human brain, tumour angiogenesis quantification, etc. In addition, the most recent microscope systems in biomedical laboratories are based on multi/hyper-spectral imaging for brightfield or fluorescence microscopy.

High throughput exploitation of these multivariate and/or multidimensional images requires advanced image processing methods and algorithms. To take into account jointly the spatial and the temporal/spectral information as well as the way to combine or to reduce the different temporal/spectral dimensions need adapted mathematical models. Moreover, extension of standard image processing approaches to 4D images leads to inefficient algorithms in terms of computational requirements (memory overload, time of computation, etc.).

In this framework, the aim of this minisymposium was to draw an overview of some recent developments in the field by French and German teams. We focus in particular on techniques which lie in mathematical morphology, multivariate data analysis, statistical classification, graph-based representations and algorithms, stochastic modelling, optic fow estimation, etc.

G. Noyel, J. Angulo, and D. Jeulin, from Mines ParisTech (France), consider automatic segmentation of DCE-MRI series in angiogenesis imaging. The approach is based on stochastic watershed segmentation for hyperspectral images and more specifically, the paper focuses on new methods to generate random germs regionalized by a previous classification in order to use probabilistic watershed on hyperspectral images. These germs are much more efficient than the standard uniform random germs. The algorithms are illustrated

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*,
Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_28,

to compare the obtained segmentation which is then needed for detecting the eventual tumours.

J. Angulo, from Mines ParisTech (France), discusses individual nucleus modelling and segmentation, from fluorescence labelled images, of cell populations growing in complex clusters. The proposed approach is based on models and operators from mathematical morphology. Cells are individually marked by the ultimate opening and then are segmented by the watershed transformation. A cell counting algorithm based on classical results of Boolean model theory is heuristically used to detect errors in segmented clustered nuclei.

J. Stawiaski, E. Decencière, and F. Bidault, from Mines ParisTech (France) and Institut Gustave Roussy (France), present a segmentation method of 3D time-series images for radiotherapy planning. The aim of this study is to propose some techniques for the segmentation of tumors surrounding or contained in the lungs. The 4D images are produced using a respiration gating procedure and computed tomography. It uses a 4D watershed algorithm, combined with graph-based techniques to delineate the tumors in the time-series.

K. Rohr, W.J. Godinez, N. Harder, S. Yang, I.-H. Kim, S. Wörz, and R. Eils, from University of Heidelberg (Germany) and German Cancer Research Center (DKFZ), summarise in their paper tracking and registration approaches developed for automatic analysis of multidimensional biomedical images. The tracking approach allows computing the trajectories of cells in fluorescence microscopy image sequences. The registration approach enables to geometrically align cell microscopy images by using elastic transformations.

Regionalized Random Germs by a Classification for Probabilistic Watershed Application: Angiogenesis Imaging Segmentation

Guillaume Noyel, Jesús Angulo, and Dominique Jeulin

MINES ParisTech, CMM – Centre de Morphologie Mathématique, Mathématiques et Systèmes, 35 rue Saint Honoré – 77305 Fontainebleau cedex, France
{noyel,angulo,jeulin}@cmm.ensmp.fr

Summary. New methods are presented to generate random germs regionalized by a previous classification in order to use probabilistic watershed on hyperspectral images. These germs are much more efficient than the standard uniform random germs.

1 Introduction

Probabilistic watershed was introduced by Angulo and Jeulin [1] to detect the contours of the widest and the most contrasted regions in images. The obtained contours are more regular and significant than these associated to the deterministic watershed. Probabilistic watershed was then extended to hyperspectral images by Noyel et al. [5].

The standard stochastic WS consists in starting from uniform random points germs as sources to flood the norm of a gradient in order to obtain the associated contours. After repeating the process a large number of times, a probability density function of contours (pdf) is computed by the Parzen kernel method [1]. The pdf is segmented by a hierarchical watershed according to a morphological criterion such as the volume (i.e. integral of intensities) [3]. For hyperspectral images, a pdf is built for each channel of the image and the flooding function is the weighted sum of the pdf of the channels. This function, called a marginal probability density function, is based on spatial information [5].

As, for hyperspectral images, a spectral classification can be computed [4, 6], it is interesting to estimate the marginal pdf *mpdf* conditionally to this previous spectral classification [6]. Therefore, this pdf represents jointly spatio-spectral information.

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*,
Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_29,

In the sequel, after presenting the results obtained using uniform random germs, we compare several approaches to compute random germs regionalized by a previous classification.

2 Prerequisites

Our results are presented on a medical image of DCE-MRI series (Dynamic Contrasted Enhanced Magnetic Resonance Imaging) of mice. The image is a series of 512 channels of size 128×128 acquired at a regular step of 1 s, in time, on mice presenting tumors [2]. This image is filtered and represented in a parameter space $\mathbf{p}$ of a smaller dimension [6]. A marginal pdf is built in this space. The pdf is segmented by a hierarchical watershed according to a volume criterion in 20 or 30 regions. An external marker is added during the computation of the pdf. The results of the segmentation are presented on a channel of the image space $\mathbf{f}_\lambda$. In Fig. 1, we notice that the pdf $mpdf(\mathbf{p}, mrk_i)$ presents a lot of contours on the background of the image. Therefore the image segmentation, $sg_R^{vol}(mpdf(\mathbf{p}, mrk_i))$, leads to an over-segmentation especially in the background.

In fact, for images presenting wide and low contrasted regions, several germs (or markers) may fall in these regions during the pdf estimation with uniform random germs. These uniform random germs create artificial contours that do not corresponds to relevant contours.

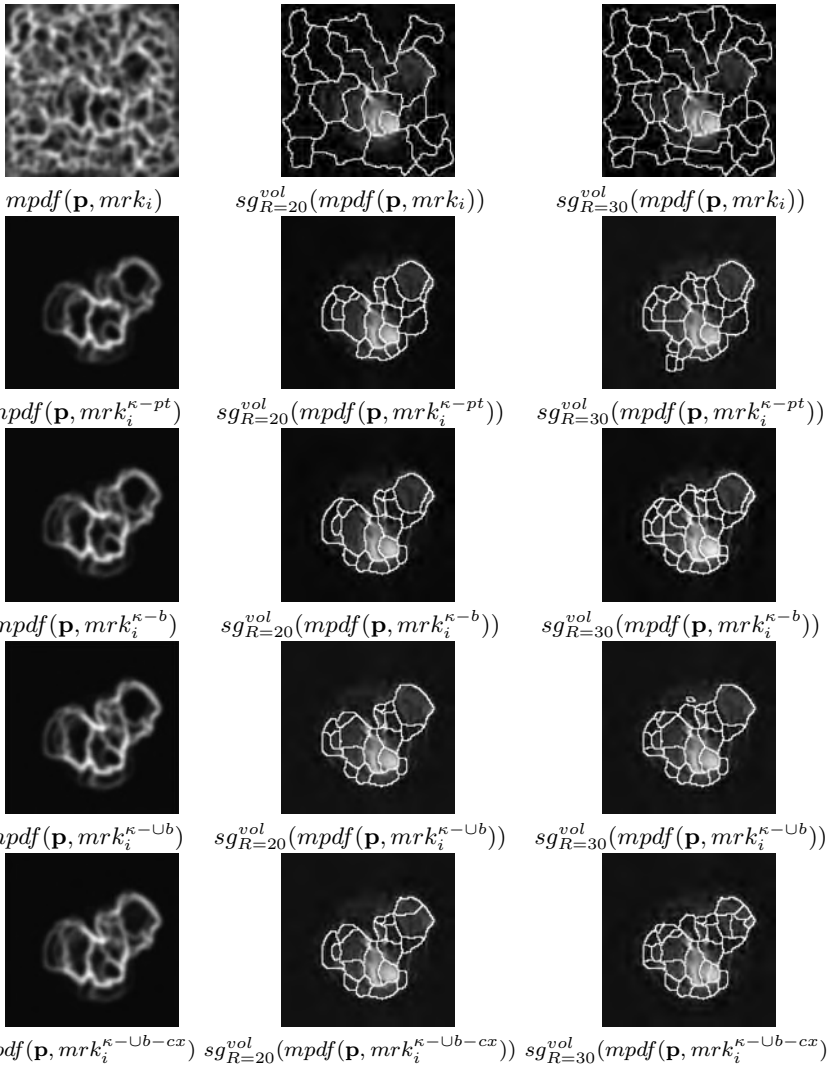
That is the reason why random germs regionalized by a previous classification are introduced. In order to do this, random germs are drawn in the classes of the classification. However, to avoid that a germ may fall on the boundary of a class, that would lead to a leak during the flooding process of the watershed on the pdf, each class of the classification κ is reduced by an erosion (i.e. an anti-extensive transformation) with a square structuring element (s.e.) with size 3×3 pixels. Therefore a new void class is introduced. Moreover the holes in each class are filled by a closing by reconstruction (s.e. 3×3 pixels). After this morphological transformation, the classes are re-labelled with a different label for each connected class. The transformed classification is written $\widehat{\kappa}$ and named pre-segmentation (Fig. 1 top). It is composed of connected classes, $\widehat{\kappa} = \cup_k C_k$ with $\cap C_k = \emptyset$. The void class is written C_0.

3 Random Points-Germs Regionalized by a Classification

Uniform random germs are drawn on the pre-segmentation $\widehat{\kappa}$. If a germ falls in a connected class C_k not yet marked by a previous germ, $k \neq 0$, then it is kept otherwise it is rejected. As we have a pre-segmentation $\widehat{\kappa}$, we use it to detect the background of the image by preventing germs from falling into the background class (Algorithm 1). Therefore not all the germs are kept.

We notice that the resulting segmentations $sg_R^{vol}(mpdf(\mathbf{p}, mrk_i^{\kappa-pt}))$ are much better than with uniform random germs $sg_R^{vol}(mpdf(\mathbf{p}, mrk_i))$ (Fig. 1).

κ $\widehat{\kappa}$



$mpdf(\mathbf{p}, mrk_i)$ $sg^{vol}_{R=20}(mpdf(\mathbf{p}, mrk_i))$ $sg^{vol}_{R=30}(mpdf(\mathbf{p}, mrk_i))$

$mpdf(\mathbf{p}, mrk_i^{\kappa-pt})$ $sg^{vol}_{R=20}(mpdf(\mathbf{p}, mrk_i^{\kappa-pt}))$ $sg^{vol}_{R=30}(mpdf(\mathbf{p}, mrk_i^{\kappa-pt}))$

$mpdf(\mathbf{p}, mrk_i^{\kappa-b})$ $sg^{vol}_{R=20}(mpdf(\mathbf{p}, mrk_i^{\kappa-b}))$ $sg^{vol}_{R=30}(mpdf(\mathbf{p}, mrk_i^{\kappa-b}))$

$mpdf(\mathbf{p}, mrk_i^{\kappa-\cup b})$ $sg^{vol}_{R=20}(mpdf(\mathbf{p}, mrk_i^{\kappa-\cup b}))$ $sg^{vol}_{R=30}(mpdf(\mathbf{p}, mrk_i^{\kappa-\cup b}))$

$mpdf(\mathbf{p}, mrk_i^{\kappa-\cup b-cx})$ $sg^{vol}_{R=20}(mpdf(\mathbf{p}, mrk_i^{\kappa-\cup b-cx}))$ $sg^{vol}_{R=30}(mpdf(\mathbf{p}, mrk_i^{\kappa-\cup b-cx}))$

Fig. 1. *Top*: classification κ by LDA in four classes in parameter space and its transformed classification (or pre-segmentation) $\widehat{\kappa}$. *Bottom*: marginal pdf *mpdf* for several kinds of germs and associated segmentations by a volumic watershed

Algorithm 1 Random points-germs regionalized by a classification $mrk_i^{\kappa-pt}(x)$

Given N the number of drawn germs m
Set the background class and the void class C_0 to *marked*
for all drawn germs m from 1 to N **do**
    if C_k, such as $m \in C_k$, is *not marked* **then**
        Keep m
        Set the class C_k to *marked*
    end if
end for

4 Random Balls-Germs Regionalized by a Classification

One of the drawback of the point germs is to enhance the contours of the small regions. With larger germs the probability of the small contours decreases. To obtain random balls germs $B(m, r)$, the centers m of the balls are drawn according to a rule. If the center is kept, a random radius r is drawn in the interval $]0, Rmax]$.

4.1 Each Connected Class May Be Hit One Time

The centers m of the balls are drawn as random points-germs regionalized by the pre-segmentation $\widehat{\kappa}$. Only the intersection between the ball $B(m, r)$ and the connected class C_k, such as $m \in C_k$, is kept as a germ. We notice that each connected class may be hit one time (Algorithm 2). The segmentations are a bit better than with regionalized random points germs $mrk_i^{\kappa-pt}(x)$ (Fig. 1).

Algorithm 2 Random balls-germs regionalized by a classification (each connected class may be hit one time) $mrk_i^{\kappa-b}(x)$

Given N the number of drawn germs m
Set the background class and the void class C_0 to *marked*
for all drawn germs m from 1 to N **do**
    if C_k, such as $m \in C_k$, is *not marked* **then**
        $r = \mathcal{U}[1, Rmax]$
        Keep as a germ $\mathrm{B}(m, r) \cap C_k$
        Set the class C_k to *marked*
    end if
end for

When a class can only be hit one time, the drawback is that only a small number of germs are effectively implanted: for $N = 100$ germs, only an average of 6 are really implanted.

4.2 Each Connected Class May Be Hit Several Times

Increasing the number of really implanted germs leads to a better detection of contours thanks to larger markers in each class. That is the reason why we introduce the possibility that several germs may fall in the same connected class.

Union of Germs in Each Connected Class

When several random balls germs fall in the same connected class C_k, their intersection is made with the class C_k. Then their union is made to obtain the germ of the class (Algorithm 3).

In Fig. 1, the regions of the segmentations $sg_R^{vol}(mpdf(\mathbf{p}, mrk_i^{\kappa-\cup b}))$ are almost the same whatever the number of regions R are.

Algorithm 3 Regionalized random balls-germs: each connected class may be hit several times and the union of balls is made in each connected class of the pre-segmentation $mrk_i^{\kappa-\cup b}(x)$

Given N the number of drawn germs m
Set the background class and the void class C_0 to *marked*
for all drawn germs m from 1 to N **do**
    if C_k, such as $m \in C_k$, is *not marked* **then**
        $r = \mathcal{U}[1, Rmax]$
        $mrk_{old}^{C_k} = mrk_i^{\kappa-\cup b}(x) \cap C_k$
        $mrk_{new}^{C_k} = (\mathrm{B}(m, r) \cap C_k) \cup mrk_{old}^{C_k}$
        Add $mrk_{new}^{C_k}$ to $mrk_i^{\kappa-\cup b}(x)$
    end if
end for

Union of Connected Germs in Each Connected Class

As for the previous germs, when several random balls germs fall in the same connected class C_k, their intersection is made with the class C_k. Then the union of the connected germs is made to obtain one of the germs of the class (Algorithm 4).

We notice that the resulting segmentations may be a bit over-segmented $sg_R^{vol}(mpdf(\mathbf{p}, mrk_i^{\kappa-\cup b-cx}))$ (Fig. 1). It can be useful to make a thinner analysis of each region.

5 Conclusion

We have shown that using regionalized random germs by a classification is better than using uniform random germs in order to segment by means of the probabilistic watershed. Moreover, the segmentations, in which all

Algorithm 4 Regionalized random balls-germs: each connected class may be hit several times and the union of connected balls is made in each connected class of the pre-segmentation $mrk_i^{\kappa-\cup b-connex}(x)$

Given N the number of drawn germs m
Set the background class and the void class C_0 to *marked*
for all drawn germs m from 1 to N **do**
    if C_k, such as $m \in C_k$, is *not marked* **then**
        $mrk_{old}^{C_k} = mrk_i^{\kappa-\cup b}(x) \cap C_k$
        $mrk_{new}^{C_k} = (\mathrm{B}(m,r) \cap C_k) \cup mrk_{old}^{C_k}$
        Add $mrk_{new}^{C_k}$ to $mrk_i^{\kappa-\cup b}(x)$
    end if
end for
Label each connected regions in the image of markers

the germs fallen in a connected region of the pre-segmentation $\widehat{\kappa}$ have the same label, have generally correct contours $sg_R^{vol}(mpdf(\mathbf{p}, mrk_i^{\kappa-pt}))$ $sg_R^{vol}(mpdf(\mathbf{p}, mrk_i^{\kappa-b}))$, $sg_R^{vol}(mpdf(\mathbf{p}, mrk_i^{\kappa-\cup b}))$. When there may be several germs in a connected region, $sg_R^{vol}(mpdf(\mathbf{p}, mrk_i^{\kappa-\cup b-cx}))$, the image is over-segmented and it can be useful to make a thinner analysis of the segmented regions [6].

Acknowledgements

The authors are indebted to Pr. C.A. Cuenod (Hôpital Européen G. Pompidou, Paris, France) for providing the MRI images.

References

1. Angulo, J., Jeulin, D.: Stochastic watershed segmentation. In Banon, G., et al. (eds.) Proceedings of the 8th International Symposium on Mathematical Morphology, vol. 1, pp. 265–276. Instituto Nacional de Pesquisas Espaciais (INPE), 2007
2. Balvay, D., Frouin, F., Calmon, G., Bessoud, B., Kahn, E., Siauve, N., Clment, O., Cuenod, C.A.: New criteria for assessing fit quality in dynamic contrast-enhanced T_1-weighted MRI for perfusion and permeability imaging. Magn. Reson. Med. **54**, 868–877 (2005)
3. Meyer, F.: An overview of morphological segmentation. Intern. J. Pattern Recognit Artif. Intell. **15**(7), 1089–1118 (2001)
4. Noyel, G., Angulo, J., Jeulin, D.: Morphological segmentation of hyperspectral images. Image Anal. Stereol. **26**, 101–109 (2007)
5. Noyel, G., Angulo, J., Jeulin, D.: Random germs and stochastic watershed for unsupervised multispectral image segmentation. In: Apolloni, B., et al. (eds.) *KES 2007/ WIRN 2007*, volume III of *LNAI 4694*, pp. 17–24. Knowledge-Based Intelligent Information and Engineering Systems. Springer, Heidelberg (2007)
6. Noyel, G., Angulo, J., Jeulin, D.: Filtering, segmentation and region classification by hyperspectral mathematical morphology of DCE-MRI series for angiogenesis imaging. In: Proceedings of the IEEE International Symposium on Biomedical Imaging ISBI 2008, pp. 1517–1520. (2008)

Nucleus Modelling and Segmentation in Cell Clusters

Jesús Angulo

CMM – Centre de Morphologie Mathématique, Mathématiques et Systèmes, MINES Paristech; 35, rue Saint Honoré – 77305 Fontainebleau cedex, FRANCE
jesus.angulo@ensmp.fr

Summary. This paper deals with individual nucleus modelling and segmentation, from fluorescence labelled images, of cell populations growing in complex clusters. The proposed approach is based on models and operators from mathematical morphology. Cells are individually marked by the ultimate opening and then are segmented by the watershed transformation. A cell counting algorithm based on classical results of Boolean model theory is heuristically used to detect errors in segmenting clustered nuclei.

1 Introduction

High content screening (HCS) refers to technological platforms for parallel cells growing in multi-well plates (or in other supports as cell on chip) and fluorescent labelling of proteins of interest (immuno-fluorescence with antibodies, GFP-tagged proteins), together with image capture by automated microscopy and subsequent cell image analysis [5]. HCS is of interest for the discovery of new cellular biology mechanisms (i.e., using siRNA), new pharmaceuticals (i.e., mass screening of potential active molecules) or for the development of new tests for diagnostic/prognostic, for toxicology tests (i.e., evaluation of different compounds at different concentrations). Cell image segmentation [3] to define individual cells is the most critical step to achieve a robust high throughput system which will be able to process thousands of cell images without needing a manual interaction. Errors in segmentation process may propagate to the feature extraction and classification.

Many image processing algorithms have been proposed for cell segmentation, however segmentation of cell populations which grow in complex clusters is still a challenging issue [6]. This paper deals with individual nucleus modelling and segmentation of cell clusters from fluorescence labelled images. The proposed approach is based on models and operators from mathematical morphology [7], a non-linear image processing methodology which is proven to be a very powerful tool in biomedical microscopy image analysis. Cell images

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*, Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_30,

used in this study represent only the nuclear content (a DNA marker is used for fluorescence labelling). This is the most interesting case study since once the nuclei are detected and segmented the other cell markers can be easily quantified.

2 Morphological Model of Cell Population

A cell population can be modelled as a realization I of Poisson points $i \in I$ of intensity θ in $\mathbb{R}^2$, i.e., points implanted in the space independently one of the others according to a constant density θ. Let us consider C^{cell} as a compact random set, centred at the origin, which represents the "individual cell". For each point $i \in I$ a realization of C^{cell} is generated and implanted at associated point i, denoted C_i^{cell}. The union C_{popul} of the C_i^{cell}:

$$C_{popul} = \cup_{i \in I} C_i^{cell},$$

is by definition a realization of a Boolean set where the different C_i^{cell} are mutually independent. They can touch each other and overlap, and consequently can constitute cell clusters. In random set theory, the complement C_{popul}^c is named "pores" set of the "grains" set C_{popul}. The cell population C_{popul} is observed by an image associated to the microscopic field Z under study. However, this model only involves binary images, and in practice, the fluorescence images are scalar functions with values in the set $\mathcal{T}$ of grey levels. Consequently the realization of a individual cell is a random function f_i^{cell} whose support is the random set C_i^{cell} and the observed fluorescence image is

$$f_{popul} = \vee_{i \in I} f_i^{cell}.$$

As it is shown below, the binary model $C_{popul} \in \mathcal{P}(E)$ is used in a heuristic way for counting the cells in segmented clusters, which are obtained by the deterministic segmentation of the scalar model $f_{popul}(x)$ ($x \in E$, where $E \subset \mathbb{R}^2$ is the pixel space of the bounded field Z).

3 Individual Nucleus Segmentation

Figure 1a gives a typical example of cell nuclei population growing in overlapped clusters. Our purpose is to use automated watershed segmentation [1] to build the contours of individual cells. For a precise segmentation, watershed transformation $wshed(g, mrk)$ needs a scalar function of contour energy g and a marker for each cell mrk. The function g is calculated using the morphological gradient, defined as the difference between the dilation and the erosion [7], i.e., $g = \delta_{B_1}(f_{popul}) - \varepsilon_{B_1}(f_{popul})$, where the structuring element B_1 is an unitary disk. The other required ingredient is the function providing the inner markers.

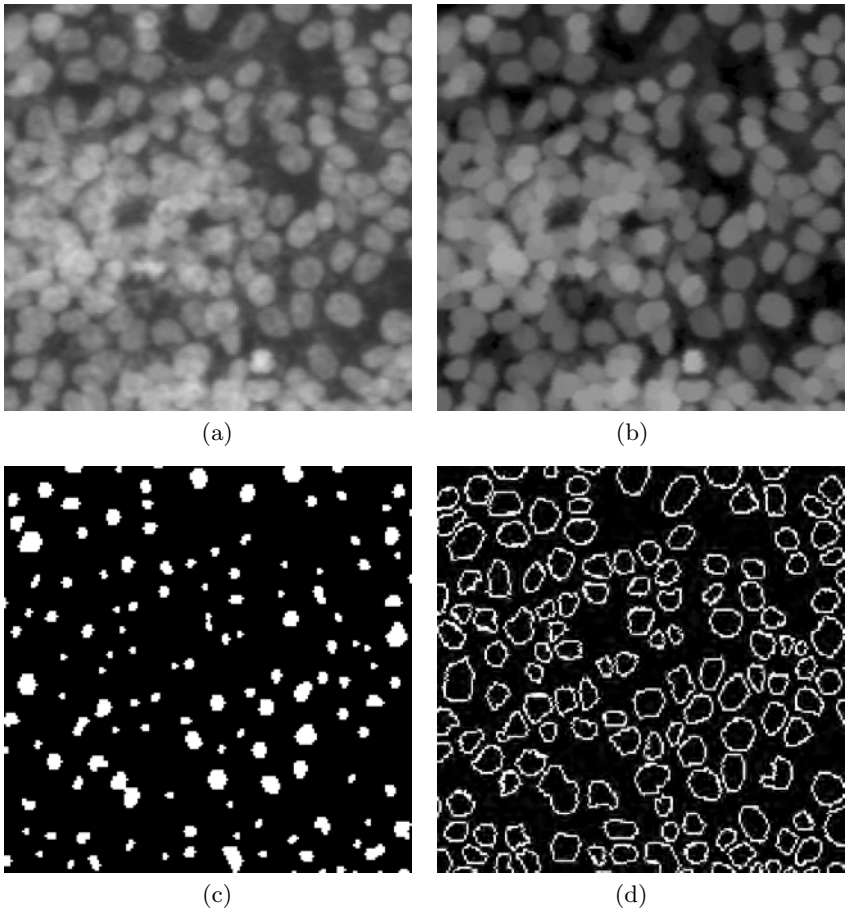


Fig. 1. Example of individual nucleus segmentation from a field Z of a fluorescence labelled cell population: (**a**) Original image f_{popul}, (**b**) Ultimate opening using hexagons Ult-$\gamma_B(f_{popul})$, (**c**) Image of regional maxima Max (Ult-$\gamma_B(f_{popul})$), (**d**) Watershed segmentation using the maxima as inner markers (the outer marker is the SKIZ of the inner markers) on the gradient of f_{popul}

As a first approximation, we consider that the cell are modelled by balls and consequently the support of the realization f_i^{cell} is a circular random set of radius r_i. According to their cell cycle phase, the nuclei have different radius but their distribution can be bounded in an appropriate interval. The size distribution of the image structures can be studied using the notion of granulometry [4]. A granulometry is a one-parameter family of openings $\Gamma = (\gamma_{B_n})_{n\geq 0}$ according to the structuring element (i.e., shape probe) B of size

n such that γ_{B_n} follows the absorption law; i.e., $\forall n \geq 0, \forall m \geq 0$, $\gamma_{B_n}\gamma_{B_m} = \gamma_{B_m}\gamma_{B_n} = \gamma_{B_{\max(n,m)}}$. The opening $\gamma_{B_n}(f) = \delta_{B_n}\varepsilon_{B_m}(f)$ is an increasing, anti-extensive and idempotent operator [7]. Based on the notion of granulometry, the ultimate opening Ult-γ_B operator has been recently introduced [2]. Let us consider the numerical residual operator associated to a discrete family of openings defined as follows

$$\text{Ult-}\gamma_B(f)(x) = \sup_{n_{\min} \leq k \leq n_{\max}} \left(\gamma_{B_k}(f)(x) - \gamma_{B_{k+1}}(f)(x)\right).$$

It replaces the initial image $f(x)$ by a union of the most significant cylinders included in the sub-graph of the initial function. A significant cylinder is the biggest and highest cylinder covering every point of the image. The application of the ultimate opening to the image f_{popul} allows to adjust a maximal cylinder for each cell of the clusters, Fig. 1b. The computation of the regional maxima of Ult-$\gamma_B(f_{popul})$ provides an appropriate inner marker for individual cells, $mrk_i(x) = \text{Max}\left(\text{Ult-}\gamma_B(f_{popul})\right)$, Fig. 1c. The outer markers mrk_o of the nuclei, which constrain the segmentation, are defined as the skeleton by influence zones [1,7] (i.e., the voronoï diagram) computed as the watershed of the distance function of the complement of inner markers image. Using both markers, $mrk(x) = mrk_i(x) \vee mrk_o(x)$, the application of the watershed lead to the final cell contours, Fig. 1c. Note that from a practical viewpoint, the balls used to approach the cells are a family of hexagons.

4 Stochastic Nucleus Counting

As we can observe from the example of Fig. 1, the algorithm described above segments properly well separated cells and most of cells in the clusters. However, to avoid obtaining clusters of various cells instead of individual cells, our approach loses some nuclei. We propose now a second alternative, which involves to *relax* the watershed segmentation, replacing the SKIZ by the image border as outer marker. The associated segmentation produces clusters of nuclei and we consider by hypothesis that the cell population is determined by the union of the detected clusters: $C_{popul} = \cup_{j \in J} C_{cluster,j}$. A full quantification involves an algorithm for counting cells in each segmented cluster $C_{cluster,j}$.

Let us start by the following classical theorem [4] from the theory of Boolean Random Closed Sets:

$$\Pr\{B \subset X^c\} = e^{-\theta A(X' \oplus B)}$$

which characterises the probability that the compact set B is contained in the pores X^c, where $A(X' \oplus B)$ is the average surface area of the primary grain X' dilated by the set B. In particular, if B is reduced to a single point, $\Pr\{B \subset X^c\}$ becomes the porosity q (proportion of pores) and the relation is:

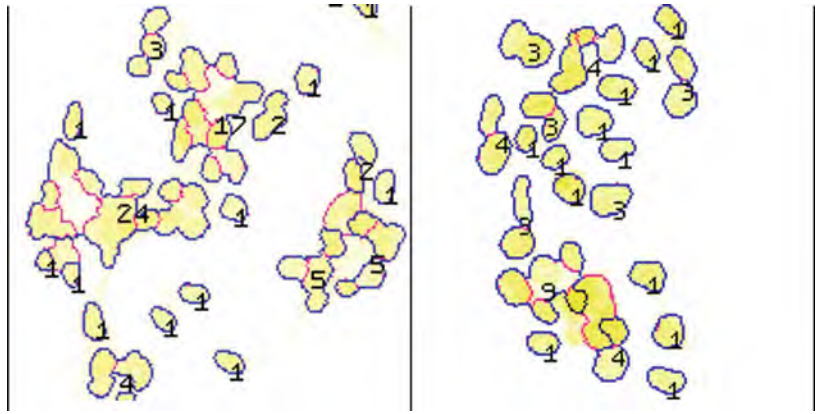


Fig. 2. Two examples of quantified population of nuclei (*in yellow*). The number close to each cluster indicated the counted nuclei by the Boolean formula: N_{Z_j}

$$q = e^{-\theta A(X')} \Leftrightarrow \theta = -\frac{\log q}{A(X')}$$

Even when the structure is not Boolean, the Central Limit Theorem suggests to use this result a priori.

Considering our problem, a probabilistic algorithm for counting partly covering nuclei in the binary set of cluster j, $C_{cluster,j}$, is given by the following formula

$$N_{Z_j} = \{\text{number of nuclei in } Z_j\} = -\frac{\mid Z_j \mid}{\overline{C^{cell}}} \, log(q)$$

where q is the porosity of set $C_{cluster,j}$ (i.e., $q = A(C^c_{cluster,j})$), $\mid Z_j \mid$ is the area of the field Z_j under study and $\overline{C^{cell}}$ mean area of individual nucleus. The equation is valid specifically if θ is constant in each cluster j. To better match the Boolean model, the image field Z_j of each cluster is the bounding box containing the set $C_{cluster,j}$. The mean area of an individual nucleus is estimated from some isolated nuclei from the population. In fact, this value can be learned and fixed from representative segmented cells of several populations. Figure 2 provides two examples of populations of segmented clusters counted by the Boolean formula.

5 Conclusions and Perspectives

A full automated segmentation algorithm for clustered nuclei in fluorescence labelled images has been presented. Any parameter is required since the application of a granulometry is able to adaptively identify each region candidate to be a nucleus, which is then segmented by watershed algorithm. In fact, the single prior datum is the shape used for the size distribution, a circle in our

case. For other cells presenting a more elongated nuclear shape, the ultimate opening can be implemented using families of ellipses of variable orientation and eccentricity, which will lead to a better nucleus adjustment.

A probabilistic algorithm for counting the number of nuclei in a cluster has been also presented. From our results, we state that the number of nuclei obtained by the Boolean model is more robust than a simple ratio of surfaces. It can be used to verify the appropriateness of the segmentation for each cluster and eventually, to detect the wrong segmented cluster.

The result of the ultimate opening, see Fig. 1b, produces a random function which describes each cell by a cylinder such as $f_i^{cell}(x) = t_i$ if $x \in C_i^{cell}$, otherwise $f_i^{cell}(x) = 0$, where t_i is the fluorescence intensity of nucleus i. Indeed, we expect to introduce in forthcoming research a direct cell modelling and counting, without passing by a binary image, using the theory of Boolean functions [8]. However, this application need a more deep modelling of scalar nuclei images, including the study of variation of the florescence intensity which seems be dependent on the DNA nucleus status but also on the effect of cell aggregation. The present algorithms are suitable for static cell culture images. Spatial modelling which includes the time dimension should be necessary for analysis of cell culture kinetics using time-lapse images. The challenging issue is to model how are formed the clusters and how do they evolved in time.

References

1. Beucher, S., Meyer, F.: The morphological approach to segmentation: the watershed transformation. In: Dougherty, E. (ed.) Mathematical Morphology in Image Processing, pp. 433–481. Marcel Dekker, New York (1992)
2. Beucher, S.: Numerical residues. Image Vis. Comput. **25**, 405–415 (2007)
3. Carpenter, A.E., Jones, T.R., Lamprecht M.R., et al.: CellProfiler: image analysis software for identifying and quantifying cell phenotypes. Genome Biol. **7**, R100 (2006)
4. Matheron, G.: Random sets and integral geometry. Wiley, New York (1975)
5. Neumann, B., Held, M., Liebel, U., et al.: High-througthput RNAi screening by time-lapse imaging of live human cells. Nat. Methods **3**, 385–390 (2006)
6. Nilsson, B., Heyden, A.: Segmentation of complex cell clusters in microscopic images: application to bone marrow samples. Cytometry A. **66**(1), 24–31 (2005)
7. Serra, J.: Image Analysis and Mathematical Morphology, vol. 1. Academic, London (1992)
8. Serra, J.: Boolean random functions. J. Microsc. **156**(1), 41–63 (1989)

Spatio-Temporal Segmentation for Radiotherapy Planning

Jean Stawiaski[1], Etienne Decencière[1], and François Bidault[2]

[1] Mines ParisTech, Mathématiques et Systèmes, Centre de Morphologie Mathématique, Fontainebleau, France, jean.stawiaski@ensmp.fr, Etienne.Deceniere@ensmp.fr
[2] Institut Gustave Roussy, Villejuif, France, bidault@igr.fr

Summary. This paper presents a segmentation method of 3D time-series images for radiotherapy planning. The aim of this study is to propose some techniques for the segmentation of tumors surrounding or contained in the lungs. The 4D images are produced using a respiration gating procedure and computed tomography. The aim of the segmentation is to follow the tumor movement while the patient is breathing, so that he does not need to hold his respiration during the radiation treatment. The proposed technique is based on mathematical morphology and graph cuts. It uses a 4D watershed algorithm, combined with graph-based techniques to delineate the tumors in the time-series. The differences between different classical spatio-temporal segmentation algorithms will be highlighted, and conclusions on the related trade-offs between speed and precision will be drawn.

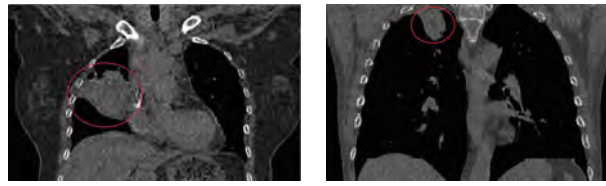
1 Introduction

Lung cancer is a disease of uncontrolled cell growth in the lungs. This growth may lead to metastasis, invasion of adjacent tissues and infiltration beyond the lungs. Lung cancer is the most common cause of cancer-related death in men and the second most common in women. This disease is responsible for 1.3 million deaths worldwide annually.[1] Early detection of lung tumors remains a challenging task of medical imaging. Radiotherapy and surgery treatments of lung cancer are highly difficult and risky tasks due to the proximity of the heart and numerous blood vessels. The presence of blood vessels leads to an important risk of dissemination of tumoral cells in the whole body of the patient which causes the apparition of multiple tumors in various locations of the body. An early and accurate treatment of the lung cancer is thus necessary to give a better chance of healing to the patient.

Selecting the best treatment for lung cancer depends on the clinician being able to identify the precise borders of the tumors. Moreover the detection of

[1] World Health Organization 2006.

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*, Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_31,



(a) (b)

Fig. 1. Lungs 3D CT images. (**a**, **b**) The tumors are indicated by light *circles*. The tumors present low contrasted boundaries with the surrounding tissues of the thorax

the tumors, in classical computed tomography (CT) images, is limited by the breathing motion of the patient. In this scenario, spatio-temporal data (time-series images) can be advantageously used to optimize the radiotherapy treatment. We give in this paper some techniques for the segmentation of CT times-series images to detect and track tumor borders during the breathing motion. Our methods are based on the computation of a minimal graph cut in the region adjacency graph of an unsupervised watershed transform [2,3,6,7]. In this scenario, the user has to roughly specify the location of the tumors and the surrounding tissues. Our strategy aims to compute a minimal surface separating the user defined markers and lying on the tumor boundaries. The minimal surface method remains a leading technique for the segmentation of low contrasted structures such as lung tumors as illustrated in Fig. 1.

2 General Description

We propose two strategies for the segmentation of tumors in 4D volumes. The first one is a direct extension of the classical techniques used to segment 3D volumes. We propose to build a 4D volume by concatenating all the 3D images. Since watershed segmentation [2] can be computed on a gradient image of any dimension, the method requires only the computation of a spatio-temporal gradient image. The user has then to specify some markers on a 3D volume of the time-series. A watershed segmentation is then computed according to the spatio-temporal gradient and finally a minimal surface is extracted from the region adjacency graph of the 4D watershed transform [6]. The main advantage of this first strategy is that the whole 4D volume is segmented in a single step and large motions are allowed by this procedure. However this method requires a huge amount of memory since the whole 4D volume has to be stored to compute the watershed transform and a large graph has to be stored to compute the minimal surface representing the evolution of the tumor boundaries. This first method is thus unusable on a classical personal computer.

The second protocol is slightly different and aims at providing a faster segmentation algorithm. We consider in this scenario a sequential segmentation

of 3D images. The user has to provide markers of the tumor and healthy tissues on the first image of the time-series. The result of the segmentation at time t is then used to produce markers for the segmentation at time $t+1$. Since the motion of the thorax is relatively small, the new markers can be easily obtained from the previous segmentation. In our approach, we eroded each region of the segmentation to ensure that the resulting eroded image can be used as markers for the next image. This procedure works because tumors are mainly compact objects. For thin objects segmentation, this procedure would fail because the erosion step will delete the thin structures and no markers could be extracted from the first segmentation. The main advantage of this protocol is that the method does not need more memory than a classical 3D segmentation technique since it is based on sequential segmentation of 3D images. However this method does not allow large motion and does not allow to segment thin structures. This second method has been chosen for our experiments.

3 Approximate Minimal Surfaces for Tumor Segmentation

We detail now how to extract the tumor boundaries by an approximate minimal surface using a region adjacency graph [6]. The combination of graph-cuts with a watershed low-level segmentation provides us with an explicit and efficient way to compute approximate minimal surfaces. Our basic assumption is that the minimal surface to be computed is embedded in the watershed low-level segmentation contours. We propose thus to solve the following combinatorial problem: finding a surface composed of a finite union of watershed contours such that the surface minimizes a given geometric functional. We solve this problem by using graph-cuts optimization on a region adjacency graph.

Following the formulation of Caselles et al. [4], we want to find a surface S defined by a finite union of watershed contours that minimizes the following energy function:

$$E(S) = \int\int_S g(\|\nabla I(x,y)\|)dxdy \tag{1}$$

where g is a positive and strictly decreasing function and $\|\nabla I(x,y)\|$ is the modulus of the gradient of the image I (image contrast). Note that Cauchy–Crofton formulae can be used to minimize the energy function $E(S)$ by computing a minimal graph cut as described by Boykov et al. in [3].

Let us consider $G = (V, E, W)$ as the pixel graph of an image I. Classically V is the set of nodes and represents the pixels of I, E is the set of edges representing neighborhood relations between pixels and W is a positive weight assigned to each edge of E. In our terminology, an edge linking two nodes i and

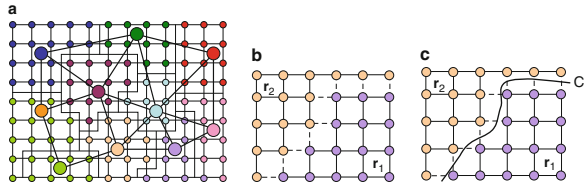


Fig. 2. (**a**) A region adjacency graph. (**b**) The set of nodes of the pixel graph considered to compute boundary properties between two regions, with a V_4 adjacency system. (**c**) A curve crossing the edges of the boundary between two regions r_1 and r_2

j is written $e_{i,j}$ and the corresponding edge weight is denoted by $w_{i,j}$. From the pixel graph, we define the region adjacency graph $G_R = (V_R, E_R, W_R)$ of the watershed transform where V_R is the set of nodes (i.e. the regions of the watershed transform), E_R is the set of edges (i.e. the neighborhood relation between regions) and W_R is the weights of the edges.

Let us define $F_{(r_i,r_j)}$ as the set of edges of the pixel graph connecting two regions r_i and r_j of the low-level watershed segmentation:

$$F_{(r_i,r_j)} = \{e_{m,n} \in E \mid m \in r_i,\ n \in r_j\} \ . \tag{2}$$

Note that the set $F_{(r_i,r_j)}$ depends on the adjacency system of the pixel graph G. The set of edges of the pixel graph describes also implicitly a set of surfaces between the regions r_i and r_j as illustrated in Fig. 2. Let $S_{(r_i,r_j)}$ denote the set of surfaces that could cross the edges of $F_{(r_i,r_j)}$. Following Cauchy–Crofton formulas with the V_6 adjacency system, the energy function $E(S_{(r_i,r_j)})$ can be approximated by:

$$E(S_{(r_i,r_j)}) \approx \sum_{(e_{m,n} \in F_{(r_i,r_j)})} g(max\left(\|\nabla I(m)\|, \|\nabla I(n)\|\right)) \ , \tag{3}$$

where $\|\nabla I(m)\|$ and $\|\nabla I(n)\|$ are the gradient magnitudes of the end points of $e_{m,n}$. In the following, we consider the strictly positive and decreasing function g:

$$g(\|\nabla I(p)\|) = \left(\frac{1}{1 + \|\nabla I(p)\|}\right)^k \ . \tag{4}$$

The parameter $k \in R^+$ is a free parameter that can be used as a smoothing term as shown by Allène et al. in [1]. In our application this parameter was set to $k = 2$. The function g works as an edge indicator of the image I and takes a small value if neighbors pixels m and n take different grey values p_m and p_n. The energy $E(S_{(r_i,r_j)})$ of the boundary between two regions is simply

Table 1. Edge weights for approximate minimal surfaces

Edge	Weight	For
w_{s,r_i}	$+\infty$	$r_i \in M_t$
$w_{r_i,t}$	$+\infty$	$r_i \in M_h$
w_{r_i,r_j}	$E(S_{(r_i,r_j)})$	$r_i \in V_R, r_j \in N_{r_i}$

obtained by summing the local contrasts along the boundaries between two regions.

The tumor boundaries are finally extracted by computing a minimal graph cut of the region adjacency graph with weights given by Table 1. The minimal cut is computed on the region adjacency graph with two additional nodes s and t, respectively connected to the markers of the liver and the markers of the external tissues. In the following, we denote the markers that specify the healthy and tumoral tissues as the set of regions M_h and M_t. Note also that additional markers can be used. In this last case a multi-terminal cut algorithm [5] is used to compute a set of multiple minimal surfaces separating each pair of markers. This last technique can be used to segment the lungs, the tumor as well as the body of the patient in a single step.

4 Results

Figure 3 illustrates a segmentation result on a 4D CT image. The second segmentation protocol was used to obtain the presented results. The segmentation was obtained by using a multi-terminal cut algorithm. The user has provided markers for the lungs, the tumor, as well as the surrounding tissues. The tumor and the lungs have been correctly delineated by using this strategy since the motion of the lungs is especially small on the upper part of the lungs, where the tumor is located.

5 Conclusion

Minimal surfaces computed on the region adjacency graph provide stable and robust segmentation results for the aimed application, the delineation of lung tumors. The method is sufficiently fast and precise to be used on large data-sets such as 4D images. In such conditions the analysis of all data-sets by a radiologist cannot be realized by manual segmentation. The proposed method is thus a first solution for the tracking of lung tumors. With the exponential growth of medical image data, it is clear that such interactive methods are good alternatives to fully manual segmentations. Up to now, fully automatic methods also fail to achieve relevant segmentation results in all the cases and often require manual corrections.

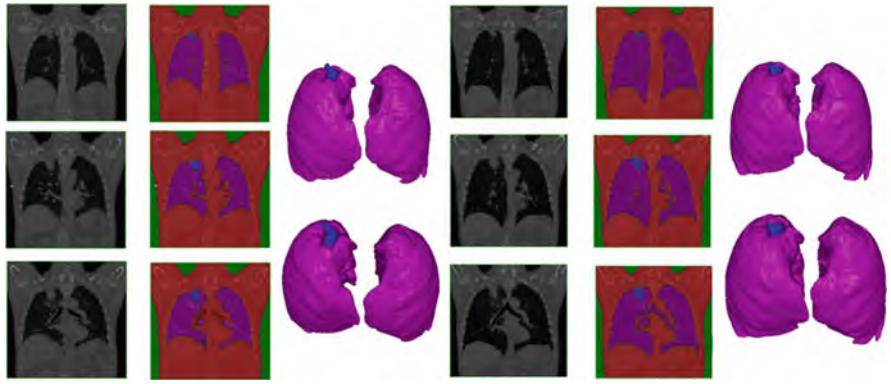


Fig. 3. 4D CT images segmentation results at two different steps of a breathing cycle. The segmentation was obtained with multi-terminal cuts

Acknowledgments

We would like to thank Région Ile-de-France and the Cancéropôle Ile-de-France for funding our research on medical image segmentation.

References

1. Allène, C., Audibert, J., Couprie, M., Cousty, J., Keriven, R.: Some links between min-cuts, optimal spanning forests and watersheds. Proceedings 8th International Symposium on Mathematical Morphology, vol. 1, pp. 253–264. (2007)
2. Beucher, S., Meyer, F.: The morphological approach to segmentation: the watershed transformation. In: Dougherty, E.R. (ed.) Mathematical Morphology in Image Processing, pp. 433–481. (1993)
3. Boykov, Y., Kolmogorov, V.: Computing geodesics and minimal surfaces via graph cuts. Proceedings of the 9th International Conference on Computer Vision, Nice, France, vol. 1, pp. 26–33. (2003)
4. Caselles, V., Kimmel, R., Sapiro, G.: Geodesic active contours. Int. J. Comput. Vis. **22**, 61–79 (1997)
5. Dahlhaus, E., Johnson, D.S., Papadimitriou, C.H., Seymour, P.D., Yannakakis, M.: The complexity of multiterminal cuts. SIAM J. Comput. **23**, 864–894 (1994)
6. Stawiaski, J., Decencière, E.: Computing approximate geodesics and minimal surfaces using watershed and graph-cuts. Proceedings of the The 8th International Symposium on Mathematical Morphology, Rio de Janeiro, Brazil, vol. 1, pp. 349–360. (2007)
7. Stawiaski, J., Decencière, E.: Combining graph-cuts and morphological segmentation. Image Anal. Stereol. **27**(1), 39–46 (2008)

Tracking and Registration for Multidimensional Biomedical Image Analysis

K. Rohr, W.J. Godinez, N. Harder, S. Yang, I.-H. Kim, S. Wörz, and R. Eils

University of Heidelberg, BIOQUANT, IPMB, and German Cancer Research Center (DKFZ), Dept. Bioinformatics and Functional Genomics, Biomedical Computer Vision Group (BMCV),
Im Neuenheimer Feld 267, 69120 Heidelberg, Germany
k.rohr@dkfz.de, w.godineznavarro@dkfz.de, n.harder@dkfz.de, s.yang@dkfz.de, i.kim@dkfz.de, s.woerz@dkfz.de, r.eils@dkfz.de

Summary. Tracking and registration approaches have been developed for automatic analysis of multidimensional biomedical images. The tracking approach allows computing the trajectories of cells in fluorescence microscopy image sequences. The registration approach enables to geometrically align cell microscopy images by using elastic transformations.

1 Introduction

The analysis of high-content multidimensional biomedical images requires the application of different techniques. In this contribution, we describe approaches for tracking and registration, which are central tasks in image analysis and have a wide spectrum of applications in the fields of biology and medicine.

In cell biology, an important application is the analysis of the motion of cellular structures. We have developed a tracking approach to determine the trajectories of cells from fluorescence microscopy image sequences. The approach can cope with splitting cells, which is important in our application. In conjunction with segmentation and classification schemes the aim is to study the influence of genes on the process of cell division and thus to identify gene function. In addition, we have introduced approaches for tracking virus particles. We have also developed registration methods which enable the geometric alignment of corresponding image data. Our approaches can cope with elastic (non-rigid) deformations between images and have been applied for the registration of cell microscopy images and gel electrophoresis images.

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*,
Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_32,

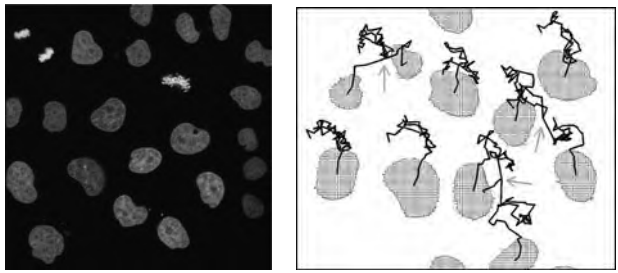


Fig. 1. Original cell microscopy image (*left*) and example of a tracking result with indicated splitting events (*right*)

2 Tracking of Cellular Structures in Fluorescence Microscopy Image Sequences

2.1 Tracking of Cells

Tracking is a central task in biomedical image analysis and allows analyzing the movement of objects (e.g., [11]). We have developed a tracking approach which determines the trajectories of cells from 2D and 3D multi-cell time-lapse images generated by high-throughput RNA interference (RNAi) experiments (Harder et al. [5,6]). RNAi is an effective technology to identify the biological function of genes by systematically knocking down genes and analyzing the resulting phenotypes [2]. The general aim of our work is to understand the process of cell division (mitosis) at a molecular level. To study the influence of genes on cell division we quantify the duration of cell cycle phases to determine whether the knockdown of a certain gene leads to a delay of certain phases. To this end we combine the tracking approach with segmentation and classification schemes.

Given 2D and 3D fluorescence microscopy image sequences of live cells, we first segment cell nuclei by an efficient region-adaptive thresholding scheme. This scheme computes local intensity thresholds in overlapping image regions using Otsu's histogram-based threshold selection scheme. Based on the segmented objects we use a tracking scheme which determines the temporal connections between cells and can handle splitting objects. We have developed the following two-step tracking approach: First, initial, non-splitting trajectories are established, and second, mitotic events are detected and the related trajectories are merged. In the first step, the initial trajectories are determined using a feature point tracking algorithm based on [1]. As feature points we use the centers of gravity of segmented cell nuclei. For each frame of an image sequence the algorithm considers the predecessor and the successor frame. In these frames, object correspondences are determined by searching for trajectories with maximum smoothness. For one feature point we determine all potential predecessor and successor feature points within a certain Euclidean distance and compute the smoothness of trajectories based on a

cost function. The cost function takes into account the angle defined by successive feature points as well as the distance between the points. Changes in direction and distance cause higher costs. In the second step, we detect mitosis events and merge related trajectories. All trajectories that do not start in the first frame are taken into account as possible mitosis events. To determine whether a mitosis event exists we exploit the distance and size of potential parent and child objects. The result of tracking are tree-structured trajectories which represent the ancestral relationships between cells (Fig. 1).

After tracking we classify cells based on a support vector machine (SVM) classifier [14] using both static and dynamic image features. Our approach distinguishes between seven cell cycle phases (interphase, prophase, prometaphase, metaphase, early anaphase, late anaphase, and telophase) and automatically determines the duration of these phases. In our approach we make use of a priori knowledge about the sequence of cell cycle phases represented by a finite state machine. The inclusion of this knowledge improves the classification and quantification result.

2.2 Tracking of Virus Particles

Besides tracking of cells we have also been working on analyzing the movement of virus particles. Viruses are much smaller than cells and appear as spots in light microscopy images. To automatically track multiple virus particles in time-lapse fluorescence microscopy images we have developed deterministic and probabilistic approaches. Whereas the deterministic approaches rely on a two-step paradigm comprising virus localization and correspondence finding, the probabilistic approaches are based on a Bayesian paradigm and formulate tracking as a sequential estimation problem.

For the probabilistic approaches, we assume that a virus particle is represented by a state vector $\mathbf{x}_t$, and that a noisy measurement $\mathbf{y}_t$ reflects the true state of $\mathbf{x}_t$. At time step t, the aim is to estimate the state $\mathbf{x}_t$ of a virus given a sequence of measurements $\mathbf{y}_{1:t}$. By modeling the temporal evolution using a *dynamical model* $p(\mathbf{x}_t|\mathbf{x}_{t-1})$ and incorporating measurements derived from the images via a *measurement model* $p(\mathbf{y}_t|\mathbf{x}_t)$, a Bayesian filter estimates the *posterior* distribution $p(\mathbf{x}_t|\mathbf{y}_{1:t})$ via stochastic propagation and Bayes' theorem:

$$p(\mathbf{x}_t|\mathbf{y}_{1:t}) \propto p(\mathbf{y}_t|\mathbf{x}_t) \int p(\mathbf{x}_t|\mathbf{x}_{t-1})\, p(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1})\, d\mathbf{x}_{t-1}.$$

An estimate of $\mathbf{x}_t$ can be obtained from the posterior $p(\mathbf{x}_t|\mathbf{y}_{1:t})$, which, in our case, is estimated using a particle filter [8]. The idea of this filter is to approximate the posterior via a set $\{\mathbf{x}_t^i; w_t^i\}_{i=1}^{N_s}$ of N_s random samples $\mathbf{x}_t^i$ (the 'particles') that are associated with importance weights w_t^i. In our case, we have developed tracking approaches based on a mixture of particle filters and based on independent particle filters (Godinez et al. [3,4]). We have successfully applied the approaches to multichannel microscopy images of HIV-1 particles and have quantified the performance based on ground truth from

manual tracking. It turned out that the probabilistic approaches outperform the deterministic schemes.

3 Non-Rigid Registration of Cell Microscopy Images

Another central task in biomedical image analysis is the normalization of image data, for example, the normalization of cell microscopy images for subsequent quantification and statistical analysis. To this end a geometric transformation needs to be computed. The task of finding an optimal geometric transformation between corresponding image data is known as image registration, and generally one has to use non-rigid or elastic deformation models which allow coping with local shape differences (e.g., [7, 12, 17]). To register 2D and 3D fluorescence microscopy images of different cell nuclei we have developed an elastic registration approach which relies on optic flow estimation (Yang et al. [16], Kim et al. [9]). This approach is based on the scheme in [13] and is driven by symmetric forces. We use either segmented images or directly exploit the image intensities. Assume that f and g are the source and target images, respectively, for which we want to compute the deformation vector field $\mathbf{U}$. With $f(\mathbf{x})$ and $g(\mathbf{x})$ denoting the intensity values at position $\mathbf{x}$, the instantaneous deformation vector field at iteration k can be written as:

$$d\mathbf{U}_k(\mathbf{x}) = \frac{2[f(\mathbf{x}) - g(\mathbf{u}_{k-1}(\mathbf{x}))][\nabla f(\mathbf{x}) + \nabla g(\mathbf{u}_{k-1}(\mathbf{x}))]}{p(\mathbf{x})}$$

where $p(\mathbf{x}) = [\nabla f(\mathbf{x}) + \nabla g(\mathbf{u}_{k-1}(\mathbf{x}))]^2 + [f(\mathbf{x}) - g(\mathbf{u}_{k-1}(\mathbf{x}))]^2$ and ∇ denotes the nabla operator. The equation is computed if $p(\mathbf{x}) \geq \epsilon$, where ϵ is a small positive constant, which is used to prevent cases where the denominator $p(\mathbf{x})$ is close to zero. If $p(\mathbf{x}) < \epsilon$, we set $d\mathbf{U}_k(\mathbf{x}) = 0$. Furthermore, $\mathbf{u}_k(\mathbf{x}) = \mathbf{x} + \mathbf{U}_k(\mathbf{x})$ is the transformed position $\mathbf{x}$ with the total deformation field, which is finally used to transform the source image, and $\mathbf{U}_k(\mathbf{x}) = \mathbf{U}_{k-1}(\mathbf{x}) + d\mathbf{U}_k(\mathbf{x})$, $\mathbf{U}_0(\mathbf{x}) = \mathbf{0}$. With this optic flow-based approach the deformation between two images is computed based on intensity differences and the gradients of the images. Our experimental results showed that this approach using symmetric forces yields better results compared to the standard approach, where the forces are not symmetric. To speed up the computation, we have proposed an adaptive step length optimization scheme and also employ a multi-resolution scheme.

The approach has been successfully applied to multi-channel 3D confocal images of different cells for shape normalization and analysis of the 3D structure of cells and chromosomes (Fig. 2). We have also investigated the registration of dynamic cell microscopy images, i.e., 4D (3D+t) images of moving cell nuclei. In this case, the task is the normalization of the shape of moving cells over time to decouple the movement and deformation of cells from the movement of protein particles for accurate determination of particle motion.

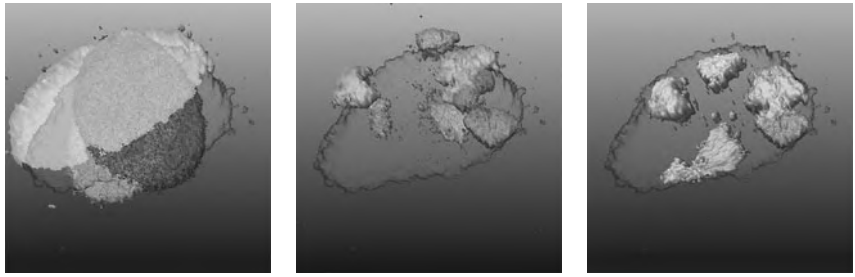


Fig. 2. Four different nuclei of HeLa cells overlaid using different *gray* tones (*left*), corresponding chromosome pairs without registration (*middle*), and after elastic registration (*right*)

Besides optic flow-based registration schemes, we have also developed spline-based registration approaches. In particular, we have introduced an elastic registration approach, which relies on analytic solutions of the Navier equation under Gaussian forces. The corresponding splines have been termed Gaussian elastic body splines (Wörz and Rohr [15], Kohlrausch et al. [10]). The approach has been successfully applied to register gel electrophoresis images and medical tomographic images.

4 Conclusion

We have described tracking and registration approaches for automatic analysis of multidimensional biomedical images. The tracking approaches allow analyzing the movement of cellular structures, and the registration approaches enable to geometrically normalize the image data for subsequent accurate quantification.

Acknowledgement

Support of the EU project MitoCheck, the BMBF FORSYS project ViroQuant, the EU project 3DGenome, and the DFG project ELASTIR is gratefully acknowledged.

References

1. Chetverikov, D., Verestoy, J.: Tracking feature points: a new algorithm. Proceedings of the 14th International Conference on Pattern Recognition, pp. 1436–1438. Brisbane, Qld., Australia, Aug. 1998
2. Friedman, A., Perrimon, N.: Genome-wide high-throughput screens in functional genomics. Curr. Opin. Genet. Dev. **14**, 470–476 (2004)

3. Godinez, W.J., Lampe, M., Wörz, S., Müller, B., Eils, R., Rohr, K.: Tracking of virus particles in time-lapse fluorescence microscopy image sequences. In: Fessler, J., Denney, T. (eds.) Proceedings of the ISBI'07. Arlington, VA, USA, 12–15 April 2007, pp. 256–259
4. Godinez, W.J., Lampe, M., Wörz, S., Müller, B., Eils, R., Rohr, K.: Probabilistic tracking of virus particles in fluorescence microscopy images. Proceedings of the ISBI'08, Paris, France, 14–17 May 2008
5. Harder, N., Mora-Bermúdez, F., Godinez, W.J., Ellenberg, J., Eils, R., Rohr, K.: Automated analysis of the mitotic phases of human cells in 3D fluorescence microscopy image sequences. In: Larsen, R., Nielsen, M., Sporring, J. (eds.) Proceedings of the MICCAI'06, Copenhagen, Denmark, Oct. 1–6, 2006. Lecture Notes in Computer Science 4190, Part I, pp. 840–848. Springer, Berlin (2006)
6. Harder, N., Mora-Bermúdez, F., Godinez, W.J., Ellenberg, J., Eils, R., Rohr, K.: Determination of mitotic delays in 3D fluorescence microscopy images of human cells using an error-correcting finite state machine. In: Fessler, J., Denney, T. (eds.) Proceedings of the ISBI'07, Arlington, VA, USA, 12–15 April 2007 , 1044–1047
7. Holden, M.: A review of geometric transformations for nonrigid body registration. IEEE Trans. Med. Imaging **27**, 111–128 (2008)
8. Isard, M., Blake, A.: CONDENSATION – conditional density propagation for visual tracking. Int. J. Comput. Vis. **29**(1), 5–28 (1998)
9. Kim, I.-H., Yang, S., Le Baccon, P., Heard, E., Chen, Y.-C., Spector, D., Kappel, C., Eils, R., Rohr, K.: Non-rigid temporal registration of 2D and 3D multi-channel microscopy image sequences of Human Cells. In: Fessler, J., Denney, T. (eds.) Proceedings of the ISBI'07, Arlington, VA, USA, 12–15 April 2007, 1328–1331
10. Kohlrausch, J., Rohr, K., Stiehl, H.S.: A new class of elastic body splines for nonrigid registration of medical images. J. Math. Imaging Vis. **23**(3), 253–280 (2005)
11. Meijering, E., Smal, I., Danuser, G.: Tracking in molecular bioimaging. IEEE Signal Process. Mag. **23**(3), 46–53 (2006)
12. Rohr, K.: Landmark-Based Image Analysis: Using Geometric and Intensity Models. Kluwer, Dordrecht (2001)
13. Thirion, J.P.: Image matching as a diffusion process: an analogy with Maxwells demons. Med. Image Anal. **2**(3), 243–260 (1998)
14. Vapnik, V.: Statistical Learning Theory. Wiley, New York (1998)
15. Wörz, S., Rohr, K.: Physics-based elastic registration using non-radial basis functions and including landmark localization uncertainties. Comput. Vis. Image Underst. **111**, 263–274 (2008)
16. Yang, S., Köhler, D., Teller, K., Cremer, T., Le Baccon, P., Heard, E., Eils, R., Rohr, K.: Non-rigid registration of 3D multi-channel microscopy images of cell nuclei. IEEE Trans. Image Process. **17**(4), 493–499 (2008)
17. Zitova, B., Flusser, J.: Image registration methods: a survey. Image Vis. Comput. **24**, 977–1000 (2003)

Patches of Finite Elements for Singularly-Perturbed Diffusion Reaction Equations with Discontinuous Coefficients

Massimiliano Culpo[1], Carlo de Falco[2], and Eugene O'Riordan[3]

[1] Bergische Universität Wuppertal, Gaußstr. 20, Wuppertal, Germany
culpo@math.uni-wuppertal.de
[2] Dublin City University, Glasnevin, Dublin 9, Ireland
carlo.defalco@dcu.ie
[3] Dublin City University, Glasnevin, Dublin 9, Ireland
eugene.oriordan@dcu.ie

Summary. We present a numerical method for solving Diffusion Reaction equations on two completely overlapping unstructured meshes which reduces the requirements on mesh generation software when strong local refinement is needed to capture features of the solution that appear on different scales.

1 Introduction

Singularly perturbed Diffusion Reaction equations with non-smooth coefficients can exhibit thin internal and boundary layers where the solution varies rapidly. Layer resolving methods for problems in one dimension are either based on fitted difference operators or on fitted meshes. Among those of the latter class, methods based on Shishkin-type meshes [2] are especially attractive because of their simplicity. Their applicability to multidimensional problems is, though, constrained by the difficulty of generating structured conforming meshes for general domain geometries. We present a numerical method for solving Diffusion Reaction equations on two completely overlapping unstructured meshes which reduces the requirements on mesh generation software when strong local refinement is required to capture features of the solution that appear on different scales. To validate the proposed method we present numerical results on a problem which can be seen as a 2d extension of the problem derived in [1] to compute the capacitance of a Metal Oxide Semiconductor (MOS) structure. In [1] it was shown, in the 1d case, that a parameter fitted mesh gives a significant improvement in the accuracy of the computed capacitance over a uniform mesh when the MOS bias is near the threshold voltage. The results shown in Sect. 4 indicate that the method

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*,
Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_33,

described below could be beneficial when dealing with modern MOS structures of non trivial geometry for which a 1d model is not appropriate [4].

2 Continuous Problem

Let $\Omega \subset \mathbb{R}^d$, $d = 1, 2$ be a bounded open domain. Let Ω be partitioned into two subdomains Ω_1, Ω_2 s.t. $\Omega_1 \cap \Omega_2 \equiv \emptyset$, $\overline{\Omega_1 \cup \Omega_2} \equiv \overline{\Omega}$ and define $\Gamma := \Omega \backslash (\Omega_1 \cup \Omega_2)$. Let $\partial\Omega \equiv \Gamma_N \cup \Gamma_D$ with $\Gamma_N \cap \Gamma_D \equiv \emptyset$ and $\Gamma_D \cap \Omega_1 \not\equiv \emptyset \not\equiv \Gamma_D \cap \Omega_2$. We consider the problem

$$\begin{cases} -\text{div } \sigma(u) + ru = f & \text{in } \Omega \setminus \Gamma \\ \sigma(u) = \varepsilon\kappa_i \nabla u,\ \kappa_i > 0 & \text{in } \Omega_i,\ i = 1, 2 \\ u|_{\Gamma_D} = g_D, & \sigma(u) \cdot n|_{\Gamma_N} = 0 \end{cases} \tag{1}$$

$$r = \begin{cases} r_1 \geq \beta > 0 \text{ in } \Omega_1 \\ r_2 \equiv 0 \text{ in } \Omega_2 \end{cases} \qquad f = \begin{cases} f_1 \text{ in } \Omega_1 \\ f_2 \text{ in } \Omega_2 \end{cases}$$

with the constraint that u and the component of $\sigma(u)$ along the direction normal to Γ be continuous in all Ω. Here $\varepsilon > 0$ is a small perturbation parameter, n denotes the outward unit normal to $\partial\Omega$ and n_i the outward unit normal to $\partial\Omega_i$. For the sake of simplicity and to prevent the appearance of boundary layers occurring in u we impose the further boundary condition

$$g_D|_{\Gamma_D \cap \partial\Omega_1} = \left.\frac{f}{r}\right|_{\Gamma_D \cap \partial\Omega_1}$$

2.1 Solution Decomposition

We restate problem (1) in a multidomain formulation as follows

$$\begin{cases} -\text{div } \sigma(v_i) + r_i v_i = f_i & \text{in } \Omega_i,\ i = 1, 2 \\ v_i|_{\Gamma_D \cap \partial\Omega_i} = g_D|_{\Gamma_D \cap \partial\Omega_i},\ \sigma(v_i) \cdot n_i|_{\Gamma_N \cap \partial\Omega_i} = 0 \\ v_i|_\Gamma = \left.\dfrac{f_1}{r_1}\right|_\Gamma =: g_\Gamma \end{cases} \tag{2}$$

$$\begin{cases} -\text{div } \sigma(w_1) + r_1 w_1 = 0 & \text{in } \Omega_1 \\ w_1|_{\partial\Omega_1 \cap \Gamma_D} = 0,\ \sigma(w_1) \cdot n_p|_{\partial\Omega_p \cap \Gamma_N} = 0 \end{cases} \tag{3}$$

$$\begin{cases} -\text{div } \sigma(w_2) = 0 & \text{in } \Omega_2 \\ w_2|_{\partial\Omega_2 \cap \Gamma_D} = 0,\ \sigma(w_2) \cdot n_2|_{\partial\Omega_2 \cap \Gamma_N} = 0 \end{cases} \tag{4}$$

$$\begin{cases} \displaystyle\sum_{i=1,2} (\sigma(v_i) + \sigma(w_i)) \cdot n_i|_\Gamma = 0 \\ w_1|_\Gamma = w_2|_\Gamma =: w_\Gamma \end{cases} \tag{5}$$

The solution u of (1) is related to v_i, w_i by

$$u|_{\Omega_i} = v_i + w_i,\ i = 1, 2;\ u|_{\Gamma} = g_{\Gamma} + w_{\Gamma} \tag{6}$$

The case where $d = 1$ is of particular interest. In this case Ω reduces to an interval $(a, b) \subset \mathbb{R}$ and without loss of generality we can assume $a = 0, b = 1$. Furthermore Γ will be a single point in $\mathbb{R}$ s.t. $0 < \Gamma < 1$ and we can write $\Omega_1 \equiv (0, \Gamma)$, $\Omega_2 \equiv (\Gamma, 1)$. Finally $\partial\Omega \equiv \Gamma_D \equiv \{0, 1\}$, $\Gamma_N \equiv \emptyset$ and $r(0)g_D(0) = f(0)$. Following the arguments in [1] we can state the following

Lemma 1. *Let $d = 1$ and k be an integer satisfying $0 \leq k \leq 4$. Then the solution u of the problem (2)–(6) satisfies the pointwise bounds*

$$\begin{cases} \left|\dfrac{d^k v_1}{dx^k}(x)\right| \leq C + C\varepsilon^{1-\frac{k}{2}} e^{-(\Gamma - x)\sqrt{\beta/\varepsilon}}, \forall x \in \Omega_1 \\ \left|\dfrac{d^k v_2}{dx^k}(x)\right| \leq C, \qquad\qquad \forall x \in \Omega_2 \end{cases}$$

$$\begin{cases} \left|\dfrac{d^k w_1}{dx^k}(x)\right| \leq C\varepsilon^{-\frac{k}{2}} e^{-(\Gamma - x)\sqrt{\beta/\varepsilon}}, \forall x \in \Omega_1 \\ \left|\dfrac{d^k w_2}{dx^k}(x)\right| \leq C, \qquad\qquad \forall x \in \Omega_2 \end{cases}$$

where C is a constant independent of ε.

Note that the term w_1 in (6) is negligible at a distance to the left of Γ, which is proportional to $\sqrt{\varepsilon}$.

3 Finite Element Discretization

To construct a parameter-uniform numerical method [2], we introduce the quantity τ_ε, which represents the *width of the interior layer*. Define the *interior layer region* as the subdomain

$$\Omega_p(\tau_\varepsilon) := \left\{ x \in \Omega_1, |\min_{y \in \Gamma} |x - y| \leq \tau_\varepsilon \right\}$$

Following very closely the presentation of [3] we introduce the Galerkin/Finite Elements discretization of problem (2), (3), (4), (5) as follows.
Define the following spaces

$$\begin{cases} S_i \equiv \left\{ u \in H^1(\Omega_i) |\ u|_{\partial\Omega_i \cap \Gamma_D} = g|_{\partial\Omega_i \cap \Gamma_D},\ u|_{\Gamma} = g_{\Gamma} \right\} \\ \mathcal{V}_i \equiv \left\{ u \in H^1(\Omega_i) |\ u|_{(\partial\Omega_i \cap \Gamma_D) \cup \Gamma} = 0 \right\} \end{cases} \quad i = 1, 2$$

$$\begin{cases} Z_2 \equiv \left\{ u \in H^1(\Omega_2) |\ u|_{\partial\Omega_2 \cap \Gamma_D} = 0 \right\} \\ Z_p \equiv \left\{ u \in H^1(\Omega_p) |\ u|_{\partial\Omega_p \setminus \Gamma_N \setminus \Gamma} = 0 \right\} \\ \mathcal{V}_p \equiv \left\{ u \in H^1(\Omega_p) |\ u|_{\partial\Omega_p \setminus \Gamma_N} = 0 \right\} \end{cases} \tag{7}$$

Let $\mathcal{T}_i^h$ indicate a quasi-uniform, conforming triangulation of Ω_i and $S_i^h \subset S_i$, $i = 1, 2$, $\mathcal{V}_i^h \subset \mathcal{V}_i, i = 1, 2, p$ and $Z_i^h \subset Z_i, i = 2, p$ be continuous finite element spaces consisting of functions that are linear on each element of $\mathcal{T}_i^h$.
Away from Γ we do not require any correspondence between the nodes of $\mathcal{T}_p^h$ and those of $\mathcal{T}_1^h$, for sake of simplicity, though, we assume the triangulations $\mathcal{T}_i^h$ to be constructed in such a way that $\mathcal{T}_1^h \cup \mathcal{T}_2^h$ be a globally conforming triangulation of Ω and $\mathcal{T}_p^h \cup \mathcal{T}_2^h$ be a globally conforming triangulation of $\Omega_p \cup \Omega_2$, so that the mesh nodes located on the interface Γ are the same in all three meshes. A consequence of this assumption is that the traces on Γ of the functions in any of the sets Z_i^h and S_i^h all belong to the same space T^h; the functions in T^h are piece-wise linear functions defined on Γ.
The discretization of problem (2), (3), (4), (5) reads:
Find $v_i^h \in S_i^h, i = 1, 2;\ w_1^h \in Z_p^h;\ w_2^h \in Z_2^h;\ \Phi_{\Gamma,i}^h, \Psi_{\Gamma,i}^h \in T^h, i = 1, 2$
such that

$$\begin{cases} B_i(\nu, v_i^h) = L_i(\nu) \quad \forall \nu \text{ in } \mathcal{V}_i^h \\ B_i(\nu, v) = (\nabla\nu, \sigma(v))_{\Omega_i} + (\nu, rv)_{\Omega_i} - (\nu, \Phi_{\Gamma,i}^h)_\Gamma \qquad i = 1, 2 \\ L_i(\nu) = (\nu, f)_{\Omega_i} \end{cases} \tag{2'}$$

$$\begin{cases} A_p(\nu, w_1^h) = 0 \quad \forall \nu \text{ in } \mathcal{V}_p^h \\ A_p(\nu, w) = (\nabla\nu, \sigma(w))_{\Omega_p} + (\nu, rw)_{\Omega_p} - (\nu, \Psi_{\Gamma,1}^h)_\Gamma \end{cases} \tag{3'}$$

$$\begin{cases} A_2(\nu, w_2^h) = 0 \quad \forall \nu \text{ in } \mathcal{V}_2^h \\ A_2(\nu, w) = (\nabla\nu, \sigma(w))_{\Omega_2} - (\nu, \Psi_{\Gamma,2})_\Gamma \end{cases} \tag{4'}$$

$$\begin{cases} \sum\limits_{i=1,2} (\nu, \Phi_{\Gamma,i} + \Psi_{\Gamma,i})_\Gamma, \quad \forall \nu \in T^h \\ w_1^h\big|_\Gamma = w_2^h\big|_\Gamma \end{cases} \tag{5'}$$

$$u^h\big|_{\Omega_1 \setminus \Omega_p} = v_1^h\big|_{\Omega_1 \setminus \Omega_p};\ u^h\big|_{\Omega_2} = v_2^h\big|_{\Omega_2} + w_2^h\big|_{\Omega_2};\ u^h\big|_{\Omega_p} = v_1^h\big|_{\Omega_p} + w_1^h\big|_{\Omega_p} \tag{6'}$$

The resulting algebraic problem consists in solving the following sequence of linear systems

$$\begin{cases} B_{II}^1 \mathbf{v}_I^1 = \mathbf{f}_I^1 - B_{I\Gamma}^1 \mathbf{g}_\Gamma \\ B_{II}^2 \mathbf{v}_I^2 = \mathbf{f}_I^2 - B_{I\Gamma}^2 \mathbf{g}_\Gamma \end{cases}, \quad \begin{cases} \boldsymbol{\Phi}^1 = \mathbf{f}_\Gamma^1 - B_{\Gamma I}^1 \mathbf{v}_I^1 - B_{\Gamma\Gamma}^1 \mathbf{g}_\Gamma \\ \boldsymbol{\Phi}^2 = \mathbf{f}_\Gamma^2 - B_{\Gamma I}^2 \mathbf{v}_I^2 - B_{\Gamma\Gamma}^2 \mathbf{g}_\Gamma \end{cases} \tag{8}$$

$$\begin{bmatrix} A_{II}^p & A_{I\Gamma}^p & 0 \\ A_{\Gamma I}^p & A_{\Gamma\Gamma}^p + A_{\Gamma\Gamma}^2 & A_{\Gamma I}^2 \\ 0 & A_{I\Gamma}^2 & A_{II}^2 \end{bmatrix} \begin{bmatrix} \mathbf{w}_I^p \\ \mathbf{w}_\Gamma \\ \mathbf{w}_I^2 \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ -\left(\boldsymbol{\Phi}^1 + \boldsymbol{\Phi}^2\right) \\ \mathbf{0} \end{bmatrix} \tag{9}$$

Where the subscript I indicates the degrees of freedom relative to internal mesh nodes while the subscript Γ denotes degrees of freedom relative to the interface nodes. To complete the definition of the discretization algorithm we need to prescribe a formula for τ_ε. To this end, let us again focus our attention on the case $d = 1$ and assume $\Omega \equiv (0, 1)$ as in Sect. 2.1. In such a case, if the standard lumping technique is adopted for the matrices corresponding to the

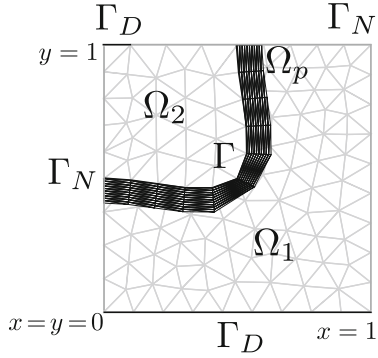


Fig. 1. Computational domain and mesh for the test case of Sect. 4: $N = 117$, $\varepsilon = 1.5e^{-3}$, $\tau_\varepsilon = 0.0916$

zero-order terms in (2')–(4'), the algebraic equations in (8)–(9) become identical to those that would arise in a Centered Finite Difference discretization. Using the methods of analysis in [1], one can establish the following.

Theorem 1. *Let* $d = 1$ *and* $\tau_\varepsilon := \min\left\{d,\ 2\sqrt{\frac{\varepsilon_1}{\beta}}\ln N\right\}$. *Then*

$$\|u - u^h\|_\infty \leq C\frac{\ln N}{N} \quad \|u^{h\prime} - u'\|_{\infty,\Omega\setminus\Omega_p} \leq \frac{\ln N}{N} \quad \sqrt{\varepsilon}\|u^{h\prime} - u'\|_{\infty,\Omega_p} \leq C\frac{(\ln N)^2}{N}$$

where C *is a constant independent of* ε *and* N.

Although the bounds in Theorem 1 have only been established in the case of $d = 1$, the numerical results in Sect. 4 suggest that similar error estimates may also hold for $d = 2$.

4 Results

As a test case we consider a problem with $d = 2$ in the domain pictured in Fig. 1 where we let $r_1 = f_1 = 1$ in Ω_1, $r_2 = f_2 = 0$ in Ω_2, $g_D = 1$ on $\Gamma_D \cap \Omega_1$ and $g_D = 0$ on $\Gamma_D \cap \Omega_2$. Figure 2a, b show the solution of the test problem as computed by the algorithm of Sect. 3 and by a standard piece-wise linear Finite Element discretization on a single quasi-uniform triangulation over the whole domain Ω, respectively. As anticipated in the introduction, we see the algorithm of Sect. 3 as a viable extension to the two dimensional case of the algorithm that in [1] was used to estimate the capacitance of a MOS structure. Given the particular application we have in mind, the quantity we are most interested in is the total flux through the Dirichlet portion of the boundary of Ω_2, *i.e.* $Q := \int_{\Gamma_D \cup \Omega_2} \sigma(u) \cdot \nu \, d\gamma$. By comparing the plots in Fig. 2c, d one may notice that, while the error produced by the standard approximation has a non-trivial dependence on both the number of degrees of freedom N and on the singular perturbation parameter ε, for the algorithm proposed here the error, at least for small enough ε is a function of N alone.

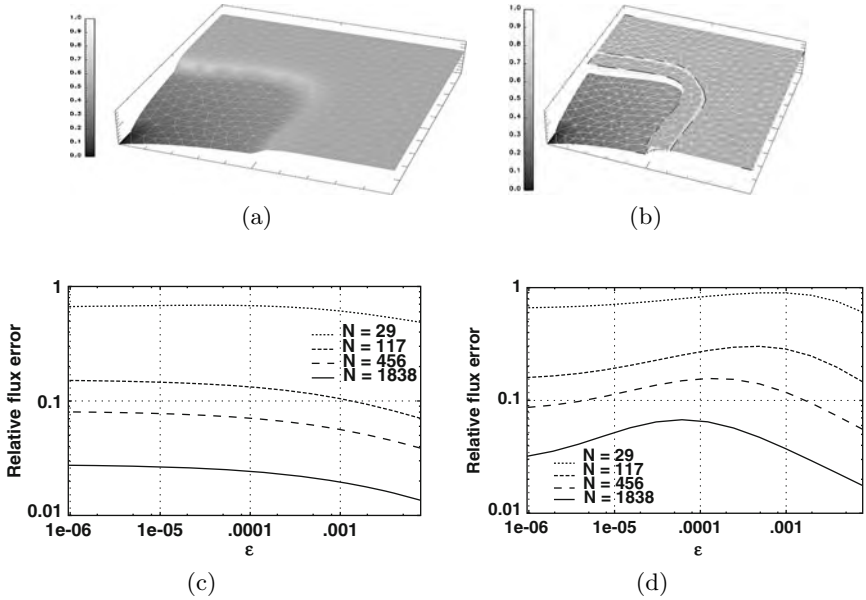


Fig. 2. (**a**) Computed solution u of the test problem of Sect. 4 for $\varepsilon = 1.5 \times 10^{-3}$ and $N = 456$. (**b**) Computed solution of the test problem for $\varepsilon = 1.5 \times 10^{-3}$ and $N = 456$ with the algorithm described in Sect. 3, note that in Ω_p both the solution u and the regular component v_1 are shown. (**c**) Relative error in the flux Q computed by the algorithm of Sect. 3. (**d**) Relative error in the flux Q computed on a quasi-uniform mesh

Acknowledgments

The first author was supported by the European Commission in the framework of the CoMSON RTN project. The second and third author were supported by the Mathematics Applications Consortium for Science and Industry in Ireland (MACSI) under the Science Foundation Ireland (SFI) mathematics initiative.

References

1. de Falco, C., O'Riordan, E.: Dublin City University preprint 2008, MS-08-04.
2. Farrell, P.A., Hegarty, A.F., Miller, J.J.H., O'Riordan, E., Shishkin, G.I.: Robust Computational Techniques for Boundary Layers. Chapman and Hall/CRC, Boca Raton (2000)
3. Hughes, T., Engel, G., Mazzei, L., Larson, M.: J. Comp. Phys. **163**, 467–488 (2000)
4. Kim, J.V., et al.: VLSI Technology, 2005. Digest of Technical Papers. 2005 Symposium on, pp. 34–35, June 2005

Upgrading the UK Broadband Infrastructure by Monte Carlo Simulation

W.T. Lee[1] and K. Mueller[2]

[1] MACSI, Department of Mathematics and Statistics, University of Limerick, Limerick, Ireland william.lee@ul.ie
[2] Informa Telecoms & Media, 37-41 Mortimer Street, London W1T 3JH, UK katja.mueller@informa.com

Summary. There is currently much interest in upgrading the UK broadband infrastructure by connecting subscribers to their local exchanges by fibre optic cables. The key to determining whether such an upgrade is feasible is cost. A method of calculating an optimal upgrade, in the sense of maximising return on investment, is described and results of its application to a test case are discussed.

1 Introduction

The topology of the UK's broadband infrastructure is illustrated in Fig. 1. Individual subscribers receive and transmit data using the asymmetric digital subscriber line technology (ADSL) through copper (or aluminium) cables which are amalgamated at nodes and end at exchanges. At the exchanges the signal is converted to optical pulses which are sent and received through fibre-optic cables that form the transport network.

A number of factors have lead to an increase in Internet subscribers' data transfer requirements with the result that the current infrastructure is becoming inadequate. These include:

- An increased reliance on Internet connectivity in everyday life
- A tremendous rise in the consumption of bandwidth intensive multimedia applications
- The popularity of Web 2.0 websites and related activities

The latter is significant since the interactivity encouraged by Web 2.0 applications requires large upload speeds. Current ADSL technology is designed to increase download speeds at the cost of reducing upload speeds.

The bottleneck in data transmission lies between the subscribers and the exchanges. Under optimal conditions, i.e. a subscriber lives within a few hundred metres of the exchange, ADSL technology allows upload speeds of 1 Mbps

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*, Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_34,

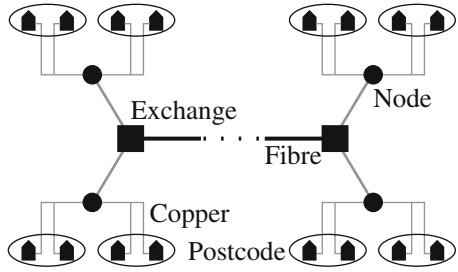


Fig. 1. Topology of UK broadband infrastructure

and download speeds of 8 Mbps. There is a strong attenuation problem, meaning that these speeds are halved for a subscriber who lives several kilometres from an exchange.

There is a consensus that some form of network upgrade is required. The extent of the upgrade is still a matter of debate. Should it be an upgrade of equipment or a complete overhaul of the network infrastructure? One option is an upgrade of ADSL technology to ADSL2+. This would introduce a moderate gain in upload and download speeds, requiring little additional investment since it would not change the underlying infrastructure. This has already been applied in the UK and other OECD countries with limited success. Usually only half of all copper access cables lend themselves to this sort of upgrade.

A more radical approach would be to replace the copper cables with fibre optic cables: the subject of this paper. In this work we consider an upgrade known as Fibre-to-the-Home (FTTH) in which all copper cables are replaced by optical fibres, allowing upload and download speeds of 10–100 Mbps. Several European operators are implementing hybrid schemes in which optical fibre cabling is used to replace only part of the access network.

The main hurdle to upgrading the UK infrastructure to fibre optic cable is cost. This is not entirely straightforward however: there is more than one way in which the upgrade can be carried out. An upgraded network will not require as many exchanges as the current network, since fibre optic cable does not suffer from the rapid attenuation of copper cables. Therefore, as part of the upgrade, a number of exchanges can be decommissioned, with a concomitant reduction in the operating costs of the network. Also a limited upgrade in which only some subscribers are upgraded is possible. We here present a way of estimating the costs of a given upgrade and the revenue potential of the upgraded network, and an algorithm based on Monte Carlo simulated annealing for finding an upgrade which maximises return on investment.

Our approach differs from other modelling studies in this area in two respects. Firstly, we use detailed small-area estimates of broadband subscriber numbers at UK postcode resolution (a UK postcode typically describes a single street), This data is taken from the Point Topic Broadband User

Survey [1]. Secondly, previous studies have made a priori choices about locations upgraded, deciding in advance which subscribers and exchanges to upgrade. In our approach the algorithm chooses the scope of the upgrade in order to maximise return on investment.

2 Cost Model

In order to find an optimal network we need a measure by which to compare upgraded networks. We take this to be ΔP, the increase in profits due to the network, summed over the number of years investors are willing to wait for a return on their expenditure. ΔP can be broken down into contributions from increased revenue from subscribers with a fibre connection R, savings made due to the lower operating costs of a fibre network S, and the initial investment needed to set up the network I.

$$\Delta P = R + S - I \tag{1}$$

In calculating these quantities we take the perspective of the incumbent operator BT Openreach, the owner of the current infrastructure.

2.1 Revenue

The increase in revenue due the provision of a fibre service is given as

$$R = \sum_i n_{\mathrm{bb},i}^{\mathrm{pc}} r \tag{2}$$

where summation is over all postcodes i that have been upgraded to fibre, $n_{\mathrm{bb},i}^{\mathrm{pc}}$ is the number of households in that postcode with broadband subscriptions and r is the increased revenue per subscriber.

2.2 Savings

There are two possible sources of savings from an upgraded network. The first is from the decommissioning of exchanges, the second is from the reduced operating costs of exchanges that have been converted to fibre. Thus the savings S are given by

$$S = S_{\mathrm{dc}}^{\mathrm{ex}} N_{\mathrm{dc}}^{\mathrm{ex}} + S_{\mathrm{ug}}^{\mathrm{ex}} N_{\mathrm{ug}}^{\mathrm{ex}} \tag{3}$$

where $S_{\mathrm{dc}}^{\mathrm{ex}}$ is the saving made from decommissioning an exchange, $N_{\mathrm{dc}}^{\mathrm{ex}}$ is the number of decommissioned exchanges, $S_{\mathrm{ug}}^{\mathrm{ex}}$ is the saving made from an upgraded exchange and $N_{\mathrm{ug}}^{\mathrm{ex}}$ is the number of upgraded exchanges.

2.3 Investment

A substantial amount of investment is needed to upgrade the network. This can be categorised into four areas

- Exchanges must be upgraded to fibre (although there is a long term saving associated with this there is also an initial investment that must be made).
- Tunnels must be dug to get the cables from the exchanges to the subscribers.
- The cable itself must be bought.
- There is an additional cost in making the final connection to the subscribers.

The investment, I, is described by the equation

$$I = C_{\mathrm{ug}}^{\mathrm{ex}} N_{\mathrm{ug}}^{\mathrm{ex}} + \sum_i \left[C_{\mathrm{dig}} d_{\mathrm{pc},i} + C_{\mathrm{cable}} d_{\mathrm{pc},i} n_{\mathrm{hh},i}^{\mathrm{pc}} + C_{\mathrm{ug},i}^{\mathrm{hh}} n_{\mathrm{bb},i}^{\mathrm{pc}} \right] \tag{4}$$

where $C_{\mathrm{ug}}^{\mathrm{ex}}$ is the cost of upgrading an exchange, $N_{\mathrm{ug}}^{\mathrm{ex}}$ is the number of upgraded exchanges, summation is over upgraded postcodes i, C_{dig} is the cost per unit length of digging tunnels, $d_{\mathrm{pc},i}$ is the total length of tunnels, C_{cable} is the cost per unit length of cable, $n_{\mathrm{hh},i}^{\mathrm{pc}}$ is the number of households in upgraded postcode i, $C_{\mathrm{ug},i}^{\mathrm{hh}}$ is the cost of upgrading a single household.

3 Optimisation Algorithm

Finding the best possible arrangement of upgraded and decommissioned exchanges and upgraded postcodes is a combinatorial optimisation problem. The Metropolis Monte Carlo algorithm excels at solving this type of problem [2,3].

Figure 2 shows how the algorithm is used to decide whether to accept or reject a proposed adjustment of the system. The probability test consists of accepting a step in which ΔP_{new} is smaller than ΔP_{old} with probability

$$\exp\left[\frac{\Delta P_{\mathrm{old}} - \Delta P_{\mathrm{new}}}{T}\right] \tag{5}$$

where T is initially taken as a large quantity and then reduced to zero over the course of the simulation. Thus, in the initial stages of the simulation, steps which substantially reduce ΔP are often accepted; whereas, in the final iterations, only those adjustments which increase ΔP are allowed.

The adjustments made to the network are shown in Fig. 3. These consist of: upgrading an exchange and its surrounding postcodes to fibre; decommissioning an exchange and attaching its surrounding postcodes by fibre to the nearest upgraded exchange; reversing the previous processes.

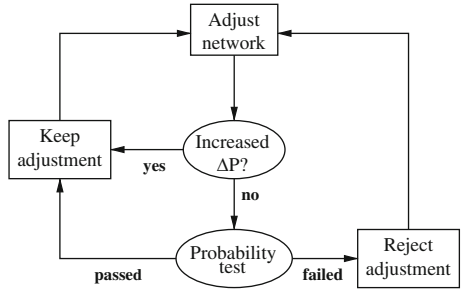


Fig. 2. The Metropolis Monte Carlo algorithm

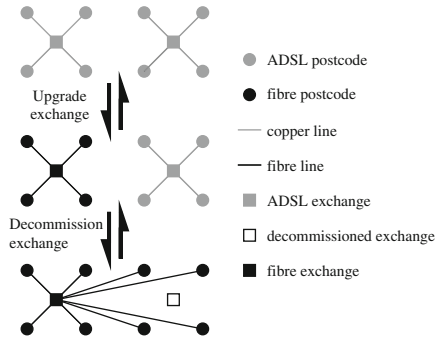


Fig. 3. Adjustments made to the network

4 A Test Case

To test the algorithm we conducted a case study: finding an optimal network for the Isle of Wight. The Isle of Wight makes a good test case: it is a self contained geographical unit and accommodates both urban and rural areas. We apply the algorithm with the parameters shown in Table 1, and assume that investors are willing to wait 5 years for a return on their initial outlay. Figure 4 shows the optimised network. Under these conditions the algorithm suggests only upgrading the exchanges and postcodes in the three largest towns.

5 Conclusions and Further Work

Upgrading the UK broadband infrastructure to fibre optic cable is a complex and expensive task. With the telecoms industry under pressure from investors, stake-holders and competitors to perform well in financial terms and to keep innovating their service applications, it is important carry out network upgrades in as cost effective a manner as possible. We have here presented an algorithm for calculating an optimal upgrade and demonstrated

Table 1. Parameter values used in the simulation

Parameter	Value	Units
$S_{\mathrm{dc}}^{\mathrm{ex}}$	200.0	10^3 \$ year^{-1}
$S_{\mathrm{ug}}^{\mathrm{ex}}$	18.0	10^3 \$ year^{-1}
$C_{\mathrm{ug}}^{\mathrm{ex}}$	150.0	10^3 \$
$C_{\mathrm{ug}}^{\mathrm{hh}}$	200.0	\$ hh^{-1}
C_{dig}	38.0	\$ m^{-1}
C_{cable}	1.5	\$ m^{-1}
r	15.0	\$ hh^{-1} month^{-1}

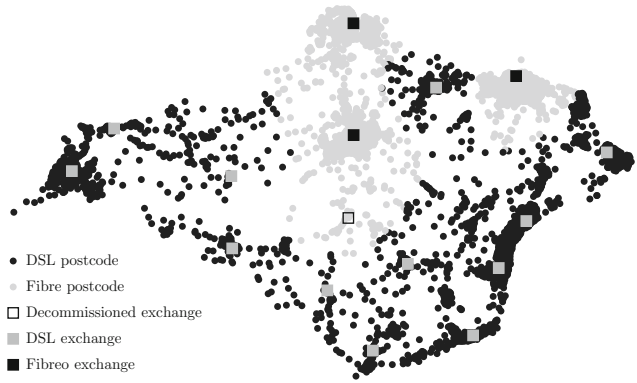


Fig. 4. Optimal network for the Isle of Wight

its application in a test case. In future work we plan to scale up the calculations to the whole of the UK.

Acknowledgments

Data for this study was provided by Point Topic Ltd. (`www.point-topic.com`). WTL wishes to acknowledge the support of the Mathematics Applications Consortium for Science and Industry (`www.macsi.ul.ie`) funded by the Science Foundation Ireland Mathematics Initiative Grant 06/MI/005.

References

1. Point Topic Ltd. Broadband User Survey, `www.point-topic.com`
2. Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., Teller, E.: J. Chem. Phys. **21** 1087–92 (1953)
3. Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P.: Numerical Recipes Cambridge University Press, Cambridge (1992)

Multi-Stepping and Anti-Icing / De-Icing Devices

J.P.F. Charpin[1] and P. Verdin[2]

[1] MACSI, Department of Mathematics and Statistics, University of Limerick, Limerick, Ireland, jean.charpin@ul.ie
[2] Applied Mathematics and Computing (AMAC), School of Engineering, Cranfield University, Cranfield, Bedfordshire MK43 OAL, United Kingdom, p.verdin@cranfield.ac.uk

Summary. A multi-step version of the aircraft icing code ICECREMO2 is used to simulate ice layers in cold and mild conditions. The models used are presented and results are compared with available data. The modifications necessary to introduce anti- or de-icing devices in the model are also discussed.

1 Aircrafts and Icing

Aircraft icing occurs in cold and wet conditions. Water droplets hitting the plane may freeze instantaneously and form rime ice. Part of the droplets may also remain liquid and form glaze ice [1]. Ice growing on key parts of a plane is a major concern for aircraft manufacturers. Computer codes are used to assist them with aircraft design and certification work. ICECREMO2 is one of the most recent codes, developed by a number of companies in the United Kingdom, see acknowledgements. Flow parameters necessary to simulate ice growth must be evaluated during a flow calculation at the start of the simulation. When ice accretes, flow-dependent parameters may vary significantly and updates are necessary to improve the accuracy of the ice prediction. This method known as multi-stepping will be detailed for ICECREMO2. Models for anti- and de-icing devices that prevent or limit ice formation will also be reviewed briefly.

2 Modelling Aircraft Icing

2.1 Icing Model

Ice grows in two phases, rime ice appears first and then glaze ice. ICECREMO2 uses its own icing model [6], this is now presented for both phases. Flow-dependent parameters will also be identified.

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*, Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_35,

Rime Ice Growth

Rime ice grows in cold conditions; the impinging droplets freeze almost instantaneously when they reach the surface. The ice growth rate may then be calculated using a mass balance:

$$\frac{\partial b}{\partial t} = \frac{\beta \varrho_w V_\infty}{\varrho_i} \ , \tag{1}$$

where b denotes the ice height and $\beta \varrho_w V_\infty$ represents the mass of water hitting the surface. The catch or collection efficiency, β, is the ratio of the mass flux hitting the surface and the mass flux that would hit if water droplets had straight line trajectories. It must be computed using the results of the flow calculation and is therefore likely to vary significantly when the ice layer increases. This parameter should be updated when using multi-step methods.

Glaze Ice Growth

In milder conditions, when the ice layer is thick enough, some of the impinging droplets may remain liquid and a water layer forms at the top of the ice accretion. This layer will remain extremely thin but it is key to accurate simulations. The mass balance (1) is then replaced by

$$\varrho_i \frac{\partial b}{\partial t} + \varrho_w \left(\frac{\partial h}{\partial t} + \nabla \cdot \mathbf{Q} \right) = \beta \varrho_w V_\infty \ , \tag{2}$$

where the fluid flux $\mathbf{Q}$ is defined by

$$\mathbf{Q} = \left(-\frac{\varrho_w g h^3}{3\mu} \mathbf{g} \cdot \mathbf{e_x} + \frac{A_x h^2}{2\mu} \ , -\frac{\varrho_w g h^3}{3\mu} \mathbf{g} \cdot \mathbf{e_y} + \frac{A_y h^2}{2\mu} \right) \ , \tag{3}$$

and h denotes the water height. The mass balance (2) accounts for the evolution of the ice layer thickness and the movement of the water layer through the flux. The shear stress (A_x, A_y) is highly dependent on the flow and should be updated during multi-step calculations.

Equation (2) is not enough to determine both the ice growth rate, $\partial b/\partial t$, and the water growth rate, $\partial h/\partial t$. It must be coupled with the Stefan condition:

$$\varrho_i L_f \frac{\partial b}{\partial t} = \kappa_i \frac{\partial T}{\partial z} - \kappa_w \frac{\partial \theta}{\partial z} = \kappa_i \frac{T_f - T_s}{b} - \kappa_w \frac{E - F\, T_f}{1 + Fh}, \tag{4}$$

where the temperature gradients in the ice and water layers, $\partial T/\partial z$ and $\partial \theta/\partial z$, are calculated assuming that the layers are thin. The temperature of the substrate is T_s and the temperature at the interface between the ice and water layers is the freezing temperature T_f. The gradient at the top of the water layer is $T_z = E - F\theta(b+h)$ where E and F reflect the physics at the top of the ice layer, see [5] for details. All the parameters involved in E and F are constant except for the collection efficiency and the heat transfer coefficient. These two coefficients must be updated when using a multi-step algorithm. This completes the model and it may now be included in an icing algorithm.

2.2 Icing Code Structure

Icing codes are generally split into four parts [4]. The flow around the body is determined first, using computational fluid dynamics techniques. Water droplet trajectories are then simulated to evaluate the collection efficiency. This parameter reflecting the quantity of liquid impinging on the wing surface is calculated during the third stage, together with the heat transfer coefficient. Finally, the ice growth is calculated. This procedure is known as a one-step calculation.

To improve the accuracy of the ice simulations the four parts of the one-step algorithm are repeated as many times as necessary. A criterion defining the start of a new cycle is specified and the corresponding ice growth rates are determined. The initial geometry is then adjusted to account for the accreted ice and a new flow field calculation is performed around the iced geometry. This procedure is iterated until the specified total icing exposure time is reached but it is potentially time consuming: the criterion triggering the start of a new cycle and a new flow calculation needs to be defined with care. Most commonly, the total icing exposure time is divided into a given number of equal time increments. This may not be the most appropriate method when glaze ice is growing and calculating the ice growth rate is less straightforward. A criterion based on the ice height may be more effective: a new flow calculation is started when the maximum ice height growing during a multi-time-step reaches a value chosen by the user. Time-steps would be longer when ice grows slowly and shorter for higher ice growth rates and this should reduce the number of multi-time-steps. Both solutions will be tested in the following.

3 Numerical Results

Ice shapes are simulated on a NACA0012 wing in rime ice conditions, $T_{ambient} = -26°\text{C}$, and glaze ice conditions, $T_{ambient} = -7°\text{C}$.

3.1 Rime Ice Conditions

Figure 1 shows the experimental ice shape calculated by Shin and Bond [7], the one step ice shape and ice shapes simulated using the step-by-step algorithm with the time and ice criteria. All simulations over-estimate the ice layer at the edge of the wing and slightly underestimate the accretion further downstream. Using the multi-step algorithm reduces the excess of ice close to the leading edge. For both multi-step criteria, convergence is reached after 5 flow calculations, corresponding to 72 s for the time criterion and a maximum height $b_{max} = 0.8\%\ c$ for the ice height criterion where c is the chord length. It appears that the criterion used does not make a significant difference here.

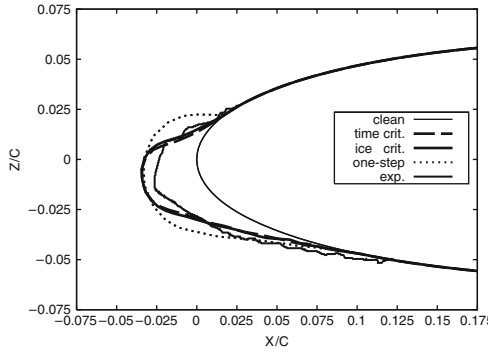


Fig. 1. Ice shapes in rime ice conditions

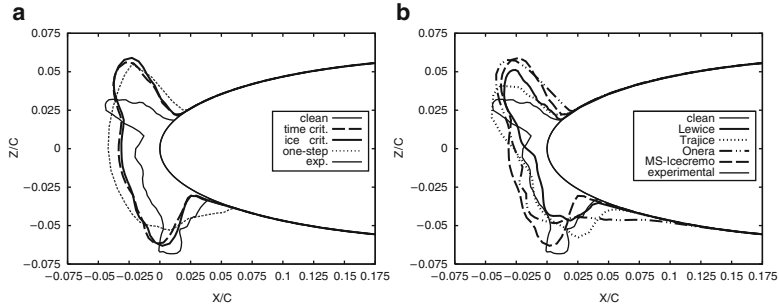


Fig. 2. Ice shapes in glaze ice conditions: (**a**) numerical simulations; (**b**) comparison with other icing codes

3.2 Glaze Ice Conditions

Figure 2 shows the experimental results [3] and the ice shape simulated for both the time and the ice height criteria. As in the rime ice situation, the accreted layer is overestimated. However, the two horns typical of glaze ice growth appear very clearly. These difficulties to match experimental results are due to the complexity of the ice shape and to the model that only allows ice to grow orthogonally to the wing surface. Here, the model prevents an accurate simulation of the lower horn.

Convergence is achieved after 12 iterations for the time criterion and the maximum height $b_{max} = 0.6\%\ c$ for the ice height criterion, corresponding to 10 flow calculations. The results are of equivalent quality, as may be seen on Fig. 2. As could be expected, when using the ice height criterion, the time-steps are significantly longer at the beginning of the simulation and become

shorter towards the end. For the present example, the ice height criterion is clearly more efficient and should be preferred to the standard time criterion.

The ICECREMO2 results compare favourably with shapes calculated using standard multi-step versions of the icing codes LEWICE, TRAJICE and ONERA in their December 2000 version [3], see Fig. 2. All codes over-estimate the ice layer and none of them matches the experimental shape perfectly. Overall, the step-by-step version of ICECREMO2 gives the best approximation for the ice shape and the algorithm used here is a significant improvement from the codes used in the NATO/RTO study [3]. The main difference between the codes is the icing model. The improved ice shape is due to the icing model, the carefully chosen criterion triggering the flow recalculation is key in reducing the time required to reach this best possible shape [8].

4 Challenges of Anti-Icing and De-Icing Models

Various systems have been developed to combat aircraft icing. Action may be taken before take off, but this will only be effective for a short period, or during the flight. Two approaches may be used: anti-icing systems try to prevent ice from accreting at all, working continuously from the start of ice growth. This can become extremely energy consuming and instead de-icing devices may be used that remove the ice accretion periodically. In this situation, a small layer of ice is allowed to form before removal. A popular solution is to use electro-thermal systems in the form of electrical heaters elements placed on the airfoil to destroy the ice adhesion, so that aerodynamic forces can remove the ice from the surface.

The model presented in Sect. 2.1 can be adapted to include heating pads. In this situation, the surface temperature, T_s, is allowed to vary and heat equations need to be solved inside the wing structure, as implemented in ICECREMO2, see [2]. The objective is then to reach the freezing temperature at the surface of the wing. The bottom of the ice layer will start melting and the top of the layer will shatter before being removed by aerodynamical forces. The geometry and position of the heating pads should be optimised to guarantee maximum efficiency and the heating strategy should be adapted to minimise the energy required to guarantee the fly safety. This enhanced model will then have to be coupled with a multi-step algorithm to guarantee a good level of accuracy.

5 Conclusions and Further Work

Ice layers were simulated using a multi-step version of the aircraft icing code ICECREMO2. The icing model improved the accuracy of the simulation while the algorithm reduced the time required to achieve this best possible shape. Extending the method to anti-icing and de-icing models requires a modification of the existing model and finding an optimal energy saving strategy.

Acknowledgements

The present research work forms part of the ICECREMO2 project. ICECREMO2 is a three-dimensional ice accretion and water flow code developed collaboratively by Airbus UK, BAe Systems, Dunlop Aerospace (Meggitt), Rolls-Royce, Westland Helicopters Ltd (now AgustaWestland), QinetiQ and Cranfield University under the auspices of the UK Department of Trade and Industry.

J.P.F. Charpin acknowledges the support of the Mathematics Applications Consortium for Science and Industry (MACSI), funded by the Science Foundation Ireland (SFI) Mathematics initiative 06/MI/005.

References

1. Gent, R.W., Dart, N.P., Cansdale, J.T.: Phil. Trans. R. Soc. Lond. A **358**, 2873–2911 (2000)
2. Harireche, O., Verdin, P., Thompson, C.P., Hammond, D.W.: J. Aircr. **45**(6), 1924–1936 (2008)
3. Kind, R.J.: NATO-RTO, **TR-038** (2001)
4. Kind, R.J., Potapczuk, M.G., Feo, A., Golia, C., Shah, A.D.: Prog. Aerosp. Sci. **34**, 257–345 (1998)
5. Myers, T.G., Charpin, J.P.F.: Int. J. Heat Mass Transf. **47**, 5483–5500 (2004)
6. Myers, T.G., Charpin, J.P.F., Chapman, S.J.: Phys. Fluids **14**, 2788–2803 (2002)
7. Shin, J., Bond, T.H.: AIAA **92-0647** (1992)
8. Verdin, P., Charpin, J.P.F., Thompson, C.P.: J. Aircr. **46**(5), 1607–1613 (2009)

A Diffusion Model for Spatially Dependent Photopolymerization

D. Mackey[1], T. Babeva[2], I. Naydenova[2], and V. Toal[2]

[1] School of Mathematical Sciences, Dublin Institute of Technology, Dublin, Ireland, dana.mackey@dit.ie
[2] Centre for Industrial and Engineering Optics, Dublin Institute of Technology, Dublin, Ireland, babeva@gmail.com, izabela.naydenova@dit.ie, vincent.toal@dit.ie

Summary. Photopolymers represent an attractive class of optical recording materials due to properties such as high refractive index modulation, dry film processing, low cost, etc. Applications include holographic data storage disks, optical interconnections, memories and filters. This paper addresses the dynamics of short-exposure holographic grating formation; a new model is proposed to explain the experimental observations of low diffraction efficiency in high spatial frequency gratings.

1 Introduction

The basic formulation of a dry photopolymer system consists of one or two monomers, photoinitiator and sensitizing dye, all dispersed in a binder matrix. Upon uniform illumination, a monomer polymerizes and the refractive index of the system changes. When material is exposed to an interference pattern more monomers are being polymerized in the bright regions than in the dark ones. This sets up a concentration gradient and the monomer diffuses from dark to bright areas. The recorded holographic grating (spatial distribution of refractive index) is a result of changes in the relative density of components.

Grating evolution in photopolymers has been studied by several authors ([4, 6], etc.). However, the common feature of most theoretical models proposed to date is that they cannot describe the experimental observation of poor diffraction efficiency at high spatial frequencies. There are two theories explaining this poor response. The "nonlocal-response diffusion model" of [5] assumes growth of polymer chains away from their initiation locations and predicts that high frequency gratings can be improved if shorter polymer chains are created during the recording. Despite the successful theoretical model no supporting experimental evidence has been reported so far for spatial frequencies higher than 3,000 lines/mm.

A second theory was proposed in [1] and [2]. It states that the counter diffusion of short-chain polymer molecules away from the bright fringes is

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*, Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_36,

responsible for the reduction in diffraction efficiency and predicts that producing short chains is not sufficient to achieve high spatial frequency resolution but, in addition, their diffusion must be suppressed. In order to achieve theoretical validation for this theory, we propose here a new mathematical model for hologram formation and compare numerical simulations of the refractive index modulation after short exposures with experimental results.

2 Problem Formulation

The photopolymer is exposed to two coherent beams of intensities I_1 and I_2 which create the following illumination pattern

$$I(x) = I_0\,(1 + V\,\cos(kx)),$$

where k is the grating wavenumber, $I_0 = I_1 + I_2$ and $V = 2\sqrt{I_1 I_2}/(I_1 + I_2)$ are the overall intensity and visibility of the interference pattern, respectively. The holographic grating formation then proceeds in three steps: initiation, propagation and termination. Upon illumination, the sensitizing dye absorbs a photon and reacts with the electron donor to produce free radicals; in the presence of monomer these free radicals initiate polymerization. During the propagation step, free radicals and monomer molecules interact and produce growing polymer chains. At the termination step, two free radicals or two polymer chains interact and the polymer chains stop growing. As stated above, the faster consumption of monomer in the illuminated areas sets up a concentration gradient so the free monomer diffuses from dark to bright fringes; in addition we now assume that short-chain polymer molecules (or radicals) can also diffuse during recording. All these processes modify the spatial modulation of the refractive index and yield a phase grating.

The refractive index of a material consisting of a mixture of components can be calculated with the well-known Lorentz–Lorenz equation:

$$\frac{n^2 - 1}{n^2 + 2} = \sum_i \Phi_i \frac{n_i^2 - 1}{n_i^2 + 2}$$

where n is the effective refractive index of the mixture, n_i are the refractive indices of the components (monomer, polymer and binder) determined separately from spectrophotometric measurements, and Φ_i are the normalized concentrations of the components (e.g. $\Phi_m = m/(b + m + p)$, where m, p and b denote concentrations of monomer, polymer and binder, respectively). The details of this calculation are not important so will not be included here.

The refractive index modulation determines the grating strength and is calculated as the difference between the values in the bright and dark fringes,

$$\Delta n(t) = n_{max}(t) - n_{min}(t)$$

In short exposure conditions $\Delta n(t)$ should ideally exhibit fast growth followed by convergence to an equilibrium state.

3 Proposed Model

We now propose a generalization to existing models which takes into account monomer and polymer diffusion, creation of short polymer chains and introduces a simple "immobilization" mechanism which mimics the growth of polymer chains to the extent where they cannot diffuse any longer. The short exposure régime is also reflected in the model, whereby all polymerization and immobilization processes stop once the light beam is terminated.

In what follows, the spatial domain is assumed to be $x \in [0, \Lambda]$, where $\Lambda = \frac{2\pi}{k}$ is the grating period (or fringe spacing). The classical model (see, for example, [6]) consists of a polymerization-diffusion equation for the monomer molecules but assumes diffusion stops once monomer is polymerized,

$$\frac{\partial m}{\partial t} = D_m \frac{\partial^2 m}{\partial x^2} - F(x)\, m. \tag{1}$$

Here $m(x,t)$ denotes monomer concentration, D_m is the monomer diffusion constant and the polymerization rate is proportional to the illumination

$$F(x) = F_0 \left(1 + V \cos(kx)\right)^a \equiv F_0\, f(x)$$

where F_0 is the polymerization constant and $a > 0$. The initial concentration of free monomer in the material is spatially uniform, $m(x,0) = m_0$.

In addition, we now assume that short polymer chains can diffuse and the diffusion coefficient is also proportional to the illumination, $D(x) = D_p f(x)$, meaning that at high intensity, more short-chains are formed, which are more mobile. We also assume that short chains are converted to long chains at a rate proportional to monomer and polymer concentrations (Γ is the conversion rate constant) and that the long chains are immobile once formed. The resulting equations are

$$\frac{\partial m}{\partial t} = D_m \frac{\partial^2 m}{\partial x^2} - \Phi(t)\, F(x)\, m, \tag{2}$$

$$\frac{\partial p_1}{\partial t} = \frac{\partial}{\partial x}\left[D(x)\frac{\partial p_1}{\partial x}\right] + \Phi(t)\ \left[F(x)\, m - \Gamma m p_1\right] \tag{3}$$

$$\frac{\partial p_2}{\partial t} = \Phi(t)\, \Gamma m p_1, \tag{4}$$

where, $p_1(x,t)$ is the concentration of short polymer chains, $p_2(x,t)$ is the concentration of long polymer chains, with initial conditions $p_1(x,0) = 0, \quad p_2(x,0) = 0$. We assume these equations are supplemented by zero-flux boundary conditions. To account for a short exposure régime in (2)–(4) we have introduced the step function

$$\Phi(t) = \begin{cases} 1, & \text{if } t \le t_e \\ 0, & \text{if } t > t_e \end{cases}$$

where t_e is the exposure time. Note that this model is only valid for exposure times which are much shorter than diffusion times, as otherwise diffusion coefficients for both monomers and polymers are known to be time dependent.

With the choice of non-dimensional variables

$$\bar{x} = \frac{x}{\Lambda}, \quad \bar{t} = tF_0, \quad \bar{m} = \frac{m}{m_0}, \quad \bar{p_i} = \frac{p_i}{m_0} \ (i = 1, 2),$$

the model becomes

$$\frac{\partial m}{\partial t} = \kappa \frac{\partial^2 m}{\partial x^2} - \Phi(t)\, f(x)\, m \tag{5}$$

$$\frac{\partial p_1}{\partial t} = \varepsilon\kappa \frac{\partial}{\partial x}\left[f(x)\frac{\partial p_1}{\partial x}\right] + \Phi(t)\ [f(x)\, m - \gamma m p_1] \tag{6}$$

$$\frac{\partial p_2}{\partial t} = \Phi(t)\, \gamma m p_1, \tag{7}$$

where

$$\kappa = \frac{D_m}{F_0 \Lambda^2}; \qquad \varepsilon = \frac{D_p}{D_m} \ll 1; \qquad \gamma = \frac{\Gamma m_0}{F_0}. \tag{8}$$

We also have the initial and boundary conditions

$$m(x, 0) = 1, \ p_i(x, 0) = 0; \tag{9}$$

$$\frac{\partial m}{\partial x}(x, t) = \frac{\partial p_1}{\partial x}(x, t) = \frac{\partial p_2}{\partial x}(x, t) = 0, \qquad \text{for } x = 0, 1. \tag{10}$$

On adding and integrating equations (5)–(7) we get the conservation law

$$\int_0^1 [m(x, t) + p_1(x, t) + p_2(t)]\ dx = 1,$$

which is to be expected, since monomer is converted into polymer while the total concentration of particles remains constant.

4 Results and Conclusions

The non-dimensional model (5)–(10) was integrated numerically using a standard finite difference method. The numerical values used for the diffusion constants are $D_m = 10^{-7}\,\mathrm{cm}^2/\mathrm{s}$, $D_p = 10^{-9}\,\mathrm{cm}^2/\mathrm{s}$ (close to the values determined in [3]), so $\varepsilon = 0.01$. The polymerization rate constant is assumed to be $F_0 = 0.3\,\mathrm{s}^{-1}$, $a = 0.5$, and Λ is varied between $2 \cdot 10^{-7}\,\mathrm{m}$ and $1 \cdot 10^{-5}\,\mathrm{m}$ (corresponding to a range of 100–5,000 lines/mm). The exposure time is $t_e = 0.2\,\mathrm{s}$, unless otherwise specified. Numerical values for γ (the "immobilization" rate constant) would be difficult to determine experimentally; however we found that $\gamma = 1$ yielded qualitatively and quantitatively satisfactory results.

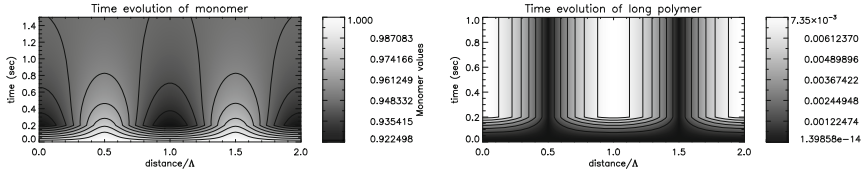


Fig. 1. Relative concentrations of monomer (m/m_0) and long polymer (p_2/m_0). Here $\Lambda = 5 \cdot 10^{-6}$, $t_e = 0.2\,\mathrm{s}$

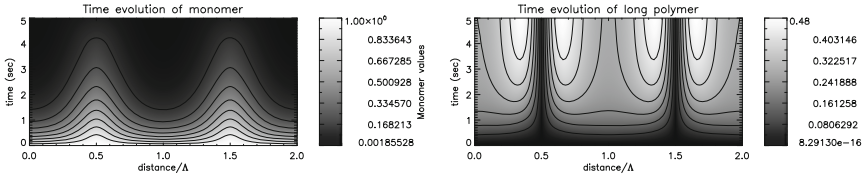


Fig. 2. Relative concentrations of monomer (m/m_0) and long polymer (p_2/m_0). Here $\Lambda = 10^{-5}\,\mathrm{m}$ and $t_e = 4\,\mathrm{s}$

Figures 1 and 2 show the evolution of the monomer and long polymer concentrations for various values of the system parameters. The spatial modulation of these species is represented over two grating periods. After the beam is stopped, the monomer and short polymer concentrations converge towards a homogeneous state under the influence of diffusion, while the spatial profile of the long polymer remains frozen. Note there is a departure from the expected sinusoidal pattern occurring for a combination of low spatial frequency and longer exposure time. A detailed analysis of how the ratio between diffusion and polymerization rates, κ, and the exposure time, t_e, affect the grating dynamics will form the subject of further study.

Figure 3 shows remarkable agreement between the qualitative behaviour of the refractive index modulation obtained from experiment and model simulation, for three values of the spatial frequency (200, 350 and 500 lines/mm). Figure 4 shows the time evolution of the refractive index modulation for a wider range of spatial frequencies, between 100 and 5,000 lines/mm. Note that the model simulations reflect the experimental observations of a drop in refractive index modulation at high spatial frequencies.

In conclusion, the model validates the assumption that the poor high spatial frequency response in photopolymers can be explained by the diffusion of short polymer chains from bright to dark fringes. An improvement strategy would require that the holographic recording conditions be chosen so as to suppress the production of short polymer chains and low permeability binders be used in order to prevent their diffusion. Based on this principle, a successful experimental strategy for producing a high diffraction efficiency reflection hologram in an acrylamide-based photopolymer was recently presented in [3].

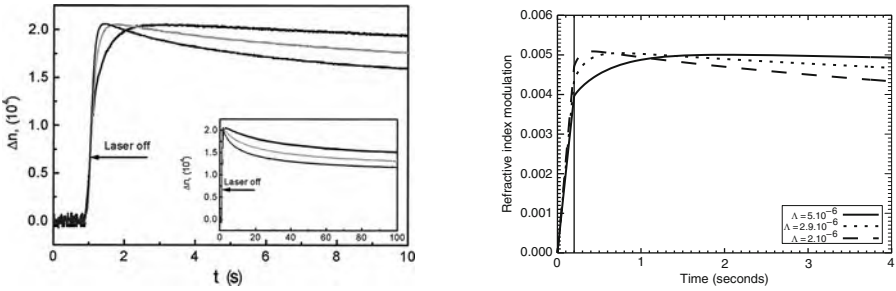


Fig. 3. Comparison of experimental (*left*) and numerical (*right*) results for the refractive index modulation. The three curves correspond to $\Lambda = 5 \cdot 10^{-6}$ m, $\Lambda = 2.9 \cdot 10^{-6}$ m and $\Lambda = 2 \cdot 10^{-6}$ m

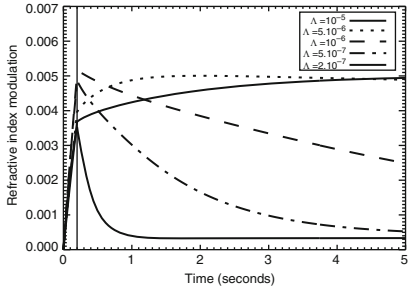


Fig. 4. Refractive index modulation for several values of Λ. Note the deterioration of the grating strength for high spatial frequencies

Acknowledgements

Financial support from the Science Foundation Ireland grant N 065/RFP/PHY085 is gratefully acknowledged.

References

1. Martin, S., Naydenova, I., Jallapuram, R., Howard, R., Toal, V.: Proc. SPIE **6252**, 62525–625217 (2006)
2. Naydenova, I., Jallapuram, R., Howard, R., Martin, S., Toal, V.: Appl. Opt. **43**, 2900–2905 (2004)
3. Naydenova, I., Jallapuram, R., Toal, V., Martin, S.: Appl. Phys. Lett., **92**, 031109 (2008)
4. Piazzola, S., Jenkins, B.: J. Opt. Soc. Am. B **17**, 1147–1157 (2000)
5. Sheridan, T., Lawrence, R.: J. Opt. Soc. Am. A **17**, 1108–1114 (2000)
6. Zhao, G., Mourolis, P.: J. Mod. Opt. **41**, 1929–1939 (1994)

Minisymposium *Interfacial Processes in Industrial and Environmental Turbulent Flows*

I. Eames and J.C.R. Hunt

University College London, Torrington Place, London WC1E 7JE
i.eames@ucl.ac.uk, jcrh@cpom.ucl.ac.uk

1 Introduction

The theme of this symposium was the study of the evolution and dynamics of interfaces that develop in flows between regions with contrasting levels of turbulence, concentration of passive scalar, of indeed fluid properties. The focus here was on high Reynolds number interfaces in inertially dominated flows. In many industrial problems, such as mixing caused by jets or wake separation from aircraft wings, these interfaces determine important properties of the flows such as entrainment mechanisms and lift/drag forces. In the natural environment, the air/sea interface generates surface waves and solitary waves with dangerous levels of energy when they interact with submerged bodies and break on beaches. Within the atmosphere and ocean, the breaking of internal waves supported on thermoclines causes vertical mixing. In these situations, the usual Reynolds averaged approach [5] to modelling or analysing structure is no longer appropriate.

In this minisymposium we invited a range of papers on recent research on many types of interfacial processes in industrial and environmental flows. They include new results, which have fundamental significance and applications for persistence of coherent structures in turbulence, strongly inhomogeneous turbulence, tsunamis, mixing at interfaces in step-stratified fluids [2] and large-scale computational models to capturing subgrid scales processes. Many common threads were identified in these presentations along with some major gaps in our understanding of critical processes, which all made for an exciting meeting.

It was clear that there was a gap in our current understanding of the complex interfacial processes associated with turbulent flows (e.g. [3, 4]). The computational and analytical challenge appears to be formidable to explain and interpret the much more developed body of experimental data. As recognised in other research areas [1], a major challenge appears to be to develop robust new concepts and diagnostic methodologies to classify and interpret

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*,
Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_37,

the underlying physical processes. Some progress appears to being made in the use of local and global conservation laws (for impulse/momentum) but this is non-trivial for some problems (e.g. waves). This approach also leads to improved modelling for practical industry and environmental problems.

The meeting was stimulating and led to long post symposium discussion on the lawn outside the conference hall. There was a consensus that there should be a follow-up meeting to discuss future progress and set out a strategy to tackle these problems, and indeed a Euromech meeting on turbulent interfacial processes will be run at University College London during the end of June 2010.

The organisers thank Qinetiq Plc for their financial support for this minisymposium.

References

1. Eames, I.: Phil. Trans. Royal. Soc. **366**, 2095–2102 (2008)
2. Guyez, E., Flor, J.-B., Hopfinger, H.J.: J. Fluid Mech. **577**, 127–136 (2007)
3. Hunt, J.C.R., Durbin, P.: P Fluid Dyn. Res. 375–404 (1999)
4. Hunt, J.C.R., Eames, I., Westerweel, J.: J. Fluid Mech. **554**, 499–523 (2006)
5. Reynolds, O.: Proc. Roy. Soc. 935–982 (1883)

Wakes of Maneuvering Body in Stratified Fluids

S.I.Voropayev[1,2] and H.J.S.Fernando[1]

[1] Environmental Fluid Dynamics Laboratories, Department of Civil Engineering & Geological Sciences, University of Notre Dame, Notre Dame, IN 46556, USA s.voropayev@nd.edu, Harindra.J.Fernando.10@nd.edu
[2] Institute of Oceanology, Russian Academy of Sciences, Moscow, 117851, Russia

Summary. When submerged self-propelled body makes a maneuver, it imparts net momentum on the surrounding fluid. Using experimental observations and theoretical arguments we show that in a stratified fluid this leads to impulsive momentum wakes with large, long-lived coherent vortices in the late flows, which may be used as a signature for identification of submarine wakes in oceanic thermocline. We also show that in a strongly stratified fluid the drag on the body due to radiating internal waves may be significant and cause the wake to receive an extra momentum.

1 Wake Classification and Theoretical Preliminaries

Wakes of self-propelled bodies in a stratified fluid continue to be of great interest in detecting and stealth of underwater bodies in the ocean thermocline. Although the information on wakes of towed bodies is voluminous, this information has limited utility in self-propelled body applications, given that there are fundamental differences between the ways where the momentum is imparted into the fluid (see Fig. 1). The wide differences between the wakes shown in Fig. 1 could be explained by the nature of the momentum forcing resulting from the body-fluid interaction. The only study that has dealt with these differences is [6], wherein a small submarine model with an externally forced jet in the aft, that is equivalent to a water-jet submarine cruising in a stratified fluid, was used. For steady motion the late wake signature was found to be weak (Fig. 1b) compared to momentum (overthrusted) wake (Fig. 1c). In some runs, the thrust (and body velocity) was rapidly increased, imparting more momentum into the wake, and the authors of [6] noted the emergence of large vortices. They agreed that these large long-lived vortices emerging during maneuvering can be a cue of identifying self-propelled bodies in stratified fluids. In our present study this concept has been explained theoretically and verified experimentally, using an actual self-propelled body.

The body-fluid interaction leads to momentum exchange between the two, which, as far as the late wake is concerned, can be considered as occurring

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*, Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_38,

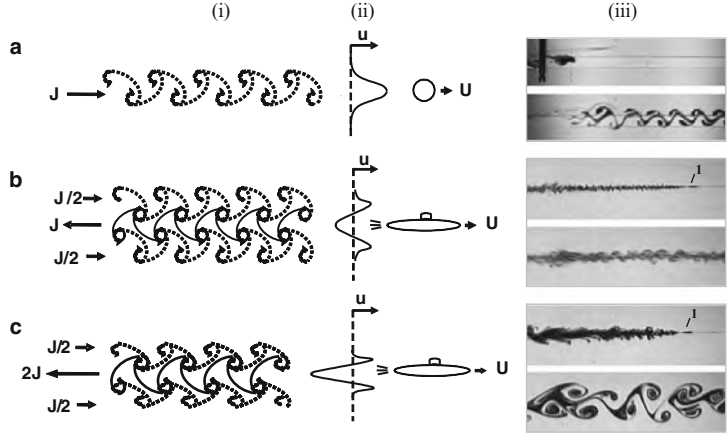


Fig. 1. In columns (*i-iii*) we show: (*i*) top view schematics, (*ii*) velocity profiles and (*iii*) wake patterns for three basic cases (**a**–**c**) of far wakes in a stratified fluid: (**a**) momentum wake behind a towed body or a jet, (**b**) zero-momentum and (**c**) momentum (*overthrusted*) wake behind a self-propelled body. Momentum fluxes J and $J/2$ are used only for simplicity and more accurately they will be defined later

in a compact area compared to the wake size. A straightforward case is the *momentum* wake of a towed body (Fig. 1a), where the momentum flux J imparted into the wake becomes the same as the drag on the body, the effect of which can be represented by a point momentum source [7]. When a self-propelled body is cruising steadily and the internal wave drag is negligible, the viscous and form drags and engine thrust are in balance and there is no net momentum imparted to the wake (Fig. 1b), resulting in a *zero momentum wake*. Forcing in such wakes is equivalent to a moving force doublet with zero net momentum [4]. Such wakes (Fig. 1b) decay much faster than the momentum wakes (Fig. 1a) and there are no large, persistent eddies formed in this case. When the body is in unsteady motion, the thrust and drag are not balanced (Fig. 1c), and forcing can be considered as a combination of a momentum source and a force doublet, the former decaying slower and leaving behind momentum wake with intense vortex street.

During a time interval $\triangle t$, when the body accelerates, it acquires a momentum $I = J\triangle t$ (J is the difference between the thrust reaction and drag forces on body) and the same momentum I with the opposite sign is transported to the fluid. The balance of (kinematic) momentum for a body of volume V moving with velocity U is

$$(1 + k)V dU/dt = J_T - J_D = J, \tag{1}$$

where k – virtual mass coefficient, J_T – thrust reaction *on the body*, $J_D = J_D^* + J_D^W$ – net drag *on the body* that includes the viscous and form drag J_D^* and the wave drag J_D^W as a result of momentum flux radiated as internal waves (which is not associated with narrow wake, in Fig. 2). For steady motion the

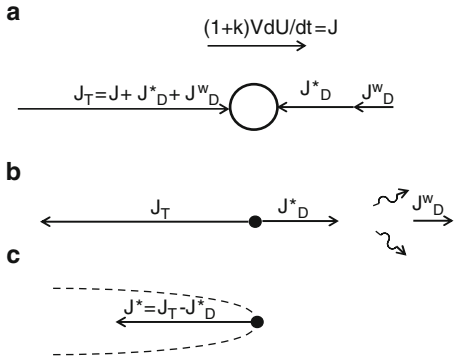


Fig. 2. Schematic showing forces acting on: (**a**) accelerating self-propelled body (*circle*), (**b**) surrounding fluid and (**c**) wake, shown by a *dashed line*

net momentum flux J applied to the fluid is zero. When the body accelerates, e.g. from U_- to U_+, it acquires momentum

$$I = \int^{\triangle t} Jdt = (1+k)V(U_+ - U_-) = (1+k)V\triangle U. \tag{2}$$

This gives the estimate $J \approx I/\triangle t$ for the momentum flux transported *to the fluid*. Note, that the *history of the body acceleration* is not important and the *flow momentum* I can be estimated if $\triangle U$ is known.

By definition, the wake behind a body is a narrow conical region (Fig. 2c) where the fluid motion is vortical, while outside this region the flow is practically potential. The wake intensity is characterized by the *wake momentum flux* J^* in this narrow region. The engine thrust J_T that needs to be supplied to overcome J_D, however, has to be imparted into the wake, which acquires a momentum flux

$$J^* = J_T - J_D^* = J + J_D^W. \tag{3}$$

Taking into account that J^* conserves and neglecting details that become unimportant at late times, one arrives at the conclusion that the action of accelerating self-propelled body on a fluid is equivalent to the action of momentum source of intensity (3) that acts impulsively during time interval $\triangle t$.

Numerous studies show that when a horizontal momentum source acts impulsively in a stratified fluid, large (compared to the source size), long living (compared to the duration of forcing) pancake-like dipolar (momentum) eddies are formed. These eddies have been extensively studied and general mechanisms of their formation and evolution explained in detail [5]. In particular, it was shown that such eddies develop in a self-similar regime and their horizontal length scale increases with time t as $D^* \approx I^{1/4}N^{1/12}t^{1/3}$ [8, 9]. To verify the analysis presented above, a series of experiments with real self-propelled body was conducted.

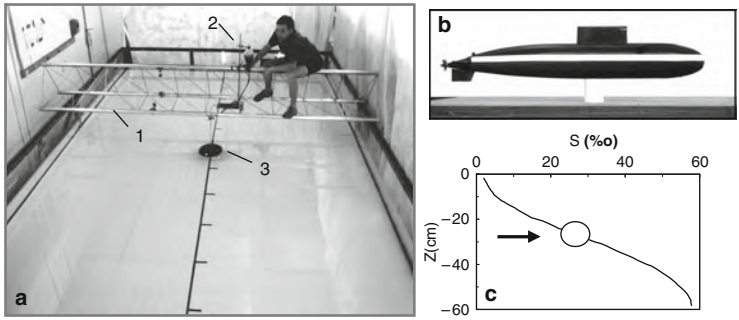


Fig. 3. In (**a**) – tank ($8 \times 4 \times 0.8$ m) with removable bridge (*1*) and device (*2*) to produce a dyed spot (*3*). In (**b**) – submarine model (length 75 cm, diameter $D_0 = 11$ cm). In (**c**) – typical vertical distribution of the density (salinity, S), *circle* – the body position, *arrow* – level where the dye spot or tracer particles (for PIV) were seeded

2 Experimental Set-Up

Experiments were conducted using scaled radio-controlled model and tank with linearly stratified ($N = 1.1 - 2.6\,\mathrm{s}^{-1}$) by salt water (Fig. 3). Dye visualization and PIV were used for flow diagnostics. For technical details see [2,8].

3 Results and Interpretation

Three basic cases were studied in which the model: (a) moves steadily, (b) starts from rest, and (c) moves steadily and then strongly/weakly accelerates.

In the experiment shown in Fig. 4, the model moves steadily. The flow pattern obtained during our previous zero-momentum wake experiments with small body [6] is shown in Fig. 1b, and one may expect a similar scenario in the present case. Nevertheless, the late-wake in a strongly stratified fluid demonstrated a different behavior, in that the PIV data show that the wake in Fig. 4 is with momentum. This can be explained by noting that under certain conditions the internal wave drag can become important [1]. Significant wave generation occurs when the body Froude number is close to one, and for Fig. 4, $Fr = U_+/D_0N = 0.8$, in which case the wave drag becomes comparable with the form and viscous drag [3]. To maintain the steady motion, therefore, the propulsion system should generate a thrust that is equal to the sum of the form, viscous and wave drags. Since internal gravity waves radiate momentum (related to the wave drag) away from the source region, the momentum flux associated with form and viscous drag remains unbalanced. Thus, the wake in accordance with (3) has a momentum flux $J^* = J_D^W$ and a vortex street (similar to that in Fig. 1c) is generated in the flow. Using the model [7], the

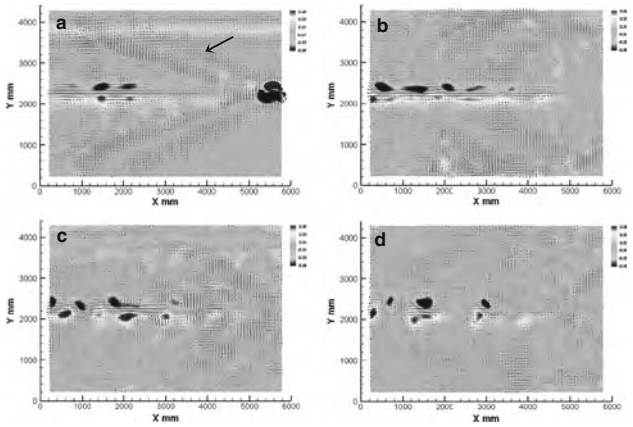


Fig. 4. PIV data showing the momentum wake behind a steady moving (from *left to right*) self-propelled body. Intense vortex street is clearly visible in the fields of velocity (*small arrows*) and vorticity (*different shades*). Internal waves (*large arrow*) are also visible in (**a**). Experimental parameters: Re=15,500, Fr =0.8, N=1.8 s^{-1}, Nt = 27 (**a**), 81 (**b**), 162 (**c**), 270 (**d**)

wake momentum flux (and thus the wave drag) can be estimated from the data of Fig. 4, which give $\lambda \approx 60$ cm for a typical wavelength of the primary vortex street. The estimates show [2] that in the considered case the wave drag is comparable ($\sim$ 80 %) to the viscous and form drag. Note, that there are no unusually large eddies in such momentum wakes. Situation, however, changes drastically when a body makes a maneuver, e.g., acceleration. This leads to generation of large eddies, which are illustrated below.

In the experiment shown in Fig. 5, the body starts from the rest, accelerates and then moves with constant velocity (Fig. 5f). During this maneuver the body deposits on the fluid a momentum of I. Initially, the propeller generates an intense jet-like flow with a sharp vorticity front that propagates away from the body. Soon, however, the buoyancy effects become important and the front collapses in the vertical direction and expands horizontally forming a pancake structure (Fig. 5b). The self-propagation velocity of the vorticity front is less than that of the fluid velocity of the jet-like flow behind the front. As a result, the (vertical) vorticity in this flow is advected to the front region, forming patches of concentrated vorticity of opposite signs in the form of conjunct 2–3 dipoles (Fig. 5c), which move in the background potential dipolar flow induced by the pressure forces and merge together (Fig. 5d) forming a large dipole of the size $D^* \approx 115$ cm (Fig. 5e, g) that is much larger than the model diameter. Similar results were obtained in the case when the body moves steadily and then strongly/weakly accelerates (not shown).

Thus, in all considered cases the acceleration of a real self-propelled body, moving with high Re number in a stratified fluid, leads to impulsive momentum wake with a system of eddies, which merge and asymptotically form large

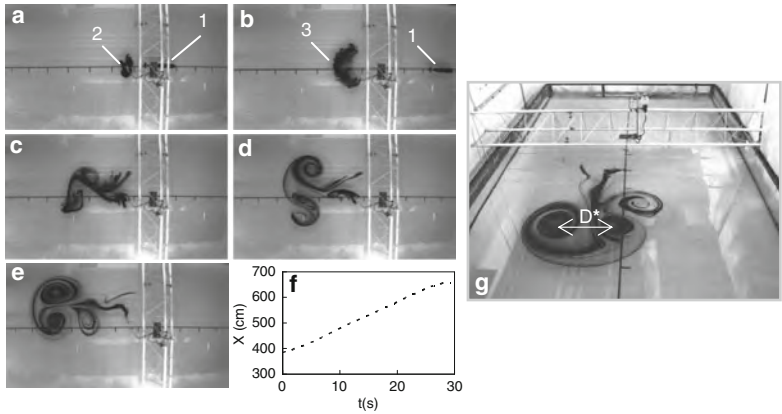


Fig. 5. In (**a**) – the body (*1*) is at rest and positioned behind a dyed spot (*2*). In (**b**) – the body (*1*) starts moving to the right, generating the vorticity front (*3*); as time progresses, after a number of bifurcations (**c**), (**d**), a large dipolar eddy is forming in the late wake (**e**). In (**f**) – the position X of the body as a function of time. In (**g**) – an enlarged image of the resulting eddy at $Nt = 630$. Experimental parameters: $Re = 10{,}500$, $I = 110{,}000\,\mathrm{cm^3 s^{-2}}$, $Nt = 0$ (**a**), 16.5 (**b**), 140 (**c**), 230 (**d**), 630 (**e**)

and long-lived dipolar vortex. Comparison of measurements with the model predictions show [2] that in all cases of the body acceleration, the size D^* of the resulting eddy at late times may be correctly calculated using the proposed parameterization.

Acknowledgements

This study was supported by the Office of Naval Research. We are grateful to J.C.R. Hunt who supported S.I.V. in attending ECMI 2008 Conference.

References

1. Gorodtsov, V.A., Teodorovich, E.V.: Fluid Dyn. **17**, 893 (1982)
2. Morrison, R.J.: Studies on wake signatures in a shallow or stratified fluid: Laboratory experiments and phenomenological modeling M.S. Thesis, Arizona State University, Tempe (2006)
3. Scase, M.M., Dalziel, S.B.: J. Fluid Mech. **498**, 289 (2004)
4. Smirnov, S.A., Voropayev, S.I.: Phys. Lett. A **307**, 148 (2003)
5. Voropayev, S.I., Afanasyev, Y.D.: Vortex Structures in a Stratified Fluid. Chapman & Hall, London (1994)
6. Voropayev, S.I., McEachern, G.B., Fernando, H.J.S., Boyer, D.L.: Phys. Fluids **11**, 1682 (1999)
7. Voropayev, S.I., Smirnov, S.A.: Phys. Fluids **15**, 618 (2003)
8. Voropayev, S.I., Fernando, H.J.S., Smirnov, S.A., Morrison, R.: Phys. Fluids **19**, 076603 (2007)
9. Voropayev, S.I., Fernando, H.J.S., Morrison, R.: Phys. Fluids **20**, 026602 (2008)

Eddy Dynamics Near Sharp Interfaces and in Straining Flows

J.C.R. Hunt[1], I. Eames[1], and J. Westerweel[2]

[1] University College London, Torrington Place, London WC1E 7JE, UK
i.eames@ucl.ac.uk, jcrh@cpom.ucl.ac.uk
[2] Delft University of Technology, Mekelweg 2, 2628 CD Delft, The Netherlands
J.Westerweel@tudelft.nl

Summary. Idealised models of eddy motion near sharp interfaces, such as shear layers and density interface, are examined. Strong shear layers can block the vertical motion of eddies with part of their impulse permanently transferred to the shear layer. Eddies which pass through the shear layer communicate a fraction of their impulse to the shear layer. Throughout these processes, the far field dipole moment (and total impulse) is conserved. For sharp density interfaces, vertical impulse is not globally conserved and the kinetic energy of the vortex goes towards generating waves. Here, both blocking and sheltering occurs, since the eddy vertical motion is constrained and the dipole moment is reduced. In straining flows, we show that the rotation and rapid amplification/suppression of the vortex impulse leads to an upscale transport in impulse, but the energy of the vortex hardly changes. Vorticity annihilation caused by diffusion can partially or completely destroy the colliding vortex patches causing the energy to decrease. The general relevance of these results is discussed.

1 Introduction

Recent experimental measurements and numerical simulations are changing our perceptions about the central mechanisms of turbulence. This affects how turbulence might be described and modeled. Following Osborne Reynolds and the influences of the kinetic theory of gases and statistical physics, the flow field is usually divided into a mean and fluctuating component, the latter being analysed as a perturbation to the former. However this is neither a good physical description nor an accurate basis for mathematical analysis of the key aspects of turbulence. These interfaces, which tend to lie at the outer edges of inhomogeneous turbulent flows Hunt et al. [2] or intermittently within turbulent flows Kaneda [6], or to be formed between layers of contrasting density, are not like interfaces betweenscalar variables (temperature etc).

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*,
Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_39,

This is because the intense vorticity in the layers, created by their distortion or displacement, affects the flow on either side of them through two main mechanisms: first kinematic blocking and distorting of the eddies tends to impede their motion through the interfaces Hunt [3]. Second, the presence of body forces tend to shield the region above the layer from turbulence beneath.

In this paper we first develop idealised models to study the interactions between eddies and the interfaces that exist between regions of a fluid with contrasting properties, such as mean velocity or density. These calculations provide a clear mechanistic view of the processes that occur near interfaces (using robust concepts such as integral invariants) or between vortices, and is a different approach than the statistical methodology usually applied in turbulence studies.

2 Vortical Structures Interacting with Thin Shear Layers and Stratification

We consider a dipolar vortex or eddy of area A, velocity $\boldsymbol{U}$ moving near an interface as shown in Fig. 1. Many of the processes that occur near interfaces such as shear layers and density stratification can be broadly understood by computing impulse and energy, especially if these quantities are invariant. The vortex impulse is $\boldsymbol{I}_M = \rho(1 + C_M)A\boldsymbol{U}$ where ρ is the fluid density and $C_M \sim 1$ is the added-mass coefficient. The impulse of the system of vortices is $\boldsymbol{I}_M = \int_{A_\infty} \rho(y, -x)\omega \mathrm{d}A$, where $\boldsymbol{I}_M = (I_{Mx}, I_{My})$. The vortex velocity has two components arising from the local velocity of the flow and a self-induced component. The dipole moment of the flow created by the vortex and the perturbation to the vortex sheet is

$$2\pi\boldsymbol{D} = \boldsymbol{I}_M/\rho + \int_{A_\infty - A} (\boldsymbol{X} - \boldsymbol{X}_0) \times \boldsymbol{\omega} \mathrm{d}A, \tag{1}$$

where the integration is taken over the region outside the vortex. The second term on the right-hand side of (1) is the impulse of the interface (vortex sheet or isopycnal surface) $\boldsymbol{I}_I/\rho$ expressed in terms of the Lagrangian displacement of fluid elements $\boldsymbol{X}$ from their initial position $\boldsymbol{X}_0$. The kinetic energy, K, is defined as $K = \int_{V_\infty} \rho\frac{1}{2}|\boldsymbol{u}|^2 \mathrm{d}A$. The interface is displaced locally by a vertical distance Y and the deformation is characterised in terms of the area of dense fluid lifted and the moment of area are defined by $\mathcal{D}_Y = \int_{-\infty}^{\infty} Y \mathrm{d}x$, $\mathcal{D}_{YY} = \int_{-\infty}^{\infty} \frac{1}{2}Y^2 \mathrm{d}x$ respectively.

The interaction between a dipolar vortex and shear layer (with Γ circulation per unit length), is characterised by $\mathcal{S} = \Gamma/R_0U_0$. The dipole moment is invariant even when the shear layer is unstable and viscous effects are present. The velocity of the vortex below the shear layer is $\boldsymbol{U} = \frac{1}{2}\Gamma\hat{\boldsymbol{x}} + \boldsymbol{U}_I$, where $\boldsymbol{U}_I$ is the self-induced velocity of the vortex. Hunt [4] studied the interaction of a weak vortex with a strong shear layer ($\mathcal{S} \gg 1$). The disturbance created

by the vortex tends to decrease the local growth rate of perturbations on the shear layer. For $\mathcal{S} \ll 1$, the eddy is sufficiently strong that it can pass through the vortex sheet and the change in the vortex impulse is

$$\boldsymbol{I}_M - \boldsymbol{I}_{M0} = -D_Y \Gamma \rho \hat{\boldsymbol{x}}. \tag{2}$$

The velocity of the vortex and direction is changed by passing through the layer. For long-time, the disturbance of the vortex sheet a component of the impulse of the vortex and can combine with the vortex, moving aay.

The disturbance caused by a vortex impacting on a thermocline is strikingly different from the case of shear layers because of the ability to store potential energy, and the rapid propagation of waves on the interface. The interaction between a vortex and a sharp density interface (where the density contrast is $\Delta\rho$) is characterised by a Froude number $\mathcal{F} = U_0/\sqrt{\Delta\rho g R_0/g}$. The action of a baroclinic torque on thin density interfaces, increases the distribution of circulation on the isopycnal surface, $\Gamma(=\int_-^+ \omega \mathrm{d}n)$, at a rate $\mathrm{d}\Gamma/\mathrm{d}t = \Delta\rho(t) g \hat{\boldsymbol{y}} \times \hat{\boldsymbol{n}}/\rho$. The impulse associated with the vorticity on the interface is $\boldsymbol{I}_I$. When the vortex moves towards the interface, $\boldsymbol{I}_I \cdot \hat{\boldsymbol{y}} < 0$ because the interface is displaced upwards. This means that the total vertical impulse decreases as the eddy approaches the interface and the dipole moment in the far field is reduced. The total momentum or impulse of the flow, $\boldsymbol{M}$, decreases at a rate, $\mathrm{d}\boldsymbol{M}/\mathrm{d}t = -D_Y \Delta\rho g \hat{\boldsymbol{y}}$ but the total energy is conserved, so that $\mathrm{d}(K + \Delta\rho g D_{YY})/\mathrm{d}t = 0$.

For $\mathcal{F} < 1$, the vortex is blocked by the interface. The maximum deflection H can be estimated by relating the initial kinetic energy of the vortex to the potential energy of the deflection when the vortex has been brought momentarily to rest:

$$\frac{1}{2}\rho(\alpha + C_M) U_0^2 A_0 \sim \Delta\rho g D_{YY} = \lambda \frac{1}{2} \Delta\rho g H^n R_0^{2-n}, \tag{3}$$

where the left-hand side represents the initial kinetic energy of the eddy and $\alpha \sim 1$. For $\mathcal{F} \sim 1$, $D_{YY} \sim H R_0^2$ and $H/R_0 \sim \mathcal{F}^2$ as confirmed by Linden [7]. But as the strength of the stratification increases $\mathcal{F} \to 0$, $D_{YY} \sim H^2 R_0$ and $H/R_0 \sim \mathcal{F}$.

When $\mathcal{F} \gg 1$, the vortex kinetic energy is reduced as it passes through the interface and its vertical speed decreases by a fraction $\Delta U/U_0 = -\Delta\rho g D_{YY}/(1 + C_M) A_0 U_0^2 \sim 0.5\mathcal{F}^{-2}$. The vertical impulse of the deformed isopycnal surface scales as $-\Delta\rho g R_0^2/U_0 \sim -|\boldsymbol{I}_M|\mathcal{F}^{-2}$ and leads to the generation of a vertical vortex (traveling downwards). The vortex which passes through the interface is now denser than the ambient fluid and decelerates. The rate of decrease of the momentum of the vortex is $\mathrm{d}\rho(1 + C_M) U A_0/\mathrm{d}t = -\Delta\rho g A_0$. There is a flux of mass from the vortex which causes its area to decrease. Maximum height the front reaches is $H/R_0 = U_0^2(1 + C_M)/\Delta\rho g R_0/\rho \sim (1 + C_M)\mathcal{F}^2$. The final state is the conversion of the deformation of the isopycnal surface to a series of waves which communicate energy to the far field and the horizontal component of impulse.

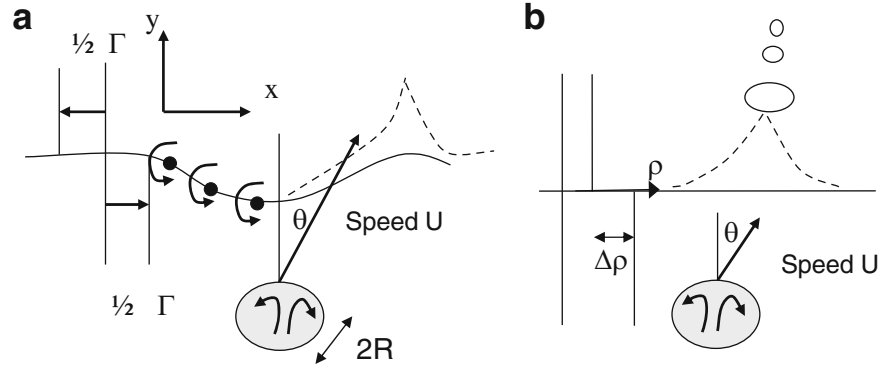


Fig. 1. Schematic of a dipolar vortex impacting on (**a**) thin shear layer and (**b**) a sharp density interface

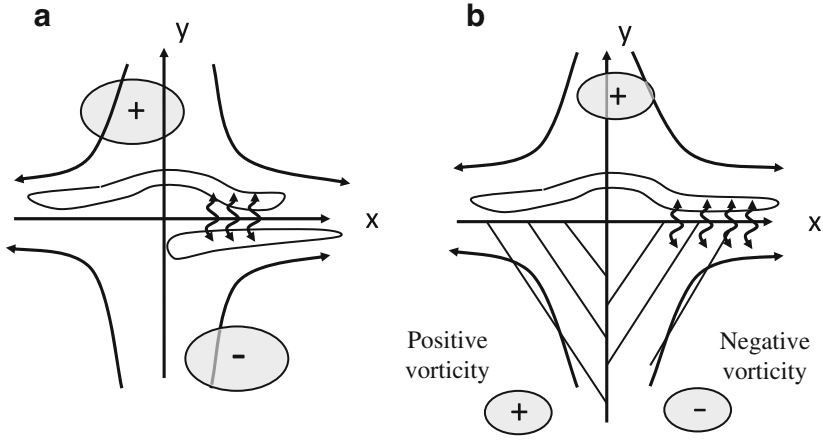


Fig. 2. Schematic of vortex patches interacting in a straining flow. In (**a**) two different signed vortices are squeezed together and diffusion leads to partial or complete annihilation. (**b**) A vortex with positive vorticity is swept past a stagnation point and partially cancels a vortex with negative vorticity

3 Vortical Interactions in Straining Regions

Decaying two-dimensional turbulence is characterised by upscale transport, usually manifested by the emergence of a small number of large-scale vortices which are well separated. The eventual impulse and angular momentum are determined by statistical variations in the initial forcing (as shown by Davidson [1]). The upward transport of energy is accomplished by the merging of like signed vorticity and the stretching and dissipation of weaker vortices. The removal of weaker vortices is due to the interplay of intervortical straining accompanied by vorticity annihilation and the degradation of the impulse

of vortex pairs is important as we shall show. This has important implications for the persistence of dipolar vortices within a decaying two-dimensional turbulent flow.

Consider an arbitrary distribution of vorticity in a straining field, $\boldsymbol{u}_E = (\alpha x, -\alpha y)$. In the absence of diffusion the rate of change of the impulse of the vortical field (when the flow induced by the weak vortices is much smaller than the external straining flow) is

$$\frac{\mathrm{d}\boldsymbol{I}_M}{\mathrm{d}t} = \int_{A_\infty} \rho \boldsymbol{u}_E \times \omega \hat{\boldsymbol{z}} \mathrm{d}A = \alpha(-I_{Mx}, I_{My}). \tag{4}$$

The impulse of the vortex changes according to

$$\boldsymbol{I}_M = (I_{Mx}(0)e^{-\alpha t}, I_{My}(0)e^{\alpha t}). \tag{5}$$

The straining flow increases the vertical component of impulse, while decreasing the horizontal impulse. The direction of impulse is rotated – this rotation is compensated by an opposite rotation of the linear impulse of the exterior flow to ensure impulse is globally conserved. Although the growth rate of impulse is exponential for $I_{Mx}(0) \neq 0$, the rate of change of kinetic energy is

$$\frac{\mathrm{d}K}{\mathrm{d}t} = \int_A \rho \boldsymbol{u} \cdot \frac{\mathrm{D}\boldsymbol{u}}{\mathrm{D}t} A. \tag{6}$$

When the flow induced by the vorticity field is small, $\mathrm{d}K/\mathrm{d}t = \int_{A_\infty} \rho \boldsymbol{u}_E \cdot (\boldsymbol{u}_E \times \omega \hat{\boldsymbol{z}}) \mathrm{d}A = 0$. So while the impulse of the flow can dramatically increase, the change in energy is negligible.

The effect of diffusion introduces additional new physics. Fig. 2a shows two patches of vorticity (of circulation Γ_1 and Γ_2) which are pushed together in a linear straining flow. The initial horizontal impulse is large and the vertical impulse is small. The total circulation in the flow, $\Gamma = \int \omega \mathrm{d}A$, is invariant even when diffusive effects are important. The diffusive interaction between these opposite signed patches of vorticity leads either to partial or complete to vorticity cancellation.

When the vortices move in a straining flow created, for instance, at the front of a large moving dipolar vortex. The straining flow below the $y = 0$ line contains vorticity (of two signs). Patches of vorticity impact the front stagnation plane and are stretched. As incident patches impact on the plane, their diffusion into the region adjacent to the stagnation plane leads either to the cancellation or addition to the vorticity field.

4 Conclusions

We have studied the interaction between eddies and interfaces that occur in inhomogeneous turbulence and environmental flows. By applying local or global integral measures, such as impulse, circulation or energy, some of

which are invariant (see [5]), we have been able to understand (from a new standpoint) the complex processes that accompany two dimensional eddies interacting with interfaces.

The analysis of two-dimensional vortices moving in a straining region has shown that the change of the kinetic energy with time of the vortices is either negligible or it decreases (when diffusive effects are important). This means, generally, that kinetic energy moves upscale to larger eddies. But, for three-dimensional flows, stretching of vortex tubes can lead to an increase of the kinetic energy of small scale structures and that both upscale and downscale transport of energy occurs.

References

1. Davidson, P.A.: J. Fluid Mech. **580**, 431–450 (2007)
2. Hunt, J.C.R., Eames, I., Westerweel, J.: J. Fluid Mech. **554**, 499–519 (2006)
3. Hunt, J.C.R.: J. Fluid Mech. **61**, 625–706 (1973)
4. Hunt, J.C.R., Durbin, P.A.: Fluid Dyn. Res. **24**, 375–404 (1999)
5. Hunt, J.C.R., Delfos, R., Eames, I., Perkins, R.J.: Flow Turbul. Combust. **79**, 155–174 (2007)
6. Kaneda, L.F.: Proceedings of the IUTAM Symposium on Computational Physics and New Perspectives in Turbulence. Springer Science, Dordrecht (2008)
7. Linden, P.F.: J. Fluid Mech. **60**, 467–480 (1973)

Evolution and Run-Up of Tsunamis

C.A. Klettner[1], I. Eames[1], J.C.R. Hunt[1], and H.J.S. Fernando[2]

[1] University College London, Torrington Place, London WC1E 7JE, UK,
christian.klettner@ucl.ac.uk, i.eames@ucl.ac.uk, jcrh@cpom.ucl.ac.uk
[2] Arizona State University, Tempe, Arizona 85287-9309, USA,
j.fernando@asu.edu

Summary. The evolution of a tsunami from generation, propagation to coastal regions and beach run-up is studied. The effect of the initial profile on how a tsunami evolves as it is propagating over uniform depth is studied numerically. As the wave moves into shallower water its form changes and by applying momentum conservation and dimensional analysis, predictions for the speed, height and run-up of surges up beaches can be made. Theoretical predictions are compared with laboratory experiments and field observations.

1 Introduction

A tsunami is a series of waves created when a body of water is rapidly displaced. The initial profile is dependant on what exactly has caused the perturbations in the sea surface. These processes include (a) when the ocean bed is displaced due to an undersea earthquake which can give rise to elevated and depressed components, (b) landslides (e.g. due to a volcano erupting) which fall into the ocean which are associated mainly with an elevated sea surface and (c) undersea landslides which result in mainly depressed sea surfaces (see Hunt [7]). The Indian Ocean tsunami on Boxing Day 2004 was caused by the Indian–Australian plate subducting under the Eurasian/Andaman plate resulting in an elevation wave hitting Sri Lanka and a leading depression wave hitting Thailand (i.e. (a) above).

The purpose of this paper is to explore new approaches to provide practical estimates of how an initial disturbance evolves and propagates into shallow waters and runs up beaches. The method of integral invariants is reviewed and related to how long waves such as tsunamis propagate. The decay rate of these waves is studied numerically by solving the Korteweg-de Vries equation. Using conservation principles and scaling arguments, the characteristics of the surge up the beach are estimated and compared with laboratory experiments carried out in a wave tank and field observations.

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*,
Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_40,

2 Integral Invariants

Integral invariants can be applied to analyse non-linear phenomena in inertially dominated flows. In the context of vortical and turbulent flows, local or global integral measures such as helicity, circulation and impulse may be conserved (even for viscous flows) and can give insight into how the flow evolves with time. For free surface waves with surface tension which generate a potential flow, Benjamin and Olver [1] established eight conserved quantities including mass (3D)/ cross-sectional area (2D) (A), momentum (M) and energy (E). For inviscid flows (with vorticity in the fluid interior), Longuet–Higgins [11] showed that the number of invariants reduced to mass/cross-sectional area, momentum and energy. For viscous flows, the two principle invariants are mass and momentum since kinetic energy is dissipated. These are defined in terms of wave amplitude (y) and horizontal velocity (u) as:

$$A = \int_{-\infty}^{+\infty} y \mathrm{d}x, \quad M = \int_{-\infty}^{+\infty} \int_{-h_0}^{y} \rho u \mathrm{d}y \mathrm{d}x, \tag{1}$$

where h_0 is the uniform water depth. Evaluating A, M and E depends on the manner in which the initial waves are generated. For rock falls or subsurface landslips, M can be estimated from the impulsive force generated by such movement, with A either determined from the volume of rock splashing into the water or is zero for landslides.

3 Generation and Propagation of an Initial Disturbance

We analyse how a perturbation, of initial amplitude a_0 and length L_0, to a two-dimensional free surface evolves as it propagates over the ocean. We simplify the problem and use a long-wave approximation which includes dispersion and wave steepening. The free surface disturbance $y(x,t)$ moving with the long-wave speed ($c_0 = \sqrt{gh_0}$) can be described by [10]:

$$\zeta_\tau + (3\zeta^2)_X + \zeta_{XXX} = 0, \tag{2}$$

where the wave height, position (relative to a frame moving with the long wave speed) and time are non-dimensionalised according to $\zeta = y/a_0$, $X = (3a_0/2h_0)^{1/2}(x - c_0t)/h_0$ and $\tau = (\sqrt{3}/2)(c_0t/h_0)(a_0/2h_0)^{3/2}$. In this formulation $\alpha \approx O(\beta) \ll 1$, where $\alpha = a_0/h_0$ and $\beta = (h_0/L_0)^2$ are a measure of the non-linearity and dispersiveness respectively. Numerical solutions to (2) were obtained using a finite difference scheme, [4].

The long-time evolution of a wave depends on its initial cross-sectional area $\tilde{A} = (A/a_0h_0)(2h_0/3a_0)^{1/2}$. For $\tilde{A}>0$, a soliton is always generated; for $\tilde{A}<0$ no soliton is generated and the initial wave ultimately develops into a dispersive wave train. It is not possible to predict a priori the long time development of $\tilde{A} = 0$, Hammack and Segur [6]. These three cases were studied

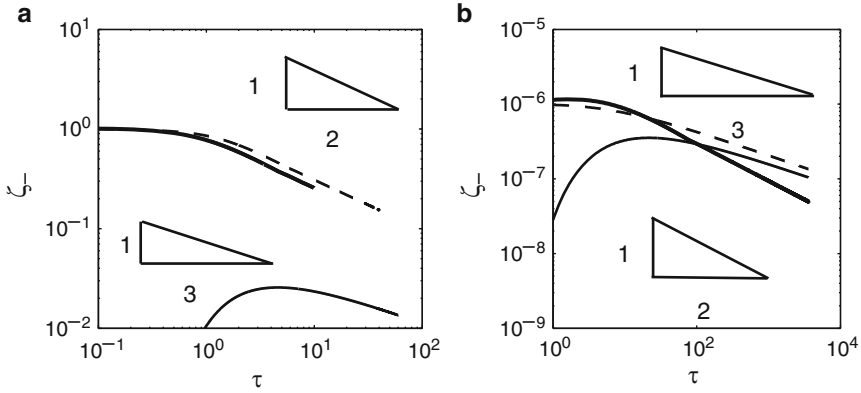


Fig. 1. Evolution of maximum depression for $\tilde{A} = 0$ (*thick line*), $\tilde{A}>0$ (*thin line*) and $\tilde{A}<0$ (*dashed line*) with $D = 3$ for (**a**) ζ_0=1 (*non-linear regime*) and (**b**) $\zeta_0 = 10^{-6}$ (*linear regime*). The *lines* with slopes $-1/3$ and $-1/2$ are indicated in each figure

using a free surface deformation of (a) $\zeta(X,0) = 2.3\zeta_0 X/D \exp(-X^2/D^2)$, (b) $\zeta_0 \exp(-X^2/D^2)$, (c) $-\zeta_0 \exp(-X^2/D^2)$ for which $\tilde{A} = 0$, $\tilde{A}>0$ and $\tilde{A}<0$ respectively. Since the aim is to understand how the wave amplitude changes, we follow the maximum positive and negative amplitudes defined by:

$$\zeta_+(\tau) = \max_{-\infty<X<\infty} \zeta(X,\tau), \qquad \zeta_-(\tau) = \left| \min_{-\infty<X<\infty} \zeta(X,\tau) \right|. \tag{3}$$

When a soliton is present, the maximum positive amplitude tends to a constant. The magnitude of the depression (ζ_-) is a useful metric for the dispersive wave component, particularly as we shall see, as the time before the tsunami interacts with the coast may be so short that the soliton has not had time to emerge. It takes a long time for this to emerge and is not generally evident in reported wave surface signatures, Constantin and Johnson [3].

Two contrasting cases were investigated by fixing $D = 3$ and varying ζ_0 with $\zeta_0 = 1$ (nonlinear regime, Fig. 1a) and $\zeta_0 = 10^{-6}$ (linear regime, Fig. 1b). In the non-linear regime, $\alpha \sim O(\beta) \ll 1$ and the numerical results show that $\zeta_- \sim \tau^{-1/2}$ for ($\tilde{A} = 0, <0$) while $\zeta_- \sim \tau^{-1/3}$ for ($\tilde{A}>0$). In the linear regime, $\zeta_- \sim \tau^{-1/2}$ (for $\tilde{A} = 0$) and $\zeta_- \sim \tau^{-1/3}$ (for $\left|\tilde{A}\right| \neq 0$).

The similarity solution to the linear KdV equation derived by Miles [12] is:

$$\zeta(X,\tau) = \tilde{A}\tau^{-1/3}\mathrm{Ai}(\tau^{-1/3}X) - \langle X\zeta_0(X,0)\rangle\, \tau^{-2/3}\mathrm{Ai}'(\tau^{-1/3}X) + ... \tag{4}$$

Here, Ai is the Airy function and $\langle f(X)\rangle = \int_{-\infty}^{\infty} f(X)\mathrm{d}X$. Equation (4) predicts that the magnitude decreases as $\tau^{-1/3}$ for $|\tilde{A}|>0$ and $\tau^{-2/3}$ for $\tilde{A} = 0$.

Our numerical calculations support Miles' analysis for long wave linear processes but not for the case of $\tilde{A} = 0$, when the similarity method fails to pick up the leading wave component.

It is more useful to look at the Fourier solution to the linear KdV equation, i.e. $\zeta = \int_0^\infty \hat{\zeta}(k) e^{i(kX+\omega\tau)} \mathrm{d}k$. Using the method of steepest descents, the dominant contribution arises from $X/\tau = -\mathrm{d}\omega/\mathrm{d}k = -3k^2$. The dispersive wave component propagates with a constant speed in the negative X-direction. In the frame moving with the dispersive wave (where $-X/\tau$ is constant), $\zeta \cong \hat{\zeta}(k)\sqrt{2\pi} e^{-i2Xk} / \sqrt{6\,|k|\,\tau}$ and $\zeta_- \sim \tau^{-1/2}$. This explains why the decay rate increases when $\tilde{A} = 0$.

4 Movement into Shallow Waters and Interaction with the Beach

As the wave train moves into a shallow coastal region, of depth h_c, conservation of energy requires that the amplitude of the wave increases by a factor $(h_0/h_c)^{1/4}$ (Green's Law), Synolakis [13]. To link the wave dynamics in the shallow water region, up to the beach run-up, we use conservation of momentum. The momentum associated with a leading wave M_L can be interpreted in terms of an added-mass coefficient (C_M) and cross-sectional area A_L through $M_L = \rho C_M A_L c$. For small amplitude waves on deep water, the added-mass coefficient $C_M \sim a_c/L$ so that $M_L \sim \rho a_c^2 c$. But for large amplitude waves on shallow water, the amplitude is so large that $C_M \sim O(1)$ and $M_L \sim \rho A_L c$.

The momentum, M_L, is approximately conserved during the generation of the bore so that the initial bore height a_{BI} and length L_{BI} are related by $M_L = M_B = \rho\sqrt{g a_{BI}} L_{BI} a_{BI}$ (for both elevated and depressed waves). Writing $\lambda = a_{BI}/L_{BI} \ll 1$, by dimensional analysis, the initial speed and amplitude of the bore are:

$$u_{BI} = \lambda^{1/5} (M_L/\rho)^{1/5} g^{2/5}, \qquad a_{BI} = \lambda^{2/5} (M_L/\rho)^{2/5} g^{-1/5}. \tag{5}$$

The component of gravity along the inclined beach (of slope α_b) slow downs the surge (amplitude a_B, length L_B and velocity u_B), as it moves up the beach (of slope α_b):

$$\frac{\mathrm{d}M_B}{\mathrm{d}x} = -\rho\alpha_b (a_B L_B) g/u_B = -\alpha_b (M_L/\rho)^{1/3} L_B^{2/3} g^{1/3}. \tag{6}$$

The 'run-up' distance B_R from where the bore was created to where the surge stops ($M_B = 0$) is:

$$B_R = \frac{5}{3} (M_L/\rho)^{2/5} g^{-1/5} / \alpha_b^{3/5}. \tag{7}$$

Similar studies have been undertaken by Carrier et al. [2] using the non-linear shallow water equations.

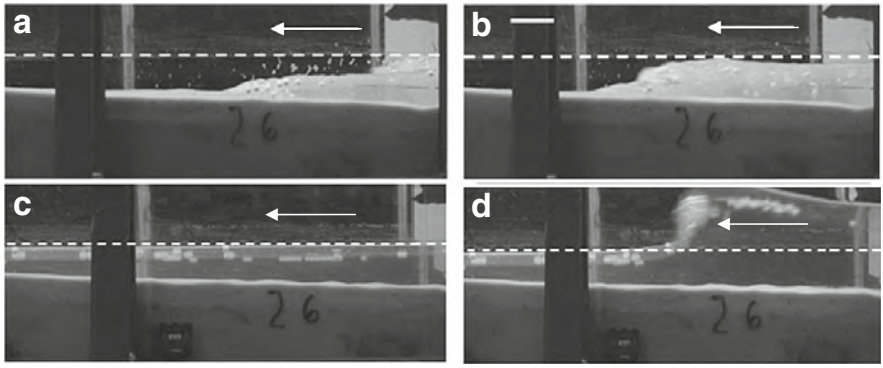


Fig. 2. (**a**) and (**b**) showing a leading depression wave surge propagating up the beach. (**c**) and (**d**) show an elevated wave just before breaking. *Dotted lines* indicate the initial water level. The *solid white line* is 5 cm

A laboratory study was undertaken to test the relationships (5) and (7). Experiments where performed in a wave tank as described in detail by Voropayev et al. [14]. Figure 2a, b shows the surge of a depression wave as it moves up the beach. The surge is below the initial height of the free surface (indicated by the dotted line) because a shoaling depression wave creates a shoreline recession before breaking. In contrast, Fig. 2c, d show an elevation wave just about to break. In this case there has been no shoreline recession and the wave breaks very close to the original shoreline. Elevated waves typically break further up the beach than depressed waves. Measurements from these experiments were favourable with the predictions given by (5) and (7) (see Klettner et al. [9]).

5 Discussion

Post event estimates of the 2004 Indian Ocean tsunami suggest that the sea bed had an uplift of $a_0 \sim 6\,\text{m}$ and subsidence of similar magnitude over a length of $\sim 1,200\,\text{km}$ and a width of $100 - 150\,\text{km}$ over a period of 500 s, Grilli et al. [5].

The wave propagated $\sim 600\,\text{km}$ to Thailand, in a time of $\sim 6,000\,\text{s}$. The characteristic timescale used in (2) is sensitive to h_0; taking into account the variation of h_0 with distance to Thailand, we estimate that the wave evolves over a time $\tau \sim 10$. During this time the wave amplitude decreases due to radial spreading and dispersive decay. As the tsunami was generated along a length of $1,200\,\text{km}$ and only travelled $600\,\text{km}$ to Thailand the decay in amplitude caused by radial spreading is negligible. The initial tsunami wave length and height corresponds in dimensionless variables to $\tilde{A} = 0$, $D = 3$ and $\zeta_0 = 1$, [5]. Figure 1a shows that the amplitude decayed to $\sim 0.3a_0$ in a time $\tau \sim 10$. Movement into shallower water of $h_c \sim 12\,\text{m}$ increased the amplitude

by a factor of 3. Taking these two factors into account the initial amplitude will be $\sim 5.6\,\mathrm{m}$. The yacht Mercator recorded an amplitude of approximately $3-4\,\mathrm{m}$ for the leading wave train components, [5].

Using the invariants discussed in Sect. 2, we relate the properties in the coastal regions to how the wave creates a bore and moves up the beach. The leading wave component, estimated from the Mercator data, had a height and length of approximately 3 m and 7.2 km respectively. The momentum associated with the leading wave component is $M_L = \rho A_L c \sim 10^7\,\mathrm{Nm}^{-1}\mathrm{s}$. From (5), the initial surge velocity $u_{BI} \sim 8\,\mathrm{ms}^{-1}$ which compares well with the $6-8\,\mathrm{ms}^{-1}$ field estimates over land by Kawata et al. [8]. The estimated vertical run-up distance of the surge ($\alpha_b B_R$) is $\sim 11\,\mathrm{m}$ which is comparable to field measurements of $3-11\,\mathrm{m}$, [8]. [It should be noted that our run-up estimate also includes the vertical distance from where the original shoreline recedes to (due to shoreline recession) and the original shoreline.]
The use of invariants in linking together the evolution of a tsunami from its generation to beach run-up provides a new method of making practical estimates of the ultimate fate of a wave disturbance. We used volume/area to classify the evolution of a tsunami, energy to link the change in height between ocean/coastal regions and momentum during run-up.

Acknowledgements

Sridhar Balasubramanian is gratefully acknowledged for his help with the experiments.

References

1. Benjamin, T.B., Olver, P.J.: J. Fluid Mech. **125**, 137–185 (1983)
2. Carrier, G.F., Wu, T.T., Yeh, H.: J. Fluid Mech. **475**, 79–99 (2003)
3. Constantin, A., Johnson, R.S.: J. Phys. A **39**, L215–L217 (2006)
4. Feng, B.F., Mitsui, T.: J. Comput. Appl. Math. **90**, 95–116 (1998)
5. Grilli, S.T., Ioualalen, M., Asavanant, J., Shi, F., Kirby, J.T., Watts, P.: J. Waterway Port Coastal Ocean Eng. **133**, 414–428 (2007)
6. Hammack, J.L., Segur, H.: J. Fluid Mech. **65**, 289–314 (1974)
7. Hunt, J.C.R.: Mathematics Today, October, 144–146 (2005)
8. Kawata, Y., Tsuji, Y., Sugimoto, Y., Hayashi, H., et al.: http://www.tsunami.civil.tohoku.ac.jp/sumatra2004/report.html (2005)
9. Klettner, C.A., Balasubramanian, S., Hunt, J.C.R., Fernando, H.J.S., Voropayev, S.I., Eames, I.: in submission
10. Korteweg, D.J., de Vries, G.: Philos. Mag. Ser. **39**, 422–443 (1895)
11. Longuet-Higgins, M.S.: J. Fluid Mech. **134**, 155–159 (1983)
12. Miles, J.W.: J. Fluid Mech. **87**, 773–783 (1974)
13. Synolakis, C.E.: Phys. Fluids **87**, 490–491 (1991)
14. Voropayev, S.I., Testik, F.Y., Fernando, H.J.S., Boyer, D.L.: Ocean Eng. **30**, 1647–1667 (2003)

Interfacial Mixing by Horizontal Vortices and Shear Turbulence

J. B. Flór, E.H. Hopfinger, and E. Guyez

Laboratoire des Ecoulement Geophysiques et Industriels LEGI-CNRS, B.P.53X, 38041 Grenoble Cedex 09, France, Jan-Bert.Flor@hmg.inpg.fr, Emil.Hopfinger@legi.grenoble-inpg.fr, E.M.C.Guyez@warwick.ac.uk

Summary. In this paper the entrainment rates of shear flows and turbulent Taylor vortices are considered in the light of the mixed-layer deepening in oceans, seas and lakes. The entrainment by Taylor vortices can be considered to represent a model for entrainment by continuously driven horizontal vortices such as Langmuir circulation. When the forcing is related to the surface wind stress, the interfacial entrainment by Langmuir vortices can be estimated from the experimental results on turbulent Taylor vortices, and compared with the entrainment rates in pure shear-flows. The results indicate that mixing by Langmuir vortices is important for all Richardson numbers, Ri_* (based on wind-induced surface stress velocity), and that for $Ri_* > 80$ Langmuir vortices dominate the mixed-layer deepening.

1 Introduction

The mixing of the surface layer of the ocean is of relevance for ocean-atmosphere exchange of heat, mass and momentum, and has been a major motivation for many studies on mixing of stratified fluids. The vertical fluid exchange and mixing of the surface layer in water basins and lakes is also relevant to the water quality in view of the transport of biological and chemical compounds. The relevance of Langmuir circulation to the mixed surface layer deepening is not fully understood (see Thorpe [11]). In this paper, experimental results on entrainment in turbulent shear flows and Taylor vortices are compared. Details of this study can be found in Flór et al. [2].

Langmuir cells consist of an array of alternating horizontal vortices at the ocean surface that are aligned with the wind direction. These cells establish due to the combined action of wind-induced shear and of Stokes drift (see reviews of Leibovich [6]; Thorpe [11]) and may have a depth between 2 and 300 m with an aspect-ratio close to 1. Typical velocities range from 10 to 20 cm/s for wind speeds of 3–5 m/s or larger (see e.g. Smith [9]; Weller and Price [12]; Thorpe [11]).

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*, Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_41,

In order to estimate the relevance of the mixing process of shear turbulence relative to Langmuir circulation, Li et al. [7] suggested that the mixing stops for a critical Froude number given by

$$Fr = \frac{w_{dn}}{(h\Delta b)^{1/2}} = C \tag{1}$$

where w_{dn} is the maximum downwelling velocity, $\Delta b = g\Delta\rho/\rho$ the buoyancy jump at the base of the mixed layer and h the mixed layer depth. Li et al. [7] suggest for the value C, 0.9 for a two-layer and 0.6 for a linearly stratified fluid. For the shear instability they employed the Price et al. model which predicts static stability of the mixed layer when

$$Ri_b = \frac{\Delta bh}{(\Delta U)^2} \geq 0.65. \tag{2}$$

with mean velocity difference across the density interface at the base of the mixed layer ΔU. Then the transition from Langmuir mixing to shear mixing was predicted to be

$$w_{dn}/C \geq \sqrt{0.65}\Delta U. \tag{3}$$

As far as the measurement resolution allows, in situ measurements confirmed that Langmuir vortices may dominate the dynamics when this criterion is fulfilled.

This criterion is based on laminar vortices, for which the mixing is arrested above the critical Froude number. For turbulent vortices, the mixing may continue on a smaller scale even for large Froude numbers because of the presence of smaller scales. In order to know how relevant this mixing is compared to shear turbulence it is essential to know the entrainment rates of both processes. In this context we consider the entrainment rates measured in different laboratory studies. A sketch of these different laboratory flows is presented in Fig. 1.

2 Experiments on Shear- and Vortex-Induced Mixing

A Taylor-Couette device consists of two concentric cylinders with the inner cylinder rotating and centrifugal instability leads to the formation of so-called Taylor vortices of the size of the gap width (see Fig. 1). Supposing that the side walls act as symmetry planes in the horizontal direction, the Taylor vortices above the interface have essential aspects in common with Langmuir vortices.

In these measurements the entrainment rate was measured from the vertical density flux (see Guyez et al. [3] for details)

$$F(z,t) = \int_z^h \frac{\partial\rho(z,t)}{\partial t} dz.$$

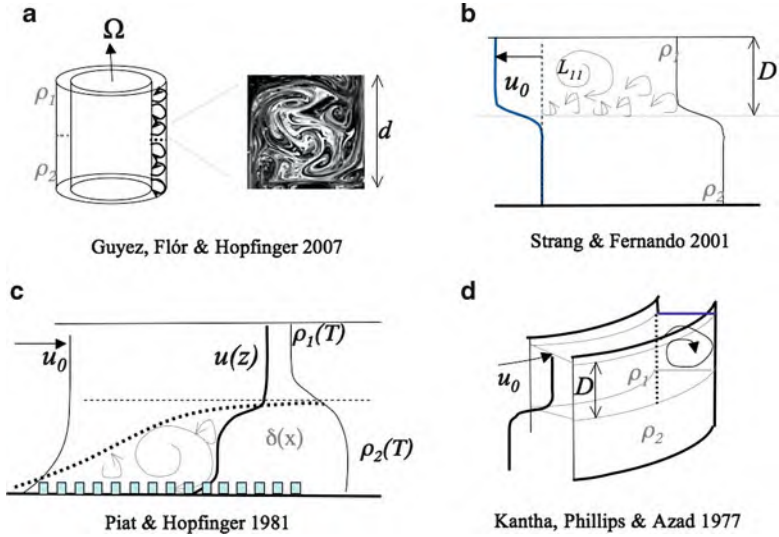


Fig. 1. Schematic diagram of the different experimental flows considered, with (**a**) the Taylor Couette flow and Taylor vortices, (**b**) the shear flow in the Odell-Kovasznay device (see Strang and Fernando [10]) (**c**) a wind induced shear flow and (**d**) the shear flow driven by an annular disk at the fluid surface

At the boundaries $z = 0$ and $z = h$ the flux is zero so that when the interface is thin and is bounded above and below by mixed layers, the flux decreases linearly above and below the interface. The flux can be expressed in terms of an entrainment velocity u_e across the interface in the form $F(z_{int}, t) = \Delta\rho(t)u_e(t)$ from which the entrainment rate u_e/u_m with maximum velocity u_m was calculated. Figure 2 (lower curve) is the entrainment rate as a function of the Richardson number here defined as $Ri_* = \Delta bd/u_m^2$ and d the size of the vortex.

To relate the entrainment rates of Langmuir vortices to that of Taylor vortices, we use the estimations of Li et al. [7] for the surface friction-velocity, u_*, induced by the wind at the surface, U_w, and the maximum down welling velocity w_{dn}. These relations are based on in situ measurements and are $u_* = 1.3 \cdot 10^{-3}U_w$, and $w_{dn} = 8.3 \cdot 10^{-3}U_w$ yielding for the maximum down-welling velocity $w_{dn} = 6.4u^*$. Relating the maximum down-welling velocity to the maximum vertical velocity in the Taylor vortices, $u_m \approx w_{dn} = 6.4u^*$ one obtains the upper curve in Fig. 2, showing the entrainment rate and Richardson number of the Taylor vortices relative to the wind friction velocity u^*.

In the Taylor–Couette flow the interface stays at approximately the same height because of the forcing symmetry below and above the interface. Since the mixing is invariant to forcing at both sides or a single side of the interface (see Turner 1968) we may consider a single vortex above the interface. If the

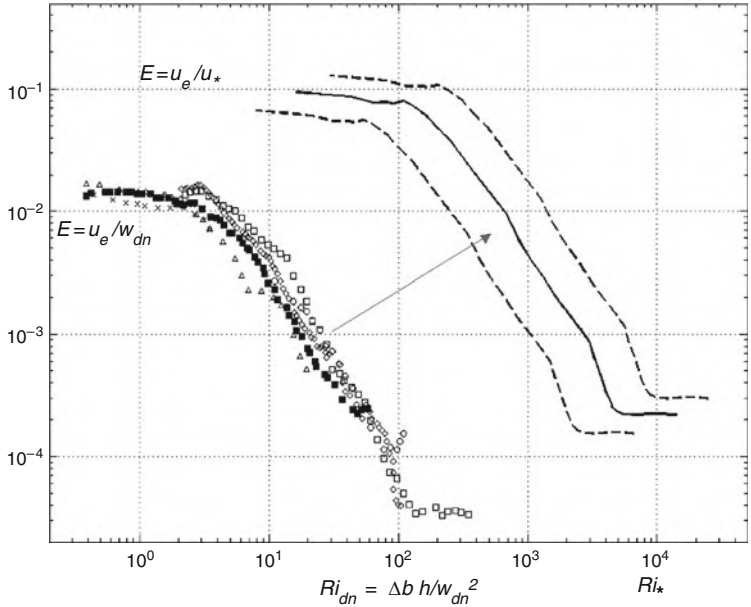


Fig. 2. Entrainment coefficient E versus Richardson number Ri_{dn} and Ri_*. Dashed lines represent the lower and upper limits of the entrainment (see Flór et al. [2])

vortices could be generated only above the interface, the mixed layer would deepen by the turbulent vortex near the interface with a rate $F/\Delta\rho$ over a depth d. This entrainment rate is therefore comparable to the entrainment rate based on the layer deepening as measured in the shear flows discussed below (for a detailed discussion on the different entrainment rates see Hunt et al. [4]).

In shear flows the fluid at the interface mixes due to shear instabilities (Kelvin Helmholtz instability and for larger Richardson numbers Holmboe instability, see Strang and Fernando [10]), and large scale motions continuously homogenize the upper layer. A sketch of a shear flow in which the upper layer is moving over a denser lower layer is shown in Fig. 1b. The entrainment rate is measured from the increase in upper-layer depth and is scaled with the *rms* velocity. This velocity is approximately equal to the surface friction velocity u_*. As typical length scale the integral length scale of the mixing eddies L_{11} is taken (see Strang and Fernando [10]), so that one obtains a Richardson number $Ri_* = \Delta b L_{11}/u_*^2$.

The configuration of a boundary layer topped by a density interface considered by Piat and Hopfinger [8] is analogue to a wind induced shear flow in a mixed surface layer limited by a pycnocline. The experiments were conducted

in a wind tunnel, with the two-layer stratification set by temperature difference. The turbulent boundary layer was generated by roughness elements at the bottom boundary so that with distance from the inlet of the tunnel, the boundary layer increased in thickness until it mixed the interface and thus increased the lower layer thickness. The entrainment velocity varied with distance and was defined as $u_e = (dD(x)/dx)U_0$, with U_0 the background velocity and $D(x)$ the mixed layer depth. The friction velocity here induced by the bottom roughness was taken as typical velocity yielding a Richardson number $Ri_* = \Delta bh/u_*^2$ and entrainment rate $E = u_e/u_*$.

Another type of shear flow is one that is driven by the rotation of an annular disk at the surface of a two-layer salt-stratified fluid (see Fig. 1d) (see Kantha et al. [5]). Because of the secondary circulation generated by the fluid accelerated in the viscous boundary layer of the annular disk, later studies rejected these results since essentially different from pure shear flows. In the present context, however, these results can be considered as intermediate between a pure shear flow, and the pure vortex driven entrainment in Taylor-Couette flow. The entrainment rate is again based on layer deepening dh/dt, whereas the Richardson number is defined as $Ri_* = \Delta bh/u_*^2$ with u_* the surface friction velocity.

3 Comparison of Results and Discussion

Figure 3 shows the different results for the entrainment rate dimensioned with the surface friction velocity as a function of Ri_*. For $Ri_* < 80$ shear induced entrainment dominates over the Langmuir vortex induced entrainment by approximately a factor 2 or less. At the crossover $Ri = 80$ shear instabilities are arrested by the stratification and the entrainment rate drops off to very low values whereas turbulent vortices continue the mixing on smaller scales, thus explaining the higher entrainment rates for Langmuir circulation. The entrainment rates obtained by Kantha et al. [5] follow the shear turbulence for $Ri < 80$ and the Langmuir turbulence for $Ri > 80$ and are coherent with the present findings. For higher Ri_* numbers it drops of slightly faster than the purely vortex driven flow because the secondary circulation in Kantha et al.'s experiments must be relatively weak compared to the rather turbulent Taylor vortices. Deardorff and Yoon [1], who considered the flow in an annular tank driven by an annular disk that did not entirely cover the fluid surface as a mean to reduce the secondary circulation. For $Ri > 80$, their results (not plotted in the figure) fall in between those of Strang and Fernando [10] and Kantha et al. [5], again in coherence with the relatively higher vortex-entrainment rates for large Ri-numbers.

Observations confirm this tendency of an initial layer deepening due to shear instability and subsequent deepening due to Langmuir cells. For a pycnocline with typically $\Delta b = 3 \cdot 10^{-3}\,\mathrm{m\,s^{-2}}$, 1 m depth and $5m/$ wind speed (i.e. $u_* = 0.65\ 10^{-2}\,\mathrm{m/s}$) this implies a Richardson number of $Ri_* \approx 72$ with

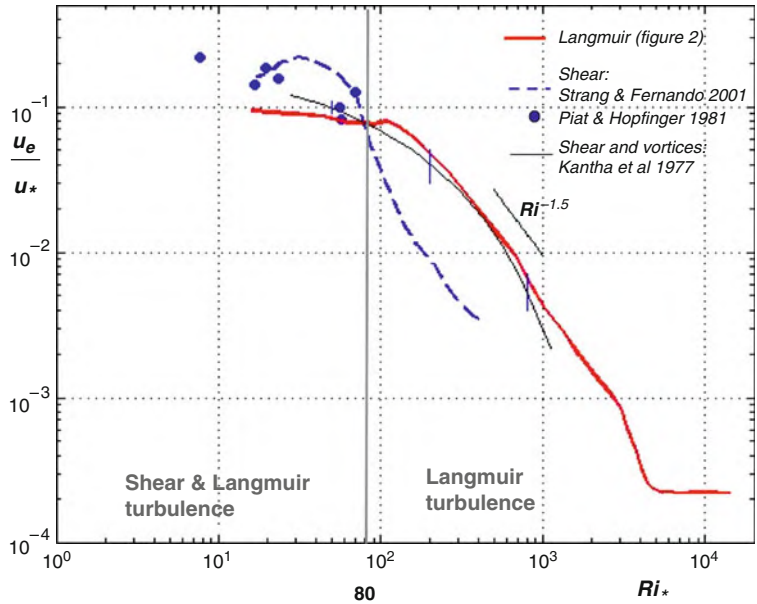


Fig. 3. Entrainment rates dimensioned with the wind-induced surface shear-stress u_* as a function of Ri_*. The error bars in the data are estimated from the data sets

deepening mainly due to Langmuir vortex cells for larger layer depths. Langmuir cells typically obtain a depth of approximately 20 m or larger, suggesting a dominant effect of Langmuir circulation on mixed layer deepening.

References

1. Deardorff, J.W., Yoon, S.-C.: J. Fluid Mech. **142**, 97–120 (1984)
2. Flór, J.B., Hopfinger, E.J., Guyez, E.J.: Deep Sea Res.: submitted (2008)
3. Guyez, E.J., Flór, J.B., Hopfinger, E.J.: J. of Fluid Mech. **46**, 11–21 (2007)
4. Hunt, J.C.R., Rottman, J.W., Britter, R.E.: IUTAM Symposium 1983 Atmospheric Dispersion of Heavy Gases and Small Particles, pp. 361–395. (1984)
5. Kantha, L.H., Phillips, O.M., Azad, R.S.: J. Fluid Mech. **79**, 753–768 (1977)
6. Leibovich, S.: Ann. Rev. Fluid Mech. **15**, 391–427 (1983)
7. Li, M., Zahariev, K., Garrett, C.: Science **270**, 1955–1957 (1995)
8. Piat, J.-F., Hopfinger, E.J.: J. Fluid Mech. **113**, 411–432 (1981)
9. Smith, J.A.: J. Geophys. Res. **103**, 12.649–12.668 (1998)
10. Strang, E.J., Fernando, H.J.S.: J. Fluid Mech. **428**, 349–386 (2001)
11. Thorpe, S.A.: Ann. Rev.Fluid Mech. **36**, 55–79 (2004)
12. Weller, R.A., Price, J.F.: Deep Sea Res. **35**, 711–47 (1988)
13. Turner, J.S., J. Fluid Mech. **33**, 639–656 (1968)

Minisymposium *Inverse Problems and Signal Processing in Industrial Applications*

R. Ramlau[1] and G. Teschke[2,3]

[1] Universität Linz, Linz, Austria, ronny.ramlau@jku.at
[2] University of Applied Sciences Neubrandenburg, Institute for Computational Mathematics in Science and Technology, Brodaer Str. 2, 17033 Neubrandenburg, Germany, teschke@hs-nb.de
[3] Junior Konrad-Zuse-Fellow, Zuse Institute Berlin (ZIB), Takustr. 7, 14195 Berlin, Germany

The overall goal of this minisymposium is to document recent mathematical developments in the field of inverse problems and signal processing that are relevant for various scientific and industrial applications. The particular focus is on scientific and industrial applications in which the desired information is only given by indirect measurements. To this end, one is facedwith two problems: First, one needs to model the connection between the observed data and the searched for information, and secondly the extraction or reconstruction has to be done in a stable way. The main difficulty for this framework is that the extraction process is rather ill-posed, and methods from regularization theory have to be employed in order to control the influence of the data noise in the extraction or reconstruction process.

In five presentations, very different scientific and industrial problems ranging from life sciences, laser optics, rotational dynamics, and the analysis of the ionosphere and atmosphere were discussed. The talk *Sparse deconvolution for peak picking and ion charge estimation in mass spectrometry* presented by T. Alexandrov was concerned with a new procedure for peak detection in mass spectrometry data using sparse deconvolution. The essential ingredient is an ℓ_p sparsity measure that lead to algorithms allowing sparse signal reconstructions. The authors show how this procedure can estimate the ion charges for isotopic patterns of overlapping peaks. The evaluation is performed on the thymosin β_4 16–38 fragment measurements. In the talk *Mathematical Imbalance Determination from Vibrational Measurements and Industrial Applications*, presented by R. Ramlau, the focus was on the detection of imbalances in rotating systems, e.g., aircraft engines, wind turbines or vacuum pumps. For the reconstruction, a model that connects the imbalance distribution to the vibration has to be derived. This can be done by using experimental data or by an FEM discretization of the partial differential equation that describes the vibration. For the inversion process, Tikhonov regularization is used. R. Pike

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*,
Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_42,

presented *A new approach to the analysis of scanning optical imaging systems using singular function expansions*, where he used the singular value decomposition of the integral imaging operator in order to update the widely used theory of optical transfer functions for scanning optical imaging systems. This also leads to the design of optical masks to increase resolution of the imaging system.

In the presentation *The Application of Wavelet Analysis for the Detection of Planetary Wave Type Oscillations in the Ionospheric Total Electron Content*, given by C. Borries, Ionospheric Total Electron Content (TEC) maps are analysed for detecting oscillations with typical periods of planetary waves (PW). The Fourier transform and the continuous wavelet transform are combined for the spectral analyses of the data set. A few ionospheric oscillations found in TEC have typical properties of PW. However, most of the zonal mean TEC variations, which dominate the ionospheric variability, are allocated to the variability of the solar influence. Propagating and standing waves are supposed to occur due to PW. The talk *Statistical Signifcance of Gabor Frames Expansions – Simple Filtering Principles for Radar Wind Profiler Data* presented by G. Teschke has discussed a new signal processing method for the suppression of intermittent clutter echoes in radar wind profilers. The technique presented makes use of discrete Gabor frame expansions in combination with a statistical significance test. The rationale of this algorithm was outlined and an example using data obtained with an operational 482 MHz wind profiler was given.

Summarizing, all presentations have shown relevant and very demanding "real world" problems that are of great importance in the mentioned scientific and industrial areas. It was demonstrated that with very recent developed analysis tools from inverse problems, e.g. such as sparsity measures, and from signal analysis, e.g. such as frame theory or wavelet theory, it is possible to generate procedures to tackle the posed problems.

Sparse Deconvolution for Peak Picking and Ion Charge Estimation in Mass Spectrometry

Kristian Bredies[1], Theodore Alexandrov[1], Jens Decker[2], Dirk A. Lorenz[1], and Herbert Thiele[2]

[1] Center for Industrial Mathematics, University of Bremen, D-28334 Bremen, Germany, {kbredies, theodore, dlorenz}@math.uni-bremen.de
[2] Bruker Daltonik GmbH, D-28359 Bremen, Germany, {jens.decker, herbert.thiele}@bdal.com

Summary. In this paper we propose a new procedure for peak detection in mass-spectrometry data using sparse deconvolution. We apply the procedure for estimation of the ion charges for isotopic patterns of overlapping peaks. The evaluation is performed on the thymosin β_4 16-38 fragment measurements. Moreover, a comparison with the Mexican hat based algorithm of peak picking is provided.

1 Introduction

For mass spectrometry (MS) the detection of m/z (i.e. mass over charge) peaks is a vital step of the data processing pipeline. The purpose of a MS peak picking algorithm is the transformation of a profile spectrum into a list of peaks. For most instruments the profile spectra are obtained by digitalization of a time dependent signal.

The main required properties of a peak detection algorithm are: (1) good m/z precision and accuracy, (2) resolution of overlapping peaks, (3) selective recognition of noisy peaks and (4) performance.

The resolution of overlapping isotopic peaks (for example see Fig. 1) is important to resolve overlapping chemical compounds and to determine the distance between the isotopic peaks of the same molecular ion. The distance between two isotopic peaks in the pattern for the charge z is approximately $1.00235/z$ Th for peptides with deviations in the milli-Thomson range depending on the exact formula. Based on this distance it is possible to determine the charge of an ion which is e.g. important for the real time selection of the most promising ions for fragmentation experiments. This is especially the case for ion trap instruments which have a typical full width half maximum of 0.2–0.5 Th depending on the measurement mode and the type of instrument. Even though the peak resolution is rather limited, these instruments are still of large interest especially due to their large sensitivity and comprehensive fragmentation capabilities.

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*, Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_43,

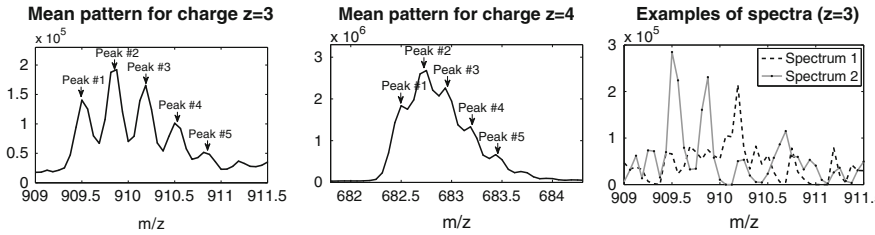


Fig. 1. Mean isotopic patterns of overlapping peaks for ion charges $z = 3$ and $z = 4$ and two examples of spectra for the ion charge $z = 3$

1.1 Mathematical Formulation of the Problem

The problem under study is (1) detection and picking of overlapping isotopic peaks and (2) estimation of the charge of the molecular ion using the distance between the isotopic peaks found. The distance between two neighboring isotope peaks can be assumed to be $1/z$ Th which allows a determination of the charge z provided that the positions of the peaks are known.

For finding the actual positions of the peaks for one isotope pattern, the model assumption is that the measured data f is composed of only few peaks of a known shape G_σ, i.e.

$$f = \sum_{i \in I} u_i G_{\sigma,i} \ .$$

We suppose $G_{\sigma,i}$ to be a Gaussian peak at position i whose area has been normalized to 1: $G_{\sigma,i}(x) = c_\sigma \exp\bigl(-\sigma(x-i)^2\bigr)$. Its width can be tuned by σ and is usually given by the characteristics of the utilized mass spectrometer and resolution. Moreover, u_i represent the corresponding coefficients and I is the (finite) collection of positions we are considering for the peaks.

This model can be interpreted as a result of convolution of several Dirac delta peaks (of heights u_i) with the Gaussian kernel that happened in the process of mass spectrometry measurements.

2 Proposed Method

2.1 Peaks Detection Through Sparse Deconvolution

For isotope patterns, it is important to suppose that the number of actual peaks, i.e. the number of non-zero coefficients u_i, is significantly less than the number of available peaks in I. Such a model assumption can be mathematically implemented by taking so-called "sparsity constraints" into account. Recovering a series of delta peaks from convolved data is known as "sparse

deconvolution" and has been studied recently [3,4]. Moreover, sparsity assumptions can serve as a regularization to reduce the sensitivity with respect to noisy data, see [6] and the references therein. Therefore, for the deconvolution of the isotope patterns, we are following a variational approach which amounts to the solution of the following minimization problem:

$$\min_u \frac{\| \sum_{i \in I} u_i G_{\sigma,i} - f \|^2}{2} + \alpha \sum_{i \in I} |u_i|$$

In the algorithm, I consists of equally sampled points covering the part of the spectrum to be deconvolved. The sampling rate $(\Delta i)^{-1}$ is chosen to be significantly higher than the sampling rate of the spectrum, usually of the factor 4. Additionally, the regularization parameter α is set to be a multiple of the area of the data, i.e. $\alpha = \tau \sum_j |f_j|$ with a $\tau > 0$. The solution of the minimization problem was done by an iterative thresholding algorithm from [2].

2.2 Charge Estimation

This part of the algorithm takes a deconvolved isotope pattern u and tries to extract its charge z by examining the distances between the peak positions. First, it extracts the N most significant peaks, i.e. an ascending sequence of i for which the u_i correspond to the N greatest values of u. In the implemented algorithm, we chose $N = 5$. Each of the positions i_k are corrected by fitting a parabola to the coordinates $(i - \Delta i, i, i + \Delta i)$ and the corresponding values. Then i_k is replaced by the position of the parabola maximum. Subsequently, all differences between the i_k are collected and weighted according to how many peaks are skipped, i.e.

$$D = \left((i_{k+1} - i_k)_k, \tfrac{1}{2}(i_{k+2} - i_k)_k, \tfrac{1}{3}(i_{k+3} - i_k)_k, \ldots, \tfrac{1}{N}(i_N - i_1) \right) .$$

For D, the mean value m as well as the variance V are computed. The charge z is then estimated by the integer closest to $1/m$.

In order to make the charge estimation more robust, the following additional step is performed. Assuming that there is one outlier in the collection of positions $(i_k)_k$, we compute the corresponding charge estimate as well as the variance for the positions where

1. i_k is left out for $k = 1, \ldots, N$,
2. i_k is replaced by $\frac{1}{2}(i_{k+1} + i_{k-1})$ for $k = 2, \ldots, N-1$.

Eventually, the charge which corresponds to the smallest variance for the above test is returned.

Algorithm

1. Given: isotope pattern f
   - Create a set of positions $I = i_{\text{start}} : \Delta i : i_{\text{end}}$
   - Compute $\alpha = \tau \sum_j |f_j|$
2. Solve the minimization problem $\min_u \ \frac{1}{2} \| \sum_{i \in I} u_i G_{\sigma,i} - f \|^2 + \alpha \sum_{i \in I} |u_i|$
3. Extract most significant peaks
   - Find the indices $i_1, \ldots, i_N$ of the N greatest u_i
   - Replace each peak i_k by the maximum of the fitting parabola
4. Create reduced peak lists $(P_p)_p$
   - P_0: original peak list
   - $P_1, \ldots, P_N$: peak list with i_p left out
   - $P_{N+1}, \ldots, P_{2N-2}$: peak list with i_{p-N+1} replaced by $\frac{1}{2}(i_{p-N} + i_{p-N+2})$
5. For each reduced peak list: extract charge z_p
   - Compute all differences $(i_k - i_l)/(k - l)$ for all $k > l$
   - Calculate mean m_p and variance V_p of differences
   - Set $z_p = \text{round}(1/m_p)$
6. Return charge z_p corresponding to the minimal V_p

3 Experiments

3.1 Data Set Description

To get spectra with known m/z values and multiple charge states a direct injection measurement of the thymosin β_4 16-38 fragment (Bachem No. H-2926) was done using 200 fmol/µl in 50% acetonitrile, 0.1% formic acid at 3 µl/min. A HCT Ultra ETD II instrument from Bruker Daltonik was used for these measurements. In total 462 spectra were accumulated in the enhanced standard mode without moving average and prefiltering.

3.2 Peak Picking Using the Mexican Hat Wavelet

For the data set given, we compared our peak picking procedure (Algorithm steps 1–3) with a procedure based on using the Mexican hat (MH) wavelet.

Traditional peak picking algorithms are looking for a zero crossing of the 1st derivative. The detection of peaks which do not give a maximum anymore is not possible in that way. The detection of overlapping peaks also becomes difficult.

One approach to overcome this problem is to use the 3rd derivative instead of the 1st [8] to be able to detect shoulder peaks. Using higher derivatives enhances the influence of noise. For Gaussian peaks and a normal distributed noise assumption it can be shown [1] that smoothing with a Gaussian kernel is the optimal filter. Combining the calculation of the 2nd derivative with a Gaussian smoothing is equivalent to convolving the data with a Mexican hat wavelet which is the 2nd derivative of a Gaussian. This approach was successfully used by [5,7] and exploited within the OpenMS framework.

3.3 Comparison Results

The evaluation of the proposed algorithm (denoted SD hereafter) and the comparison with the Mexican hat wavelet based procedure (MH) was organized as follows. For all the spectra we detected peaks positions in two m/z regions $\mathcal{I}_3 = [909, 911.5]$ Th and $\mathcal{I}_4 = [681.5, 684.5]$ Th containing the isotopic patterns for the charges $z = 3$ and $z = 4$, respectively. For each region the distance between the peaks was calculated using our Algorithm (steps 4–6) and converted to the charge value. Then, for each region ($\mathcal{I}_3$ and $\mathcal{I}_4$) the error rate (E_3 and E_4, respectively) was calculated, i.e. the ratio (in percentage) of the correctly evaluated charges to the number of spectra. The algorithms are rated according to the mean rate $E = (E_3 + E_4)/2$.

Both SD and MH have parameters. SD, besides the number of iterations (10,000 iterations are used), has the parameters σ (manages the supposed width of the peaks) and τ (the regularization parameter). The MH algorithm has the two parameters n and c. The parameter c is defining the width of the Mexican hat function in units of the data point distance and is equal to the standard deviation of the Gaussian function defining the Mexican hat function (but not directly equal to σ in SD); $2n + 1$ is the number of data points for which the Mexican hat function is defined (0 beyond that range).

A grid search has been performed where for each pair of parameters the mean error rate was calculated. The calculated mean rates are presented in

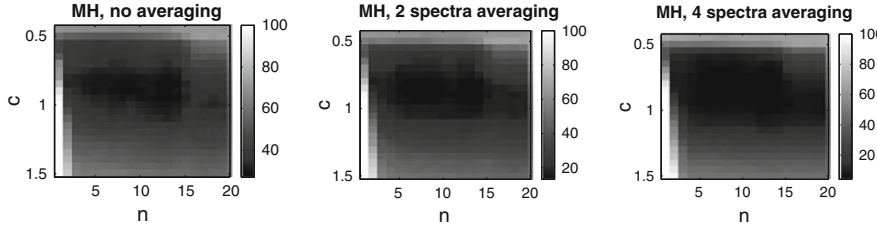


Fig. 2. Grid search results for MH: the mean error rate E (in percentage) for different pairs of parameters n and c

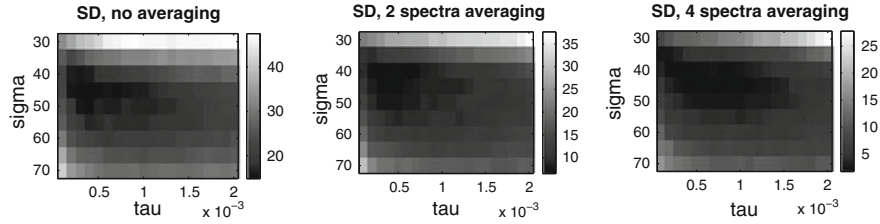


Fig. 3. Grid search results for SD: the mean error rate E (in percentage) for different pairs of parameters σ and τ

Table 1. The minimum mean error rate E (in percentage, over all tested parameters) of our algorithm (SD) and of the Mexican hat based algorithm (MH) calculated for 462 original spectra ("no averaging"), for 461 spectra resulting after averaging of each two neighbor spectra ("2 spectra averaging") and for 459 spectra after averaging of each four neighbor spectra ("4 spectra averaging")

	No averaging	2 Spectra averaging	4 Spectra averaging
SD	14.8%	6.8%	2.5%
MH	27.2%	14.3%	4.8%

Figs. 2 (MH) and 3 (SD). Table 1 (column "no averaging") contains the minimal values of the mean error rates over all tested pairs of parameters. The SD algorithm significantly outperforms the MH peak picking algorithm.

The results can be enhanced by averaging several spectra previous to the peak picking as the data is very noisy (see Fig. 1 for examples of spectra). Though the averaging requires additional measurements (technical replicates), this operation is often used in MS. We simulated the averaging by taking means of two and four neighbor spectra that reduces the data set size to 461 and 459 spectra, respectively. Table 1 contains the mean error rates computed after averaging. The averaging of four spectra for example improves the error rates by the factor of 7. Both procedures (MH and SD) provide low error rates but SD is significantly better than MH for all types of averaging used.

The only disadvantage of SD is its runtime which is 17 min versus 7 s for MH (on an Intel 2.66 GHz PC, for fixed parameters).

Acknowledgements

Thanks to Markus Lubeck, Bruker Daltonik, for doing the MS measurements.

References

1. Andreev, V., Rejtar, T., et al.: Anal. Chem. **75**, 6314–6326 (2003)
2. Bredies, K., Lorenz, D.: SIAM J. Sci. Comput. **30**, 657–683 (2008)
3. Dahlke, S., Maass, P., et al.: Mathematical Methods in Time Series Analysis and Digital Image Processing, 75–109 (2007)
4. Klann, E., Kuhn, M. et al.: Inverse Probl. **23**, 2231–2248 (2007)
5. Lange, E., Gröpl, C., et al.: Proc. Pacific Symp. on Biocomp. 243–245 (2006)
6. Lorenz, D.: J. Inverse Ill-Posed Probl. **16**, 463–478 (2008)
7. Sturm, M., Bertsch, A., et al.: BMC Bioinformatics **9**, 163 (2008)
8. Vivó-Truyols, G., Torres-Lapasió, J., et al.: J. Chromatrogr. A **1096**, 146–155 (2005)

Mathematical Imbalance Determination from Vibrational Measurements and Industrial Applications

Jenny Niebsch[1] and Ronny Ramlau[1,2]

[1] Radon Institute for Computational and Applied Mathematics, Austrian Academy of Science, A-4040 Linz, Austria, jenny.niebsch@oeaw.ac.at
[2] Industrial Mathematics Institute, Johannes Kepler University of Linz, Altenbergerstrasse 69, A-4040 Linz, Austria, ronny.ramlau@jku.at

Summary. The paper focuses on the identification of imbalances from vibrational measurements in rotating machinery and its application to industrial problems. Since it is an ill-posed inverse problem the reconstruction is based on regularization techniques. To handle the direct problem, a model of the rotor under consideration has to be provided. We have employed the imbalance reconstruction principle to several industrial applications of linear and nonlinear nature.

1 Introduction

Imbalances in rotating machinery are a major problem since they can lead to an insecure operation and an early abrasion or even to the destruction of the engine. A direct impact of imbalances are vibrations of the engine which also are the only measurable information on imbalances. In most cases they are only available at some positions at the bearing of the engine where sensors can be mounted. In practice, the usual technique for balancing an engine consists in several vibration measurements: one for the original run with the existing unknown imbalance, and one or several runs with test weights which have to be placed on defined balancing positions. Since the engine has to be at least partly demounted for placing the test weights, this is an expensive and time consuming issue.

Mathematically, imbalances and the resulting vibrations are connected via an operator A acting between Hilbert spaces X and Y that maps an imbalance $f \in X$ to a vibrational signal $g \in Y$:

$$Af = g. \tag{1}$$

The computation of an unknown imbalance distribution for given vibrational data is called the Inverse problem. It is ill-posed since the solution f does not

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*, Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_44,

depend continuously on the data g. In fact, if we only have noisy data with noise level δ, i.e., $\|g^\delta - g\| \leq \delta$, then the measured data g^δ might not even belong to the range of A, and standard algorithms for the computation of a solution of (1) from g^δ might produce an arbitrarily bad approximation to the solution. To obtain a stable solution, one has to use regularization methods, see Sect. 3. Furthermore, we have to deal with incomplete data. That is we can not obtain vibrational data for a frequency range that covers all eigenfrequencies of the engine and additionally not for every point (or model node) of the engine but only for a few positions where vibrational sensors can be mounted. In practice, the number should be minimized due to the cost of sensors and the data processing. Hence the solution of the problem might not be unique.

The mathematical solution of this Inverse Problem enables us to reconstruct the imbalances from vibration measurements without additional test runs. In this way, a lot of money and time can be saved. We have applied the method to several different engines during cooperation projects with industrial companies, e.g. large generators of the Siemens AG Berlin, Germany, Wind power plants (cooperation with Fielax GmbH Bremerhaven, Germany), Vacuum pumps of the Oerlikon Leybold Vacuum GmbH Köln, Germany, and Ultra precision machine tools (Research project with the University of Bremen.

2 The Solution of the Direct Problem

So far we have used two ways of determining A in our applications:

2.1 Experimental Method

A is determined as the influence coefficient matrix. This method requires an approximate linear behavior of the engine. In a first step the vibrational data $\mathbf{u}_{pr}$ for the primary imbalance state $\mathbf{p}_{pr}$ are measured. Afterwards a test mass (imbalance) is attached at the first (balancing) plane. The resulting imbalance is denoted by $\mathbf{p}_1 + \mathbf{p}_{pr}$. After the associated vibrations $\mathbf{u}_1$ are measured, the weight will be placed under a different angle, and the measurement will be repeated. Then the unit mass will be removed and attached in the same way to the next balancing plane and so on. As an example, we have here four imbalance states $\mathbf{p}_{1+\mathbf{p}_{pr}}, \cdots, \mathbf{p}_4 + \mathbf{p}_{pr}$ and the associated vibrations $\mathbf{u}_1, \cdots, \mathbf{u}_4$. Since $\mathbf{u}_{pr}(\omega) = \mathbf{A}(\omega)\mathbf{p}_{pr}$ we have $\mathbf{u}_j(\omega) - \mathbf{u}_{pr}(\omega) = \mathbf{A}(\omega)\mathbf{p}_j$, i.e. the vibrational data for the primary imbalance state have to be subtracted from all the other measurement vectors. The resulting data vectors are collected in a matrix. If we have a sample of K frequencies the influence coefficient matrix is computed as

$$\mathbf{A} = \begin{bmatrix} (\mathbf{u}_1 - \mathbf{u}_{ur})(\omega_1) & \cdots & (\mathbf{u}_4 - \mathbf{u}_{ur})(\omega_1) \\ \vdots & \ddots & \vdots \\ (\mathbf{u}_1 - \mathbf{u}_{ur})(\omega_K) & \cdots & (\mathbf{u}_4 - \mathbf{u}_{ur})(\omega_K) \end{bmatrix} [\mathbf{p}_1, \cdots, \mathbf{p}_4]^{-1}.$$

The experimental method is easier to apply but due to the noisiness of the measurement process the matrix A is already subjected to errors. Additionally, the measuring effort is very large.

2.2 Mathematical Method

The starting point for a mathematical construction of the system matrix is to idealize a rotating system as a flexible shaft which is divided in sections. The motion of each section can be described by a partial differential equation.We transform this PDE in an ordinary differential equation (ODE) via the Finite Element Method and arrive at

$$M\ddot{\mathbf{u}}(t) + D\dot{\mathbf{u}}(t) + S\mathbf{u}(t) = \mathbf{p}(t). \tag{2}$$

Here $\mathbf{u}$ is the system displacement vector where the degrees of freedom of each boundary points of the elements are collected, S is the system stiffness matrix, D the damping matrix, M the system mass matrix, and $\mathbf{p}$ the system load vector.

Now we have to describe how the imbalance is related to the load vector $\mathbf{p}$, and how we derive A from (2). A rotor imbalance originates from or can be described as an inhomogeneous mass distribution, i.e. a mass Δm which is eccentric with radius vector $\mathbf{r}$ from the barycenter and an angle φ from a zero angle mark. Hence an imbalance f_0 is described by $f_0 = \Delta m\ r\ \exp(i\varphi)$.

An imbalance rotating with the frequency ω causes a harmonic load

$$p = \omega^2\ f_0 \exp(i\omega t) = \omega^2\ \Delta m\ r\ \exp(i(\omega \cdot t + \varphi)). \tag{3}$$

If there is more than one possible imbalance position, the load p becomes a vector $\mathbf{p}$. From physical reasons we assume a harmonic vibration with the same frequency $\mathbf{u}(t) = \mathbf{u_0} \exp(i\omega t)$. Inserting this in (2) we arrive at

$$\mathbf{u_0} = (-M + i\omega^{-1}D + \omega^{-2}S)^{-1}\ \mathbf{f_0}. \tag{4}$$

Let Q be a matrix that extracts $\mathbf{u_0}$ at the sensor positions where vibration can be measured and let the measured data be denoted by $\mathbf{g} = Q\mathbf{u_0}$. Now the operator A is given by

$$\begin{aligned} \mathbf{g} &= A\ \mathbf{f_0}, \\ A &= Q\ (-M + i\omega^{-1}D + \omega^{-2}S)^{-1}. \end{aligned} \tag{5}$$

3 Inverse Problem Solution

The Inverse Problem was solved by using the well-known Tikhonov regularization. The regularizer is defined as minimizing element of the Tikhonov functional

$$J_\alpha(f) = \|g^\delta - Af\|^2 + \alpha\|f - \bar{f}\|^2. \tag{6}$$

For the determination of the regularization parameter α, we can use the so called Morozov's discrepancy principle where α is chosen s.t.

$$\delta \leq \|g^\delta - Af_\alpha^\delta\|^2 \leq c\delta \tag{7}$$

holds. Details and convergence results for this method are given in [4,7]. For linear operators, the minimizer of the Tikhonov functional can be computed by solving a linear system.

3.1 Imbalance Reconstruction Algorithm

For the numerical realization we have used the Tikhonov regularization in combination with Morozov's discrepancy principle. The minimizer f_α^δ for a linear operator is simply computed by solving the linear system

$$(A^*A + \alpha I)f = A^*g^\delta + \alpha\bar{f}\ . \tag{8}$$

This applies to all experimentally derived models (generators, vacuum pumps). Here we assume linear behavior at least approximately. The wind turbine problem is linear, too. In the case of a high precision cutting machine we have to consider the damping behavior of an air bearing between shaft and housing. This might be nonlinear.

We want to remark that due to the L_2-Norm in the regularization term of the Tikhonov functional (6), $\|f-\bar{f}\|^2$, the solution with minimal L_2-norm was determined. It did not take into account the sparse character of an imbalance distribution, like point imbalances at certain positions. We have solved this problem with a multiple step algorithm. It is described in [1]. There is still another possibility to avoid this problem: As the number of imbalance positions is finite, the L_2 norm of the imbalance distribution is equivalent to the l_2 norm of the associated imbalance vector. Now, in order to obtain a sparse imbalance distribution directly, we propose to replace the l_2-penalty by an l_p-penalty with $1 \leq p < 2$ and consider the functional

$$J_\alpha(f) = \|g^\delta - Af\|_{L_2}^2 + \alpha\|f - \bar{f}\|_{l_p}^p \tag{9}$$

instead. In particular, if $p = 1$ is chosen, then we can expect a sparse reconstruction if the underlying solution is sparse, see e.g. [8]. The minimization of the functional with $p < 2$ is more challenging than for the case $p = 2$, as the penalty might not be differentiable. However, e.g. for $p = 1$, minimization algorithms have been designed in [8] for linear operators, and for nonlinear operators in [9,10], where also regularization results have been presented. The application of these algorithms is currently under investigation.

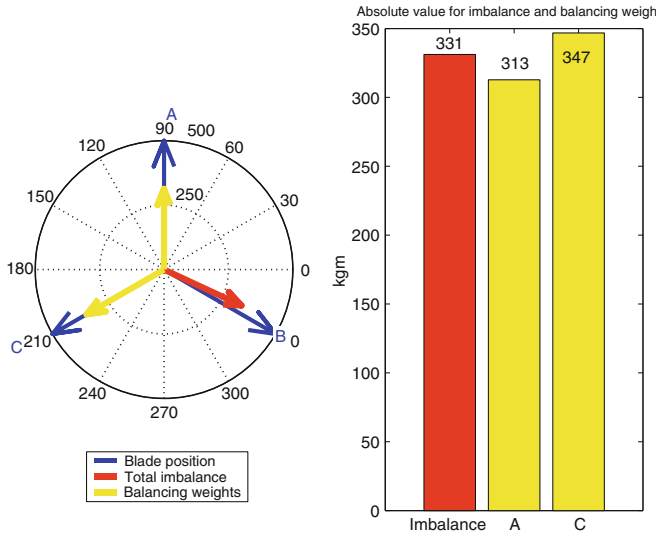


Fig. 1. Reconstruction of a 350 kgm imbalance at blade B with 10% data error (*dark*) and computation of the balancing weights (*light*)

4 Application Examples

As one application we present an imbalance reconstruction for a wind energy plant. A mathematical model was derived for a plant of the type Vestas V80-2MW. We assumed an imbalance of 350 kgm at the blade B, which is a realistic value. Most companies consider this amount as threshold for imbalances. The vibration data for such an imbalance were produced by forward computation and disturbed with a data error of 10%. The reconstruction and the related balancing weights are shown in Fig. 1.

For the high precision cutting machinery we reconstructed given imbalances setting at the two balancing rings of the machine using the experimental method for the solution of the forward problem. One result is shown in Fig. 2.

5 Conclusion

The safely and economic operation of a rotating machinery requires a well balanced system. Therefore the detection of imbalances and their removal is an important point in the machine diagnosis. If we have to rely on vibrational measurements at the casing of the engine, the problem is ill-posed and can not be solved with common techniques. Presently, the balancing process requires extensive and time consuming measurement procedures.

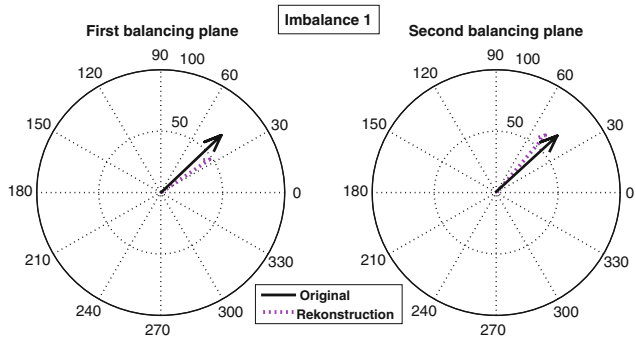


Fig. 2. Reconstruction (*solid line*) of imbalance settings (*dashed*) at the balancing planes of a High Precision Cutting Machine

Our new imbalance reconstruction method reduces the effort for balancing significantly. It uses recently developed techniques for solving inverse ill-posed problems in combination with a model of the rotating system which either has to be developed or is provided by the client. The method was developed on the basis of several practical examples and was successfully tested with artificial and real data.

References

1. Dicken, V., Maass, P., Menz, I., Niebsch, J. and Ramlau, R.: Nonlinear inverse unbalance reconstruction in rotor dynamics. Inverse Probl. Sci. Eng. **13**(5), 507–543 (2005)
2. Dicken, V., Maass, P., Ramlau, R., Rienäcker, A., Streller, C.: Inverse imbalance reconstruction in rotordynamics. ZAMM**86**(5), 385–399 (2006)
3. Gasch, R., Knothe, K.: Strukturdynamik 2. Springer, Berlin (1989)
4. Ramlau, R.: Morozov's discrepancy principle for Tikhonov regularization of nonlinear operators. Numer. Funct. Anal. Optim. **23**(1&2), 147–172 (2002)
5. Ramlau, R.: TIGRA–an iterative algorithm for regularizing nonlinear ill–posed problem. Inverse Probl. **19**(2), 433–467 (2003)
6. Ramlau, R., Niebsch, J.: Verbesserte Auswuchtung von Rotoren. *Industrial report.* (2004)
7. Scherzer, O.: The use of Morozov's discrepancy principle for Tikhonov regularization for solving nonlinear ill–posed problems. Computing **51**, 45–60 (1993)
8. Daubechies, I., Defrise, M., DeMol, C.: An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. Commun. Pure Appl. Math. **51**, 1413–1541 (2004)
9. Ramlau, R., Teschke, G.: A Tikhonov-based projection iteration for nonlinear Ill-posed problems with sparsity constraints. Numer. Math. **104**, 177–203 (2006)
10. Ramlau, R. Regularization properties of Tikhonov regularization with sparsity constraints. ETNA vol. 30, 54–78 (2008)

An Update of Hopkins' Analysis of the Optical Disc Player Using Singular-System Theory

Roy Pike

Clerk Maxwell Professor of Theoretical Physics, King's College, London, UK
roy.pike@kcl.ac.uk

Summary. In this paper we describe a new approach to the analysis of scanning optical imaging systems which uses singular function expansions, rather than Fourier optics, to update the well-known low-aperture treatment of Hopkins of 1979 (J. Opt. Soc. Am. 69:4–24). This new approach can also be used to update the widely used theory of optical transfer functions for general imaging systems at arbitrary numerical apertures.

1 Introduction

We consider an optical system with an illumination lens and an objective lens combination as depicted schematically in Fig. 1.

The same lens serves for both illumination and imaging functions in a reflective system. This is a special case of the general partially coherent optics described, for example, in early work of Hopkins [3]. We will use $\mathbf{x}$, $\mathbf{k}$ and $\mathbf{y}$ as the two-dimensional disc-plane, pupil-plane and image-plane coordinates, respectively. The action of each lens is described by a linear integral equation, relating object $f(\mathbf{x})$ to image $g(\mathbf{y})$,

$$g(\mathbf{y}) = \int d\mathbf{x} W(\mathbf{y} - \mathbf{x}) f(\mathbf{x}), \tag{1}$$

where W is its point-spread function (PSF). Following Hopkins, even in the reflective case, we allow the incoming and outgoing illumination to be described using different PSFs, W_{in}, and W_{out}, respectively, which can include a non-uniform beam profile and aberration corrections, particularly on the high-aperture side in an optical-disc system. It is normally adequate to use a paraxial approximation on the detector side for W_{out}. The supports of both object and image in $\mathbf{x}$ and $\mathbf{y}$ respectively are, in theory, infinite but in practice will be defined by the rapid fall-off of the illumination or by a finite detector aperture.

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*,
Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_45,

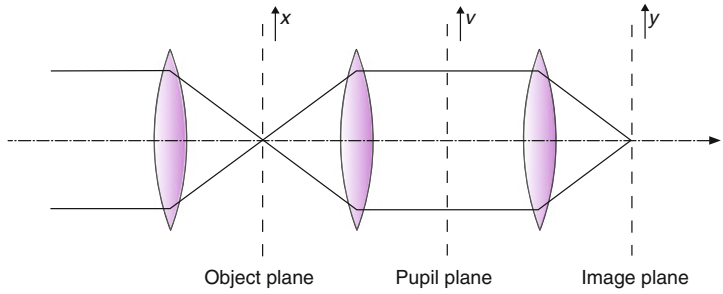


Fig. 1. Object, pupil and image planes

In contrast to the theory of Hopkins, which only applies to low-aperture (paraxial) scalar theory, our theoretical treatment will not depend formally on the numerical apertures, which will occur simply as numerical parameters in the calculation. Nevertheless, the cylindrical symmetry of paraxial systems can be used to improve the numerical efficiency of our calculations when this approximation may be made. The optics of scanning systems of high numerical aperture are discussed in terms of singular function expansions in [4].

The description of the imaging process uses (1) twice, first with the illumination of the disc surface as the image plane of W_{in}, and then using the reflected light from the disc using W_{out} to form the image seen by the detector. Thus, using $\mathbf{s}$ for the scanning variable, the objective lens sees a (complex) field, $f(\mathbf{x})$, as its object, equal to $W_{in}(\mathbf{x})R(\mathbf{x}-\mathbf{s})$, where $R(\mathbf{x})$ is the disc reflectance (or transmittance). The field in the image plane of the objective (ignoring magnification) is thus defined by the linear integral operator, A, where

$$g(\mathbf{y},\mathbf{s}) = (AR)(\mathbf{y},\mathbf{s}) = \int d\mathbf{x} W_{out}(\mathbf{y}-\mathbf{x}) W_{in}(\mathbf{x}) R(\mathbf{x}-\mathbf{s})). \tag{2}$$

2 Hopkins' Analysis

Using Fourier optics, which is applicable in the paraxial approximation, Hopkins writes the pupil-plane field amplitude as

$$\begin{aligned} E(\mathbf{k},\mathbf{s}) &= \mathcal{F}[R(\mathbf{x}-\mathbf{s}).W_{in}(\mathbf{x})] = \hat{R}(\mathbf{k},\mathbf{s}) \otimes \hat{W}_{in}(\mathbf{k}) \\ &= \int dk' \hat{R}(\mathbf{k}') \hat{W}_{in}(\mathbf{k}-\mathbf{k}') e^{i\mathbf{k}'.\mathbf{s}}, \end{aligned} \tag{3}$$

where $\mathcal{F}$ denotes the Fourier transform, $\otimes$ denotes convolution, the overhat denotes Fourier-transformed functions and we have used the Fourier shift theorem. The pupil-plane intensity is given by

$$\begin{aligned} I(\mathbf{k},\mathbf{s}) &= |E(\mathbf{k},\mathbf{s})|^2 \\ &= \int\int \mathbf{dk}'\mathbf{dk}'' \hat{R}(\mathbf{k}') \hat{R}(\mathbf{k}'') \hat{W}_{in}(\mathbf{k}-\mathbf{k}') \hat{W}_{in}(\mathbf{k}-\mathbf{k}'') e^{i(\mathbf{k}'-\mathbf{k}'').\mathbf{s}} \end{aligned} \tag{4}$$

and the integrated intensity over the pupil-plane is

$$\begin{aligned} I(\mathbf{s}) &= \int_{pupil} \mathbf{dk} I(\mathbf{k}, \mathbf{s}) \\ &= \int\int \mathbf{dk}' \mathbf{dk}'' \hat{R}(\mathbf{k}') \hat{R}(\mathbf{k}'') e^{i(\mathbf{k}'-\mathbf{k}'').\mathbf{s}} D(\mathbf{k}', \mathbf{k}''), \end{aligned} \tag{5}$$

where

$$D(\mathbf{k}', \mathbf{k}'') = \int_{pupil} \mathbf{dk} \hat{W}_{in}(\mathbf{k} - \mathbf{k}') \hat{W}_{in}(\mathbf{k} - \mathbf{k}''). \tag{6}$$

We put $\mathbf{k}' - \mathbf{k}'' = \mu$ so that

$$\begin{aligned} I(\mathbf{s}) &= \int d\mu e^{i\mu\mathbf{s}} \int_{pupil} \mathbf{dk}' \hat{R}(\mathbf{k}') \hat{R}(\mathbf{k}' + \mu) D(\mathbf{k}', \mathbf{k}' + \mu) \\ &= \int d\mu I(\mu) e^{i\mu\mathbf{s}}, \end{aligned} \tag{7}$$

where

$$I(\mu) = \int_{pupil} \mathbf{dk} \hat{R}(\mathbf{k}) \hat{R}(\mathbf{k} + \mu) D(\mathbf{k}, \mathbf{k} + \mu). \tag{8}$$

We normalise by

$$\int_{pupil} \mathbf{dk} D(0, 0) = \int_{pupil} \mathbf{dk} |\hat{W}_{in}(\mathbf{k})|^2. \tag{9}$$

To calculate the output signal from an extended image-plane square-law detector we use Parseval's theorem

$$\int_{image} d\mathbf{y} I(\mathbf{y}) = \int_{pupil} d\mathbf{k} I(\mathbf{k}), \tag{10}$$

so that it is given directly by (7). Calculations are performed by constructing reflection functions for various periodic arrangements in two euclidean local dimensions (along and across track) of specified pits on the disc surface, with a sufficiently fine discretisation for numerical integration in those two dimensions in the $\mathbf{x}$ and $\mathbf{k}$ planes.

3 Singular Function Analysis

To economise on notation from here on we will use the Dirac bra-ket notation for wave-amplitude functions in the $\mathbf{x}$, $\mathbf{k}$ and $\mathbf{y}$ planes and automatic summation on repeated indices. The image-plane amplitude is given by the operator form of the imaging equation (2)

$$|g(\mathbf{s})> = A|R(\mathbf{s})>, \tag{11}$$

where the operator A maps the **x** plane into the **y** plane (considered as L^2 function spaces). The singular value decomposition of A is

$$A = \alpha_i |v_i >< u_i|, \tag{12}$$

where the singular values α_i are real and the singular functions $|u_i >$ and $|v_i >$ are orthonormal basis functions in the **x**- and **y**-planes, respectively. Using this decomposition and the following singular function expansions of R and g:

$$\begin{aligned} |R(\mathbf{s}) > &= < u_i|R(\mathbf{s}) > |u_i >= R_i(\mathbf{s})|u_i > \\ |g(\mathbf{s}) > &= < v_i|g(\mathbf{s}) > |v_i >= g_i(\mathbf{s})|v_i >, \end{aligned} \tag{13}$$

we find that

$$|g(\mathbf{s}) >= \alpha_i R_i(\mathbf{s})|v_i >, \tag{14}$$

The integrated intensity over the image plane is then

$$\begin{aligned} I(\mathbf{s}) &= < g|g >= |\alpha_i R_i(\mathbf{s})|^2 \\ &= \alpha_i^2 R_i(\mathbf{s}) R_i^*(\mathbf{s}), \end{aligned} \tag{15}$$

When scanning we need to recover only the axial values of $R(\mathbf{s})$.

In the paraxial approximation the calculation of the two-dimensional singular system may be performed in one dimension by using the axial symmetry of the optical system [1]; otherwise a full two-dimensional calculation is needed. The singular functions and singular values are precomputed and the calculation then simply needs the scalar product of $R(s)$ with the desired very small number of $u_i s$ over the disc plane. Column 1 of Table 1 of Bertero et al. [1] shows that in the paraxial approximation the values of α_i^2 fall off as shown in Fig. 2; it can be seen that the contribution of the second function in the expansion is less than 4% and the third less that 1% of that of the first.

In contrast to the Hopkins method, in which the R are all coupled together by the functions D of (6), the object components are completely decoupled from each other in our calculation due to the orthonormality of the singular functions. This allows the simplification which may be seen between (15) and (7). Of course, the small number of terms required in the new method is also helpful.

Acknowledgements

I would like to acknowledge gratefully collaboration over a number of years with Drs. Jan Grochmalicki and Deeph Chana at King's College, London and with the members of the optical group at Philips Laboratories, Eindhoven, particularly Drs. Sjoerd Stallinga and Paul Urbach on theoretical questions. This work has been supported by EPSRC Grant No GR/L13964 and by the European Union FET programme 2000-26479 "Super Laser Array Memory".

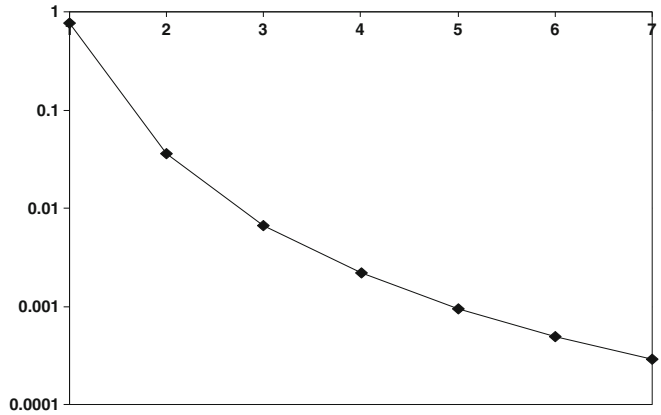


Fig. 2. The squared spectrum of singular values in the paraxial approximation

References

1. Bertero, M., Boccacci, P., Davies, R.E., Pike, E.R.: Inverse Probl. **7**, 655–74 (1991)
2. Hopkins, H.: J. Opt. Soc. Am. **69**, 4–24 (1979)
3. Hopkins, H.: Proc. R. Soc. Lond. **A208**, 263–277 (1951)
4. Grochmalicki, J., Pike, E.R.: App. Opt. **39**, 6341–6349 (2000)

The Application of Wavelet Analysis for the Detection of Planetary Wave Type Oscillations in the Ionospheric Total Electron Content

C. Borries

University of Applied Sciences Neubrandenburg, Brodaer Str. 2, Neubrandenburg, Germany, and
DLR, Institute of Communications and Navigation, Kalkhorstweg 53, Neustrelitz, Germany, claudia.borries@dlr.de

Summary. Ionospheric Total Electron Content (TEC) maps are analysed for detecting oscillations with typical periods of planetary waves (PW). The Fourier transform and the continuous wavelet transform are combined for the spectral analyses of the data set. A few ionospheric oscillations found in TEC have typical properties of PW. However, most of the zonal mean TEC variations, which dominate the ionospheric variability, are allocated to the variability of the solar influence. Propagating and standing waves are supposed to occur due to PW.

1 Introduction

Planetary waves (PW) are large scale waves, which emerge in the lower and middle atmosphere. They contribute essentially to the atmospheric dynamics, because they transport energy and momentum. In winter they dominate the dynamics of the middle atmosphere. PW are able to penetrate upwards, but due to strong changes in temperature and winds in the turbopause region at about 110 km altitude, most of these waves break or dissipate in this region. This is approved by numerical modelling [8, e.g.].

Nevertheless, PW type oscillations (PWTO) can be found in the ionosphere [7, e.g.]. Their contribution to the ionospheric variability was estimated in [2] with 15–20%. The PWTO might be an indicator for a vertical coupling between the middle atmosphere and ionosphere. In this case indirect processes like the modulation of upward propagating tides or atmospheric gravity waves through PW could be responsible for the vertical transport of the PW energy [7, e.g.].

Regional hemispheric Total Electron Content (TEC) maps are a relatively new data base characterizing the variability of the ionosphere. In this work they are used to analyse oscillations in TEC with scales of PW. The Fourier transform and the continuous wavelet transform are applied complementary

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*,
Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_46,

for the spectral analyses. A typical occurrence of the PWTO in the ionosphere is estimated by characterizing the waves found in the TEC maps referring to their zonal wavelength, period and zonal propagation direction. Comparisons to solar wind measurements and stratospheric analyses are used to investigate the probable origin of the PWTO.

2 Data Base

Maps of the vertical TEC, which is the vertically integrated electron content (estimated in electrons/m^2), are used to investigate periodic variations in the ionosphere. Regional TEC-maps are regularly produced by the DLR Neustrelitz [5] using ground based GNSS measurements for the estimation of TEC. After determining the slant TEC along a number of ray paths by using a special calibration technique for the ionospheric delay on GPS signals [5], the slant TEC is mapped to the vertical by using a single layer approximation of the ionosphere at 400 km height. To ensure a high reliability of the TEC maps, the measured data are combined with the empirical TEC model NTCM2. For each grid point value a weighting process between nearest measured and model values is carried out. The absolute accuracy of the so-generated TEC maps has been estimated to lie in the order of a few 10^{16} electrons/m^2 [5]. This accuracy is high enough to monitor large scale perturbation processes.

The North Pole TEC-Maps covering the northern hemisphere from the polar cap down to 50°N with a regular grid (spacing is 2.5°/7.5° in latitude/longitude) and a time resolution of 1 h are available since 2002. These maps are suitable for the analyses of large scale wave phenomena because of their hemispheric coverage. It is useful to calculate relative differences ($\Delta TEC_{rel} = (TEC - TEC_{med})/(TEC_{med})$) to monthly median values (TEC_{med}) in order to reduce the major influence of the sun on the analysis results [4].

3 Methods of Analyses

Spectral analyses will be applied on the $\Delta\mathrm{TEC}_{rel}$-maps in order to get information about the horizontal scale and propagation of the waves. Because of the relatively little meridional extent of the maps, a possible meridional wave propagation will be neglected. This is suitable for the PW analyses, because PW mainly propagate zonally in the middle and lower atmosphere.

Hence, a two dimensional spectral analysis is necessary. The frequency-wavenumber-analyses, a well-known procedure for the space-time spectral analyses, described in [3], is applied. The waves are assumed to be harmonic plain waves with a pure zonal propagation $f(x,t) = \int_{-\infty}^{\infty} c(k,\omega) e^{i(kx-\omega t)} d\omega dk$.

At first, the Fourier transformation is applied in the space dimension on the signal $f(x,t)$

$$F(k,t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(x,t) e^{-ikx} dx = C(k,x) + iS(k,x) \tag{1}$$

The Fourier coefficients are decomposed into their real ($C(k,x)$) and imaginary ($S(k,x)$) part, in order to keep the phase information. The Fourier analyses in time dimension is applied on C and S.

$$P_c(k,\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} C(k,t) e^{-i\omega t} dt \tag{2}$$

$$P_s(k,\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} S(k,t) e^{-i\omega t} dt \tag{3}$$

The two Fourier spectra P_c and P_s contain the information about the present wavenumbers and frequencies. The power spectrum dependent on the direction of the wave propagation can be calculated with

$$4P(k,\pm\omega) = P_c^* P_c + P_s^* P_s \pm 2Q_{cs} \tag{4}$$

(see [3]) where Q_{cs} represents the quadrature spectrum ($P_c^* P_s = K_{cs} + iQ_{cs}$) and the asterisk indicates the complex conjugate. The power spectrum of the eastward propagating waves is calculated with the positive sign and the westward waves respectively with the negative sign. Two waves with the same wavenumber and frequency propagating in opposite directions describe a standing wave.

In order to get a better localization of the signal in time, the Fourier analysis in time dimension can be replaced e.g. by a short time Fourier or a wavelet analysis. Because the wavelet analysis has a better resolution concerning the Heisenberg uncertainty principle, the continuous wavelet transform(CWT) will be applied in this paper.

$$W_\psi f(s,\tau) = \frac{1}{\sqrt{c_\psi}} \int_{-\infty}^{\infty} f(t) \psi_{s,\tau}^*(t) dt \tag{5}$$

The CWT is basically a convolution of the signal $f(t)$ and the complex conjugate of a scaled and translated version of a mother wavelet ($\psi_{s,\tau}(t) = |s^{-0.5}| \psi_0 \left(\frac{\tau - t}{s}\right)$). The morlet wavelet ($\psi_0(\eta) = \pi^{-0.25} e^{i\omega_0 \eta} e^{-\eta^2/2}$) with a center frequency $\omega_0 = 6$ is used as mother wavelet because of its good correspondence to a cosine oscillation. Equation 5 has to be modified to get the true amplitudes $\tilde{W}_\psi f(s,\tau) = \sqrt{c_\psi} (2\pi s)^{-0.5} W_\psi f(s,\tau)$.

4 Results

To demonstrate the estimation of the ionospheric variability derived from the North Pole TEC-maps, $\Delta\mathrm{TEC}_{rel}$ is analysed along the 55°N latitude for the zonal mean variations and waves with wavenumber 1. The results of the CWT

are not discussed for single wave events. Instead, a global wavelet spectrum is calculated, which is the root of the integrated squared amplitudes (RMS).

$$RMS(s) = \sqrt{\int \tilde{W}_{\psi}^{(95)} f(s,\tau)^2 d\tau}$$

Because also noise may cause amplitudes in the wavelet spectrum it is necessary to use only the 95% significant amplitudes ($\tilde{W}_{\psi}^{(95)} f$). Thus, the global wavelet spectrum is a good representation of the dominant oscillations in $\Delta\mathrm{TEC}_{rel}$. However, the results presented here have to be treated carefully, because the relatively small number of available North Pole TEC-maps (available for 6 years, complete) is not appropriate to reliably represent the typical PWTO in the ionosphere. The typical periods of the zonal mean variation of $\Delta\mathrm{TEC}_{rel}$ are shown in Fig. 1, separately for winter (upper left panel) and summer season (upper right panel). Both demonstrate a very clear 27-day period. Regarding the strong dependence of TEC on the solar radiation, this periodicity can be assigned to the 27-day solar cycle, which occurs due to the rotation of the sun. The correlation between the ionospheric electron content and the 10.7 cm radio flux $F_{10.7}$, which is a proxy of the solar EUV radiation, is described in [4]. Furthermore, the 13.5-day period has to be allocated to the solar cycle, too, as it is its second harmonic.

Disregarding the solar rotation periods, the spectra of winter and summer months differ a lot. While strong oscillations with periods of 5 and 9 days occur during winter, there are only weak variations with periods between 2 and 10 days during summer. The high winter activity corresponds to the activity

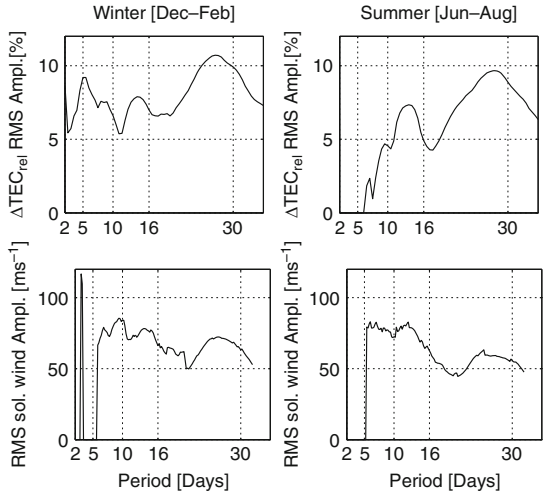


Fig. 1. *Upper panels*: zonal mean variation of $\Delta\mathrm{TEC}_{rel}$ (55°N, 2002–2007); *lower panels*: global wavelet spectra of the absolute solar wind speed (SWE, 2002–2007)

of PW in the middle atmosphere. Due to the wind systems, which act like a filter for PW, in the middle atmosphere the PW mainly occur during winter. However, the amplitude of the zonal mean variations in the stratosphere is low.

The solar wind, which has a higher variability than the $F_{10.7}$, has a significant influence on the ionosphere, too. The global wavelet spectrum of the absolute wind speed measured with the solar wind experiment (SWE) is shown in Fig. 1 (lower panels). A dominant 9-day period, which correlates with the 9-day period in $\Delta\mathrm{TEC}_{rel}$, can be found in the variation of the solar wind. This periodic variation was present at almost the same time in the solar wind and $\Delta\mathrm{TEC}_{rel}$ during several months in 2005. A 9-day period was also found in [6] in the thermospheric infrared data derived from the SABER instrument and was associated with the recurrence of coronal holes on the sun. A direct coupling between the sun and the infrared energy budget of the thermosphere was stated.

The origin of the dominant 5-day period in $\Delta\mathrm{TEC}_{rel}$ during winter stays an open question. The wavelet amplitude spectrum reveals that this wave was very strong in the winter 2004/2005. Neither the solar signal nor the stratospheric PW show a similar significant signal. Further investigations are necessary to clarify its origin. The global wavelet spectra for the standing and propagating waves with wavenumber 1 observed in $\Delta\mathrm{TEC}_{rel}$ during winter are shown in Fig. 2. In good agreement to stratospheric PW the wavenumber 1 PWTO-activity in $\Delta\mathrm{TEC}_{rel}$ is almost absent during summer (not shown here). During winter the wavenumber 1 PWTO contribute up to 7% to the ionospheric variability. Significant peaks can be often found at typical periods of PW which are at 5, 10, 16 and 30 days. Despite this similarity, concurrently observed oscillation in the stratosphere and the ionosphere, as described in a case study in [1], are seldom. Just like the results of numerical modelling, this comparison can not approve a direct correlation between stratospheric PW and ionospheric PWTO. But, it has to be considered that the wave properties may change through non linear interaction with e.g. tides or gravity waves.

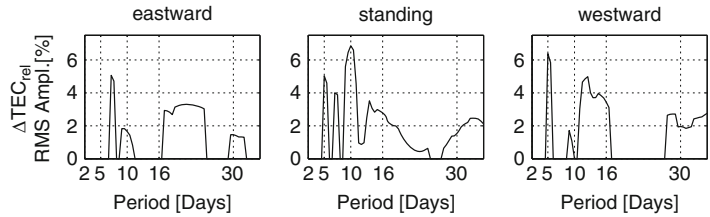


Fig. 2. Analyses of the wavenumber 1 in $\Delta\mathrm{TEC}_{rel}$ (55°N, 2002–2007) for the winter month (Dec–Feb). *Left panel*: eastward propagating waves; *middle panel*: standing waves; *right panel*: westward propagating waves

5 Summary and Conclusions

Regional hemispheric TEC maps have been analysed for the occurrence of PWTO in the ionosphere. The frequency-wavenumber-analyses using a combination of the Fourier and the wavelet analysis, which were applied for the signal decomposition, have demonstrated to be very suitable for the derivation of the wave parameters.

The zonal mean variations showed the highest amplitudes of all PWTO found in $\Delta\mathrm{TEC}_{rel}$. Most of the observed zonal mean oscillations were allocated to variations in the solar influence (EUV and solar wind), due to the rotation of the sun. This emphasises the major influence of the sun on the ionospheric variability, which includes the period range of PW.

The propagating and standing PWTO found in $\Delta\mathrm{TEC}_{rel}$ revealed a few typical properties of stratospheric PW. Nevertheless, a comparison could not approve a direct correlation between stratospheric PW and ionospheric PWTO. However, before dissipating at the turbopause height, the PW might modulate other waves like gravity waves or tides, which are able to propagate up to F2 Layer heights around 250 km. Indirect mechanisms like these might be able to transport the PW energy to higher altitudes. Such mechanisms have to be analysed in order to investigate the origin of the ionospheric PWTO.

References

1. Borries, C., Jakowski, N., Jacobi, C., Hoffmann, P., Pogoreltsev, A.: J. Atmos. Sol. Terr. Phys. **69**(17–18), 2442–2451 (2007)
2. Forbes, J.M., Palo, S., Zhang, X.: J. Atmos. Sol. Terr. Phys. **62**, 685–693(2000)
3. Hayashi, Y.: J. Meteorol. Soc. Jpn. **49**(2), 125–128 (1971)
4. Jakowski, N., Fichtelmann, B., Jungstand, A.: J. Atmos. Terr. Phys. **53**(11/12), 1125–1130 (1991)
5. Jakowski, N., Heise, S., Wehrenpfennig, A., S. Schlüter, Reimer, R.: J. Atmos. Sol. Terr. Phys. **64**(5–6), 729–735 (2002)
6. Mlynczak, M.G., Martin-Torres, F.J., Mertens, C.J., Marshall, B.T., Thompson, R.E., Kozyra, J.U., Remsberg, E.E., Gordley, L.L., Russell, J.M., Woods, T.: Geophys. Res. Lett. **35**, L05808 (2008)
7. Pancheva, D., Mitchell, N.J., Clark, R., Drobjeva, J., Lastovicka, J.: Ann. Geophys. **20**, 1807–1819 (2002)
8. Pogoreltsev, A., Vlasov, A.A., Fröhlich, K., Jacobi, C.: **69**(17–18), 2083–2101 (2007)

Statistical Significance of Gabor Frames Expansions: Simple Filtering Principles for Radar Wind Profiler Data

G. Teschke[1,2] and V. Lehmann[2]

[1] University of Applied Sciences Neubrandenburg, Institute for Computational Mathematics in Science and Technology, Brodaer Str. 2, 17033 Neubrandenburg, Germany, teschke@hs-nb.de

[2] Junior Konrad-Zuse-Fellow, Zuse Institute Berlin (ZIB), Takustr. 7, 14195 Berlin, Germany, and
Deutscher Wetterdienst, Meteorologisches Observatorium Lindenberg, D-15848 Lindenberg, Germany, Volker.Lehmann@dwd.de

Summary. A new signal processing method is presented for the suppression of intermittent clutter echoes in radar wind profilers. The technique presented makes use of a discrete Gabor frame expansion in combination with a statistical significance test. The rationale of this algorithm is outlined and an example using data obtained with an operational 482 MHz wind profiler is given.

1 Introduction

Radar wind profilers (RWP) were developed from MST-Radars and have meanwhile become standard instruments for measuring wind velocities in the atmosphere. Overviews of the technical and scientific aspects of RWP including its signal processing have been provided, among others, by e.g. [1]. Especially the routine application by weather services and the assimilation of the data in Numerical Weather Prediction Models is an indicator for the degree of maturation that this technology has achieved, see e.g. [6]. However, it is a matter of fact that sometimes large and unacceptable differences are observed between the profiler data and independent reference measurements. In many cases these differences are clearly attributable to either clutter echoes or Radio Frequency interference. Especially the problem of bird contamination has been well-known for more than a decade and it still is a research topic in RWP signal processing. There exist many attempts to reduce bird contamination, e.g. [5]. However, the disadvantage of all these methods is that the mitigation processing builds upon the Doppler spectra (either before or after spectral integration). Given the highly non-stationary characteristics of the intermittent clutter signal, it is necessary to deal with the problem before the Doppler

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*,
Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_47,

spectrum is estimated, because Fourier methods are generally inadequate for nonstationary signals. Further approaches that have tried to overcome these deficiencies by using wavelet representations were suggested by [2] and further by [4]. However, nonredundant wavelet filtering is in several cases also not best suited and causes undesired artifacts leading to erroneous filtering results. In this paper, we discuss a new signal-clutter separation method that circumvents these problems. It is based on a *Gabor frame decomposition* of the time series followed by the *statistical filtering approach* suggested by [5]. For an extensive description of the presented approach and a discussion in much greater detail we refer the interested reader to [3].

2 Classical Signal Model and Its Limitations

The classical RWP signal model assumption can be written as

$$\mathbf{S}[k] = \mathbf{I}[k]e^{i\omega k\Delta t} + \mathbf{N}[k], \tag{1}$$

where $\mathbf{I}[k] \sim N(0, \sigma^2_{\mathbf{I}})$ and $\mathbf{N}[k] \sim N(0, \sigma^2_{\mathbf{N}})$ are independent complex zero-mean Gaussian random vectors describing the atmospheric signal and the receiver noise, Δt is the sampling interval of the sequence and ω the mean Doppler frequency. Furthermore $\mathbf{I}[k]$ is narrowband compared to the receiver bandwidth and $|\omega| \leq \pi/\Delta t$ (Nyquist criterion). Because $\mathbf{S}[k]$ is the result of the demodulation of a real valued zero-mean and stationary Gaussian random process, the resulting Gaussian complex random process is also wide-sense stationary and zero-mean. Furthermore, the sequence has a vanishing pseudo-covariance, that is we have $\mathsf{E}(\mathbf{S}[k]\mathbf{S}[l]) = 0$. Such a process is usually called proper, circular or phase-invariant. Therefore,

$$\begin{aligned}(\mathbf{R})_{k,l} &= \mathsf{Cov}(\mathbf{S}[k], \mathbf{S}[l]) = \mathsf{E}(\mathbf{I}[k]\bar{\mathbf{I}}[l])e^{i\omega(k-l)\Delta t} + \mathsf{E}(\mathbf{N}[k]\bar{\mathbf{N}}[l]) \\ &= \sigma^2_{\mathbf{I}}\varrho[k-l]e^{i\omega(k-l)\Delta t} + \sigma^2_{\mathbf{N}}\delta_{k-l,0},\end{aligned}$$

where ϱ is specified below. While this is a classical assumption in radar signal processing, it is unknown for which maximal time series length this assumption can be made safely. We found that bird clutter signals are significantly nonstationary over typically used dwell times of about 30–60 s. The associated autocovariance function can be expressed as follows

$$\mathsf{ACov}(k) = \sigma^2_{\mathbf{I}}\varrho[k]e^{i\omega k\Delta t} + \sigma^2_{\mathbf{N}}\delta_{k,0} = \sigma^2\rho[k]\ , \tag{2}$$

where we set $\sigma^2 := \sigma^2_{\mathbf{I}} + \sigma^2_{\mathbf{N}}$ and $\rho[k] := \frac{\sigma^2_{\mathbf{I}}\varrho[k]e^{i\omega k\Delta t} + \sigma^2_{\mathbf{N}}\delta_{k,0}}{\sigma^2_{\mathbf{I}} + \sigma^2_{\mathbf{N}}}$, while assuming $\varrho[k] = e^{-2\pi^2 w^2 k^2 \Delta t^2}$. In reality, however, there is sometimes a third component contributing to the signal, namely clutter [7], so that the signal model must be written as:

$$\mathbf{S}[k] = \mathbf{I}[k]e^{i\omega k\Delta t} + \mathbf{N}[k] + \mathbf{C}[k]\ . \tag{3}$$

Clutter is the totality of undesired echoes and interfering signals, therefore it is impossible to generalize the properties of $\mathbf{C}[k]$. In the case of RWP, clutter includes in particular echoes from airborne objects such as aircraft and birds as well as returns from the ground. Interfering signals may be caused by other radio transmitters operating in the RWP receiver band. In the remainder of the paper, we restrict ourselves to intermittent clutter signals and it removal from $\mathbf{S}$.

3 Gabor Frame Expansions for Discretely Sampled Signals

Assume we are given some discrete and finite time (periodic) signal $\tilde{\mathbf{S}}$ with sampling points $n = 0, \ldots, N-1$, that is $\tilde{\mathbf{S}}[n] = \tilde{\mathbf{S}}[n+N]$. We therefore have to periodize the analysis and synthesis windows as well,

$$\tilde{\mathbf{h}}[n] = \sum_l \mathbf{h}[n + lN], \quad \tilde{\mathbf{g}}[n] = \sum_l \mathbf{g}[n + lN].$$

Slightly abusing the notation, we omit the tilde denoting periodic (finite) functions in the following. The signal $\mathbf{S}$ can be discretely represented by

$$\mathbf{S}[n] = \sum_{m=0}^{M-1} \sum_{k=0}^{K-1} a_{m,k} \mathbf{h}_{m,k}[n], \tag{4}$$

whereas the Gabor coefficients can be derived from

$$a_{m,k} = \sum_{n=0}^{N-1} \mathbf{S}[n] \bar{\mathbf{g}}_{m,k}[n]. \tag{5}$$

Introducing integers ΔM and ΔK and the toral component $W_N = \exp[2\pi i/N]$, the discrete analysis and synthesis windows can be rewritten as

$$\mathbf{h}_{m,k}[n] = \mathbf{h}[n - m\Delta M] W_N^{nk\Delta K},$$

$$\mathbf{g}_{m,k}[n] = \mathbf{g}[n - m\Delta M] W_N^{nk\Delta K}.$$

As can be seen, ΔM denotes the time and ΔK the frequency step size. They correspond to T and Ω. In our setting they are constrained by $\Delta M \cdot M = \Delta K \cdot K = N$. The reconstruction formula becomes

$$\mathbf{S}[j] = \sum_{m=0}^{M-1} \sum_{k=0}^{K-1} a_{m,k} \mathbf{h}_{m,k}[j] = \sum_{l=0}^{N-1} \mathbf{S}[l] \sum_{m=0}^{M-1} \sum_{k=0}^{K-1} \bar{\mathbf{g}}_{m,k}[l] \mathbf{h}_{m,k}[j],$$

where we have assumed that the following biorthogonality relation is fulfilled,

$$\sum_{m=0}^{M-1}\sum_{k=0}^{K-1} \bar{\mathbf{g}}_{m,k}[l]\mathbf{h}_{m,k}[j] = \delta_{l,j}.$$

It can be shown that the biorthogonality relation is satisfied if

$$\sum_{j=0}^{N-1} \mathbf{h}[j+qK]W_N^{-jpM}\bar{\mathbf{g}}[j] = \frac{N}{MK}\,\delta_{p,0}\delta_{q,0} \tag{6}$$

for $0 \le p \le \Delta M - 1$ and $0 \le q \le \Delta K - 1$. System (6) can be rewritten in matrix form: Let $\mathbf{v} = (N/(MK), 0, \ldots, 0)^T$ be a vector of length $\Delta M \Delta K$ and $\mathbf{g} = (\mathbf{g}[0], \ldots, \mathbf{g}[N-1])$ the vector representing the discretely sampled dual frame, and let $\mathbf{A}$ be the matrix of size $\Delta M \Delta K \times N$ with entries $\mathbf{A}_{(p,q),j} = \bar{\mathbf{h}}(j+qK)W_N^{jpM}$, then the dual frame atom $\mathbf{g}$ is the solution of the linear system

$$\mathbf{A}\mathbf{g} = \mathbf{v}. \tag{7}$$

For oversampling $\Delta M \Delta K < N$, system (7) is under-determined, and the solution is no longer unique and therefore there is a variety of possible dual frame atoms $\mathbf{g}$. One suitable choice (beside optimal localizing window functions) is given by the minimum norm solution $\mathbf{g}_{min} = \mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{v}$.

4 Statistical Significance, Filtering and a Real Example

First, we observe that

$$|a_\lambda|^2 = \sum_{n=0}^{N-1} \mathbf{S}[n]\mathbf{g}_\lambda[n] \sum_{l=0}^{N-1} \bar{\mathbf{S}}[l]\bar{\mathbf{g}}_\lambda[l].$$

With $\mathsf{E}\mathbf{S}[n] = 0$ and $\mathsf{E}\mathbf{S}[n]\bar{\mathbf{S}}[n+l] = \sigma^2\rho[l]$ we obtain $\mathsf{E}|a_\lambda|^2 = \sigma^2\langle\rho * \mathbf{g}_\lambda, \mathbf{g}_\lambda\rangle$ and $\mathsf{Cov}(|a_\lambda|^2, |a_\eta|^2) = \sigma^4|\langle\rho * \mathbf{g}_\lambda, \mathbf{g}_\eta\rangle|^2$. The '$*$'-symbol stands here for the discrete convolution. Therefore,

$$\mathsf{Var}|a_\lambda|^2 = \sigma^4|\langle\rho * \mathbf{g}_\lambda, \mathbf{g}_\lambda\rangle|^2 = (\mathsf{E}|a_\lambda|^2)^2 \text{ and thus } \frac{(\mathsf{E}|a_\lambda|^2)^2}{\mathsf{Var}|a_\lambda|^2} = 1 \tag{8}$$

which holds true for independent as well as dependent samples $\mathbf{S}[n]$ that follow a distribution which is determined by its moments. In order to construct a statistical test that verifies property (8), we have to find optimal estimators for $\mathsf{E}|a_\lambda|^2$ and $\mathsf{Var}|a_\lambda|^2$ that are based on a finite number of observations. To this end, we introduce an index subset $\Omega_\lambda \subset \Lambda$ containing λ and $L-1$ further different indices η, i.e. $|\Omega_\lambda| = L$. As an estimator for $\mathsf{E}|a_\lambda|^2 = \sigma^2\langle\rho * \mathbf{g}_\lambda, \mathbf{g}_\lambda\rangle$, which is based on L neighboring observation variables, we define

$$\hat{E}(\Omega_\lambda) := \frac{1}{C_{\Omega_\lambda}} \sum_{\eta\in\Omega_\lambda} |a_\eta|^2 \text{ with } C_{\Omega_\lambda} = \sum_{\eta\in\Omega_\lambda} \frac{\langle\rho * \mathbf{g}_\eta, \mathbf{g}_\eta\rangle}{\langle\rho * \mathbf{g}_\lambda, \mathbf{g}_\lambda\rangle} > 1. \tag{9}$$

Assuming there exists some small $\varepsilon > 0$ with $\sum_{\eta',\eta\in\Omega_\lambda} |\langle \rho * \mathbf{g}_{\eta'}, \mathbf{g}_\eta\rangle|^2 \leq C_{\Omega_\lambda}^{2-\varepsilon}$, estimator (9) consistent, for details see [3]. By the same reasoning, we define a consistent estimator for variance,

$$\hat{V}(\Omega_\lambda) := C \sum_{\eta\in\Omega_\lambda} (|a_\eta|^2 - \hat{E}(\Omega_\lambda))^2, \tag{10}$$

where the constant is defined by

$$C^{-1} := 2 \sum_{\eta\in\Omega_\lambda} \frac{c_\eta^2}{c_\lambda^2} + (L - 2C_{\Omega_\lambda}) \left(1 + \frac{1}{(\sum_\eta c_\eta)^2} \sum_{\xi,\alpha\in\Omega_\lambda} c_{\xi,\alpha}^2\right). \tag{11}$$

To identify now intermittent clutter (the nonstationary signal component), we proceed as follows: In a first step, define the index set representing the k-th row, which we denote by $\Omega_k = \{(m,k) : m = 0,\ldots,M-1\}$, and sort for each k the sequence $\{|a_{m,k}|^2\}_{(m,k)\in\Omega_k}$ in decreasing order. That is, we derive the order statistic of $\{|a_{m,k}|^2\}_{(m,k)\in\Omega_k}$ which we denote by $\{|[a]_{m,k}|^2\}_{(m,k)\in\Omega_k}$ ($[\cdot]$ stands for the order statistic map). Therefore, we have $|[a]_{m,k}|^2 \geq |[a]_{m+1,k}|^2$ for all $(m,k) \in \Omega_k$. For $l = 0,\ldots,M-1$, we define subsets $\Omega_k(l) = \{(m,k) : m = l,\ldots,M-1\}$. The largest coefficients are stepwise discarded, which has the goal of eliminating the clutter signal component. Using the quantities $\hat{E}(\Omega_k(l))$ and $\hat{V}(\Omega_k(l))$ of the subset, the test statistics ϑ is computed for $l = 0,\ldots,M-1$ as long as $\vartheta(|[a]_{l,k}|^2) := \frac{(\hat{E}(\Omega_k(l)))^2}{\hat{V}(\Omega_k(l))} < 1$ holds. The largest coefficient of the first subset for which the test (positive for clutter) is not satisfied (a clutter-free subset) is then taken as a threshold for a frequency-dependent identification of the clutter component. All coefficients $|a_{m,k}|^2$ greater than the threshold are regarded as clutter. Based on this test, we introduce a clutter index set as $\Omega_k^c := \{(m,k) : \vartheta(|[a]_{m,k}|^2) < 1\ ,\ m = 0,\ldots,M-1\}$. The coefficients $a_{m,k} \in \Omega_k^c$ are finally set to $t_k e^{i\arg a_{m,k}}$, where t_k is the average value of the remaining coefficients, $t_k = \frac{1}{|\Omega_k\setminus\Omega_k^c|} \sum_{(m,k)\in\Omega_k\setminus\Omega_k^c} |a_{m,k}|$. Consequently, the filtered signal $\mathbf{S}$ is given by

$$\Phi(\mathbf{S})[n] = \sum_{k=0}^{K-1} \left\{ \sum_{(m,k)\in\Omega_k\setminus\Omega_k^c} a_{m,k}\mathbf{h}_{m,k}[n] + \sum_{(m,k)\in\Omega_k^c} t_k e^{i\arg a_{m,k}} \mathbf{h}_{m,k}[n] \right\}. \tag{12}$$

To show the performance of the proposed algorithm, we process data that were obtained during routine operation of a 482 MHz wind profiler radar of the Deutscher Wetterdienst at Bayreuth, Germany in the fall of 2005. We consider data taken in the south beam of the radar wind profiler at range gate 9 (1.6 km height agl, dwell at 00:09:45 UTC). Figure 1 shows a time series in which strong intermittent clutter (bird echo) can be recognized. The results of the filtering procedure illustrate that the method completely eliminates the nonstationary signal component.

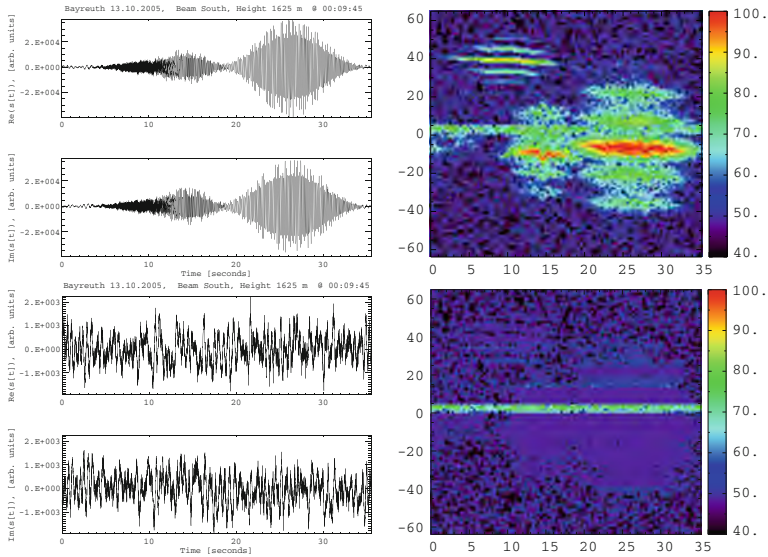


Fig. 1. *Top left*: real time series; *top right*: Gabor spectrum of this time series; *bottom right*: filtered Gabor spectrum; *bottom left*: reconstructed time series

References

1. Doviak, R.J., Zrnic, D.S.: Doppler Radar and Weather Observations. Academic, San Diego (1993)
2. Jordan, J.R., Lataitis, R.J., Carter, D.A.: Removing ground and intermittent clutter contamination from wind profiler signals using wavelet transforms. J. Atmos. Ocean. Tech. **14**, 1280–1297 (1997)
3. Lehmann, V., Teschke, G.. Advanced intermittent clutter filtering for radar wind profiler: signal separation through a gabor frame expansion and its statistics. Ann. Geophys. **26**(4), 759–783 (2008)
4. Lehmann, V., Teschke, G.: Wavelet based methods for improved wind profiler signal processing. Ann. Geophys. **19**, 825–836 (2001)
5. Merritt, D.A.: A statistical averaging method for wind profiler Doppler spectra. J. Atmos. Ocean. Tech. **12**(5), 985–995 (1995)
6. Monna, W.A., Chadwick, R.B.: Remote-sensing of upper-air winds for weather forecasting: Wind-profiler radar. B. World Meteorol. Organ. **47**(2), 124–132 (1998)
7. Muschinski, A., Lehmann, V., Justen, L., Teschke, G.: Advanced radar wind profiling. Meteorol. Z. **14**(5), 609–626 (2005)

Minisymposium *Multirate Time Integration for Multiscaled Systems*

E. Jan W. ter Maten[1] and Michael Günther[2]

[1] NXP Semiconductors and TU Eindhoven, the Netherlands
E.J.W.ter.Maten@tue.nl
[2] Bergische Universität Wuppertal, Wuppertal, Germany
Michael.Guenther@math.uni-wuppertal.de

Multiphysical[1] and network modeling usually lead to coupled systems that exhibit largely different timescales. In time domain it is called multirate, in space multiscale. Efficient algorithms need to take these phenomena into account. Such methods are specially requested by industry. As a practical example one can think of a cellular phone, which consists of coupled digital and analog subcircuits that operate at vastly different frequencies. The inclusion of memory cells extends the difference in dynamics even more. One can find other examples in astrophysics; in multibody systems of vehicles [1]; in robotics (where multiple time scales and hierarchy in the dynamics may exhibit); in coupling mechanical and electrical systems; in reaction and diffusion processes involving multiple chemical components; in combustion engines with chain drives. Similar effects have to be taken into account when designing Integrated Circuits (ICs) [5] and Power Systems [2].

To speed up numerical integration of ordinary differential equations (ODEs), three research directions have been followed in the last decades, all based on exploiting multirate behaviour in time:

- Multi-method schemes (for systems containing both non-stiff and stiff parts): here a partitioning is done on the level of the discretization scheme, i.e., an explicit scheme is used for the non-stiff parts, and an implicit method for the stiff ones.
- Multi-order schemes (for non-stiff problems only): here the same explicit method, and the same step size is used for all parts, but the order of the method is chosen according to the activity level of the subsystem.
- Multirate schemes (for both stiff and non-stiff problems): here the same (implicit or explicit method) with the same order is applied to all subsystems, but the step size is chosen according to the activity level.

[1] This minisymposium was an event of the ECMI Special Interest Group on Scientific Computing in Electronics Industry

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*,
Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_48,

Research on Multirate Time Integration (MTI) for DAEs for semiconductor industries started based on Backward Differentiation Formula and on Rosenbrock-Wanner methods [5, 6]. For reaction diffusion equations and for systems of hyperbolic conservation laws MTI was applied to the system of ODEs that arise after semidiscretization of these PDEs – also starting with Rosenbrock methods [3, 4]. The multirate time stepping caused deeper analysis in order to guarantee that stability, monotonicity, interface, stiff source term and conservation law constraints are met.

This minisymposium addressed particular aspects of the various multirate time integration methods

- A ROW-based hierarchical mixed multirate method that can deal with an arbitrary amount of subcircuits for differential-algebraic equations of index 1 [5] [Univ. of Wuppertal & Infineon AG & Qimonda AG].
- A self-adjusting multirate time stepping strategy, where the partitioning into different levels of slow to fast components is performed automatically during the time integration [4] [Univ. of Amsterdam & CWI Amsterdam].
- A multi-step multirate implementation, where the interpolation error and the coarse discretisation error is controlled by the macro stepsize, while the micro stepsize controls the fine discretisation errors for the fast state part [6] [TU Eindhoven & Philips & NXP Semiconductors].
- Two families of explicit multirate time discretization methods based on Adams–Bashforth and partitioned Runge–Kutta schemes for hyperbolic conservation laws, which avoid the necessity to take small global time steps restricted by the largest CFL number [3] [Virginia Tech, Blackburg, VA].

References

1. Arnold, M.: Multi-rate time integration for large scale multibody system models In: Eberhard, P. (ed.) IUTAM Symposium on Multiscale Problems in Multibody System Contacts, pp. 1–10. Springer, Berlin (2007)
2. Chen, J., Crow, M.L.: A variable partitioning strategy for the multirate method in power systems. IEEE Trans. Power Syst. **23-2**, 259–266 (2008)
3. Constantinescu, E.M.: Adaptive numerical methods for large scale simulations and data assimilation. PhD-Thesis Virginia Polytechnic Institute and State University (2008)
4. Savcenco, V.: Multirate numerical integration for ordinary differential equations. PhD-Thesis University of Amsterdam (2008)
5. Striebel, M.: Hierarchical mixed multirating for distributed integration of DAE network equations in chip design. PhD-Thesis University of Wuppertal (2006). See also: Fortschritt-Berichte VDI, Reihe 20 Rechnerunterstützte Verfahren Nr 404, VDI Verlag GmbH, Düsseldorf (2006)
6. Verhoeven, A.: Redundancy reduction of IC models by multirate time integration and model order reduction. PhD-Thesis TU Eindhoven (2008)

Domain Decomposition Based Multirating and its Perspective in Circuit Simulation

Michael Striebel[1], Andreas Bartel[2], and Michael Günther[2]

[1] Chemnitz University of Technology, Research Group Mathematics in Industry and Technology, D-09107 Chemnitz, Germany, michael.striebel@mathematik.tu-chemnitz.de

[2] Bergische Universität Wuppertal, Department of Mathematics, Chair of Applied Mathematics/Numerical Analysis, D-42097 Wuppertal, Germany, bartel@math.uni-wuppertal.de, guenther@math.uni-wuppertal.de

Summary. Based on domain decomposition, multirate time integration takes into account largely different timescales. In this class, a mixed multirate scheme and its application to an arbitrary number of subsystems is outlined. Moreover, the matter of activity change and the connection to model order reduction is discussed.

1 Domain Decomposition as Modular Modelling

In PDE domain decomposition, a domain is split into different sub-domains. For the consistency of the overall problem, these sub-domains have to be linked via artificial boundary conditions and Lagrangian multipliers to match the solution at the boundaries. A similar approach is used naturally in circuit simulation packages. Here complex circuits are decomposed into different parts with respect to their function. This approach enables to model the sub-circuits separately. In contrast to spatial (sub-)domains within the PDE case, we have to deal here with (sub-)circuits described by their topology, and hence

- Matching boundary conditions are transferred to artificial voltage sources, which match node potentials at the boundaries of the sub-circuits.
- Branch currents through artificial voltage sources play the role of Lagrangian multipliers linking the sub-domains.

Applying charge oriented modified nodal analysis [4, 9] to a decomposed circuit with $r \in \mathbf{N}$ subsystems, the mathematical models yields the following type of coupled differential-algebraic equations (DAEs):

$$\left.\begin{aligned} 0 &= \mathcal{A}_\lambda \dot{y}_\lambda + f_\lambda(x_\lambda, t) + \mathcal{A}_{w_\lambda} w \\ 0 &= y_\lambda - q_\lambda(x_\lambda) \end{aligned}\right\} \quad \text{for } \lambda = 1, \ldots, r, \tag{1a}$$

$$0 = \mathcal{A}_{w_1}^t x_1 + \cdots + \mathcal{A}_{w_r}^t x_r \,. \tag{1b}$$

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*, Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_49,

Here $x_\lambda \in \mathbf{R}^{n_\lambda}$ denotes the unknown node potentials and branch currents of voltage defining elements, $y_\lambda \in \mathbf{R}^{m_\lambda}$ contains the unknown charges and fluxes, and the functions q_λ, f_λ describe the contribution of the reactive and nonreactive components of the λth subsystem, whose topology determines the incidence matrix $\mathcal{A}_\lambda \in \{-1, 0, 1\}^{n_\lambda \times m_\lambda}$. The coupling quantity $w \in \mathbf{R}^M$ denotes the set of branch currents through artificial voltage sources. These are implicitly defined by the coupling equation (1b), the constitutive equations for the artificial voltage sources. Finally, the incidence matrices $\mathcal{A}_{w_\lambda} \in \{-1, 0, 1\}^{n_\lambda \times M}$ $(\lambda = 1, \ldots, r)$ relate w to the corresponding terminal of the respective subcircuit as input source.

In the following, we restrict ourself to the case where (1) states an index-1 problem, in the sense that:

(C1) The overall system (1a,1b) has index 1 w.r.t. $x_1, \ldots, x_r, w$.
(C2) All systems (1a) define index-1 systems w.r.t. x_λ (given w).

The latter holds, if there are neither CV-loops nor LI-cutsets in the subcircuits [3]. Virtual voltage sources can be associated with the coupling and we can show that (C1) holds if there are also no loops of capacitors, voltage sources and virtual voltage sources in the overall circuit [1,9].

Under these conditions (1) is equivalent [9] to the semi-explicit system:

$$\left.\begin{aligned} \dot{y}_\lambda &= f_\lambda(z_\lambda, w), \\ 0 &= h_\lambda(y_\lambda, z_\lambda, w), \end{aligned}\right\} \quad \text{for } \lambda = 1, \ldots, r, \tag{2a}$$

$$0 = g(z_1, \ldots, z_r)\,, \tag{2b}$$

where f_λ, h_λ are linear in w and y_λ, w, respectively and g is linear in $z_1, \ldots, z_r$. Notice the abuse of notation in y_λ, f_λ.

2 Mixed Multirate

The functional diversity in modularly modelled systems causes a heterogeneous distribution of activity. At each time the quantities "node potential" and "element currents" may show the tendency to change rapidly in some regions of the circuit whereas only minor fluctuation can be recognised in other parts. *Multirate methods* accommodate this behaviour and reduce computational expenses. The basic idea of these schemes is to prevent parts to be integrated more often than necessary to meet prescribed error tolerances. To this end, we associate activity levels to step size proposals and are able to split large systems in subsystems which operate on different time scales with differing optimal step sizes. Now, domain decomposition based multirating means to use these (differing) optimal step sizes for the respective subsystems to define an overall method.

For simplicity, we formulate the multirate method first for a coupled system (of *latent* y_L and *active* y_A variables) for ordinary differential equations:

$$\dot{y}_L = f_L(y_L, y_A), \qquad \dot{y}_A = f_A(y_L, y_A).$$

At the current time point t_0 (with given $y(t_0) = (y_L, y_A)^t = y_0$), we suppose that the latent part (subscript L) can be integrated with one macrostep $\mathcal{H}_L$ whereas a sequence of q microsteps $\mathcal{H}_{A,1}, \ldots, \mathcal{H}_{A,q}$ is needed for the active part (subscript A) to reach $t_0 + \mathcal{H}_L$. Numerically, a subset of systems proposing a large individual step size is identified as latent; the others demand a small step, and are therefore active.

Various approaches are being developed [4, 8, 11]. We concentrate on the application of Rosenbrock-Wanner (ROW), i.e., one-step, methods. A detailed discussion can be found in [9]. Here we just outline the basic principles. In its most general way the one-step formalism of this procedure is given by:

$$y_{L,1} = y_{L,0} + \sum_{i=1}^{s_L} b_i^L \cdot l_i^L, \tag{3a}$$

$$y_{A,\mu} = y_{A,\mu-1} + \sum_{i=1}^{s_A} b_i^A \cdot l_i^{A,\mu} \quad (\mu = 1, \ldots, q), \tag{3b}$$

$$l_i^L = \Phi_L(\mathcal{H}_L;\, y_{L,0}, Y_i^A, l_1^L, \ldots, l_{s_L}^L) \quad (i = 1, \ldots, s_L), \tag{3c}$$

$$l_i^{A,\mu} = \Phi_A(\mathcal{H}_{A,\mu};\, y_{A,\mu-1}, Y_i^{L,\mu}, l_1^{A,\mu}, \ldots, l_{s_A}^{A,\mu}) \quad (i = 1, \ldots, s_A), \tag{3d}$$

where $\Phi_\mathcal{L}$ denotes an $s_\mathcal{L}$ stage ROW scheme with coefficients $b^\mathcal{L}$, $A^\mathcal{L}$, $B^\mathcal{L}$, $\Gamma^\mathcal{L}$ ($\mathcal{L} = L, A$). For ROW schemes, (3c,d) determine the increments l_i^L, $l_i^{A,\mu}$ by linear relations. The coupling of latent and active subsystems is performed by the terms Y_i^A and $Y_i^{L,\mu}$, which will be defined from the increments l^L and l^A, respectively. Y_i^A and $Y_i^{L,\mu}$ signify the sampling of the fast variables on the coarse grid and vice versa.

Mixed multirate [2] is characterised by a "compound step" and a series of "later microsteps". In the former, the macrostep (3a) and the first microstep (3b) ($\mu = 1$) are computed at once. Y_i^A, $Y_i^{L,1}$ are determined in RK-like manner, which employs additionally: coupling coefficients $D^{AL}, D^{LA}, N^{AL}, N^{LA}$, and scaling of increments $l_i^{A,1}$ and l_i^L by the *step size ratio* $\mathfrak{m} = \frac{\mathcal{H}_L}{\mathcal{H}_{A,1}}$ and $\mathfrak{m}^{-1}$, respectively. For the later microsteps *dense output* formulae [6] are applied to get reasonable values $Y_i^{L,2}, \ldots, Y_i^{L,q}$.

2.1 Mixed Multirate for Circuit Simulation

In the case of coupled index-1 networks (C1-C2) with active and latent variables, we have to solve the equivalent semi-explicit DAE:

$$\begin{aligned} \dot{y}_L &= f_L(z_L, w) & \dot{y}_A &= f_A(z_A, w) \\ 0 &= h_L(y_L, z_L, w) & 0 &= h_A(y_A, z_A, w) \end{aligned} \tag{4}$$
$$0 = g(z_L, z_A),$$

where the coupling quantity w is assumed to be latent, too. Using a ROW scheme like (3) for (4), we have to add increments for the algebraic variables. For $s(=s_{\mathcal{L}})$ stages, weights $b^{\mathcal{L}}$, and increments $l^{\mathcal{L}}$, $k^{\mathcal{L}}$, p, we have (sloppily):

$$\begin{pmatrix} y_{L,1} \\ z_{L,1} \\ w_1 \end{pmatrix} = \begin{pmatrix} y_{L,0} \\ z_{L,0} \\ w_0 \end{pmatrix} + (b^L)^t \begin{pmatrix} l^L \\ k^L \\ p \end{pmatrix}, \quad \begin{pmatrix} y_{A,1} \\ z_{A,1} \end{pmatrix} = \begin{pmatrix} y_{A,0} \\ z_{A,0} \end{pmatrix} + (b^A)^t \begin{pmatrix} l^A \\ k^A \end{pmatrix}. \quad \text{(5a)}$$

According to (3c,d) the stage increments are defined by the linear system

$$M^\star \cdot \left(l_i^L,\ k_i^L \mid l_i^A, k_i^A \mid p_i\right)^t = \mathrm{RHS}_i, \quad \text{for } i = 1, \ldots, s \quad \text{(5b)}$$

with the system matrix $M^\star =$

$$\left(\begin{array}{cc|cc|c} \mathrm{I}_{y_L} & -\mathcal{H}_L\gamma^L\frac{\partial f_L}{\partial z_L} & & & -\mathcal{H}_L\gamma^L\frac{\partial f_L}{\partial w} \\ -\gamma^L\frac{\partial h_L}{\partial y_L} & -\gamma^L\frac{\partial h_L}{\partial z_L} & & & -\gamma^L\frac{\partial h_L}{\partial w} \\ \hline & & \mathrm{I}_{y_A} & -\mathcal{H}_A\gamma^A\frac{\partial f_A}{\partial z_A} & -\frac{1}{\mathfrak{m}}\cdot\mathcal{H}_A\nu^{AL}\frac{\partial f_A}{\partial w} \\ & & -\gamma^A\frac{\partial h_A}{\partial y_a} & -\gamma^A\frac{\partial h_A}{\partial z_A} & -\frac{1}{\mathfrak{m}}\cdot\nu^{AL}\frac{\partial h_A}{\partial w} \\ \hline & -\gamma^L\frac{\partial g}{\partial z_L} & & -\mathfrak{m}\cdot\nu^{LA}\frac{\partial g}{\partial z_A} & \end{array}\right)$$

and a right-hand side RHS_i depending on stepsizes $\mathcal{H}_L, \mathcal{H}_A$, step size ratio $\mathfrak{m}$, increments $l_j^{\mathcal{L}}, k_j^{\mathcal{L}}, p_j$ of the former steps $j = 1, \ldots, i-1$. As above, two ROW coefficient sets (labels L, A) and additional coupling coefficients are employed.

In the *later microsteps*, it remains to solve $[\dot{y}_A = f_A, 0 = h_A]$ for unknown y_A, z_A, where $w(t)$ enters the right-hand-side. As we introduced the coupling quantity w as an additional latent unknown, it is already computed (in the compound step) and a cheap approximation to $w(t_0 + \theta \cdot \mathcal{H}_L)$ is obtained via dense output formulae. For a detailed definition, see [9].

The *method's coefficients* have to be determined such that the accuracy of the local approximation is of a prescribed order. To this end, B-series for ODEs [5] are adapted to our coupled problem (2). As for mixed multirate methods, the order conditions depend on the step size ratio $\mathfrak{m}$. Therefore, the coefficients depend on this quantity and have to be computed during integration.

2.2 Hierarchical Mixed Multirate

Aiming a multirate method that can deal with an arbitrary amount of activity levels, *hierarchical mixed multirate* seems to be the most feasible approach. The main idea is to nest compound steps and later micro-steps in a way, that at each time merely a two-level multirate scheme is engaged. At any time point of integration each subsystem has either the status *asleep* or *latent* or *active*. A part is *asleep* if the last time point at which an approximation is

available is beyond the current one. The set of non-sleeping subsystems is split into latent and active subsets. Due to this decomposition, a *compound step* can be applied to the non-sleeping part. Otherwise (only active variables are present) *later microsteps* are executed. The sleeping subsystems contribute to the current step via dense-output.

2.3 Implementation and Trapping Events

For electric network descriptions (1), a hierarchical mixed multirate method of order 2 has been embedded into Qimonda's in-house simulator `titan`. Step size control is performed with an embedded scheme of order 1. Linear transformations are applied to this multirate algorithm (5) such that the resulting method can be used for the network problem (1) directly.

Great importance is attached to the problem of traversing signals, which can force sleeping subsystems to "wake up" during a macrostep (t_{n-1} to $t_n = t_{n-1} + \mathcal{H}_L$). This causes an a-posteriori rejection of that macrostep. The detection of such situations is based on comparing pin voltages of connected subsystems: any time point $t_{\text{wup}} \in (t_{n-1}, t_n)$ where the difference of the voltages computed from the non-sleeping part and the corresponding voltage computed by a dense-output formula applied to the sleeping part becomes too large, is considered a wake up point. Moreover, not the whole macrostep is restored, but a re-initialisation at t_{wup} is performed using again dense-output formulae to get appropriate initial values. For a detailed description of an industrial test case (chain of inverters), we refer to [10].

3 Connection of Multirate to Model Order Reduction

To quickly get evidence of the behaviour of complex circuits, simulation techniques need to be adapted. *Multirate* tackles this task from an algorithmic point of view by incorporating subsystems' behaviour in the numerical procedure (sampling) for the overall system (1a)–(1b). *Model order reduction* (MOR) starts on the modeling level. It seeks to replace the r subsystems (1a) of presumably high order (large number of unknowns) with order reduced models. This is achieved in the following way: given the input w, the λth substitute model with essential state variable $\hat{x}_\lambda$ returns an output $\hat{\mathcal{A}}^t_{w_\lambda}\hat{x}_\lambda$ which approximates the corresponding output $\mathcal{A}^t_{w_\lambda}x_\lambda$ of the full system sufficiently accurate. For nonlinear problems, MOR basically is done by scanning the full system in a training phase and extracting dominant information that determines the reduced substitute model; this is realised, e.g., in the trajectory piecewise linearisation approach (TPWL) [7].

There are some analogies of multirate and MOR that could be used to improve or merge both strategies. In multirate the latent part contributes to the later microsteps just in terms of the terminal quantities, i.e., the output

$\mathcal{A}^t_{w_\lambda} x_\lambda$. The corresponding values on the fine time grid are derived from information gathered on the coarse one. This can be viewed as training of a current- or voltage source replacing the large latent subcircuit. If the latent part was replaced by a reduced order model incorporating more dynamical effects, the procedure could become more stable and we can hope to act out the full multirate behaviour. Moreover, the combination of compound step and later microsteps can be regarded as on-the-fly training. If this can be transferred to MOR, it may pave the way to construct models that are less sensitive to varying input signals and which could be produced whenever needed.

4 Conclusion

A multirate scheme for circuit simulation that can deal with an arbitrary number of subsystems has been derived, where domain decomposition of large electrical circuits is achieved by introducing extra variables. The hierarchical multirate method has been embedded in a sophisticated industrial simulator.

Future tasks are a partitioning strategy and step size control tailored to multirate needs. Step size control should be improved as we want to combine very large with small steps. Higher order schemes and extensions to higher index problems are desirable. The perspective is to use analogies of multirate and model order reduction to combine and enhance both approaches.

References

1. Arnold, M., Günther, M.: Preconditioned dynamic iteration for coupled differential-algebraic systems. BIT, **41**(1), 1–25 (2001)
2. Bartel, A., Günther, M., Kværnø, A.: Multirate methods in electrical circuit simulation. In: Anile, A.M., Capasso, V., Greco, A. (eds.) Progress in Industrial Mathematics at ECMI 2000, Mathematics in Industry, vol. I, pp. 258–265. Springer, Berlin (2002)
3. Estévez Schwarz, D., Tischendorf, C.: Structural analysis for electric circuits and consequences for MNA. Int. J. Circ. Theor. Appl. **28**, 131–162 (2000)
4. Günther, M., Feldmann, U., ter Maten, E.J.W.: Modelling and discretization of circuit problems. In: Ciarlet, P.G., Schilders, W.H.A., ter Maten, E.J.W. (eds.) Numerical Methods in Electromagnetics, vol. XIII of Handbook of Numerical Analysis, pp. 523–659. Elsevier, North Holland (2005)
5. Hairer, E., Nørsett, S.P., Wanner, G.: Solving Ordinary Differential Equations I – Nonstiff Problems, Second revised Edition, Springer, Berlin (2000)
6. Hairer, E., Wanner, G.: Solving Ordinary Differential Equations II – Stiff and Differential-Algebraic Problems. Second revised edition, Springer, Berlin (1996)
7. Rewieński, M.J., White, J.: A trajectory piecewise-linear approach to model order reduction and fast simulation of nonlinear circuits and micromachined devices. IEEE Trans. CAD Int. Circ. Syst. **22**(2), 155–170 (2003)
8. Savcenco, V., Hundsdorfer, W.H., Verwer, J.G.: A multirate time stepping strategy for stiff odes. BIT **47**, 137–155 (2007)

9. Striebel, M.: Hierarchical Mixed Multirating for Distributed Integration of DAE Network Equations in Chip Design. Number 404 in Fortschritt-Berichte VDI Reihe 20. VDI-Verlag Düsseldorf (2006)
10. Striebel, M., Bartel, A., Günther, M.: A Multirate ROW-Scheme for Index-1 Network Equations, Appl. Numer. Math. **59**, pp. 800–814 (2009)
11. Verhoeven, A., El Guennouni, A., ter Maten, E.J.W., Mattheij, R.: Multirate methods for the transient analysis of electrical circuits. PAMM, Proceedings to GAMM 2005 **5**, 821–822 (2005)

Multirate Numerical Integration for Stiff ODEs

V. Savcenco and R.M.M. Mattheij

Technische Universiteit Eindhoven, P.O. Box 513, 5600 MB Eindhoven,
The Netherlands, V.Savcenco@tue.nl, R.M.M.Mattheij@tue.nl

Summary. This paper contains an overview of a self-adjusting multirate method. A simple extension which allows the improvement of the efficiency of the method is introduced. The performance of the extended and the original method is compared for a test problem.

1 Introduction

For the numerical solution of systems of ODEs there are many methods available. These methods use time steps that are varying in time, but are constant over the components. However, there are many problems of practical interest, where the temporal variations have different time scales for different sets of the components. For example, cellular phones consist of coupled digital and analogue sub-circuits, which operate in nano- and micro-seconds, respectively. The motion of the particles around a star, which attracts mass from a secondary star, in astrophysics is described by a large system of ordinary differential equations. In this system the components, that correspond to the particles near the center, are much faster than those corresponding to the distant ones. To exploit these local time scale variations, one needs *multirate methods* that use different, local time steps over the components. In these methods big time steps are used for the slow components and small time steps are used for the fast ones.

Various multirate methods were developed for solving systems with different time scales. The first descriptions of multirate schemes were given by Gear and Wells [4] for multistep methods. Sand and Skelboe [9] studied the stability of backward Euler multirate methods. Multirate methods for non-stiff problems have been examined by Engstler and Lubich [3]. A multirate scheme based on the partitioned Runge–Kutta methods was introduced by Günther et al. [5]. In [1,13,14] multirate methods have been applied to the modeling of electrical networks. Multirate methods for hyperbolic conservation laws were studied by Constantinescu and Sandu [2].

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*,
Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_50,

A multirate method based on the Rosenbrock methods, together with a self-adjusting partitioning strategy was introduced and analyzed in [7, 10–12]. In this paper we present an overview of this method and suggest a way to improve it. The comparison of the numerical results obtained with the original and extended strategy is presented.

The paper is organized as follows. In Sect. 2 we will introduce the Rosenbrock methods which will be used as our basic numerical integration methods and describe the multirate time stepping strategy. The performance of the extended version of the multirate strategy for a test problem is discussed in Sect. 3. Finally, Sect. 4 contains the conclusions.

2 A Multirate Time Stepping Strategy

We will consider multirate methods for solving systems of ODEs

$$w'(t) = F(t, w(t)), \qquad w(0) = w_0, \tag{1}$$

with given initial solution $w_0 \in \mathbb{R}^m$. The approximations at the global time levels t_n will be denoted by w_n.

Our multirate time stepping strategy is based on local temporal error estimation. For a given time step $\Delta t_n = t_n - t_{n-1}$, we compute a first, tentative approximation at the new time level for all components. For those components for which the error estimator indicates that the local temporal error is larger than a given tolerance Tol, the computation is redone with smaller steps. The refinement is recursively continued until the error estimator is below Tol for all components. Schematically, with components horizontally and time vertically, the multirate time stepping is displayed in Fig. 1.

In the original strategy [12], the refinement is performed by recalculating the required components with halved steps. For many problems, the time steps needed for the active components are much smaller than those needed for the slow ones. In such cases it is more efficient to immediately recompute the active components with more than two smaller steps instead of doing several halving recursive refinements. Therefore in this paper we will extend the original strategy and will assume that the number of smaller time steps at the refinement stage can be also larger than two. Using ideas from [12], it is possible to design an adaptive procedure of choosing the size of the time slabs for the extended strategy.

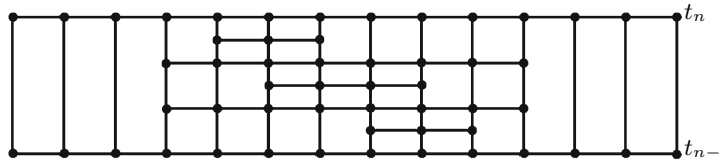


Fig. 1. Multirate time stepping for a time slab $[t_{n-1}, t_n]$

2.1 Main Time Integration Methods

In this paper we will use the Rosenbrock methods [6] as our basic numerical integration methods. To proceed from t_{n-1} to $t_n = t_{n-1} + \tau$, an s-stage Rosenbrock method calculates

$$w_n = w_{n-1} + \sum_{i=1}^{s} b_i k_i \,, \tag{2}$$

$$k_i = \tau F\left(t_{n-1} + \alpha_i \tau, w_{n-1} + \sum_{j=1}^{i-1} \alpha_{ij} k_j\right) + \tau \frac{\partial F}{\partial w}(t_{n-1}, w_{n-1}) \sum_{j=1}^{i} \gamma_{ij} k_j$$
$$+\gamma_i \tau^2 \frac{\partial F}{\partial t}(t_{n-1}, w_{n-1}), \qquad i = 1, \ldots, s \,, \tag{3}$$

where $\alpha_{ij}, \gamma_{ij}, \alpha_i, \gamma_i, b_i$ are real parameters defining the method and τ denotes the step size. For the local error estimation within the variable step size control we use the embedded formula

$$\overline{w}_n = w_{n-1} + \sum_{i=1}^{s} \overline{b}_i k_i \,, \tag{4}$$

which uses the same k_i-values as (2), but has different weights.

2.2 Interface Treatment

During the refinement stage, values at the intermediate time levels of components which are not refined might be needed. These values can be obtained by use of dense output built in the time integration method

$$w_I(t_{n-1} + \theta\tau) = w_{n-1} + \sum_{i=1}^{s} b_i(\theta) k_i, \qquad 0 \le \theta \le 1 \,. \tag{5}$$

Proper interface treatment is very important for multirate schemes. Use of dense output of order lower than the order of the main time integration method can lead to order reduction.

It is well known that use of Rosenbrock methods for problems with stiff source terms can lead to order reduction. During the refinement step, sub problems with stiff source terms have to be solved. An easy applicable technique, to avoid the order reduction, was proposed in [11]. Assuming that $g(t)$ is a component of $F(t, w(t))$, this technique suggests that the source terms $g(t_{n-1} + \alpha_i \tau) + \gamma_i \tau g'(t_{n-1})$ in a Rosenbrock method (3) of order p shall be replaced by $g_{n-1,i}$ with $\boldsymbol{g}_{n-1} = [g_{n-1,i}]$ chosen as

$$\boldsymbol{g}_{n-1} = \sum_{k=0}^{p} \boldsymbol{B}^k \boldsymbol{e} \tau^k g^{(k)}(t_{n-1}) \,, \tag{6}$$

where $\boldsymbol{B} = [\alpha_{ij} + \gamma_{ij}] \in \mathbb{R}^{s \times s}$ and $\boldsymbol{e} = [1] \in \mathbb{R}^s$.

3 Numerical Test

In this section we will present numerical results for a test problem. We consider the behavior of both strategies: original [12] and extended (where refinement with more than two steps is possible). The results are compared to the single-rate approach, also using the same basic time integration method. As a measure for the amount of work we consider the total number of components at which solutions are computed over the complete time integration interval, multiplied by the number of stages of the method. The fact that with the multirate approach some solution components are computed several times at certain time levels is taken into account. As the basic time integration method, for solving this problem, we use the two-stage second-order Rosenbrock ROS2 method [8].

3.1 An Inverter Chain Problem

An inverter is an electrical sub-circuit which transforms a logical input signal to its negation. The inverter chain is a concatenation of several inverters, where the output of an inverter serves as input for the succeeding one. An inverter chain with an even number of inverters will delay a given input signal and will also provide some smoothing of the signal.

The model for m inverters consists of the equations

$$\begin{cases} w_1'(t) = U_{\text{op}} - w_1(t) - \Upsilon g\big(u_{\text{in}}(t), w_1(t)\big)\,, \\ w_j'(t) = U_{\text{op}} - w_j(t) - \Upsilon g\big(w_{j-1}(t), w_j(t)\big)\,, \qquad j = 2, \ldots, m\,, \end{cases} \tag{7}$$

where

$$g(u, v) = \big(\max(u - U_{\text{thres}}, 0)\big)^2 - \big(\max(u - v - U_{\text{thres}}, 0)\big)^2. \tag{8}$$

The coefficient Υ serves as stiffness parameter. We solve the problem for a chain of $m = 500$ inverters with $\Upsilon = 100$, $U_{\text{thres}} = 1$ and $U_{\text{op}} = 5$. The initial condition is

$$w_j(0) = 6.247 \cdot 10^{-3} \text{ for } j \text{ even}, \quad w_j(0) = 5 \text{ for } j \text{ odd}. \tag{9}$$

The input signal is given by

$$u_{\text{in}}(t) = \begin{cases} t - 5 & \text{for} \quad 5 \le t \le 10\,, \\ 5 & \text{for} \quad 10 \le t \le 15\,, \\ \frac{5}{2}(17 - t) & \text{for} \quad 15 \le t \le 17\,, \\ 0 & \text{otherwise.} \end{cases} \tag{10}$$

An illustration of the solution for some of the even components is given in Fig. 2.

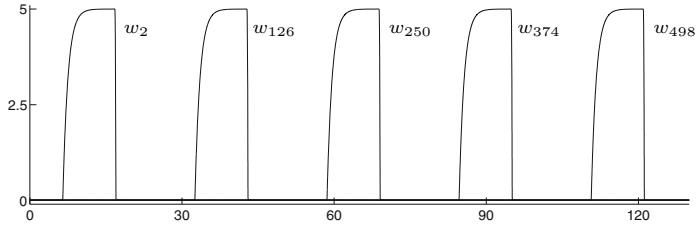


Fig. 2. Solution components $w_j(t)$, $j = 2, 126, 250, 374, 498$, for problem (7)–(10)

Table 1. Errors and work amount for problem (7)–(9)

	Single-rate		Multirate (extended)		Speedup	
Tol	Error	Work	Error	Work	Original	Extended
$5 \cdot 10^{-4}$	$1.44 \cdot 10^{-1}$	45649872	$1.47 \cdot 10^{-2}$	2846068	8.7	16.0
10^{-4}	$3.94 \cdot 10^{-2}$	94524592	$7.16 \cdot 10^{-3}$	5512400	13.0	17.1
$5 \cdot 10^{-5}$	$1.37 \cdot 10^{-2}$	131413560	$3.24 \cdot 10^{-3}$	6980676	13.5	18.8
10^{-5}	$2.04 \cdot 10^{-3}$	287207252	$9.22 \cdot 10^{-4}$	14332486	11.1	20.0

In Table 1 the errors at output time $T = 130$ (measured in the maximum norm with respect to an accurate reference solution) together with the amount of work are presented for several tolerances for the single-rate method and the extended multirate strategy. The speedup for both original and extended strategies is calculated.

It is seen from the table that a substantial improvement in amount of work is obtained for this problem. For the single-rate scheme, the amount of work is almost 18 times larger than for the extended multirate scheme. Moreover, the error behavior of the multirate scheme is very good. We can also see that for this problem we get a considerably larger speedup for the extended strategy compared to the original strategy.

4 Conclusions

In this paper we made on overview and extended the multirate time stepping strategy introduced in [7, 10–12].

As seen from the numerical tests, the efficiency of time integration methods can be significantly improved by using large time steps for inactive components, without sacrificing accuracy. Comparing the results obtained for the original and the extended strategies, we do have preference for the extended approach, where the values of the active components can be recalculated by the use of more than two smaller time steps.

References

1. Bartel, A., Günther, M.: A multirate W-method for electrical networks in state space formulation. J. Comp. Appl. Math. **147**, 411–425 (2002)
2. Constantinescu, E.M., Sandu, A.: Multirate time stepping methods for hyperbolic conservation laws. J. Sci. Comput. **33**, 239–278 (2007)
3. Engstler, C., Lubich, C.: Multirate extrapolation methods for differential equations with different time scales. Computing **58**, 173–185 (1997)
4. Gear, C., Wells, D.: Multirate linear multistep methods. BIT **24**, 484–502 (1984)
5. M. Günther, A. Kværnø, Rentrop, P.: Multirate partitioned Runge-Kutta methods. BIT **41**, 504–514 (2001)
6. Hairer, E., Wanner, G., Solving Ordinary Differential Equations II – Stiff and Differential-Algebraic Problems. Springer, Berlin (1996)
7. Hundsdorfer, W., Savcenco, V.: Analysis of a multirate *theta*-method for stiff ODEs. Appl. Num. Math. **59**, 693–706 (2009)
8. Hundsdorfer, W., Verwer, J.G.: Numerical Solution of Time-Dependent Advection-Diffusion-Reaction Equations. Springer, Berlin (2003)
9. Sand, J., Skelboe, S.: Stability of backward Euler multirate methods and convergence of waveform relaxation. BIT **32**, 350–366 (1992)
10. Savcenco, V.: Comparison of the asymptotic stability properties for two multirate strategies. J. Comp. Appl. Math. **220**, 508–524 (2008)
11. Savcenco, V.: Construction of a multirate RODAS method for stiff ODEs. J. Comp. Appl. Math. **225**, 323–337 (2009)
12. Savcenco, V., Hundsdorfer, W., Verwer, J.G.: A multirate time stepping strategy for stiff ordinary differential equations. BIT **47**, 137–155 (2007)
13. Striebel, M., Günther, M.: A charge oriented mixed multirate method for a special class of index-1 network equations in chip design. Appl. Numer. Math. **53**, 489–507 (2005)
14. Verhoeven, A., Guennouni, A.El., ter Maten, E.J.W., Mattheij, R.M.M.: A general compound multirate method for circuit simulation problems. In: Anile, A.M., Ali, G., Mascali, G. (eds.) Scientific Computing in Electrical Engineering, 143–150. Springer, Berlin (2006)

Terminal Current Interpolation for Multirate Time Integration of Hierarchical IC Models

A. Verhoeven[1], E.J.W. ter Maten[1,2], J.J. Dohmen[2], B. Tasić[2], and R.M.M. Mattheij[1]

[1] Eindhoven University of Technology (CASA), Den Dolech 2, 5612 AZ Eindhoven, The Netherlands, Arie.Verhoeven@na-net.ornl.gov, www.casa.tue.nl
[2] NXP Semiconductors (Design Methods), Eindhoven, The Netherlands www.nxp.com

1 Introduction

Multirate time-integration methods [3–5] appear to be attractive for initial value problems for DAEs with latency or multirate behaviour. Latency means that parts of the circuit are constant or slowly time-varying during a certain time interval, while multirate behaviour means that some variables are slowly time-varying compared to other variables. In both cases, it would be attractive to integrate these slow parts with a larger timestep than the other parts. This saves the computational workload while the accuracy is preserved. A nice property of multirate is that it does not use any linear structure, in contrast to MOR, but only a relaxation concept. If the coupling is sufficiently monitored and the partitioning is well chosen, multirate can be very efficient.

In this paper we will show how multirate time integration can be applied to hierarchical circuit models. Besides the classical interpolation variants, also some new implicit variants are discussed.

2 Hierarchical Circuit Models

Integrated Circuits can be modeled by a hierarchical system of differential-algebraic equations [1, 2]:

$$\frac{d}{dt}[\mathbf{q}(t,\mathbf{x})] + \mathbf{j}(t,\mathbf{x}) = \sum_{i=1}^{N} \mathbf{B}_{(i)}^{T} \left[\frac{d}{dt}[\mathbf{q}^{(i)}(t,\mathbf{x}^{(i)})] + \mathbf{j}^{(i)}(t,\mathbf{x}^{(i)}) \right] = \mathbf{0}. \quad (1)$$

Clearly this circuit model with global state vector $\mathbf{x} \in \mathbb{R}^d$ consists of N coupled subcircuit models. Each local state vector $\mathbf{x}^{(i)} \in \mathbb{R}^{d_i}$ (voltages,currents) consists of a terminal ($\hat{\mathbf{x}}^{(i)}$) and an internal ($\check{\mathbf{x}}^{(i)}$) part:

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*, Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_51,

$$\mathbf{x}^{(i)} = \mathbf{B}_{(i)}\mathbf{x} = \begin{bmatrix} \hat{\mathbf{x}}^{(i)} \\ \check{\mathbf{x}}^{(i)} \end{bmatrix}.$$

The matrices $\mathbf{B}_{(i)} \in \{0,1\}^{d_i \times d}$, defined by

$$\mathbf{B}_{(i)} = \begin{bmatrix} \hat{\mathbf{B}}_{(i)} & \\ & \check{\mathbf{B}}_{(i)} \end{bmatrix}, \tag{2}$$

are used to select the proper parts $\hat{\mathbf{x}}^{(i)} = \begin{bmatrix} \hat{\mathbf{B}}_{(i)} & \end{bmatrix}\mathbf{x}$ and $\check{\mathbf{x}}^{(i)} = \begin{bmatrix} & \check{\mathbf{B}}_{(i)} \end{bmatrix}\mathbf{x}$ of $\mathbf{x}^{(i)}$ from $\mathbf{x}$. This even allows for a hierarchical structure.

Similarly, also the functions $\mathbf{q}^{(i)}$ (charges, fluxes) and $\mathbf{j}^{(i)}$ (currents, voltages) have a similar structure: $\mathbf{q}^{(i)} = \mathbf{B}_{(i)}\mathbf{q} = \begin{bmatrix} \hat{\mathbf{q}}^{(i)} \\ \check{\mathbf{q}}^{(i)} \end{bmatrix}$, $\mathbf{j}^{(i)} = \mathbf{B}_{(i)}\mathbf{j} = \begin{bmatrix} \hat{\mathbf{j}}^{(i)} \\ \check{\mathbf{j}}^{(i)} \end{bmatrix}$. We can rewrite (1) in a part consisting of the collected equations for the terminal unknowns (3) and a part consisting of the remaining equations for the internal unknowns (4):

$$\sum_{i=1}^{N} \hat{\mathbf{B}}_{(i)}^{T} \left[\frac{d}{dt}[\hat{\mathbf{q}}^{(i)}(t, \mathbf{x}^{(i)})] + \hat{\mathbf{j}}^{(i)}(t, \mathbf{x}^{(i)}) \right] = \mathbf{0}, \tag{3}$$

$$\frac{d}{dt}[\check{\mathbf{q}}^{(i)}(t, \mathbf{x}^{(i)})] + \check{\mathbf{j}}^{(i)}(t, \mathbf{x}^{(i)}) = \mathbf{0}, \quad i = 1, \ldots, N. \tag{4}$$

Each subcircuit model can again be further decomposed in this manner.

3 Multirate Transient Analysis

For single-rate time integration all equations are discretised simultaneously by the same time step. If the time constants per subcircuits are quite different, it is attractive to perform multirate time integration. Then the fast subcircuits can be integrated on a local, fine, time-grid. Especially when the fast subcircuits are small in size, the additional costs for synchronisation and partitioning can be overcome and the overall multirate procedure becomes much more efficient than the single-rate time integration. An attractive multirate method is the Compound-Fast version [5,6,8], which first integrates the whole system at the new coarse time gridpoint and after that re-integrates only the active part at the fine time-grid. We will denote the coarse and fine time gridpoints by $\{T_n, 0 \leq n \leq N\}$ and $\{t_{n-1,m}, 1 \leq n \leq N, 0 \leq m \leq q_n\}$ with macro-steps $H_n := T_n - T_{n-1}$, and micro-steps $h_{n,m} := t_{n,m} - t_{n,m-1}$ and multirate factors q_n such that $t_{n-1,0} = T_{n-1}, t_{n-1,q_n} = T_n$. For a partitioning in a latent (slow) and an active (fast) part, $\mathbf{x}^{(L)} \in \mathbb{R}^{d_L}$ and $\mathbf{x}^{(A)} \in \mathbb{R}^{d_A}$, we typically get:

$$\hat{\mathbf{B}}_{(L)}^T \left[\frac{d}{dt}[\hat{\mathbf{q}}^{(L)}(t,\mathbf{x}^{(L)})] + \hat{\mathbf{j}}^{(L)}(t,\mathbf{x}^{(L)}) \right] + \hat{\mathbf{B}}_{(A)}^T \left[\frac{d}{dt}[\hat{\mathbf{q}}^{(A)}(t,\mathbf{x}^{(A)})] + \hat{\mathbf{j}}^{(A)}(t,\mathbf{x}^{(A)}) \right] = \mathbf{0},$$

$$\frac{d}{dt}[\check{\mathbf{q}}^{(L)}(t,\mathbf{x}^{(L)})] + \check{\mathbf{j}}^{(L)}(t,\mathbf{x}^{(L)}) = \mathbf{0}, \qquad \frac{d}{dt}[\check{\mathbf{q}}^{(A)}(t,\mathbf{x}^{(A)})] + \check{\mathbf{j}}^{(A)}(t,\mathbf{x}^{(A)}) = \mathbf{0}.$$

Voltage Interpolation A first approach is to integrate only the internal part of the active subcircuit at the fine time-grid. Then we get the following active circuit model for $\check{\mathbf{x}}^{(A)}$

$$\frac{d}{dt}[\check{\mathbf{q}}^{(A)}(t,\mathbf{x}^{(A)})] + \check{\mathbf{j}}^{(A)}(t,\mathbf{x}^{(A)}) = \mathbf{0}, \quad \mathbf{x}^{(A)} = \begin{bmatrix} \hat{\mathbf{x}}^{(A)} \\ \check{\mathbf{x}}^{(A)} \end{bmatrix}. \tag{5}$$

In practise $\hat{\mathbf{x}}^{(A)}$ will also behave latently, and in this case it is preferable to use voltage interpolation of the terminal voltages $\hat{\mathbf{x}}^{(A)}$. From the hierarchical linear solver in Pstar (the in-house analogue circuit simulator provided by NXP Semiconductors) [1] we know that (5) is solvable for $\check{\mathbf{x}}^{(A)}$. However, stability is now not automatically preserved from the original model. Furthermore the DAE-index can be larger than one, which typically leads to sawtooth-like shapes of $\check{\mathbf{x}}^{(A)}$.

Current Interpolation A second approach is to integrate the complete active subcircuit (i.e. using $\mathbf{q}^{(A)}$ rather than $\check{\mathbf{q}}^{(A)}$, etc) at the fine time-grid. Then we get the following active circuit model for $\mathbf{x}^{(A)}$

$$\frac{d}{dt}[\mathbf{q}^{(A)}(t,\mathbf{x}^{(A)})] + \mathbf{j}^{(A)}(t,\mathbf{x}^{(A)}) = -\hat{\mathbf{B}}_{(A)}\hat{\mathbf{B}}_{(L)}^T \left[\frac{d}{dt}[\hat{\mathbf{q}}^{(L)}(t,\mathbf{x}^{(L)})] + \hat{\mathbf{j}}^{(L)}(t,\mathbf{x}^{(L)}) \right]. \tag{6}$$

This leads to a more stable situation including the conservation of Kirchhoff's Current Law at the terminals and preservation of the DAE-index. In this case it is preferable to interpolate the terminal currents

$$\mathbf{i}_{L\to A} = -\hat{\mathbf{B}}_{(A)}\hat{\mathbf{B}}_{(L)}^T \left[\frac{d}{dt}[\hat{\mathbf{q}}^{(L)}(t,\mathbf{x}^{(L)})] + \hat{\mathbf{j}}^{(L)}(t,\mathbf{x}^{(L)}) \right]. \tag{7}$$

This can be done by adding $\mathbf{i}_{L\to A}$ as unknown or by calculating it explicitly. Direct interpolation of the slow voltages $\mathbf{x}^{(L)}$ is not attractive because often $d_L \gg d_A$. In Pstar for each subcircuit the corresponding terminal current $\frac{d}{dt}[\hat{\mathbf{q}}^{(i)}(t,\mathbf{x}^{(i)})] + \hat{\mathbf{j}}^{(i)}(t,\mathbf{x}^{(i)})$ is stored. Then the vector $\mathbf{i}_{L\to A}$ can be constructed for each multirate-partitioning.

4 Implicit Interpolation

Interpolation of the currents $\mathbf{i}_{L\to A}$ causes solvability problems for the active part if the active subcircuits are not grounded (so one may have to ground the most latent coupled terminal unknown). Stability and the differential index are only preserved if all subcircuits are stable DAEs of index one. In general

this property can not be assumed for a circuit simulator. An alternative could be a modified BDF multirate algorithm with implicit interpolation.

In (7) we already introduced the terminal current $\mathbf{i}_{L\to A}$ from latent-to-active. We also introduce the terminal current $\mathbf{i}_{A\to L}$ from active-to-latent. Then one can also write the hierarchical circuit model of (6) like

$$\begin{cases} \frac{d}{dt}[\mathbf{q}^{(L)}(t,\mathbf{x}^{(L)})] + \mathbf{j}^{(L)}(t,\mathbf{x}^{(L)}) = \mathbf{i}_{A\to L}, & i \\ \mathbf{i}_{A\to L} = -\hat{\mathbf{B}}_{(L)}\hat{\mathbf{B}}_{(A)}^{T}\left[\frac{d}{dt}[\hat{\mathbf{q}}^{(A)}(t,\mathbf{x}^{(A)})] + \hat{\mathbf{j}}^{(A)}(t,\mathbf{x}^{(A)})\right], & ii \\ \mathbf{i}_{L\to A} = -\hat{\mathbf{B}}_{(A)}\hat{\mathbf{B}}_{(L)}^{T}\left[\frac{d}{dt}[\hat{\mathbf{q}}^{(L)}(t,\mathbf{x}^{(L)})] + \hat{\mathbf{j}}^{(L)}(t,\mathbf{x}^{(L)})\right], & iii \\ \frac{d}{dt}[\mathbf{q}^{(A)}(t,\mathbf{x}^{(A)})] + \mathbf{j}^{(A)}(t,\mathbf{x}^{(A)}) = \mathbf{i}_{L\to A}. & iv \end{cases} \tag{8}$$

Here $\mathbf{i}_{A\to L}$ and $\mathbf{i}_{L\to A}$ are the terminal currents that couple both subcircuits. If the vector $\mathbf{i}_{L\to A}$ is given it is possible to perform the refinement for $\mathbf{x}^{(A)}$. For the Slow-Fast multirate method $\mathbf{i}_{L\to A}$ is approximated at the coarse time-grid, based on $\mathbf{x}^{(L)}$. However, it is also possible to approximate $\mathbf{i}_{L\to A}$ by a different approach. Note that $\mathbf{i}_{L\to A}$ and $\mathbf{i}_{A\to L}$ are related by the Kirchhoff's Current Law

$$\hat{\mathbf{B}}_{(L)}^{T}\mathbf{i}_{A\to L} + \hat{\mathbf{B}}_{(A)}^{T}\mathbf{i}_{L\to A} = \mathbf{0}.$$

Variant I Let us discretise (8i) by Euler Backward with step $H_{n,m} = t_{n-1,m} - t_{n-1,0}$, where $t_{n-1,0} = T_{n-1}$

$$\begin{aligned} \mathbf{q}^{(L)}(t_{n-1,m},\mathbf{x}_{n-1,m}^{(L)}) - \mathbf{q}^{(L)}(t_{n-1,0},\mathbf{x}_{n-1,0}^{(L)}) + H_{n,m}\,\mathbf{j}^{(L)}(t_{n-1,m},\mathbf{x}_{n-1,m}^{(L)}) \\ = H_{n,m}\mathbf{i}_{A\to L}(t_{n-1,m}). \end{aligned} \tag{9}$$

We assume that just one Newton step is needed to correct the prediction $\mathbf{y}_{n-1,m}^{(L)}$ of $\mathbf{x}_{n-1,m}^{(L)}$, which is acceptable if $\mathbf{x}^{(L)}$ behaves latently. Thus

$$\mathbf{J}_{n-1,m}^{(L)}\big(\mathbf{x}_{n-1,m}^{(L)} - \mathbf{y}_{n-1,m}^{(L)}\big) = H_{n,m}\mathbf{i}_{A\to L}(t_{n-1,m}) - \mathbf{f}_{n-1,m}^{(L)}, \text{ where}$$

$\mathbf{J}_{n-1,m}^{(L)} = \mathbf{C}^{(L)}(t_{n-1,m},\mathbf{y}_{n-1,m}^{(L)}) - \mathbf{C}^{(L)}(t_{n-1,0},\mathbf{x}_{n-1,0}^{(L)}) + H_{n,m}\mathbf{G}^{(L)}(t_{n-1,m}, \mathbf{y}_{n-1,m}^{(L)})$, $\mathbf{f}_{n-1,m} = \mathbf{q}^{(L)}(t_{n-1,m},\mathbf{y}_{n-1,m}^{(L)}) - \mathbf{q}^{(L)}(t_{n-1,0},\mathbf{x}_{n-1,0}^{(L)}) + H_{n,m}\mathbf{j}^{(L)}(t_{n-1,m},\mathbf{y}_{n-1,m}^{(L)})$, $\mathbf{C}^{(L)} = \frac{\partial \mathbf{q}^{(L)}}{\partial \mathbf{X}^{(L)}}$, and $\mathbf{G}^{(L)} = \frac{\partial \mathbf{j}^{(L)}}{\partial \mathbf{X}^{(L)}}$. The matrix $\mathbf{J}_{n-1,m}^{(L)}$ is invertible if the latent part (8i) is solvable, which is a reasonable assumption. Hence

$$\mathbf{x}_{n-1,m}^{(L)} = \mathbf{y}_{n-1,m}^{(L)} + \mathbf{J}_{n-1,m}^{-1}\big(H_{n,m}\mathbf{i}_{A\to L}(t_{n-1,m}) - \mathbf{f}_{n-1,m}\big).$$

We do not want to compute $\mathbf{J}_{n-1,m}^{-1}\mathbf{f}_{n-1,m}$ for all m, so we use linear interpolation of $\mathbf{J}_{n-1,0}^{-1}\mathbf{f}_{n-1,0} = \mathbf{J}_{n-1,0}^{-1}\mathbf{f}(\mathbf{x}_{n-1,0}^{(L)})$ and $\mathbf{J}_{n,0}^{-1}\mathbf{f}_{n,0} = \mathbf{J}_{n,0}^{-1}\mathbf{f}(\mathbf{x}_{n,0}^{(L)})$. Thus we obtain

$$\mathbf{J}_{n-1,m}^{-1}\mathbf{f}_{n-1,m} := \lambda_m\mathbf{J}_{n-1,0}^{-1}\mathbf{f}_{n-1,0} + \mu_m\mathbf{J}_{n,0}^{-1}\mathbf{f}_{n,0}.$$

Here $\lambda_m = 1 - \frac{m}{q}$ and $\mu_m = 1 - \lambda_m = \frac{m}{q}$. For $m = 0$ we have that $\mathbf{x}^{(L)}_{n-1,m} = \mathbf{y}^{(L)}_{n-1,0}$ solves (9), hence we have that $\mathbf{f}_{n-1,0} = \mathbf{0}$. After interpolating the operator $\mathbf{J}^{-1}$ applied to $H_{n,m}\mathbf{i}_{A\to L}(t_{n-1,m})$, we can approximate $\mathbf{x}^{(L)}_{n-1,m}$ by:

$$\begin{aligned}\mathbf{x}^{(L)}_{n-1,m} &\approx \mathbf{y}^{(L)}_{n-1,m} + (\lambda_m \mathbf{J}^{-1}_{n-1,0} + \mu_m \mathbf{J}^{-1}_{n,0}) H_{n,m}\mathbf{i}_{A\to L}(t_{n-1,m}) - \mu_m \mathbf{J}^{-1}_{n,0}\mathbf{f}_{n,0} \\ &\approx \mathbf{a}(t_{n-1,m}) + \mathbf{A}(t_{n-1,m})\mathbf{i}_{A\to L}(t_{n-1,m}),\end{aligned} \tag{10}$$

where $\mathbf{a}(t_{n-1,m}) = \mathbf{y}^{(L)}_{n-1,m} - \mu_m \mathbf{J}^{-1}_{n,0}\mathbf{f}_{n,0}$ and $\mathbf{A}(t_{n-1,m}) = (\lambda_m \mathbf{J}^{-1}_{n-1,0} + \mu_m \mathbf{J}^{-1}_{n,0}) H_{n,m}$. Hence, one can write:

$$\begin{cases} \mathbf{i}_{A\to L} = -\hat{\mathbf{B}}_{(L)}\hat{\mathbf{B}}^T_{(A)}\left[\frac{d}{dt}[\hat{\mathbf{q}}^{(A)}(t,\mathbf{x}^{(A)})] + \hat{\mathbf{j}}^{(A)}(t,\mathbf{x}^{(A)})\right], \\ \mathbf{i}_{L\to A} = -\hat{\mathbf{B}}_{(A)}\hat{\mathbf{B}}^T_{(L)}\left[\frac{d}{dt}[\hat{\mathbf{q}}^{(L)}(t,\mathbf{a}(t)+\mathbf{A}(t)\mathbf{i}_{A\to L})] + \hat{\mathbf{j}}^{(L)}(t,\mathbf{a}(t)+\mathbf{A}(t)\mathbf{i}_{A\to L})\right], \\ \frac{d}{dt}[\mathbf{q}^{(A)}(t,\mathbf{x}^{(A)})] + \mathbf{j}^{(A)}(t,\mathbf{x}^{(A)}) = \mathbf{i}_{L\to A}. \end{cases} \tag{11}$$

Next we can compute $\mathbf{x}^{(L)}$ by evaluating formula (10). Thus we get a multirate method of Fastest First type instead of Slowest First type. In contrast to the Compound-Fast multirate method we do not need a compound step now to predict $\mathbf{i}_{L\to A}$.

Variant II In a similar way as for $\mathbf{x}^{(L)}$, $\mathbf{i}_{L\to A}$ satisfies

$$\begin{aligned}\mathbf{i}_{L\to A}(t_{n-1,m}) &= -\hat{\mathbf{B}}_{(A)}\hat{\mathbf{B}}^T_{(L)}\left[\frac{d}{dt}[\hat{\mathbf{q}}^{(L)}(t_{n-1,m},\mathbf{x}^{(L)}_{n-1,m})] + \hat{\mathbf{j}}^{(L)}(t_{n-1,m},\mathbf{x}^{(L)}_{n-1,m})\right] \\ &= -\hat{\mathbf{P}}\left[\frac{d}{dt}[\mathbf{q}^{(L)}(t_{n-1,m},\mathbf{x}^{(L)}_{n-1,m})] + \mathbf{j}^{(L)}(t_{n-1,m},\mathbf{x}^{(L)}_{n-1,m})\right],\end{aligned}$$

where $\hat{\mathbf{P}} = \hat{\mathbf{B}}_{(A)}\hat{\mathbf{B}}^T_{(L)}\left[\hat{\mathbf{B}}_{(L)} \;\; \mathbf{O}\right]$. In a similar way as for $\mathbf{x}^{(L)}$ in (10) we can derive the following feedback law for $\mathbf{i}_{L\to A}$:

$$\mathbf{i}_{L\to A}(t) \approx \mathbf{b}(t) + \mathbf{B}(t)\mathbf{x}^{(L)}(t). \tag{12}$$

Expressing $\mathbf{x}^{(L)}(t)$ as in (10) we can derive $\mathbf{c}(t)$ and $\mathbf{C}(t)$ such that

$$\mathbf{i}_{L\to A}(t) \approx \mathbf{c}(t) + \mathbf{C}(t)\mathbf{i}_{A\to L}(t).$$

This reduces the system (8) even further to the following system for $\mathbf{i}_{A\to L}, \mathbf{x}^{(A)}$:

$$\begin{cases} \mathbf{i}_{A\to L} = -\hat{\mathbf{B}}_{(L)}\hat{\mathbf{B}}^T_{(A)}\left[\frac{d}{dt}[\hat{\mathbf{q}}^{(A)}(t,\mathbf{x}^{(A)})] + \hat{\mathbf{j}}^{(A)}(t,\mathbf{x}^{(A)})\right], \\ \frac{d}{dt}[\mathbf{q}^{(A)}(t,\mathbf{x}^{(A)})] + \mathbf{j}^{(A)}(t,\mathbf{x}^{(A)}) = \mathbf{c}(t) + \mathbf{C}(t)\mathbf{i}_{A\to L}. \end{cases} \tag{13}$$

We can eliminate $\mathbf{i}_{A\to L}$, which results in the following system for $\mathbf{x}^{(A)}$

$$\frac{d}{dt}[\mathbf{q}^{(A)}(t,\mathbf{x}^{(A)})] + \mathbf{j}^{(A)}(t,\mathbf{x}^{(A)}) = \mathbf{c}(t) - \mathbf{C}(t)\hat{\mathbf{B}}_{(L)}\hat{\mathbf{B}}^T_{(A)}\left[\frac{d}{dt}[\hat{\mathbf{q}}^{(A)}(t,\mathbf{x}^{(A)})] + \hat{\mathbf{j}}^{(A)}(t,\mathbf{x}^{(A)})\right]. \tag{14}$$

From its structure it can be seen that only the terminal active equations that are directly coupled to the latent part are modified. In fact they are multiplied by a linear transformation. This linear transformation is such that the dynamical behaviour of the original system has been preserved. Again, the vector-valued function $\mathbf{c}(t)$ is an interpolation-based current source.

Variant III In a similar way as for $\mathbf{i}_{L\to A}$ we can derive the formula

$$\mathbf{i}_{A\to L}(t) \approx \mathbf{f}(t) + \mathbf{F}(t)\mathbf{x}^{(A)}(t). \tag{15}$$

Combining all three formulae (10), (12) and (15) enables us to express $\mathbf{i}_{L\to A}$ directly in terms of $\mathbf{x}^{(A)}$:

$$\mathbf{i}_{L\to A}(t) = \mathbf{g}(t) + \mathbf{G}(t)\mathbf{x}^{(A)}(t).$$

Then we get the following system for $\mathbf{x}^{(A)}$:

$$\left\{ \frac{d}{dt}[\mathbf{q}^{(A)}(t, \mathbf{x}^{(A)})] + \mathbf{j}^{(A)}(t, \mathbf{x}^{(A)}) = \mathbf{g}(t) + \mathbf{G}(t)\mathbf{x}^{(A)}. \right. \tag{16}$$

5 Conclusions

We described a multirate method for hierarchical IC models. It is analysed and tested in [5, 7, 8]. For IC models with many slowly-varying unknowns it is possible to achieve a good speed-up while the accuracy is maintained. We also proposed a new implicit type of interpolation that can solve some typical problems with solvability and stability for the active part. Variant I needs to evaluate all terminal equations for the slow models and solves all terminal currents, which leads to a second order system and can be expensive. But it can also be applied for fast terminal currents $\mathbf{i}_{L\to A}, \mathbf{i}_{A\to L}$. Variant II only needs to evaluate active elements but it still needs $\mathbf{i}_{A\to L}$ as additional unknown. Therefore it still can be applied for active $\mathbf{i}_{A\to L}$. Variant III really reduces to a system for only the active part. It is only allowed if all terminal currents behave slowly. For the third variant it is clear that $\mathbf{i}_{L\to A}$ is replaced by a combination of current sources and resistors. In fact this is model reduction of the large latent part.

References

1. Fijnvandraat, J.G., Houben, S.H.M.J., ter Maten, E.J.W., Peters, J.M.F.: Time domain analog circuit simulation. J. Comput. Appl. Math. **185**(2), 441–459 (2006)
2. ter Maten, J., Verhoeven, A., El Guennouni, A., Beelen, Th.: Multirate hierarchical time integration for electronic circuits. In: PAMM, Virtual Annual Meeting Proceedings of the GAMM conference, pp. 819–820. Luxembourg (2005)
3. Savcenco, V.: Multirate Numerical Integration for Ordinary Differential Equations. PhD thesis, CWI, Amsterdam (2008)
4. Striebel, M.: Hierarchical Mixed Multirating for Distributed Integration of DAE Network Equations in Chip Design. PhD thesis, Bergische University of Wuppertal, Wuppertal, Germany (2006)

5. Verhoeven, A.: Redundancy Reduction of IC Models by Multirate Time Integration and Model Order Reduction. PhD thesis, Eindhoven University of Technology, Department of Mathematics and Computer Science, Eindhoven (2008)
6. Verhoeven, A., El Guennouni, A., ter Maten, E.J.W., Mattheij, R.M.M.: A general compound multirate method for circuit simulation problems. In: Anile, A.M., Ali, G., Mascali, G. (eds.) Scientific Computing in Electrical Engineering, pp. 143–150. Capo d'Orlando, Italy. Springer, Berlin (2006)
7. Verhoeven, A., ter Maten, E.J.W., Mattheij, R.M.M., Tasić, B.: Stability analysis of the BDF slowest first multirate methods. Int. J. Comput. Math. **84**, 895–923 (2007)
8. Verhoeven, A., Tasić, B., Beelen, T.G.J., ter Maten, E.J.W., Mattheij, R.M.M.: BDF Compound-Fast multirate transient analysis with adaptive stepsize control. J. Num. Anal. Ind. Appl. Math. **3**(3–4) (2008)

On Extrapolated Multirate Methods

Emil M. Constantinescu and Adrian Sandu

[1] Argonne National Laboratory, Mathematics and Computer Science Division, 60439, USA, emconsta@mcs.anl.gov
[2] Virginia Tech, Department of Computer Science, Blacksburg, VA 24061, USA asandu@cs.vt.edu

Summary. In this paper we construct extrapolated multirate discretization methods that allow to efficiently solve problems that have components with different dynamics. This approach is suited for the time integration of multiscale ordinary and partial differential equations and provides highly accurate discretizations. We analyze the linear stability properties of the extrapolated multirate explicit and linearly implicit methods. Numerical results with multiscale ODEs illustrate the theoretical findings.

1 Introduction

In this study we develop *multirate* time integration schemes using *extrapolation methods* for the efficient simulation of multiscale ODEs and PDEs via the method of lines. In multirate time integration, the time step can vary across the solution components and has to satisfy only the local stability conditions, resulting in substantially more efficient overall computations.

Previous work in multirate methods includes [1,9,14]. Engstler and Lubich [6] developed multirate schemes based on extrapolated forward Euler methods (MURX). The components with slow dynamics are inactivated at certain time levels, while the fast components are evaluated every time step. Our work extends this strategy to extrapolated compound multirate explicit and implicit steps. In this case the extrapolation procedure operates on multirate time stepping schemes.

In this paper we investigate the following initial value problem

$$\mathbf{y} = \begin{pmatrix} y'(t) \\ z'(t) \end{pmatrix} = \begin{pmatrix} f\left(x, y(x), z(x)\right) \\ g\left(x, y(x), z(x)\right) \end{pmatrix}; \quad \left[y(x_0)\ z(x_0)\right]^T = \left[y_0\ z_0\right]^T, \ x > x_0, \tag{1}$$

where $\mathbf{y}$ is the solution vector partitioned into two components that have their own particular time scales (y represents the slow component and z the fast one). These types of problems occur naturally in electric circuit simulations [1].

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*, Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_52,

We seek to apply time discretization methods with a different time step length for each dynamic characteristic to (1) and consider the extrapolation methods [4,10] with multirate explicit and implicit base schemes.

Extrapolation Methods

Consider a sequence n_i of positive integers with $n_i < n_{i+1}$, $1 \le i \le E$ and define corresponding step sizes $h_i = H/n_i$. Further, define the numerical approximation of (1) at $x_0 + H$ using step size h_i and a p^{th} order *base method*

$$T_{i,1} := \mathbf{y}_{h_i}(x_0 + H),\ 1 \le i \le E. \tag{2a}$$

By using E approximations to (2a) with different h_i's, one can eliminate the truncation error terms by using Richardson extrapolation. In general, high order approximations of (1) can be obtained by solving a linear system with E equations, with the k^{th} solution being a numerical approximation of order $p + k - 1$ [10, Chap. II, Theorem 9.1] using the Aitken–Neville formula [7]:

$$T_{j,k+1} = T_{j,k} + \frac{T_{j,k} - T_{j-1,k}}{(n_j/n_{j-1}) - 1},\ j = 1 \dots k. \tag{2b}$$

Scheme (2) is called the *extrapolation method*. The most economical choice for the sequence n_j is the harmonic sequence, $n_j = 1, 2, 3, \dots$ [3].

A popular base method is the linearly implicit Euler [4,5] which can be derived from the implicit Euler method applied to problem (1) under smoothness assumptions: $(I - hJ)(\mathbf{y}_{i+1} - \mathbf{y}_i) = hf(x_i, \mathbf{y}_i)$, where $J = \frac{\partial f}{\partial \mathbf{y}}(x_i, \mathbf{y}_i)$.

2 Multirate Base Methods

We propose the following multirate base methods for solving (1). The *multirate explicit Euler method* is given by

$$\begin{aligned} y_{n+1} &= y_n + h\, f(y_n, z_n) \\ z_{n+\frac{i}{m}} &= z_{n+\frac{i-1}{m}} + \frac{h}{m}\, g\left(Y_{n+\frac{i-1}{m}}, z_{n+\frac{i-1}{m}}\right),\ i = 1, \dots, m, \end{aligned} \tag{3a}$$

where m is a positive integer and $Y_{n+\frac{i}{m}}$ is an approximation of y at $x_{n+\frac{i}{m}}$. Forward Euler is first order accurate and hence the zeroth order interpolation can be used to approximate Y, the first order interpolation can also be considered:

$$Y_{n+\frac{i-1}{m}} = y_n,\ Y_{n+\frac{i-1}{m}} = y_{n+1},\ \text{or } Y_{n+\frac{i-1}{m}} = \frac{m-i+1}{m} y_n + \frac{i-1}{m} y_{n+1},\ i = 1, \dots, m.$$

Linearly implicit Euler method can also be considered as a candidate for the base methods used in the extrapolation procedure. The *multirate linearly implicit method* is given by

$$\begin{bmatrix} I - hf_y(0) & -hf_z(0) \\ -\dfrac{h}{m}g_y(0) & I - \dfrac{h}{m}g_z(0) \end{bmatrix} \cdot \begin{bmatrix} y_{n+1} - y_n \\ z_{n+\frac{1}{m}} - z_n \end{bmatrix} = \begin{bmatrix} h\, f\,(y_n, z_n) \\ \dfrac{h}{m}\, g\,(y_n, z_n) \end{bmatrix}, \tag{3b}$$

$$\left(I - \frac{h}{m}g_z(0)\right)\left(z_{n+\frac{i}{m}} - z_{n+\frac{i-1}{m}}\right) = \frac{h}{m}\, g\left(Y_{n+\frac{i-1}{m}}, z_{n+\frac{i-1}{m}}\right),\ i = 2, \ldots, m,$$

where the notation $f_{\{y,z\}}(0)$ and $g_{\{y,z\}}(0)$ denotes the derivatives evaluated at x_0, the initial extrapolation time in (2a).

Consistency of the Extrapolated Multirate Methods

In Henrici's notation [12], one step methods are expressed as $y^{n+1} = y^n + h\,\Phi\,(x^n, y^n, h)$. It is easy to see that methods (3) can be represented in this notation. It follows [8,10] that schemes (3) can be extrapolated using (2) (see [2], [10, Chap. II, Theorem 3.6]). Next we illustrate this theoretical aspect on a numerical example.

Numerical Consistency Investigation of the Extrapolated Multirate Methods

Consider the following initial value problem

$$\begin{pmatrix} y(x) \\ z(x) \end{pmatrix}' = \begin{pmatrix} \Gamma & \varepsilon \\ \varepsilon & -1 \end{pmatrix} \begin{pmatrix} \left(-1 + y^2 - \cos(x)\right)/(2y) \\ \left(-2 + z^2 - \cos(\omega x)\right)/(2z) \end{pmatrix} - \begin{pmatrix} \sin(x)/(2y) \\ \omega \sin(\omega x)/(2z) \end{pmatrix}, \tag{4}$$

with the exact solution $[y(x)\ z(x)]^T = \left[\sqrt{1 + \cos(x)}\ \sqrt{2 + \cos(\omega x)}\right]^T$ shown in Fig. 1. This problem was adapted from [1] and the scalar Prothero–Robinson [11]. We illustrate the theoretical findings by integrating (4) with $0 \le x \le H$, $\varepsilon = 0.5$, $\Gamma = -2.0$, $m = \omega = 20.0$ and schemes (3) with successively smaller steps H using the extrapolation procedure (2). The observed orders based on the numerical error both in L_1 and L_2 norms are presented in Table 1 and confirm the theoretical expectations.

3 Linear Stability Analysis of the Extrapolated Multirate Methods

Following the analysis done by [13], we investigate the extrapolated schemes with base methods (3) applied to the scaled system

$$\begin{pmatrix} y(x) \\ z(x) \end{pmatrix}' = \begin{pmatrix} -1 & \varepsilon \\ \omega & -m \end{pmatrix} \begin{pmatrix} y(x) \\ z(x) \end{pmatrix} = A \begin{pmatrix} y(x) \\ z(x) \end{pmatrix} = \begin{pmatrix} f\,(y(x), z(x)) \\ g\,(y(x), z(x)) \end{pmatrix}, \tag{5}$$

Table 1. The local discretization order of the extrapolation method (2a), (2b) with the multirate base methods

2								
2	3							
2	3	4						
2	3	4	5					
2	3	4	5	6				
2	3	4	5	6	7			
2	3	4	5	6	7	8		
2	3	4	5	6	7	8	9	
...	...	...	...	...	...	...	...	...

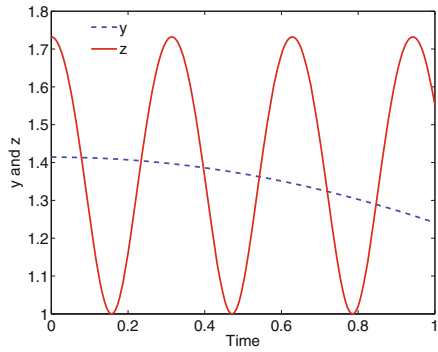


Fig. 1. The exact solution of the modified nonlinear Prothero–Robinson equation (4) with $\varepsilon = 0.5$, $\Gamma = -2.0$, $\omega = 20.0$

where we assume that m is a fixed integer that represents the scale difference between the slow and the fast components, and ε and ω represent coupling parameters. System (5) is stable if the real part of the eigenvalues of A is negative, which gives $\omega\varepsilon \leq m$.

The stability function $R(\dots hA_{ij} \dots)$ for a numerical discretization of (5) is defined by the quantity that verifies $\mathbf{y}_{n+1} = R(\dots hA_{ij} \dots)\mathbf{y}_n$. The method is stable if $\rho(R(\dots hA_{ij} \dots)) \leq 1$. The stability functions of extrapolated (3) can be easily calculated using (2b) as in [11, Chap. IV].

We take a practical approach and ask the following question: How does the stability region of a multirate method with ratio m applied to (5) compare to the stability region of the single-rate method with the time step length of the fastest component (i.e., H/m)? We note that the multirate method is more efficient in this case by taking fewer steps on the slow components.

Numerical Linear Stability Investigation of the Extrapolated Multirate Methods

We next investigate the linear stability properties of the multirate extrapolation method (2), (3) applied to problem (5) with fixed ratio $m = 2$.

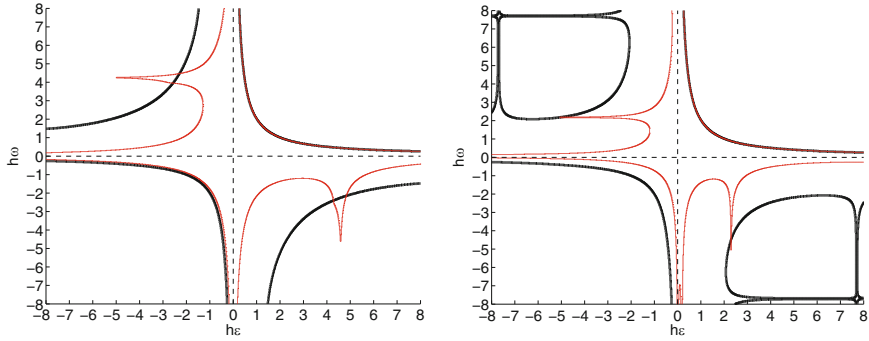


Fig. 2. The linear stability region for problem (5) using the explicit multirate ($m = 2$) method (3a) (*thin red line*) and the corresponding single-rate explicit method (*thick dark line*) for T_{32} and T_{44}

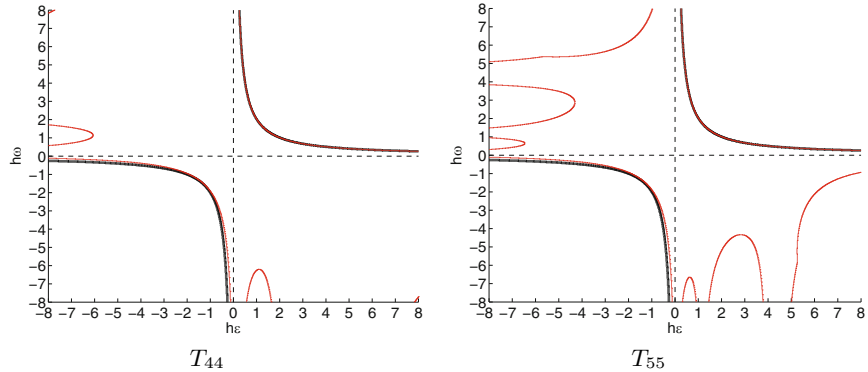


Fig. 3. The linear stability region for problem (5) using the implicit multirate ($m = 2$) method (3b) (*thin red line*) and the corresponding single-rate explicit method ($m = 1$) (*thick dark line*) for T_{44} and T_{55}

In Fig. 2 we show the stability regions in the $h\omega$-$h\varepsilon$ plane for the extrapolated multirate explicit method (3a) for the extrapolation terms in positions T_{32} and T_{44} (see Table 1). The stability region of the multirate method is slightly degraded; however, for practical purposes, it is acceptable.

In Fig. 3 we show the stability regions for the extrapolated multirate implicit method (3b) for the extrapolation terms in positions T_{44} and T_{55}. Experimentally, we determine that on the first column of the extrapolation tableau the multirate implicit methods preserve the "unconditional" stability of the implicit base (single-rate) method; i.e., the stability region extends to (∞, ∞) and $(-\infty, -\infty)$ in the $h\omega$-$h\varepsilon$ plane. However, when the multirate

solution is extrapolated, the stability region shrinks in quadrants II and IV (see Fig. 3). This aspect needs to be investigated further.

4 Concluding Remarks

In this manuscript we construct extrapolated multirate implicit and explicit discretization methods that allow to efficiently solve problems that have multiple scales. We propose two methods that are based on multirate forward and linearly implicit Euler schemes. The cost of implementing these methods is very small and can easily reach very high orders of accuracy.

The proposed multirate extrapolation methods represent a sequence of embedded methods which can be used for step size control and variable order approaches due to their trivial extension to higher orders. Extrapolation methods are less efficient than the popular Runge–Kutta or linear multistep schemes; however, they can be parallelized very easily [15]. Each entry on the first extrapolation tableau column ($T_{i,1}$) can be computed independently, and are well suited for multiprocessor/multicore architectures.

The extrapolated multirate forward Euler method shows only a slight degradation of the linear stability region. The multirate linearly implicit method performs very well for nonstiff problems or for stiff problems with relaxed component coupling. The linear stability region does not resemble the unconditional stability of the single-rate counterpart. This aspect needs to be investigated further.

References

1. Bartel, A., Günther, M.: A multirate W-method for electrical networks in state-space formulation. J. Comput. Appl. Math. **147**(2), 411–425 (2002). ISSN 0377-0427. doi: 10.1016/S0377-0427(02)00476-4
2. Constantinescu, E.M., Sandu, A.: On extrapolated multirate methods. Technical Report TR-08-12, Computer Science, Virginia Tech, (2008). URL http://eprints.cs.vt.edu.
3. Deuflhard, P.: Order and stepsize control in extrapolation methods. Numer. Math. **41**(3), 399–422 (1983) doi: 10.1007/BF01418332
4. Deuflhard, P.: Recent progress in extrapolation methods for ordinary differential equations. SIAM Rev. **27**(4), 505–535 (1985). ISSN 0036-1445
5. Deuflhard, P., Hairer, E., Zugck, J.: One-step and extrapolation methods for differential-algebraic systems. Numer. Math. **51**(5), 501–516 (1987). doi: 10.1007/BF01400352
6. Engstler, C., Lubich, C.: Multirate extrapolation methods for differential equations with different time scales. Computing **58**(2), 173–185 (1997). ISSN 0010-485X
7. Gasca, M., Sauer, T.: Polynomial interpolation in several variables. Adv. Comput. Math. **12**(4), 377–410 (2000). doi: 10.1023/A:1018981505752

8. Gragg, W.B.: On extrapolation algorithms for ordinary initial value problems. J. Soc. Ind. Appl. Math. Ser. B Numer. Anal. **2**(3), 384–403 (1965). doi: 10.1137/0702030
9. Günther, M., Kværnø, A., Rentrop, P.: Multirate partitioned Runge-Kutta methods. BIT **41**(3), 504–514 (2001)
10. Hairer, E., Norsett, S.P., Wanner, G.: Solving Ordinary Differential Equations I: Nonstiff Problems. Springer, Berlin (1993a)
11. Hairer, E., Norsett, S.P., Wanner, G.: Solving Ordinary Differential Equations II: Stiff and Differential-Algebraic Problems. Springer, Berlin (1993b)
12. Henrici, P.: Discrete Variable Methods in Ordinary Differential Equations. Wiley, New York (1962)
13. Kværnø, A.: Stability of multirate Runge-Kutta schemes. Int. J. Differ. Equ. Appl. **1**(1), 97–105 (2000)
14. Kværnø, A., Rentrop, P.: Low Order Multirate Runge-Kutta Methods in Electric Circuit Simulation. (1999)
15. Rauber, T. Rünger, G.: Load balancing schemes for extrapolation methods. Concurrency Pract. Exp. **9**(3), 181–202 (1997)

Is Geometry or Dynamics More Important in Cardiac Arrhythmogenesis?

Arun V. Holden, Stephen H. Gilbert, and Alan P. Benson

Institute of Membrane and Systems Biology & Multidisciplinary Cardiovascular Research Institute, University of Leeds, Leeds LS2 9JT, UK
a.v.holden@leeds.ac.uk, s.h.gilbert@leeds.ac.uk, a.p.benson@leeds.ac.uk

Summary. We examine the effects of cardiac geometry and architecture on the excitable media paradigm, and illustrate the effect of cardiac structure on the dynamics of arrhythmias by investigating scroll wave filament dynamics in two biophysically-detailed heterogeneous models of the human left ventricular free wall.

1 Introduction

During ventricular fibrillation (VF), rapid, self-sustained and spatio-temporally highly irregular electrical excitation waves in the ventricles results in loss of their normal synchronised rhythmic beating (Fig. 1). Both experimental [11] and computational [10] evidence supports the idea that VF is sustained by re-entrant wave propagation, in which a wave of excitation propagates through, away from, and back into, the same piece of tissue. Re-entrant waves have been mathematically idealised in extensive homogeneous isotropic excitable media by 2D spiral and 3D scroll waves [17]. Virtual cardiac tissues have proved to be an effective tool for simulating cardiac propagation patterns, and for proposing hypotheses that can be tested experimentally [6]. This excitable medium paradigm provides a simple explanation for the development of monomorphic ventricular tachycardia (VT) into VF: the normal sinus rhythm is a repetitive sequence of wavefronts propagating though the myocardium; a wavebreak leads to VT (analogous to a spiral or scroll wave), which breaks down into the spatio-temporal irregularity of VF. This oversimplified and seductive cartoon is illustrated in Fig. 1, where arrhythmogenesis is explained in terms of wave stability. This reaction diffusion framework has been remarkably successful in providing simple mathematical explanations for arrhythmic behaviours, e.g. meander of spiral waves in terms of Hopf bifurcations [1] and its control by resonant drift [5]. However, it fails to address some details and major problems of the clinical phenomenology. Clinicians talk of substrate for arrhythmias, not as the properties of the cardiac tissue within which arrhythmias occur, but as their heterogeneity.

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*,
Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_53,

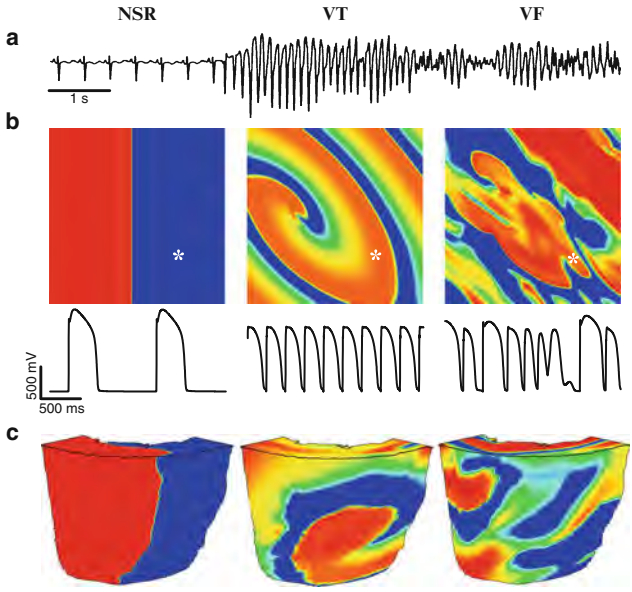


Fig. 1. (**a**) Electrocardiogram showing degeneration into cardiac arrhythmia – from normal sinus rhythm (NSR) to ventricular tachycardia (VT) then ventricular fibrillation (VF). (**b**) In two dimensions, propagation of a continuous wavefront (idealised as a plane wave) represents one excitation of NSR, wavebreak leading to a pair of re-entrant waves (idealised by a single spiral wave) underlies VT, while spatio-temporal irregularity underlies VF. Excited tissue is lighter, resting tissue is darker. Also shown are membrane potential recordings from the sites indicated by asterisks. (**c**) In three-dimensional tissue with orthotropic (fibre and sheet) structure extracted from a diffusion tensor imaged human heart, the qualitative dynamics of propagation underlying the behaviours are the same, but quantitative differences exist

Clinically, re-entrant arrhythmias are more likely to occur when there is an increase in spatial heterogeneity, in either the excitability (dynamics) or coupling (geometry/architecture) components underlying propagation. Heterogeneities in excitability can be mapped by molecular mapping techniques [21], and architectural heterogeneities by diffusion tensor magnetic resonance imaging (DT-MRI) [3]. Mathematically, these heterogeneities emerge as space scales, surface (endo- and epicardial) and internal (scar tissue and blood vessels) boundary conditions, and as spatial changes in excitation and coupling parameters.

2 Ventricular Wall Structural Models

Anisotropic fibre orientation and possible orthotropic sheet structure throughout the ventricular myocardium [9], along with the tissue geometry and heterogeneous cell electrophysiology, underlie both the spatio-temporal pattern

of the spread of electrical excitation, and the mechanical properties. Propagation of electrical activity is orthotropically anisotropic [14], being fastest in the direction of the long axis of the fibre due to the presence of gap junctions that are principally located at the ends of the myocytes, and slowest across any sheet planes due to the small number of muscle branches connecting otherwise electrically-insulated muscle sheets [14, 16]. Contraction of myocytes occurs in the long axis direction and, together with transmural shear along sheet planes, results in transmural thickening and apex-base shortening. In DT-MRI [2], theory suggests that the primary eigenvector of the measured diffusion tensor will be along the predominant direction of myocyte long-axis orientation [13, 15, 18] and that the secondary eigenvector will lie in the cleavage plane between sheets [12].

3 The Human Virtual Ventricular Wall

One possible mechanism for the transition from VT to VF is when a single re-entrant wave of excitation (a scroll wave) that rotates around a phase singularity (a filament) with a high frequency breaks down into multiple wavelets. We used the human ventricular electrophysiology model of Ten Tusscher et al. [19] and constructed two heterogeneous models of the left ventricular free wall in order to investigate the influence of tissue geometry and architecture on filament dynamics: (1) a simple cuboid model with dimensions of $60 \times 60 \times 20\,\text{mm}$, where the fibre direction always pointed parallel to the endocardial and epicardial surfaces and rotated 120° across the ventricular wall at a rate of 6°/mm; and (2) a wedge model with geometry and architecture obtained from a DT-MRI data set of the human ventricles. The wedge dimensions are similar to those of the cuboid. In all cases, we set the diffusion coefficient in the fibre direction to give a conduction velocity for a solitary plane wave of $0.7\,\text{m}\,\text{s}^{-1}$. For isotropic propagation we set the diffusion coefficients in the fibre, sheet and sheet normal directions the same. To introduce fibre orientation we set the diffusion coefficients with the ratio 4:1:1 such that conduction velocity is twice as fast along the fibre as across it, i.e. cylindrically anisotropic. To introduce sheet structure and orthotropic propagation, the diffusion coefficients were set with the ratio 36:9:1 to give a conduction velocity ratio of 6:3:1. For filament tracking and quantification we used the method described by Fenton and Karma [8]. We integrated equations using a Forward Time Centred Space method, with an operator splitting and adaptive time step technique [20] utilising a minimum time step of $\Delta t_{min} = 0.02\,\text{ms}$ and a maximum time step $\Delta T = 0.2\,\text{ms}$. Space steps in the cuboid model were $\Delta x = \Delta y = \Delta z = 0.33\,\text{mm}$. In the wedge model, space steps were $\Delta x = \Delta y = 0.425\,\text{mm}$ and $\Delta z = 0.5\,\text{mm}$ as defined by the DT-MRI dataset, to give approximately 4×10^5 nodes inside the tissue. See [4] for more detailed information on the models.

Figure 2 shows snapshots at $t = 2\,\text{s}$ of membrane potential on the surface of the model geometries and corresponding filament locations, for both models

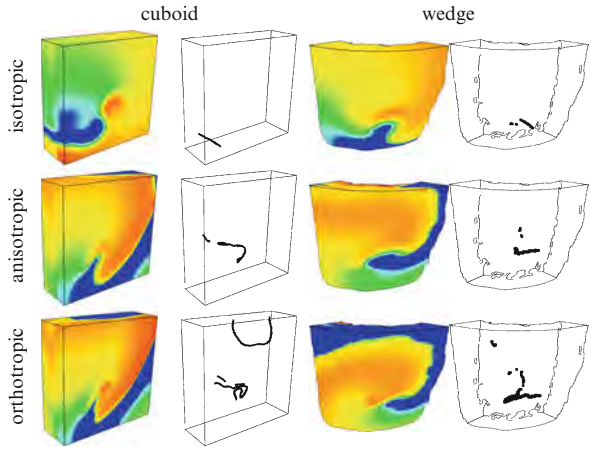


Fig. 2. Snapshots of membrane potential and filament locations after 2 s of re-entrant activity in isotropic, anisotropic and orthotropic cuboid and wedge models. For both models the snapshots are from an epicardial aspect, with the scroll wave rotating clockwise

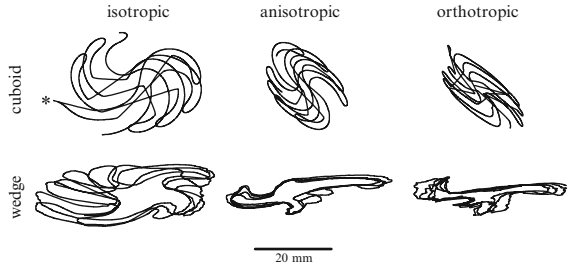


Fig. 3. Filament trajectories on the epicardial surface of the cuboid and wedge models during 1 s of simulation, under isotropic, anisotropic and orthotropic conditions. The asterisk on the isotropic cuboid trajectory indicates where the filament moved off the epicardial surface of the geometry

under isotropic, anisotropic and orthotropic conditions. For the isotropic cuboid, the scroll wave dies out soon after as the filament reaches the boundary. The multiple filaments for the orthotropic cuboid show the beginning of scroll wave breakup – numerous wavelets form soon after and the activation patterns in the tissue represent the complex patterns seen during VF. Note the numerous filaments present in the wedge model under all three conditions. Filament trajectories on the epicardial surfaces of the geometries are shown in Fig. 3. For the cuboid model, changing from isotropy through anisotropy to orthotropy has the effect of rescaling the meander of the filament in the direction perpendicular to the fibre axis, which on the epicardial surface is the sheet normal direction. Scroll wave filament length during 1 s of simulation

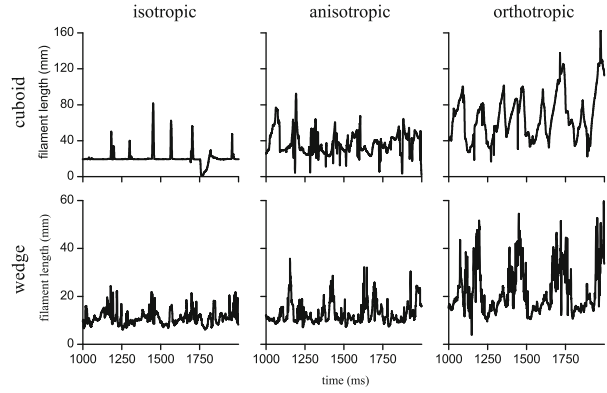


Fig. 4. Scroll wave filament length during 1 s of simulation in the cuboid and wedge models, under isotropic, anisotropic and orthotropic conditions. Note the different scales on the ordinate for the cuboid and wedge models

is shown in Fig. 4. Note the different scales on the ordinate for the cuboid and wedge models. For all conditions, the filament length in the cuboid is longer than in the wedge, a result of tissue geometry (i.e. size) rather than architecture. Oscillations of filament length are evident in all simulations, a consequence of filament twist which is due to the heterogeneous excitation kinetics in the models – see [7]. For both models, filament length increases as anisotropy and then orthotropy are introduced. These effects – due to rotational anisotropy – further increase the effects of the transmural excitation heterogeneity. Although filament curvature increases with anisotropy then orthotropy in the wedge model, the same pattern is not seen in the cuboid. The maximum twist along a single filament increases in both models as anisotropy and then orthotropy are introduced.

4 Conclusions

The normal sinus rhythm of the heart, re-entrant arrhythmias and fibrillation can all be described by the propagation of nonlinear waves in an excitable medium. Physiological and pathological patterns can be explained in terms of nonlinear wave properties – the dependence of velocity on rate by nonlinear dispersion, and breakdown from spatio-temporal patterned activity into irregularity by interactions between waves and by changes in wave stability. However, this emphasis on nonlinear wave dynamics neglects the overall architecture of the heart and its heterogeneities. By combining all these structural (or parametric) heterogeneities into computational models of excitation, propagation can be explored and the resultant functional heterogeneities, that are produced by slow recovery processes, emerge. Although the types of possible wave behaviours follow from the physics of excitable media, the details of the

initiation and subsequent evolution of patterns of excitation in cardiac muscle depends on the details of geometry, anisotropic and orthotropic architecture, and heterogeneities.

Acknowledgements

We thank Drs. Patrick A. Helm and Raimond L. Winslow at the Center for Cardiovascular Bioinformatics and Modeling and Dr Elliot McVeigh at the National Institute of Health for provision of the DT-MRI data sets. This work was supported by the European Union through the Network of Excellence BioSim (contract LHSB-CT-2004-005137), the Dr Hadwen Trust for Humane Research, and the Medical Research Council.

References

1. Barkley, D., Kness, M., Tuckerman, L.S.: Phys. Rev. A **42**, 2489–2492 (1990)
2. Basser, P.J., Mattiello, J., LeBihan, D.: Biophys. J. **66**, 259–267 (1994)
3. Benson, A.P., Gilbert, S.H., Li, P., Newton, S.M., Holden, A.V.: Math. Model. Nat. Phenom. (in press)
4. Benson, A.P., Halley, G., Li, P., Tong, W.C., Holden, A.V.: Chaos **17**, 015105 (2007)
5. Biktashev, V.N., Holden, A.V.: Phys. Lett. A **181**, 216–224 (1993)
6. Biktashev, V.N., Holden, A.V., Mirnov, S.F., Pertsov, A.M., Zaitsev, A.V.: Int. J. Bifurcat. Chaos **9**, 695–704 (1999)
7. Clayton, R.H., Holden, A.V.: Prog. Biophys. Mol. Biol. **85**, 473–499 (2004)
8. Fenton, F., Karma, A.: Chaos **8**, 20–47 (1998)
9. Gilbert, S.H., Benson, A.P., Li, P., Holden, A.V.: Eur. J. Cardiothorac. Surg. **32**, 231–249 (2007)
10. Gray, R.A., Jalife, J., Panfilov, A.V., Baxter, W.T., Cabo, C., Davidenko, J.M., Pertsov, A.M.: Science **270**, 1222–1223 (1995)
11. Gray, R.A., Pertsov, A.M., Jalife, J.: Nature **392**, 75–78 (1998)
12. Helm, P.A., Tseng, H.-J., Younes, L., McVeigh, E.R., Winslow, R.L., Magn. Reson. Med. **54**, 850–859 (2005)
13. Holmes, A.A., Scollan, D.F., Winslow, R.L., Magn. Res. Med. **44**, 157–161 (2000)
14. Hooks, D.A., Trew, M.L., Caldwell, B.J., Sands, G.B., LeGrice, I.J., Smaill, B.H.: Circ. Res. **101**, e103–e112 (2007)
15. Hsu, E.W., Muzikant, A.L., Matulevicius, S.A., Penland, R.C., Henriquez, C.S.: Am. J. Physiol. **274**, H1627–H1634 (1998)
16. LeGrice, I.J., Smaill, B.H., Chai, L.Z., Edgar, S.G., Gavin, J.B., Hunter, P.J.: Am. J. Physiol. **269**, H571–H582 (1995)
17. Panfilov, A.V., Holden, A.V.: Computational Biology of the Heart Wiley, Chichester (1997)
18. Scollan, D.F., Holmes, A., Winslow, R.L., Forder, J.: Am. J. Physiol. **275**, H2308–H2318 (1998)
19. ten Tusscher, K.H.W.J.,, Noble, D., Noble, P.J., Panfilov, A.V., Am. J. Physiol. **286**, H1573–H1589 (2004)
20. Qu, Z., Garfinkel, A.: IEEE Trans. Biomed. Eng. **46**, 1166–1168 (1999)
21. Zhang, H.G., Dobrzynski, H., Holden, A.V., Boyett, M.R.: Lect. Notes Comp. Sci. **2674**, 132–140 (2003)

A Bidomain Numerical Validation for Assessing Times of Fast and Ending Repolarization from Monophasic Action Potentials

P. Colli Franzone[1], L.F. Pavarino[2], S. Scacchi[1], and B. Taccardi[3]

[1] Dip. di Matematica, Università di Pavia, Via Ferrata 1, 27100 Pavia, Italy
peiro.collifranzone@unipv.it, simone.schacchi@unimi.it
[2] Dip. di Matematica, Università di Milano, Via Saldini 50, 20133 Milano, Italy
luca.pavarino@unimi.it
[3] CVRTI, University of Utah, Salt Lake City, UT 84112, USA
taccardi@cvrti.utah.edu

Summary. 3D numerical simulations of unipolar electrograms (EGs) and hybrid monophasic action potentials (HMAPs) were performed by using the cardiac Bidomain model with homogeneous and heterogeneous Luo–Rudy I membrane models. While estimating local recovery times from EGs can be difficult in case of flat T-waves or linear ST ramps, the HMAP signal always displays a monophasic downstroke as does the transmembrane action potential (TAP) and contains valuable information for assessing repolarization time. The simulation results show that: (a) the HMAP fast repolarization time is a reliable estimate of the TAP fast repolarization time; (b) the HMAP ending (90%) repolarization time is a less reliable estimate of the TAP ending repolarization time; (c) analogous conclusions hold for the associated action potential durations APD and APD90.

1 Introduction

While methods for determining cardiac activation times from electrographic signals recorded directly from the heart have been firmly established, see e.g. [11] and the references therein, there are still uncertainties and controversies about the best method for determining cardiac recovery times. The repolarization time at a given point x of the cardiac domain is related to some time markers associated with the downstroke of the transmembrane action potential (TAP, considered to be the gold standard) recorded at x, or with the T wave of the extracellular unipolar recording (EG) recorded at x. An alternative extracellular technique is based on the downstroke of the hybrid monophasic action potential (HMAP; see [6, 8]) at x, obtained by taking as a reference the potential at a fixed permanently depolarized (PD) site and

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*,
Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_54,

reversing its polarity. The HMAP is equivalent to the difference between the EG at the PD site and the EG at the exploring site x. The HMAP information content during the repolarization phase has been recently questioned and the HMAP genesis has been a controversial subject. In this paper, we present the results of a Bidomain – LR1 3D simulation study showing that HMAPs contains valuable information for assessing both the fast and ending TAP repolarization times, confirming our recent study [4].

2 Methods

2.1 The Bidomain: LR1 Model

Our simulation study is based on the macroscopic bidomain representation of the cardiac tissue coupled with Luo–Rudy I [7] ionic membrane model. This system allows us to compute the intra and extracellular potentials $u_i(x,t), u_e(x,t)$, the transmembrane potential $v(x,t) = u_i(x,t) - u_e(x,t)$, the gating variables $w(x,t)$ and the ionic concentrations $c(x,t)$, as the solutions of the reaction-diffusion system

$$\begin{cases} c_m \dfrac{\partial v}{\partial t} - div(D_i \nabla u_i) + i_{ion}(v,w,c) = -i^i_{app} & \text{in } H \times (0,T) \\ -c_m \dfrac{\partial v}{\partial t} - div(D_e \nabla u_e) - i_{ion}(v,w,c) = i^e_{app} & \text{in } H \times (0,T) \\ \dfrac{\partial w}{\partial t} - R(v,w) = 0, \qquad \dfrac{\partial c}{\partial t} - S(v,w,c) = 0 & \text{on } H \times (0,T) \\ \mathbf{n}^T D_{i,e} \nabla u_{i,e} = 0 & \text{in } \partial H \times (0,T) \\ v(\mathbf{x},0) = v_0(\mathbf{x}), \qquad w(\mathbf{x},0) = w_0(\mathbf{x}), \qquad c(\mathbf{x},0) = c_0(\mathbf{x}) & \text{in } H. \end{cases} \tag{1}$$

Here $c_m = \chi C_m$ and $i_{ion} = \chi I_{ion}$ denote the capacitance and the ionic current of the membrane per unit volume, and $i^{i,e}_{app}$ the applied intra- and extracellular currents per unit volume satisfying the compatibility condition $\int_H i^i_{app} = \int_H i^e_{app}$. χ denotes the surface membrane area per unit volume. The cardiac volume H is considered fully insulated, since we have imposed zero normal fluxes of intra- and extracellular currents. The extracellular potential u_e, defined apart from an independent constant determined by the choice of the reference potential, is determined by the condition $\int_H u_e(x,t)dx = 0$.

2.2 Numerical Methods

The cardiac domain H considered in this study is a cartesian slab of dimensions $1.92 \times 1.92 \times 0.48\,\mathrm{cm}^3$, modeling a portion of the left ventricular wall. In all computations, a structured grid of $192 \cdot 192 \cdot 48$ hexahedral isoparametric Q_1 finite elements of size $h = 0.1\,\mathrm{mm}$ is used in space, while the time discretization is an Implicit-Explicit Euler method. The linear solver at each

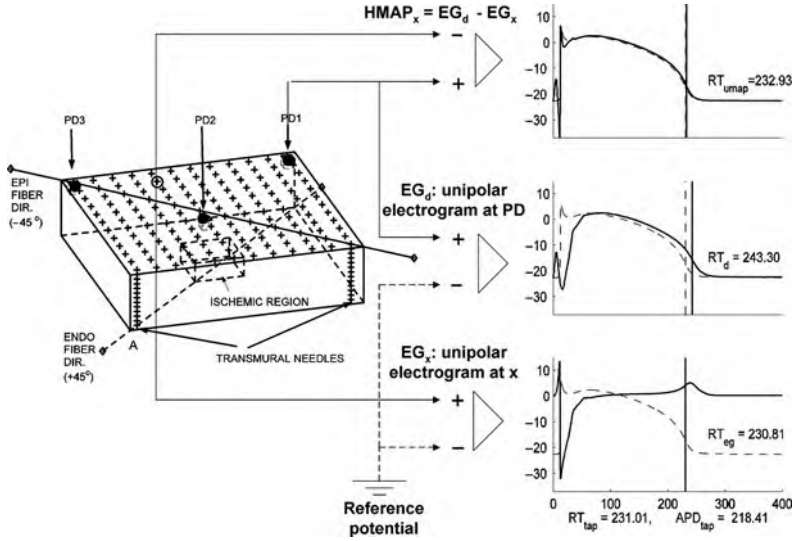


Fig. 1. *Left*: cardiac slab H, permanently depolarized sites PD1, PD2, PD3, ischemic region, transmural needles. *Right*: unipolar electrograms EG_x at the exploring site x (*bottom*) and EG_d at the PD site (*middle*), hybrid monophasic action potential HMAP_x at x (*top*)

time step is the conjugate gradient method, preconditioned by a hybrid multilevel Schwarz preconditioner. We use the PETSc parallel library [1] in order to ensure the parallelization and portability of our code, run on a Linux Cluster with 56 Opteron AMD processors and Infiniband network. Each simulation required about 8 h on 36 processors; further numerical details on our parallel solver can be found in [2, 3, 9, 12].

2.3 Multi-Electrode Array

In our cardiac slab H, we consider a matrix of 12×12 exploring multielectrode needles spaced 1.6 mm from each other and 0.8 mm from the slab boundary, as shown in Fig. 1. Each needle carries 13 recording sites, spaced 0.4 mm along the shank. We then have 12×12 sites on each of the 13 intramural planes, for a total of $12 \times 12 \times 13 = 1{,}872$ recording sites in the slab, each recording the intra and extracellular potentials.

2.4 Potential Waveform and Repolarization Markers

At each recording site x we store the following waveforms:

$\mathrm{EG}_x(t) = u_e(x,t)$: unipolar electrogram at the exploring site x,

$\mathrm{TAP}_x(t) = u_i(x,t) - u_e(x,t)$: transmembrane potential at x,

$\mathrm{HMAP}_x(t) = \mathrm{EG}_d(t) - \mathrm{EG}_x(t)$: hybrid monophasic action potential at x, where $\mathrm{EG}_d(t)$ is the unipolar electrogram at a permanently depolarized site PD (see Fig. 1 and below). From these waveforms, we then compute the following markers of fast repolarization time:

$\mathrm{RT}_{tap}(x)$ = time of min $\partial_t \mathrm{TAP}_x(t)$ during downstroke,

$\mathrm{RT}_{hmap}(x)$ = time of min $\partial_t \mathrm{HMAP}_x(t)$ during downstroke,

and the following markers of ending repolarization time:

$\mathrm{RT90}_{tap}(x)$: instant when $\mathrm{TAP}_x(t) = 90\%$ of its resting value during downstroke,

$\mathrm{RT90}_{hmap}(x)$: instant when $\mathrm{HMAP}_x(t) = 90\%$ of its resting value during downstroke.

2.5 Transmural Heterogeneity

We consider three different types of transmural APD distribution, one homogeneous (H-slab) and the other two heterogeneous (3-slab and W-slab), while in any plane parallel to the epicardium all cells have the same intrinsic APD. In the heterogeneous slabs, the intrinsic APD of the cells is obtained by multiplying the potassium current I_K in the LR1 model by a factor fact_{I_K}, as detailed in Table 1. In this way, we mimic the experimental transmural APD profile with M-cell layers as in [13] (3-slab) or as in [14, Fig. 4], [10, Fig. 5] (W-slab); see [3] for more details.

2.6 Subendocardial Ischemia

Two simulations with subendocardial moderate (MI-slab) and severe (SI-slab) ischemic regions are performed. The ischemic region has dimensions $0.4 \times 0.4 \times 0.16\,\mathrm{cm}^3$ and is located as shown in Fig. 1. In the LR1 model, the current I_K is scaled by a factor 2.325, yielding TAPs with $\mathrm{APD90} = 250\,\mathrm{ms}$. Inside the ischemic region, the extracellular potassium concentration $[K]_o$ is increased from 5.4 mM (control) to 10.5 mM (MI-slab) and 18 mM (SI-slab).

Table 1. Parameter calibration for modeling the transmural heterogeneities in the three cardiac slabs H-slab, 3-slab, W-slab

	H-slab	3-slab			W-slab			
# of layers	1	3			4			
		Endo	Mid	Epi	Endo	Sub-endo	Mid	Epi
thickness (cm)	0.48	0.16	0.16	0.16	0.058	0.096	0.254	0.072
fact_I_K	1	2.62	1.95	2.88	2.71	1.95	2.47	2.88
APD (ms)	266	235	272	225	232	272	242	225

2.7 Permanently Depolarized (PD) Volume

In order to generate an almost constant TAP inside a small volume, denoted by permanently depolarized (PD) volume, we assign the extracellular potassium concentration equal to the intracellular one, i.e. I_{K1} is zero in the small PD volume with dimensions $0.8 \times 0.8 \times 0.8\,\mathrm{mm}^3$. We considered three PD sites labeled PD1, PD2, PD3 in Fig. 1, but for brevity we only report the results with PD3.

2.8 Stimulation Site

Inside the PD volume the transmembrane potential values are above threshold thus generating a first excitation-recovery TAP that sweeps the cardiac slab H. We wait for 500 ms and take the steady state reached by the bidomain system as the initial condition for our simulations. An extracellular stimulus ($i^e_{app} = -250\,\mathrm{mA/cm}^3$ for 1 ms) is then applied in a small volume (3 mesh points in each direction) at the locations A in Fig. 1 and an intracellular stimulus $i^i_{app} = i^e_{app}$ is also applied in order to satisfy the compatibility condition for the solvability of the bidomain system 1.

3 Results

The results of Table 2 show a very high correlation coefficient (≥ 0.98) between both the markers of fast repolarization RT_{hmap} and RT_{tap} and the markers of ending repolarization $\mathrm{RT90}_{hmap}$ and $\mathrm{RT90}_{tap}$. This good global matching, confirmed by the regression lines of Fig. 2, top, assures a high reliability of the markers in terms of localizing regions that repolarize first and last and in terms of repolarization patterns. RT_{hmap} provides good estimates of RT_{tap} with average discrepancy of about 2 ms, while $\mathrm{RT90}_{hmap}$ provides less accurate estimates of $\mathrm{RT90}_{tap}$ with average discrepancy of about 4 ms. This lower accuracy is also confirmed by the histograms reported in Fig. 2, bottom.

Table 2. Recovery times and action potential duration markers discrepancies

	RT_{hmap} vs RT_{tap}			$\mathrm{RT90}_{hmap}$ vs $\mathrm{RT90}_{tap}$			ARI_{hmap} vs APD			$\mathrm{ARI90}_{hmap}$ vs APD90		
	Mean	Std	Corr	Mean	Std	Corr	Mean	Std	Corr	Mean	Std	Corr
H-slab	1.83	1.60	0.99	3.04	2.15	0.99	1.83	1.60	0.77	3.04	2.15	0.72
3-slab	2.39	2.22	0.98	5.27	2.45	0.98	2.39	2.22	0.94	5.27	2.45	0.92
W-slab	2.07	2.00	0.99	5.15	1.97	0.99	2.07	2.00	0.92	5.15	1.97	0.95
MI-slab	2.05	2.99	0.98	3.07	2.24	0.98	2.05	2.99	0.93	3.07	2.24	0.90
SI-slab	1.81	1.53	0.99	2.95	1.87	0.98	1.81	1.53	0.84	2.95	1.87	0.63
Global	2.03	2.15	0.98	3.90	2.40	0.98	2.03	2.15	0.92	3.90	2.40	0.88

Mean Average absolute difference between two markers, *Std* Standard deviation of the absolute difference between two markers, *Corr* Correlation coefficient between two markers

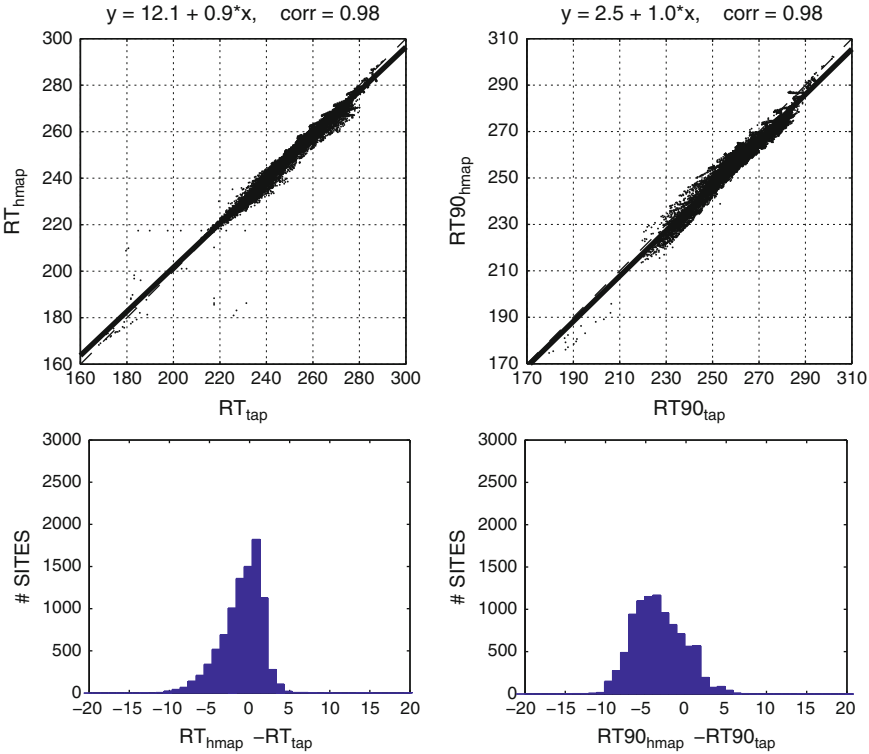


Fig. 2. *Top*: regression lines of RT_{hmap} vs RT_{tap} (*left*) and $\mathrm{RT90}_{hmap}$ vs $\mathrm{RT90}_{tap}$ (*right*), for all 5 slabs of Table 2. *Bottom*: histograms of discrepancies $\mathrm{RT}_{hmap}-\mathrm{RT}_{tap}$ (*left*) and $\mathrm{RT90}_{hmap}-\mathrm{RT90}_{tap}$ (*right*) with 1 ms bins, for all 5 slabs of Table 2

Despite this qualitatively good global performance, the extracellular RT markers do not always yield an accurate estimate of the spatial distribution of TAP-based repolarization time, because some local large discrepancies might ensue. A preliminary study of the reliability of EG-based RT markers has been presented in our recent work [5], here extended to include also the presence of ischemic regions. Nevertheless, HMAP-based markers represent a reliable alternative for estimating RT_{tap} ($\mathrm{RT90}_{tap}$) at recording sites located inside or near the borders of the ischemic region where the classical EG-based markers may fail because of linear ST ramp or absence of T wave.

References

1. Balay, S., et al.: PETSc home page. http://www.mcs.anl.gov/petsc, 2001
2. Colli Franzone, P., Pavarino, L.F.: Math. Mod. Meth. Appl. Sci. **14**(6), 883–911 (2004)

3. Colli Franzone, P., Pavarino, L.F., Taccardi, B.: Math. Biosci. **204**, 132–165 (2006)
4. Colli Franzone, P., Pavarino, L.F., Scacchi, S., Taccardi, B.: Am. J. Physiol. Heart Circ. Physiol. **293**, H2771–H2785 (2007)
5. Colli Franzone, P., Pavarino, L.F., Scacchi, S., Taccardi, B.: FIMH07. In: Sachse, F.B., Seemann, G. (eds.) LNCS, vol. 446, pp. 139–149. Springer, Berlin (2007)
6. Franz, M.R.: Monophasic Action Potentials: Bridging Cells to Bedside. Futura Publishing Company, New York (2000)
7. Luo, C., Rudy, Y.: Circ. Res. **68**(6), 1501–1526 (1991)
8. Nesterenko, V.V., Weissenburger, J., Antzelevitch, C.: J. Cardiovasc. Eletrophysiol. **11**, 948–951 (2000)
9. Pavarino, L.F., Scacchi, S.: SIAM J. Sci. Comp. **31**(1), 420–443 (2008)
10. Poelzing, S., Rosenbaum, D.S.: Am. J. Physiol (Heart Circ. Physiol.) **286**, H2001–H2009 (2004)
11. Punske, B.B., et al.: Ann. Biomed. Engrg. **31**(7), 781–792 (2003)
12. Scacchi, S.: Comp. Meth. Appl. Mech. Engrg. **197**(45–48), 4051–4061 (2008)
13. Viswanathan, P.C., Shaw, R.M., Rudy, Y.: Circulation **99**, 2466–2474 (1999)
14. Yan, G.X., et al.: Circulation **98**, 1921–1927 (1998)

Framework for Modular, Flexible and Efficient Solving the Cardiac Bidomain Equations Using PETSc

G. Seemann[1], F.B. Sachse[2], M. Karl[1], D.L. Weiss[1], and V. Heuveline[3], and O. Dössel[1]

[1] Institute of Biomedical Engineering, Universität Karlsruhe (TH), Karlsruhe Institute of Technology, Karlsruhe, Germany, Gunnar.Seemann@kit.edu, Meike.Karl@ibt.uni-karlsruhe.de, Weiss@ibt.uni-karlsruhe.de, Olaf.Doessel@kit.edu

[2] Nora Eccles Harrison Cardiovascular Research and Training Institute, University of Utah, Salt Lake City, Utah, USA, fs@cvrti.utah.edu

[3] Institute for Applied and Numerical Mathematics, Universität Karlsruhe (TH), Karlsruhe Institute of Technology, Karlsruhe, Germany, vincent.heuveline@kit.edu

Summary. In this work, a new framework is presented that is suitable to solve the cardiac bidomain equation efficiently using the scientific computing library PETSc. Furthermore, the framework is able to modularly combine different ionic channels and is flexible enough to include arbitrary heterogeneities in ionic or coupling channel density. The ability of this framework is demonstrated in an example simulation in which the three-dimensional electrophysiological heterogeneity was adjusted in order to get a positive T-wave in the body electrocardiogram (ECG).

1 Introduction

The cardiac electrophysiological modeling got more and more quantitative in the last years. This is mainly based on new measurement results concerning e.g. electrophysiology, heterogeneity, and fiber orientation. Furthermore, additional complex approaches like Markovian models replace the traditional Hodgkin–Huxley type formulations [3]. Cardiologists got interested in the modeling. This implies the use of whole heart models of humans or even ECG simulations. These interests lead to the necessity of a modular, flexible and efficient framework which is presented in this work.

The system includes the individual integration of each channel equation to consider time constants of the different processes. It is able to plug the different channels flexible and modularly into new electrophysiological models. This procedure can be done for a variety of tissue types and on arbitrary geometries. Any fiber orientation can be considered to account for the intra- and

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*, Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_55,

extracellular anisotropy. Different modules exist to generate system matrices within different methods (e.g. FDM, FEM) to solve them using PETSc.

2 Methods

In order to describe the electrophysiological processes in the heart by mathematical models, several microscopic and macroscopic approaches exist. In this work, the macroscopic reaction-diffusion system called bidomain model [2] was used as basis for the developed framework. This approach consists of a reaction part describing the electrophysiology of each mathematical cell and a diffusion part responsible for the excitation conduction process in a tissue model. Using the bidomain model, macroscopic electrophysiological processes in the heart can be investigated.

The tissue model describes the geometrical basis for the modeling. It could either be a schematic representation of the heart or based on segmented image data. Fiber orientation can be integrated using diffusion tensor MRI data or integrated with rule based methods. Figure 2 shows an example of a geometry extracted from the MEET Man project (www.ibt.kit.edu/meetman.php). For a realistic modeling, both the ventricles and the atria should be considered.

Cardiac electrophysiological models are capable to describe single cardiac cell behavior. They describe the transmembrane voltage V_m, gating processes and the change of ion concentrations with a set of non-linear, coupled ODEs. The modeling is based on the Hodgkin–Huxley (HH) equations:

$$\frac{\mathrm{d}V_m}{\mathrm{d}t} = -\frac{1}{C_m}\left(\sum I_{ion} - I_m\right) \tag{1}$$

$$I_{ion} = g_{max}\prod p_i(V_m - E_{ion}) \qquad \text{with} \quad \frac{\mathrm{d}p_i}{\mathrm{d}t} = \alpha(1-p_i) - \beta p_i \tag{2}$$

whereas C_m the membrane capacitance, I_{ion} ionic currents of different types, I_m the membrane current, g_{max} the maximum conductance of a channel, E_{ion} the Nernst potential and p_i the gates describing the opening and closing process of a gate with V_m dependent α forward and β backward rate constants. Different ionic channel currents I_{ion} have different gating properties leading to the specific course of the action potential in cardiac cells.

The bidomain model that is based on two coupled Poisson's equations can describe tissue electrophysiology based on these single cell models:

$$\nabla\cdot\left(\left(\sigma_i+\sigma_e\right)\nabla\Phi_e\right) = -\nabla\cdot\left(\sigma_i\nabla V_m\right) \tag{3}$$

$$\nabla\cdot\left(\sigma_i\nabla V_m\right) + \nabla\cdot\left(\sigma_i\nabla\Phi_e\right) = \beta\left(C_m\frac{\mathrm{d}V_m}{\mathrm{d}t} + \sum I_{Ion}\right) - I_{si} \tag{4}$$

Here, σ_i and σ_e are the intra- and extracellular conductivity tensors describing anisotropic properties due to the fiber orientation, Φ_e the extracellular potential and I_{si} the stimulus current. Equation 3 is an elliptical PDE which

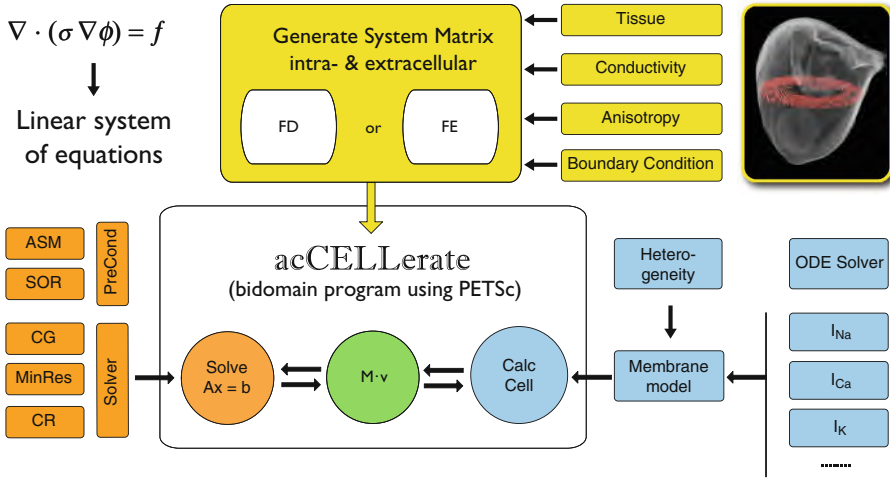


Fig. 1. Structure of the bidomain framework. (*top*) A matrix generator creates the system matrix using different input information. (*middle*) The main program (acCELLerate) calculates the bidomain equations. (*left*) The pre-conditioners and solvers provided by PETSc can be combined. (*right*) The membrane models were plugged together from different ion channel models and solved by ODE methods. Arbitrary heterogeneity can be set for each ion channel for each computational cell

consumes the most time solving the bidomain model. The non-linear parabolic PDE in (4) involves the HH equations. To consider different media like tissue and blood, boundary conditions need to be included [2]. Under assumption of equal anisotropy ratios between intra- and extracellular space, (3) can be neglected and the approach reduces to the monodomain model.

3 Results

The bidomain framework consists of four components. The first is a general system matrix generator. The second a library of channel models that can be integrated into membrane models. A third that uses PETSc to pre-condition as well as solve the LSE iteratively. Finally, the main component that integrates modularly the other components and does the time stepping (see Fig. 1).

The system matrix generator is compiling a matrix into the PETSc format. This matrix describes the operator $\nabla \cdot \sigma \nabla$. The matrix generator considers different tissue types like cardiac cell and blood. It can assign spatially varying conductivity values describing gap junction heterogeneity. Orthotropic anisotropy is considered using arbitrary fiber orientation for each computational cell. At this stage, boundary conditions are inserted into the matrix. Depending on the data format, the accuracy and the computational equipment, finite differences or finite element methods were used to form the LSE.

Solving the LSE of the left part of (3) is done by standard pre-conditioners and solvers of PETSc (Portable, Extensible Toolkit for Scientific Computation, www.mcs.anl.gov/petsc). Computational tests have shown that preconditioning with successive over-relaxation (SOR) in combination with solving the LSE using conjugate gradient (CG) with parameter adjustments lead to the fastest solution [4]. Additional decrease in computational time can be achieved when solving the LSE not for every time step than the ODE since the extracellular processes are slower than those of the cell processes.

Most of the flexibility, modularity and efficiency is included in the third component. With this part, the electrophysiological properties of cardiac cells are reconstructed. Computational efficiency is achieved by several mechanisms. Pre-calculating constants is performed only during the initialization of the model. Additionally, look-up tables were generated for those variables that are only dependent on V_m. During the calculation, the Rush–Larsen scheme [6] is used reducing the number of mathematical functions significantly. Each ion channel is calculated independently from the others. This has the advantage that it's ODEs can be treated separately by e.g. forward Euler or more advanced and time adaptive methods provided e.g. by CVODE (www.llnl.gov/CASC/sundials). Furthermore, this approach eases the adaption of existing and the design of new electrophysiological models. This is possible since each channel model can be plugged modularly into a membrane model that only provides information about the membrane conductance and the intra- and extracellular ion concentrations [12]. Another important part that increases the flexibility is the strict division of parameters and equations using these parameters. The advantage is that all parameters can be changed without recompiling the code. A further important part to generate a flexible framework is that each parameter of the channel model can have arbitrary values spatially distributed in order to reconstruct electrophysiological heterogeneity. This is achieved by over-loading the corresponding parameter from outside the library.

The main component of the framework is "acCELLerate" (www.ibt.kit.edu/acCELLerate.php). The software internally uses the so-called operator splitting method [8,11]. The advantage of this process is that the cell models, the solving of the LSE and the matrix-vector operations can be performed sequentially and the software is more modular. The benefit is that only one software for the bidomain model is necessary even if the data format is varying. Additionally, the main component is responsible for IO, to communicate with the user or additional software using the generated data and to parallelize the process using the message passing interface (MPI) within PETSc.

To illustrate the features of this simulation framework an example simulation was performed investigating the effects of different electrophysiological heterogeneity on T-wave morphology. Heterogeneity was described three-dimensional for several ionic channels in transmural, intraventricular and apico-basal direction as summarized in [13]. The heterogeneity was assumed to be distributed gradually so that almost each cell is calculated with different

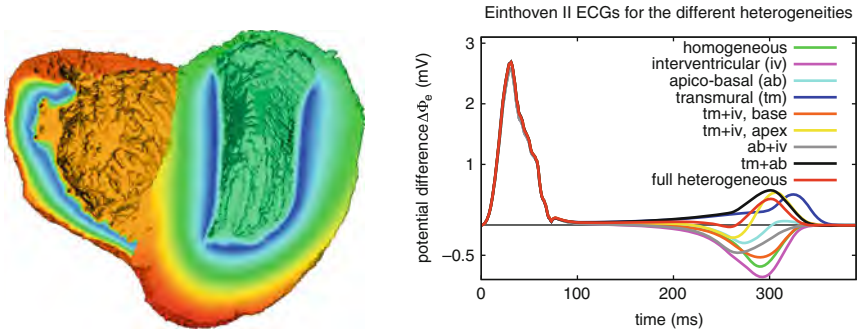


Fig. 2. Electrophysiological heterogeneity and ECG waves. (*left*) Distribution of the maximum conductance g_{Ks} (*blue*: 0.049 nS/pF; *red*: 0.49 nS/pF). The density of this channel is distributed heterogeneously in transmural, intraventricular and apico-basal direction. (*right*) Resulting ECGs of different heterogeneous configurations. The configuration "full heterogeneous" generates the most realistic T-wave

parameters. As an example, the distribution of the maximum conductance g_{Ks} of the slow delayed rectifier potassium channel is shown in Fig. 2. For several configurations from homogeneous over only transmural or only apico-basal heterogeneity towards full heterogeneity simulations of the excitation propagation in the heart were calculated. The corresponding body ECGs were calculated using the bidomain model by extracting the extracellular field at standard derivation points on the MEET Man surface. The most realistic T-wave was generated only for the full heterogeneous configuration (see Fig. 2).

To calculate these results 48 2.8 GHz Intel Xeon "Harpertown" CPUs in a Xserve cluster were used. The heart model had 4.4 Mio. active ten Tusscher models [10] with a spatial resolution of 0.4 mm. The temporal discretization was 0.02 ms. The calculation of each heart cycle took approximately three hours using SOR pre-conditioner and the CG solver. The ECG calculation was performed on a tetrahedral mesh of the thoracic geometry consisting of 113,839 nodes. This calculation needed another hour.

4 Discussion and Conclusions

The new tool "acCELLerate" was developed solving large scale electrophysiological reaction-diffusion problems. It has an extensible program structure. By using PETSc including parallelization, time and memory efficient tools were implemented. The modular structure discloses the potentialities of extension and enables the integration of different cell models with variable parameters to achieve highly realistic simulations of electrophysiological processes.

Other groups have also developed efficient bidomain frameworks [1, 5, 7–9, 11, 14]. Mostly, they focused on optimizing the solution for (3) being the

most time consuming part. The advantage of the presented framework is the efficient solving of (4). The framework is modular and flexible enough to insert a huge variety of measurement data. Thus, it is applicable for solving the monodomain model and still has advantages for the bidomain model.

Further increase in computation time can be achieved by using spatially adaptive grids [1, 14]. Here, the grid resolution can be lower in areas with small variations of V_m. The disadvantage of this approach is that for each time step the system matrix has to be recompiled. This could superimpose the benefits of this approach, especially when simulating arrhythmia with only few situations of small V_m changes. The use of specific pre-conditioners [5, 7] could increase the calculation speed of the elliptical PDE. When using a large amount of CPUs this approach seems to be inevitable.

In future work, the advantage of a linearly implicit solution for (4) [1] has to be tested. At least the stability of the solution is higher with larger time steps since the ODE of (4) is stiff. On the other hand, another PDE needs to be solved inserting additional computation time. This procedure needs to be investigated if this implicit scheme is applicable to the bidomain model.

Cardiac dysfunction is the most common reason for death. To understand pathologies, insight into the electrophysiology is necessary. Additional to clinical experiments, computer simulations of the heart will be used to gain this knowledge. New measurement data has to be incorporated generating accurate electrophysiological models of the diseases. The presented flexible program is capable of using this data that is spatially and temporally varying. By this the framework will be able to complement experiments and will help to better understand the heart, as well as to support diagnosis and therapy.

References

1. Colli Franzone, P., Deuflhard, P., Erdmann, B., Lang, J., Pavarino, L.F.: SIAM J. Sci. Comput. **28/3**, 942–962 (2006)
2. Henriquez, C.S., Muzikant, A.L., Smoak, C.K.: J. Cardiovasc. Electrophysiol. **7**, 424–444 (1996)
3. Hodgkin, A.L., Huxley, A.F.: J. Physiol. **177**, 500–544 (1952)
4. Karl, M., Seemann, G., Sachse, F.B., Dössel, O., Heuveline, V.: Proc. MBEC, accepted (2008)
5. Plank, G., Liebmann, M., dos Santos, R.W., Vigmond, E.J., Haase, G.: IEEE Trans. Biomed. Eng. **54/4**, 585–596 (2007)
6. Rush, S., Larsen, H.: IEEE Trans. Biomed. Eng. **25/4**, 389–392 (1978)
7. Scacchi, S.: Comput. Methods Appl. Mech. Engrg. **197/45**, 4051–4061 (2008)
8. Sundnes, J., Lines, G.T., Tveito, A.: Math. Biosc. **172/2**, 55–72 (2001)
9. Skouibine, K., Trayanova, N.A., Moore, P.: Math. Biosc. **166/1**, 85–100 (2000)
10. ten Tusscher, K.H.W.J., Noble, D., Noble, P.J., Panfilov, A.V.: Am. J. Physiol. **286**, H1573–H1589 (2004)
11. Vigmond, E.J., Aguel, F.N., Trayanova, N.A.: IEEE Trans. Biomed. Eng. **49/11**, 1260–1269 (2002)

12. Weiss, D.L., Seemann, G., Dössel, O.: Biomed. Technik **50**, 566–567 (2005)
13. Weiss, D.L., Seemann, G., Keller, D.U.J., Farina, D., Sachse, F.B., Dössel, O.: Comput Cardiol **34**, 49–52 (2007)
14. Whiteley, J.: Ann. Biomed. Engin. **36/8**, 1398–1408 (2008)

On Efficiency and Accuracy in Cardioelectric Simulation

M. Weiser[1], B. Erdmann[1], and P. Deuflhard[1,2]

[1] Zuse Institute Berlin, 14195 Berlin, Germany, weiser@zib.de, erdmann@zib.de
[2] Freie Universität Berlin, 14195 Berlin, Germany, deuflhard@zib.de

Summary. Reasons for the failure of adaptive methods to deliver improved efficiency when integrating monodomain models for myocardiac excitation are discussed. Two closely related techniques for reducing the computational complexity of linearly implicit integrators, deliberate sparsing and splitting, are investigated with respect to their impact on computing time and accuracy.

1 Introduction

The excitation of myocardial cells is the basis for heart contraction and thus has attracted research in modelling as well as simulation. The propagation of a depolarization front of the transmembrane potential through the myocardium ultimately leads to the release of Ca^{2+} and thus a contraction of the muscle fibers. The evolution of the transmembrane potential is described by a set of reaction-diffusion equations modeling the ion transport by anisotropic diffusion between cells as well as between intra- and extracellular space (cf. [10–12]).

Under the simplifying assumption of identical intra- and extracellular diffusion tensor, myocardial excitation is described by the monodomain equation linking the transmembrane potential v to gating variables w and ion concentrations c:

$$c_m \partial_t v = \operatorname{div}(D_M \nabla v) + I_{\text{ion}}(v, w, c) \tag{1}$$

$$\partial_t w = R(v, w) \tag{2}$$

$$\partial_t c = S(v, w, c) \tag{3}$$

The reactions I_{ion}, R, and S are specified by membrane models. Here we restrict our attention to a very small phenomenological model by Aliev and Panfilov [1] and a physiological model by Luo and Rudy [7] of moderate complexity.

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*,
Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_56,

2 Adaptive Integration of Reaction-Diffusion Equations

The most common approach to spatial discretization of (1)–(3) is to use finite element methods on a fixed, quasi-uniform mesh. Due to the small width of the depolarization front relative to the heart geometry, rather fine meshes are needed in order to obtain sufficiently accurate solutions. Recently, adaptive FE methods have been proposed for simulating the myocardiac excitation [2, 3, 5, 13].

The results reported e.g. in [5] using the fully adaptive, linearly implicit FE code KARDOS [6] are mixed. On one hand, error control works just as expected and the number of vertices encountered in adaptive mesh refinement is a factor 150 below the number of vertices in a uniform mesh achieving the same local resolution (see Fig. 1 for illustration). On the other hand, the reduction in the number of degrees of freedom is not translated into savings of computing time, which is unacceptably high.

There seem to be several reasons for this effect. First of all, as long as a depolarization front is traversing the domain, the time step is limited by front speed and width. Only when the whole domain is covered by the plateau phase, the time step increases significantly. In the fibrillation example computed in [5], at any point in time there is a depolarization front somewhere in the domain, such that the time step remains small. Second, error control and mesh adaptation require the computation of an error estimator, which takes a significant part of the computational work. Third, mesh modifications require the frequent assembly of stiffness and mass matrices, up to a few times each time step. Finally, mesh modifications themselves and the resulting non-locality of data structures take their toll.

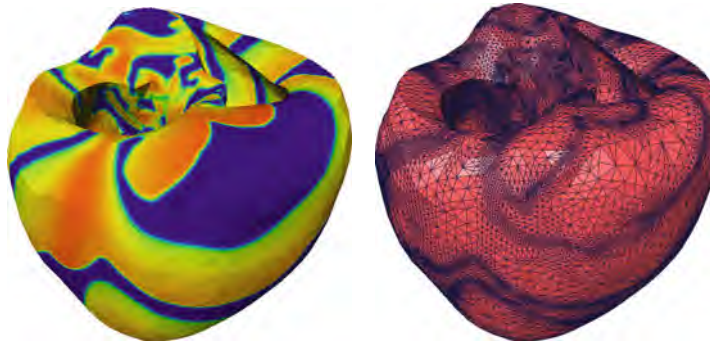


Fig. 1. Typical front of potential in ventricular fibrillation and the corresponding adaptive mesh

3 Deliberate Sparsing

Rosenbrock methods, which are linearly implicit Runge–Kutta methods, are used in KARDOS for time stepping. The lowest order method is the linearly implicit Euler scheme

$$(I - \tau(J + \nabla \cdot D\nabla))u_{k+1} = u_k + \tau(f(u_k) - Ju_k) \quad \text{with } J = f'(u_k) \tag{4}$$

for solving $\partial_t u = \operatorname{div}(D\nabla u) + f(u)$.

When applied to the monodomain equations (1)–(3), the linear system (4) has to be solved with a nonsymmetric block matrix

$$J_{\mathrm{AP}} = \begin{bmatrix} M - \tau(\partial_v f_v + A) & -\tau\partial_w f_v \\ -\tau\partial_v f_w & M - \tau\partial_w f_w \end{bmatrix} \quad \text{or} \quad J_{\mathrm{LR}} = \begin{bmatrix} * & * & * & * & * & * & * & * \\ * & * & & & & & & \\ * & & * & & & & & \\ * & & & * & & & & \\ * & & & & * & & & \\ * & & & & & * & & \\ * & & & & & & * & \\ * & & & & & * & * & * \end{bmatrix}$$

for the Aliev–Panfilov and Luo–Rudy membrane models, respectively. A denotes the stiffness matrix whereas M stands for the mass matrix. $*$ denotes a non-zero matrix with the sparsity structure of M. One problem with Rosenbrock methods is that their convergence order is reduced if the linear systems corresponding to (4) is not solved exactly.

A subset of linearly implicit methods, so-called W-methods (cf. [4]), allows to use arbitrary matrices $J \neq f'$ without affecting the order of convergence. This enables *deliberate sparsing* [9], a technique to drop certain parts of f' in order to decrease the computational complexity in computing J and solving the system. Even though in principle the error constant is affected by the approximation error $J - f'$, in practice the step sizes depend mostly on how well the large negative eigenvalues are captured. This is because stability limits the step size for explicit methods. Numerical experiments indicate that for normal heart beat cycle and both models, the system's stiffness is dominated by the diagonal blocks, such that dropping all off-diagonal blocks is possible without decreasing step size. The remaining block diagonal B is not only smaller, but also symmetric, which allows to use more efficient methods for symmetric matrices. Nevertheless, time savings given in Table 1 are disappointing. The reason for this is not yet clear and under investigation.

4 Splitting and Mass Lumping

Another sparsing opportunity on the element level comes from the fact that no spatial derivatives are involved in the reactions R and S. Instead of using the FE framework for propagating the gating variables and ion concentrations,

their values can be computed spatially decoupled by solving the ODE at each mesh vertex. This is known as *splitting* [8], leaving just a 1×1 FE "block" system to be solved.

From a different point of view, *mass lumping* by using quadrature rules with nodes only at the element vertices when assembling the mass matrices is an established method to obtain diagonal mass matrices. Of course, diagonal mass matrices are easily stored and trivially inverted. A closer inspection reveals that splitting leads to diagonal blocks in the notation of Sect. 3 as well. Moreover, splitting and mass lumping are mathematically equivalent, which permits a seamless interpretation of splitting in the framework of FE. Due to the lower accuracy of the vertex-based quadrature, the a-priori error estimates are worse for mass lumping than for Gaussian quadrature with nodes in the interior of the elements. While the FE convergence order is the same, the discretization error may be increased by a constant factor. In different contexts, factors of 4–6 are usually observed. Since an L^2-error reduction of four requires one additional level of uniform refinement, splitting may be expected to require an eight times larger discretization than using a full FE approach with more accurate quadrature.

A closer look at the spatial discretization errors of the gating variables and ion concentration reveals a more subtle influence of splitting on the total accuracy. Variables with slow dynamics are spatially smooth, whereas fast variables follow the depolarization front quickly and exhibit strong local features. On the same spatial grid, the discretization error in the slow variables is therefore very small, and an error increase by a factor of 4 has little effect on the overall solution. In contrast, for fast variables the effect is clearly visible (see Fig. 2). Comparable results are obtained with adaptive computations on 3D domains. Unfortunately, when mass lumping is only done for a subset

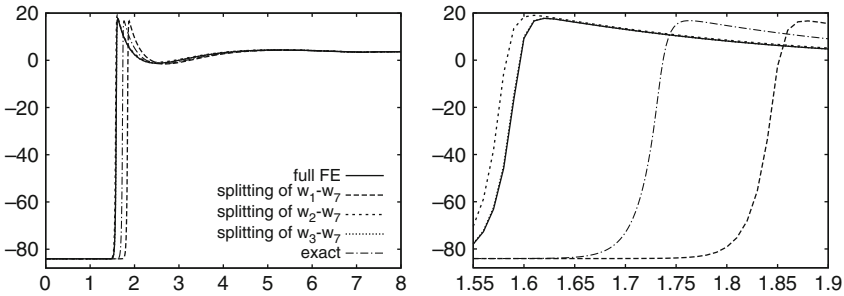


Fig. 2. Depolarization front positions 0.1 s after ignition at the right hand side boundary of the 1D domain for splitting of different groups of gating variables in the Luo–Rudy model. *Left*: whole domain. *Right*: zoom. The full FE solution travels faster than the exact solution. Mass lumping for the fastest gating variable w_1 has the largest effect and slows down the front even behind the exact solution

Table 1. Wall-clock runtime reduction factors due to algorithmic improvements in KARDOS

Alg. feature	Aliev–Panfilov	Luo–Rudy
Deliberate block sparsing	1.05	1.25
Splitting w_1–w_7	2	30
Splitting w_2–w_7		15
Splitting w_3–w_7		4

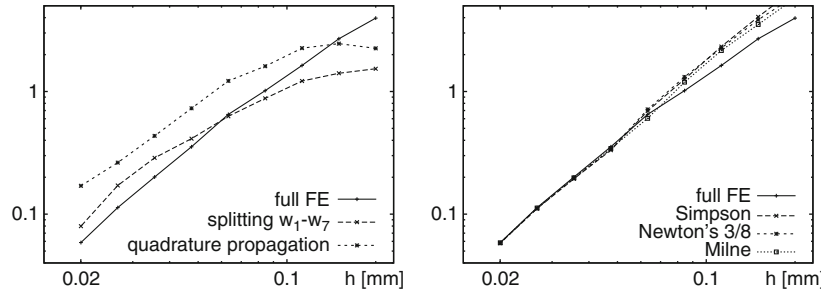


Fig. 3. Depolarization front position error on a 1D domain after 0.1 s of simulating the Luo–Rudy model with different mesh sizes. *Left*: full FE, mass lumping, and propagating at Gauß quadrature nodes. The errors, closely related to the front speed errors, span a range of about factor 3. *Right*: Effect of different quadrature rules

of the gating variables, the runtime improvements are less pronounced (see Table 1).

Surprisingly, the integration of gating variables without any spatial discretization error is actually possible – at least on fixed grids. The key observation is that with a given quadrature rule for assembling the reaction terms in (1), the gating variables are only evaluated at a finite set of points in the domain Ω. Propagating the gating variables just at these spatial positions yields exact values, up to time discretization, as far as the transmembrane potential is affected. Even more surprisingly, the overall accuracy of the front speed obtained with quasi-exact gating values can be worse than a full FE approach using the same quadrature rule, see Fig. 3 left. Increasing the accuracy of the quadrature rule gives results which are quite similar to the full FE approach, see Fig. 3 right, which indicates that the dominating discretization error is due to the transmembrane potential.

5 Conclusions

Adaptive discretization of cardioelectric excitation yields reliable results with a relatively small number of degrees of freedom, but the overhead of error estimation, mesh adaptation and frequent assembly on modified grids outweighs the efficiency gains. Deliberate sparsing and splitting techniques can

improve the situation to some extent, but their effect on accuracy needs to be investigated in more detail.

References

1. Aliev, R.R., Panfilov, A.V.: Chaos, Solitons Fractals **7**, 293–301 (1996)
2. Belhamadia, Y.: IEEE Transact. Biomed. Eng. **55**, 443–452 (2008)
3. Colli Franzone, P., Deuflhard, P., Erdmann, B., Lang, J., Pavarino, L.F.: SIAM J. Sci. Comput. **28**, 942–962 (2006)
4. Deuflhard, P., Bornemann, F.A.: Scientific Computing with Ordinary Differential Equations. Texts in Applied Mathematics, vol. 42, 2nd edn. Springer, New York (2002)
5. Deuflhard, P., Erdmann, B., Roitzsch, R., Lines, G.T.: Adaptive Finite Element Simulation of Ventricular Fibrillation Dynamics. Computing and Visualization in Science **12**, 201–205 (2009)
6. Lang, J.: Adaptive Multilevel Solution of Nonlinear Parabolic PDE Systems, volume 16 of Lecture Notes in Computational Science and Engineering. Springer, Berlin (2001)
7. Luo, C., Rudy, Y.: Circ. Res. **68**, 1501–1526 (1991)
8. Munteanu, M., Pavarino, L.F.: Decoupled Schwarz algorithms for implicit discretizations of nonlinear monodomain and bidomain systems. Math. Models Methods Appl. Sci. **19**(7), 1065–1097 (2009)
9. Nowak, U.: IMPACT Comput. Sci. Engrg. **5**, 53–74 (1993)
10. Panfilov, A.V., Holden, A.V. (eds.) Computational Biology of the Heart. Wiley, Chichester (1997)
11. Sachse, F.B.: Computational Cardiology, volume 2966 of Lecture Notes in Computer Science. Springer, Heidelberg (2004)
12. Sundnes, J., Lines, G.T., Cai, X., Nielsen, B.F., Mardal, K.-A., Tveito, A.: Computing the Electrical Activity in the Heart, volume 1 of Monographs in Computational Science and Engineering. Springer, New York (2006)
13. Whiteley, J.P.: Ann. Biomed. Engrg. **35**, 1510–1520 (2007)

Computational and Numerical Methods for the Efficient and Accurate Solution of the Bidomain Equations

J.P. Whiteley

Oxford University Computing Laboratory, Wolfson Building, Parks Road, Oxford OX1 3QD, UK, Jonathan.Whiteley@comlab.ox.ac.uk

Summary. Two previously published algorithms for solving the bidomain equations are combined to yield an algorithm that efficiently computes an accurate numerical solution of the bidomain equations. The first of these algorithms utilises the multiscale nature of the governing equations by solving the more rapid processes at a much higher resolution than the slower processes, and is ideal for use when a steep action potential wavefront is propagating across tissue. The second algorithm is suitable when no fast processes are taking place. This combined algorithm results in a threefold increase in computational efficiency over the most efficient algorithm that it is compared to for the simulation presented here.

1 The Bidomain Equations

The bidomain equations, or a simplification known as the monodomain equations, are the most commonly used mathematical model of tissue level cardiac electrophysiology. The bidomain equations may be written as [2]:

$$\chi\left(\mathcal{C}_m\frac{\partial V_m}{\partial t}+I_{\text{ion}}(\mathbf{u},V_m)\right)-\nabla\cdot(\sigma_i\nabla(V_m+\phi_e))=I_{sv_i}, \tag{1}$$

$$\nabla\cdot((\sigma_i+\sigma_e)\nabla\phi_e+\sigma_i\nabla V_m)=I_{sv_e}, \tag{2}$$

$$\frac{\partial\mathbf{u}}{\partial t}=\mathbf{f}(\mathbf{u},V_m), \tag{3}$$

where $V_m(\mathbf{x},t)$ is the transmembrane potential, $\phi_e(\mathbf{x},t)$ is the extracellular potential, $\mathbf{u}(\mathbf{x},t)$ is a vector containing various gating variables and chemical concentrations, σ_i is the intracellular conductivity tensor, σ_e is the extracellular conductivity tensor, χ is the surface to volume ratio, $\mathcal{C}_m$ is the membrane capacitance per unit area, I_{ion} is the ionic current, I_{sv_i} is the intracellular stimulus current applied within the tissue volume, and I_{sv_e} is the extracellular stimulus current applied within the tissue volume. Functional forms for I_{ion} and $\mathbf{f}$ are prescribed by an electrophysiological cell model – see [3] for a collection of these.

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*, Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_57,

Boundary conditions are required for (1) and (2) – these are given by

$$\mathbf{n} \cdot (\sigma_i \nabla (V_m + \phi_e)) = I_{sa_i}, \qquad \mathbf{n} \cdot (\sigma_e \nabla \phi_e) = I_{sa_e}, \tag{4}$$

where $\mathbf{n}$ is the outward pointing unit normal vector to the tissue, I_{sa_i} is the intracellular stimulus current applied across the boundary, and I_{sa_e} is the extracellular stimulus current applied across the boundary. The system (1)–(3) subject to boundary conditions (4) is then closed by specifying initial conditions for V_m and $\mathbf{u}$. We note that ϕ_e is only required to be defined up to an additive constant.

Under certain conditions, ϕ_e may be eliminated from (1) and (2) – the resulting equations are known as the monodomain equations [2]. The techniques described here may be applied to the monodomain equations as well as the bidomain equations.

2 Solving the Bidomain Equations Numerically

Computing an accurate solution of the bidomain equations on a realistic three-dimensional computational geometry is a significant computational challenge. This is mainly due to the multiscale nature of the problem – if one computational mesh is used for all dependent variables then this mesh must be sufficiently fine that it captures the fastest processes. Slower processes are then computed at a much higher resolution than is needed. A further complication caused by the multiscale nature of the problem is numerical stability: multiscale processes result in stiff differential equations, and the numerical solution of stiff equations is notoriously prone to numerical instabilities [1].

We begin the description of our numerical algorithm by describing a semi-implicit numerical algorithm for solving the bidomain equations that has good stability properties. We then explain how this algorithm may be adapted to efficiently handle the multiscale processes that are observed when an action potential propagates across cardiac tissue. Finally, we discuss how both the original algorithm and the multiscale modification to this algorithm may be combined to increase the efficiency of simulations without compromising on accuracy.

2.1 The Basic Numerical Algorithm

The semi-implicit algorithm on which this work is based is described in [5] – we only describe it briefly here. The dependent variables V_m, ϕ_e, $\mathbf{u}$ are calculated a collection of discrete times $t_0, t_1, \ldots, t_N$. We write

$$V_m^n(\mathbf{x}) = V_m(\mathbf{x}, t_n), \qquad \phi_e^n(\mathbf{x}) = \phi_e(\mathbf{x}, t_n), \qquad \mathbf{u}^n(\mathbf{x}) = \mathbf{u}(\mathbf{x}, t_n).$$

We discretise the partial differential equations (1) and (2) by treating the conduction terms implicitly and the reaction terms explicitly. This results in the following discretisation in time:

$$\frac{\chi \mathcal{C}_m}{\Delta t_n} V_m^n - \nabla \cdot (\sigma_i \nabla (V_m^n + \phi_e^n)) = \frac{\chi \mathcal{C}_m}{\Delta t_n} V_m^{n-1} + I_{sv_i} - \chi I_{\text{ion}}(V_m^{n-1}, \mathbf{u}^{n-1}), \tag{5}$$

$$\nabla \cdot ((\sigma_i + \sigma_e) \nabla \phi_e^n + \sigma_i \nabla V_m^n) = I_{sv_e}, \tag{6}$$

where $\Delta t_n = t_n - t_{n-1}$. Equations (5) and (6) may be discretised in space using the finite difference method or finite element method, yielding a matrix equation that must be solved on each timestep:

$$A \begin{pmatrix} \mathbf{V_m^n} \\ \boldsymbol{\phi}_\mathbf{e}^\mathbf{n} \end{pmatrix} = \begin{pmatrix} \mathbf{b_v} \\ \mathbf{b_e} \end{pmatrix}, \tag{7}$$

where A is a matrix arising from the discretisation of (5), (6), $\mathbf{V_m^n}$ is a vector of unknowns that arise from the discretisation of V_m^n, $\boldsymbol{\phi}_\mathbf{e}^\mathbf{n}$ is a vector of unknowns that arise from the discretisation of ϕ_e^n, $\mathbf{b_v}$ arises from the right-hand-side of (5), and $\mathbf{b_e}$ arises from the right-hand-side of (6). Note that A is the same on every timestep, and so this matrix need only be computed once at the start of the simulation.

Having solved (7) to compute $\mathbf{V_m^n}$ and $\boldsymbol{\phi}_\mathbf{e}^\mathbf{n}$, we then solve the ordinary differential equations given by (3) using the backward Euler method to ensure stability [1].

2.2 The Multiscale Algorithm

We now briefly describe the multiscale modification of the semi-implicit algorithm described in Sect. 2.1. For more details see [7]. The key to this algorithm is the use of two meshes as shown in Fig. 1. The fine mesh – denoted by thin lines has a nodal spacing h_{fast}, whilst the coarse mesh – denoted by thick lines – has a nodal spacing h_{slow} Rapidly varying quantities are computed using the fine mesh shown in this figure, whilst other variables are computed using the coarser mesh. When required, the variables calculated on the coarse mesh may be interpolated onto the fine mesh.

When using many cardiac electrophysiological models – including the model used in our simulations [4] – the most rapidly varying quantities are those directly related to the fast sodium current I_{Na}, namely the transmembrane potential, V_m, the extracellular potential ϕ_e, the sodium m-gate and the sodium h-gate. We therefore compute these variables on the fine mesh, and all other variables on the coarse mesh. Splitting the right-hand-side of (5) and (6) into quantities computed on the fine mesh and quantities computed on the coarse mesh we have

$$\frac{\chi \mathcal{C}_m}{\Delta t_n} V_m^n - \nabla \cdot (\sigma_i \nabla (V_m^n + \phi_e^n)) = \Gamma_1 + \Gamma_2, \tag{8}$$

$$\nabla \cdot ((\sigma_i + \sigma_e) \nabla \phi_e^n + \sigma_i \nabla V_m^n) = \Gamma_3, \tag{9}$$

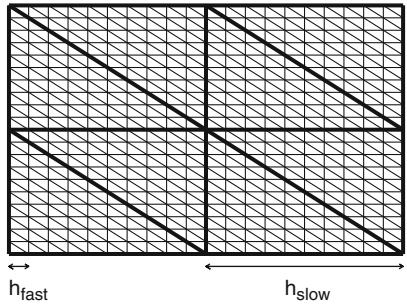


Fig. 1. The two meshes used in the multiscale algorithm. The fine mesh is denoted by *thin lines*, the coarse mesh is denoted by *thick lines*

where

$$\begin{aligned} \Gamma_1 &= \frac{\chi \mathcal{C}_m}{\Delta t_n} V_m^{n-1} + I_{sv_i} - \chi I_{Na}, \\ \Gamma_2 &= -\chi (I_{\text{ion}} - I_{Na}), \\ \Gamma_3 &= I_{sv_e}. \end{aligned}$$

Note that Γ_1 and Γ_3 are fast processes, and Γ_2 is a slow process. This allows us to write (7) as

$$A \begin{pmatrix} \mathbf{V_m^n} \\ \boldsymbol{\phi}_\mathbf{e}^\mathbf{n} \end{pmatrix} = \begin{pmatrix} \mathbf{b_v^{fast}} \\ \mathbf{b_e^{fast}} \end{pmatrix} + \begin{pmatrix} \mathbf{b_v^{slow}} \\ \mathbf{0} \end{pmatrix}, \tag{10}$$

where $\mathbf{b_v^{fast}}$ arises from Γ_1, $\mathbf{b_v^{slow}}$ arises from Γ_2 and $\mathbf{b_e^{fast}}$ arises from Γ_3. We then compute the right-hand-side of (10) on each timestep as follows. $\mathbf{b_v^{fast}}$ and $\mathbf{b_e^{fast}}$ are computed as usual using the fine mesh shown in Fig. 1. The quantities required to calculate $\mathbf{b_v^{slow}}$ are calculated at the nodes of the coarse mesh and then interpolated onto the fine mesh. Equation (10) is solved using a timestep Δt_{fast}. However, as the quantities included in $\mathbf{b_v^{slow}}$ vary on a slower timescale, this vector is updated less frequently using a timestep Δt_{slow}. This results in a significant computational saving – computing $\mathbf{b_v^{slow}}$ is generally the most computationally expensive part of the right-hand-side of (10), and so computing it less often will generate a significant saving in volume of computing.

We now turn our attention to computing the numerical solution of the ordinary differential equations given by (3). Fortunately we only have to compute two of these – those for the sodium m-gate and the sodium h-gate – at each node of the fine mesh. All other quantities are only required at the nodes of the coarse mesh, and so the ordinary differential equations representing these variables need only be solved at the nodes of the coarse mesh. We see in Fig. 1 that there are many fewer nodes in the coarse mesh, thus allowing yet another significant computational saving.

2.3 Combining the Algorithms

The multiscale algorithm described in Sect. 2.2 has been shown to give an increase in computational efficiency of two orders of magnitude with negligible loss of accuracy for a simulation of an action potential wavefront travelling across tissue [7]. However, in many simulations the propagation of the steep action potential wavefront occupies only a small portion of the whole simulation. This has been utilised in [6] where, in the absence of a propagating action potential wavefront, the whole problem is solved with all variables computed on the coarse mesh shown in Fig. 1 and a longer timestep Δt_{slow} using the semi-implicit algorithm described in Sect. 2.1. In this study we combine these algorithms – we use the multiscale algorithm described in Sect. 2.2 when the steep action potential wavefront is propagating across the tissue, and then switch to using the semi-implicit algorithm described in Sect. 2.1 on the coarse mesh shown in Fig. 1 with timestep Δt_{slow} at other times. In the next section we demonstrate the performance of this combined algorithm.

3 Computational Results

In this section we verify the accuracy and efficiency of the algorithm described in Sect. 2.3.

3.1 Description of Simulations

We perform an identical simulation to that used in an earlier study [7]. A square occupying the region $0 < x, y < 20\,\text{mm}$, with fibres running perpendicular to the x-axis, was stimulated at one corner at time $t = 0.001\,\text{s}$. A period of time $0 < t < 0.35\,\text{s}$ was simulated. Intracellular conductivities are $0.13\,\text{mS}\,\text{mm}^{-1}$ along the fibre and $0.026\,\text{mS}\,\text{mm}^{-1}$ perpendicular to the fibre. Extracellular conductivities are $0.13\,\text{mS}\,\text{mm}^{-1}$ along the fibre and $0.065\ \text{mS}\ \text{mm}^{-1}$ perpendicular to the fibre. In common with [7] we use $h_{\text{fast}} = 0.1\,\text{mm}$, $h_{\text{slow}} = 1.0\,\text{mm}$, $\Delta t_{\text{fast}} = 0.1\,\text{ms}$, $\Delta t_{\text{slow}} = 1.0\,\text{ms}$. We use the multiscale algorithm described in Sect. 2.2 initially. In common with [6] we switch to solving the equations on the coarse mesh with timestep Δt_{slow} when the fast sodium current has dropped below $10\ \text{pA}\ \text{pF}^{-1}$ at all points in the computational domain. V_m was recorded at the central point of the square.

To assess the accuracy and efficiency of the algorithm described in Sect. 2.3 the simulation described above was repeated: (a) using the basic algorithm described in Sect. 2.1, the fine mesh and timestep Δt_{fast}; and (b) the multiscale algorithm described in Sect. 2.2.

3.2 Results of Simulations

The action potential at the central point of the square calculated using all three algorithms is shown in Fig. 2. We see that the plots are visually indistinguishable, thus verifying the accuracy of the combined algorithm. We now turn

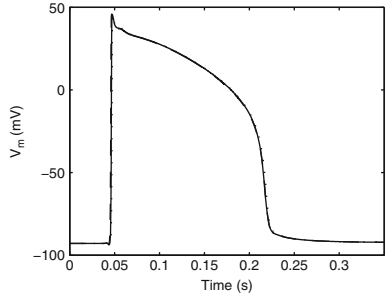


Fig. 2. The action potential at the central point of the square calculated using all three algorithms

our attention to the efficiency of the three algorithms. The basic algorithm using a fine mesh required 13,784 s of computation time. The multiscale algorithm required 786 s of computation time. The combined algorithm required 253 s of computation time. We therefore conclude that, for the simulation presented here, the combined algorithm described in Sect. 2.3 allows an increase in computational efficiency by a factor of roughly 3.

4 Discussion

Two previously published algorithms have been combined, allowing a computational speedup by a factor of around three for the simulation considered here. Although a physiologically detailed electrophysiological cell model [4] was used the geometry was very simple, being a two-dimensional square with regular fibre orientation. Future work is currently being directed towards implementing this algorithm in a realistic, irregular three-dimensional cardiac geometry with irregular fibre orientation.

References

1. Iserles, A.: A First Course in the Numerical Analysis of Differential Equations. Cambridge University Press, Cambridge (1996)
2. Keener, J.P., Sneyd, J.: Mathematical Physiology. Springer, New York (1998)
3. Nickerson, D.P.: Modelling Cardiac Electro-mechanics: From cellML to the Whole Heart. PhD Thesis, University of Auckland (2004)
4. Noble, D., Varghese, A., Kohl, P., Noble, P.: Can. J. Cardiol. **14**, 123–134 (1998)
5. Whiteley, J.P.: IEEE Trans. Biomed. Eng. **53**, 2139–2147 (2006)
6. Whiteley, J.P.: Ann. Biomed. Eng. **35**, 1510–1520 (2007)
7. Whiteley, J.P.: Ann. Biomed. Eng. **36**, 1398–1408 (2008)

Minisymposium *Operational Applications of Data Assimilation*

J.P. Argaud[1,2] and B. Bouriquet[2]

[1] Electricité de France, Research and Development, 1 avenue du Général de Gaulle, 92141 Clamart Cedex, France, jean-philippe.argaud@edf.fr
[2] Sciences de l'Univers au CERFACS, URA CERFACS/CNRS No 1875, 42 avenue Gaspard Coriolis, 31057 Toulouse Cedex, France, bertrand.bouriquet@cerfacs.fr

Data Assimilation is a general set of methods, of various complexities, for computing the optimal estimate of the true state of a system over time. It uses values obtained both from observations and *a priori* models, and information about their errors. Its main improvements have resulted from its wide spread use in meteorological and ocean models applied to weather forecast, although its come originally from control theory.

An important aim of this symposium was to show that there are other fields in science and technology, where the effective use of observed but incomplete data is crucial. Applications in robotics and nuclear sciences, as well as meteorology were featured in the symposium. Indeed Data Assimilation is becoming a cross-domain tool. In this symposium, experts from various fields presented Data Assimilation for specific applications, and compare recent advances and new ideas emerging from their different points of view.

The goal of modern Data Assimilation methods is to make optimal estimates of the initial or time developing state of variables in a model, by putting together information from all available observations and from previous forecasts. Basically, Data Assimilation can be considered as an extension of the least square method. The method of the Best Linear Unbiased Estimation (BLUE) filter is the simplest, both theoretically and computationally. But Data Assimilation is also valid for dynamical or time-varying data sets and models. The Kalman Filter is one of the general methods that provides optimal evaluation of data in relation to a model. Both BLUE and Kalman Filter are very efficient as long as the space of observed data has a limited size, but other methods need to be used for large set of data. Four-dimensional Variational Data Assimilation (4DVAR) based on this principle is currently the most widely used assimilation method for operational weather prediction. Using 4DVAR in meteorology improves accuracy of the forecasts typically by up to a few percent for 2–5 days. Instead of 4DVAR, it is equivalent applying a Kalman Filter to the data as the observation window-length increases.

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*,
Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_58,

As in many complex systems, some (but not all) dynamics of the atmosphere are very sensitive to the initial state, and the model errors. It is helpful that the dynamical equations limit the magnitude of errors in this case, because the equations effectively correlate the variables. The errors and their correlations can be further reduced by ensuring that the initial data, when it is introduced into the processes, satisfies dynamic of balance relationships given by previous equations.

These mathematical and computations procedures for making the best evaluation of incomplete data are not limited to meteorology. An autonomous robot needs to collect as much data as possible to evaluate its position and status in real time. As explained, to build real-time precise representations (maps) of its environment, it uses different filtering methods, and switches from one to another to improve the overall system fault-tolerance. Data Assimilation is applied in the same way as in meteorology and oceanography, but here the technique is called filtering and/or data fusion.

A new application has emerged in the critical field of modelling and evaluating of the core of a nuclear reactor. In this case, Data Assimilation improves operational security and optimal utilisation of resources. Two kinds of applications have been demonstrated. The aim of the first one is to collect information coming from several instruments, in order to estimate the state of the whole system. The second one is to evaluate optimally the parameters of the nuclear core model. In both applications the essential elements is to use data and modelling to make optimal estimates of the required evaluations, and continually reduce error on them.

The Data Assimilation techniques are evolving along time to new techniques in operational computation, and in design. Striking examples are the recent developments using ensemble methods for very large dimension models and for imprecise systems. Data Assimilation improves the reliability of these models, including those where the models for the same process (e.g. weather) are based on different parametrisation, and where many computation are performed simultaneously to allow for noisy predictions in highly non linear processes. At the same time, Data Assimilation schemes are becoming more accurate and faster by incorporating greater physical understanding into the simpler models.

In all fields of science and technology the objectives are to make better and faster use of the increasing volume of measured data, even though it is always incomplete. In order to improve the accuracy of the models with increasing power of computation, Data Assimilation is becoming progressively more essential in computation and observation, and also in control and design of complex systems.

Acknowledgement

The authors thank Prof. Julian Hunt for fruitful discussions and exchanges on Data Assimilation.

Data Fusion in the Navigation of Robots: Assessing Tools

Robin Jaulmes

DGA Paris Expertise Center, 16 bis avenue Prieur de la Cote d'Or, 94114 Arcueil, France, robin.jaulmes@dga.defense.gouv.fr

Summary. Popular methods used for the navigation of robots simultaneously evaluate the localization of the robot and map the environment. Most of these methods are based on the well-known Kalman data fusion filter and its equivalents. Propositions for the evaluation and comparison of implementations of these techniques are presented.

1 Introduction

Simultaneous Localization and Mapping (SLAM) techniques are one of the tools that can be used by robots to navigate in poorly known environments: simultaneously, these techniques map the environment with the robots sensors, and the same map is used to improve the quality of the navigation.

For the moment, no widely accepted methodology exist to assess the performance of SLAM methods, even if public data sets exist [1]. But, in order to measure the progresses made by the scientific community and to normalize the domain, quantitative measurements need to be defined. Our work is a step in this direction: we use quantitative metrics to measure the quality of maps produced by SLAM algorithms on given data sets and compare them.

We present the main mathematic tools that are used in SLAM techniques and give criteria that can be used to specify these techniques for industrial purposes and how these criteria can be evaluated.

2 SLAM Techniques

SLAM techniques can be of various natures [6]: different sensors can be used: monocular cameras, stereo cameras, radars, lasers, ... Several types of map can also be produced: occupation grids, superposition of scans or positions of beacons As maps built from single cameras are hard to read by a user, maps built from telemetric sensors like the classical rotating LIDAR we have on our

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*,
Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_59,

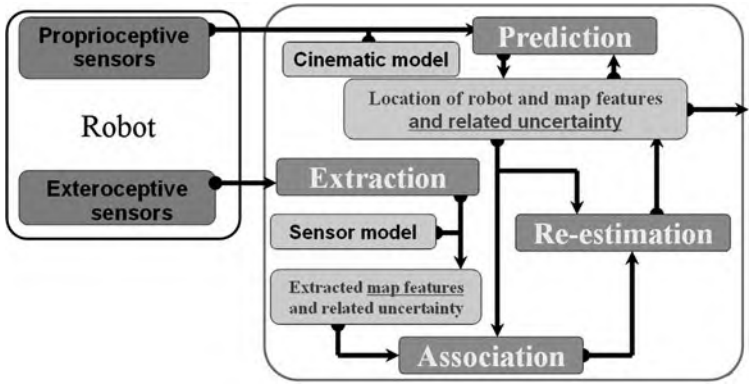


Fig. 1. Main steps of a SLAM technique

robots are easier to read and we have decided to mainly focus on them in our study.

2.1 Principle

Most of the SLAM techniques go through the steps that are illustrated in Fig. 1.

The position of the robot is predicted, according to the given command and to the dynamic model of the robot. Note that these sensors drift and cannot be used alone to navigate. Once the position has been predicted, it is described by a probability distribution. The telemetric measurements are acquired, and compared to the existing (probabilistic) world model. What is seen is associated to what has been seen previously. The position of the robot and the position of the world model are then re-estimated.

The structures used to model the probability distribution on the position of the robot, and on the position of the features of the world, and the equations used to predict and estimate these probability distributions are the elements which change between two SLAM techniques: usually, the equations of classical navigation/data fusion filters are used and the techniques differ by implementation choices.

2.2 The Kalman Filter

Let $\mathbf{X}_k$ be the true state at time k: it contains a set of variables that describe the position and attitude of the robot and all the variables that describe the map being built (for instance, the position of environment landmarks).

One of the main filter used to solve the SLAM problem is the Kalman filter. This filter uses a Bayesian assumption: the true state $\mathbf{X}_k$ at time k is evolved from the state at $k-1$ according to:

$$\mathbf{X}_k = \mathbf{F}_k \mathbf{X}_{k-1} + \mathbf{B}_k \mathbf{u}_k + \mathbf{w}_k$$

where $\mathbf{F}_k$ is the state transition model, which is a function of δt, the time interval between times $k-1$ and k.

$\mathbf{B}_k$ is the control input model applied to $\mathbf{u}_k$ and $\mathbf{w}_k$ is the noise, drawn from a zero mean multivariate normal distribution with covariance $\mathbf{Q}_k$.

At time k an observation $\mathbf{z}_k$ of the true state $\mathbf{x}_k$ is made according to:

$$\mathbf{z}_k = \mathbf{H}_k \mathbf{x}_k + \mathbf{v_k}$$

where $\mathbf{H}_k$ is the observation model which maps the true state space into the observed space (which can include measurements from all the sensors), and $\mathbf{v}_k$ is the observation noise which is also assumed to be drawn from a zero mean multivariate normal distribution with covariance $\mathbf{R}_k$.

The state of the filter is represented by $\hat{\mathbf{x}}_{k|k}$ which is the estimate of the state at time k given the initial prior and all the observations up to time k and by $\mathbf{P}_{k|k}$, which is the current error covariance matrix and an indicator of the precision of the estimation.

At each time step, the state of the filter is updated according to the following rules:

$$\hat{\mathbf{x}}_{k|k-1} = \mathbf{F}_k \hat{\mathbf{x}}_{k-1|k-1} + \mathbf{B}_k \mathbf{u}_k$$

$$\mathbf{P}_{k|k-1} = \mathbf{F}_k \mathbf{P}_{k-1|k-1} \mathbf{F}'_k + \mathbf{Q}_{k-1}$$

$$\mathbf{K}_k = \mathbf{P}_{k|k-1} \mathbf{H}'_k [\mathbf{H}_k \mathbf{P}_{k|k-1} \mathbf{H}'_k + \mathbf{R}_k]^{-1}_k$$

$$\hat{\mathbf{x}}_{k|k} = \hat{\mathbf{x}}_{k|k-1} + \mathbf{K}_k (\mathbf{z}_k - \mathbf{H}_k \hat{\mathbf{x}}_{k|k-1})$$

$$\mathbf{P}_{k|k} = (I - \mathbf{K}_k \mathbf{H}_k) \mathbf{P} k|k-1$$

2.3 EKF

Note that we have here assumed that the state transition models and the observation model were linear, which is rarely actually the case.

If we have $\mathbf{x}_k = f(\mathbf{x}_{k-1}, \mathbf{u}_k) + \mathbf{w}_k$ and $\mathbf{z}_k = h(\mathbf{x}_k) + \mathbf{v}_k$

Let $\mathbf{F}_k = \frac{\partial f}{\partial \mathbf{x}}|_{\hat{\mathbf{x}}_{k-1|k-1}, \mathbf{u}_k}$ and $\mathbf{H}_k = \frac{\partial h}{\partial \mathbf{x}\, \hat{\mathbf{x}}_{k|k-1}}$. If we keep the update equations if the Kalman filter, we obtain the Extended Kalman Filter, which is often used but lack theoretical optimality.

2.4 Particle Filtering

Another solution that can be used when state transition models and observation models are not linear is to use a sampling based approach.

The principle of these approach is that we maintain P estimates of the state of the environment. Each of these estimates is updated by sampling a value for the noise in the evolution equation of the system. The following equation, that becomes exact if each particle is drawn according to the true probability distribution and if there is an infinite number of particles, is exploited:

$$E(f|k) \simeq \frac{1}{P} \sum_{i=1}^{P} f(w_k(i))$$

However, it is not an infinite number of particle that is used. Furthermore, each particle is associated to a weight, that is representative of the likelihood of the hypothesis. The update rule for the weights is the following:

$$w_k(i) = w_{k-1}(i) \frac{p(y_k|x_k(i)p(x_k(i)|x_{k-1}(i))}{\prod(x_k(i)|x_{0:k-1}, y_{0:k}))}$$

This update is followed by a normalization of all the weights. When particles have very low weights, they are replaced by new particles. The rules that decide when and how to replace particles often vary between two implementations of the filter, along with the number of particle that is used.

2.5 FastSLAM

Many variations and combinations in the mentioned techniques exist. A popular method is FastSLAM [5] that combines particle filtering for estimating the path of the robot and small Kalman filters for the estimation of the landmarks of the map.

3 Assessment

Classical methods that assess the quality of the functions of robotics systems [4] and that are also used in the image processing and speech processing fields involve:

- Defining the evaluation criteria: the metrics are defined according to these criteria, as are the contents of the data set to build.
- Building the data set, that needs to include ground truth data that allows the evaluation.
- Measuring the performance of the evaluated method on the data set, for different set of parameters.

The criteria we investigated to evaluate SLAM techniques are the following:

- Processing time.
- Allocated resources (processors, memory).
- Precision of the localization (and drift).
- Precision of the produced map.
- Robustness to noise in the wheel encoders.
- Robustness to noise in the LIDAR scans.
- Ability to work in cluttered environments.
- Ability to correctly map loops.
- The *pertinence* of the estimated error with respect to the real error could also be measured.

Assessing the precision of the localization brought by SLAM can be done like this: the error between the real position and the estimated position is measured [7], the drift being this error divided by the distance travelled. However, assessing the quality of the produced map is harder and no quantitative approach exists yet.

We defined the metric Q to measure between 0 and 1 the quality of a produced map. N control points are defined at every corner of the ground truth map. The produced map is scaled and superposed to the ground truth map, and the error on the position of each ground truth point d_i is measured (as showed in Fig. 2). Let W be the width of the ground truth map and H be its height; let k be a constant. Then,

$$Q = \exp\left(-\frac{k}{\sqrt{W^2 + H^2}} \frac{1}{N} \sum i = 1N d_i\right)$$

Note that in order to measure Q, an operator has to designate, for each produced map, the position of each of the control point. When two or more produced points correspond to the same point on the ground truth map (which happens in poorly mapped loops), the farthest point is selected, and we also record the maximum distance between them, which is an indicator of the

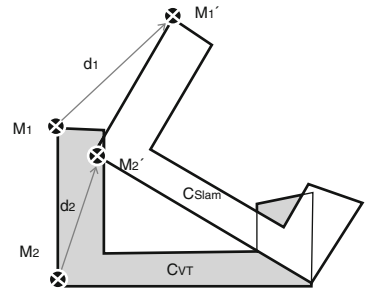


Fig. 2. Notations for the definition of Q

poorness of the algorithm to map loops. When no corresponding point can be found for one of the control points, $Q = 0$.

We have defined a user interface to do a systematic assessment of the quality of the produced map.

To assess the robustness to noise in the wheel encoders, we recommend adding several times on one of the simulated log a controlled noise. We have added an additive Gaussian noise of standard deviation σ at each step for heading and movement, and from one experiment to the following we have increased the value of σ.

We also recommend the use of specific maps to assess the abilities to map in cluttered environment and to correctly map loops. Maps of the public robotic data repository website [9] can also be used for comparison.

4 Discussion and Conclusion

We have presented the main data fusion methods that are used for Simultaneous Localization and Mapping and presented a method that allows the assessment of the performance of these techniques: the method we have used is coherent with what has been written about the topic in the literature [1–3].

The European project RAWSEEDS [10] that develop a benchmark data set should allow a good use of our proposed method and allow the comparison of SLAM techniques that can use a whole array of sensors.

References

1. Amigoni, F., Gasparini, S., Gini, M.: Good Experimental Methodologies for Robotic Mapping: A Proposal, GEMBENCH 2008
2. Collins, T., Collins, J.J., Ryan, C.: Occupancy Grid Mapping: An Empirical Evaluation, MED 2007
3. Fontana, G., Matteucci, M., Neira, J., Sorrenti, D.: Proposals for Benchmarking SLAM, GEMBENCH 2008
4. Lambert, M., Jaulmes, R., Godin, A., Molin, E., Dufourd, D.: A methodology for Assessing Robot Autonomous Functionalities, IAV 2007
5. Montemerlo, M., Thrun, S.: FastSLAM: A Factored Solution to the Simultaneous Localization and Mapping Problem, AAAI 2002
6. Thrun, S.: Robotic Mapping: A Survey, Exploring Artificial Intelligence in the New Millenium, Morgan Kaufmann 2002
7. Wulf, O., Nuchter, A., Hertzberg, J., Wagner, B.: Ground Truth Evaluation of Large Urban 6D SLAM, IROS 2007
8. Stachniss, C., Frese, U., Grisetti, G. Since January 2007. URL http://openslam.org/
9. Howard, A., Roy, N., Vincent, R., Fox, D., Vaughan, R. Since May 2003. URL http://radish.sourceforge.net

10. Ceriani, S., Fontana, G., Giusti, A., Marzorati, D., Matteucci, M., Migliore, D., Rizzi, D., Sorrenti, D.G., Taddei, P.: Rawseeds Ground Truth Collection Systems for Indoor Self-localization and Mapping: Autonomous Robots, **27**(4), 353–371, (2009). URL http://rawseeds.elet.polimi.it/home/

The Role of Balance in Data Assimilation

R.N. Bannister

Department of Meteorology, University of Reading, Reading RG6 6BB, UK,
r.n.bannister@reading.ac.uk

Summary. Data assimilation estimates the initial conditions of a weather forecast model by bringing together data from observations, and a forecast from a previously known atmospheric state. The forecast error covariance matrix is part of the assimilation and is very important in the way that the assimilation treats the data. This article shows how these error covariances for large-scale weather systems are represented using balance relationships. An example of how this method can be improved at large-scale is introduced, and contemporary issues are raised concerning how it can be adapted to model error covariances of small-scale phenomena, such as convection, where the balance approach breaks down.

1 Introduction

The motion of the atmosphere is well described by a set of prognostic momentum, energy and moisture equations on a rotating sphere, e.g. [7]. These equations allow the state of the atmosphere 'tomorrow' to be predicted providing that we know the state of the atmosphere 'today'. This is a classic initial value problem. These equations provide the basis of the forecast procedure, but when reduced to a so-called 'balanced' form, also guide how the initial conditions can be determined. The balanced form of the equations leads to a set diagnostic equations called balance relations, and the process of determining accurate and useful initial conditions consistent with measurements of the atmosphere is called data assimilation. This paper introduces the balance relations used in Meteorology to describe the large-scale atmosphere (Sect. 2) and shows how they are used in the data assimilation problem as it is solved at many weather forecasting centres (Sect. 3). This conventional data assimilation problem has its limitations and can be refined to help better solve the large-scale atmospheric data assimilation problem (Sect. 4). A modified approach, however, is being sought to tackle the small-scale data assimilation problem required by new high-resolution weather forecasting models, where the usual balance relations break down (Sect. 5).

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*,
Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_60,

2 Balance Relations Between Meteorological Variables

A balance relation is a diagnostic equation linking Meteorological variables, and under conditions that it is valid, describes a slowly evolving component of the weather. There is a hierarchy of balance relations in geophysical fluid dynamics [9] and each is approximate to a different order of accuracy, but here we use the simplest. The essential parts of the two horizontal and one vertical momentum equations are as follows

$$\frac{\mathrm{d}u}{\mathrm{d}t} = fv - \frac{\sec\phi}{\rho a}\frac{\partial p}{\partial \lambda} - D_\lambda, \quad \frac{\mathrm{d}v}{\mathrm{d}t} = -fu - \frac{1}{\rho a}\frac{\partial p}{\partial \phi} - D_\phi, \quad \frac{\mathrm{d}w}{\mathrm{d}t} = -\frac{1}{\rho}\frac{\partial p}{\partial z} - g - D_z, \tag{1}$$

where u, v are the horizontal components and w is the vertical component of the wind, p is pressure, ρ is density, $\mathrm{d}/\mathrm{d}t$ is Lagrangian derivative, λ, ϕ and z are longitude, latitude and height, $f \equiv 2\Omega \sin\phi$ is the Coriolis parameter (accounting for the Earth's rotation where Ω is the Earth's rotation rate), a is the Earth's radius, g is the acceleration due to gravity, and $D_{\lambda,\phi,z}$ are drag forces in each direction. By defining characteristic values U, W, L, H and P for horizontal wind, vertical wind, horizontal lengthscale, vertical lengthscale and pressure respectively, and the Coriolis parameter at a representative midlatitude, f_0, the variables can be scaled as follows: $u = U\tilde{u}$, $v = U\tilde{v}$, $w = W\tilde{w}$, $\mathrm{d}/\mathrm{d}t = U/L\mathrm{d}/\mathrm{d}\tilde{t}$, $\delta p = P\delta\tilde{p}$, $\delta\lambda = (L\sec\phi/a)\delta\tilde{\lambda}$, $\delta\phi = (L/a)\delta\tilde{\phi}$, and $\delta z = H\delta\tilde{z}$. Variables with a tilde are then of order unity and (1) become

$$Ro\frac{\mathrm{d}\tilde{u}}{\mathrm{d}\tilde{t}} = \frac{f}{f_0}\tilde{v} - \frac{P}{f_0\rho U L}\frac{\partial\tilde{p}}{\partial\tilde{\lambda}} - \frac{D_\lambda}{f_0 U}, \quad Ro\frac{\mathrm{d}\tilde{v}}{\mathrm{d}\tilde{t}} = -\frac{f}{f_0}\tilde{u} - \frac{P}{f_0\rho U L}\frac{\partial\tilde{p}}{\partial\tilde{\phi}} - \frac{D_\phi}{f_0 U}, \quad Ro\frac{W}{U}\frac{\mathrm{d}\tilde{w}}{\mathrm{d}\tilde{t}} = -\frac{P}{f_0\rho U H}\frac{\partial\tilde{p}}{\partial\tilde{z}} - \frac{g}{f_0 U} - \frac{D_z}{f_0 U}, \tag{2}$$

where $Ro \equiv U/f_0 L$ is the dimensionless Rossby number. For mid-latitude and large-scale flow, $Ro \sim \mathcal{O}(10^{-1})$ and $W/U \sim \mathcal{O}(10^{-2})$. This scaling justifies a balanced form of (1) by neglecting the Lagrangian derivatives since they scale as Ro. Assuming that the drag terms may also be neglected gives

$$0 = f\rho v - \frac{\sec\phi}{a}\frac{\partial p}{\partial\lambda}, \quad 0 = -f\rho u - \frac{1}{a}\frac{\partial p}{\partial\phi}, \quad 0 = -\frac{\partial p}{\partial z} - \rho g. \tag{3}$$

The first two equations describe a balance between the Coriolis terms and horizontal pressure gradients (called geostrophic balance), and the third describes a vertical balance between gravity and vertical pressure gradients (called hydrostatic balance). Forms of these equations more convenient to data assimilation are found by writing (a) the horizontal divergence of the geostrophic

equations and (b) the hydrostatic equations in terms of potential temperature, θ (θ is an adiabatically conserved form of temperature, T, $\theta = T/\Pi$, where Π is exner pressure, $\Pi = (p/1{,}000\text{hPa})^{\kappa}$, and $\kappa = 0.286$)

$$\nabla_{\mathrm{h}}^2 \delta p = \nabla_{\mathrm{h}} \cdot (f \rho_0 \nabla_{\mathrm{h}} \delta \psi), \tag{4}$$

$$\delta\theta = -\theta_0 \left(\frac{\partial \Pi_0}{\partial z} \right)^{-1} \frac{\partial}{\partial z} \left(\frac{\kappa \Pi_0}{p_0} \delta p \right). \tag{5}$$

Here ∇_{h} is the horizontal differential operator and ψ is stream function (see Sect. 3). The equation of state, $p = R\rho\Pi\theta$ (where R is the specific gas constant), has been used in (5), and both (4) and (5) are expressed in perturbation form and linearized about reference state variables (subscript 0).

3 Balance in Meteorological Data Assimilation

Meteorological data assimilation is the process of determining the best possible set of initial conditions for a forecast model which are consistent with the available observations, with the equations of motion (e.g. (1)) and with prior information provided in the form of a short forecast started from a known state at an earlier time. The initial conditions must also be close to a balanced state to avoid spurious unbalanced modes disturbing the forecast quality. Currently four-dimensional variational data assimilation (4D-VAR) is the method of choice for large-scale weather prediction [10].

4D-VAR finds the initial conditions, represented by the meteorological fields $\mathbf{x} = (u, v, p, \theta)$ (w is missing because it can be diagnosed from u and v via the continuity equation and in reality a humidity variable is included), by minimizing a cost function, J, that measures the distance between the 'model observations' predicted by $\mathbf{x}$ and the actual observations, and between $\mathbf{x}$ and the short forecast, denoted $\mathbf{x}_{\mathrm{b}}$ (the short forecast is sometimes called a background state). Distances are measured with respect to error covariance matrices, which play a very important role in 4D-VAR, and it is the application of the balance relations to this problem that helps to define the error covariance matrix that measures the departure from $\mathbf{x}_{\mathrm{b}}$, denoted $\mathbf{B}_x$. This part of J is J_{b}

$$J_{\mathrm{b}}[\delta\mathbf{x}] = \frac{1}{2} \delta\mathbf{x}^{\mathrm{T}} \mathbf{B}_x^{-1} \delta\mathbf{x}, \tag{6}$$

and an observation term is added to J_{b} to give J. In (6), $\delta\mathbf{x}$ is the departure from $\mathbf{x}_{\mathrm{b}}$ of the state to be determined, i.e. $\mathbf{x} = \mathbf{x}_{\mathrm{b}} + \delta\mathbf{x}$. $\mathbf{B}_x$ is a large matrix which contains multivariate covariances that account for the near balance between the incremental fields in $\delta\mathbf{x}$. The method used to model these covariances using (4) and (5) reformulates (6) in terms of new fields in $\delta\boldsymbol{\chi}_1$

$$J_{\mathrm{b}}[\delta\boldsymbol{\chi}_1] = \frac{1}{2} \delta\boldsymbol{\chi}_1^{\mathrm{T}} \mathbf{B}_{\chi_1}^{-1} \delta\boldsymbol{\chi}_1, \tag{7}$$

where $\delta\boldsymbol{\chi}_1 = (\delta\psi, \delta\chi, \delta p_\mathrm{r})$. Fields $\delta\psi$ and $\delta\chi$ represent the rotational and divergent parts of δu and δv via the Helmholtz relation, i.e. $(\delta u, \delta v) = \mathbf{k} \times \nabla_\mathrm{h}\delta\psi + \nabla_\mathrm{h}\delta\chi$ (where $\mathbf{k}$ is the unit vector normal to the sphere), and δp_r is the residual pressure that is not in balance with $\delta\psi$ (see below). $\mathbf{B}_{\chi_1}$ is the error covariance matrix of these new variables; it is taken to be univariate meaning that it is a block diagonal matrix, and so is thus simpler to represent than $\mathbf{B}_x$. Let $\delta\mathbf{x}$ and $\delta\boldsymbol{\chi}_1$ be related via the linear transform $\mathbf{K}_1$ as follows

$$\begin{gathered}\delta\mathbf{x} = \mathbf{K}_1\delta\boldsymbol{\chi}_1 \\ \begin{pmatrix}\delta u\\ \delta v\\ \delta p\\ \delta\theta\end{pmatrix} = \begin{pmatrix}-(1/a)\partial/\partial\phi & (\sec\phi/a)\partial/\partial\lambda & 0\\ (\sec\phi/a)\partial/\partial\lambda & (1/a)\partial/\partial\phi & 0\\ \mathbf{H} & 0 & 1\\ \boldsymbol{\Theta}\mathbf{H} & 0 & \boldsymbol{\Theta}\end{pmatrix}\begin{pmatrix}\delta\psi\\ \delta\chi\\ \delta p_\mathrm{r}\end{pmatrix},\end{gathered} \tag{8}$$

where $\mathbf{H}\delta\psi \equiv \nabla_\mathrm{h}^{-2}\nabla_\mathrm{h}\cdot(f\rho_0\nabla_\mathrm{h}\delta\psi)$ and $\boldsymbol{\Theta}\delta p \equiv -\theta_0(\frac{\partial\Pi_0}{\partial z})^{-1}\frac{\partial}{\partial z}(\frac{\kappa\Pi_0}{p_0}\delta p)$ are the linear balance and hydrostatic operators (4) and (5). The first two rows give δu and δv using the Helmholtz relation (see above), expanded in spherical geometry. In (8) the winds are not affected by δp_r. The third row gives δp as a sum of a part that is in linear balance with $\delta\psi$, $\mathbf{H}\delta\psi$, and a residual, δp_r. The fourth line is the hydrostatic balance operator, $\boldsymbol{\Theta}$ acting on this combined pressure. In this formulation, geostrophic balance is applied weakly – as a (geostrophically) 'unbalanced' pressure is allowed (δp_r) – but hydrostatic balance is applied strongly – as no (hydrostatically) unbalanced potential temperature is allowed.

Equations (6) and (7) are equivalent representations of the same problem where $\mathbf{B}_x = \mathbf{K}_1\mathbf{B}_{\chi_1}\mathbf{K}_1^\mathrm{T}$. Thus, given the transform (8), and the given form of $\mathbf{B}_{\chi_1}$ (in this case block diagonal), minimizing the cost function whose J_b term is given by (7) implies this $\mathbf{B}_x$, even though it is not explicitly calculated. This methodology is regarded as a way of modelling the $\mathbf{B}_x$-matrix using such a change of variables [6]. The remaining task is to determine the $\mathbf{B}_{\chi_1}$-matrix, which may be done by analysing a sample population of estimated errors of the atmosphere to find the error covariances within each field of $\delta\boldsymbol{\chi}_1$, e.g. [3]. This matrix is usually static and so does not change from day-to-day.

4 Refining the Method for Large-Scale Meteorological Systems

A transform similar to (8) is used for operational data assimilation (e.g. [5,8]), which makes the assumption that the first field in $\delta\boldsymbol{\chi}_1$, $\delta\psi$ is a wholly 'balanced' field since it is used with $\mathbf{H}$ to calculate the balanced pressure. Although, $\delta\psi$ is *largely* balanced, the assumption that it is *wholly* balanced is not supported by theory which expects $\delta\psi$ to acquire an unbalanced component in some dynamical regimes, namely those associated with large horizontal

and small vertical scales [11]. Work is underway to account for this by making modifications to (8) [1], which builds on earlier work [4].

The replacement for $\delta\boldsymbol{\chi}_1$ is $\delta\boldsymbol{\chi}_2$, which (a) replaces $\delta\psi$ with its balanced component only, $\delta\psi_{\mathrm{b}}$, and (b) replaces δp_{r} with a 'true' unbalanced pressure, δp_{u}. The replacement for $\mathbf{K}_1$ is $\mathbf{K}_2$, turning (8) into

$$\delta\mathbf{x} = \mathbf{K}_2\delta\boldsymbol{\chi}_2$$
$$\begin{pmatrix}\delta u\\ \delta v\\ \delta p\\ \delta\theta\end{pmatrix} = \begin{pmatrix}-(1/a)\partial/\partial\phi & (\sec\phi/a)\partial/\partial\lambda & -(1/a)\partial/\partial\phi\,\bar{\mathbf{H}}\\ (\sec\phi/a)\partial/\partial\lambda & (1/a)\partial/\partial\phi & (\sec\phi/a)\partial/\partial\lambda\,\bar{\mathbf{H}}\\ \mathbf{H} & 0 & 1\\ \boldsymbol{\Theta}\mathbf{H} & 0 & \boldsymbol{\Theta}\end{pmatrix}\begin{pmatrix}\delta\psi_{\mathrm{b}}\\ \delta\chi\\ \delta p_{\mathrm{u}}\end{pmatrix}. \quad (9)$$

Elements of $\mathbf{K}_2$ are the same as those of $\mathbf{K}_1$ except for two new elements which allow δp_{u} to contribute to δu and δv. These work by first calculating from δp_{u} an unbalanced streamfunction (using the new 'anti-balance operator' $\bar{\mathbf{H}}$) and then using the standard Helmholtz operators to compute δu and δv from this unbalanced streamfunction. The form of $\bar{\mathbf{H}}$ is outlined in other works, [1, 4]. As is evident, these unbalanced winds are ignored in the standard implementation, but this new balanced/unbalanced partitioning of fields allows the linear balance operator, $\mathbf{H}$ to act (now appropriately) with the balanced field $\delta\psi_{\mathrm{b}}$. The forecast error covariance matrix for $\delta\boldsymbol{\chi}_2$ is $\mathbf{B}_{\chi_2}$. Then $\delta\boldsymbol{\chi}_2$ and $\mathbf{B}_{\chi_2}$ replace $\delta\boldsymbol{\chi}_1$ and $\mathbf{B}_{\chi_1}$ in (7), and will lead to the new refined implied forecast error covariance matrix for model variables $\mathbf{B}_x = \mathbf{K}_2\mathbf{B}_{\chi_2}\mathbf{K}_2^{\mathrm{T}}$.

5 Data Assimilation for High-Resolution Model Forecasts

Weather forecasting models are now applied at high-resolution and are capable of resolving features that have a horizontal lengthscale of a few km. These features are small compared to the lengthscales of large-scale weather systems of many hundreds of km. Additionally, the models are capable of resolving convective storms which are associated with relatively fast vertical winds. These modelling capabilities mean that for these features, Ro and W/U are no longer expected to be small as they must be to justify (3). A new 4D-VAR data assimilation system is being sought to find accurate initial conditions for such high-resolution models. Such a system requires an appropriate forecast error covariance matrix, but unlike the assimilation for large-scale weather systems using (8) or (9), one cannot rely on the application of balance relations like (4) and (5) in its formulation [2]. Important challenges relating to balance issues that need to be overcome include the following:

- Following the change of variable formulation used above, a set of new incremental fields needs to be chosen (i.e. in $\delta\boldsymbol{\chi}_3$) that allow an appropriate description of the small-scale flow. These should be largely uncorrelated

(allowing the use of a block-diagonal $\mathbf{B}_{\chi_3}$) and should not rely on balance relations (4) and (5) holding in their version of the transform, $\mathbf{K}_3$.
- The new implied forecast error covariance matrix, $\mathbf{B}_x = \mathbf{K}_3\mathbf{B}_{\chi_3}\mathbf{K}_3^{\mathrm{T}}$, must change from day-to-day (flow-dependency) since the nature of error covariances at small-scales is expected to be more changeable than for the large-scale flow. This may be modelled with a flow-dependent $\mathbf{K}_3$ transform and/or a $\mathbf{B}_{\chi_3}$-matrix that changes from day-to-day.
- Large-scale features will still remain even in a high-resolution model that permits small-scale features. This means that the conventional (balance-based) forecast error covariance modelling techniques of Sects. 3 and 4 should still apply for the large-scale part of the flow.

How such a forecast error covariance matrix can be realistically modelled is an unsolved problem in Meteorology, but is the focus of current research.

6 Summary

Data assimilation is an essential part of the numerical weather forecasting problem, which finds the initial conditions by finding the state that is most consistent with observations and a forecast which had started from an earlier time. The cost function, which is minimized in 4D-VAR, relies on a realistic error covariance matrix which describes the uncertainty associated with the earlier forecast. For large-scale flow, this matrix may be constructed with the aid of geostrophic and hydrostatic balance relationships, which have been reviewed in this article. The standard way of achieving this, by making a change of variables, has been shown, together with a refinement of the method that allows the geostrophic balance relation to be applied in a more appropriate way. Issues that arise from the need to do data assimilation for high-resolution models, where the balance relations are not appropriate, have been discussed.

Acknowledgements

This work has benefited from discussions with Mike Cullen and Mark Dixon of the Met Office, and Stefano Migliorini of the University of Reading. The author is supported by the Natural Environment Research Council under the National Centre for Earth Observation.

References

1. Bannister, R.N., et al.: Int. J. Numer. Meth. Fluids **56**, 1147–1153 (2008)
2. Berre, L.: Mon. Wea. Rev. **128**, 644–667 (2000)
3. Berre, L.: Tellus **58A**, 196–209 (2006)

4. Cullen, M.J.P.: Q.J.R. Meteorol. Soc. **129**, 2777–2796 (2003)
5. Derber, J., Bouttier, F.: Tellus **51A**, 195–221 (1999)
6. Fisher, M.: ECMWF Semin. Proc. 45–64 (2003)
7. Holton, J.R.: An Introduction to Dynamic Meteorology, 4th edn. Elsevier Academic, San Diego (2004)
8. Lorenc, A., et al.: Q.J.R. Meteorol. Soc. **126**, 2991–3012 (2000)
9. McIntyre, M.E.: Encyclopaedia of Atmospheric Sciences, pp. 680–685. Academic, London (2003)
10. Rabier, F.: Q.J.R. Meteorol. Soc. **131**, 3215–3233 (2005)
11. Wlasak, M.A., Nichols, N.K., Roulstone I.: Q.J.R. Meteorol. Soc. **132**, 2867–2886 (2006)

Data Assimilation in Nuclear Power Plant Core

J.P. Argaud[1,2], B. Bouriquet[2], P. Erhard[1], S. Massart[2], and S. Ricci[2]

[1] Electricité de France, 1 av. du Général de Gaulle, 92141 Clamart Cedex, France
jean-phillippe.argaud@edf.fr
[2] Sciences de l'Univers au CERFACS, URA CERFACS/CNRS No 1875, 42 av. Gaspard Coriolis. 31057 Toulouse Cedex, France
bertrand.bouriquet@cerfacs.fr

Summary. The use of Data Assimilation is fairly recent in the field of nuclear core modelling. This paper is focused on field reconstruction and parameters estimation, based on the standard simulation of neutron fields which are already very accurate. In one application, these methods are used to investigate how to gather information coming from several instruments and to evaluate the impact of instrument loss. In a second application, it leads to best parameter estimation through processing a large set of data, with the aim of improving the simulation for upcoming nuclear core.

1 Introduction

This paper is a review of data assimilation applications in nuclear core physics. In a first part, we will describe the context of the data assimilation in nuclear core physics. Here we insist specifically on the core physics, data assimilation basis and core instrumentation. Then a second part will be focused more precisely on some industrial applications of data assimilation.

2 Data Assimilation in Nuclear Core

2.1 The Physics of Nuclear Core in a Power Plant

A key consideration for nuclear core physics in a power plant is the description of neutron density inside the core. The neutrons are produced during uranium fission process, induced by neutrons themselves. The aim of simulation is to describe the neutron density, which must remain in a steady state, in each configuration and in the whole domain. To ensure the critical steady state of the core, neutrons are slowed down (moderated) in order to have an optimum reaction with the uranium. Generally the material used for this moderation (moderator) is water. At the same time, the water is used as a transportation medium of the thermal energy produced in the core. As a consequence, the temperature of the material has to be known everywhere inside the core.

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*,
Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_61,

A real nuclear core is a very complex medium, including uranium fuel, metals, mechanical frame of the core, fission products, moderator... Moreover, within the core some strong feedbacks between the various physical phenomena exist. The basic example is the moderator temperature, that is driven by the neutron density, itself driven by the temperature of the moderator. All those effects have to be taken into account in the modelling in order to reach the required simulation accuracy. To summarise, core physics is a multi-physics problem involving, among others, nuclear reaction physics, thermohydraulics, heat transfer physics and thermics.

2.2 Data Assimilation

The main goal of data assimilation is to find the true state, denoted $\mathbf{x_t}$, of an observable system (see [4] for an introduction). However, this true state cannot be determined. Thus data assimilation proposes to combine all the system information, in order to obtain the "*best-estimated true state*". The information about the system can come from measurements as well as from the simulation. All the available information on the system is stored in a vector $\mathbf{z}$. These data are uncertain, and can be estimated as a function of $\mathbf{x_t}$ as:

$$\mathbf{z} = \mathbf{\Gamma x_t} + \xi \tag{1}$$

where $\mathbf{\Gamma}$ is the transformation operator from the state space to the data space, and ξ is the error in the data space. In a least square (LS) sense, the optimal true state $\mathbf{x_t}$ is the one that minimises the following objective function:

$$J_{LS}(\mathbf{x}) = [\mathbf{\Gamma x} - \mathbf{z}]^{\mathbf{T}}[\mathbf{\Gamma x} - \mathbf{z}] \tag{2}$$

The uncertainties ξ can be characterised by a distribution for each component and by a covariance matrix $\mathbf{S}$. To use this information, a new objective function can be written:

$$J(\mathbf{x}) = [\mathbf{\Gamma x} - \mathbf{z}]^{\mathbf{T}}\mathbf{S}^{-\mathbf{1}}[\mathbf{\Gamma x} - \mathbf{z}] \tag{3}$$

When $\mathbf{\Gamma}$ is linear, the minimum value of the J function is called the BLUE (for *Best Linear Unbiased Estimator*) and denoted $\mathbf{x_{BLUE}}$.

As quoted before, the system information $\mathbf{z}$ can be split in two distinct parts: one from the observation $\mathbf{y_o}$, and another one coming from everything else (for example the model), named *background*, and denoted $\mathbf{x_b}$. As a vector, it can be denoted as $\mathbf{z} = (\mathbf{x_b}\ \mathbf{y_o})^{\mathbf{T}}$. In consequence, the $\boldsymbol{\Gamma}$ operator can be rewritten as $\mathbf{\Gamma} = (\mathbf{I}\ \mathbf{H})^{\mathbf{T}}$, where $\mathbf{I}$ is the identity matrix, and $\mathbf{H}$ the matrix associated to the observation operator H. If H is non-linear, $\mathbf{H}$ is the tangent linear matrix associated to H. Assuming background and observation to be independent, the covariance matrix $\mathbf{S}$ can be written as follows:

$$\mathbf{S} = \begin{pmatrix} \mathbf{B} & 0 \\ 0 & \mathbf{R} \end{pmatrix}, \tag{4}$$

where $\mathbf{B}$ is the background-error covariance matrix, and $\mathbf{R}$ the observation-error covariance matrix. Finally, the objective function J is then:

$$J(\mathbf{x}) = [\mathbf{x} - \mathbf{x_b}]^{\mathbf{T}}\mathbf{R}^{-1}[\mathbf{x} - \mathbf{x_b}] + [\mathbf{y_o} - \mathbf{H}.\mathbf{x}]^{\mathbf{T}}\mathbf{B}^{-1}[\mathbf{y_o} - \mathbf{H}.\mathbf{x}] \tag{5}$$

In order to minimise this function, we can calculate the gradient ∇J of J with respect to $\mathbf{x}$. It can formally be written as follows:

$$\nabla J(\mathbf{x}) = \mathbf{2R^{-1}x} + \mathbf{2H^{T}B^{-1}H}.\mathbf{x} \tag{6}$$

The minimum of J is obtained when the required (but not sufficient) condition $\nabla J(\mathbf{x}) = \mathbf{0}$ is true, for one particular $\mathbf{x}$. This minimising value is named the analysis and denoted $\mathbf{x_a}$. It can be obtained either by a direct calculation or by a minimisation procedure. In the direct case, the solution is:

$$\mathbf{x_a} = \mathbf{x_b} + \mathbf{K}(\mathbf{y_o} - \mathbf{H}.\mathbf{x_b}) \tag{7}$$

with the linear operator $\mathbf{K}$ (the "*gain*" of the analysis, or the "*Kalman gain*"):

$$\mathbf{K} = \mathbf{BH^{T}}(\mathbf{HBH^{T}} + \mathbf{R})^{-1} \tag{8}$$

This expression (8) represents the Kalman filter in the simplest static case. More details on this could be find for example in reference [1].

We can make some interesting remarks to better interpret the equation (5). If we assume background $\mathbf{x_b}$ to be completely wrong (thus its covariance $\mathbf{B}$ is infinity), then the best estimated value is given by $\mathbf{x}$ coming from observation, such that $\mathbf{y_o} - \mathbf{H}.\mathbf{x} = \mathbf{0}$. On contrary, assuming measurements $\mathbf{y_o}$ to be completely wrong (and then $\mathbf{R}$ to be infinity), the best estimated value is the background $\mathbf{x_b}$.

2.3 Incore Measurement

The nuclear core plant is monitored along its entire lifespan. In order to do that, several measurements are used. There are three kinds of instruments that are used as standard to monitor nuclear power cores:

- Movable In-core Detector (MID), which are mobile fission chambers.
- Thermocouples.
- Fixed ex-core detector, which are named "*external chambers*".

Those instrumentation can be found on any power plants used by Electricité de France (EDF). Data coming from the ex-core detectors is very efficient for security purpose, which is their main goal. Nevertheless, they are too crude for being reliable in a fine evaluation of the neutron state. Thus, we will not use in this study information coming from those devices. All other instruments will be used for core state evaluation.

For the purpose of this study, we will add an artificial extra kind of detector, named Low granularity Movable In-core Detector (LMID): measurements

coming from the LMID are artificially built. Evaluation of the LMID response uses neutron flux calculation as for the MID, but assuming a different physical process and a lower granularity of the measure. The lower granularity assumption induces a partial integration of the standard MID response over a given area. Of course, the physical process assumed to make a measurement in LMID being different, the resolution and the accuracy of LMID instruments are different from the other ones.

3 Application of Data Assimilation

3.1 Multi-Instruments Data Assimilation Scheme

Multi-instruments data assimilation try to reconstruct the whole true state using information coming from various kind of instruments, as described in Sect. 2.3. The purpose is to obtain some state values (and errors on them) that are more accurate than inputed ones. The increase of accuracy results from combination of various available sources of information. A schematic view of the method is presented on Fig. 1.

In this case, observation operator represents information given by an instrument and to be compared to simulated data. This scheme represents the BLUE method described in Sect. 2.2. In our case, equations can be solved in a direct way because the size of the problem is fairly small (around 7,000–8,000 points) by comparison to meteorological ones (more than 10^5). The originality here is the use of heterogeneous information, as in meteorology, coming from different in-core instruments.

The Fig. 2 displays the quality of the reconstructed neutron state, versus the measurements, as a function of the number of various instruments taken into account for the analysis. Successive improvements of the RMS (Root Mean Square) misfit are presented with respect to the number of instruments used to build the analysis $\mathbf{x_a}$. This RMS is calculated in comparison to reference measures that was not used to reconstruct the core state.

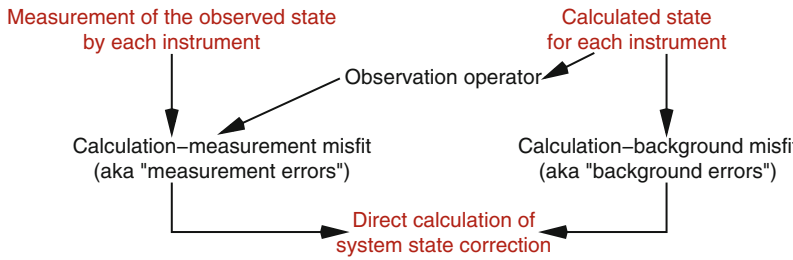


Fig. 1. Schematic diagram of multi-instruments data assimilation

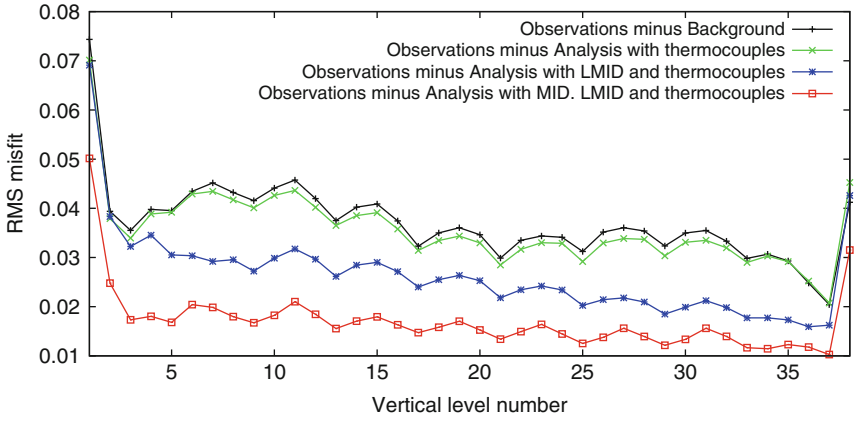


Fig. 2. Improvement of the RMS misfit with respect to the addition of instruments

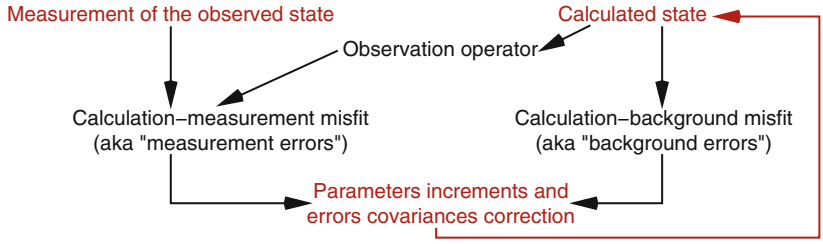


Fig. 3. Schematic diagram of parameter identification with data assimilation

As expected, the reduction of the misfit is stronger when additional informations are used for the reconstruction analysis. Taking into account thermocouples measurements, that are fully integral measurements, with a rather big representative error, only improves slightly the reconstruction. Then going further with the LMID and MID, that have some good or very good granularity, respectively, leads to a significant decrease of the RMS.

3.2 Parameter Identification Scheme

This method can be described as a multi-level BLUE method, as schematically presented on Fig. 3. In this case, the purpose of the observation operator is to give a linearised output of the model, with respect to some parameters. The parameters we focused on, in this study, are related to the parametrisation of the neutron reflector surrounding the core [3].

To obtain optimal parameters, several measurement campaigns are processed. Thus this algorithm is multi-level in the sense that it is iterative on the

parameters calculation. Moreover, there is an upper iterative level to optimise the **B** and **R** matrix with a Desroziers–Ivanov method [2], in order to correct the covariance matrices with the information coming from the observations.

Such a method leads to very interesting information on model parameters. A detailed study, of the parameters and their fitting, put in light its dependence with respect to the burning level (irradiation) of the nuclear fuel in the core. These properties, already mentioned for studied parameters, can be demonstrated using data assimilation. Another advantage is the capacity of the method to use all measurements, in a coherent framework.

4 Conclusions and Perspectives

With data assimilation methods, we manage to build an efficient scheme for two kinds of applications in nuclear core sciences.

In the *multi-instrument* processing, the main result is to be able to use all available measurements from various instruments in a unique framework. Moreover this method shows a slight but systematic improvement with respect to standard interpolation method. For *parameter evaluation*, the used method allows for an optimal identification of parameters in a single coherent scheme. Moreover, a new kind of dynamic for these parameters was shown.

As a perspective, these methods can be applied to the study of whatever model parameters, as long as their number still remains small (less than 100). For more parameters, intensive computations will require high performance computing (HPC). Thus, a wide use of Data Assimilation seems to be very promising to take advantage of all available nuclear data.

References

1. Bouttier, F., Courtier, P.: Meteorological Training Course, ECMWF, 1999
2. Desroziers, G., Ivanov, S.: Q.J.R. Meteorol. Soc. **127**(574, Part B), 1433–1452 (2001)
3. Massart, S., Buis, S., Erhard, P., Gacon, G.: Nucl. Sci. Eng. **155**(3), 409–424 (2007)
4. Talagrand, O.: J. Meteorol. Soc. Jpn. **75**(1B), 191–209 (1997)

An Iterative Method for Transport Equations in Radiotherapy

Bruno Dubroca[1] and Martin Frank[2]

[1] Université de Bordeaux, 33405 Talence, France, dubroca@math.u.bordeaux.fr
[2] University of Kaiserslautern, 67653 Kaiserslautern, Germany, frank@mathematik.uni-kl.de

Summary. Treatment with high energy ionizing radiation is one of the main methods in modern cancer therapy that is in clinical use. During the last decades, two main approaches to dose calculation were used, Monte Carlo simulations and semi-empirical models based on Fermi–Eyges theory. A third way to dose calculation has only recently attracted attention in the medical physics community. This approach is based on the deterministic kinetic equations of radiative transfer. In this work, we study a full discretization of the transport equation, whose solution is supposed to serve as a benchmark for simplified methods. The computational challenge is that scattering is forward-peaked, which makes a fine resolution and thus a very large linear system of equations necessary. Traditional methods like source iteration are inefficient or fail in this case. Therefore we propose a new method which combines an incomplete factorization of the scattering matrix and several iterative steps to obtain a fast and accurate solution. Numerical examples are given.

1 Introduction

The history of external beam radiation therapy starts with a remarkable anecdote: Literally two weeks after their discovery, X-rays were already used for cancer therapy. Röntgen discovered X-rays on December 28, 1895. Emil Grubbe, an undergraduate student at a medical school in Chicago, heard of Röntgen's work and obtained a vacuum discharge tube. He started experiments with the new rays, by producing X-ray images of himself. Obvious for us today, he started to suffer from radiation dermatitis. He realized the harmful effect of X-rays on tissue and on January 12, 1896, at the suggestion of one of his colleagues, he used his experimental setup to treat previously untreatable carcinoma. In February of 1896, he founded the first radiation therapy facility in Chicago.

Besides surgery and chemotherapy, the use of ionizing radiation is one of the main tools in the therapy of cancer today. According to WHO data, in the year 2007, there were about 11.3 million new cancer cases. More than half

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*,
Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_62,

of the patients that are treated receive radiation therapy at one point during their treatment.

There are many challenges facing an applied mathematician in this field. One is optimal treatment planning which aims at ensuring that enough energy is deposited in cancer cells so that they are destroyed, while at the same time healthy tissue around the cancer cells should be harmed as little as possible and some regions at risk should receive almost no radiation at all. In this work, we focus on a different aspect, namely methods for dose calculation.

Most dose calculation algorithms in clinical use rely on the Fermi–Eyges theory of radiation. In recent work [6], it has been shown that these can produce errors of up to 12% near inhomogeneities. In this work, we consider dose calculation using a Boltzmann transport equation. Similar to Monte Carlo simulations it relies on a rigorous model of the physical interactions in human tissue that can in principle be solved exactly. Monte Carlo simulations are widely used, but it has been argued that a grid-based Boltzmann solution should have the same computational complexity [2]. Electron and combined photon and electron radiation were recently studied in [5,7–9]. For a review on neutral particle codes that have been applied to the dose calculation problem we refer the reader to [4].

Here we study a full discretization of the transport equation, whose solution is supposed to serve as a benchmark for simplified methods. The computational challenge is that scattering is forward-peaked, which makes a fine resolution in energy and angle and thus a very large linear system of equations necessary. Traditional methods like source iteration are inefficient or fail in this case. Therefore we propose a simple iterative method which combines an incomplete factorization of the scattering matrix and several iterative steps to obtain a fast and accurate solution.

2 The Radiative Transfer Equation

The transport of particles that undergo inelastic scattering in a medium can be described by a Boltzmann transport equation

$$\begin{aligned}\mu\partial_x\psi(x,\epsilon,\mu) = \rho(x)\int_0^\infty\int_{-1}^1 s(x,\epsilon',\epsilon,\mu',\mu)\psi(x,\epsilon',\mu')d\mu' d\epsilon' \\ - \rho(x)\int_0^\infty\int_{-1}^1 s(x,\epsilon',\epsilon,\mu',\mu)\psi(x,\epsilon,\mu)d\mu' d\epsilon' + q(x,\epsilon,\mu)\end{aligned} \tag{1}$$

Here, ψ can be thought of being the number of particles at $x \in \mathbb{R}^3$ with energy ϵ, and direction $\mu \in [-1,1]$. To simplify the following presentation, we have written the radiative transfer equation in slab geometry (one-dimensional in both space and direction). However, our method can be easily extended and we show a two-dimensional (in both space and angle) result in the end. Scattering

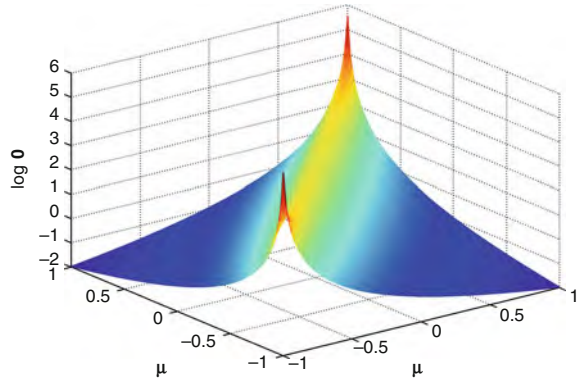


Fig. 1. Model Henyey–Greenstein scattering kernel ($g = 0.8$) as a function of μ and μ'

is determined by the density ρ of the medium and by the scattering kernel s, which can be seen as the probability that a particle with initial energy ϵ' and initial direction μ' has energy ϵ and direction μ after the scattering event. A model kernel is shown in Fig. 1. Note that the scale is logarithmic, which means that small angle changes are very likely. In order to resolve these small angle changes by a direct discretization, a large number of angles is necessary. As is well known and as we will again demonstrate later, traditional source iteration methods are inefficient for large scattering coefficients. Thus we propose a new iterative scheme. We should note that there exist several approximate methods to treat forward-peaked scattering (cf. [3]). These, however, introduce an additional approximation error. Our purpose here is to show that the equations can be solved by a direct method whose error can be controlled by the discretization only.

3 An Iterative Scheme

We discretize the unknown ψ in all variables. Let the index i denote direction, j energy, l space and n time, i.e.

$$\psi^n_{i,j,l} \sim \psi(t_n, x_l, \epsilon_j, \mu_i).$$

Here we have introduced an artificial time which we use as relaxation. Consider an implicit discretization in time, an upwind discretization in space and some discretization in energy and angle (e.g. finite differences or finite volume):

$$\frac{\psi_{i,j,l}^{n+1}-\psi_{i,j,l}^{n}}{\Delta t}+\mu_i^+\frac{\psi_{i,j,l}^{n+1}-\psi_{i,j,l-1}^{n+1}}{\Delta x}+\mu_i^-\frac{\psi_{i,j,l+1}^{n+1}-\psi_{i,j,l}^{n+1}}{\Delta x}$$
$$=\rho_l\left(\sum_{i',j'}\sigma_{i,i',j,j'}\psi_{i',j',l}^{n+1}-\sum_{i',j'}\sigma_{i,i',j,j'}\psi_{i,j,l}^{n+1}\right).$$

We have neglected the source q and defined $\mu^+=\max(\mu,0)$, $\mu^-=\min(\mu,0)$. Write this as

$$\left(\frac{1}{\Delta t}+\frac{|\mu_i|}{\Delta x}+\sum_{i',j'}\rho_l\sigma_{i,i',j,j'}\right)\psi_{i,j,l}^{n+1}-\sum_{i',j'}\rho_l\sigma_{i,i',j,j'}$$
$$=\frac{1}{\Delta t}\psi_{i,j,l}^{n}+\frac{\mu_i^+}{\Delta x}\psi_{i,j,l-1}^{n+1}-\frac{\mu_i^-}{\Delta x}\psi_{i,j,l+1}^{n+1}.$$

The left hand side is a matrix-vector multiplication in the i,j variables, the first term being a diagonal part. If we arrange the i,j in a suitable way into a vector ψ_l^n, we can write this as

$$M_l\psi_l^{n+1}=\frac{1}{\Delta t}\psi_{i,j,l}^{n}+\frac{\mu_i^+}{\Delta x}\psi_{i,j,l-1}^{n+1}-\frac{\mu_i^-}{\Delta x}\psi_{i,j,l+1}^{n+1}.$$

The symmetric matrix M_l is strongly diagonal-dominant but it is not sufficient to consider only its diagonal in an iterative scheme. The key idea is to factorize it as

$$M_l=A_l+B_l,$$

B_l containing the diagonal and a to be specified number of sub-/super-diagonals, and A_l containing the remainder. We want to invert the B_l part of M_l, thus we write

$$B_l\psi_l^{n+1}=\frac{1}{\Delta t}\psi_l^n+\frac{\mu_i^+}{\Delta x}\psi_{l-1}^{n+1}-\frac{\mu_i^-}{\Delta x}\psi_{l+1}^{n+1}-A_l\psi_l^{n+1}.$$

This is an implicit equation for ψ^{n+1}, which we solve iteratively by sweeping in the l variable. Let l run from 0 to $l_{\max}$, i.e. from left to right. The new iterate $\psi_l^{n+1,k+1}$ is given by

$$B_l\psi_l^{n+1,k+1}=\frac{1}{\Delta t}\psi_l^n+\frac{\mu_i^+}{\Delta x}\psi_{l-1}^{n+1,k+1}-\frac{\mu_i^-}{\Delta x}\psi_{l+1}^{n+1,k}-A_l\psi_l^{n+1,k}$$

and if we sweep from right to left

$$B_l\psi_l^{n+1,k+1}=\frac{1}{\Delta t}\psi_l^n+\frac{\mu_i^+}{\Delta x}\psi_{l-1}^{n+1,k}-\frac{\mu_i^-}{\Delta x}\psi_{l+1}^{n+1,k+1}-A_l\psi_l^{n+1,k}.$$

For $k\to\infty$ this gives us ψ^{n+1}. It only remains to iterate over the time index $n\to\infty$.

Table 1. Number of iterations as a function of time relaxation parameter

Relaxation Δt	10^1	10^2	10^3	10^4	10^5	10^6	10^7
Iterations	605	78	19	10	9	9	8

Table 2. Number of iterations for 64 directions and different matrix decompositions

Diagonals	0	1	2	4	8	12	14	16	32
Iterations	124	108	93	67	31	14	11	9	8

4 Numerical Results

First we study the convergence of our method. To that end we vary the relaxation time Δt and, more importantly, study the influence of the matrix decomposition. As a test case we choose an example from the medical physics literature [1]. It consists of a layered medium with depth 120 mm, consisting of three layers of 40 mm each, out of which the first and third are optically thick and the second is optically thin. It was sufficient to take 64 directions for the angle discretization. Table 1 shows the number of external iterations as a function of the time relaxation parameter, which should be chosen to be greater than 10^4. Table 2 shows how the performance of the algorithm depending on the matrix decomposition. The traditional source iteration method corresponds to the case of taking zero diagonals. We observe a significant decline in external iterations of more than one order of magnitude when we take several diagonals into matrix B_l. Of course the computational effort increases with the number of diagonals, since a larger linear system has to be solved. However, this increase is set off by the decrease in external iterations. Thus taking 16 diagonals (a quarter of the matrix dimension) is a sensible choice here.

We conclude with a test case in a two-dimensional quadratic domain which contains a void-like layer, shown in gray in Fig. 2. The physical parameters are detailed in [1]. An isotropic source of particles is placed on the left boundary. The propagation into the medium, as well as the void-like layer are equally well resolved in the numerical solution. The simulation used 64 directions and ran for about 10 min on 32 processors.

Acknowledgments

This work was supported by the French Ministry of Foreign Affairs under EGIDE contract 17852SD, by German Academic Exchange Service DAAD under grant D/0707534. and by the German Research Foundation DFG under grant KL 1105/14/2,

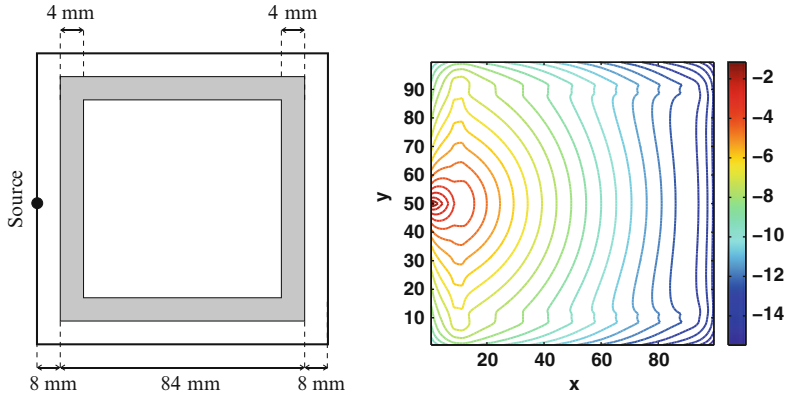


Fig. 2. Medium with void-like layer (*left*) and corresponding particle flux (*right*)

References

1. Aydin, E.D., Oliveira, C.R.E., Goddard, A.J.H.: Med. Phys. **29**, 2013–2023 (2002)
2. Börgers, C.: Phys. Med. Biol. **43**, 517–528 (1998)
3. Edström, P.: SIAM Rev. **47**, 447 (2005)
4. Gifford, K.A., Horton, J.L. Jr., Wareing, T.A., Failla, G., Mourtada, F.: Phys. Med. Biol. **51**, 2253–2265 (2006)
5. Hensel, H., Iza-Teran, R., Siedow, N.: Phys. Med. Biol. **51**, 675–693 (2006)
6. Krieger, T., Sauer, O.A.: Phys. Med. Biol. **50**, 859–868 (2005)
7. Tervo, J., Kolmonen, P.: Math. Models. Methods. Appl. Sci. **12**, 109–141 (2002)
8. Tervo, J., Kolmonen, P., Vauhkonen, M., Heikkinen, L.M., Kaipio, J.P.: Inv. Probl. **15**, 1345–1361 (1999)
9. Tervo, J., Vauhkonen, M., Boman, E.: Lin. Alg. Appl. **428**, 1230–1249 (2008)

Boundary Control of Radiative Transfer Equations for Application in Radiotherapy Planning

Martin Frank[1] and Michael Herty[2]

[1] University of Kaiserslautern, D-67653 Kaiserslautern, Germany
frank@mathematik.uni-kl.de
[2] RWTH Aachen University, D-52056 Aachen, Germany
herty@mathc.rwth-aachen.de

Summary. A radiotherapy treatment planning problem is formulated as a boundary control problem constrained by the radiative transfer equation. Optimality conditions for the radiative transfer equation as well as for the SP_1 approximation are stated. The latter are solved numerically.

1 Introduction

Mathematical methods play an increasing role in medicine, especially in cancer therapy. Several special journal issues have been devoted to cancer modeling and treatment, cf. [2–4, 7] among others. While until recently, treatment planning was done by an experienced physician "by hand", computer-aided treatment planning systems based on optimization algorithms currently enter into clinical practice, cf. [14] and references therein.

The use of ionizing radiation is one of the main tools in the therapy of cancer. The aim of radiation treatment is to deposit enough energy in cancer cells so that they are destroyed. On the other hand, healthy tissue around the cancer cells should be harmed as little as possible. Furthermore, some regions at risk, like the spinal chord, should receive almost no radiation at all. Most dose calculation algorithms in clinical use rely on the Fermi–Eyges theory of radiation which is insufficient at inhomogeneities, e.g. void-like spaces like the lung. We start with a Boltzmann transport model for the radiation which accurately describes all physical interactions, and based on this model we develop a direct optimization approach based on adjoint equations. Until recently, dose calculation using a Boltzmann transport equation has not attracted much attention in the medical physics community. This access is based on deterministic transport equations of radiative transfer. Similar to Monte Carlo simulations it relies on a rigorous model of the physical interactions in human tissue that can in principle be solved exactly. Monte Carlo simulations are

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*,
Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_63,

widely used, but it has been argued that a grid-based Boltzmann solution should have the same computational complexity [5]. Electron and combined photon and electron radiation were studied in the context of inverse therapy planning cf. [17,18] and most recently [19]. Furthermore, several neutral particle codes have been applied to the dose calculation problem, see [12] for a review.

This work is part of an ongoing project on dose calculation methods and optimal treatment planning based on Boltzmann transport equations. A consistent model of combined photon and electron radiation was developed [13] that includes the most important physical interactions. In [8], an approximate partial differential equation model was designed. A step toward time-dependent control in the case of moving patients was done in [9]. In the present work we extend results from [10], where distributed control was considered analytically and numerically, to boundary control.

2 Radiotherapy Planning as a Boundary Control Problem

Consider a part of the patient's body which contains the region of the cancer cells. We assume that this part of the body can be described as a convex, open, bounded domain Z in $\mathbb{R}^3$. Furthermore, we assume that Z has a smooth boundary with outward normal vector n. The direction, into which the electron is moving is given by $\omega \in S^2$, where S^2 is the unit sphere in three dimensions. To formulate boundary conditions, we define the in- and outgoing boundaries as

$$\partial Z^{\pm} := \{(x,\omega) \in \partial Z \times S^2 : n(x) \cdot \omega > (<) 0\}.$$

We consider particle transport modeled by the Boltzmann equation for the particle density $\psi(x,\omega)$ as

$$\omega \cdot \nabla_x \psi(x,\omega) + \sigma_t(x,t)\psi(x,\omega) = \sigma_s(x) \int_{S^2} s(x, \omega \cdot \omega')\psi(x,\omega')d\omega' \qquad (1)$$

with

$$\psi(x,\omega) = q(x,\omega) \ \text{ on } \partial Z^-.$$

For the sake of simplicity, we neglect the energy dependence of ψ. Here, $\psi(x,t,\omega)\cos\theta dA d\omega$ is the number of electrons that pass through an area dA at point x into a solid angle $d\omega$ around ω at time t. The angle θ is the angle between ω and dA. The total cross section $\sigma_t(x,t)$ is the sum of absorption cross section $\sigma_a(x,t)$ and total scattering cross section $\sigma_s(x,t)$. The scattering phase function is normalized,

$$2\pi \int_{-1}^{1} s(x,\mu)d\mu = 1. \qquad (2)$$

From the physical interpretation, we have that ψ, q, σ_t, σ_s and s are non-negative quantities. The detailed interactions of electrons with atoms give rise to complicated explicit formulas for the scattering coefficient, see e.g. [13].

The problem of external beam radiotherapy is to determine a boundary condition $q \geq 0$ in some optimal way. A number of functionals and methods have been devised to describe the effect of radiation on biological tissue, cf. the extensive lists of references in the reviews [6] and [16]. It is clear that the amount of destroyed cells in a small volume, be they cancer or healthy cells, is not directly proportional to the dose

$$D(x) = \int_{S^2} \psi(x, \omega) d\omega \tag{3}$$

deposited in that volume. However, no single accepted type of model has emerged yet. Moreover, current biological models require input parameters which are not known exactly [16]. This is why the authors of [16] opted not to investigate these models but rather to focus on some general mathematical cost functionals. A quadratic objective function together with nonlinear constraints was identified as the most versatile model. Thus we try to find a boundary value q such that the quadratic deviation from a prescribed dose $\bar{D}$ becomes minimal. The ideal dose curve of course has a certain fixed value in the tumour tissue and is zero elsewhere. We write

$$\bar{D}(x) = \int_{S^2} \bar{\psi}(x, \omega) d\omega \tag{4}$$

and introduce a weight function α depending on space, which penalizes deviations from the desired dose in normal tissue, tumour tissue and regions at risk differently. We want to minimize

$$J_1(D) = \int_Z \frac{\alpha}{2} (D - \bar{D})^2 dx \tag{5}$$

Furthermore, we include a penalty term proportional to the applied external source in the minimization process to prevent trivial solutions. We use a simple penalization as

$$J_2(q) := \int_{\partial Z^-} \frac{\beta}{2} (n \cdot \omega) q(s, \omega)^2 ds d\omega. \tag{6}$$

The problem of teletherapy reads then: Find $q \geq 0$ such that

$$J_1(D) + J_2(q) \tag{7}$$

is minimal, subject to the radiative transfer equation (1), and the relation (3) between ψ and the dose. This is an optimization problem constrained by an integro-differential equation.

Along the lines of [10], we can introduce a Lagrangian to this problem and formally derive the following first-order optimality conditions: The radiative transfer equation is

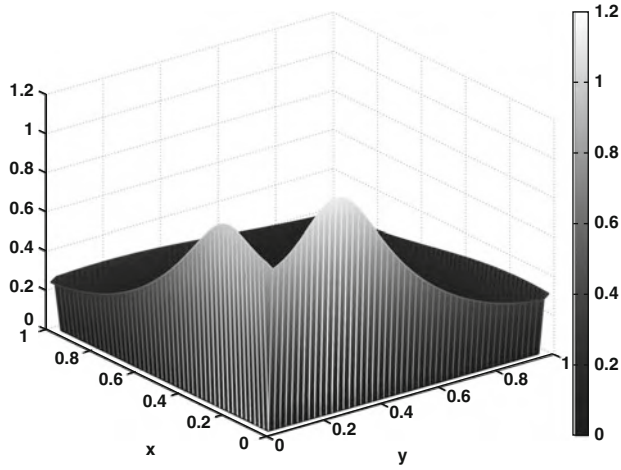


Fig. 1. Boundary control q

$$\omega \nabla_x \psi + \sigma_t \psi = \sigma_s \int_{S^2} s\psi d\omega' \quad \text{in } Z \text{ with } \psi = q \text{ on } \partial Z^-. \tag{8a}$$

The Lagrange multiplier λ satisfies a backward transfer equation

$$-\omega \nabla_x \lambda + \sigma_t \lambda = \sigma_s \int_{S^2} s\lambda + \alpha(\psi - \bar{\psi}) d\omega \quad \text{in } Z \text{ with } \lambda = 0 \text{ on } \partial Z^+. \tag{8b}$$

The gradient information, which is necessary for any efficient optimization algorithm in this case, is encoded in λ. It can be used directly in the optimality condition for q:

$$q = \left(q + \lambda - \alpha \int_{Z^-} n \cdot \omega q d\omega \right)^+ \text{ on } \partial Z^-. \tag{8c}$$

Here, $\xi^+ = \max(\xi, 0)$.

3 Numerical Results

We apply the optimize-then-discretize approach to the control law. In our 2D simulations, we use the Simplified P_1 (SP_1) approximation [15]. The unknown is the dose

$$D(x) = \int_{S^2} \psi(x, \omega) d\omega. \tag{9}$$

The SP_1 approximation for the radiative transfer equation [11] reads for isotropic scattering

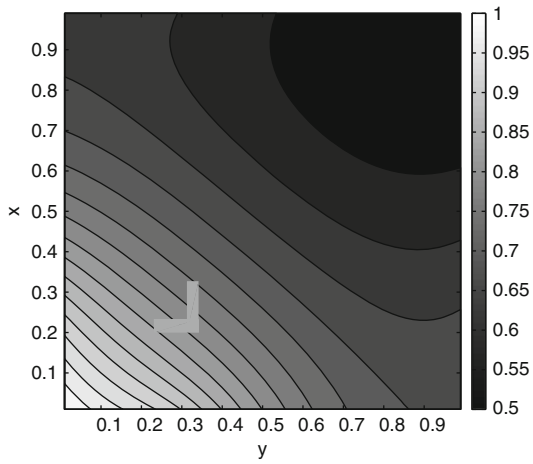


Fig. 2. Contour plot of the optimal dose distribution

$$-\nabla\frac{1}{3\sigma_t}\nabla D+(\sigma_t-\sigma_s)D=0. \tag{10}$$

The boundary conditions are

$$n\cdot\nabla D=\frac{3}{2}\sigma_t\left(l_1(q)-D\right),\quad \text{where}\quad l_1(q)=-4\int_{\partial Z^-} n\omega q(s,\omega)dsd\omega. \tag{11}$$

From the SP_1 approximation and the cost functional (5) and (6), we obtain the optimality system using SP_1 asymptotic as

$$\nabla\frac{1}{3\sigma_t}\nabla\lambda^{(0)}-(\sigma_t-\sigma_s)\lambda^{(0)}=-4\pi\alpha(D-\bar{D}),\ \text{with}\ n\cdot\nabla\lambda^{(0)}=-\frac{3}{2}\sigma_t\lambda^{(0)}\ \text{on}\ \partial Z^+. \tag{12}$$

The gradient equation is

$$\lambda^{(0)}-\frac{2}{3\sigma_t}n\cdot\nabla\lambda^{(0)}=-2\pi\beta l_1(q). \tag{13}$$

To demonstrate the feasibility of our approach, we consider the unit square $[0,1]^2$ with the parameters [1] $\sigma_s=0.05$ and $\sigma_t=0.5$. The domain contains an L-shaped tumour (Fig. 1). We numerically compute the solution to the SP_1-approximation of the optimality system. We consider the functional given by (5) and (6). When solving the optimal control problem we set $\beta=0$ (no regularization for the boundary control) and $\alpha=1$. All computations are done on a 50×50 grid. We solve the optimality system by using a projected gradient method. The optimal boundary control and the dose in the optimal state are shown in Figs. 1 and 2, respectively. Since the tumour is located in the bottom left corner, both maxima of the control are reasonable. The

dose basically falls off exponentially from the boundary, thus here a better dose distribution cannot be expected. In terms of performance, because of the high dimensionality of the optimization problem, we expect our adjoint-based method to outperform all black box methods using numerical gradients.

Acknowledgments

This work was supported by the German Research Foundation DFG under grants SPP 1253 and KL 1105/14/2, and from the Rheinland-Pfalz Excellence Cluster "Dependable Adaptive Systems and Mathematical Modeling".

References

1. Aydin, E.D., Oliveira, C.R.E., Goddard, A.J.H.: Med. Phys. **29**, 2013–2023 (2002)
2. Bellomo, N., Maini, P.K.: Math. Mod. Math. Appl. Sci. **15**, iii–viii (2005)
3. Bellomo, N., Maini, P.K.: Appl. Sci. **16**, iii–vii (2006)
4. Bellomo, N., Maini, P. K.: Appl. Sci. **17**, iii–vii (2007)
5. Börgers, C.: Phys. Med. Biol. **43**, 517–528 (1998)
6. Börgers, C.: IMA Volumes in Mathematics and its applications **110** (1999)
7. Censor, Y.: Lin. Alg. Appl. **428**, 1203–1205 (2008)
8. Frank, M., Hensel, H., Klar, A.: SIAM J. Appl. Math. **51**, 582–603 (2007)
9. Frank, M., Herty, M., Hinze, M.: Time-dependent Closed Loop Control of the Radiative Transfer Equations with Applications in Radiotherapy, preprint
10. Frank, M., Herty, M., Schäfer, M.: Math. Mod. Meth. Appl. Sci. **18**, 573–592 (2008)
11. Frank, M., Klar, A., Larsen, E.W., Yasuda, S.: J. Comput. Phys. **226**, 2289–2305 (2007)
12. Gifford, K.A., Horton, J.L. Jr., Wareing, T.A., Failla, G., Mourtada, F.: Phys. Med. Biol. **51**, 2253–2265 (2006)
13. Hensel, H., Iza-Teran, R., Siedow, N.: Phys. Med. Biol. **51**, 675–693 (2006)
14. Kuefer, K.-H., et al.: in Handbook of optimization in medicine. Springer (2009)
15. Larsen, E.W., Morel, J.E., McGhee, J.M.: Asymptotic derivation of the multigroup P_1 and simplified P_N equations with anisotropic scattering. Nucl. Sci. Eng. **123**, 328 (1996)
16. Shepard, D.M., Ferris, M.C., Olivera, G.H., Mackie, T.R.: SIAM Rev. **41**, 721–744 (1999)
17. Tervo, J., Kolmonen, P.: Math. Models. Methods. Appl. Sci. **12**, 109–141 (2002)
18. Tervo, J., Kolmonen, P., Vauhkonen, M., Heikkinen, L.M., Kaipio, J.P.: Inv. Probl. **15**, 1345–1361 (1999)
19. Tervo, J., Vauhkonen, M., Boman, E.: Lin. Alg. Appl. **428**, 1230–1249 (2008)

Model Hierarchies and Optimal Control of Radiative Transfer

R. Pinnau[1] and G. Thömmes[2]

[1] TU Kaiserslautern, Department of Mathematics, Kaiserslautern, Germany, pinnau@mathematik.uni-kl.de
[2] Fraunhofer ITWM, Kaiserslautern, Germany

Summary. Optimal control problems in radiative transfer are solved by means of the space mapping technique. Exploiting a hierarchy of approximate models, this allows for the construction of fast numerical algorithms. The performance of the algorithms is underlined by numerical experiments.

1 Introduction

We consider an optimal control problem in the realm of radiative transfer with a tracking-type cost functional for given functions $\bar{\varphi}, \bar{Q} : D \to \mathbb{R}$,

$$F(\varphi, Q) = \frac{\alpha_1}{2} \int_D (\varphi - \bar{\varphi})^2 dx + \frac{\alpha_2}{2} \int_D (Q - \bar{Q})^2 dx. \tag{1}$$

Here, $\varphi(x) = \int_{S^2} I(x, \omega) d\omega$ denotes total flux corresponding to the space and direction dependent intensity I.

The intensity $I(x, \omega) : \mathbb{R}^d \times S^2 \to \mathbb{R}$ is computed by solving the radiative transfer equation,

$$\varepsilon \omega \cdot \nabla I + (\sigma_s + \sigma_a) I = \frac{\sigma_s}{4\pi} \int_{S^2} I \, d\omega' + Q(x), \tag{2a}$$

where d is the space dimension of the underlying domain, S^2 is the sphere in $\mathbb{R}^3$ and σ_s and σ_a are problem dependent scattering and absorption parameters and ε is a scaling factor of the equation, i.e. $\varepsilon = x^{ref}/(\sigma_a^{ref} + \sigma_s^{ref})$. This equation is supplemented with appropriate Dirichlet data

$$I(x, \omega) = A, \qquad n \cdot \omega < 0, \tag{2b}$$

for ingoing directions. The equation contains the source term $Q(x)$, which can be interpreted as an exterior source or sink of radiation energy. This external source is the control variable of our problem.

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*, Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_64,

The corresponding optimisation problem for determining a distributed control $Q(x)$ reads

$$\min_{\varphi,Q} F(\varphi, Q) \text{ subject to (2).} \tag{3}$$

This optimisation problem was first analysed in [4], where the existence and uniqueness of an optimal control Q is proved. Here, we exploit the approximate SP_N model hierarchy [6] and the space mapping technique [2, 3] to construct a fast optimisation algorithm. This model hierarchy was already used studied in the context of optimal control in radiative transfer in [5], where an asymptotic analysis of the first order optimality system is performed.

In the following we describe the approximate model hierarchy and the space mapping technique and discuss some numerical results.

2 The SP_N Models

The SP_N approximations for the radiative transfer equation (2) are derived in a formal manner applying the Neumann series to an unbounded operator when the medium is assumed to be optically thick, i.e. the mean free path ϵ is small (for a detailed derivation see [8]). The continuous SP_1 approximation of (2) is given by

$$-\frac{\varepsilon^2}{3(\sigma+\kappa)}\nabla^2\phi + \kappa\phi = 4\pi Q, \tag{4a}$$

and the corresponding SP_2 approximation is

$$-\varepsilon^2\frac{5(\sigma+\kappa)+4\kappa}{15(\sigma+\kappa)}\nabla^2\xi + \kappa\xi = 4\pi Q, \tag{4b}$$

where $\xi = \phi + \frac{4\kappa}{5(\sigma+\kappa)}(\phi - 4\pi Q/\kappa)$, and the SP_3 equations are

$$-\frac{\varepsilon^2}{3(\sigma+\kappa)}\nabla^2(\phi+2\phi_2) + \kappa\phi = 4\pi Q, \tag{4c}$$

$$-\frac{9\varepsilon^2}{35(\sigma+\kappa)}\nabla^2\phi_2 + (\sigma+\kappa)\phi_2 - \frac{2}{5}\kappa\phi = -\frac{2}{5}\,4\pi Q. \tag{4d}$$

In all cases, ϕ approximates the mean intensity given by (2). For the validity of the SP_N approximations we refer to [8]. This model hierarchy yields the so-called coarse models, which are used in the setup of the following space mapping algorithm.

3 Aggressive Space Mapping

The solution of the full optimisation problem (3) is rather time consuming, since it requires the discretisation of the spatial and of the angular variable.

Now, we want to define a algorithm which only requires solves of the *fine* radiative transfer problem (2) and not the solution of the full optimisation problem. Instead, we will only solve optimisation problems based on the *coarse* SP_N models, for which one can easily implement an optimisation algorithm using the adjoint information for the computation of descent directions (for details we refer to [5]).

In particular, we want to approximate the solution of the fine model by an appropriate solution of the coarse model for which we define the misalignment function (here we follow [3])

$$r(R, Q) = \|\phi(R) - \varphi(Q)\|,$$

where $\phi(R)$ is the coarse model output of a SP_N model for a given source R and $\varphi(Q)$ is fine model output computed by (2) for a given source Q. For a given Q we look for R such that $r(R, Q)$ is minimal, i.e., we define the space mapping function

$$p(Q) = \text{argmin}_R r(R, Q).$$

Since we want to evaluate p only a few times, we assume $\varphi(Q^*) \approx \phi(R^*)$, such that

$$p(Q^*) = \text{argmin}_R r(R, Q^*) \approx R^*.$$

Hence, we first determine R^* and then solve for $p(Q^*) = R^*$. But in general it holds $p(Q^*) \neq R^*$, such that we solve instead for

$$Q^* = \text{argmin}_Q \|p(Q) - Q^*\|.$$

This is done iteratively and the space mapping p is updated using a Broyden–rank–1 update yielding the so–called ASM (aggressive space mapping) algorithm (for details we refer to [3]):

1. Evaluate $Q_0 = R^* = \text{argmin}_R \|c(R)\|^2$ and let B_0 be the identity matrix.
2. While $\|p(Q_k) - R^*\| / \|\zeta^*\| >$ tolerance
   a) Evaluate $\varphi(Q_k)$ by solving the fine model (2)
   b) Determine $R_k = p(Q_k) = \text{argmin}_R \|c(R) - f(Q)\|^2$
   c) Solve $B_k h_k = -(p(Q_k) - R^*)$ for h_k
   d) Set $Q_{k+1} = Q_k + h_k$
   e) Update $B_{k+1} = B_k + \frac{(p(Q_{k+1}) - R^*) h_k^T}{h_k^T h_k}$
   f) Set $k \to k + 1$.

Here, we have rewritten the cost functional (1) by defining

$$c(R) = \begin{pmatrix} \sqrt{\frac{\alpha_1}{2}}(\varphi(R) - \bar{\varphi}) \\ \sqrt{\frac{\alpha_2}{2}}(R - \bar{Q}) \end{pmatrix}, \quad \text{and} \quad f(Q) = \begin{pmatrix} \sqrt{\frac{\alpha_1}{2}}(\phi(Q) - \bar{\varphi}) \\ \sqrt{\frac{\alpha_2}{2}}(Q - \bar{Q}) \end{pmatrix}.$$

Remark 1. On each iteration level we need one evaluation of the fine model and one solve of the optimal control problem for the coarse model. That is, it is sufficient to implement an adjoint code for the coarse model [5].

4 Numerical Results

We implemented test cases in 1D using the DSA iterative scheme for the transport equations of the forward and adjoint equations on the *fine* level [1,7]. The radiative transfer equation was discretised on an equidistant space grid using the diamond differencing scheme by evaluating intensity I and source q at the nodes $x_i = i\Delta x,\ i = 0, \dots, M$ and using averages $I_{i+\frac{1}{2}} = (I_{i+1} + I_i)/2$ and $q_{i+\frac{1}{2}} = (q_{i+1} + q_i)/2$. The iteration is started by choosing an initial iterate I_{ij}^0 and computing the flux $\varphi_i^0 = \sum_{j=1}^N I_{ij}^0 w_j$. Then, for $k \geq 0$, the iteration proceeds in two substeps. First, the following transport equation with given right side is solved for the intermediate intensity $I_{ij}^{k+\frac{1}{2}}$

$$\varepsilon\mu_j \frac{I_{i+1,j}^{k+\frac{1}{2}} - I_{ij}^{k+\frac{1}{2}}}{\Delta x} + \sigma_t I_{i+\frac{1}{2},j}^{k+\frac{1}{2}} = \frac{\sigma_s}{2}\varphi_{i+\frac{1}{2}}^k + q_{i+\frac{1}{2}},$$
$$\text{with b.c.} \quad I_{0,j}^{k+\frac{1}{2}} = A,\ \mu_j > 0, \quad I_{M,j}^{k+\frac{1}{2}} = A,\ \mu_j < 0.$$

This corresponds to the transport sweep in the source iteration method. Note that the sweep is done from left to right when $\mu_j > 0$, and from right to left when $\mu_j < 0$. Then the flux difference $\varphi_i^{k+\frac{1}{2}} = \sum_{j=1}^N (I_{ij}^{k+\frac{1}{2}} - I_{ij}^k) w_j$ is taken as source term for the computation of the correction $\delta\varphi^{k+\frac{1}{2}}$:

$$-\frac{\varepsilon^2}{3\sigma_t}\frac{\delta\varphi_{i+1}^{k+\frac{1}{2}} - 2\delta\varphi_i^{k+\frac{1}{2}} + \delta\varphi_{i-1}^{k+\frac{1}{2}}}{\Delta x^2} + \sigma_a \frac{\delta\varphi_{i+1}^{k+\frac{1}{2}} + 2\delta\varphi_i^{k+\frac{1}{2}} + \delta\varphi_{i-1}^{k+\frac{1}{2}}}{4}$$
$$= \sigma_s \frac{\varphi_{i+1}^{k+\frac{1}{2}} - \varphi_{i+1}^k}{2} + \sigma_s \frac{\varphi_i^{k+\frac{1}{2}} - \varphi_i^k}{2},$$

with homogeneous boundary conditions on the left and right of the interval. The new iterate for the flux is eventually updated

$$\varphi_i^{k+1} = \varphi_i^{k+\frac{1}{2}} + \delta\varphi_i^{k+\frac{1}{2}}.$$

After the iteration has stopped, we obtain a numerical solution for the intensity by performing an additional sweep with the final flux. The coarse level SP_N approximations corresponding to the one-dimensional transfer equation were discretised using standard finite differences.

In the examples presented in the following, the radiative transfer problem had an absorption cross section $\sigma_a = 1$ and a scattering cross section $\sigma_s = 1$, and the non-dimensional parameter $\varepsilon = 1$ was used. The equation was discretised with $N_x = 51$ nodes $x_i = i/N_x$ in space. For the angular variable, $N_g = 32$ nodes given by the quadrature points of double Gaussian quadrature on $[-1, +1]$ were used. The RTE equation was solved for homogeneous Dirichlet boundary conditions $A = 0$ for ingoing directions at the left and right, respectively. The tracking part of the functional had the weight $\alpha_1 = 1$

Table 1. Comparison of the optimisation methods

	Iterations	Evals	Runtime $[s]$	F_{End}
Gradient	12	44	0.7	$5.3540 \cdot 10^{-8}$
Trust region	14	15	0.5	$7.5886 \cdot 10^{-6}$
ASM	3	3(40)	0.3	$2.9682 \cdot 10^{-6}$

Table 2. Comparison of the optimisation methods

	Iterations	Evals	Runtime $[s]$	F_{End}
Gradient	11	42	0.8	$7.2731 \cdot 10^{-3}$
Trust region	4	5	0.2	$7.2733 \cdot 10^{-3}$
ASM	2	2(14)	0.2	$7.4130 \cdot 10^{-3}$

and the regularisation the weight $\alpha_2 = 1$. The same functional was also used for the coarse model.

The performance of a classical gradient and a trust region algorithm for the fine model was compared with aggressive space mapping based on the SP_1 model. The space mapping $p(Q)$ was evaluated using a trust region method for the coarse functional optimisations [3]. We used a stopping criterion based on the functional values and the size of the gradient of the functional. These parameters were also used for the standard trust region method. The standard gradient method included an Armijo line search.

4.1 Box-Shaped Source

In the first numerical test we use a prescribed reference source q_{ref} as shown in Fig. 1, for which φ_{ref} is computed. Then, we seek the optimiser starting from a vanishing source. The performance of the algorithms can be seen from Table 1 and the corresponding final results are depicted in Fig. 1. The aggressive space mapping needs just three evaluations of the fine model and 40 function evaluations on the coarse level. Hence, we get comparable results with less numerical effort.

4.2 Unknown Reference Source

In the second test case we want to reconstruct an unknown source and set $q_{ref} = 0$. The reference intensity belongs to a Gaussian source term (center 0.5, std. dev. 0.1) and we start the algorithms again from a vanishing source. The performance of the algorithms can be seen from Table 2. Again, the aggressive space mapping needs just very few evaluations of the fine model and yields results comparable to the full optimisation.

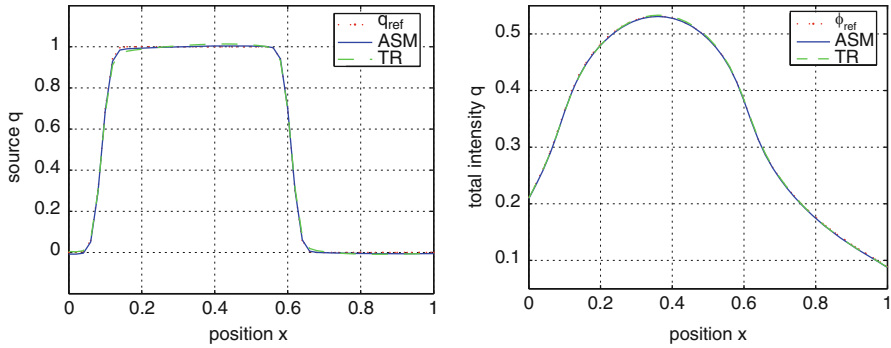


Fig. 1. *Left*: Control after optimisation using space mapping. The results of aggressive space mapping and a trust region solve are compared with the reference solution *Right*: The corresponding state given by the flux, i.e., the angle-integrated intensity

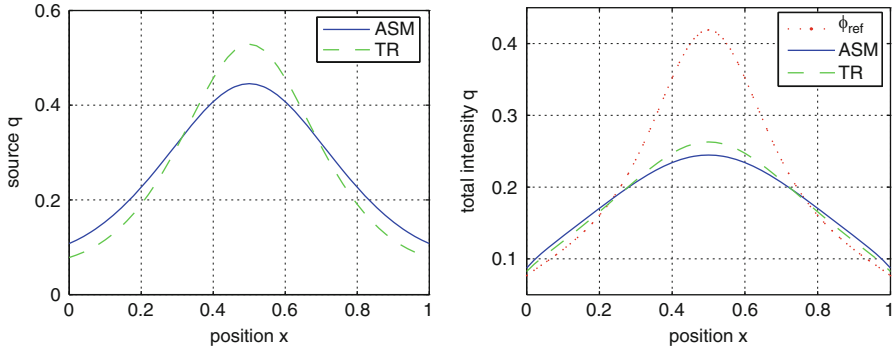


Fig. 2. *Left*: Control after optimisation using space mapping. The results of aggressive space mapping and a trust region solution are compared. *Right*: The corresponding state given by the flux, i.e., the angle-integrated intensity and the reference intensity

References

1. Alcouffe, R.E.: Nucl. Sci. Eng. **64**, 344–355 (1977)
2. Bandler, J.W., Biernacki, R.M., Chen, S.H., Grobelny, P.A., Hemmers, R.H.: IEEE Trans. Microw. Theory Tech. **43**, 2874–288 (1995)
3. Echeverría, D., Hemker, P.W.: Comput. Methods Appl. Math. **5**(2), 107–136 (2005)
4. Herty, M., Pinnau, R., Seaid, M.: OMS **22**(6), 917–936 (2007)
5. Herty, M., Pinnau, R., Thömmes, G.: ZAMM **87**(5), 333–347 (2007)
6. Larsen, E.W., Thömmes, G., Klar, A., Seaid, M., Götz, T.: JCP **183** (2002)
7. Lewis, E.E., Miller, W.F. Jr.: Comput. Methods of Neutron Transport, Wiley New York (1984)
8. Thömmes, G.: Radioactive Heat Transfer for Glass Cooling Problems, PhD Thesis, Tu Darmstadt (2002)

Minisymposium *Optimization and Model Order Reduction in Circuit Design*

Giuliana Gangemi

STMicroelectronics, Stradale Primosole 50, Catania, Italy
Giuliana.Gangemi@st.com

While during the last decades the great enhancements in the field of digital design methodologies and tools have allowed to design larger digital circuits in less time, the analog circuit design methods have not progressed at the same rate. The design of analog electrical circuits needs electronic engineers with a long experience and a wide knowledge of the theories that rule this kind of circuits. However, experimental optimization tools exist; they search the space of solutions for optimal configurations of variables sets, given a circuit netlist provided by the designers. Typical analog integrated circuit optimization problems are computationally hard and require the handling of multiple, conflicting, and non-commensurate objectives having strong nonlinear interdependence. In general it is possible to reformulate integrated circuit design as constrained multi-objective optimization problems defined in a mixed integer/discrete/continuous domain. The hereby employed traditional numerical techniques are becoming too much time-consuming for circuits of industrial complexity. The long computation time required for the optimization of a complete circuit cannot be tolerated especially in the early design stages. For tackling this complexity problem model reduction methods are a promising approach in order to achieve a faster performance evaluation in order to obtain more robust devices within a more efficient design process.

The minisymposium focused on the usage of model reduction techniques in combination with optimization methods. The results are developed in the EU Marie Curie projects SymTecO (Symbolic Techniques for Circuit Optimization) and O-Moore-Nice! (Operational Model Order Reduction for Nanoscale IC Electronics). Both projects address Transfer of Knowledge on Mathematics for Industry.

Paola Barrera from STMicroelectronics in Catania, with Thomas Halfmann and Jochen Broz from the Fraunhofer Institute (ITWM) in Kaiserslautern, presented a talk from SYMTECO on *"A Netlist Reduction Algorithm to Symbolic Circuit Analysis"*, in which new reduction algorithm in the area of symbolic circuit analysis was described. The reduction of a netlist as well as of the model order complexity are important modelling issues which help

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*,
Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_65,

to speed up the process of integrated circuit design [5]. The proposed method eliminated nodes from a netlist topology assuring a user-given accuracy margin. The algorithm was based on the decision diagram derived from the circuit topology and considered low memory storage issues in order to efficiently carry out the simplification. Starting from the application of a spiral inductor test case [1] efficiency was evaluated. The reduced system complexity in terms of netlist nodes and model order encouraged the application to other industrial test cases.

Alberto Venturi from the Fraunhofer Institute (ITWM) in Kaiserslautern, with the contribution of A. Ciccazzo, S. Rinaudo from STMicroelectronics in Catania (Italy) gave a talk from SYMTECO on *"Application of optimization and model order reduction techniques"* in which he explained how given the computation time required for the analysis of a complete circuit can be too long for an adequate use of optimization methods in industrial circuit design, the use of symbolic analysis together with model order reduction techniques could reduce the computational cost and hence make optimization a practicable way in the circuit design. To evaluate the possibilities offered by this technique, a linear test case had been considered: the problem of an inductor simulation had been analyzed by introducing simplified analytical expressions and different optimization algorithms in the fitting/optimization process. Then the technique was applied to a real circuit, a voltage reference, trying to improve the stability of the reference over the temperature.

Jan ter Maten of NXP Semiconductors presented *"Parameterized Model Order Reduction for nonlinear IC models"*. This work was in cooperation with Joost Rommes (NXP) and Michael Striebel (TU Chemnitz) of the O-MOORE-NICE! project and with Tamara Bechtold(NXP), Kasra Mohaghegh (Univ. of Wuppertal) and Zoran Ilievski (TU Eindhoven) of the COMSON RTN-project. He demonstrated Model Order Reduction for a nonlinear system of differential-algebraic equations of a diode chain. While the Trajectory PieceWise Linear method (TPWL) is very fast it also is very sensitive to the change of input signals. The weighting procedure of linear models was pointed out as a key ingredient that needs further research in order to further improve the method. Proper Orthogonal Decomposition (POD) much better preserves nonlinearity, but needed significant adaptions (called Adapted Missing Point Estimation) to become comparable in speed to TPWL [7]. The resulting method also is much more accurate than TPWL and behaves better to changes of the input. The snapshots collected in POD can also be used to efficiently obtain a first impression of sensitivities of objective functions [6].

Luciano De Tommasi from Antwerp and Ghent University – IBBT, Belgium, gave a talk entitled *"Optimization in surrogate model building for RF circuit blocks"* (joint O-MOORE-NICE! project work with D. Gorissen, J. Croon and T. Dhaene). Surrogate models, also known as response surface models, have become a cost effective alternative for replacing expensive computer simulations when exploring the design space, performing what-if analysis, optimization and sensitivity analysis. Relevant aspects which have

been investigated include model type selection [2, 3], adaptive sampling [2] and optimization of model parameters [2–4] (adaptive modeling).

Acknowledgements

Several people contributed to the presentations: E.J.W. ter Maten, J. Rommes, T. Bechtold (NXP Semiconductors), A. Verhoeven (TU Eindhoven), M. Striebel (TU Chemnitz), P. Barrera, A. Ciccazzo, S. Rinaudo (STMicroelectronics), A. Venturi, Th. Halfmann, J. Broz (Fraunhofer Institut, ITWM Kaiserslautern) L. De Tommasi (Univ. of Antwerp), D. Gorissen, T. Dhaene (Ghent University), J. Croon (TSMC-NXP, Eindhoven)

References

1. Ciccazzo, A., Greco, G., Rinaudo, S.: Coupled EM and Circuit Simulation Flow for Integrated Spiral Inductor. In: Anile, A.M., Alì, G. Mascali, G. (eds.): Scientific Computing in Electrical Engineering, Series Mathematics in Industry, vol. 9, pp. 317–322. Springer (2006)
2. De Tommasi, L., Gorissen, D., Crombecq, K., Dhaene, T.: Introduction to Surrogate Modeling of Narrowband Weakly Nonlinear Low Noise Amplifiers, Technical Report, 2008
3. Gorissen, D., De Tommasi, L., Croon, J., Dhaene, T.: Automatic Model Type Selection with Heterogeneous Evolution: An application to RF circuit block modeling, Proceedings of IEEE World Congress on Computational Intelligence (WCCI), Hong Kong, 1–6 June 2008
4. Gorissen, D., De Tommasi, L., Croon, J., Dhaene, T.: RF Circuit Block Modeling via Kriging Surrogates. Proceedings of 17th International Conference on Microwave, Radar, and Wireless Communications (MIKON), Wroclaw, 19–21 May 2008
5. Halfmann, T., Wichmann, T.: Symbolic Methods in Industrial Analog Circuit Design. In: Anile, A.M., Alì, G., Mascali, G. (eds.): Scientific Computing in Electrical Engineering, Series Mathematics in Industry, vol. 9, pp. 87–92. Springer (2006)
6. Ilievski, Z., Xu, H., Verhoeven, A., ter Maten, E.J.W., Schilders, W.H.A., Mattheij, R.M.M.: Adjoint Transient Sensitivity Analysis in Circuit Simulation. In: Ciuprina, G., Ioan, D. (eds.) Scientific Computing in Electrical Engineering SCEE 2006, Series Mathematics in Industry, vol. 11, pp. 183–189. Springer (2007)
7. Verhoeven, A., ter Maten, E.J.W., Striebel, M., Mattheij, R.M.M.: Model Order Reduction for Nonlinear IC Models. In: Korytowski, A., Malanowski, K., Mitkowski, W., Szymkat, M. (eds.): System Modeling and Optimization, Series IFIP AICT 312, Springer, pp. 476–491 (2009)

A Netlist Reduction Algorithm to Symbolic Circuit Analysis

Paola Barrera[1], Jochen Broz[2], and Thomas Halfmann[2]

[1] STMicroelectronics, Stradale Primosole 50, Catania, Italy,
paola.barrera@st.com
[2] Fraunhofer, ITWM Kaiserslautern, Germany,
jochen.brot@itwm.fraunhofer.de, thomas.halfmann@itwm.fraunhofer.de

Summary. A new reduction algorithm in the area of symbolic circuit analysis is presented. The reduction of a netlist as well as of the model order complexity are important modeling issues which help to speed up the process of integrated circuit design. The proposed method eliminates nodes from a netlist topology assuring a user-given accuracy margin. The algorithm is based on the decision diagram derived from the circuit topology and considers low memory storage issues in order to efficiently carry out the simplification. Starting from the application of a spiral inductor test case, the efficiency is evaluated. The reduced system complexity in terms of netlist nodes and model order encourage the application to other industrial test cases.

1 Introduction

Analog and mixed-signal design is of great importance in microelectronics applications, like automotive and telecommunication. The traditional design of analog integrated circuits is based on a mixture of expertise, some manual calculations, and numerical circuit simulations [4]. In order to improve a faster performance evaluation as well as a deeper understanding of complex circuits, the symbolic analysis and simplification techniques have been introduced in the electronic design community. Symbolic analysis is a formal technique to calculate the behavior of a circuit, writing and solving the circuit equations, in which the variables and the circuit parameters are represented by symbols. Symbolic analysis is complementary to the numerical analysis (where the variables and the circuit elements are represented by numbers), that even if allows an accurate simulation of the circuit is not able to carry out which elements are critical for the circuit behaviour [2]. If several evaluations of the circuit with different sets of parameters are necessary, for example to handle an optimization problem, the process is time consuming. So an alternative could be to solve once the circuit equations symbolically and verify which terms are relevant or need to be optimized. On the other hand the exact symbolic analysis

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*,
Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_66,

yields expressions which increase in complexity with the increase of the components in the circuit. For example, even a simple common-emitter amplifier consisting of only one BJT transistor already results in a symbolic transfer function for the small-signal voltage gain with more than 130 terms [4]. It appears clear that in order to use the advantages of the symbolic analysis it is necessary to simplify the generated expressions by keeping the dominant terms only. In this way it is possible not only to simplify the solutions and put in evidence which parts or parameters are dominants but also to speed up the optimization and evaluations time of the circuit behaviour. To achieve this goal a symbolic simplification or symbolic approximation which uses a whole family of hybrid symbolic/numeric algorithms for expression simplification is taken into account. An user-fixed error, evaluated on the base of the neglected terms, is introduced with this approach, but the reduction of the complexity of the symbolic expression is obtained [2, 3]. However the simplification algorithms, are applied to the equations of the circuit obtained by standard analysis technique, like the MNA (Modified, Nodal, Analysis) and the Sparse Tableau, which give rise to a matrix constructed on the basis of the circuit netlist. So the proposed approach, is based on the netlist reduction of the circuit under test, in order to obtain a simplified equations representation. The developed symbolic/numeric algorithms can be then applied to the simplified equations. In the next section, the technique of symbolic circuit representation is first introduced with the illustration of a practical example. The basic methodology of how symbolic analysis works is explained, then an indication of the advantages of the application of the netlist reduction algorithm is given. The description of the algorithm flow and the preliminary results are also introduced. Finally, concluding remarks and points out directions for future research.

2 Principles of Symbolic Analysis

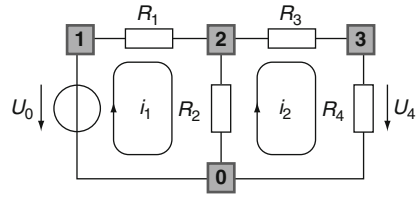
2.1 Circuit Equations Analysis

With increasing frequencies and faster signal transition times, on chip inductive effects, which describe the interconnection of one macro circuit to another one, are critically important in the design and verification of integrated circuits. On-chip interconnects are typically modeled by linear elements. Consequently, many commercial extraction tools, generate RLC circuits for high performance designs. These circuits together with the non-linear drivers are then analyzed by fast timing simulators or tools using linearized models of the drivers. The extraction tools generate a large amount of data, and as a consequence the analysis tools require significant resource in term of CPU calculation time and memory occupation [1]. On the other hand, as well as for numerical analysis tools, the input of symbolic analysis, is an extracted netlist which describes the circuit under test. In order to speed up together

the symbolic and the numerical simulations, a netlist reduction algorithm has been developed. The netlist of a circuit is a list which contains the name, the node connections, the symbolic and the numerical value of each element. Following the example illustrated in Fig. 1 the element R1 of the list can be represented like: [A,(1,2), R1, 10]. In order to apply the symbolic analysis algorithms, the first step is to set up the equations associated to the circuit, using e.g. the MNA or the Sparse Tableau representations. If the MNA representation is used, for the circuit in Fig. 1, the result is a 4×4 matrix. It is possible to observe that each resistance (not connected to the node labeled as zero) introduces in the matrix four terms and two corresponding equations (in order to calculate the potentials at the nodes at which it is connected). If the circuit under test is more complex, if are present controlled source, for example, or in industrial applications, the number of terms and equation in the matrix increase drastically. In a symbolic circuit analysis, the second step is the output calculation and its simplification with a user-given error, or the simplification of the matrix with a fixed error and the consequent output evaluation. It appears clear that a reduced matrix can help the simplification process, in either cases. The basic idea, of the developed algorithm is: starting from the netlist which describes the circuit under test, take into account the series and parallel connections and reduce at first the circuit topology and then according with the numerical values of the element simplify the circuit topology. In either case the number of elements in the matrix will be reduced. In the first case the resulting output will be exact, in the second case, there will be an user-given error. In the following section it is introduced the algorithm flow.

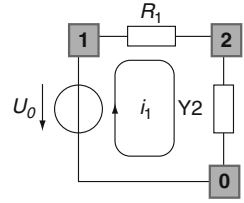
2.2 Netlist Reduction Algorithm

The developed algorithm is based on the analysis of the circuit topology. From a circuit point of view, components connected in series or in parallel can be replaced with unique elements. As a result there will be the deletions of the nodes that the elements in series have in common and the compression in one branch of all the components which are in parallel. The effect of the application of these simplifications will preserve the exact circuit behavior as well as the number of parameters, but a significant reduction in the number of variables and so in the equations and in the matrix terms which describe the circuit is obtained. We will refer to this topological reduction like a series and/or parallel reduction. As the designer knows the range of variations of the circuit parameters, it is possible not only to apply a topological reduction but also operate circuit simplification with a user-given error. We will refer to this simplification like a netlist reduction. To better explain the algorithm flow the circuit depicted in Fig. 1 is considered. So looking at it, if the question is the evaluation of the voltage across the resistance R1 it is possible to replace the resistances R3 and R4, with an unique element RA, equal to the sum of the components R3 and R4, because they are in series. The



$$\begin{bmatrix} \frac{1}{R_1} & -\frac{1}{R_1} & 0 & 1 \\ -\frac{1}{R_1} & \frac{1}{R_1}+\frac{1}{R_2}+\frac{1}{R_3} & -\frac{1}{R_3} & 0 \\ 0 & -\frac{1}{R_3} & \frac{1}{R_3}+\frac{1}{R_4} & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} V\$1 \\ V\$2 \\ V\$3 \\ I\$Vin \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ Vin \end{bmatrix}$$

Fig. 1. Schematic circuit representation and corresponding MNA matrix



$$\begin{bmatrix} \frac{1}{R_1} & -\frac{1}{R_1} & 1 \\ -\frac{1}{R_1} & \frac{1}{R_1}+\frac{R_2+R_3+R_4}{R_2(R_3+R_4)} & 0 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} V\$1 \\ V\$2 \\ I\$Vin \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ Vin \end{bmatrix}$$

Fig. 2. Schematic circuit representation and corresponding MNA matrix after the algorithm application

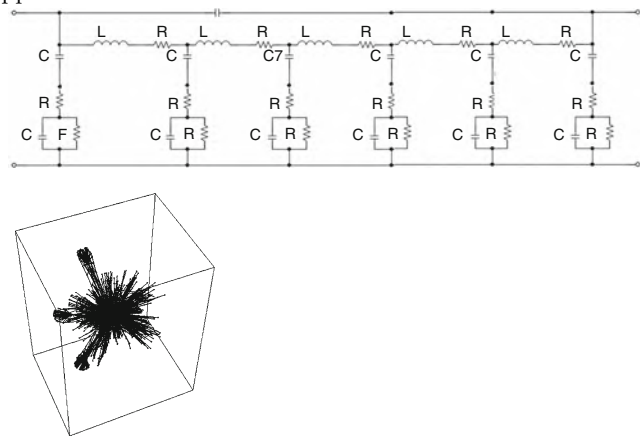


Fig. 3. Schematics of the inductor circuit and second test case

obtained resistance RA is now in parallel with the resistance R2, so they can be replaced by an unique element RB evaluated with the classical topological rules. The resulting circuit and the corresponding matrix are depicted in Fig. 2. Due to the topology reduction the circuit and so the netlist associated to it, is reduced of two components, the corresponding matrix has a row and a column less (a variable less), however, the solution will be still exacts. Looking at the components in series, on the base of the knowledge of the order of magnitude of the components, all the elements less than a fixed value can be described by a short circuit whereas all components connected in parallel having a value higher than the fixed one can be replaced with an open circuit. If a component can be represented like a short circuit the nodes at which it is connected will be compressed in a unique node which connects directly the components linked to it. Because the component disappears it disappears also the parameter associated to it. An open circuit imply instead only the deletion of the parameter associated to it. The obtained circuit will have a reduced number of branches and nodes but also of parameters. As a result not only the order and complexity of the analytical model will be reduced but also the number of parameters to be analyzed, like in a sensitivity analysis.

3 Results

The algorithm has been applied to the equivalent circuit of the inductor considered in Fig. 3. If the MNA representation is used, it yields a 25×25 matrix and 27 parameters. After the application of the algorithm, it is obtained an 8×8 matrix and a number of parameters equal to 4. A more complex circuit, used in industrial application, which represents the interconnection of one macro to another macro-circuit has been then considered (see Fig. 3). The original number of components which are resistances and capacitances, from about 8,900, thanks to the application of the short circuit concept has been reduced to about 5,400, applying the open circuit definition from 5,400 to 4,500 and after the compression of series and parallel in an unique elements to 4,100. The obtained reduction is about 55.

4 Conclusions and Further Work

It has been proposed a netlist reduction algorithm based on the elimination of nodes and branches of a circuit, assuring a user-given accuracy margin. It is based on the decision diagram derived from the circuit topology. After the application of the algorithm the matrix which describes the circuit has a number of parameters as well as of variables and so of equations drastically reduced. The application to different industrial test cases has been introduced, resulting in a reduced system complexity in terms of netlist nodes and model order. The obtained results encourage the application to other industrial test

cases. Other electric circuit properties (such as the T and delta connections, for example)are in implementation and the application to non-linear elements will be exploited.

References

1. Chowdhury, M.H., Amin, C.S., Ismail, Y.I., Kashyap, C.V., Krauter, B.L.: Realizable Reduction of RLC circuit using node elimination IEEE. **3**, 494–497 (2003)
2. Gielen, G., Wambacq, P., Sansen, W.M.: Proc. IEEE **82**(2), 287–304 (1994)
3. Henning, E.: Symbolic Approximation and Modeling Techniques for Analysis and Design of Analog Circuit, Shaker Verlag, Germany (2000)
4. Halfmann, T., Wichmann, T.: Symbolic Methods in Industrial Analog Circuit Design, Scientific Computing in Electrical Engineering (SCEE 2004), Capo D'Orlando Italy, Sep. (2004)

Introduction of Symbolic Simplified Expressions in Circuit Optimization

Angelo Ciccazzo[1], Thomas Halfmann[2], Angelo Marotta[1], Salvatore Rinaudo[1], and Alberto Venturi[2]

[1] ST Microelectronics, Stradale Primosole 50, 95121, Catania, Italy
angelo.marotta@st.com, salvatore.rinaudo@st.com, www.st.com
[2] Fraunhofer-Institut für Techno- und Wirtschaftsmathematik, Fraunhofer-Platz 1 D-67663, Kaiserslautern, Germany
alberto.venturi@itwm.fraunhofer.de,
thomas.halfmann@itwm.fraunhofer.de, www.itwm.fraunhofer.de

Summary. The computation time required for the analysis and optimization of a complete circuit can be too long. The use of symbolic analysis together with model order reduction techniques can reduce it and make the optimization a practicable way in the circuit design. To evaluate the possibilities offered by the technique, firstly a linear test case has been considered. The problem of an inductor simulation has been analyzed by introducing simplified analytical expressions and different optimization algorithms in the fitting/optimization process. Then the optimization technique has been applied to a real circuit, a voltage reference, trying to improve the stability of the reference over the temperature.

1 Introduction

Optimization techniques, symbolic analysis and simplification techniques have been applied to two examples of electronics circuits: firstly to an equivalent lumped circuit of a micro inductor, and then to a band voltage reference. Both of them are important circuital blocks really employed in the nowadays electronics. A band voltage reference (BVR) is a voltage reference that gives an output proportional to the band gap energy of a transistor. The BVR circuit balances the negative temperature coefficient of a pn junction with the positive temperature coefficient of the "thermal voltage": Vt = kT\q. Integrated inductors improve both reliability and efficiency of silicon-integrated RF cells; they can offer circuit solutions with superior performance and contribute to a higher level of integration The inductance of an integrated inductor can be computed exactly by solving Maxwell's equations [1] but to facilitate the design of such components, significant work has gone into modelling spiral inductors using lumped circuit models that takes into account the parasitic resistors and capacitors. Generally, the adopted method to extract model

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*, Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_67,

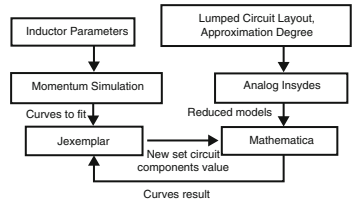


Fig. 1. Inductor Simulation Flow diagram

parameters from an event, is based on the Least Squares Method, that is, minimizing the l^2 norm of the difference between the output and the measured (or required) value, and consists in an optimization problem. In our research work six different optimization methods, both deterministic and stochastic have been tested, in combination with symbolic analysis and simplification techniques, for the fitting of inductor Y parameters. Symbolic analysis is a formal technique to write and solve equations describing circuits behaviour without introducing the numerical value of variables which are symbolically represented. Dimensions of symbolic expression of a circuit actually increase rapidly with the complexity of the network, and in order to make use of the symbolic technique not limited to small circuits, it is necessary to simplify generated expressions by keeping the *dominant terms* only.

2 Inductor Simulation Flow

A flow to permit designers to apply precise inductor model in their integrated circuits simulations was developed by the STMicroelectronics CAD group [8]. Following Fig. 1 it is possible to see how the simulation flow has been modified by introducing the use of JEXEMPLAR, MATHEMATICA and ANALOG INSYDES software in order to obtain and employ an approximated Y parameters symbolic expression. This expression is used by the fitting process in order to find the appropriate set of circuit component values and to reproduce the *target curves*. Now the employed optimization software is JEXEMPLAR; it reads the file with the numerical expression of Y parameters and compares it with the target file, then it generates a new set of component values or ends the process if the fixed approximation is rejoined. MATHEMATICA has the duty to calculate and save the numerical value of Y parameters using the approximated symbolic Y parameters expression obtained by ANALOG INSYDES. ANALOG INSYDES, making use of the MATHEMATICA calculus routines and, receiving as input the description of the inductor lumped circuit, provides the approximated Y parameters expressions. The approximation software offers the possibility to approximate equation systems both before and after the symbolic solution of the systems [10], and both these approximation techniques have been applied.

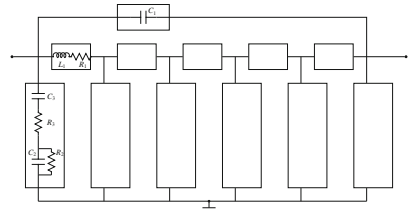


Fig. 2. Inductor layout

Table 1. Lower and upper bounds of the Inductor problem variables

Var.	Lower bound	Upper bound	Var.	Lower bound	Upper bound
Rb1	1.00×10^{-8}	1.00×10^{-4}	Cox1	1.00×10^{-17}	1.00×10^{-13}
Rb2	1.00×10^{-4}	1	Cox2	1.00×10^{-17}	1.00×10^{-13}
Rox1	1.00×10^{-6}	1.00×10^{-2}	Rl	1.00×10^{-4}	1
Rox2	1.00×10^{-4}	1	LL	1.00×10^{-13}	1.00×10^{-9}
Cb1	1.00×10^{-26}	1.00×10^{-22}	Cl	1.00×10^{-21}	1.00×10^{-17}
Cb2	1.00×10^{-25}	1.00×10^{-21}			

The structure of the lumped circuit is fixed, the inductor is represented as an equivalent distributed inductor model with a variable cell number to better describe the inductor behaviour at high frequencies [1,9]. The complete inductor model has the layout illustrated in Fig. 2.

The computer platform used was based on Xeon 3.2 Ghz Intel processor 32 bit, RedHat Enterprise Linux release 3 update 4 operating system. We have tested six of the state-of-the-art optimization algorithms for real world applications; in particular, we use *Controlled Random Search* (CRS) [4], *Controlled Random Search Enhanced* (CRS-E) [5], the *immune algorithm* (optIA) [6] *Powell's algorithm* [2], *Direct method* (Direct) [3] and *Differential Evolution*, DE [7].

3 Simulations

3.1 Inductor

The optimization process has been executed with both approximated and not approximated Y parameter expressions. Our black box function takes as input 11 variables which bounds are presented in Table 1.

Three different simplified expressions of the Y parameters obtained with a maximum relative error of 10^{-1}, 10^{-4} and 10^{-6} have been tested. For each algorithm, we fixed the number of objective function evaluation to 10^4 and a tolerance factor value of $\epsilon = 10^{-12}$; if the algorithm find a configuration with a Y value that differs less than ϵ from the target, it will be stopped. In addiction, for the Direct algorithm, we stop the algorithm if the volume of the hyper rectangle is less than 10^{-12}. Moreover, for the optIA algorithm, we use a population of ten candidate solutions, $d = 10$, a duplication parameter value

equal to two, $dup = 2$, and the elitist selection operator. For DE algorithm, after a series of test to individuate the best option, we set the real factor which controls the amplification of the differential variation to 0.2 and the crossover probability to 0.5. Finally, for DIRECT and POWELL algorithms, we set a initial point where each variable is centred into the given lower and upper bound, instead for CRS, CRS-E and OPTIA we start totally from random points.

By inspecting the results we can note that the best found solution for the circuit using no approximation is clustered at $Y = 4.86 \times 10^{-3}$; excluding DIRECT, every algorithm reaches this solution and the difference between each algorithm is negligible.

Analyzing the results using the approximated function at different level of tolerance, Table 2, we can observe that the algorithms are robust and the results scale accordingly to the fixed tolerance; in Fig. 3, we show the convergence plot. Each algorithm found a solution that generally differs from the one obtained using a non approximated model of 0.8×10^{-3} and this suggests that the model is accurate enough to perform a complete optimization of the sizing of the circuit using only the symbolic model.

From an optimization point of view, we can infer that the produced approximated expression can be optimized with both deterministic and stochastic algorithms obtaining comparable results; instead, by inspecting the convergence plots, we can note that the stochastic approach guarantees a better speed of convergence than the deterministic ones.

3.2 Band Voltage Reference

The objective of the optimization with the BVG circuit was to stabilize, working on the project parameters, the reference voltage to 650 mv over a temperature interval between −40 and +125°C, and with two different values of the voltage supply 0.9 and 1.2 V. The optimization variables are the physical dimensions of the circuit components; designers suggested a first list of 32 parameters, then a sensitivity analysis made with the functionalities of ELDO by MENTOR GRAPHICS highlighted the three most important variables to set the output voltage. The variables have been considered continuous. The result of the sensitivity analysis is a consequence of the circuit structure: the voltage

Table 2. Performance of the optimization algorithms using the symbolic model with different tolerance settings

Algorithm	10^{-6}	10^{-4}	10^{-1}
CRS	$\mathbf{4.871 \times 10^{-3}}$	$\mathbf{5.111 \times 10^{-3}}$	5.621×10^{-3}
CRS-E	$\mathbf{4.871 \times 10^{-3}}$	5.113×10^{-3}	5.621×10^{-3}
POWELL	4.871×10^{-3}	5.137×10^{-3}	5.624×10^{-3}
DIRECT	9.828×10^{-3}	8.477×10^{-3}	9.144×10^{-3}
OPTIA	4.872×10^{-3}	5.134×10^{-3}	$\mathbf{5.616 \times 10^{-3}}$
DE	4.875×10^{-3}	5.308×10^{-3}	$\mathbf{6.016 \times 10^{-3}}$

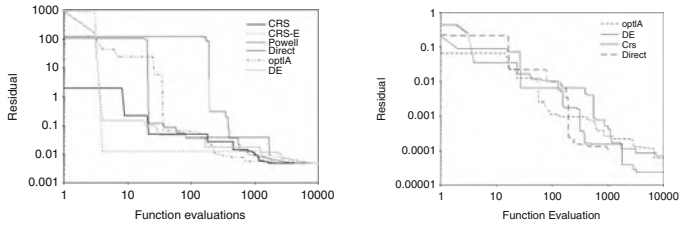


Fig. 3. *Left*: Inductor convergence plot with the error set to 10^{-6}. CRS algorithm obtains the best convergence curve. *Right*: BVR convergence plot

Table 3. Inductor : performance of the optimization algorithms without approximations

Algorithm	Best Residual	Algorithm	Best Residual
DE	$\mathbf{2.402 \times 10^{-5}}$	CRS	6.9564×10^{-5}
optIA	6.0517×10^{-5}	Direct	1.1019×10^{-4}

reference is generated by injecting a controlled current on the resistors. In this configuration the role of the resistors is crucial.

By inspecting the results, Table 3, it is possible to see that the performances of all the algorithms, with the exception of Direct, are good; DE showed to be able to get a results slightly better. In Fig. 3.1 is showed the convergence plot for the algorithms tested. In Fig. 4 is plotted the curve voltage vs temperature for the circuit both optimized that not optimized. It is possible to see how the The optimization has reduced the peak to peak value of the curve but has also moved maximum versus lower temperature, instead of keeping it at 27°C.

4 Conclusion

The possibility to introduce simplification techniques in the optimization flow of an inductor simulation has been evaluated; as a test case was used the spiral inductor. The symbolic circuit simulator Analog Insydes was linked with an optimization framework, which take into account deterministic and stochastic optimization algorithms. The results show that it has been possible to find solutions using the simplified parameter expressions comparable to the solutions obtained without approximated expressions. It has been proved that the time is dramatically decreased. In order to extend the same strategy to more complicate circuits, we started by verifying the utility of the application of optimization techniques to a band voltage reference circuit. The result obtained has showed the possibility to improve the designers job; future works are led to complete the flow by applying the symbolic analysis also to similar circuits.

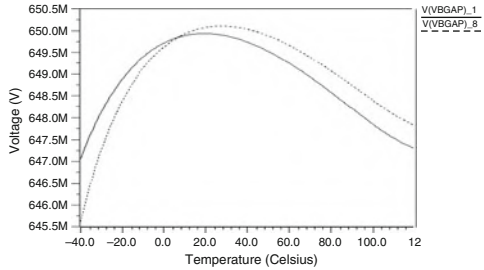


Fig. 4. Voltage vs temperature, comparison between optimized solution (*continuous line*)and starting one *dotted line*)

Acknowledgments

The present research work was supported by Symbolic Techniques for Circuit Optimization – Symteco Marie Curie Fellowships for the Transfer of Knowledge (ToK). ST Microelectronics Catania – Italy, and Fraunhofer Institute for Industrial Mathematics (ITWM) in Kaiserslautern – Germany.

References

1. Mohan, S.S., del Mar Hershenson, M., Boyd, S.P., Lee, T.H.: IEEE J. Solid State Circuits **34**(10), 1419–1424 (1999)
2. Powell, M.J.D.: Chapter of Nonlinear Programming, pp. 31–65. Academic, New York (1970)
3. Jones, D.R., Perttunen, C.D., Stuckman, B.E.: J. Optim. Theory Appl. **1**(79), 157–181 (1993)
4. Price, W.L.: Comput. J. Br. Comput. Soc. **4**(20), 367 (1977)
5. Brachetti, P., De Felice Ciccoli, M., Di Pillo, G., Lucidi, S.: J. Global Optim. **2**(10), 165–184 (1997)
6. Cutello, V., Nicosia, G., Pavone, M.: Proceedings of the 2006 ACM Symposium on Applied Computing. ACM, New York (2006)
7. Storn, R., Price, K.: INTERNATIONAL COMPUTER SCIENCE INSTITUTE-PUBLICATIONS-TR, 1995 - icsi.berkeley.edu, (1995)
8. Ciccazzo, A., Greco, G., Rinaudo, S.: Scientific Computing in Electrical Engineering, vol. 9, pp. 317–322. Springer, New York (2006)
9. Nieuwoudt, A., Massoud, Y. Proceedings of the 42nd Annual Conference on Design automation, pp. 648–651. New York, NY, USA (2005)
10. Sommer, R., Hennig, E., Droge, G., Horneber: Alta Frequenza-RIVISTA diELETTRONICA, E.-H.: **5**(6), 29–37 (November 1993)

Proper Orthogonal Decomposition Model Order Reduction of Nonlinear IC Models

A. Verhoeven[1], M. Striebel[2], J. Rommes[3], and E.J.W. ter Maten[1,3], and T. Bechtold[3]

[1] Eindhoven University of Technology, The Netherlands, Arie.Verhoeven@na-net.ornl.gov
[2] Technische Universität Chemnitz, Germany, Michael.Striebel@nxp.com
[3] NXP Semiconductors, Eindhoven, The Netherlands, Jan.ter.Maten@nxp.com, Joost.Rommes@nxp.com, Tamara.Bechtold@nxp.com

Summary. We demonstrate Model Order Reduction for a nonlinear system of differential-algebraic equations of a diode chain by Proper Orthogonal Decomposition with Adapted Missing Point Estimation. The collected time snapshots also allow for an efficient impression of the sensitivity of objective functions.

1 Introduction

Future simulation for nanoelectronics requires that circuit equations can be coupled to electromagnetics, to semiconductor equations, and to heat transfer. The consequence is that one has to deal with large systems. Model Order Reduction (MOR) is a means to speed up simulation of large systems. Existing MOR techniques mostly apply to linear problems and even then they have to be generalized to become applicable to a resulting system of (Partial) Differential-Algebraic Equations (DAEs, PDAES). To make MOR applicable to industrial applications one has to address nonlinearity and parameterization. Here we consider Proper Orthogonal Decomposition (POD) to reduce the system size. An adaption is presented to also reduce the complexity in evaluating functions and Jacobian matrices.

The problem of reducing nonlinear systems can be described as follows. Given a, possibly large-scale, nonlinear time-invariant dynamical system $\Sigma = (\mathbf{g}, \mathbf{f}, \mathbf{h}, \mathbf{x}, \mathbf{u}, \mathbf{y}, t)$

$$\Sigma = \begin{cases} \frac{d\mathbf{g}(\mathbf{x}(t))}{dt} = \mathbf{f}(\mathbf{x}(t), \mathbf{u}(t)) \\ \mathbf{y}(t) = \mathbf{h}(\mathbf{x}, \mathbf{u}) \end{cases}$$

where $\mathbf{x}(t) \in \mathbb{R}^n$, $\mathbf{u}(t) \in \mathbb{R}^m$, $\mathbf{y}(t) \in \mathbb{R}^p$, $\mathbf{f}(\mathbf{x}(t), \mathbf{u}(t)), \mathbf{g}(\mathbf{x}(t)) \in \mathbb{R}^n$, $\mathbf{h}(\mathbf{x}(t), \mathbf{u}(t)) \in \mathbb{R}^p$, find a reduced model $\widetilde{\Sigma} = (\tilde{\mathbf{g}}, \tilde{\mathbf{f}}, \tilde{\mathbf{h}}, \tilde{\mathbf{x}}, \mathbf{u}, \tilde{\mathbf{y}}, t)$

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*, Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_68,

$$\widetilde{\Sigma} = \begin{cases} \frac{d\tilde{\mathbf{g}}(\tilde{\mathbf{x}}(t))}{dt} = \tilde{\mathbf{f}}(\tilde{\mathbf{x}}(t), \mathbf{u}(t)) \\ \tilde{\mathbf{y}}(t) = \tilde{\mathbf{h}}(\tilde{\mathbf{x}}, \mathbf{u}) \end{cases}$$

where $\tilde{\mathbf{x}}(t) \in \mathbb{R}^r$, $\mathbf{u}(t) \in \mathbb{R}^m$, $\tilde{\mathbf{y}}(t) \in \mathbb{R}^p$, $\tilde{\mathbf{f}}(\tilde{\mathbf{x}}(t), \mathbf{u}(t)), \tilde{\mathbf{g}}(\tilde{\mathbf{x}}(t)) \in \mathbb{R}^r$, $\tilde{\mathbf{h}}(\tilde{\mathbf{x}}(t), \mathbf{u}(t)) \in \mathbb{R}^p$, such that $\tilde{\mathbf{y}}(t)$ can be computed in much less time than $\mathbf{y}(t)$ and the approximation error $\mathbf{y}(t) - \tilde{\mathbf{y}}(t)$ is small.

In the context of circuit simulation the dynamical systems we are dealing with are circuit blocks or subcircuits. Connection to and communication with a block's environment is done via its terminals, i.e. external nodes. Therefore, we can assume that the currents or voltages are always injected linearly into the circuit under consideration. A similar reasoning applies for the determination of the output signal $\mathbf{y}(t)$, which is also assumed to be not explicitly dependent on the input $\mathbf{u}(t)$. Hence, in the remainder of this document, we assume the dynamical systems to be of the form

$$\Sigma = \begin{cases} \frac{d\mathbf{g}(\mathbf{x}(t))}{dt} = \mathbf{f}(\mathbf{x}(t)) + \mathbf{B}\mathbf{u}(t) \\ \mathbf{y}(t) = \mathbf{C}^T\mathbf{x} \end{cases}$$

where $\mathbf{B} \in \mathbb{R}^{n\times m}$ and $\mathbf{C} \in \mathbb{R}^{n\times p}$.

The two best-known methods for reduction of nonlinear systems are Proper Orthogonal Decomposition (POD), and Trajectory PieceWise-Linear techniques (TPWL) [4, 6, 7] (and references cited there).

2 Proper Orthogonal Decomposition (POD)

Proper Orthogonal Decomposition extends the Petrov–Galerkin projection based methods that are used for linear systems to nonlinear system. By choosing a suitable $\mathbf{V} \in \mathbb{R}^{n\times r}$ and a test matrix $\mathbf{W} \in \mathbb{R}^{n\times r}$, where $\mathbf{W}$ and $\mathbf{V}$ are biorthonormal, i.e., $\mathbf{W}^T\mathbf{V} = \mathbf{I}_{r\times r}$, $r \leq n$, the reduced system is given by

$$\begin{cases} \mathbf{W}^T \frac{d\mathbf{g}(\mathbf{V}\tilde{\mathbf{x}}(t))}{dt} = \mathbf{W}^T\mathbf{f}(\mathbf{V}\tilde{\mathbf{x}}(t)) + (\mathbf{W}^T\mathbf{B})\mathbf{u}(t) \\ \tilde{\mathbf{y}}(t) = (\mathbf{C}^T\mathbf{V})\tilde{\mathbf{x}} \end{cases}$$

Similar to linear model order reduction, the idea is that $\mathbf{V}$ captures the dominant dynamics, i.e., the states of the original system are approximated well by $\mathbf{V}\tilde{\mathbf{x}} \approx \mathbf{x}$. The test matrix $\mathbf{W}$ is chosen such that the Petrov–Galerkin condition $\mathbf{r} = \frac{d\mathbf{g}(\mathbf{V}\tilde{\mathbf{x}}(t))}{dt} - \mathbf{f}(\tilde{\mathbf{x}}(t)) - \mathbf{B}\mathbf{u}(t) \perp \mathbf{W}$ is met.

POD constructs the matrix $\mathbf{V}$ as follows. A time domain simulation of the complete system is done and snapshots of the states at suitably chosen times t_i are collected in the state matrix $\mathbf{X}$

$$\mathbf{X} = [\mathbf{x}(t_0),\ \mathbf{x}(t_1),\ \mathbf{x}(t_2),\ \cdots \mathbf{x}(t_{N-1})] \in \mathbb{R}^{n\times N},$$

where N is the number of time points t_i. To extract the subspace that represents that dominant dynamics, the singular value decomposition of $\mathbf{X}$ is computed $\mathbf{X} = \mathbf{U}\Sigma\mathbf{T}$ where $\mathbf{U} \in \mathbb{R}^{n\times n}$, $\Sigma = [\mathrm{diag}(\sigma_1, \ldots, \sigma_n)\ 0_{n\times(N-n)}] \in$

$\mathbb{R}^{n\times N}$ (if $N > n$), and $\mathbf{T} \in \mathbb{R}^{N\times N}$. Let the singular values $\sigma_1 \geq \sigma_2 \cdots \sigma_r \gg \sigma_{r+1} > \cdots > \sigma_n$ be ordered in decreasing magnitude. POD chooses the matrix $\mathbf{V}$ to have as its columns the left singular vectors corresponding to the $r \ll n$ largest singular values

$$\mathbf{V} = [\mathbf{u}_1,\ \mathbf{u}_2, \cdots, \mathbf{u}_r] \in \mathbb{R}^{n\times r}.$$

The number k of vectors to choose can depend on a tolerance based criterion like $\sigma_{k+1} < \epsilon$, or on the relative difference between σ_k and σ_{k+1}. The test matrix $\mathbf{W}$ is taken as $\mathbf{W} = \mathbf{V}$, i.e., the residual is orthogonal to the reduced state space.

We stress that the reduction obtained from POD and similar projection based methods is solely in the number of states: r for the reduced systems vs. n for the original system and $r \ll n$. However, the costs for evaluating nonlinear terms such as $\mathbf{W}^T\mathbf{f}(\mathbf{V}\tilde{\mathbf{x}}(t))$ (and associated Jacobian matrices) will be larger than for the original system. Hence with respect to simulation times no reduction may be obtained unless additional measures are taken.

3 Missing Point Estimation/Adapted POD

We will present some results computed with the Missing Point Analysis/Adapted POD approach described in [3–5]. We reflect the basic idea with the case of a simple ODE

$$\frac{d}{dt}\mathbf{x} = \mathbf{f}(\mathbf{x}),$$

of dimension n with nonlinear right hand side $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^n$. The singular value decomposition $\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^T$ of a matrix $\mathbf{X} \in \mathbb{R}^{n\times N}$ of N snapshots is computed, giving n singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n$. The orthogonal matrix $\mathbf{L} = \mathbf{U} \cdot \mathrm{diag}(\sigma_1, \ldots, \sigma_n) \in \mathbb{R}^{n\times n}$ is introduced, with its columns $\mathbf{l}_1, \ldots, \mathbf{l}_n$ spanning the complete space $\mathbb{R}^n$. Hence, one can change to the new basis, i.e., $\mathbf{x} = \mathbf{L}\mathbf{y}$ and apply a Galerkin-like projection to the system

$$\mathbf{L}^T \frac{d}{dt}(\mathbf{L}\mathbf{y}) = \mathbf{L}^T\mathbf{f}(\mathbf{L}\mathbf{y}). \tag{1}$$

Strictly speaking we do not apply Galerkin projection as the columns of L are orthogonal, but not orthonormal.

Classical POD reduction acts on $\mathbf{x} = \mathbf{L}\mathbf{y}$ in the sense that the expansion of $\mathbf{x}$ in the basis $\mathbf{l}_1, \ldots, \mathbf{l}_n$ where $(\mathbf{l}_1, \ldots, \mathbf{l}_n) = \mathbf{L} = (\sigma_1 \cdot \mathbf{v}_1, \ldots, \sigma_n \cdot \mathbf{v}_n)$ with $(\mathbf{v}_1, \ldots, \mathbf{v}_n) = \mathbf{U}$ is truncated with respect to the magnitude of the singular values $\sigma_1, \ldots, \sigma_n$:

$$\begin{aligned}
\mathbf{x} = \mathbf{L}\mathbf{y} &= (\sigma_1\mathbf{v}_1)\cdot y_1 + \cdots + (\sigma_r\mathbf{v}_r)\cdot y_r + (\sigma_{r+1}\mathbf{v}_{r+1})\cdot y_{r+1} + \cdots + (\sigma_n\mathbf{v}_n)\cdot y_n \\
&\approx (\sigma_1\mathbf{v}_1)\cdot y_1 + \cdots + (\sigma_r\mathbf{v}_r)\cdot y_r + 0\cdot y_{r+1} + 0\cdot y_n \\
&= (\mathbf{l}_1, \ldots, \mathbf{l}_r, 0, \ldots, 0)\cdot\mathbf{y} \\
&= (\mathbf{L}\mathbf{P}_r^T\mathbf{P}_r)\cdot\mathbf{y}, \quad \text{with } \mathbf{P}_r = \left(\mathbf{I}_{r\times r}\ \mathbf{0}_{r\times(n-r)}\right) \in \{0,1\}^{r\times n} \\
&= (\mathbf{L}\mathbf{P}_r^T)\cdot(\mathbf{P}_r\mathbf{y}) = (\mathbf{L}\mathbf{P}_r^T)\cdot\mathbf{z}_r \quad \text{with } \mathbf{z}_r = (y_1, \ldots, y_r)^T \in \mathbb{R}^r
\end{aligned}$$

where r usually is chosen in such a way that $\sigma_{r+1} < \text{TOL}$ or $\sigma_{r+1} \ll \sigma_r$.

This procedure can also be interpreted as keeping the r most "dominant" columns of L and neglecting the rest, where a column's norm is taken as a criterion. That means, $\mathbf{L}$ is approximated by

$$\mathbf{L} \approx \mathbf{L}\mathbf{P}_r^T\mathbf{P}_r, \quad \text{with } \mathbf{P}_r \in \{0,1\}^{r\times n}. \tag{2}$$

where $\mathbf{P}_r = \left(\mathbf{I}_{r\times r}\ \mathbf{0}_{r\times(n-r)}\right)$ selects these columns. By construction of $\mathbf{L} = \mathbf{U}\cdot\text{diag}(\sigma_1,\ldots,\sigma_n)$, where $\mathbf{U}^T\mathbf{U} = \mathbf{I}_{n\times n}$, we have $\|\mathbf{v}_i\|_2 = \sigma_i$ for $i = 1,\ldots,n$. In this respect the r most dominant columns are therefore $\mathbf{l}_1,\ldots,\mathbf{l}_r$.

In the adapted POD presented in [4] this perception is carried over to the transposed $\mathbf{L}^T$. That means, one selects, again based on the norms, the $g \in \mathbb{N}$ most dominant columns $\{\tilde{\mathbf{l}}_{\mu_1},\ldots,\tilde{\mathbf{l}}_{\mu_g}\}$ of $\mathbf{L}^T = (\tilde{\mathbf{l}}_1,\ldots,\tilde{\mathbf{l}}_n)$ and neglects the rest:

$$\mathbf{L}^T \approx \mathbf{L}^T\mathbf{P}_g^T\mathbf{P}_g, \quad \text{with } \mathbf{P}_g \in \{0,1\}^{g\times n}. \tag{3}$$

First, these approximations to $\mathbf{L}$ and $\mathbf{L}^T$ from (2) and (3), respectively, are inserted into (1):

$$\mathbf{L}^T\mathbf{P}_g^T\mathbf{P}_g\frac{d}{dt}(\mathbf{L}\mathbf{P}_r^T\mathbf{P}_r\mathbf{y}) = \mathbf{L}^T\mathbf{P}_g^T\mathbf{P}_g\mathbf{f}(\mathbf{L}\mathbf{P}_r^T\mathbf{P}_r\mathbf{y}) \tag{4}$$

From (2) and (3) it follows that

$$\mathbf{L}^T \approx \mathbf{P}_r^T\mathbf{P}_r\mathbf{L}^T\mathbf{P}_g^T\mathbf{P}_g,$$

and multiplying with $\mathbf{P}_r$ (consider $\mathbf{P}_r\mathbf{P}_r^T = \mathbf{I}_{r\times r}$), the system (4) turns into

$$\mathbf{P}_r\mathbf{L}^T\mathbf{P}_g^T\mathbf{P}_g\frac{d}{dt}(\mathbf{L}\mathbf{P}_r^T\mathbf{P}_r\mathbf{y}) = \mathbf{P}_r\mathbf{L}^T\mathbf{P}_g^T\mathbf{P}_g\mathbf{f}(\mathbf{L}\mathbf{P}_r^T\mathbf{P}_r\mathbf{y})$$

As $\mathbf{L}\mathbf{P}_r^T = (\sigma_1\mathbf{v}_1,\ldots,\sigma_r\mathbf{v}_r) = \mathbf{U}_r\Sigma_r$ (for $\mathbf{U}_r = (\mathbf{v}_1,\ldots,\mathbf{v}_r)$, $\Sigma_r = \text{diag}(\sigma_1,\ldots,\sigma_r)$) we get

$$\Sigma_r\mathbf{U}_r^T\mathbf{P}_g^T\frac{d}{dt}[\mathbf{P}_g\mathbf{U}_r\Sigma_r\mathbf{P}_r\mathbf{y}] = \Sigma_r\mathbf{U}_r^T\mathbf{P}_g^T\mathbf{P}_g\mathbf{f}(\mathbf{U}_r\Sigma_r\mathbf{P}_r\mathbf{y}), \quad \mathbf{L}\mathbf{y} = \mathbf{x}.$$

The above equation states a system of dimension r for $\mathbf{y} \in \mathbb{R}^n$. Therefore, we introduce the reduced state vector $\mathbf{y}_r = \Sigma_r\mathbf{P}_r\mathbf{y} \in \mathbb{R}^r$ from which we can approximately reconstruct the coefficients of the full state in the basis spanned by the columns of $\mathbf{L}$ by $\mathbf{y} \approx \mathbf{P}_r^T\Sigma_r^{-1}\mathbf{y}_r$. This in turn lets us approximate the full state in the original basis $\mathbf{x} \approx \mathbf{U}_r\mathbf{y}_r$, because $\mathbf{x} = \mathbf{L}\mathbf{y} \approx \mathbf{L}\mathbf{P}_r^T\Sigma_r^{-1}\mathbf{y}_r = \mathbf{U}_r\Sigma_r\Sigma_r^{-1}\mathbf{y}_r$. This part is consistent with the classical POD.

In addition to the reduction in the state space the adapted POD downsizes $\mathbf{f}(\cdot)$ by considering that the term $\mathbf{P}_g\mathbf{f}(\cdot)$ corresponds to just including the g components $f_{\mu_1}(\cdot),\ldots,f_{\mu_g}(\cdot)$ of $\mathbf{f}(\cdot) = (f_1(\cdot),\ldots,f_r(\cdot))^T$. Hence, it suffices to evaluate the g-dimensional function

$$\bar{\mathbf{f}} : \mathbb{R}^n \to \mathbb{R}^g : \mathbf{x} \mapsto (f_{\mu_1}(\mathbf{x}),\ldots,f_{\mu_g}(\mathbf{x}))^T.$$

After scaling with Σ_r^{-1} the reduced system for the reduced state vector $\mathbf{y}_r \in \mathbb{R}^r$ becomes

$$\mathbf{U}_r^T \mathbf{P}_g^T \frac{d}{dt}[\mathbf{P}_g \mathbf{U}_r \mathbf{y}_r] = \mathbf{U}_r^T \mathbf{P}_g^T \bar{\mathbf{f}}(\mathbf{U}_r \mathbf{y}_r), \quad \mathbf{x} = \mathbf{U}_r \mathbf{y}_r \tag{5}$$

For the general case of having not an ODE (1) but a DAE

$$\frac{d}{dt}\mathbf{g}(\mathbf{x}) = \mathbf{f}(\mathbf{x}) + \mathbf{B}\mathbf{v}$$

to deal with, one gets a reduced problem

$$\mathbf{U}_r^T \mathbf{P}_g^T \frac{d}{dt}\bar{\mathbf{g}}(\mathbf{U}_r \mathbf{y}_r) = \mathbf{U}_r^T \mathbf{P}_g^T \bar{\mathbf{f}}(\mathbf{U}_r \mathbf{y}_r) + \mathbf{U}_r^T \mathbf{B}\mathbf{v}. \tag{6}$$

with $\bar{\mathbf{g}} : \mathbb{R}^n \to \mathbb{R}^g : \mathbf{x} \mapsto (g_{\mu_1}(\mathbf{x}), \ldots, g_{\mu_g}(\mathbf{x}))^T$.

We end this section with the observation that the collected time snapshots for POD also allow for an efficient first impression of the sensitivity of several objective functions (like consumed power) even in the case of many parameters [2].

4 POD Testcase: Diodechain

We consider the diode chain model shown in Fig. 1 (with the parameters I_s, V_T, R, C). Here the diode functionality is modelled by the current function $g(V_a, V_b)$ and the input function by $U_{\text{in}}(10^9 t)$, for $t \leq 70\,\text{ns}$, see [3–5],

$$g(V_a, V_b) = \begin{cases} I_s(e^{\frac{V_a - V_b}{V_T}} - 1) & \text{if } V_a - V_b > 0.5 \\ 0 & \text{otherwise} \end{cases} \quad U_{\text{in}}(\tau) = \begin{cases} 20 & \text{if } \tau \leq 10 \\ 170 - 15\tau & \text{if } 10 < \tau \leq 11 \\ 5 & \text{if } \tau > 11 \end{cases}$$

The state of the diode chain model consists of 302 elements but there is a lot of redundancy. The numerical solution (nodal voltage in each node) on the time interval $[0, 70\,\text{ns}]$ is computed by the Euler Backward method with fixed stepsizes of 0.1 ns. The full system was run in 42.01 s. Classic POD needed 35.51 s. The POD with Adapted MPE (reducing the state space to $r = 30$ and downsizing evaluations to $g = 35$), only required 5.12 s. No visible error can be seen in the approximative results (Fig. 2 (left)).

If the input changes to $7.5\cos(\frac{2\pi t}{60 \cdot 10^{-9}}) + 12.5$ this impression is confirmed (full system 40.22 s, Classic POD even 45.34 s, POD with Adapted POD 6.28 s; Fig. 2 (right)). This makes POD ca five times slower then TPWL, but much more accurate and more robust [3]. If we further increase the amplitude of the cosine to 9.5 POD is not able to properly recover the regions with higher amplitudes (but neither is TPWL) [5].

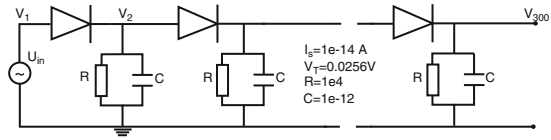


Fig. 1. Schematic of diode chain

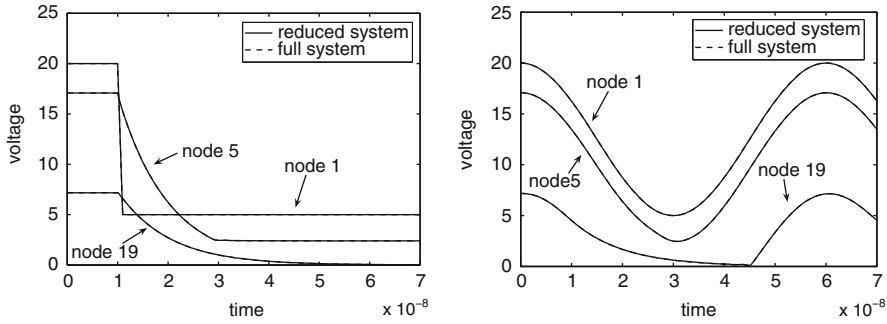


Fig. 2. *Left*: identical input; *Right*: changed input

References

1. Ciuprina, G., Ioan, D. (eds.): Scientific Computing in Electrical Engineering SCEE 2006, Series Mathematics in Industry, vol. 11. Springer, Berlin (2007)
2. Ilievski, Z., Xu, H., Verhoeven, A., ter Maten, E.J.W., Schilders, W.H.A., Mattheij, R.M.M.: Adjoint Transient Sensitivity Analysis in Circuit Simulation. In: [1], pp. 183–189. (2007)
3. Verhoeven, A., Voss, T., Astrid, P., ter Maten, E.J.W., Bechtold, T.: Model order reduction for nonlinear problems in circuit simulation, Proc. ICIAM-2007 in PAMM (Proc. in Appl. Maths. and Mech.), vol. 7(1), pp. 1021603–1021604, 2007/2008, DOI: 10.1002/pamm.200700537 (publ. online Sept. 18 2008)
4. Verhoeven, A., ter Maten, J., Striebel, M., Mattheij, R.: Model order reduction for nonlinear IC models. In: Korytowski, A., Malanowski, K., Mitkowski, W., Szymkat, M. (eds.): System Modeling and Optimization, Series IFIP AICT 312, pp. 476–491. Springer (2009)
5. Verhoeven, A., Striebel, M., ter Maten, E.J.W.: Model Order Reduction for nonlinear IC models with POD. In: Roos, J., Costa, L.R.J. (eds.): Scientific Computing in Electrical Engineering SCEE 2008, Series Mathematics in Industry, vol. 14, Springer (2010)
6. Voss, T., Pulch, R., ter Maten, E.J.W., El Guennouni, A.: Trajectory Piecewise Linear Approach for Nonlinear Differential-Algebraic Equations in Circuit Simulation. In: [1], pp. 167–173. (2007)
7. Voss, T., Verhoeven, A., Bechtold, T., ter Maten, J.: Model Order Reduction for Nonlinear Differential Algebraic Equations in Circuit Simulation. In: Bonilla, L.L., Moscoso, M., Platero, G., Vega, J.M. (eds.): Progress in Industrial Mathematics at ECMI 2006, Series Mathematics in Industry, vol. 12, pp. 518–523. Springer, Berlin (2007). ISBN 978-3-540-71991-5

Surrogate Modeling of RF Circuit Blocks

Luciano De Tommasi[1], Dirk Gorissen[2], Jeroen A. Croon[3], and Tom Dhaene[2]

[1] Energy Research Center of the Netherlands,NL-1755ZG Petten, the Netherlands, luciano.de.tommasi@gmail.com.
[2] Ghent University – IBBT, Department of Information Technology (INTEC), Gaston Crommenlaan 8 Bus 201, B-9050 Ghent, Belgium, {dirk.gorissen,tom.dhaene}@ugent.be.
[3] NXP-TSMC Research Center, High Tech Campus 37, PostBox WY4-01, NL-5656AE Eindhoven, the Netherlands, jeroen.croon@nxp.com.

Summary. Surrogate models are a cost-effective replacement for expensive computer simulations in design space exploration. Literature has already demonstrated the feasibility of accurate surrogate models for single radio frequency (RF) and microwave devices. Within the European Marie Curie project O-MOORE-NICE! (Operational Model Order Reduction for Nanoscale IC Electronics) we aim to investigate the feasibility of the surrogate modeling approach for entire RF circuit blocks. This paper presents an overview about the surrogate model type selection problem for low noise amplifier modeling.

1 Introduction

Design space exploration of RF circuit blocks involves the solution of constrained multiobjective optimization problems in order to fulfill the performance specifications and perform what-if and sensitivity analysis. Optimization demands a large amount of circuit simulations so making the whole process very expensive.

We aim to develop scalable (parametrized) models of (non-linear) RF circuit blocks. This problem is too hard to be addressed applying model order reduction techniques. Furthermore, model order reduction is a model-driven approach: it needs the mathematical description of the system, which is not available when model equations are embedded into the circuit simulator. On the other hand, surrogate modeling is a data-driven approach, which does not make any analytical assumptions upon the model which has to be reduced. The simulator is seen as a 'black box' which accepts input samples and provides output samples, see Fig. 1. Basing upon such samples, a cheap-to-evaluate surrogate model is trained. The surrogate has to be able to predict the outputs given by the simulator when a new input (which has not been

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*, Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_69,

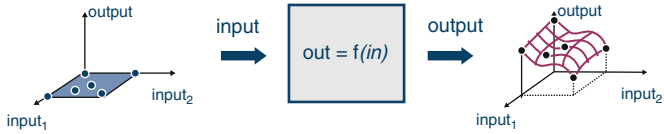


Fig. 1. Surrogate modeling approach

used in the model training) is applied. This means that surrogate modeling is equivalent to construct the surface which fits the samples in Fig. 1 (in fact, surrogate models are also known as response surface models).

Accurate surrogate models for single RF and microwave components have been already developed (e.g. using ANNs [1]). In this research activity, we aim to model complete RF circuit blocks. The first considered circuit block is a low noise amplifier (LNA) [2]. Other RF circuit blocks (e.g. mixers, VCOs, etc) can be analyzed following the same approach.

The behavior of an LNA is described by means of the admittance and noise functions, which are evaluated via accurate transistor-level simulations. Such functions are used to compute the performance figures (gain, input impedance, noise figure and power consumption) used by designers.

Each circuit simulation typically requires 1–2 min, which is a too long time to effectively explore how performance figures of LNA scale with key circuit-design parameters, such as the dimensions of transistors, passive components, signal properties and bias conditions. Therefore, the transistor level model can be usefully replaced with an accurate surrogate model (based on transistor level simulations) which is much cheaper to evaluate.

As first step towards an effective surrogate modeling process, the user of a modeling software environment has to find out which model type (among those ones available [3]) work better with his problem.

In this paper we summarize the results of a surrogate model type comparison for the LNA modeling problem.

2 Software Environment

The surrogate modeling approach developed in this paper, is based on the SUrrogate MOdeling (SUMO) Matlab Toolbox [6]. The modeling flow is shown in Fig. 2. It is based on *adaptive modeling* and *adaptive sampling* loops.

The surrogate modeling process starts with the evaluation of an initial design (e.g. Latin hypercube, Box–Behnken, etc) which uniformly fills the design space (the number of samples is specified by the user). Based on this initial set of samples, one or more surrogate models are constructed. Adaptive modeling [7] implies that a suitable optimization algorithm (e.g. hill climbing, particle swarm, genetic algorithm, DIRECT, etc) is used to tune relevant model hyperparameters, in order to minimize the error between model and

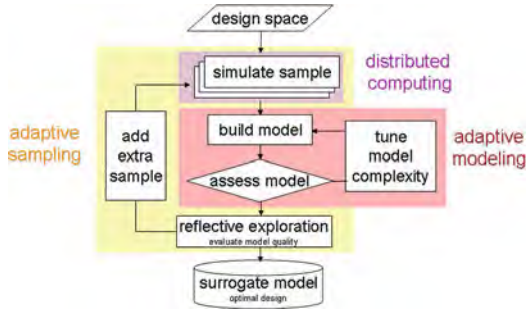


Fig. 2. Surrogate modeling flow

data. Model error is evaluated according to one or more measures and functions. Afterwards, the models are ranked according to their score, and the best model is selected.

In order to improve the accuracy, an adaptive sampling algorithm selects new samples based on the best performing models and the behavior of the reference function. In this work we applied the *gradient-based* method [6] because it has shown good performances with the LNA modeling problem.

After each sampling iteration, an adaptive modeling iteration including the new samples is started, and the whole process repeats itself until one of the following three conditions is satisfied: (1) the maximum allowed number of samples (specified by the user) has been reached, (2) the maximum allowed modeling time has been exceeded, or (3) the user required accuracy has been met.

3 Surrogate Model Type Selection

Our feasibility study about surrogate modeling of low noise amplifiers aims to determine how many design variables can be included in a model and how many samples (number of simulations) are needed to generate the model.

A model is considered sufficiently accurate when its root relative square error is lower than 0.05. The maximum number of samples allowed is 1,500.

Several surrogate model types have been compared (as implemented in the SUMO Toolbox): artificial neural networks, rational functions, radial basis functions, least squares support vector machines and kriging [4,5].

In order to reduce the computational cost of such comparison, transistor level simulations have been replaced with an analytical model of the LNA[4] [4]. Although the accuracy of such analytical model is obviously not sufficient to replace the circuit simulator in the design process, it satisfactorily reproduces the shape of the simulator outputs.

[4]Moreover, a cluster of PC has been used to build the surrogates.

Design parameters (input parameters) are, in order of importance: transistor width W, source inductance L_s, frequency f, transistor length L, gate threshold voltage V_{GT}, gate series inductance L_m. The modeling software works with the following normalized parameters (characterized by the subscript 'n') which lie in the interval [−1,1]: $W = 100 \cdot 10^{-6} \cdot 10^{W_n}\ m$, $L_s = 0.5 \cdot 10^{-9} \cdot 10^{L_{sn}}\ H$, $f = (11 + 10 \cdot f_n) \cdot 10^9\ Hz$, $L = (90 + 30 \cdot L_n) \cdot 10^{-9}$ m, $V_{GT} = 0.275 + 0.2 \cdot V_{GTn}\ V$, $L_m = 1 \cdot 10^{-9} \cdot 10^{L_{mn}}\ H$. As output parameters, we consider the admittances y_{11}, y_{12}, the input/output noise currents $\sqrt{\overline{i_{in}^2}}$, $\sqrt{\overline{i_{out}^2}}$ and their correlation ρ. If a (normalized) input parameter is not taken into account in the modeling, it is clamped to 0, the exception being L which is clamped to -1.

Examples of accurate surrogate models are shown in Figs. 3, 4 and 5.

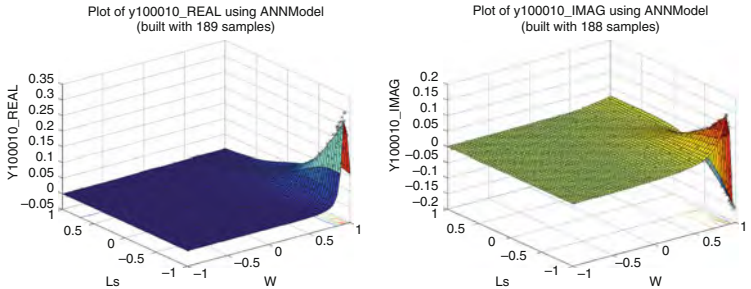


Fig. 3. ANN models of y_{11} admittance with respect to inputs W and L_s. *Left*: real part. *Right*: imaginary part

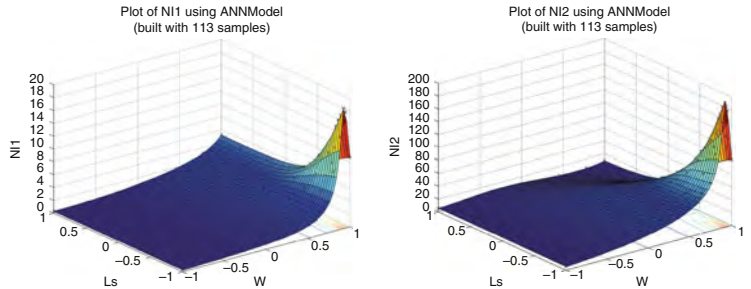


Fig. 4. ANN models of noise currents with respect to inputs W and L_s. *Left*: input noise current. *Right*: output noise current

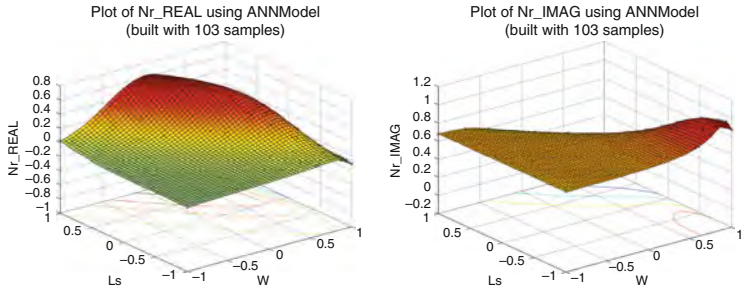


Fig. 5. ANN models of correlation between input and output noise currents with respect to inputs W and L_s. *Left*: real part. *Right*: imaginary part

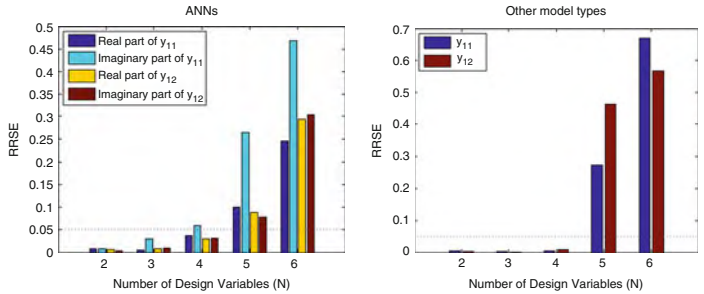


Fig. 6. Root relative square error of models of y_{11} and y_{12} as function of the number of design variables. *Left*: Artificial Neural Networks. *Right*: best model types other than ANN

Table 1 and Figs. 6 and 7 summarize the results[5]. Models which reach the accuracy level $RRSE < 0.05$ are highlighted in bold. It is seen that accurate surrogate models of LNA can be obtained using at most four input parameters. Best model types are rational functions for admittances and ANNs for noise.

As future work, the best model types identified in this study will be used with transistor level simulations. In addition, further investigations will be aimed to improve the accuracy of models including five and six input parameters.

Acknowledgement

This work was supported by the European Commission through the Marie Curie Actions of its Sixth Program under the contract number MTKI-CT-2006-042477. The authors thank Dr. Joost Rommes (NXP semiconductors, Eindhoven) for several stimulating discussions. The first author executed this work at University of Antwerp and NXP semiconductors.

[5]HC stands for 'hill climbing' optimization of hyperparameters [6]. GA stands for 'Genetic Algorithm' based optimization [8].

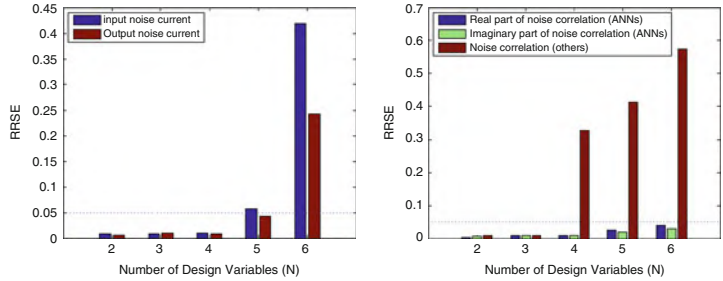


Fig. 7. Root relative square error of noise models as function of the number of design variables. *Left*: input and output noise currents (Artificial Neural Networks). *Right*: correlation between input and output currents

Table 1. Best model type and hyperparameter optimization algorithms

N	y_{11}	y_{12}	Input noise	Output noise	Noise correlation
2	**Rational HC**	**Rational GA**	**ANN GA**	**Rational HC**	**Rational GA, ANN GA**
3	**Rational HC**	**Rational HC**	**ANN GA**	**ANN GA**	**Rational HC, ANN GA**
4	**Rational GA**	**Rational GA**	**ANN GA**	**ANN GA**	**ANN GA**
5	Rational GA	RBF GA	ANN GA	**ANN GA**	**ANN GA**
6	RBF GA	RBF GA	ANN GA	ANN GA	**ANN GA**

References

1. Zhang, Q.J., Gupta, K.C.: Neural Networks for RF and Microwave Design. Artech House, Boston, London (2000)
2. Lee, T.H.: The Design of CMOS Radio-Frequency Integrated Circuits (Second Edition). Cambridge University Press, Cambridge (2003)
3. Hendrickx, W., Gorissen, D., Dhaene, T.: Grid Enabled Sequential Design and Adaptive Metamodeling. In: Proceedings of the 2006 Winter Simulation Conference, WSC 2006, pp. 872–881. 3–6 December 2006
4. Gorissen, D., De Tommasi, L., Croon, J., Dhaene, T.: Automatic Model Type Selection with Heterogeneous Evolution: An application to RF circuit block modeling. In: Proceedings of IEEE World Congress on Computational Intelligence, WCCI 2008, pp. 989–996. Hong Kong, China, 1–6 June 2008
5. Gorissen, D., De Tommasi, L., Crombecq, K., Dhaene, T.: Sequential Modeling of a Low Noise Amplifier with Neural Networks and Active Learning. Neural Computing and Applications, vol. 18(5). Springer, New York (2009)
6. The SUrrogate MOdeling Toolbox Wiki Page. URL http://www.sumowiki.intec.ugent.be/.
7. Hendrickx, W., Dhaene, T.: Sequential design and rational metamodelling. In: Proceedings of the 2005 Winter Simulation Conference, WSC 2005, pp. 290–298. 4–7 December 2005
8. The Mathworks: Genetic Algorithm and Direct Search Toolbox. URL http://www.mathworks.com/products/gads/.

Minisymposium *Precipitation, Deposition and Sedimentation of Particles in Fluid Flow*

L.L. Bonilla and Y. Farjoun

G. Millán Institute, Fluid Dynamics, Nanoscience & Industrial Mathematics, Universidad Carlos III, 28911 Leganés, Spain, bonilla@ing.uc3m.es, yfarjoun@ing.uc3m.es

In many industrial processes, particles transported by gas flows may deposit on confining walls due to different mechanisms. Often times, heterogeneous condensation of vapors on particles transported by the flow or homogeneous condensation of vapors occurs and the resulting droplets move towards cold walls. The inertia of larger particles carried by turbulent flows may also be important in deposition processes. Examples include vapor deposition from combustion gases, fouling and corrosion in biofuel plants, chemical vapor deposition, vapor condensation and aerosols capture by cold plates or rejection by hot ones, deposition of particles in the lungs during breathing, etc. In this minisymposium, several important examples of deposition processes were presented, modeled and their governing equations analyzed and solved numerically.

The paper by J.L. Castillo et al. studies the structure of granular deposits formed by aerosol particles transported by fluid streams. Aerosols are solid particles carried by gas streams which are present in many practical applications, such as heterogeneous nucleation of vapors on pre-existing particles, evolution of clouds and production of artificial rain, pollution dispersion, chemical vapor deposition processes, etc. Understanding aerosol dynamics is needed to control such processes. Deposition of aerosol particles often gives rise to granular materials whose structure is important to characterize and control because it affects the chemical, optical and mechanical properties of the product. The main morphological features of these granular deposits are their bulk properties (density, porosity and structure) and their interface properties (roughness and thickness of the active region). These features depend on the way that new particles arrive to form the deposit and should be tailored for new materials applications: nanostructured deposits, catalytic surfaces, layered materials and others. The paper presents dynamical Monte Carlo simulations of particle deposit growth relating the structure of the granular deposit to the characteristics of particle motion near the surface. The control parameter in these simulations is the Péclet number which measures

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*, Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_70,

the relative importance of deterministic motion (characterized by an average particle velocity towards the wall) and random motion characterized by the diffusion parameter divided by particle size. The authors simulate deposits on attracting or slightly repelling surfaces (positive or small negative Péclet numbers, respectively) and compare their results to experiments.

The paper by Y. Farjoun considers the homogeneous nucleation and growth of clusters in a progressively cooled vapor. Typically, temperature is considered to be constant in this type of problems and then the homogeneous condensation of vapor in droplets and their later growth are analyzed. In this paper, Y. Farjoun assumes that the system is cooled uniformly at a constant rate and that there is no flow of the carrier gas. The nucleation rate is given by Zeldovich's formula and cluster growth is described by the Becker–Döring equations. The clusters interact only through the chemical potential and the temperature of the carrier gas and the diluted vapors (both considered to be ideal gases) is not affected by the condensation process. He then derives the governing equations of the model comprising and advection equation and an integral constraint (conservation of vapor) plus a boundary condition. In this model, the undercooling increases linearly with time. An asymptotic analysis of the mathematical model yields a simple ODE problem which is then numerically solved. The study of homogeneous nucleation in this simple setting is an important first step to understand more complex phenomena of homogeneous condensation in hydrodynamic flows.

The paper by J. Neu et al. studies heterogeneous condensation of vapors on small particles mixed with a carrier gas and their motion by thermophoresis towards a cold wall. The paper explains a simplified model in which vapors are diluted and the concentration of suspended particles is so small that the flow of the carrier gas is not affected. The carrier gas is incompressible and the vapor is in local equilibrium with the condensate on the wall and on the surface of the vapor coated droplets, whose growth is diffusion limited. This model gives rise to a free boundary problem for the dew point interface: there is a region far from the wall in which there is no vapor condensation and a condensation region close to the wall in which vapor condenses on suspended particles. Once the free boundary problem is solved, the deposition rates of droplets and vapor condensation at the wall can be calculated. The paper presents the condensation model, explains how to obtain the free boundary problem and interprets the governing parameters and illustrates how the position of the dew point interface is shifted due to the flow according to the solution of the free boundary problem.

Structure of Granular Deposits Formed by Aerosol Particles Conveyed by Fluid Streams

J.L. Castillo, D. Rodríguez-Pérez, and S. Martín, A. Perea, and P.L. García-Ybarra

Dept. Física Matemática y de Fluidos, Facultad de Ciencias, UNED, Madrid, 28040 Spain
jcastillo@ccia.uned.es, daniel@dfmf.uned.es, smartin@dfmf.uned.es, aperea@ccia.uned.es, pgybarra@ccia.uned.es

1 Introduction

Aerosols (particles carried by gas streams) appear in many practical applications and the understanding of their dynamics [1, 4] is needed to control processes such as, for instance, heterogeneous nucleation of vapors on pre-existing particles, evolution of clouds, pollution dispersion and production of new materials from powders [3]. In this area, there is a need of controlling and characterizing the structure of granular materials formed by depositing aerosol particles. The main morphological features of these granular deposits as their bulk properties (density, porosity and structure) and interface properties (roughness and thickness of the active region) depend on the way that new particles arrive to form the deposit. The granular structure affects the chemical, optical and mechanical properties of the product and it should be tailored for new materials applications: nanostructured deposits, catalytic surfaces, layered materials and others.

The goal of this work is to relate the structure of the granular deposit to the characteristics of the particle motion near the surface.

2 Monte Carlo Simulation of Particle Motion

The method employed is an on-lattice dynamical Monte Carlo simulation [8,9] for the growth of particle deposits by advection and diffusion of particles towards a (initially clean and flat) surface. The model allows to follow the evolution of deposit formation and to determine the main morphological and structural properties of the generated deposits, depending on the transport properties of the arriving particles.

The Monte Carlo method is used to simulate the particle motion over the deposits until the particle reaches the deposit (becomes in contact with

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*, Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_71,

any previously deposited particle). Then the new particle attaches there and contributes to the deposit growth. Sintering or restructuring of the deposited particles is precluded so that the analysis focuses on the relevance of the mechanism of particle arrival. For the Monte Carlo simulation, the motion of an aerosol particle is split in two contributions: a mean deterministic velocity (normal to the flat surface) and a random motion. The deterministic contribution can be due to any transport phenomena [1, 4] that drives the particle along well defined trajectories, such as inertia, advection, some phoretic motions (thermophoresis, electrophoresis), particle sedimentation or external fields, whereas the random motion accounts for the particle diffusive motion (Brownian diffusion, turbulent diffusion or the effect of random fields). The Péclet number is the dynamical parameter that measures the relative importance of the deterministic motion to the random contribution,

$$Pe \equiv \frac{va}{\mathcal{D}} \tag{1}$$

relating the average particle velocity toward the wall, v, the particle diffusion coefficient, $\mathcal{D}$, and the particle diameter, a. In these simulations the average particle velocity, v, is considered to be normal to the (initially clean) surface where the deposit builds up. Time and space are discretized and a cubic lattice is used as the basic domain in the simulation with periodic boundary conditions on the lateral walls. A particle is introduced above the deposit structure at a random horizontal location and its motion is tracked until the particle either reaches the deposit or it moves far away from it. Then, a new particle is introduced and the process is repeated until the deposit height reaches a given maximum height.

The purpose of this work is to analyze the structure of deposits formed on attracting surfaces (positive values of v) and on slightly repelling surfaces (negative values of v); that is, when the Péclet number can be taken as negative, in the sense that the particle mean velocity pushes the particles away from the surface, but this repulsion is weak and Brownian diffusion is still able to bring some particle to deposit on the wall.

3 Deposits on Attracting Surfaces

The formation of deposits collected from aerosol particles which are attracted toward a wall (positive values of v in our simulations) has been extensively studied [5, 8–10]. The arriving particles form layered granular deposits with a density profile which depends on the particle Péclet number. Figure 1 shows the density profile for $Pe = 0.1$ and different deposit heights, $h_{\max}$ (which is related to the total number of deposited particles, i.e. to the total growth time). The collected deposits have a (frozen) denser region in contact with the clean wall, a (frozen) middle region of mean porosity and constant mean density, and an (active) upper region where new particles still deposit [8].

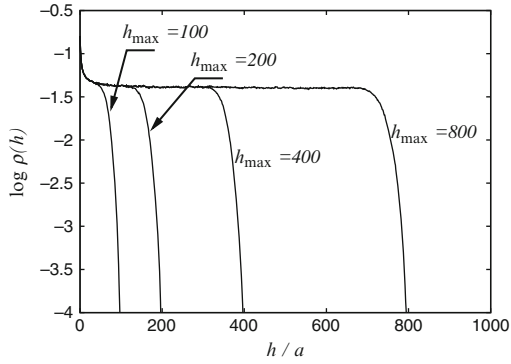


Fig. 1. Density profile versus height for different values of the maximum height, for particles with $Pe = 0.1$

In this active region, the mean density decreases from the middle density, vanishing at the top, $h = h_{\max}$.

The average deposit density in the frozen middle region (relative number of lattice sites occupied by particles in this consolidated region) shows a dependence with the Péclet number given by [8, 9]

$$\bar{\rho}(Pe) = \rho_\infty \left(1 + \frac{Pe_0}{Pe}\right)^{-B} \tag{2}$$

With $\rho_\infty = 0.302$, $Pe_0 = 4.8$, and $B = 0.52$. This region presents a fractal-like structure on the short length scales [8], up to a Pe-dependent scale given by the quantity inside the brackets in (2), with a fractal dimension $D_F = 3 - B$. The limit of large Péclet numbers corresponds to ballistic deposition when particles drift towards the wall and Brownian diffusion is absent. Then, the deposits are denser and the porosity is low. However, for small values of Pe, the deposits are fractal with "particle trees" of all the allowed sizes being limited by lattice size, growth time or the Pe-dependent scale.

Moreover, the pure diffusion limit (vanishing Pe) when the particle motion is purely diffusive and the mean velocity vanishes, corresponds to a singular limit as indicated by (2) because the fractal structure of the deposit extends to all scales and the fractal cut-off (Pe-dependent) scale goes to infinity. The open but highly branched structure of the deposit restricts the penetration of new incoming particles deep into the deposit. In this limit, although some reminiscence of the three regions for attracting surfaces still remains, the deposit density decreases continuously with height.

4 Deposits on Weakly Repelling Surfaces

Even in the case of a mean (weak) particle motion away from the wall, diffusion may bring some particles to the surface [2, 4, 7] and form a granular deposit. A *negative value of the Péclet number* characterizes these deposits, with the minus sign indicating that the mean particle velocity v is directed away from the deposit.

In the limit of small (and negative) Péclet numbers, the formed deposits are initially fractal (as in the pure diffusion limit, $Pe = 0$). But, as the deposit evolves in time the dispersion in the deposit height increases, and a new structure appears dominated by the presence of relatively large particle trees. The larger trees are more effective in collecting the particles that approach to the surface by diffusion, avoiding the growth of the shorter trees. Finally, isolated trees emerge from a fractal deposit baseline.

Therefore, at long times, these deposits present two characteristic regions (Fig. 2): a base region which retains the same structure of the pure diffusion deposits, and a second region with spikes emerging from the base. The baseline structure is shorter for stronger repulsion fields, thus the height of the base region decreases as $|Pe|$ increases, according to the law

$$h_c = \left(1 + \frac{4.8}{Pe}\right)^{1/4} \tag{3}$$

Indicating that the same crossover length (the bracketed quantity) remains for repelling surfaces, see (2). On the other hand, the spike becomes thinner as $|Pe|$ increases.

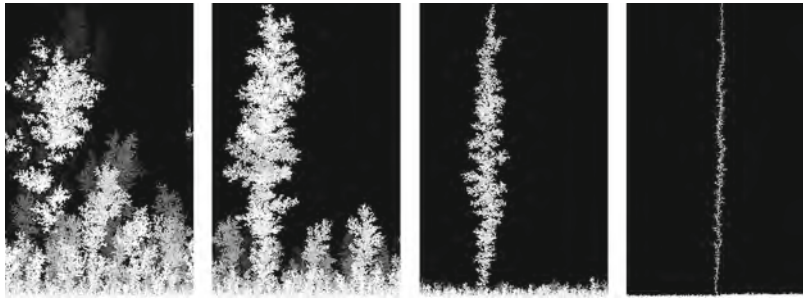


Fig. 2. Side view of the granular deposit, for weakly repealing surfaces and different Péclet numbers (a *grey scale* is used to represent the distance from the frontal wall)

5 Comparison with Experimental Results

In the laboratory, granular deposits are grown using the technique of electrohydrodynamic atomization [6] to disperse a liquid. The liquid is a suspension of carbon nanoparticles in ethanol with a small amount of dispersant to reduce the agglomeration of the nanoparticles. The liquid is pumped through a needle at a fixed flow rate, whereas a high voltage is applied between the needle and a flat collector located below the needle. The electrosprayed liquid forms a quite monodisperse cloud where the size of the primary droplet depends on the liquid properties and flow rate [6]. In our experiments the mean droplet size is of the order of $10\,\mu$m, and each droplet includes $\sim 10^3$ carbon nanoparticles. The fragmentation and evaporation of the charged droplets leaves dry carbon nanoparticles which retain the electrical charge and the nanoparticles are attracted by the collector and form a granular deposit. The charge of the particles in the cloud contributes to the particle dispersion and promotes an effective particle diffusion.

SEM images of the deposit are used to measure the surface roughness as a function of the applied voltage (proportional to the particle Péclet number). A *Shape-from-Focus* (SFF) technique is used to obtain a reconstruction of the deposit surface that can be treated by image processing. The measured roughness decreases with increasing applied voltage, in accordance with the numerical simulations that predict denser deposits as the Péclet number increases. The larger particle agglomerates (with larger inertia and Péclet number) deposit on the central part of the collector whereas at the collector edge some smaller particles deposit. Then, the characteristic particle size is not uniform and a direct comparison with the numerical simulation is not yet possible. To this end, new experiments are being conducted to get sprays containing fewer particles per droplet.

6 Conclusions

The structure of granular deposits formed by the deposition of aerosol particles has been related to the characteristic of the particle motion near the collecting surface. A Monte Carlo method has been implemented to simulate the particle motion [8,9] and the main features of the growth deposits were obtained as a function of the particle Péclet number which measures the relative importance of the mean particle velocity normal to the surface with respect to the particle Brownian motion, see (1).

For attracting surfaces (mean particle velocity towards the wall), the deposit is structured in three differentiated regions: a denser bottom layer in contact with the collecting wall, a middle region with mean constant density and the active region at the top where new particles may still attach. The deposit density decreases as the importance of Brownian diffusion increases (Pe decreases), with the density in the middle region correlated by (2). In the

ballistic limit (large Pe), the deposit is denser, whereas when Pe is reduced the deposit becomes lighter, very porous and highly branched.

In the pure diffusive limit, the branches at the top of the deposit are very effective in collecting particles. Then, the penetration of the incoming particles into the deposit is reduced and longer times are needed to achieve the fractal limited deposits. The deposits grow taller but fragile with open structures, leading to materials which are suitable for some applications where the ratio of area to volume plays a relevant role.

Furthermore, for weakly repelling surfaces the particle Brownian motion is able to bring some particles against the mean drift and a deposit is still formed on this surface. These deposits present two different regions: a base region with the same density profile as the pure diffusive deposit (as the growing mechanism is the same, Brownian diffusion) and large isolated spikes that emerge from this base region. The thickness of the base regions is reduced and the spikes become thinner as the repulsion intensity increases.

Experiments with atomized liquid suspensions are being performed to check the results of the simulations and the preliminary results shows a good qualitative agreement.

The simulation and experimental procedure used in his work open the possibility of making tailored granular materials for specific applications with defined bulk and surface morphologies.

Acknowledgement

Work supported by the Ministerio de Ciencia e Innovacion (Spain) under Grants ENE-2008-06683 and ENE-2008-06515 (partially with FEDER funds) and by the Comunidad de Madrid project COMLIMAMS, S-505/ENE/0229.

References

1. Castillo, J.L., Garcia-Ybarra, P.L.: Transport of particles and vapors in flue gases and deposition on cold surfaces. In: Bonilla, L.L., Moscoso, M., Platero, G., Vega, J.M. (eds.) Progress in Industrial Mathematics at ECMI2006, pp. 284–289. Springer, Berlin (2008)
2. Castillo, J.L., Garcia-Ybarra, P.L.: Diffusive leakage of small particles towards blowing surfaces. J. Aerosol Sci. **29**, S1107 (1998)
3. Friedlander, S.K.: Smoke, Dust and Haze. Fundamentals of Aerosol Dynamics. Oxford University Press, New York Oxford (2000)
4. Garcia-Ybarra, P.L., Castillo, J.L.: Mass transfer dominated by thermal diffusion in laminar boundary layers. J. Fluid Mech. **336**, 379–409 (1997)
5. Konstandopoulos, A.G.: Deposit growth dynamics: particle sticking and scattering phenomena Powder Technol. **109**, 262–277 (2000)
6. Loscertales, I.G., Barrero, A., Guerrero, I., Cortijo, R., Marquez, M., Gañan-Calvo, A.M.: Micro/Nano encapsulation via electrified coaxial liquid jets. Science, **295**, 1695 (2002)

7. Perea, A., Castillo, J.L., Garcia-Ybarra, P.L.: Fickian leakage on boundary layers with blowing. J. Aerosol Sci. **34**, S117–118 (2003)
8. Rodriguez-Perez, D., Castillo, J.L., Antoranz, J.C.: Relationship between particle deposit characteristics and the mechanism of particle arrival. Phys. Rev. E **72**, 021403 (2005)
9. Rodriguez-Perez, D., Castillo, J.L., Antoranz, J.C.: Density scaling laws for the structure of granular deposits. Phys. Rev. E **76**, 011407 (2007)
10. Tassopoulos, M., O'Brien, J.A., Rosner, D.E.: Simulation of microstructure mechanim relationships in particle deposition. AICHE J. **35**, 967–980 (1989)

Creation of Clusters via a Thermal Quench

Yossi Farjoun

G. Millán Institute, Fluid Dynamics, Nanoscience & Industrial Mathematics, Universidad Carlos III, 28911 Leganés, Spain, yfarjoun@ing.uc3m.es

Summary. The nucleation and growth of clusters in a progressively cooled vapor is studied. The chemical-potential of the vapor increases, resulting in a rapidly increasing nucleation rate. The growth of the newly created clusters depletes monomers, and counters the increase in chemical-potential. Eventually, the chemical potential reaches a maximum and begins to decrease. Shortly thereafter the nucleation of new clusters effectively ceases. Assuming a slow quench rate, asymptotic methods are used to convert the non-linear advection equation of the cluster-size distribution into a fourth-order differential equation, which is solved numerically. The distribution of cluster-sizes that emerges from this *creation* era of the quench process, and the total amount of clusters generated are found.

1 Introduction

While studying a simplified model of nucleation the temperature is often *assumed* to be held constant (see, e.g., the review papers by Wu [6] and Oxtoby [5]). However, time dependence of the temperature can play an important role in the aggregation process. In this paper we study the effect that a thermal quench has on aggregation. Together with J. Neu, we previously studied [3] a similar problem with *constant* temperature. For a less terse literature summary, the reader is referred to the review articles cited above.

Our system comprises a dilute, condensable vapor in a carrier gas. Starting from a temperature corresponding to zero chemical-potential, the system is cooled uniformly at a constant rate. We assume a quench rate that is slow relative to the molecular timescale. This implies that the maximal chemical-potential is small and therefore the nucleation rate and cluster growth can be assumed to follow the Zeldovich formula [7] and the Becker–Döring (BD) equations [1] respectively. We use the BD equations and not diffusion limited growth, as per Lifshitz–Slyozov [4], since the final size of the clusters is relatively small. To find the value of the chemical-potential we use conservation of particles and the Clausius–Clapeyron relation together with the ideal-gas

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*, Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_72,

law. To simplify the model and calculation, we assume that the latent heat of evaporation is large relative to the thermal energy, $k_B T$.

Initially, there are very few clusters and so the chemical-potential increases as the temperature drops (through the resulting reduction of the equilibrium monomer density). As the chemical-potential increases, so does the nucleation rate. Eventually, enough clusters have been created that their growth causes a large enough combined drain on the monomer density so that the increase in chemical-potential is stopped. After this, the nucleation rate drops quickly and approaches zero.

The paper is organized as follows: In Sect. 2 we present a short derivation of the aggregation model we use. In Sect. 3 we solve the resulting nonlinear advection PDE using an asymptotic approximation, the method of characteristics, and a numerical solver.

1.1 Assumptions

Throughout the paper we make the following assumptions.

1. Nucleation occurs at the Zeldovich rate.
2. The clusters that form are small, and grow according to BD dynamics.
3. The process is spatially uniform.
4. The only interaction between clusters is via the chemical-potential.
5. The temperature is not affected from the condensation of clusters.
6. The carrier gas and the condensable gas are ideal gasses.

2 The Model

We briefly derive the constituent equations of the quenching process. We assume basic familiarity with standard nucleation, BD growth, and the monomer conservation argument that leads to the determination of the chemical-potential from the distribution of cluster sizes. For these we follow the notation in Wu's review article [6]. First, we derive the relationship between the chemical-potential and the undercooling.

2.1 Chemical-Potential and the Undercooling

The chemical-potential η is the free energy of a monomer in condensed liquid relative to that in vapor:

$$\eta = k_B T \log \frac{c}{c_e}. \tag{1}$$

Here, k_B is the Boltzmann constant, T is the temperature, c is the ambient monomer concentration, and c_e is the equilibrium monomer concentration.

The dependence of c_e on the temperature can be found through the Clausius–Clapeyron relation and the ideal-gas law as in [2].

$$\frac{c_e}{c_0} = \frac{T_0}{T} \mathrm{e}^{\frac{\Lambda}{k_B T_0} - \frac{\Lambda}{k_B T}}, \tag{2}$$

where Λ is the latent-energy of condensation (per monomer.) In (2), T_0 and c_0 are the initial equilibrium temperature and monomer-concentration. That is, at concentration c_0 and temperature T_0 the vapor phase is in equilibrium with the liquid phase. Thus, setting $T = T_0$ gives $c_e = c_0$.

We assume that $\Lambda \gg k_B T_0$ and define $\frac{1}{\varepsilon} \equiv \frac{\Lambda}{k_B T_0}$. The non-dimensional versions of c, T, and η are:

$$\tilde{c} = \frac{c}{c_0}, \qquad \tilde{T} = \frac{T}{T_0}, \qquad \tilde{\eta} = \frac{\eta}{k_B T_0}. \tag{3}$$

Equations (1) and (2) now are as follows

$$\tilde{\eta} = \tilde{T} \log \frac{\tilde{c}}{\tilde{c}_e}, \qquad \tilde{c}_e = \tfrac{1}{\tilde{T}} \mathrm{e}^{\frac{1}{\varepsilon} - \frac{1}{\varepsilon \tilde{T}}}. \tag{4}$$

The undercooling is defined as $\tau \equiv \frac{1-\tilde{T}}{\varepsilon}$.

Using the above definition of τ we find an asymptotic approximation of $\tilde{c}_e$ for small ε:

$$\tilde{c}_e = \mathrm{e}^{-\tau} + O(\varepsilon\tau), \tag{5}$$

and the leading order approximation of $\tilde{\eta}$:

$$\tilde{\eta} \approx \log \tilde{c} + \tau, \qquad \text{for } \varepsilon \ll 1. \tag{6}$$

2.2 The Zeldovich Nucleation Rate and the Growth Rate of Clusters

Two important components of our model are the rate at which new clusters come into existence, and the rate at which existing clusters grow. The BD equations of growth specify that the size n of a cluster follows the growth "law" for clusters much larger than the critical size:

$$\dot{n} = \omega \tilde{\eta} n^{\frac{2}{3}}. \tag{7}$$

Where $\tilde{\eta}$ is as in (1), and ω is a "escape rate" constant so that $\omega n^{2/3}$ is the rate at which monomers leave the cluster.

The Zeldovich formula give the nucleation rate of new clusters:

$$j = \omega c_0 \tilde{c}_e \sqrt{\frac{\sigma}{6\pi}} \mathrm{e}^{-\frac{T_0}{T} \frac{\sigma^3}{2\tilde{\eta}^2}}. \tag{8}$$

Where σ is the "surface energy" constant of a cluster,

$$\frac{\text{surface energy}}{k_B T_0} = \frac{3}{2} \sigma n^{\frac{2}{3}}. \tag{9}$$

Using the definition of τ and (5) we have

$$j = \omega c_0 \sqrt{\frac{\sigma}{6\pi}} \mathrm{e}^{-\frac{1}{1-\varepsilon\tau} \frac{\sigma^3}{2\tilde{\eta}^2} - \tau} \tag{10}$$

2.3 Growth Dynamics via Advection PDE

We are interested in finding the evolution of the density of cluster sizes. Let $r(n,\, t)$ be the density of clusters of size n at time t (also referred to as the "distribution of cluster-sizes"). From (7) we derive an advection PDE for the distribution of clusters in the space of their size,

$$\partial_t r + \omega\tilde{\eta}\partial_n\left(n^{\frac{3}{2}} r\right) = 0, \qquad \text{in } n > 0. \tag{11}$$

The flux of clusters $\omega\tilde{\eta}n^{\frac{2}{3}} r$ asymptotes to the Zeldovich rate as $n \to 0$, thus

$$\omega\tilde{\eta}n^{\frac{2}{3}} r \to j = \omega c_0 \sqrt{\frac{\sigma}{6\pi}} \mathrm{e}^{-\frac{\sigma^3}{2\tilde{\eta}^2} - \tau} \text{ as } n \to 0. \tag{12}$$

2.4 Determination of the Chemical-Potential

The chemical-potential can be inferred from the distribution of cluster sizes and the initial concentration c_0.

$$c + \int_0^\infty nr(n,\, t)\, dn = c_0. \tag{13}$$

For small values of $\tilde{\eta}$ we have from (6) that $c = c_0 e^{\tilde{\eta} - \tau}$. Thus

$$\mathrm{e}^{\tilde{\eta} - \tau} + \int_0^\infty n\tilde{r}(n,\, t)\, dn = 1, \text{ where } \tilde{r} = \frac{r}{c_0}. \tag{14}$$

3 The Mathematical Problem

We now drop all tildes and refer only to non-dimensional variables. The mathematical problem is therefore,

$$\partial_t r + \eta\partial_n(n^{\frac{2}{3}} r) = 0, \tag{15}$$

$$\eta n^{\frac{2}{3}} r \to \sqrt{\frac{\sigma}{6\pi}} \mathrm{e}^{-\frac{\sigma^3}{2\eta^2} - \tau}, \quad \text{as } n \to 0, \tag{16}$$

$$\mathrm{e}^{\eta - \tau} + \int_0^\infty nr\, dn = 1. \tag{17}$$

Here, time t is non-dimensionalized with the scaling unit $1/\omega$.

In the current paper we consider an undercooling which increases linearly with time t:

$$\tau = \Omega(t + t_0). \tag{18}$$

The parameter Ω is externally specified, and t_0 is a time-lag needed so that the "interesting" behavior happens near $t = 0$.

3.1 Asymptotic Solution of the Creation Era

The equations needed to find the relevant scales of the creation era are mostly straightforward dominant balances of (15)–(17). There is one that is not: The change in chemical-potential must be such that the reduction in the nucleation rate is comparable to the nucleation rate itself. This implies that the change in chemical-potential is small (relative to η itself) and we use $\delta\eta \equiv \eta(0) - \eta(t)$ to follow the *change* in the chemical potential. To save space, we omit the derivation and proceed to the resulting scales. They are

$$[\delta\eta] = \left(\frac{\Omega t_0}{\sigma}\right)^3 \quad [n] = \Omega^3 t_0^3 \left(\frac{1}{\sigma^3 E}\right)^{\frac{3}{4}} \quad [t] = \left(\frac{1}{\sigma^3 E}\right)^{\frac{1}{4}} \quad [r] = \left(\frac{\sigma E}{\Omega^2 t_0^2}\right)^{\frac{3}{2}}.$$

While t_0 and Ω are connected by $\Omega^2 t_0^3 E^{\frac{1}{4}} = \sigma^{\frac{9}{4}}$. We introduced E, a measure of nucleation rate: $E \equiv \sqrt{\frac{\sigma}{6\pi}} \mathrm{e}^{-\frac{\sigma^3}{2(t_0\Omega)^2} - \Omega t_0}$. Using these scales results in the following system for $r(n,\, t)$ and η:

$$\partial_t r + \partial_n (n^{\frac{2}{3}} r) = 0, \tag{19}$$

$$n^{\frac{2}{3}} r \to \mathrm{e}^{\delta\eta} \text{ as } n \to 0, \tag{20}$$

$$\delta\eta(t) - t + \int_0^\infty nr\, dn = 0. \tag{21}$$

Since $n^{\frac{2}{3}} r$ is constant along the level curves of $3n^{\frac{1}{3}} - t$, we can write

$$r(n,t) = n^{-\frac{2}{3}} \mathrm{e}^{\delta\eta(t - 3n^{\frac{1}{3}})}. \tag{22}$$

Substituting this into (21) yields an integral equation for $\delta\eta$, and using the change of variable $t' = t - 3n^{\frac{1}{3}}$ we arrive at

$$\delta\eta(t) - t + \int_{-\infty}^{t} \left(\frac{t - t'}{3}\right)^3 \mathrm{e}^{\delta\eta(t')}\, dt' = 0. \tag{23}$$

From here we derive a fourth-order ordinary differential equation (ODE) for $\delta\eta$, by differentiating four times:

$$\dddot{\delta\eta} = -\frac{2}{9} \mathrm{e}^{\delta\eta}, \qquad \text{where } \delta\eta(t) \sim t, \text{ as } t \to -\infty. \tag{24}$$

To solve this ODE (numerically), we start with $\delta\eta(t_i) = t_i$, $\dot{\delta\eta}(t_i) = 1$, $\ddot{\delta\eta}(t_i) = 0$, $\dddot{\delta\eta}(t_i) = 0$, and $t_i = -10$.

By integrating $\mathrm{e}^{\delta\eta}$ we get the (scaled) density of clusters ρ that were formed, $\rho \approx 5.1$. The resulting density is shown in Fig. 1.

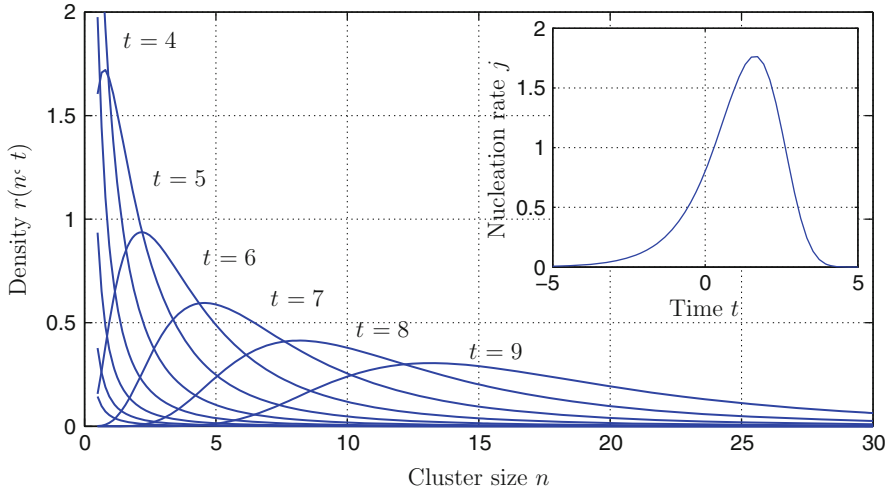


Fig. 1. The cluster size distribution $r(n, t)$ for t ranging from 0 to 9 and, inset, the nucleation rate as a function of (shifted and scaled) time

4 Conclusions

Thermal quench is a standard trigger for nucleation. However, normally the quench process itself is ignored and only the outcome (i.e., a super-saturated monomer solution) is considered. In this paper we have shown that during a slow quench, enough clusters nucleate so that no more nucleate after the quench. Presumably, the clusters that have nucleated simply grow after the quench and eventually coarsen.

Acknowledgements

The support by the National Science Foundation is acknowledged. The author was partially supported by grant DMS-0703937.

References

1. Becker, R., Döring, W.: Ann. Phys. **24**, 719 (1935)
2. Castillo, J.L., Rosner, D.E.: Chem. Eng. Sci. **44**(4), 939–956 (1989)
3. Farjoun, Y., Neu, J.C.: Phys. Rev. E. vol. 78 doi: 10.1103/PhysRevE.78.051402 051402 (2008)
4. Lifshitz, I.M., Slyozov, V.V.: J. Phys. Chem. Solids **19**, 35–50 (1961)
5. Oxtoby, D.W.: J. Phys Condens. Matter **4**, 7627–7650 (1992)
6. Wu, D.T.: Solid State Physics, vol. 50, pp. 37–187. Academic, San Diego, CA (1996)
7. Zeldovich, J.B.: Acta Physiochim, URSS **18**, 1–22 (1943)

Theory of Surface Deposition from Boundary Layers Containing Condensable Vapor and Particles

J.C. Neu[1], A. Carpio[2], and L.L. Bonilla[3]

[1] Department of Mathematics, University of California, Berkeley, CA, USA
neu@math.berkeley.edu
[2] Departmento di Matemática Aplicada, Universidad Complutense de Madrid, Madrid, Spain, ana_carpio@mat.ucm.es
[3] G. Millián Institute of Fluid Dynamics, Nanoscience and Industrial Mathematics, Universidad Carlos III de Madrid, Leganes, Spain, bonilla@ing.uc3m.es

Summary. Hot gas containing condensible vapor and small particles "blows" against a "cold" wall. The model computes the rate at which liquefied vapor accumulates on the wall due to direct condensation plus a "rain" of liquid-covered particles. The latter results from heterogeneous nucleation in undercooled vapor near the wall.

1 Physical Background

Figure 1 is a cartoon of the surface deposition process. In explaining the figure, we set forth the physical ideas of the simplest model. First, the whole process is steady state, so all the state variables are time-independent. Dilute vapor is transported by convection-diffusion in the carrier gas with velocity field $\boldsymbol{u}(\boldsymbol{x})$. We have in mind boundary layer flows whose streamlines asymptote to the wall. Far from the wall, the vapor has ambient uniform concentration c_∞. The temperature field $T(\boldsymbol{x})$ has an ambient uniform value T_∞ far from the wall, and a fixed value $T_w < T_\infty$ at the wall. We assume that the vapor is so dilute that the velocity and temperature fields are decoupled from condensation. In particular, the latent heat of vaporization has a negligible effect on the temperature field. If we also neglect thermal expansion of the carrier gas, we further decouple the velocity and temperature fields from each other. Hence, the boundary layer flow $\boldsymbol{u}(\boldsymbol{x})$ satisfies incompressible Navier–Stokes equations, and the temperature field $T(\boldsymbol{x})$ satisfies the usual convection-diffusion boundary value problem in the given incompressible flow $\boldsymbol{u}(\boldsymbol{x})$.

Next, the input from thermodynamics: There is a local *equilibrium vapor concentration* $c_e(\boldsymbol{x})$ determined from the local temperature $T(\boldsymbol{x})$ by the Clausius–Clapeyron formula,

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*,
Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_73,

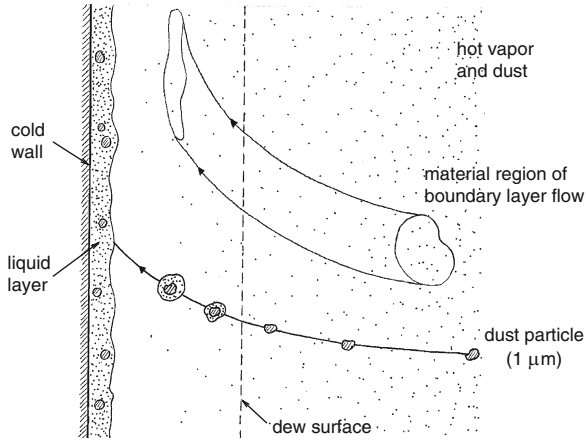


Fig. 1. Physical picture of surface deposition

$$\frac{c_e(\boldsymbol{x})}{c_\infty} = \frac{T_d}{T(\boldsymbol{x})} \exp\left(\frac{\Lambda}{k_B T_d} - \frac{\Lambda}{k_B T(\boldsymbol{x})}\right). \tag{1}$$

Here, Λ is the latent heat of vaporization and T_d is the *dew temperature* so that $c_e = c_\infty$ when $T = T_d$. The vapor is locally undersaturated if $c(\boldsymbol{x}) < c_e(\boldsymbol{x})$, and oversaturated if $c(\boldsymbol{x}) > c_e(\boldsymbol{x})$. If $T_\infty > T_d$, we have $c_\infty < c_e(\boldsymbol{x} = \infty)$ and the vapor far from the wall is undersaturated. If the wall temperature is sufficiently low, there is a *condensation layer* near the wall of oversaturated vapor, bounded by a *dew surface* where $c(\boldsymbol{x}) = c_e(\boldsymbol{x})$.

Small particles entering the condensation layer are the "seeds" for heterogeneous nucleation. Our kinetic model of the heterogeneous nucleation is the simplest possible: a diffusion-limited condensation, neglecting capillary effects. As mentioned above, liquid accumulates on the wall by direct condensation, and also as a "rain" of liquid-covered drops. There is an important piece of physics governing the "rain". If droplets are simply convected by the carrier gas, they (in principle) never reach the wall because the streamlines of $\boldsymbol{u}(\boldsymbol{x})$ asymptote to the wall and never 'dive in." What really happens is *thermophoresis*: Small droplets actually have a velocity *relative* to the carrier gas in the direction of decreasing temperature. The model of thermophoresis is that the velocity field of droplets is

$$\boldsymbol{v}(\boldsymbol{x}) = \boldsymbol{u}(\boldsymbol{x}) - \alpha \frac{\nabla T(\boldsymbol{x})}{T}, \tag{2}$$

where α is a constant, typically some small fraction of the kinematic viscosity. The streamlines of the droplet velocity field do dive into the wall, and the "rain" happens.

2 Model

We present a quantitative model based on the aforementioned physics in a specific geometry: The velocity field of the carrier gas is the Blasius stagnation point flow in the half-space $x > 0$ bounded by the wall $x = 0$. The streamlines are incoming from $x = +\infty$ and diverge to $z = +\infty$ or $z = -\infty$ as $x \to 0$, as depicted in Fig. 1. The state variables are the x-velocity u of the carrier gas, the temperature field T, the droplet concentration ρ, the vapor concentration c, and n, the volume of droplet in units of monomer volume in the condensed phase. All are functions of x only.

Nondimensionalization is based on units in the *scaling table*:

$$[x] = \sqrt{\frac{\nu}{\gamma}}, \qquad [u] = \sqrt{\nu\,\gamma}, \qquad [T] = T_\infty = 1713\ \mathrm{K},$$
$$[\rho] = \rho_\infty = 10^5\ \mathrm{cm}^{-3}, \qquad [n] = n_* = 4.72 \times 10^{10},$$
$$[c_\infty] = 1.9 \times 10^{13}\ \mathrm{cm}^{-3}.$$

Here, ν is the kinematic viscosity of the carrier gas, and γ is the strain rate of the Blasius flow at $x = \infty$, a control parameter. Units with actual numbers come from Castillo and Rosner's work [1, 2] on the fouling of a chimney by coal smoke containing sodium sulfate (the condensible vapor) and soot (the particles). In particular, $\rho_\infty = 10^5\ \mathrm{cm}^{-3}$ is the incoming concentration of soot particles at $x = \infty$. The unit $n_* = 4.72 \times 10^{10}$ is the number of vapor monomers required to make a pure liquid drop the same size as incoming soot particles, 1 μm in radius. The ambient vapor concentration $c_\infty = 1.9 \times 10^{13}\ \mathrm{cm}^{-3}$ derives from the Clausius–Clapeyron formula (1) assuming the dew temperature to be 1400 K, less than $T_\infty = 1713$ K. The wall temperature in Castillo and Rosner's example is $T_w = 1000\ \mathrm{K} < T_d$, and this forces the existence of a condensation layer.

In dimensionless variables, $u(x)$ is the parameter-free solution of the classic Blasius boundary value problem. Given $u(x)$, the dimensionless temperature field is the solution of the ODE

$$Pr\, u\, T' - T'' = 0$$

in $x > 0$, subject to boundary conditions $T = T_\infty$ at $x = \infty$ and $T = T_w$ at $x = 0$. Here $Pr = 0.7$ is the Prandtl number, the ratio of kinematic viscosity and thermal diffusivity.

Given $u(x)$ and $T(x)$, the droplet concentration $\rho(x)$ is computed: The steady-state continuity equation for conservation of droplets is $\nabla \cdot (n\,\boldsymbol{v}) = 0$, where $\boldsymbol{v}$ is the droplet velocity field in (2). In the Blasius flow case it reduces to an ODE whose dimensionless form is

$$\left(u - \alpha \frac{T'}{T}\right) \rho' = \alpha\, \rho \left(\frac{T'}{T}\right)'$$

in $x > 0$, subject to dimensionless boundary condition $\rho \to 1$ as $x \to \infty$.

The remaining state variables $n(x)$, $c(x)$ quantify the condensation kinetics. Their ODEs are

$$\left(u - \alpha \frac{T'}{T}\right) n' = \begin{cases} N\, n^{1/3}\,(c - c_e) \text{ in } x < x_*, \\ 0 \quad \text{in } x > x_*, \end{cases} \tag{3}$$

$$Sc\, u\, c' - c'' = \begin{cases} -R\, \rho\, n^{1/3}\,(c - c_e) \text{ in } x < x_*, \\ 0 \quad \text{in } x > x_*. \end{cases} \tag{4}$$

Here, c_e is the dimensionless equilibrium concentration determined by the dimensionless Clausius–Clapeyron formula,

$$c_e(x) = \frac{T_d}{T} \exp\left[\frac{1}{\varepsilon}\left(\frac{1}{T_d} - \frac{1}{T(x)}\right)\right], \tag{5}$$

and $x = x_*$ marks the dew surface where $c = c_e$. $x > x_*$ is the "dry" region of undersaturated vapor where $c < c_e$, and $0 < x < x_*$ is the oversaturated condensation layer where $c > c_e$.

The dimensionless parameters in (3-5) are

$$\begin{aligned} \alpha &= 0.1 \qquad \text{(dimensionless thermophoresis coefficient)} \\ \frac{1}{\varepsilon} &= \frac{\Lambda}{k_B\, T_\infty} = 19.42 \qquad \text{(dimensionless latent heat)} \\ Sc &=: \frac{\nu}{D} = 1.8 \qquad \text{(Schmidt number, kinematic viscosity/vapor diffusivity)} \\ R &=: \frac{\nu\, l\, \rho_\infty\, n_*^{1/3}}{\gamma} \qquad \text{(control parameter)} \\ N &=: \frac{c_\infty}{\rho_\infty\, n_*\, Sc}\, R = 0.0022362\, R. \end{aligned}$$

In the formula for R, l is a length constant proportional to the effective radius of monomer in the condensed phase.

We explain the physical content of (3,4). The convective derivative in the LHS of (3) is the time rate of growth of n as the droplet moves with (dimensionless) x-velocity $u - \alpha T'/T$. In $x > x_*$, no condensation and no growth. We assume $n(x) \equiv n_* =$ positive constant in $x > x_*$, so the particles which serve as seeds of heterogeneous nucleation have no liquid covering. In $0 < x < x_*$, droplet growth is diffusion limited, so rate of growth is proportional to linear size (proportional to $n^{1/3}$) times the local oversaturation $(c - c_e)$.

The LHS of (4) represents convection-diffusion of vapor in the carrier gas. The support of the sink term on the RHS is confined to the condensation layer $0 < x < x_*$ where nucleation is happening. There, the sink per unit volume is the individual droplet growth rate $n^{1/3}(c - c_e)$ times the droplet concentration ρ. Two obvious boundary conditions on c are $c \to c_\infty$ far from the wall, and $c = c_w$ on the wall, where c_w is the equilibrium vapor concentration at wall temperature T_w, as given by the Clausius–Clapeyron formula (1). In addition,

$c = c_e$ on the dew surface, and the derivative c' is continuous across the dew surface $x = x_*$. If we take the dew surface as *given* then there are unique solutions for $n(x)$ and $c(x)$ with $n \to n_\infty$, $c \to c_\infty$ away from wall, and $c = c_w$ on wall. Hence, the derivatives c' on "dry" and "wet" sides of the dew surface are functions of x_*. It remains to adjust x_*, so continuity of the derivative is achieved. In this sense, we have a *free boundary problem* to determine $n(x)$, $c(x)$, and the location x_* of dew surface.

Given the solutions for $\rho(x)$, $n(x)$, and $c(x)$, we may compute the rate of accumulation of condensed vapor on the wall. Direct condensation from vapor at the wall is given by the usual diffusive flux. Using $c_\infty \sqrt{\nu\gamma}$ as the unit of flux (recall $\sqrt{\nu\gamma}$ is the unit of velocity of the Blasius flow), the direct condensation flux is

$$J_\nu = \frac{c'(0)}{Sc}. \tag{6}$$

The rain of liquid drops contributes flux

$$J_c = \frac{R}{N\,Sc}\left(\rho\,(n-1)\,\frac{\alpha\,T'}{T}\right)(0). \tag{7}$$

3 Results

Numerical approximation to solutions of the free boundary value problem is elementary, and serves as a benchmark for asymptotic analysis. That asymptotic analysis is presented in our expanded paper "Surface deposition inhibited by heterogeneous condensation," submitted to the *Journal of Fluid Mechanics* [3]. One point of interest: To obtain the good agreement with the numerical results, it is necessary to go beyond the obvious matched asymptotic expansions, and deal skillfully with strong exponentials that greatly influence the predicted deposition rates on the wall.

Here, we concentrate on physical results, starting with the net deposition rate $J =: J_\nu + J_c$ which is, after all, the "bottom line" of the whole modeling effort. To see the effect of heterogeneous nucleation, we can look at J as a function of the control parameter R. Since R is proportional to ρ_∞, we see that $R \to 0$ is the limit of *no* heterogeneous nucleation, and $R \to \infty$ is the limit of strong heterogeneous nucleation. In fact, we think of $R \to \infty$ as the "equilibrium limit" since it is clear from (4) that $R \to \infty$ enforces $c(x)$ very close to $c_e(x)$ in the condensation layer. The table below shows preliminary numerical results on deposition rates:

R	J (in units of $c_\infty \sqrt{\nu\,\gamma}$)
0 (no particles!)	0.38
4.93	0.2285
739.5	0.0929
∞ (equilibrium limit)	0.0523

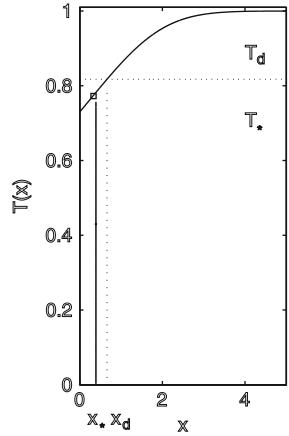


Fig. 2. Illustration of the dew point shift

These results demonstrate that lots of soot in the smoke is really your friend, if you want to reduce deposition on the wall.

Another significant point of the free boundary problem is the displacement of the dew surface away from the $T = T_d$ isotherm and closer to the wall. Figure 2 shows a rather dramatic example (using a higher wall temperature $T_w = 1250\,\mathrm{K}$). Here is the physical explanation: On the dew surface, $c(x) < c_\infty$ because the condensation layer is a vapor sink and diffusion induces a "vapor deficit" $c(x) < c_\infty$ on the "dry" side of the dew surface. Hence, $c(x) = c_e(x) < c_\infty$ on the dew surface, and the Clausius–Clapeyron formula (a monotone increasing relation between c_e and T) implies $T < T_d$ on the dew surface. Hence, the dew surface is closer to the wall than the $T = T_d$ isotherm.

Acknowledgment

This work has been supported by the National Science Foundation Grant DMS-0515616 (JCN), Ministry of Science and Innovation grant FIS2008-04921-C02 (AC and LLB) and by the Autonomous Region of Madrid under grant S-0505/ENE/0229 (JCN and LLB).

References

1. Castillo, J.L., Rosner, D.E.: Int. J. Multiphase Flow **14**, 99–120 (1988)
2. Castillo,J.L., Rosner, D.E.: Chem. Eng. Sci. **44**, 925–937 (1989)
3. Neu, J.C., Carpio, A., Bonilla, L.L.: J. Fluid Mech. **626**, 183–210 (2009)

Comparing League Formats with Respect to Match (Un)importance: A Case Study in Belgian Soccer

Dries R. Goossens[1] and Jeroen Belien[2,3]

[1] PostDoc researcher for Research Foundation – Flanders, Center for Operations Research and Business Statistics (ORSTAT), Faculty of Business and Economics, K.U.Leuven, Belgium, dries.goossens@econ.kuleuven.be
[2] Research Center for Modelling and Simulation, H.U.Brussel, Belgium, jeroen.belien@hubrussel.be
[3] Affiliated researcher Operations Management Group, Department of Decision Sciences and Information Management, Faculty of Business and Economics, K.U.Leuven, Belgium

Summary. Recently, most clubs in the highest Belgian football division have become convinced that the format of the league should be changed. Moreover, the TV station that broadcasts the league is pleading for a more attractive competition. This paper discusses the current league format, and two other formats that are presently being considered by the Royal Belgian Football Association. The attractiveness of each of the formats is measured by the number of unimportant matches: the less unimportant matches, the more attractive the competition.

1 Introduction

For decades, the first (and highest) division in Belgian football has been organized as a double round robin tournament, i.e. a tournament in which each team plays each other team twice, once at home and once away. During the last years, most clubs in the first division have become convinced that changes in the way the competition is played are needed. There is however little agreement on what these changes should be, since arguments and preferences of the teams depend on their (aspired) role in the competition. Apart from the clubs, the TV station that broadcasts the competition is looking for a league format that is as attractive as possible. In other words, the TV station wants to avoid matches that have no importance at all, since these are the games that don't attract viewers. Given the fact that the money from the broadcasting contract is the main source of income for many clubs, the wishes of the TV station carry a considerable weight.

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*, Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_74,

In Sect. 2, we discuss a number of league formats that are currently being considered to be adopted for the first division. Our measure of match unimportance is detailed in Sect. 3, and in Sect. 4, we compare the various league formats using this measure.

2 Reforming the Belgian Football League

Currently, the highest Belgian football league is played as a straightforward double round robin tournament with 18 teams, spread over 34 rounds. The winner of this league is the champion, and qualifies for the (qualification stage of the) Champions League. The second and third in the league also qualify for European football. The team that ends up last relegates to the second division. The one but last team can however remain in the highest division if it wins a double round robin tournament with three second division clubs. This means that 12 additional games are played after the regular competition, resulting in 318 matches in total (see [2]). In the remainder of this section, we discuss two formats that are under consideration by the Royal Belgian Soccer League to be used for the season 2009–2010.

The first alternative that we consider is loosely based on the competition in The Netherlands (see e.g. [4]). The format is a double round robin tournament with 16 teams, where the first in the league is the champion, and the last relegates to the second division. Two post-season play-off tournaments decide which teams qualify for European football, and which teams relegate. The European play-offs are played with the teams ranked 2–5 in the league; the relegation play-offs are played with the teams ranked 14th and 15th, and six teams from the second division. Both play-offs are organized as a direct knock-out tournament, with a home and an away game in each stage. The regular competition, together with the two play-offs results in 260 games per season. Since almost all outcomes in this competition are decided by a play-off stage, we will refer to it as the *play-off league*.

The second league format splits the competition into two parts: an autumn competition and a spring competition. Each of these competitions consists of two series, A and B, of ten teams each, that play a double round robin tournament. The competition starts with the autumn competition, consisting of 18 rounds. The winner of the A series of this competition qualifies for European football. The best five teams of the B series replace the worst five teams in the A series of the subsequently played spring competition, which is again a double round robin tournament for both the A and the B series. The winner of the A series in the spring competition is the league champion, and also qualifies for European football. The two worst teams in the B series relegate to the second division. The final ticket for European football is awarded to the winner of a direct knock-out tournament with a single game per stage, played with the 4 teams that were second in the A series or first in the B series in the autumn or the spring competition. For the next season, the worst five teams of

the A series are again replaced by the best five teams of the B series, and two teams promote from the second division, replacing the two last of the spring B series. This league format has 363 games in total, and is referred to as the *Wijnants league*, named after its inventor Herman Wijnants, chairman of the football club Westerlo VV. With his league format, Wijnants attempted to find a compromise, reconciling the various clubs with conflicting interests. We refer to [3] for more details on the Wijnants league.

3 Measuring Match (Un)Importance

The concept of match importance has been discussed before in a number of papers, but the most commonly accepted measure for match importance is what Schilling [7] calls the conditional importance $S_i(X)_{t,t+k}$ of match scheduled to be played at time $t+k$ for a team i at time t with respect to outcome X, and is defined as follows.

$$S_i(X)_{t,t+k} = p(X_i|W_{i,t+k}, H_t) - p(X_i|L_{i,t+k}, H_t) \tag{1}$$

We use the notation X_i for an outcome X that is achieved by team i. This outcome may be the league championship, but just as well qualification for European football or relegation. The event where team i wins its game scheduled at time $t+k$ is represented by $W_{i,t+k}$; the event of team i losing this game by $L_{i,t+k}$. Finally, H_t represents the history of games that have already been played up to time t. Note that this measure does not take into account a draw; extending this definition to include draws is not straightforward.

We propose to measure the attractiveness of a competition format through match unimportance. We define a match as unimportant for a team and with respect to some outcome, if it can have no influence on the outcome for that team. The underlying idea is that a game between two teams that have nothing to gain or to lose is no longer interesting for a TV station to broadcast, and will attract less fans to the stadium. Bojke [1] confirms that this is the case in the English Premier League. Thus, we suggest that the lower the number of unimportant games in a league is, the more attractive this league is.

In order to know whether a game still matters for a team t with respect to some outcome, we need to know the highest and the lowest ranking that team can still reach at the end of the season, before the game is played. We use the following notation. We define T as the set of teams in the competition, and $G(m)$ as the set of games that are yet to be played, given that there are m rounds remaining. We define the variable w_{ij} to be 1 if i wins its home game against j, and 0 otherwise. Furthermore, we say that l_{ij} is 1 if i loses its home game against j, otherwise $l_{ij} = 0$. The remaining decision variables are p_i, the number of points that a team i has at the end of the season, and r_i, which is 1 if team i is ranked higher than team t at the end of the season. The highest position a team t can possibly reach, given that m rounds remain

to be played, and that team t collected a_t points from rounds already played is given by an optimal solution of the following formulation.

Minimize

$$1 + \sum_{i \in T \setminus \{t\}} r_i \tag{2}$$

subject to

$$a_i + m + \sum_{j \in T \setminus \{i\} : ij \in G(m)} (2.01 w_{ij} - l_{ij}) + \sum_{j \in T \setminus \{i\} : ji \in G(m)} (2.01 l_{ji} - w_{ji}) = p_i, \forall i \in T \tag{3}$$

$$w_{ij} + l_{ij} \leqslant 1, \forall ij \in G(m) \tag{4}$$

$$p_t \geqslant p_i - B r_i, \forall i \in T \setminus \{t\} \tag{5}$$

$$w_{ij}, l_{ij} \in \{0, 1\}, \forall ij \in G(m) \tag{6}$$

$$r_i \in \{0, 1\}, \forall i \in T \setminus \{t\} \tag{7}$$

The goal function minimizes the number of teams that are ranked before team t, and is scaled with the term 1 to indicate that the highest ranking it can obtain is the first place. The first set of constraints states that the number of points a team i has at the end of the season equals the points this team already has, plus 3.01 points for each win and 1 point for each draw in the games that are yet to be played. Since in case of an equal number of points, the team that won the highest number of games is to be ranked first, 3.01 points are added instead of 3. Notice that in case of a draw, both w_{ij} and l_{ij} equal 0. The situation where both w_{ij} and l_{ij} equal 1 is not allowed by the second set of constraints. Finally, we need to make sure that a team i will be ranked higher than t, if it obtained more points than t. When the parameter B is chosen equal to the total number of points that can be won in the competition, the final set of constraints will do just this.

With a limited number of changes, the above formulation can be used to determine the lowest ranking that this team t could still end up with. This allows us to determine whether a game is important for some team for some outcome or not. For instance, a game is important for the league title, if the highest position this team can still reach is the first, and if the lowest position this team can still drop to is lower than the first. Indeed, if the former was not the case, the team would no longer be able to win the championship, and if only the latter was not the case, the league title could no longer escape them. A similar reasoning can be made for the other relevant outcomes: qualification for European football and relegation.

We point out that our approach is in line with the Schilling measure (1), since when we find through optimization that a match is unimportant, this means that $p(X_i|W_{i,t+k}, H_t) = 0$ and thus $S_i(X)_{t,t+k} = 0$. The probabilities in the Schilling measure are however usually determined using a Monte

Carlo simulation (see [6]). Notice also that matches for which $S_i(X)_{t,t+k} = 0$ according to a Monte Carlo simulation need not be unimportant using our optimization approach.

4 Results

For each of the three league formats discussed in Sect. 2, we developed schedules and simulated match results for 10 series of 5 consecutive seasons. In Belgium, a calendar committee is responsible for collecting the wishes of the various stakeholders (e.g. clubs, police, TV station) and creating an acceptable schedule. As the number of wishes is considerable and sometimes conflicting, finding such a schedule is quite a hard nut to crack. Our schedules were developed according to calendar committee guidelines using a two-phase method for which we refer to [2]. In order to simulate match results, we estimated a trinomial probability distributions starting from 10 years historical data. If the result of a particular match, let's say a home game of team A against team B, was five times a win for team A, two times a draw, and three times a win for team B, the resulting probability distribution for that match would be: 50% chance team A wins, 20% chance of a draw, and 30% chance team B wins. In order to test the accuracy of these probability distributions, we used the results of the first half of season 2007–2008 (151 matches), and predicted the result corresponding with the highest probability in the trinomial distribution. For instance, in the example above, we would predict team A to win. This allowed us to correctly predict for the result of 51.0% of the matches. Compared to the more complex models presented in literature where accuracies between 40% and 50% are reported (see e.g. [5]), our simple approach based solely on historical match results performs remarkably well. When possible we have used the same match results in the various league formats (i.e. common random numbers).

Table 1 shows the percentage of unimportant games for each outcome and each of the league formats. The current league has a low number of unimportant games with respect to the league title and relegation. The Wijnants league offers the teams many ways to take part in the play-off for European qualification, which results in a very low percentage of unimportant games for

Table 1. Percentage of unimportant games

Outcome	Current league(%)	Play-off league(%)	Wijnants league(%)
League Champion	12.17	16.15	28.55
European football	9.05	8.45	4.82
Relegation	4.65	8.34	27.69
A series	–	–	3.34
Any outcome	1.51	1.25	0.53

this outcome. On the other hand, the number of unimportant games for the league title and relegation are quite high, which can be explained by the fact that teams in the A series of the spring competition cannot relegate to the second division, and that teams in the B series of the spring already lost their chances on the league title half-way the season. This alone results in almost 25% unimportant games for these outcomes. The Wijnants league however offers an extra objective to play for, namely promotion to (or maintaining a place in) the A series. This explains why the number of games that don't matter for any outcome is very low in this league format. The play-off league scores in between the other leagues for all outcomes. In general, the results show that both new leagues can increase the attractiveness, since overall less games will be unimportant, which is the main concern for the TV station.

5 Conclusions and Future Work

In this paper, we compared the current league format in Belgium's first division with two formats under consideration. We have presented a tool to evaluate the attractiveness of these league formats by measuring the number of unimportant games that is to be expected. We hope this paper can contribute to a well-founded choice for a league format and help to overcome the fear for change.

This topic leaves space for quite some future research. It would be very interesting to use our optimization approach to develop a measure for match importance (instead of unimportance), and compare it with other measures, as e.g. the one by Schilling [7]. Linking this research to the expected number of spectators would make the financial aspects of adopting a new league format more tangible.

References

1. Bojke, C.: The impact of post-season play-off systems on the attendance at regular season games. In: Albert, J., Koning, R.H. (eds.) Statistical Thinking in Sports, pp. 179–202. Chapman & Hall/CRC, Boca Raton (2007)
2. Goossens, D., Spieksma, F.: Scheduling the Belgian soccer league. Interfaces **39**(2), 109–118 (2010)
3. Hauspie, J.: Het plan-Wijnants voor de hervorming van de Jupiler League: simpele rekensom (in Dutch). Sport Voetbalmagazine **7**(48), 24–25 (2007)
4. Koning, R.: Post-season play and league design in Dutch soccer. In: Rodriguez, P., Garcia, J., Kesenne, S. (eds.) League Governance, Competition and Professional Sports, University of Oviedo, pp. 183–207 (2007)
5. McHale, I., Scarf, P.: Modelling soccer matches using bivariate discrete distributions with general dependence structure. Stat. Neerl. **61**, 432–445 (2007)
6. Scarf, P.A., Shi, X.: The importance of a match in a tournament. Comput. Oper. Res. **35**(7), 2406–2418 (2008)
7. Schilling, M.F.: The importance of a game. Mathematics Magazine **67**(4), 282–288 (1994)

Modelling Batting Strategy in Test Cricket

P. Scarf[1], X. Shi[2], and S. Akhtar[1]

[1] Salford Business School, University of Salford, Salford, M5 4WT, UK
p.a.scarf@salford.ac.uk, s.akhtar@pgr.salford.ac.uk
[2] Manchester Metropolitan University Business School, Manchester M1 3GH
x.shi@mmu.ac.uk

Summary. During the third innings, we suppose that the batting team selects a run-rate and target to optimise the match outcome probabilities. Outcome probabilities are calculated using a model for the outcome given the end of third innings position, and a model for the target set given the current position and the chosen run-rate. While the run-rate is not wholly in the control of the batting side, the approach described may allow a decision-maker to consider outcome probabilities if the team is able to bat the remainder of its third innings at a particular run-rate. This can then indicate whether an aggressive or defensive batting strategy is desirable.

1 The Decision Problem

During the third innings of a test match, a batting captain whose team is in a strong position will be thinking about how large a target to set his opponents and the ideal time for his opponents to begin their final innings. The captain is thus faced with a decision problem: what is the optimum target and what is the optimum time to set it. Because test match cricket is time-limited, if the target is large and the time it is set is late in the match, a draw is likely. Furthermore, the batting side is faced with the problem about how it should bat during the remainder of its innings: quickly with a high run-rate in order to set a competitive target with sufficient time to win the match, or slowly at a lower run rate in order to ensure that the match is not lost. Thus, in the third innings, taking a simple view of the problem, the match outcome depends on the values of the target aimed for, T, and the batting strategy, S (aggressive or defensive), in the third innings. Mathematically, we may attempt to choose T and S to maximise some objective such as the probability of a win. We consider this problem in the paper.

The problem is not straightforward. Firstly, generally speaking, the win and loss probabilities tend to move together, and maximising the probability of a win may also maximise the probability of a loss. Secondly, batting strategy is not observed. We therefore assume that the run-rate in the remainder of the

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*,
Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_75,

third innings is a surrogate for the strategy. To model the decision problem, we describe two sub-models: 1. for the match outcome probabilities given the end of third innings position, 2. for the runs scored in the remainder of the third innings. Target setting has been considered in one-day matches [3,11]. However, test matches are different because, in one-day matches, there is no notion of playing out the time remaining for a draw.

2 A Decision Model for Third Innings Batting Strategy

We label the team batting third as the reference team. Suppose the reference team aims to set a target for their opponents who bat last in the match. Call this target aimed for T. We assume that when this target is reached, the batting side declare-that is, they forfeit the remainder of the innings as allowed within the rules of cricket [7]. Further, suppose the reference team aims to bat towards this target at run-rate X. Thus, T and X are the decision variables in this formulation. The probability of a win (for the reference team) will depend on T, X and the current position.

Denote the current position by $P = (s, V_s, w)$, where s is the lead (for the reference team), V_s the overs remaining in the match and w the third innings wickets lost. Let t be the actual target set. This will be at most the target aimed for, and less if the reference team are all out before they reach their desired target. Thus $t \leq T$. Let Y denote the match outcome. Then, using $prob(Y = y|P, X, T)$ to denote the match outcome probabilities given the current position P and the choice of the decision variables, it follows that

$$prob(Y = y|P, X, T) = \sum_{t=s+1}^{T} prob(Y = y|W) prob(t|P, X, T), \qquad (1)$$

where $prob(t|P, X, T)$ is the probability distribution of the target established given the current position P and choice of the decision variables, and $prob(Y = y|W)$ is the probability of outcome Y given the covariates W that describe the match state at the end of the third innings and which include the target set, t, and overs remaining at the end of the third innings, V_t. Thus $W = (t, V_t, W')$, where W' denotes other covariates that do not vary over the remainder of the third innings. Note V_t is determined by V_s, t and $X : V_t = V_s - \{(t - s - 1/X\}$.

In order to proceed with the probability calculation in (1), we require suitable models for $prob(Y = y|W)$ and $prob(t|P, X, T)$. The first of these is developed in the next section. For the second, we consider the distribution of the further runs added in the third innings, and we suppose that the mean of this distribution depends on X. Then, we can determine the probability distribution of the actual target set t given the chosen target aimed for T and run-rate, X. The overs remaining in the match at the end of the third innings is a deterministic function of the overs remaining at the current position and t and X, and so $prob(Y = y|P, X, T)$ can be calculated. Of course, factors

other than just the current position and batting strategy in the remainder of the third innings will influence match outcome. The model developed can therefore only guide captains as they make decisions about batting strategy, and we expect the model outcomes to be modified in the light of experience regarding local conditions.

2.1 Outcome Probabilities Given the End of Third Innings Position

Scarf and Shi [15] develop a model for match outcome probabilities given the end of third innings position. Denoting the set of outcomes (win, draw, loss) by $(1, 0, -1)$, the covariates by W, and taking a draw as a reference category, nominal logistic regression assumes

$$prob(Y = y|W) = \begin{cases} e^{A_1}/(1 + e^{A_1} + e^{A_{-1}}), & y = 1 \\ 1/(1 + e^{A_1} + e^{A_{-1}}), & y = 0 \\ e^{A_{-1}}/(1 + e^{A_1} + e^{A_{-1}}), & y = -1 \end{cases} \tag{2}$$

where $A_1 = \alpha_1 + \beta_1^T W$, $A_{-1} = \alpha_{-1} + \beta_{-1}^T W$. The win and loss probabilities depend on the covariate W in different ways through β_1 and β_{-1} respectively. This model is equivalent in a sense to fitting two binary logistic regression models, the first for the win-draw probability comparison, the second for the loss-draw probability comparison. This model can be contrasted with the ordinal logistic regression model or proportional odds model [8].

The factors which impact on outcomes in cricket are extensive [1, 2, 14]. We are concerned principally with match state covariates, and data on the end of third innings position and match outcome for 301 test matches over the period $1998 - 2007$ have been collected (Table 1). This is a larger dataset than considered by Scarf and Shi [15]. Outline model statistics for various fitted models are shown in Table 2. Estimates for the highlighted model are shown in Table 3. With the run-rate in the first two innings, RR_{12}, we are attempting to capture the quality of the batting conditions. A covariate that represents the deterioration of a pitch would be of interest, and further work and perhaps more data would be beneficial to consider this. The pre-match strength of teams is considered through a variable that measures the difference in win percentage for the two teams over the last 10 years. The explanatory power of the declaration indicator variable is good and not surprising since it may well incorporate many factors, possible unmeasured, which lead to a captain declaring or otherwise. However, it would not make sense to use it as a covariate in a model to support decision making regarding declaration. Figure 1 shows the win probability from the fitted model as a function of target set and overs remaining. Note that the win probability increases to a peak and then decreases-if a very large target is set, the team batting last will not attempt to play for a win and a draw becomes more likely.

Table 1. Test match data (extract of 301 test matches, Feb 1998–Dec 2007)

Date	Home	Rankings	Away	Ranking	Venue	1st Innings	2nd Innings	Team batting 3rd (team A)	Follow-on (Y/N)	Target Set	Declaration $Y = 1, N = 0$	Overs remaining	Overs used	Results
13-2-98	WI	2	E	6	Port of Spain, Trinidad	159	145	WI	N	225	0	207	108	−1
27-2-98	WI	2	E	6	Georgetown, Guyana	352	170	WI	N	380	0	152	62	1
12-3-98	WI	3	E	6	Bridgetown, Barbados	403	262	E	N	375	1	109	37	0
30-1-98	A	1	SA	3	Adelaine	517	350	SA	N	361	1	109	108	0
07-1-98	SL	7	Z	9	Kandy	469	140	Z	Y	10	0	77	2	−1

Table 2. Results of fitting nominal logistic regression model to 301 test match outcomes (Feb 1998–Dec 2007) for various sets of predictors

Model	Parameter	Likelihood	AIC	Nag. $R^2(\%)$
$T + OR + RR_{12} + W\%D + D$	12	−125.35	274.70	81.58
$T + OR + RR_{12} + W\%D$	10	−131.31	282.62	80.26
$T + OR + RR_{12}$	8	−138.87	293.73	78.51
$T + OR + W\%D$	8	−137.37	290.74	78.87
$T + OR$	6	−142.13	296.27	78.24
$T + OR(ordinal)$	4	−198.08	404.16	61.32
$T + OR + T^2$	8	−140.65	297.30	78.10
$T + OR + OR^2$	8	−140.57	297.15	78.12
$T + OR + T * OR$	8	−140.42	296.84	78.14
$T + OR + D$	8	−134.86	285.71	79.45

Akaike information criterion (AIC) and Nagalkerke's R^2 [9] shown. with covariates lead (T), overs remaining (OR), run-rate in first two innings (RR_{12}), win percentage difference (W%D), and declaration indicator (D)

2.2 The Distribution of Target Set

To model $prob(t|P, X, T)$, we proceed as follows. Let $Z|P, X$ be the total further runs added by the reference team in their third innings from the current position, P, if they complete each remaining partnership and bat at run-rate X. (For notational convenience we let $Z = Z|P, X$). Then $t = min(Z+s+1, T)$. That is, $Z+s$ would be the lead if the reference team batted until all ten wickets were lost. So, given the distribution of Z, we can determine the distribution of $t|P, X, T$. At the current position, there are w wickets down, and so the further runs added is given by $Z = Z'_{w+1}|X + \sum_{k=w+2}^{10} Z_k|X$, where $Z'_{w+1}|X$ is the additional runs added in the current partnership and $Z_k|X$ is

Table 3. Fitted parameter estimates for minimum AIC nominal logistic regression model (2) with covariates lead (T), overs remaining (OR), run-rate in first two innings (RR12), and win percentage difference ($W\%D$), with standard errors and p-values

		Coefficient	s.e	p-value
Win/draw	Intercept	−4.8841	1.6633	0.003
(1/0)	Over remaining (OR)	0.0518	0.0087	0.000
	Target (T)	−0.0069	0.0033	0.037
	$run - rate_{12}$ (RR_{12})	0.7260	0.4947	0.142
	$win\%diff$ ($W\%D$)	−0.0167	0.0122	0.170
Loss/draw	Intercept	−2.0650	1.9776	0.296
(−1/0)	Over remaining (OR)	0.0538	0.0093	0.000
	Target (T)	−0.0290	0.0038	0.000
	$run - rate_{12}$ (RR_{12})	1.6975	0.6258	0.007
	$win\%diff$ ($W\%D$)	−0.0276	0.0156	0.077

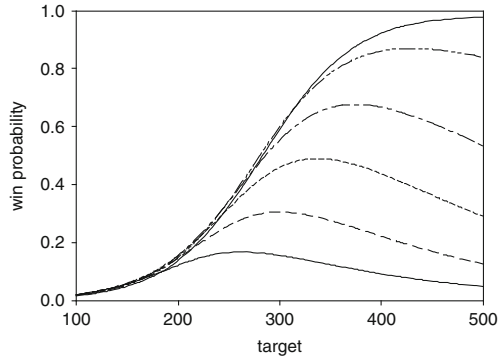


Fig. 1. Win probability for the team batting third as a function of target established and overs remaining: bottom curve 60 overs remaining; next 80 overs; 100 overs; 120 overs; 150 overs; top 200 overs. 301 test matches (Feb 1998–Dec 2007). $RR_{12} = 3.12$, $W\%D = 0$

the runs scored in the k^{th} partnership, $k = w + 2, \ldots, 10$. Now we require a model for the distributions of $Z_k|X$ and $Z'_{w+1}|X$.

Two distributions offer themselves naturally as candidates: the geometric and the negative binomial [5,10,13]. Scarf et al. [16] proposed the zero-inflated negative binomial distribution to explain the excessive number of zero scores, and found a reasonable fit. Kimber and Hansford [6], on the other hand, argue that no simple parametric distribution provides a wholly convincing fit. The zero-inflated negative binomial distribution, denoted $ZINB(\pi, \theta, p_0)$, is given by:

$$prob(Z = z) = \begin{cases} p_0 & z = 0, \\ (1 - p_0)\Gamma(z + \pi)\theta^{\pi}(1 - \theta)^{y}/\{z!\Gamma(\pi)(1 - \theta^{\pi})\} & z \geq 0 \end{cases} \quad (3)$$

$(0 < p_0 < 1)$. This implies that $E(Z) = \mu_z = (1 - p_0)\pi(1 - \theta)/\{\theta(1 - \theta^{\pi})\}$. We assume that $Z_k|X \sim ZINB(\pi_k(X), \theta_k, P_{0,k})$, that is, the parameters vary by partnership number, and π is a function of the run-rate, X. It then follows that the mean partnership score is a function of partnership number and X. A gamma function for $\pi_k(X)$ allows a maximum mean score for a finite value of X, as suggested by Fig. 2. Various forms for $\pi_k(X)$ are shown in Table 4. The parameterisation (3) ensures that $p_{0,k}$ does not depend on X. If such a dependence existed it could not be estimated. The run-rate is zero (no runs scored) in approximately 8% of partnerships. We use the minimum AIC model in Table 4 for the calculation of match outcome probabilities.

If a partnership is some way through, we assume a lack-of-memory property, so that $Z^{'}_{w+1} = Z_{w+1}$ - equivalent to the notion that the current position is at the fall of a wicket. We approximate the distribution of Z by $ZINB(\pi_z, \theta_z, p_z)$, with π_z and θ_z and p_z obtained by equating moments. Thus, setting $E(Z) = \mu_z = \sum_{k=w+1}^{10} \mu_k$, $Var(Z) = \sigma_z^2 = \sum_{k=w+1}^{10} \sigma_k^2$, and $prob(Z|P, X = 0) = p_z = \prod_{k=w+1}^{10} p_{0,k}$, we can solve for (π_z, θ_z, p_z). As previously stated, since $t = min(Z + s + 1, T)$, we can determine $prob(t|P, X, T)$ from the distribution of Z. The approximations that we use imply that the match outcome probability calculations are not exact, but we would anticipate only a small error arising here.

Table 4. AIC for various models of the distribution of run scored in a partnership, $Z_k \sim ZINB(\pi_k(X), \theta_k, p_{0,k})$

Models	LL	NP	AIC
$\pi_k = \alpha_k, \theta_k = \theta, p_{0,k} = p_0$	−6017.771	12	12059.542
$\pi_k = \alpha_k, x^{\beta} exp(-\gamma x), \theta_k = \theta, .$	−5797.870	14	11623.740
$\pi_k = \alpha_k, x^{\beta} exp(-\gamma x), \theta_k$ varying, $p_{0,k} = p_0$	−5768.538	23	11583.076
$\pi_k = \alpha_k, x^{\beta} exp(-\gamma x), \theta_k$ varying, $p_{0,k}$ varying	−5762.722	32	11589.444
$\pi_k = \alpha_k, x^{\beta} exp(-\gamma x), \theta_k = \theta,$ negative binomial	−5993.870	13	12013.740
$\pi_k = \alpha_k, x^{\beta} exp(-\gamma x), \theta_k = \theta, P_{0,k} = p_0$	−5866.749	5	11743.498

Data comprise all partnerships in third innings ($n = 1412$). *NP* number of parameters, *LL* log-likelihood

3 Example

For brevity, we present just one example, Table 5. It can be seen here that, with South Africa trailing in the series, if they wanted to maximize their win probability, and throw caution to the wind, they should bat recklessly and aim to set a target between 280 and 300. Ultimately, they batted more cautiously

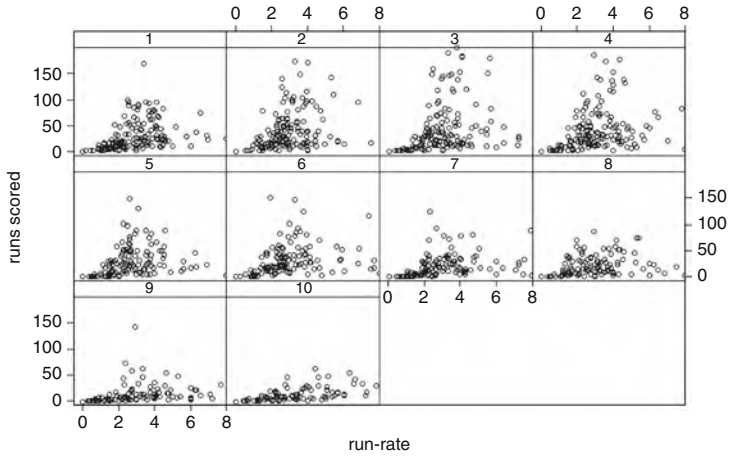


Fig. 2. Scatter plots of runs scored vs. run-rate in a partnership by partnership number

and still lost. Note, the entries in these tables have been implemented on a spreadsheet that allows for the updating of the calculations as the current position changes. Thus, it is implied that the decision support is provided continuously; this allows for 'over-by-over' and 'run-by-run' updating.

4 Discussion

The aim of this paper is to model optimum batting strategy in the third innings in test cricket. We would like to be able to model strategy given any match position. Looking at the third innings has two benefits. Firstly, some progress can be made with the mathematical problem. Secondly, batting is perhaps more strategic during this innings than in others. In the second and first typically teams will just try and score as much as possible, and in the final innings a team will be either trying to win or save a game. We approach the mathematical problem by supposing that the third innings run-rate and the target that the side batting third aims to set its opponent are decision variables. That is, we suppose that these are within the control of the batting side, and the batting side will, given the current match state, choose a run-rate and a target that are most desirable, be it to maximise the probability of a win or to minimise the probability of a loss, or some combination of the two. Of course, the run-rate is not strictly in the control of the batting side. The run-rate is merely a random variable that depends to some (unknown) extent on the batting strategy. Therefore, the output from the decision support model that we propose should be used, by a team batting third, to consider how match outcome probabilities vary with run-rate in the remainder of the third innings and target aimed for, broadly indicating how the side should try

Table 5. Match outcome probabilities (as percentages) given current position as a function of target aimed for, T, and (projected) run-rate, X

	Projected Run-Rate, X																	
	2			3			4			5			6			8		
Target, T	% win	% draw	% loss	% win	% draw	% loss	% win	% draw	% loss	% win	% draw	% loss	% win	% draw	% loss	% win	% draw	% loss
220	16	36	48	17	30	53	18	27	55	18	26	56	18	25	57	19	23	58
240	15	58	27	18	48	34	20	42	38	21	39	40	22	37	41	23	34	44
260	11	75	14	17	64	19	20	57	23	22	53	25	23	50	27	25	44	31
280	8	84	8	14	77	10	18	70	13	21	65	15	23	61	17	25	52	22
300	6	88	6	10	85	5	15	79	6	18	74	8	21	69	10	24	58	18
320	5	89	5	8	90	3	12	85	3	16	80	5	19	75	6	23	61	15
340	5	90	5	6	92	2	9	89	2	13	84	3	16	79	4	22	63	14
360	5	90	5	5	94	1	8	91	1	11	87	2	15	82	3	22	64	14
380	5	90	5	4	94	1	6	93	1	9	89	2	13	84	3	21	65	14
400	5	90	5	4	95	1	5	94	1	8	90	1	12	85	3	21	65	14

Australia vs South Africa, 2006, 3rd test of 3, Australia leading series $1-0$. South Africa 1st innings 451 in 155 overs, Australia 1st innings 359 in 95 overs ($RR_{12} = 3.24, W\%D = -18$), current position (start of final day, South Africa 94/3): reference team South Africa; lead 186; overs remaining 90; 3rd innings wickets down 3. SA added 100 in 20 overs, setting Australia a target of 287 in 68. Australia reached 288/2 in 61 to win by eight wickets

to bat. A captain would of course take account of other factors such as the state of the series, the state of the pitch, and possibly the weather. Since test matches are always played as part of a series, typically comprising three or five matches between the same two teams, the attitude of the side batting third to risk will depend very much on the state of the series. Generally, declaring captains act conservatively.

The problem addressed is a special case of the general problem of determining playing strategy given the match state. To date, the most general approach to this problem is described in the context of one-day internationals [11,12]. The problem can be stated generally as follows: if $X(t)$ is the match state at time t, and Y is the match outcome, what is $prob(X(t_1)|X(t_0), S)(t_0 < T)$, and so what playing strategy S should be adopted in the period (t_0, T)? The run-rate is used as a surrogate for S in this paper. In football, one might attempt to use the positions of players on the pitch, and modern data collection systems may allow the calculation of the "centre of gravity" of a team over time [4]. Alternatively, the decision maker might explore different $X(t_1)$ scenarios (which are plausible given $X(t_0)$ and S) by considering $prob(Y|X(t_1))$ and the decision maker's own subjective transition probabilities if strategy S is adopted. Opponents also make strategic choices, and so modelling matches as dynamic games would be interesting.

It remains an open question as to whether a quantitative approach can provide a competitive edge. Perhaps decision-makers possess an intuition about match outcomes that is more than sufficient for their purpose, and perhaps factors that are not quantified, such as the state of the pitch, and weather conditions, are so influential that they render our analysis too simple to be useful. On the other hand, the analysis in this paper might provide a tool that allows a decision-maker to explore various options quickly, while subjectively adjusting the model outputs to accommodate local conditions.

References

1. Allsopp, P.E., Clarke, S.R.: J. R. Statist. Soc. A. **167**, 657–667 (2004)
2. Brooks, R.D., Faff, R.W., Sokulsky, D.: Appl. Econ. **34**, 2853–2365 (2002)
3. Clarke, S.R.: J. Opl. Res. Soc. **39**, 331–337 (1988)
4. Salvo, D., Collins, V., McNeill, B., Cardinale, M.: Int. J. Perform. Anal. Sport **6**, 108–119 (2006)
5. Elderton, W.P.: J. R. Statist. Soc. A. **108**, 1–11 (1945)
6. Kimber, A.C., Hansford, A.R.: J. R. Statist. Soc. A. **156**, 443–455 (1993)
7. MCC: www.lords.org.uk/cricket/laws.asp (2003)
8. McCullagh, P., Nelder, J.A.: Generalized Linear Models, 2nd edn. Chapman and Hall, London, (1989)
9. Nagalkerke, N.J.D.: Biometrika. **78**, 691–692 (1991)
10. Pollard, R., Benjamin, B., Reep, C.: In Optimal Strategies in Sport. In: Ladany, S.P., Machol, R.E. (eds.) pp. 118–195. North Holland, New York (1977)
11. Preston, I. Thomas, J.: Statistician **49**, 95–106 (2000)
12. Preston, I., Thomas, J.: Statistician **51**, 189–202 (2002)
13. Reep, C., Pollard, R., Benjamin, B.: J. R. Statis. Soc. A. **134**, 623–629 (1971)
14. Ringrose, T.: J. R. Statis. Soc. A. **169**, 903–911 (2006)
15. Scarf, P.A., Shi, X.: IMA J. Man. Math. **16**, 161–178 (2005)
16. Scarf, P.A., Shi, X., Akhtar, S.: Technical Report 336/10, University of Salford (2010)

Minisymposium *Interactions between Structure and Process in Manufacturing Systems*

D. Hömberg

Weierstrass Institute for Applied Analysis and Stochastics, Mohrenstr. 39, 10117 Berlin, Germany, hoemberg@wias-berlin.de

The goal of this mini-symposium was to shed light on the interactions between processes and structures in modern production facilities. The knowledge of these interactions will enable a reliable and reproducible process control including ramp-up and the systematic design of production equipment – tasks that currently involve a high amount of empirical knowledge. The reason for this is the close connection of the properties of the (machine) structure and the (manufacturing) process with regard to the result of the manufacturing process. This connection has not yet been investigated in detail.

The applications considered in the minisymposium range from metal chipping, milling, and grinding to robot-guided laser material treatments. The mathematical tasks covered are important modelling issues arising in the coupling of process and structure, including mechanical interactions and the role of heat production and release, numerical methods for an efficient simulation of the arising coupled systems and, last but not least, their optimal control.

The talk by Matthias Maischak from Brunel University dealt with the simulation of metal chipping. Here, the focus lay on the efficient numerical simulation using a boundary element and finite element coupling procedure for the elastoplastic thermo-mechanical contact problem with a linear elastic work tool and an elastoplastic work piece. Unfortunately, Matthias' presentation is not included in the proceedings. Further information about his research can be found on his webpage.[1]

Oliver Rott from the Weierstrass Institute, Berlin, studies the influence of machine and structure on the stability of milling processes. The model consists of a harmonic oscillator equation for the dynamics of the cutter and a linear viscoelastic workpiece model. The coupling through the cutting force adds delay terms and further nonlinear effects. Numerical results show that the model is capable of predicting instabilities due to a lack of workpiece stiffness.

[1] www.brunel.ac.uk/about/acad/siscm/maths/people/acad/matthiasmaischak

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*, Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_76,

In the long run this work may lead to a more precise prediction of stability charts.

The contribution by Heribert Blum and Andreas Rademacher from Technische Universität Dortmund is related to the NC-shape grinding process. They use an empiric force model in conjunction with a geometric kinematical simulation to model the process. The machine model is based on a finite element simulation, in which the spindle and the grinding wheel are explicitly considered. The remaining parts of the grinding machine are modelled by elastic bearings. The simulations are coupled by the exchange of the predicted grinding force, which is used as Neumann type boundary condition in the finite element simulation, and of the displacement of the grinding wheel, which changes the contact conditions in the geometric-kinematical simulation. Because of the varying length scales, the diameter of the grinding wheel is about 100 mm and the depth of cut is less than 1 mm, adaptive finite element algorithms are an appropriate tool to obtain an efficient simulation. The main focus of their paper is the derivation of a goal-oriented error estimation for the linear wave equation and a corresponding adaptive refinement algorithm.

Andreas Steinbrecher from Weierstrass Institute, Berlin, considers robot-guided laser material treatments. Up to now, mathematical models for laser treatments usually assume the path of the laser on the workpiece to be known. However, depending on the workpiece geometry and the desired production goal the necessary laser path can not always be realized. To this end Andreas studies different strategies of coupling the optimal control of the laser heat treatment with the path-planning of the laser-guiding industrial robot.

Modelling, Analysis and Stability of Milling Processes Including Workpiece Effects

D. Hömberg and O. Rott

Weierstrass Institute for Applied Analysis and Stochastics, Mohrenstr. 39, 10117 Berlin, Germany, hoemberg@wias-berlin.de, rott@wias-berlin.de

Summary. A common model for milling dynamics, i.e. a harmonic oscillator, was extended by a continuous, linear thermo-elastic model describing the dynamic behaviour of the workpiece. A widely studied, empirical cutting force model is used to describe the coupling of both systems. Finally, a numerical solution strategy for the coupled system is outlined and complemented by numerical simulations that show the work piece effect on the stability of the cutting process.

1 Introduction

A milling machine is a machine tool for the shaping of metal or other solids. Its basic components are a rotating cutter, a spindle, a z-slider that is attached to a moving portal and a table on which the workpiece is mounted. The modelling of milling dynamics, the determination of stable cutting conditions and the design of more efficient milling machines are important research fields in production technology. Effective methods to predict stable processes have been developed in recent years [2, 6]. An essential part of these methods is an abstract dynamical model that reproduces the local characteristics of the actual milling system in terms of the dynamics at the tip of the cutter. In combination with a process model to describe the cutting forces it leads to a delay-differential equation (DDE), whose stability characteristics have been widely studied in the last decades [6, 8].

The focus of this paper is a detailed study of the machine work piece interactions. Hence, in addition to the DDE model for the cutter the workpiece is accounted for by a visco-elastic material model. The coupling is realised through the cutting force. This approach allows for a refined stability analysis and will eventually lead to a refined prognosis of stable cutting conditions.

The paper is organised as follows: In Sect. 2 we derive the model equations. An algorithm for the numerical approximation of the new milling model is outlined in Sect. 3. Numerical results for different scenarios corresponding to stable and unstable cutting conditions are included. The last section is devoted to some concluding remarks.

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*, Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_77,

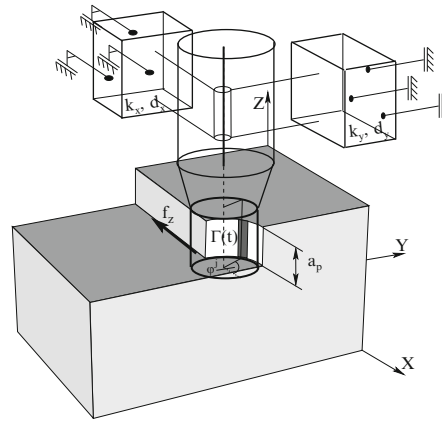


Fig. 1. Schematic representation of the milling process

2 Modelling

2.1 Model Equations

To avoid technicalities we represent the machine dynamics by an harmonic oscillator equation for the centre of mass $q = (q_x, q_y, q_z)^T$, of the cutter with mass m_c, written in the inertial reference frame (x, y, z), cf. Fig. 1. Note that this model reproduces at least one relevant mode of the milling machine. Furthermore, the cutter oscillates only in the x, y-plane. The additional z-component has been introduced to ease the coupling with the workpiece model. The coordinates in the oscillator reference frame are related to those of the workpiece reference frame by a linear, time dependent transformation $(x, y, z) = (X, Y, Z) - b(t)$, where $b(t) = (X_0 - f_z \frac{t}{\tau}, Y_0, Z_0)$ denotes the translation vector given in the workpiece frame. Hence, the equation of motion for the cutter model reads:

$$m_c \ddot{q} + D\dot{q} + Kq = [F_x, F_y, 0]^T, \tag{1}$$

where D denotes the damping and K the stiffness matrix. The right hand side of (1) takes into account the cutting force, a sum of the forces acting on each tooth in cut.

We assume that the largest part of the workpiece behaves like a visco-elastic solid. Only in the vicinity of the cutting edge, visco-elasto-plastic effects have to be taken into account. To some extend, these effects are already included in the empirical cutting force model. Therefore, we focus here on the visco-elastic behaviour of the workpiece, which we model with the standard equations of linear visco-elasticity with Kelvin–Voigt damping (see, e.g., [5]) assuming small strains.

2.2 Coupling

Boundary Conditions

The basis for the coupling of workpiece and machine model is the cutting force. Usually, the latter is computed in terms of the so-called *uncut chip thickness* (see, e.g., [1]), which describes the thickness of the material to be removed by the tooth which is in cut . Here we use the following algebraic relation between uncut chip thickness and the cutting forces due to Weck [9]:

$$\hat{F} = \left(\hat{F}_R, \hat{F}_T, \hat{F}_Z\right)^T = a_P \hat{K}(v_C) \max(h, 0), \tag{2}$$

where $\hat{K}(v_C)$ denotes the vector of cutting constants which can be a function of the cutting speed v_C [6]. Note that the precise form of $\hat{K}(v_C)$ has to be found experimentally.

An expression for the uncut chip thickness h can be derived by considering a two dimensional, independently vibrating work-piece-cutter system, for details we refer to [3]:

$$h = -f_z \cos\varphi^j + (q(t) - q(t-\tau)) \cdot e_r^j - (u(t, R_P) - u(t-\tau, R_Q)) \cdot e_r^j. \tag{3}$$

Here, R_P and R_Q denote the two work piece material points being currently machined. We notice that the uncut chip thickness consists of three different parts. The first one represents just the cutter displacement due to the given feed f_z. Projected on the radial direction, it yields the stationary uncut chip thickness. The second part represents the machine oscillations and produces the modulation of the chip thickness that has been identified to be the main reason for chatter. The third contribution to the uncut chip thickness is related to the workpiece deformation.

On the cutting edge the forces act in three directions: perpendicular to the cutting velocity, in opposite direction to the cutting velocity and parallel to the rotation axis of the cutter. Note that the z-component of $\hat{K}(v_C)$ vanishes for orthogonal cutting. We transform (2) into the workpiece reference frame and sum up for all teeth to obtain:

$$F = (F_x, F_y, F_z)^T = -\sum_{j=1}^{N_z} g(\varphi^j(t)) O(\varphi^j(t)) \hat{F}. \tag{4}$$

Here, $g = 1$ if the corresponding tooth 'j' is in cut and $g = 0$, otherwise. The orthogonal matrix $O(\varphi^j)$ transforms the forces $\hat{F}$ into the workpiece reference frame. N_z denotes the number of cutter teeth.

Now we define the boundary condition for the momentum balance equation as:

$$\begin{aligned} u &= 0 \qquad \text{on} \qquad \Gamma_D, \\ \sigma \cdot n &= \begin{cases} \frac{F}{|\Gamma(t)|} & \text{on } \Gamma(t) \times (0, t_e). \\ 0 & \text{otherwise.} \end{cases} \end{aligned} \tag{5}$$

$|\,\Gamma(t)\,|$ denotes the measure of the area, where the cutting force acts on the workpiece (cf. Fig. 1).

3 Numerical Simulations

3.1 The Algorithm

We present a straightforward numerical algorithm to compute an approximate solution of the coupled system in time domain. Recall that in a milling process with nonzero feed material is removed from the workpiece. Since we use a visco-elastic work piece representation that cannot take the chip removal into account directly, we introduce the following approximation of this process. At first we divide the time interval $[0, t_e]$ into subintervals with the length τ, tacitly assuming t_e to be a multiple of τ. Based on these subintervals, we construct a sequence of space-time-cylinders $Q_l = \Omega_l \times (k\tau, (k+1)\tau]$ with an update of the work piece shape according to the theoretical tooth path given by the static chip thickness (i.e. the first term in (3)).

This approach allows us to use the methods of steps to solve the coupled PDE/DDE system: For given initial data on the interval $[-\tau, 0]$ we solve the system in $[0, \tau]$. Then iteratively, we use the solution in $((l-1)\tau, l\tau]$ as initial data for the following tooth period and perform the analysis for the interval $(l\tau, (l+1)\tau]$. With the help of this technique we proved the existence of a unique weak solution of the coupled system in the entire time interval $[0, t_e]$, see e.g. [3].

With the above considerations we may now introduce a time integration scheme for an arbitrary tooth period $(l\tau, (l+1)\tau]$. To this end, we discretise the pde part of the coupled system with linear finite elements in space. Thus, we obtain a system of ordinary differential equations with delay. In our time stepping strategy we make use of an incremental decoupling such that the momentum balance can be solved in each time step with a Newmark scheme [7]. We integrate the remaining DDE with a standard ode-solver, i.e. Runge–Kutta-54 [4]. All retarded and coupling terms are provided by means of interpolation. The data transfer from one tooth period to another is also carried out by interpolation.

Note that the presented method is restricted to cutting conditions where only one tooth is in cut. However, form a practical point of view this poses no severe limitation.

3.2 Simulation Results

The system of equations has unstable and stable solutions depending on the parameters rotation speed n, number of teeth N_z, axial depth of cut a_p and radial depth of cut. For the simulations we fix the cutting parameters $f_z = 0.2\,\text{mm}$, $N_z = 4$, $D = 15\,\text{mm}$ and $n = 7{,}500\,\text{rpm}$. For the first example the

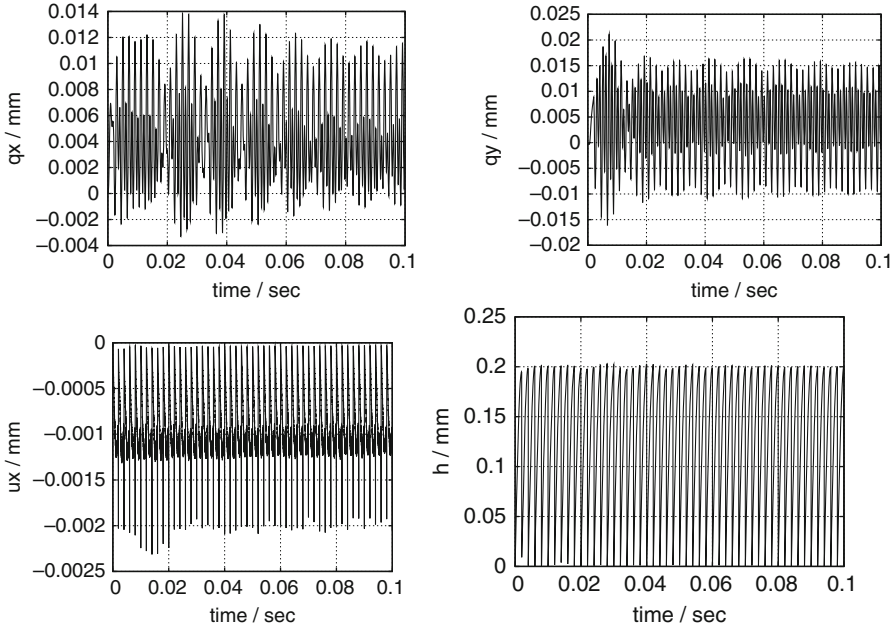


Fig. 2. x,y-component of the cutter vibrations (q_x, q_y), mean workpiece deformations (u_x) at the cutting edge and the uncut chip thickness (h)

axial depth of cut is $a_p = 1\,\text{mm}$, the entry and exit angles are $\varphi_s = \pi/2$, $\varphi_E = \pi$ and we have chosen a rather rigid workpiece geometry, corresponding to the work piece depicted in Fig. 1. We set $t_e = 50\tau$, i.e., for the choice of n and N_z, $t_e = 0.1\,\text{s}$. In Fig. 2, we see the results of the first simulation run. Since the uncut chip thickness converges to a stationary state, we identify this milling process as stable. The induced deformations do not interfere with the stable cutting conditions. The workpiece deformations are one order of magnitude smaller than the cutter vibrations.

The second example shows the work piece effect on the stability of milling process. The cutter was assumed to be rigid, which means that only the work piece may cause unstable cutting conditions. We performed the simulations for a rather rigid work piece geometry, as shown in Fig. 1 and a beam-like work piece which has a low bending stiffness. The cutting parameters are $a_p = 15\,\text{mm}$, $\varphi_s = \pi/2$ and $\varphi_E = 3\pi/4$ and we simulated 20 tooth periods, i.e. for the choice of n and N_z, $t_e = 0.04\,\text{s}$. While the uncut chip thickness converges for the stable workpiece geometry to the stationary evolution, we observe for the beam like, unstable workpiece geometry a divergence of the uncut chip thickness indicating the occurrence of unstable cutting conditions, i.e. chatter.

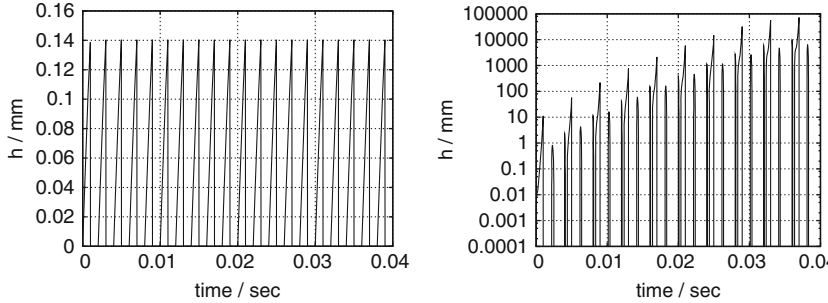


Fig. 3. Uncut chip thickness for a stable and an unstable work piece

4 Conclusions

The goal of this paper was to enhance existing models of the milling process to allow for the consideration of the workpiece influence. The simulations in Sect. 3 show that the model is capable of reproducing instability effects due to a lack of workpiece stiffness.

The results are promising and open up various directions for future research, such as the development of an efficient numerical tool for the systematic derivation of stability diagrams for the coupled system.

References

1. Altintas, Y.: Manufacturing Automation Cambridge University Press, Cambridge (2000)
2. Altintas, Y., Weck, M.: Ann. CIRP **53/2**, 619–642 (2004)
3. Chelminski, K., Hömberg, D., Rott, O.: WIAS Preprint No. **1392**, (2008)
4. Deuflhard, P., Bornemann, F.: Scientific Computing with Ordinary Differential Equations. Springer, New York (2002)
5. Dill, E.H.: Continuum Mechanics. CRC, New York (2007)
6. Faassen, R.P.H., van de Wouw, N., Oosterling, J.A.J., Nijmeijer, H.: Int. J. Mach. Tool Manufact. **43**, 1437–1446 (2003)
7. Hughes, T.J.R.: The Finite Element Method. Dover Publications, New York (2000)
8. Insperger, T., Stépán, G.: Int. J. Numer. Methods Eng. **61**, 117–141 (2004)
9. Weck, M., Teipel, K.: Dynamisches Verhalten spanender Werkzeugmaschinen. Springer, Berlin (1977)

Adaptive Finite Element Discretisation of the Spindle Grinding Wheel System

H. Blum and A. Rademacher

Institute of Applied Mathematics, Technische Universität Dortmund, 44221 Dortmund, Germany, Heribert.Blum@mathematik.tu-dortmund.de, Andreas.Rademacher@mathematik.tu-dortmund.de

Summary. In the simulation of the NC-shape grinding process, a finite element model of the grinding machine is included. To enhance the accuracy and efficiency of the finite element computation, a posteriori error estimation and resulting adaptive mesh refinement techniques are used. In this note, a dual weighted a posteriori error estimate for a linear second order hyperbolic model problem is derived. Numerical results illustrate the performance of the presented approach.

1 Introduction

To model the interaction between the grinding process and the machine structure is indispensable in the simulation of the NC-shape grinding process. The coupling of separate machine and process simulations is a common simulation approach. We use an empiric force model in conjunction with a geometric-kinematical simulation to model the process [7]. The machine model is described in [15]. It is based on a finite element simulation, in which the spindle and the grinding wheel are explicitly considered. The remaining parts of the grinding machine are modelled by elastic bearings. The simulations are coupled by the exchange of the predicted grinding force, which is used as Neumann type boundary condition in the finite element simulation, and of the displacement of the grinding wheel, which changes the contact conditions in the geometric-kinematical simulation. Because of the varying length scales, the diameter of the grinding wheel is about 100 mm and the depth of cut is less than 1 mm, adaptive finite element algorithms are an appropriate tool to obtain an efficient simulation.

In general, a posteriori error estimates for second order hyperbolic problems are possible for two different discretisation approaches. One of them uses space time Galerkin methods for discretisation and applies similar techniques for error control as in the static case [2, 3, 10, 12]. The other one is based on finite differences in time and finite elements in space. Here, separate error estimators are used for the space and time direction [9, 11, 16] or error estimates for the whole problem [1, 6] are derived.

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*, Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_78,

Starting from the weak formulation of the wave equation, a space time Galerkin discretisation is introduced in Sect. 2. In Sect. 3, the goal-oriented a posteriori error estimate is derived and an adaptive refinement algorithm based on it is discussed. Numerical results, which illustrate the performance of the developed approach, are presented in Sect. 4. The article concludes with a discussion of the results and an outlook.

2 Continuous Formulation and Space Time Galerkin Discretisation

Based on the weak formulation of the linear wave equation, a space time nonconforming Petrov Galerkin scheme with continuous basis functions in time is introduced. We consider the linear wave equation

$$\rho\ddot{u} - \operatorname{div}(\kappa\nabla u) = f \tag{1}$$

on the domain $\Omega \subset \mathbb{R}^2$ and the time interval $I = [0,T]$ with the initial conditions $u(0) = u_s$ and $\dot{u}(0) = v_s$ as well as homogeneous Dirichlet boundary conditions. For notational simplicity, the density ρ is set equal to 1. The parameter κ describes the elasticity coefficient.

Rewriting (1) as a first order system, multiplying by suitable test functions, and spatial integration by parts lead to the weak formulation:

$$\forall\varphi = (\psi,\chi) \in U \times V: \quad A(w,\varphi) = 0 \tag{2}$$

Here, $w = (u,v) \in U \times V$ is the weak solution and

$$\begin{aligned} U &:= \left\{u | u \in L^2(I;H_0^1(\Omega)),\ \dot{u} \in L^2(I;L^2(\Omega))\right\}, \\ V &:= \left\{v | v \in L^2(I;L^2(\Omega)),\ \dot{v} \in L^2(I;H^{-1}(\Omega))\right\} \end{aligned}$$

are the appropriate trial spaces, which are continuously embedded into $C\left(I;L^2(\Omega)\right)$. The bilinear form A is given by

$$\begin{aligned} A(w,\varphi) := {} & ((\dot{u}-v,\psi)) + ((\dot{v},\chi)) + (a(u)(\chi)) - ((f,\chi)) \\ & + (u(0)-u_s,\psi(0)) + (v(0)-v_s,\chi(0)) \end{aligned}$$

with $((\psi,\chi)) := \int_0^T \int_\Omega (\psi\chi)\,dx\,dt$ and $a(u,\chi)) := \int_0^T (\kappa\nabla u, \nabla\chi)\,dt$.

The time interval I is decomposed into M subintervals $I_m := (t_{m-1}, t_m]$ with $0 = t_0 < t_1 < \ldots < t_{M-1} < t_M = T$ and $k_m := t_m - t_{m-1}$. The finite element trial space in time step m, V_h^m, is based on the spatial mesh $\mathbb{T}_h^m$ and on bilinear basis functions. In time, piecewise linear continuous basis functions are used for the trial space and piecewise constant functions for the test space:

$$\begin{aligned} V_{kh} &:= \left\{ v_{kh} \in C(I;H_0^1(\Omega)) \,\middle|\, v_{kh|_{I_m}} \in \tilde{\mathcal{P}}_1(I_m,V_h^m),\ v_{kh}(0) \in V_h^0 \right\} \\ W_{kh} &:= \left\{ v_{kh} \in L^2(I;H_0^1(\Omega)) \,\middle|\, v_{kh|_{I_m}} \in \mathcal{P}_0(I_m,V_h^m),\ v_{kh}(0) \in V_h^0 \right\} \end{aligned}$$

The space $\tilde{\mathcal{P}}_1(I_m, V_h^m)$ is a slight modification of the space of linear polynomials (see [14]). Eventually, the discrete problem is to find $w_{kh} = (u_{kh}, v_{kh}) \in V_{kh} \times V_{kh}$ with

$$\forall \varphi_{kh} = (\psi_{kh}, \chi_{kh}) \in W_{kh} \times W_{kh} : \quad A(w_{kh}, \varphi_{kh}) = 0. \tag{3}$$

3 A Posteriori Error Estimation

In this section, the a posteriori error estimate is derived. In the first step an abstract result from [5] is applied on the present situation. Then, the error estimate is transformed into a computable estimate by well-tested approximations. Afterwards, it is used as basis for an adaptive refinement process.

Functionals of interest of the form $J(w) := \int_0^T J_1(w(t))\,dt$ are considered, where J_1 is an arbitrary three times continuously differentiable functional. The Lagrangian is defined by $\mathcal{L}(w, z) := J(w) - A(w)(z)$. We say $(w, z) \in (U \times V) \times (V \times U)$ is a stationary point of $\mathcal{L}$, if

$$\forall (\delta w, \delta z) \in (U \times V) \times (V \times U) : \quad \mathcal{L}'(w, z)(\delta w, \delta z) = 0.$$

The discrete stationary point $(w_{kh}, z_{kh}) \in (V_{kh} \times V_{kh}) \times (W_{kh} \times W_{kh})$ is given analogously. Following the results in [5,13], we obtain the abstract error representation

$$\begin{aligned} J(w) - J(w_{kh}) &= \frac{1}{2}\mathcal{L}'(w_{kh}, z_{kh})(w - \tilde{w}_{kh}, z - \tilde{z}_{kh}) + \mathcal{R}_{kh} \\ &= \frac{1}{2}\rho(w_{kh})(z - \tilde{z}_{kh}) + \frac{1}{2}\rho^\star(w_{kh}, z_{kh})(w - \tilde{w}_{kh}) + \mathcal{R}_{kh}, \end{aligned}$$

with arbitrary $\tilde{w}_{kh} \in V_{kh} \times V_{kh}$ and $\tilde{z}_{kh} \in W_{kh} \times W_{kh}$. In the proof, we have to pay attention to the fact that a nonconforming Petrov Galerkin discretisation scheme is used. Here, the primal and the dual residual are given by

$$\begin{aligned} \rho(w)(\varphi) &:= \mathcal{L}'_z = -A(w, \varphi) \\ \rho^\star(w, z)(\varphi) &:= \mathcal{L}'_w = J'(\varphi) - A(\varphi, z), \end{aligned}$$

respectively. The remainder term $\mathcal{R}_{kh}$ is bounded above by the third power of the error.

The weights, which represent the interpolation error, are approximated by $w - \tilde{w}_{kh} \approx \Pi_{kh} w_{kh}$ and $z - \tilde{z}_{kh} \approx \Pi_{kh} z_{kh}$. Here, the operator Π_{kh} is given by $\Pi_{kh} := i_{kh} - \mathrm{id}$, where i_{kh} is a patchwise interpolation of higher order [4]. The operator Π_{kh} approximates the interpolation error in space and time. The spatial counterpart is defined by $\Pi_h := i_h - \mathrm{id}$ and the temporal one by $\Pi_k := i_k - \mathrm{id}$. Eventually, we obtain the computable error representation

$$J(w) - J(w_{kh}) \approx \frac{1}{2}\left[\rho(w_{kh})(\Pi_{kh} z_{kh}) + \rho^\star(w_{kh}, z_{kh})(\Pi_{kh} w_{kh})\right].$$

Using the identity

$$\Pi_{kh}\varphi_{kh} = i_k \Pi_h \varphi_{kh} + \Pi_k \varphi_{kh},$$

which holds true for tensor product trial functions, the error estimate is split into a temporal part η_k and a spatial part η_h:

$$\begin{aligned} J(w) - J(w_{kh}) &= \frac{1}{2}\left[\rho(w_{kh})(\Pi_k z_{kh}) + \rho^\star(w_{kh}, z_{kh})(\Pi_k w_{kh})\right] \\ &\quad + \frac{1}{2}\left[\rho(w_{kh})(i_k \Pi_h z_{kh}) + \rho^\star(w_{kh}, z_{kh})(i_k \Pi_h w_{kh})\right] \\ &=: \eta_k + \eta_h. \end{aligned}$$

The spatial residual terms are integrated by parts to localise the error estimate as basis for an adaptive refinement process. This process consists of several steps. In the first step, a space time refinement strategy decides, whether a refinement in spatial or temporal direction or in both directions is performed. We use an equilibration strategy, which was developed in [14]. The temporal refinement strategy is a simple fixed fraction strategy [4]. In space, a more complex global fixed fraction strategy [14] is used. There, all refinement indicators of all mesh cells are compared. After the adaptive refinement, the meshes are regularised to ensure a suitable structure, which includes only single hanging nodes in space and time and a patch structure property [8].

4 Numerical Results

The domain of the spindle grinding wheel system contains several re-entrant corners. Furthermore, the material is varying throughout the domain. A model example for this difficulties is an L-shape domain with varying material, which is considered here. The data of the example is chosen as:

$$\begin{aligned} \Omega \times I &:= ([-0.5, 0] \times [-0.5, 0.5]) \cup ([0, 0.5] \times [-0.5, 0]) \times [0, 1] \\ \kappa &:= 1 + \min\{1, 10(x_1 - 0.05)_+\} \\ f &:= 100 \mathbb{I}_{x_1 \geq \frac{1}{4} \wedge t \in ([0, \frac{1}{4}] \cup [\frac{1}{2}, \frac{3}{4}])} \\ J(w) &:= \frac{1}{|I||B|} \int_I \int_B u(x,t)\, dx\, dt, \quad B := B^\infty_{\frac{1}{8}}\left(-\frac{1}{4}, \frac{1}{4}\right)^T \end{aligned}$$

In Fig. 1, the spatial meshes of different time steps are depicted. In the beginning, the cells in the area of the acting force are refined. Along the outgoing wave, the mesh is refined. The inner edge of the L-shape domain is especially well resolved. At the end, the domain of interest B gets more and more refined. The second impulse is not considered, since the arising wave does not reach B. In Fig. 2, the development of the error in the functional of interest is shown over the complete number of mesh cells. The calculation with dynamic meshes

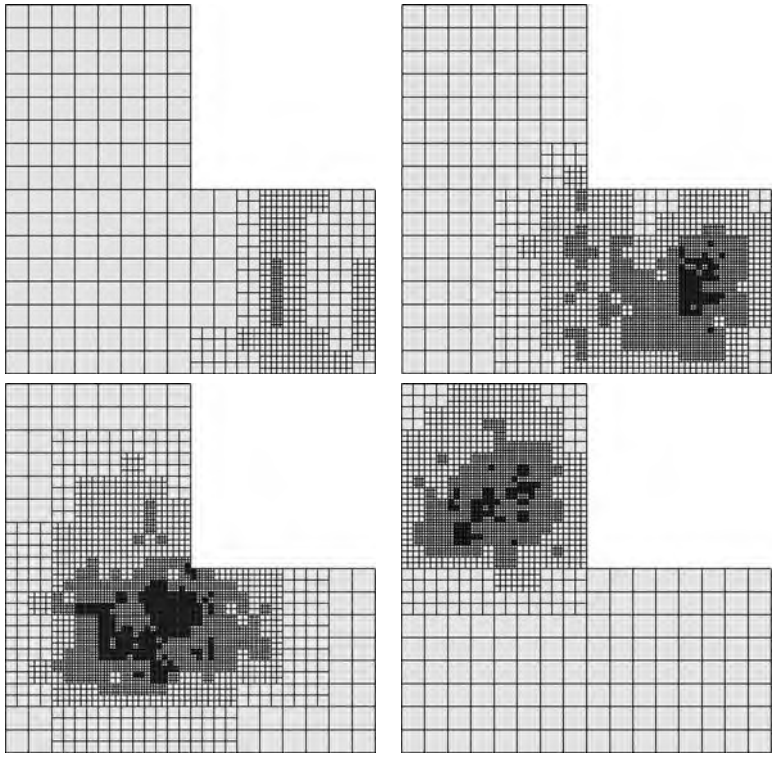


Fig. 1. Meshes for different time steps ($n = 1$, $n = 50$, $n = 100$, $n = 150$)

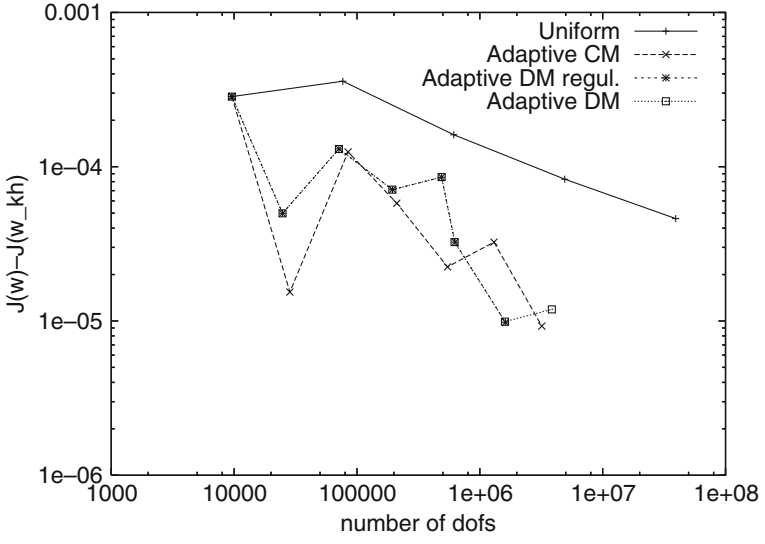


Fig. 2. Convergence of the adaptive method

is most efficient, followed by the calculation with an adaptively refined mesh, which is kept constant during one refinement cycle. The graph of the calculation without temporal mesh regularisation shows the need of the algorithm to ensure the proper convergence of the adaptive method. The effectivity indices are in the range of 1.

5 Conclusions and Further Work

In this article, we have presented a new approach to goal-oriented error estimation for the linear wave equation. It leads to well adaptively refined meshes and enhances the efficiency of the finite element discretisation.

The extension of the presented approach to nonlinear second order hyperbolic problems will be considered in a separate article. The mesh refinement and regularisation algorithms will be enhanced further and elaborately analysed.

References

1. Adjerid, S.: Comput. Methods. Appl. Mech. Eng. **191**, 4699–4719 (2002)
2. Aubry, D., Lucas, D., Tie, B.: Comput. Methods Appl. Mech. Eng. **176**, 41–50 (1999)
3. Bangerth, W., Rannacher, R.: East-West J. Numer. Math. **7**, 263–282 (1999)
4. Bangerth, W., Rannacher, R.: Adaptive Finite Element Methods for Differential Equations. Birkhäuser, Basel (2003)
5. Becker, R., Rannacher, R.: An Optimal Control Approach to Error Estimation and Mesh Adaptataion in Finite Element Methods, Acta Numerica 2000. Cambridge University Press, Cambridge (2001)
6. Bernardi, C., Süli, E.: Math. Models Methods Appl. Sci. **15**, 199–225 (2005)
7. Biermann, D., Blum, H., Jansen, T., Rademacher, A., Scheidler, A., Weinert, K.: Proceedings of the 1st Cirp International Conference on Process Maschine Interaction, pp. 279–288. (2008)
8. Rademacher, A.: Adaptive Finite Element Methods for Nonlinear Hyperbolic Problems of Second Order. Ph.D. Thesis, Technische Universität Dortmund (2010)
9. Bornemann, F.A., Schemann, M.: Comput. Visual. Sci. **1**, 137–144 (1998)
10. Johnson, C.: Comput. Methods Appl. Mech. Eng. **107**, 117–129 (1993)
11. Li, X., Wiberg, N.-E., Zeng, L.F.: Earth. Eng. Struct. Dyn. **21**, 555–571 (1992)
12. Li, X., Wiberg, N.-E.: Comput. Methods Appl. Mech. Eng. **156**, 211–229 (1998)
13. Meidner, D.: Adaptive Space-Time Finite Element Methods for Optimization Problems Governed by Nonlinear Parabolic Systems. Ph. D. Thesis, Ruprecht-Karls-Universität Heidelberg (2008)
14. Schmich, M., Vexler, B.: SIAM J. Sci. Comput. **30**, 369–393 (2008)
15. Weinert, K., Blum, H., Jansen, T., Rademacher, A.: Prod. Eng. **1**, 245–252 (2007)
16. Xie, Y.M., Zienkiewicz, O.C.: Earth. Eng. Struct. Dyn. **20**, 871–887 (1991)

Optimal Control of Robot Guided Laser Material Treatment

Andreas Steinbrecher

Weierstrass Institute for Applied Analysis and Stochastics, Germany,
anst@wias-berlin.de

1 Introduction

Laser material treatments such as hardening or welding have become a basic part of the process chain for sophisticated metal workpieces. Mounted on industrial robots, laser treatment devices become increasingly important in automated manufacturing, especially in automotive industry.

For the employment of single robots a number of planning tools is available, considering issues like path-planning, control, collision detection, etc. but disregarding the specific task the robot has to perform. Up to now it is always assumed that the track along which the laser light impinges on the workpiece surface is precisely known. But the most natural criterion to decide whether the employment of a robot has been successful is not the tracking of a prescribed path but the question if the robot has achieved its production goal.

In this article we will consider the optimal control of robot guided laser material treatments, where the multibody system model of a robot is coupled with a PDE model of the laser treatment. We will present and discuss several optimization approaches in view of a robust and suitable numerical solution. We will illustrate the approaches in an application to the surface hardening of steel.

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*,
Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_79,

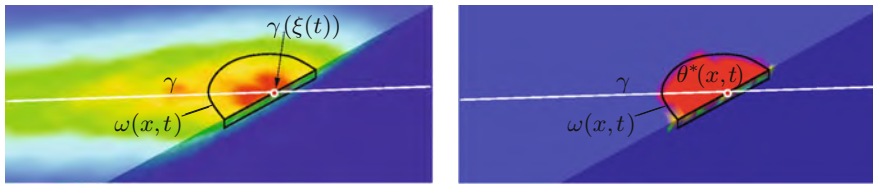


Fig. 1. (**a**) target curve γ and target region ω, (**b**) target region ω and desired temperature profile θ^*

2 The Mathematical Model

The *production goal* in the case of surface hardening is to achieve a desired temperature profile θ^* inside a moving target region described by ω in a solid body Ω. This target region is given as moving flat cylinder sliding under the surface Γ with constant velocity along the target curve γ with $\gamma(\xi) \in \Gamma$ for all $\xi \in [0,1]$. The parameter $\xi = \xi(t)$ depends on the time $t \in \mathbb{I} = [0,T]$ and determines the movement of the target region ω (Fig. 1).

The heat conduction on $\Omega \times \mathbb{I}$ will be modeled by use of the heat equation

$$\rho c \frac{\partial}{\partial t}\theta - \kappa \Delta \theta = F \quad \text{with } \theta(x,0) = \theta_0 \text{ on } \Omega \text{ and } \frac{\partial}{\partial \nu}\theta(x,t) = 0 \text{ on } \Gamma \times \mathbb{I}. \quad (1)$$

The temperature distribution is denoted by θ, the mass density by ρ, the specific heat by c, and κ denotes the heat conductivity. The laser is modeled as a distributed heat source in the right-hand-side $F = F(x,t,l,u_l)$ of the heat equation and depends on the laser position l (to be specified later) and the laser power u_l. Furthermore, the motion of the robot in general will be modeled by use of the equations of motion in descriptor form, see [3]. Since our considerations are based on serial robots the equations of motion have the form $\dot{p} = v$ and $\dot{v} = f(p,v,t) + u_r$ in $\mathbb{I}$ with $p(0) = p_0$, $v(0) = v_0$. which can be written as

$$\dot{q} = k(q,u_r,t), \quad q(0) = q_0 \quad (2)$$

with $q^T = [p^T, v^T]$ and $q^T(0) = [p_0^T, v_0^T]$. Here, p denotes the position (joint angles, joint displacements), v the according velocity, and u_r the control of the robot. For more details in the modeling we refer to [1–3].

For the numerical solution of the state equations (1) and (2) and its adjoint equations we use FEM tools and grid generator provided by `pdelib`[1] combined with Runge-Kutta methods. The right-hand-side k of the model equations (2) of the used robot are provided by INVISION[2].

[1] `pdelib` is a collection of software components for solving PDEs. In particular finite volume and finite element methods are supported. `pdelib` is developed by Weierstrass Institute for Applied Analysis and Stochastics (WIAS) in Berlin.

[2] INVISION is a software package for the real-time simulation of manufacturing-plants developed and supported by Rücker EKS. Its data base includes all relevant industrial robots currently used in industry.

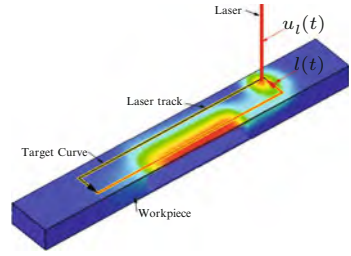
In the following we will consider several approaches for the optimization problem to minimize the objective functional of the form $J = J_\theta + J_R$, where

$$J_\theta = \frac{1}{2}\int_0^T \int_\Omega (\omega(x,t)(\theta(x,t) - \theta^*(x,t)))^2 \mathrm{d}x \ \mathrm{d}t$$

evaluates the obtained temperature profile θ within the moving target region ω and J_R represents some regularization and/or penalty terms which vary depending on the approach. The optimization is done using the gradient method.

3 Optimization Approaches and Numerical Results

Laser Power and Laser Position Optimization: In this approach we let the laser follow the target curve γ such that the laser track l is given by $l(t) = \gamma(s(t))$. Then for $t \in \mathbb{I}$ we want to find an optimal *laser power* $u_l(t)$ and *laser position* $s(t) \in [0,1]$ achieving the production goal by minimizing



$$J(u_l, s) = J_\theta + \frac{\alpha}{2}\int_0^T ||u_l(t)||^2 \ \mathrm{d}t + \frac{\beta}{2}\int_0^T ||\ddot{s}(t)||^2 \ \mathrm{d}t \tag{3}$$

subject to the heat equation (1). Using the Lagrange approach we can derive the first order optimality conditions consisting of (1), (4), (5), (6):

$$-\rho c \frac{\partial}{\partial t}\mu_\theta - \kappa \Delta \mu_\theta = \omega^2(\theta - \theta^*) \text{ on } \Omega \times \mathbb{I} \text{ with } \mu_\theta(x,T) = 0 \text{ on } \Omega \tag{4}$$
$$\text{and } \frac{\partial}{\partial \nu}\mu_\theta(x,t) = 0 \text{ on } \Gamma \times \mathbb{I},$$

$$\int_0^T \Big(\alpha u_l + \int_\Omega \mu_\theta \frac{\partial}{\partial u_l} F(x,t,l,u_l) \ \mathrm{d}x\Big)(u_l - \bar{u}_l) \ \mathrm{d}t \geq 0 \quad \text{for all } \bar{u}_l \in U_{ad}^{u_l}, \tag{5}$$

$$\int_0^T \Big(\beta s^{(4)} + \int_\Omega \mu_\theta \frac{\partial}{\partial l} F(x,t,l,u_l) \frac{\partial}{\partial s}\gamma(s) \ \mathrm{d}x\Big)(s - \bar{s}) \ \mathrm{d}t \geq 0 \text{ for all } \bar{s} \in \mathcal{U}_{ad}^s, \tag{6}$$

where $\mathcal{U}_{ad}^{u_l}$ and $\mathcal{U}_{ad}^s$ are the admissible sets of u_l and s, respectively.

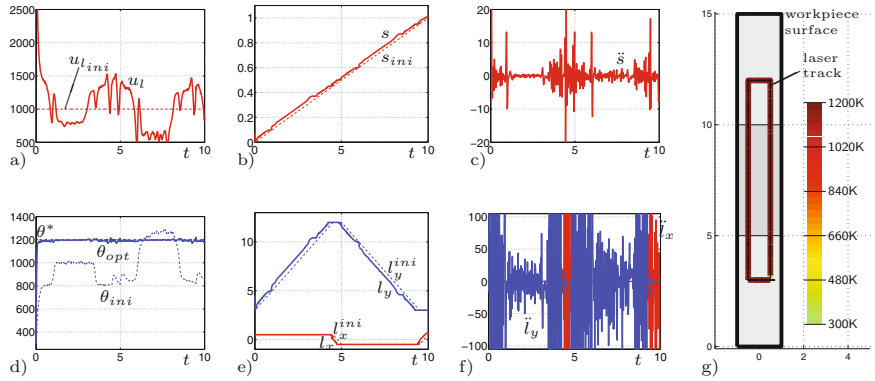


Fig. 2. (**a**) laser control u_l, (**b**) laser position s along the target curve γ, (**c**) laser acceleration $\ddot{s}$ along the target curve γ, (**d**) temperature in target region $\omega(t)$, (**e**) laser track l , (**f**) laser acceleration $\ddot{l}$, (**g**) laser treatment result in the target curve γ

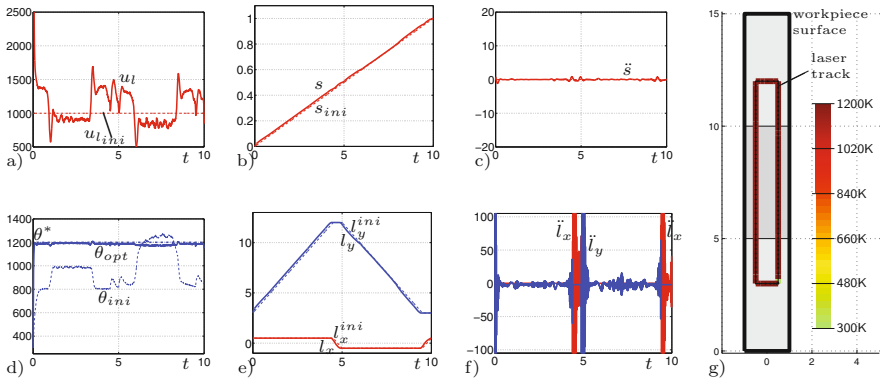


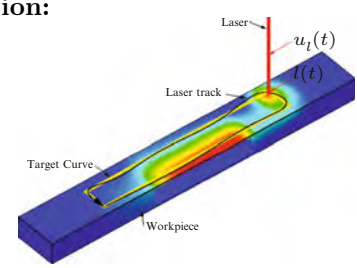
Fig. 3. (**a**) laser control u_l, (**b**) laser position s along the target curve γ, (**c**) laser acceleration $\ddot{s}$ along the target curve γ, (**d**) temperature in target region $\omega(t)$, (**e**) laser track l , (**f**) laser acceleration $\ddot{l}$, (**g**) laser treatment result in the target curve γ

The numerical results of the optimization are illustrated in Fig. 2. The achieved temperature in the target region (Fig. 2d) matches the desired one very well. But the use of $\beta = 0$ allows large oscillating acceleration $\ddot{s}$ of the laser within the target curve because large values of $\ddot{s}$ are not penalized in the objective functional. In general, this laser path is not realizable by an industrial robot.

To reduce the oscillations in $\ddot{s}$ and to improve the realizability we increase the parameter β to 100 and get the results shown in Fig. 3. Although the high acceleration of the laser is reduced as long as the target curve is smooth, the acceleration of the laser passing the corners in the target curve is still infinitely large (Fig. 3f) and therefore, again not realizable.

This approach is suitable as long as the curvature of the target curve remains in certain bounds depending on the acceleration bounds of the robot.

Laser Power and Laser Track Optimization: In this approach we will use the *laser track* $l = [l_x, l_y]$ on the workpiece surface itself as optimization variables in addition to the *laser power* u_l. Therefore, we want to determine optimal functions u_l and l satisfying the production goal by minimizing



$$J(u_l, l) = J_\theta + \frac{\alpha}{2} \int_0^T ||u_l(t)||^2 \, \mathrm{d}t + \frac{\beta}{2} \int_0^T ||\ddot{l}(t)||^2 \, \mathrm{d}t$$

subject to the heat equation (1). We obtain the first order optimality conditions (1), (4), (5), (7):

$$\int_0^T \Big(\beta l^{(4)} + \int_\Omega \mu_\theta \frac{\partial}{\partial l} F(x, t, l, u_l) \, \mathrm{d}x \Big) (l - \bar{l}) \, \mathrm{d}t \geq 0 \text{ for all } \bar{l} \in \mathcal{U}_{ad}^l. \tag{7}$$

The numerical optimization result is shown in Fig. 4. The obtained temperature in the target region (Fig. 4c) fits almost the desired one, such that the production goal is reached. Note that the additional freedom in the choice of the laser track yields a smoother laser track (Fig. 4f and g). Furthermore,

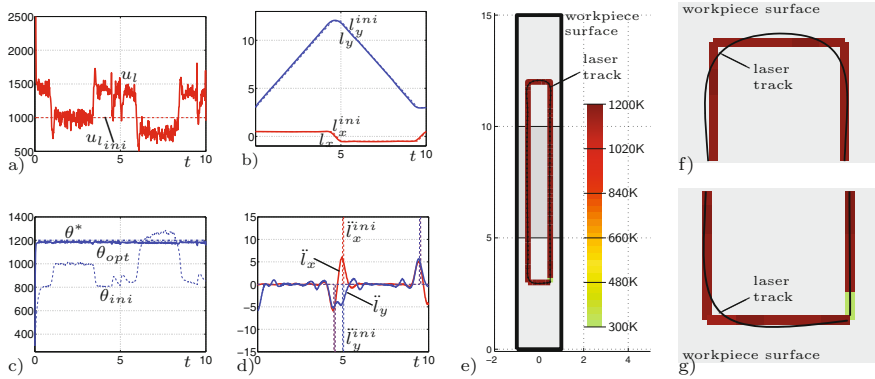


Fig. 4. (**a**) laser control u_l, (**b**) laser track l on the surface, (**c**) temperature in target region $\omega(t)$, (**d**) laser acceleration $\ddot{l}$, (**e**) laser treatment result in the target curve γ, (**f**) laser treatment result (zoom), (**g**) laser treatment result (zoom)

the acceleration of the laser (Fig. 4d) is smoother and smaller than in the previous example, cf. Fig. 3f).

Hence, we may conclude that this laser track is realizable by industrial robots even in the case of nonsmooth target curves. The β acts as a regularization parameter to achieve realizable laser tracks l.

Laser Power and Robot Control Optimization: In the following we are looking for the *robot control* u_r and the *laser power* u_l satisfying the production goal, i.e., we consider the minimization of

$$J(u_l, u_r) = J_\theta + \frac{\alpha}{2} \int_0^T ||u_l(t)||^2 \, \mathrm{d}t + \frac{\beta}{2} \int_0^T ||u_r(t)||^2 \, \mathrm{d}t$$

subject to (2), describing the robot, coupled with (1), describing the laser treatment. The laser track is given as function of q, i.e, $l = l(q)$. Then, the first order optimality conditions are given by (1), (2), (4), (8), (5), (9):

$$\dot{\mu}_q^T = -\int_\Omega \mu_\theta \frac{\partial}{\partial q} F(x, t, l(q), u_l) \, \mathrm{d}x - \mu_q^T \frac{\partial}{\partial q} k(q, u_r, t) \quad \text{with } \mu_q^T(T) = 0, \tag{8}$$

$$\int_0^T \Big(\beta u_r^T + \mu_q^T \frac{\partial}{\partial u_r} k(q, u_r, t) \Big)(u_r - \bar{u}_r) \, \mathrm{d}t \geq 0 \text{ for all } \bar{u}_r \in \mathcal{U}_{ad}^{u_r}. \tag{9}$$

The numerical results are depicted in Fig. 5. As shown in Fig. 5c) the desired temperature θ^* is reached in the middle of the time interval $\mathbb{I}$, but it close to its starting and end points. In the beginning, first the robot has to be accelerated such that the laser hits the moving target region in an appropriate way. Unfortunately, the gradient method reacts too sensitive on changes in the robot control. Small changes in the control in the beginning of the interval $\mathbb{I}$ have great influence on the laser track at the end of $\mathbb{I}$. Therefore,

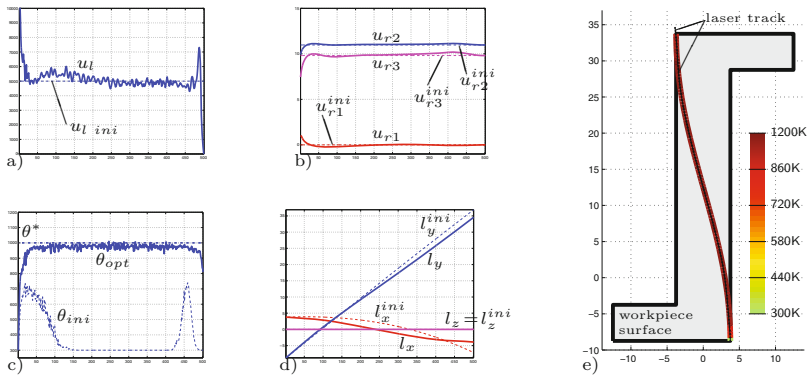


Fig. 5. (**a**) laser control u_l, (**b**) robot control u_r, (**c**) temperature in target region $\omega(t)$, (**d**) laser track l , (**e**) laser treatment result in the target curve γ

the gradient method has to perform very small gradient steps to improve the objective functional over the whole interval $\mathbb{I}$ leading to an extremely slow convergence. For more details we refer to [1].

4 Summary

We introduced and discussed several approaches for the optimal control of robot guided laser material treatments. The *laser power and laser position optimization* is suitable for smooth target curves since the laser position is restricted to the target curve. The *laser power and laser track optimization* is suitable for arbitrary target curves since the laser position is freely choosable on the surface. The relation to a robot guiding the laser is possible by use of the acceleration penalty term. Furthermore, the gradient method applied to the *laser power and robot control optimization* converges too slowly since the model reacts too sensitive on perturbations/changes of the robot control.

We suggest a hybrid optimization approach, the *laser power and laser track optimization* followed with a standard path planning approach to compute the robot control.

References

1. Hömberg, D., Steinbrecher, A., Stykel, T.: Optimal control of robot guided heat treatment. Technical report, DFG Research Center Matheon, WIAS Berlin, Berlin, Germany. in preparation
2. Hömberg, D., Weiss, W.: PID-control of laser surface hardening of steel. IEEE Transactions on Control Systems Technology, 2006
3. Steinbrecher, A.: Numerical Solution of Quasi-Linear Differential-Algebraic Equations and Industrial Simulation of Multibody Systems. PhD thesis, Technische Universität Berlin, 2006

Minisymposium *Mathematical Models for Supply Chains*

S. Göttlich and A. Klar

TU Kaiserslautern, Department of Mathematics, Postfach 3049, 67653 Kaiserslautern, Germany, goettlich@mathematik.uni-kl.de, klar@itwm.fraunhafer.de

Modeling of supply chains covers a broad mathematical spectrum which allows for application-oriented as well as more theoretical results. Nowadays, where simulation and optimization issues are of interest, mathematicians, engineers and economists focus on the computational validation of those models. From a mathematical point of view, one can mainly distinguish two classes of models: continuous and discrete ones. The latter are either common in particle-based simulations of complex production systems or optimization problems. However, continuous models are used for the simulation of large-scaled supply chains where not only feasible solutions and predictions come first but also fast computing times. The following articles exactly pick up all these features and contribute new and current results in this direction.

Ute Ziegler, from RWTH Aachen University, presents a discrete optimization model which supports the decision-making process in planning new production lines. The objective therein is to minimize investment, production and transportation costs while a multiple set of time-dependent constraints must be fulfilled. To find feasible and optimal solutions, several starting heuristics are implemented and efficiently tested on sample examples.

Simone Göttlich et al. describe a continuous model based on a coupled set of ordinary differential equations involving customer demands, order policies and money flows. The reformulation as an ODE-restricted optimization model has been proposed to determine suitable order and distribution strategies. It is furthermore shown that maximizing the profit of liquid suppliers in this model is independent of internal pricing and does not effect any policies.

Kathrin Padberg, from TU Dresden, and her co-workers focus on rate equations propagating material flows in push or pull (supply or demand-oriented) supply chains. The performance of a stability analysis provides the well-known Bullwhip effect (oscillating demand blow-up) as a mathematical instability which can be only prevented by a mixed push-pull-strategy.

Laurent Navoret, from the Université Paul Sabatier of Toulouse, and his collaborators concentrate on the interdisciplinary issue of economics and

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*, Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_80,

biology. The idea is to analyze the limiting procedure from microscopic (fine scale) to macroscopic (coarse scale) congestion models. Using the movement of sheep herds as example, where transition regimes from dilute to gregarious phases play an important role, the authors point out possible links for ongoing research in the field of supply chains.

Marco Laumanns, from ETH Zurich, describes a stochastic optimization model for transshipments of goods under uncertain demand in inventory-distribution networks. The question is how to choose a cost-effective alternative of either additional transportation costs or high inventory costs. One way to quantify this relation is to determine optimal control policies for each alternative and to compute the resulting average cost savings under these policies as the value of the transshipment option.

Design Network Problem and Heuristics

U. Ziegler[1] and S. Göttlich[2]

[1] Lu7 Mathematik-Kontinuierliche Optimieruny, RWTH Aachen, Templergraben 55, 52062 Aachen, Germany, ziegler@mathc.rwth-aachen.de
[2] Department of Mathematics, TU Kaiserslautern, Postfach 3049, 67653 Kaiserslautern, Germany, goettlich@mathematik.uni-kl.de

Summary. A mixed-integer model based on a coupled system of differential equations is presented in order to optimize design and material distribution of production networks. Due to many binary variables arising in this model and in order to guarantee feasible solutions several starting heuristics, which provide incumbents for the branch and cut algorithm, are developed and compared.

1 Introduction

In general, the manufacturing of goods requires a multitude of different operations accomplished by a vast number of miscellaneous suppliers and processors. The system of all feasible consecutive production steps is comprised by a production network, describing the manufacturing of goods, either starting from the required raw material or from semi-finished parts that pass through several processors until the finished product is obtained. Each processor is specified by a set of several properties, including a limited production capacity. This causes the need of a buffer where arriving parts queue up until they can be treated. On this scale we work with a continuous approach where the evolution inside the network is described as a continuous flow. This is advantageous for large scale problems, since the computation time is independent of the number of parts in the system as described in [1], amongst others. The task of finding the optimal dynamic distribution of flow through the network due to a certain objective function leads to a PDE-& ODE-constraint optimization problem, see [4]. After several discretization and linearization steps we end up with a linear mixed-integer programming formulation (MIP) which has been derived in [2]. Finding the optimal network design is a typical question in the context of planning and managing production systems. Hence, the original MIP model is expanded by the task of selecting optimal processor configurations out of a prescribed set. The resulting enlarged MIP uses a specific objective function where the focus is on minimizing diverse costs, see Sect. 2. Further details are explained in [3]. Instances of the model are solved

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*,
Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_81,

by Cplex developed by ILOG, see [5], a commercial solver using a LP-based branch and cut algorithm, which is the currently most reliable solving method for mixed integer programming problems. Due to the vast number of binary variables the design model is highly complex which means that the computation time explodes already for small sized instances. The solver often runs several hours without finding a single feasible solution. Consequently, the use of fast heuristical algorithms that provide feasible starting solutions for the branch and cut algorithm is advisable. Several approaches are suggested in Sect. 3. In Sect. 4 the typical course of the optimization algorithm with and without prepended heuristics is compared using some sample instances.

2 Design Model

A production network is described by a directed graph $G = (V, E)$, where V denotes the set of vertices and E the set of edges. Each vertex v possesses a set of incoming edges, denoted by δ_v^{in}, and a set of outgoing edges δ_v^{out}. We assume the network to have exactly one inflow edge e_{in}, where the material is induced into the system. All remaining edges represent locations where a processor possibly can be installed. The set of valid processor configurations for each edge is denoted by C_e, $\forall e \in E$. Every processor is described by several properties concerning the production capacity $\mu_{e,c}$, the production velocity $u_{e,c}$, the length $L_{e,c}$, setup costs $\zeta_{e,c}^{setup}$ and running costs $\zeta_{e,c}^{run}$, $\forall c \in C_e, e \in E$. After discretization of time and space of the material flow we end up with the following variables for each edge. The flow entering an edge at timestep t is denoted by z_e^t. The entering material either has to be stored in a queue, which is referred to as q_e^t, or it enters the processor. The flow entering the processor is denoted by x_e^t and the flow leaving a processor is referred to as y_e^t. The timestep is denoted by t going from 0 to n_t, and the timestep size is given by Δt. As initial condition we assume the network to be empty at $t = 0$. In the sequel, the design model is stated and shortly explained. All constraints hold for all $e \in E$. $c \in C_e$ and $t = 0, \dots n_t$:

$$\min w^{setup} \cdot p^{setup} + w^{run} \cdot p^{run} + w^{queue} \cdot p^{queue} - w^{out} \cdot p^{out} \qquad \text{(obj)} \quad (1)$$

subject to

$$y_e^t \lesseqgtr y_e^{t-1} + \frac{\Delta t}{L_{e,c}} u_{e,c} \cdot (x_e^{t-1} - y_e^{t-1}) \pm B_e \cdot (1 - \gamma_{e,c}) \qquad \text{(flow inside pr.)} \quad (2)$$

$$q_e^t = q_e^{t-1} + \Delta t (z_e^{t-1} - x_e^{t-1}) \qquad \text{(queue)} \quad (3)$$

$$\sum_{e \in \delta_v^{out}} z_e^t = \sum_{\tilde{e} \in \delta_v^{in}} y_{\tilde{e}}^t \qquad \text{(coupling constraints)} \quad (4)$$

$$\mu_{e,c} \kappa_e^t - \mu_{e,c}(1 - \gamma_{e,c}) \leq h_{e,c}^t \leq \mu_{e,c} \cdot \gamma_{e,c} \qquad \text{(flow entering processor)} \quad (5)$$

$$x_e^t = \sum_{c \in C_e} h_{e,c}^t \qquad \text{(flow entering processor)} \quad (6)$$

$$\frac{q_e^t}{\epsilon} - M\kappa_e^t \leq x_e^t \leq \frac{q_e^t}{\epsilon} \quad \text{(flow entering processor) (7)}$$

$$0 \leq z_e^t \leq D_e \cdot \sum_{c \in C_e} \gamma_{e,c} \quad \text{(configuration selection) (8)}$$

$$\sum_{c \in C_e} \gamma_{e,c} \leq 1 \quad \text{(configuration selection) (9)}$$

$$0 \leq x_e^t \leq \mu_e^{max}, 0 \leq y_e^t \leq \mu_e^{max}, 0 \leq q_e^t \quad \text{(box constraints) (10)}$$

$$\kappa_e^t, \gamma_{e,c} \in \{0,1\} \quad \text{(integrality constraints) (11)}$$

The binary variable $\gamma_{e,c}$ is a flag indicating whether configuration c is selected for edge e or not, by setting $\gamma_{e,c}$ to one or to zero respectively. Constraint (9) ensures, that at most one processor may be selected for each edge. In (8) it is requested that one processor has to be activated as soon as material is lead to the corresponding edge. Constraint (2) originates from a linear advection equation discretized by a one-sided Upwind scheme and describes the transport of flow inside a processor. The boundary condition is given by the induced material flow at the inflow edge, $x_{e_{in}}^t$, which is given in advance. The evolution of the queue in front of each edge is discretized with the explicit Euler method and stated in (3). The flow distribution at branching points is indirectly formulated in (4) by demanding that the amount of material arriving at vertex v must be equal to the amount of material leaving this vertex at every timestep. In order to avoid unnecessary queuing costs, we want the processor to work at full capacity as soon as the corresponding queue is not empty. To avoid a discontinuous dependence of the inflow x_e^t from the queue size, a relaxation parameter $\epsilon \leq 1$ is introduced as proposed in [2]. Constraints (5)–(7) result from linearizing the equation $x_e^t = \min(\mu_{e,c}, \frac{q_e^t}{\epsilon})$ and generalizing it for all active and inactive configurations. The objective function of the design model consists of four weighted cost terms which are given by setup costs $p^{setup} := \sum_{e,c} \gamma_{e,c} \zeta_{e,c}^{setup}$, running costs $p^{run} := \sum_{t,e,c} \Delta t \cdot \zeta_{e,c}^{run} \gamma_{e,c} x_{e,c}^t$, queuing costs $p^{queue} := \sum_{t,e} \Delta t \cdot q_e^t$ and outflow benefit $p^{out} := \sum_t \Delta t \cdot y_{e_{out}}^t$, where e_{out} denotes the outflow edge, i.e. the edge without successor. For most of the constraints, the big-M formulation has been used for linearization, which is a very popular technique in discrete optimization issues. That means μ_e^{max}, B_e, M and D_e must be set to sufficiently large constants. Further details are described in [3].

3 Starting Heuristics

Due to the fact, that the branch and cut algorithm often takes a long time ($\gg 1\,\text{h}$) to find an incumbent, heuristical algorithms that find good feasible solutions within a short computation time are required. To generate a feasible solution, at most one configuration has to be selected for each edge

and the distribution rates at all branching points of the network have to be fixed. In this section several heuristic approaches are shortly raised.

3.1 Shortest Path Heuristic

The idea of the shortest path heuristic is to find the cheapest path through the system due to certain heuristical cost values, which typify estimates of production costs arising when the corresponding processor is active. The material flow is then exactly restricted to this path. For the calculation of the heuristical cost values, processor properties as well as an estimation of the flow amount possibly entering the processor are required. More details can be found in [3]. The algorithm works as follows.

1. Compute heuristical cost values of each processor and select the cheapest processor of all available configurations for each edge.
2. Use Dijkstra's algorithm to compute the cheapest path from inflow to outflow edge.
3. Set the distribution rates in a way that solely the cheapest path is used for the flow.

Since one single path is active in the heuristical solution, setup and running costs are quite low. Consequently, this heuristic gives quite good results, as long as the queuing cost weight of the objective function is small enough. Otherwise it would be more convenient to activate more edges for the evolution of flow, such that queues in front of low capacitated processors are reduced. For such instances an alternative heuristic is developed.

3.2 Flow Simulation Heuristic

The so-called flow simulation heuristic distributes flow amongst all edges of the network. The distribution rates, depending on the heuristical cost values of the corresponding processors, are set in such a way that cheaper edges are used more intensely than more expensive ones. The algorithm executes the edges of the whole graph in a specific order which allows for a more exact estimation of the arriving flow and therewith a more precise heuristical cost value. In short, the algorithm can be described as follows:

1. Choose the vertex whose incoming edges are already checked (start with the end vertex of the inflow edge).
2. Compute heuristical cost values of its outgoing edges using the flow estimation value of their predecessors.
3. Select the cheapest configuration for all outgoing edges.
4. Set rates depending on the heuristical cost values of the selected configurations for all outgoing edges.
5. Compute flow estimates of all outgoing edges using the flow estimates of the predecessors and the distribution rates.

Due to the distribution of flow amongst all edges of the network, the appearance of queues is reduced. However, since the algorithm lacks an anticipatory component concerning the computation of distribution rates, there can still be found instances, where unnecessary queues arise. For these cases an additional algorithm can be appended which changes the distribution rates of the heuristical solution leading to a stronger queue reduction.

3.3 Modified Flow Simulation Heuristic

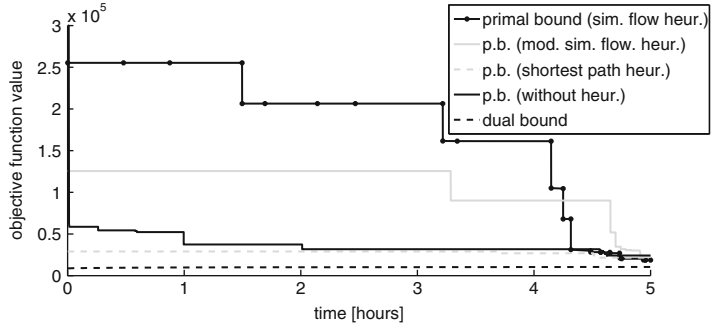
The modification mainly compares the flow estimates with the capacities of the corresponding processors and solves local linear optimization problems which reduce the queues by redistributing the flow in certain regions of the network where the emergence of queues is significant. Therefore the so-called accumulation value is computed for each vertex, which is nothing else than the difference between the sum of flow estimates traversing the vertex and the sum of processor capacities of all outgoing edges. A negative accumulation value means that queues can be avoided by redistributing the flow over all outgoing edges. At vertices with positive accumulation value, the flow already has to be redistributed at the predecessor nodes by simultaneously taking care that no new accumulation values arise at the neighbor vertices. Control parameters work as balances between the conservation of the original rates and the reduction of accumulation points.

4 Computational Results

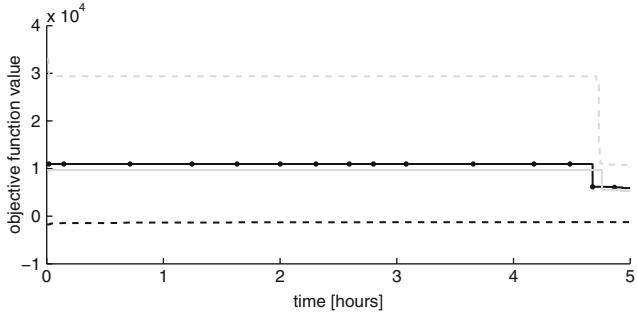
As already mentioned in the previous section the quality of heuristical solutions mainly depends on the choice of cost term weights. Figure 1 shows the typical behaviour of the optimization algorithm when different heuristics are prepended.

The two figures originate from a highly interconnected network with 90 edges and 5 configurations per edge. In Fig. 1(a) the emphasis of optimization is put on the setup and running costs. As supposed, the shortest path heuristic provides by far the best incumbent. In Fig. 1(b) the same instance is used, but this time the queuing cost weight and the outflow benefit dominate. In this case the flow simulation heuristics provides a better incumbent than the shortest path heuristic which can even be slightly improved by the distribution rate modification. This is an example, where the branch and cut algorithm is not able to find a feasible solution during the entire runtime of 5 h when no heuristic is used.

In general, even though the results concerning the optimization time and the quality of the incumbents by the use of different starting heuristics differ heavily from one instance to another, some observations can be pointed out for certain network constellations. Especially for instances with a vast number of configurations per edge it often takes the commercial optimization software



(a) The cost term weights of objective function are set to $\omega^{setup} = \omega^{run} = 100$ and $\omega^{queue} = \omega^{out} = 10$.



(b) $\omega^{setup} = \omega^{run} = 1$ and $\omega^{queue} = \omega^{out} = 100$

Fig. 1. Primal and dual bounds in the course of a 5 h capped optimization algorithm with different starting heuristics prepended

several hours to find the first feasible solution, whereas the heuristical algorithms take less than a second. Furthermore, the use of starting heuristics for instances with a fine timescale is recommendable, since rounding error often lead to the fact, that the optimization algorithm terminates without finding any feasible solution.

References

1. Armbruster, D., Degond, P., Ringhofer, C.: SIAM J. Appl. Math. **66**, 896–920 (2006)
2. Fügenschuh, A., Göttlich, S., Herty, M., Klar, A., Martin, A.: SIAM J. Sci. Comp. **30**, 1490–1507 (2008)
3. Göttlich, S., Dittel, A., Ziegler, U.: preprint (2009)
4. Göttlich, S., Herty, M.: Strategy and Tactics in Supply Chain Event Management, pp. 249–266. Springer, New York (2007)
5. ILOG CPLEX Division, 889 Alder Avenue, Suite 200, Incline Village, NV 89451, USA. URL http://www.cplex.com, accessed in 2008

Time-Dependent Order and Distribution Policies in Supply Networks

S. Göttlich[1], M. Herty[2], and Ch. Ringhofer[3]

[1] Department of Mathematics, TU Kaiserslautern, Postfach 3049, 67653 Kaiserslautern, Germany, goettlich@mathematik.uni-kl.de
[2] Department of Mathematics, RWTH Aachen University, Templergraben 55, 52056 Aachen, Germany, herty@mathc.rwth-aachen.de
[3] Department of Mathematics, Arizona State University, Tempe, AZ 85287-1804, USA, ringhofer@asu.edu

Summary. The dynamic of a production network is modeled by a coupled system of ordinary differential delay equations. Distribution and order policies are determined by an optimization problem for maximizing the profit of the production line.

1 Introduction

We consider a network of suppliers which order goods from each other, process a product according to orders, and receive payments according to a pricing policy. The dynamics of supply chains has been investigated in recent years (see cf. [1–3,5,7]) and extended to include money flows and bankruptcy, e.g. [2]. We extend existing results in the following ways: We consider general networks, represented by an arbitrarily connected graph. Each node in the network has a finite production or cycle time and a finite production capacity, as well as a front-end and back-end inventory. It is therefore possible that a supplier orders more than can be produced and stockpiles supplies. It is also possible that a supplier produces more than is ordered and stockpiles the output. Each supplier receives payments according to a dynamically determined pricing policy. Bankruptcy occurs if payments made exceed payments received beyond a certain available credit limit. Distribution and order policies are chosen in order to maximize the total profit. Mathematically, this problem is formulated as a mixed–integer programming problem.

2 The Model

The supply chain is modeled as a network of $S_1, \ldots, S_J$ nodes (suppliers) which order and deliver goods according to given (dynamic or static) policies,

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*, Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_82,

and receive payments according to certain pricing policies. We suppose that each supplier can choose his own policy. We furthermore assume that each supplier S_j has two inventories, a front-end (or input) inventory with an inventory position $p_j(t)$, and a back-end (or output) inventory with inventory position $q_j(t)$. Items are taken from the input inventory, processed by a given time τ_j, put into the output inventory and instantaneously delivered according to demand. Each supplier cannot take more than $\mu_j\, dt$ items in every infinitesimal time interval dt, i.e., the supplier has a maximal capacity of μ_j on the production process. There is no restriction on the inventories. This allows for storing overproduction and buffering for shortages in the supplies. Adjacent suppliers are directly connected by the rate of orders and flow, respectively. That means, supplier k orders products from supplier j at a rate Ω_{kj} and supplier k sends products to supplier j at a rate Φ_{jk}.

2.1 The Inventory Model

We turn to the mathematical equations describing the above model. We model the input inventory position p_j by a simple ordinary differential equation (ODE) and put a constraint such that the inventory cannot become negative.

$$\frac{dp_j}{dt} = \begin{pmatrix} f_j^{in} - \mu_j & \text{for } p_j > 0 \\ 0 & \text{for } p_j = 0 \end{pmatrix}. \tag{1}$$

The ODE (1) has a discontinuous right hand side and is therefore not guaranteed to have a solution. We replace (1) by the smooth dynamics

$$\frac{dp_j}{dt} = \begin{pmatrix} f_j^{in} - \mu_j & \text{for } p_j > \varepsilon\mu_j \\ f_j^{in} - \frac{p_j}{\varepsilon} & \text{for } p_j \le \varepsilon\mu_j \end{pmatrix} = f_j^{in} - g_j \tag{2a}$$

where

$$g_j = \min\{\mu_j, \frac{p_j}{\varepsilon}\}, \quad 0 < \varepsilon << 1. \tag{2b}$$

It is easy to see that (2) exhibits, for small ε, asymptotically the same behavior as (1) and the differential equation (2) has a Lipschitz continuous right hand side, and has therefore a well defined solution.

The evolution of the output inventory position q_j is modeled in the same way, where the influx into the output inventory is the time delayed outflux of the input inventory, i.e. $f_j^{in}(t) \to g_j(t-\tau_j)$ holds, where τ_j is the time it takes the supplier to process an item. The capacity μ_j is replaced by the demand w_j, i.e. the supplier cannot ship at a rate greater than the current demand from the output inventory. Therefore, the output inventory position q_j will satisfy $\frac{dq_j}{dt} = g(t-\tau_j) - \min\{w_j, \frac{q_j}{\varepsilon}\}$. So, in summary, the dynamics of each supplier S_j, $j = 1 : J$, are given by

$$\begin{aligned} &(a)\ p_j'(t) = f_j^{in}(t) - g_j(t), \quad (b)\ g_j(t) = \min\{\mu_j, \frac{p_j}{\varepsilon}\}, \\ &(c)\ q_j'(t) = g_j(t-\tau_j) - f_j(t), \quad (d)\ f_j(t) = \min\{w_j, \frac{q_j}{\varepsilon}\}. \end{aligned} \tag{3}$$

For notational convenience we include external suppliers of raw materials and final customers as suppliers in the network by formally defining raw material suppliers as suppliers with an infinite output inventory and customers as suppliers with zero production capacity.

It remains to connect the dynamics of the individual suppliers through the fluxes f_j and f_j^{in} in (3). This is done by Kirchhoff's law. We define a $J \times J$ distribution matrix A with nonnegative entries A_{jk}, and set

$$f_j^{in} = \sum_{k=1}^{J} A_{jk} f_k, \quad j = 1 : J. \tag{4}$$

So the product flux Φ_{jk} from S_k to S_j is given by $\Phi_{jk} = A_{jk} f_k$, and A_{jk} is the percentage of the output of supplier k sent to supplier j. We denote the column sums of the matrix A by a and by a_k the percentage of product shipped back into the system by supplier S_k; hence $a_k = \sum_{j=1}^{J} A_{jk}$, $a^T = \mathbf{1}^T A$. If there is no loss of product during shipping (which assumed to be instantaneous), then the column sums a_k will equal unity, except for those nodes S_k corresponding to final customers.

The demand w_j on supplier S_j in (3)(d) is given by the orders placed. The modeling is analogously as for the distribution rates. Assuming that the supplier S_k places orders to supplier S_j at a rate Ω_{kj}, i.e.,

$$d_j = \sum_{k=1}^{J} \Omega_{jk}, \ j = 1 : J, \quad w = \Omega \mathbf{1}. \tag{5}$$

The matrix Ω defines the topology of the network, since each supplier will only place orders to a limited number of other suppliers. The elements of the matrices A and Ω are the distribution and order policy decision variables. However, the distribution policy cannot be chosen independently of the order policy: Supplier S_k cannot ship more from the output inventory than the total demand w_k (as enforced by the form of the outflux function f in (3)(d)). He also cannot ship to the individual supplier S_j at a rate greater than S_j is ordering. That is, we have the additional constraints

$$\Phi_{jk} = A_{jk} f_k \leq \Omega_{kj}, \quad j, k = 1 : J. \tag{6}$$

The constraint (6) is a dynamic and nonlinear constraint, since all terms in (6) depend on time and the fluxes f_k are given by (3). The constraint can be satisfied by choosing an a priori rule for the distribution policy (which is not dependent on the dynamics): $A_{jk} = \frac{1}{w_k} \Omega_{kj}$, This reduces (6) to $f_k \leq w_k$ which is enforced by (3) automatically. This choice can be understood as follows.

1. If the output inventory is non-empty ($q_k = O(1)$, $f_k = d_k$): Satisfy all demands $\Phi_{jk} = \Omega_{kj}$.
2. If the output inventory is empty $q_k < \varepsilon d_k, f_k = \frac{q_k}{\varepsilon} \approx g_k(t - \tau_k) < d_k$: Distribute the output proportionally to the demand. Set $\Phi_{jk} = A_{jk} f_k = \frac{f_k}{d_k} \Omega_{kj}$.

Given a certain order policy, any other distribution policy will have to use information about the current state of the system (the values of $f_k(t)$ and $d_k(t)$) to enforce the constraint (6).

Remark 1. Up to this point we assumed that, if supplier S_j cannot satisfy all orders at any point in time, these orders are lost. A more realistic model allows for the orders to be filled at a later point in time (incurring a penalty which has to be included in a cost functional for optimization). We remove the constraint that the output inventory position $q_j(t)$ in (3) remains non-negative, and define the inventory position as q_j for positive q_j and the backlog as $-q_j$ for negative q_j. In this case (3)(c)(d) have to be replaced by

$$q_j'(t) = g_j(t - \tau_j) - f_j(t), \quad f_j = H(q_j)w_j$$

for H being the Heaviside function.

2.2 The Money Flow

One of the purposes of the model developed so far is the study of the evolution of bankruptcies in a given network. To this end, it is necessary to include cash flow into the model. Money flows run in the opposite direction of product in the network, and are weighted by a price. We denote by r_j the price per product unit supplier S_j charges to deliver. Therefore, the flow of money ψ_{jk} from supplier S_k to supplier S_j is given by $\psi_{jk} = \Phi_{kj} r_j$. Furthermore, we assume that each supplier S_j has certain production costs for delivering the product. The production costs per product unit supplier S_j are denoted by $\mathbf{r}_j$. Hence, the money supply σ_j of supplier S_j evolves according to

$$\sigma_j'(t) = \sum_{k=1}^{J} \psi_{jk} - \psi_{kj} - \sum_{k=1}^{J} \mathbf{r}_k f_k = a_j f_j r_j - \sum_{k=1}^{J} A_{jk} f_k r_k - \sum_{k=1}^{J} \mathbf{r}_k f_k. \quad (7)$$

If we assume that no credit is extended, i.e. the supplier S_j has to stop ordering once its money supply is exhausted, we obtain the condition

$$\Omega_{kj} = 0 \text{ if } \sigma_j = 0. \quad (8)$$

The above model allows for the simulation of the flow of product and payments on arbitrary complex networks, given a certain order and pricing policy, i.e. once the order rates $\Omega_{kj}(t)$ and the prices $r_j(t)$ are chosen. Choosing optimal policies by solving a constrained optimization problem is subject of Sect. 3.

3 Computing Optimal Order and Distribution Policies

We are interested in an optimal choice of the given order policies Ω_{kj} and distribution policies A_{jk}. We consider the choice to be optimal, if the profit of the supply chain $\sum_{j=1}^{J} \sigma_j(T)$ at some final time T is maximal. Constraints

to this optimization problem are the dynamics of the supply chain as well as constraints on order flows, production capacities, inventories and possible bankruptcy.

This is described by the following maximization problem for Ω_{kj} and A_{jk} :

$$\max_{\Omega_{kj}, A_{jk}} \sum_{j=1}^{J} \sigma_j(T) \text{ subject to (3), (4), (5), (6), (7), (8).} \tag{9}$$

Numerically, this problem is solved using a mixed–integer programming formulation as derived e.g. in [4,6]. We now study some analytical properties of the maximization problem.

Lemma 1. *As long as there is no bankruptcy, the internal pricing has no influence on the order Ω_{kj} and distribution policy A_{jk}.*

In fact, due to (4), we have $\sum_j A_{jk} = 1$ and hence $A_{jk} f_k = 0 \ \forall k$, for customers j, and $A_{jk} f_j = 0 \ \forall j$, for raw material suppliers k, respectively. We introduce the index sets $\mathcal{A}_C, \mathcal{A}_M$ and $\mathcal{A}_S$ for customers, raw material suppliers and the remaining suppliers. Then, we obtain

$$\sum_{j,k} A_{kj} f_j r_j - A_{jk} f_k r_k = \sum_{j \in \mathcal{A}_C} f_j r_j - \sum_{j \in \mathcal{A}_M} f_j r_j.$$

For an initial money supply of $\sigma_j(t=0) = 0$ we obtain by integrating (7)

$$\sum_j \sigma_j(T) = \int_0^T \left(\sum_j -\mathbf{r}_j f_j + \sum_{j \in \mathcal{A}_C} f_j r_j - \sum_{j \in \mathcal{A}_M} f_j r_j \right) dt. \tag{10}$$

Hence, only the production costs but *not* the internal pricing effect the cost functional (as long as there is no bankruptcy).

Second, we reformulate (3) in order to obtain a partial differential equation. This approach is analogously to the presentation in [6]: We assume each supplier as unit length. The delay in τ_j is then modeled by a function $\rho(x,t)$ satisfying

$$\begin{aligned} p'(t) = f^{in} - \rho(0,t), \quad q'(t) = \rho(1,t) - f(t), \\ \rho_t + \frac{1}{\tau} g_x = 0, \quad \rho(0,t) = \min\{\mu, \frac{p}{\varepsilon}\}, \quad f = \min\{w, \frac{q}{\varepsilon}\}, \\ g(t) = \rho(0,t). \end{aligned}$$

The latter set of equations is in fact an Upwind discretization of a partial differential equation for the part densities ρ given by

$$\partial_t \rho + \partial_x \min\{\nu, v\rho\} = 0, \quad g(0,t) = f^{in} \tag{11}$$

and where $\nu(x) = \mu\ \chi_{x\leq\frac{1}{2}} + w\ \chi_{x\geq\frac{1}{2}}$, $v = \frac{1}{\tau}$ and $p = \epsilon\rho(0-,t)$ and $q = \epsilon\rho(1+,t)$. This equation has to be solved for ρ_j, i.e., the part density for supplier S_j.

Third, if we denote by $\psi_j := \min\{\nu_j, v_j\rho_j\}$, we obtain $f_j = \psi_j(1,t)$ and $f_j^{in} = \psi_j(0,t)$. Assuming an initially empty supply chain $\rho_j(x,0) = 0$ for all suppliers j, and equal prices per product $r_j \equiv r$, we obtain from (11) upon integrating

$$\begin{aligned}
\sum_{j=1}^{J}\sigma_j(T) &= -\int_0^T\sum_{j=1}^{J}\mathbf{r}_j f_j dt + \int_0^T\sum_{j\in\mathcal{A}_C} f_j r - \sum_{j\in\mathcal{A}_M} f_j r dt \\
&= -\int_0^T\sum_{j=1}^{J}\mathbf{r}_j\psi_j(1,t)dt + \int_0^T r\left(\sum_{j\in\mathcal{A}_C}\psi_j(1,t) - \sum_{j\in\mathcal{A}_M}\psi_j(1,t)\right)dt \\
&= -\int_0^T\sum_{j=1}^{J}\mathbf{r}_j\psi_j(1,t)dt + \sum_{j=1}^{J}\int_0^T r\left(\psi_j(0,t) - \psi_j(1,t)\right)dt \\
&= -\int_0^T\sum_{j=1}^{J}\mathbf{r}_j\psi_j(1,t)dt - r\sum_{j=1}^{J}\int_0^1\rho_j(x,T)dx.
\end{aligned}$$

Summarizing, we proved:

Lemma 2. *If $\mathbf{r}_j \equiv 0$, then maximizing the costs at time T using cost functional (10), is equivalent to minimizing the load in the complete network at time $t = T$ where the load is the number of parts $\int \rho_j\,dx$ in supplier S_j.*

Acknowledgements

This work was supported by NSF grant DMS-0604986, the DAAD project D/06/28176 and the DFG grant HE 5386/6-1.

References

1. Armbruster, D., Degond, P., Ringhofer, Ch.: SIAM J. Appl. Math. **66**, 896–920 (2006)
2. Battiston, S., Delli Gatti, D., Gallegati, M., Greenwald, B., Stiglitz, J.E.: J. Eco. Dyn. Control **31**, 2061–2084 (2007)
3. Daganzo, C.: A Theory of Supply Chains. Springer, Berlin (2003)
4. Fügenschuh, A., Göttlich, S., Herty, M., Klar, A., Martin, A.: SIAM J. Sci. Comp. **30**, 1490–1507 (2008)
5. Göttlich, S., Herty, M., Klar, A.: Comm. Math. Sci. **3**, 545–559 (2005)
6. Herty, M., Ringhofer, Ch. Physica A **380**, 651–664 (2007)
7. Helbing, D., Armbruster, D., Mikhailov, A., Lefeber, E.L.: Physica A **363**, 1–60 (2006)

Dynamics of Supply Chains Under Mixed Production Strategies

R. Donner[1], K. Padberg[1,2], J. Höfener[3], and D. Helbing[4]

[1] Institute for Transport and Economics, TU Dresden, Germany, donner@vwi.tu-dresden.de, kathrin.padberg@tu-dresden.de
[2] Center for Information Services and High Performance Computing, TU Dresden, Germany
[3] Max Planck Institute for the Physics of Complex Systems, Dresden, Germany, hoefener@pks.mpg.de
[4] Department of Humanities and Social Sciences, ETH Zurich, Switzerland, dhelbing@ethz.ch

Summary. This contribution focusses on the dynamics of material flows in supply chains under pull, push and mixed production strategies. For this purpose, a mathematical input–output model of commodity flows is generalised and analysed in some detail for the case of linear supply chains. In particular, it is investigated under which conditions the effect of instabilities like the Bullwhip effect can be minimised. The presented results allow some new insight into the dynamics of manufacturing systems, which will be of importance for the development of new approaches for production planning and control.

1 Introduction

Logistics is one of the fastest growing economic sectors today. In connection with this development, the optimisation of interacting production and transportation processes has become a problem of increasing relevance, as it offers the possibility to gain efficiency and reduce costs by a sophisticated planning and control of the corresponding networks. During the last years, increasing efforts have been made to mathematically model and analyse the dynamics of supply chains and similar logistic systems [1, 5]. Apart from external factors like demand and supply variations, it has been revealed that the particular strategies for producing and ordering goods are of major importance for the stability of commodity flows between manufacturers organised in a supply network. In this work, we study analytically how purely demand- or supply-driven production strategies lead to an amplification of initial variations along a supply chain, a phenomenon known as the Bullwhip effect [3]. A combination of traditional push and pull mechanisms is suggested for minimising this effect. Based on a simple mathematical input–output model,

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*, Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_83,

we use a graph-theoretical framework for studying the dynamics of supply chains under mixed push-pull strategies and briefly describe the emergence of additional instabilities in case of an improper choice of strategic parameters.

2 Description of the Model

In order to describe the flow of commodities in a supply chain, we adapt a recently introduced fluid-dynamic input–output model [6]. For convenience, we make a number of assumptions simplifying the mathematical formulation and analysis: (a) We neglect the influence of price variability on production and order strategies of firms, which may however have considerable effects in real-world systems. (b) Every manufacturer is able to produce only one specific kind of product, which may be either sold on an external market or supplied as a commodity to other manufacturers. (c) For every commodity j, there is exactly one producer which also gets the index j.

2.1 Pull and Push Strategies

In their original work, Helbing et al. [6] formulated the input–output model only for a pull strategy, i.e. a production that is exclusively determined by the demand of the respective customers. This relationship is described by the production rates $Q_j(t)$ and the available stocks $N_j^{out}(t)$ of the corresponding finished goods. The total demand for the associated product j is given by the sum of the external market demand $D_j(t)$ and the demand of other manufacturers k. These work with production rates $Q_k(t)$ and need a certain relative fraction $C_{jk} \leq 1$ of the commodity produced by manufacturer j for their own production. Therefore, $\mathbf{C} = (C_{jk})$ is referred to as the input–output matrix of commodities. The change of the available stocks of a product j is then given by

$$\frac{dN_j^{out}}{dt} = Q_j(t) - \sum_k C_{jk} Q_k(t) - D_j(t). \tag{1}$$

In contrast to a pull strategy, with a push strategy, the production rate is exclusively determined by the supply of commodities. Hence, the quantities of interest for the individual manufacturers are the locally available stocks of the commodities $N_j^{in}(t)$. Given the total supply being the sum of a factory-specific external supply $S_j(t)$ and the supply of other members of the network, the changes of the corresponding inventories read

$$\frac{dN_j^{in}}{dt} = S_j(t) + \sum_k T_{jk} Q_k(t) - Q_j(t), \tag{2}$$

where the matrix $\mathbf{T} = (T_{jk})$ is called production matrix and determines the fraction of different products j that are produced by a manufacturer k. In the

case of more complicated production strategies, the separate consideration of in- and output buffers may be necessary, whereas both types of buffers can be usually identified with each other in a linear supply chain configuration.

2.2 Adaptation of Production Rates

The factors that determine the adjustment of the production rates are specific for every production strategy. A general observation is that manufacturers typically want to avoid strong fluctuations in the available commodities as well as in the finished goods storages. Moreover, there are desired levels $\hat{N}_j^{in,out}$ for all inventories (incoming and outgoing material) as well as production rates $\hat{Q}_j$ corresponding to an optimal use of machine capacity. A general production strategy $Q_j(t)$ realising the successive adaptation of production rates to changing economic conditions can then be formulated as

$$\frac{1}{Q_j(t)}\frac{dQ_j}{dt} = \hat{\nu}_j^{out}\left(\frac{\hat{N}_j^{out}}{N_j^{out}(t)} - 1\right) - \hat{\mu}_j^{out}\frac{1}{N_j^{out}(t)}\frac{dN_j^{out}}{dt} - \hat{\nu}_j^{in}\left(\frac{\hat{N}_j^{in}}{N_j^{in}(t)} - 1\right) + \hat{\mu}_j^{in}\frac{1}{N_j^{in}(t)}\frac{dN_j^{in}}{dt} + \alpha_j\left(\frac{\hat{Q}_j}{Q_j(t)} - 1\right), \quad (3)$$

with adaptation rates $\hat{\nu}_j^{in,out}, \hat{\mu}_j^{in,out}, \alpha_j \geq 0$. For example, $\hat{\nu}_j^{in} = \hat{\mu}_j^{in} = 0$ corresponds to a pull strategy, while $\hat{\nu}_j^{out} = \hat{\mu}_j^{out} = 0$ is characteristic for a system subject to a push principle.

3 Linear Stability of Supply Chains

A basic production unit is composed of a production line and its associated in- and output buffers, which are filled and cleared by the different suppliers and customers. In order to study the dynamics of a linear supply chain, i.e. a sequential alignment of $n \geq 1$ production units where each commodity is produced and consumed by only one manufacturer, we may identify the supply and demand terms for every unit with the production rates of the previous and subsequent units, respectively. This means that the production matrix $\mathbf{T}$ (input–output matrix $\mathbf{C}$) has non-zero entries only on the first lower (upper) subdiagonal. Moreover, the output buffer of producer j can be identified with the input buffer of the production unit $j+1$, i.e. $N_j(t) := N_j^{out}(t) = N_{j+1}^{in}(t)$, and we set $Q_0(t) := S(t)$ and $Q_{n+1}(t) := D(t)$ for the external demand and supply. Under the assumption of material conservation, we may linearise all quantities about their optimal values ($X(t) = \hat{X} + x(t)$) yielding

$$\frac{dn_j}{dt} = q_j(t) - q_{j+1}(t), \quad (4)$$

$$\frac{dq_j}{dt} = -\mu_j^{out}\frac{dn_j}{dt} - \nu_j^{out}n_j(t) + \mu_j^{in}\frac{dn_{j-1}}{dt} + \nu_j^{in}n_{j-1}(t) - \alpha_j q_j(t). \quad (5)$$

These equations describe the dynamics of $n+1$ buffers and n producers, where the strategic parameters reduce to

$$\nu_j^{in,out} := \hat{Q}_j \hat{\nu}_j^{in,out} / \hat{N}_j^{in,out}, \quad \mu_j^{in,out} := \hat{Q}_j \hat{\mu}_j^{in,out} / \hat{N}_j^{in,out}.$$

In particular, deriving the equations for q_j with respect to t, we obtain the differential equation of a forced and damped harmonic oscillator for the dynamics of the residual production rate of each producer j:

$$\frac{d^2 q_j}{dt^2} + 2\gamma_j \frac{dq_j}{dt} + \omega_j^2 q_j(t) = f_j(t) \tag{6}$$

with

$$\gamma_j = \frac{1}{2}\left(\alpha + \mu_j^{in} + \mu_j^{out}\right), \quad \omega_j^2 = \nu_j^{in} + \nu_j^{out} \tag{7}$$

and

$$f_j(t) = \mu_j^{in} \frac{q_{j-1}}{dt} + \mu_j^{out} \frac{q_{j+1}}{dt} + \nu_j^{in} q_{j-1}(t) + \nu_j^{out} q_{j+1}(t). \tag{8}$$

This relationship clearly demonstrates that in a linear approximation, pull and push strategies lead to essentially the same kind of dynamics. In particular, in the case of a single autonomous production unit with constant external forcing (i.e. $q_0(t) = q_2(t) \equiv 0$), the production is generally stable for push, pull as well as mixed strategies, as the associated eigenvalues of the damped oscillator have negative real parts due to the positivity of γ_1.

3.1 Linear Stability for General Production Strategies

For the linear stability analysis of the supply chain we use a graph-theoretic ansatz. Let $\mathbf{J}$ be the Jacobian matrix of the system consisting of n production units as introduced above. The supply chain is stable *iff* all eigenvalues of $\mathbf{J}$, i.e. the roots of the characteristic polynomial $P(\lambda) = \lambda^m + a_1\lambda^{m-1} + a_2\lambda^{m-2} + \ldots + a_m$, $m = 2n + 1$, have negative real part. A necessary condition for $P(\lambda)$ to have exclusively stable roots is that the coefficients $a_i > 0$ for all $i = 1, \ldots, m$ [4]. $\mathbf{J}$ has a graph-theoretic interpretation (a weighted directed graph, [7]) that corresponds to the linearised dynamics of the supply chain, see Fig. 1. Based on the results in [7], it is possible to derive expressions that relate the coefficients a_i to cycles (or feedback loops, i.e. paths connecting a node with itself) in the digraph (see [8] for an application to biochemical networks). A feedback loop is positive if the product of edge weights belonging to the cycle is positive, else negative. Positive feedback loops may destabilise the system as they can contribute negative terms when computing the coefficients. For a single production unit the characteristic polynomial reads

$$P(\lambda) = \lambda(\lambda^2 + (\alpha_1 + \mu_1^{in} + \mu_1^{out})\lambda + (\nu_1^{in} + \nu_1^{out}))$$

where the term $(\alpha_1 + \mu_1^{in} + \mu_1^{out})$ is related to the negative 1-cycle and $(\nu_1^{in} + \nu_1^{out})$ to the two negative 2-cycles. Note that for structural reasons,

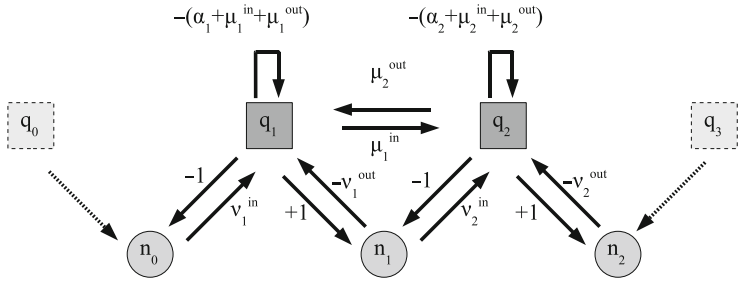


Fig. 1. Network and linearised dynamics of a linear supply chain with two producers

all characteristic polynomials of our supply chains derived in this way have a zero root, which however does not influence the stability properties of the system. For pure pull systems the concatenation of production units in a linear supply chain does not produce any additional cycles in the underlying graph. One can show that the characteristic polynomials are given by

$$P(\lambda) = \lambda \cdot p_1(\lambda) \cdot p_2(\lambda) \cdots p_n(\lambda)$$

where $p_i(\lambda) = \lambda^2 + (\alpha_i + \mu_i^{out})\lambda + \nu_i^{out}$. Equivalently, for a linear chain of push systems we obtain $p_i(\lambda) = \lambda^2 + (\alpha_i + \mu_i^{in})\lambda + \nu_i^{in}$, respectively. As the positivity of the coefficients of quadratic polynomials is sufficient for the roots to have negative real parts, a pure-strategy system is linearly stable. Mixed strategies may however cause instabilities resulting from positive feedback loops. In the case of two production units with the same mixed strategy, the contributions of positive feedback loops cancel out and we obtain $P(\lambda) = \lambda \cdot p(\lambda)$, where $p(\lambda)$ is a polynomial of degree four with positive coefficients. As this is not sufficient for its roots to lie in the left complex half-plane, the Routh–Hurwitz criterion [4] may be used to derive conditions on the parameters to guarantee stability. In fact, it turns out that in a certain parameter range, the system may undergo a Hopf bifurcation [2].

3.2 The Bullwhip Effect

Although pure push or pull systems are linearly stable we can often observe an undesired amplification of the production rates along the supply chain. For the following considerations, let us assume that the production process of the j-th unit in the chain is weakly periodic as $q_j(t) = q_j^0 \cos(\beta t + \theta_j)$. In the case of a pull strategy, only demand terms matter in (8). Evaluating these terms, one may easily see that the production rate $q_{j-1}(t)$ of the previous unit then varies as

$$q_{j-1}(t) = q_{j-1}^1 e^{-\gamma_{j-1} t} \cos(\sqrt{\omega_{j-1}^2 - \gamma_{j-1}^2}\, t + \psi_{j-1}) + q_{j-1}^0 \cos(\beta t + \theta_{j-1}) \quad (9)$$

where q^1_{j-1}, ψ_{j-1} and θ_{j-1} depend on the initial conditions and

$$q^0_{j-1} = q^0_j \left\{ 1 + \frac{\beta^4 - 2\beta^2\omega^2_{j-1}}{\omega^4_{j-1} + 4\beta^2\gamma^2_{j-1}} \right\}^{-1/2} . \tag{10}$$

After a suitably long transient time $t \gg \gamma^{-1}_{j-1}$, the production rate is then completely determined by forced oscillations due to the periodic demand. Equation (10) describes an amplification of short-term variations with $\beta^2 < 2\omega^2_{j-1}$, with a maximum factor at

$$\beta^2_{max} = \frac{\omega^4_{j-1}}{4\gamma^2_{j-1}} \left(\sqrt{1 + \frac{8\gamma^2_{j-1}}{\omega^2_{j-1}}} - 1 \right) . \tag{11}$$

This convective instability can be identified as the Bullwhip effect that has been known in management science for about 50 years [3]. Due to the symmetry between supply and demand terms in (6), all above statements remain true under a push strategy if the previous unit $j-1$ is replaced by the subsequent one $j+1$. Hence, while for a pull strategy, initial demand variations propagate and amplify upstream, a push strategy causes an amplification of supply variations downstream with the material flow. In contrast to these pure strategies, it is possible to prove that under certain conditions, systems with mixed push-pull strategies may suppress the corresponding amplification significantly [2].

4 Conclusions and Outlook

In this work, we have summarised some fundamental results from an analytical study of a simple fluid-dynamic input–output model for supply networks. The model introduced in [6] has been extended to allow the incorporation of push, pull and mixed strategies. We have shown that a linear supply chain managed by a push or a pull principle is generally linearly stable. However, these pure-strategy systems can exhibit a convective instability known as the Bullwhip effect, which causes an undesired amplification of supply or demand variations along the chain. This amplification can be suppressed by allowing heterogeneous strategies, which however may induce linear instabilities under some circumstances. A detailed analytical and numerical treatment, where we consider general supply networks as well as an application to a real-world production system is in preparation [2].

Acknowledgements

This work has been partly supported by the Deutsche Forschungsgemeinschaft (DFG project no. He 2789/8-1), the EU project MMCOMNET (contract no. 12999), the Gottlieb Daimler- und Karl Benz Stiftung (project "BioLogistics"), and the VolkswagenStiftung (project no. I/82697).

References

1. Daganzo, C.F.: A Theory of Supply Chains. Springer, Berlin (2003)
2. Donner, R., Padberg, K., Höfener, J., Lämmer, S., Seidel, T., Helbing, D.: in prep.
3. Forrester, J.W.: Harv. Bus. Rev. **36**, 37–66 (1958)
4. Gantmacher, F.R.: Applications of the Theory of Matrices. Interscience, New York (1959)
5. Helbing, D.: New J. Phys. **5**, 90.1–90.28 (2003)
6. Helbing, D., Witt, U., Lämmer, S., Brenner, T.: Phys. Rev. E **70**, 056118 (2004)
7. Maybee, J.S., Olesky, D.D., van den Driessche, P., Wiener, G.: SIAM J. Matrix Anal. Appl. **10**, 500–519 (1989)
8. Mincheva, M., Roussel, M.R.: J. Math. Biol. **55**, 61–86 (2007)

Analogies Between Social Interaction Models and Supply Chains

Laurent Navoret[1], Richard Bon[2], Pierre Degond[3], Jacques Gautrais[4], David Sanchez[5], and Guy Theraulaz[6]

[1] Institut de Mathématiques de Toulouse, Université Paul Sabatier, 118 route de Narbonne, 31062 Toulouse Cedex 9 France, laurent.navoret@math.univ-toulouse.fr
[2] Centre de Recherche en Cognition Animale, Université Paul Sabatier, 118 route de Narbonne, 31062 Toulouse Cedex 9 France, rbon@cict.fr
[3] Institut de Mathématiques de Toulouse, pierre.degond@math.univ-toulouse.fr
[4] Centre de Recherche en Cognition Animale, gautrais@cict.fr
[5] Institut de Mathématiques de Toulouse, david.sanchez@math.univ-toulouse.fr
[6] Centre de Recherche en Cognition Animale, theraula@cict.fr

Congestion is a major issue in the modeling of both animal aggregation and supply chains. Indeed, both systems face obvious physical limits : capacities of machine in supply chains [1] and non-overlapping constraints between individuals in social groups. As a paradigm of crowding and social group movement, we are interested in sheep herds. The behaviour of this gregarious species is experimentally studied [5]. Let us focus on the displacement period of a sheep herd, where all animals move with the same speed.

A macroscopic model for herds including speed and congestion constraints is derived from an individual-based model. In order to enlight the congestion part in the dynamics, a singular limit of this macroscopic model is taken and leads to two phases in the herd : a congested and a non-congested one. We finally analyse the spatial transition between these two phases. Such a study of the congestion in self-organized systems could be translated to supply chains context.

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*, Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_84,

1 Derivation of a Macroscopic Model with Speed and Congestion Constraints

1.1 Microscopic Model

The following microscopic model aims to describe the interactions of N particles labeled by $k \in \{1, \dots, N\}$ and with position $\mathbf{X}_k \in \mathbb{R}^2$ with two main constraints.

The first constraint consists in supposing that all the particles have the same magnitude of velocity, here equal to 1. In first approximation, this assumption is satisfied by the sheep in a moving herd [5] (or schools of fish [4]). Thus the velocity of the k-th particle is given by $\boldsymbol{\omega}_k$, where $\boldsymbol{\omega}_k \in \mathbb{S}^1 = \left\{\boldsymbol{\omega} \in \mathbb{R}^2, |\boldsymbol{\omega}| = 1\right\}$ is an unitary vector. Therefore the time derivative of $\boldsymbol{\omega}_k$ is orthogonal to $\boldsymbol{\omega}_k$. The second constraint is the congestion one. The particles are supposed to have a finite volume (equal to πd^2) and can not overlap, hence the existence of a maximal density ϱ^*. So the repulsive interaction has to be singular so as to prevent the density from exceeding ϱ^*.

We propose here a simple continuous model for the evolution of positions and velocities via attractive-repulsive binary interactions :

$$\frac{\mathrm{d}\mathbf{X}_k}{\mathrm{d}t} = \boldsymbol{\omega}_k, \tag{1}$$

$$\frac{\mathrm{d}\boldsymbol{\omega}_k}{\mathrm{d}t} = \nu_k^a(\mathrm{Id} - \boldsymbol{\omega}_k \otimes \boldsymbol{\omega}_k)\boldsymbol{\xi}_k^a - \nu_k^r(\mathrm{Id} - \boldsymbol{\omega}_k \otimes \boldsymbol{\omega}_k)\boldsymbol{\xi}_k^r\,, \tag{2}$$

where $\boldsymbol{\xi}_k^{a,r}$ are the attractive-repulsive forces, $\nu_{a,r}$ their respective interaction frequencies. The matrix $(\mathrm{Id} - \boldsymbol{\omega}_k \otimes \boldsymbol{\omega}_k)$ is the orthogonal projector on the orthogonal direction to $\boldsymbol{\omega}_k$ and enables to satisfy the speed constraint. The attractive force is chosen to drive the particles to the centre of mass inside an interaction disc of radius R_a, while the repulsive force is chosen to drive them to the opposite direction of the centre of mass inside an interaction disc of radius R_r (lower than R_a):

$$\boldsymbol{\xi}_k^a = \frac{\displaystyle\sum_{j,|\mathbf{X}_j - \mathbf{X}_k| \le R_a} \mathbf{X}_j - \mathbf{X}_k}{\displaystyle\sum_{j,|\mathbf{X}_j - \mathbf{X}_k| \le R_a} 1}\,, \quad \boldsymbol{\xi}_k^r = \frac{\displaystyle\sum_{j,|\mathbf{X}_j - \mathbf{X}_k| \le R_r} \mathbf{X}_j - \mathbf{X}_k}{\displaystyle\sum_{j,|\mathbf{X}_j - \mathbf{X}_k| \le R_r} 1}\,. \tag{3}$$

The attractive interaction is a constant $\nu_k^a = \nu_a$ and ν_r is taken so as to satisfy the congestion constraint:

$$\nu_k^r = \nu_r \left(\frac{\pi d^2 \sum_{j,|\mathbf{X}_j - \mathbf{X}_k| \le R_r} 1}{\pi R_r^2} \right)\,, \quad \nu_r(\varrho) = \varrho p'(\varrho)\,, \quad p(\varrho) = \left(\frac{1}{\varrho^*} - \frac{1}{\varrho} \right)^{-k}. \tag{4}$$

Note that $p(\varrho)$ tends to $+\infty$ when ϱ goes to ϱ^*. The form of the function ν_r is explicitly given only for the convenience of the following study.

1.2 Kinetic Model, Hydrodynamic Rescaling and Macroscopic Model

Mean-field Limit: $N \to +\infty$

To describe the dynamics of a large number of particles, it is usual in mathematical physics to introduce a distribution function $f(\mathbf{x}, \mathbf{v}, t)$ defined on the phase space: $f(\mathbf{x}, \mathbf{v}, t)\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{v}$ is the number of particles in the volume $[\mathbf{x}, \mathbf{x} + \mathrm{d}\mathbf{x}] \times [\mathbf{v}, \mathbf{v} + \mathrm{d}\mathbf{v}]$. From the equations satisfied by the empirical distribution $f^N(\mathbf{x}, \boldsymbol{\omega}, t) = \frac{1}{N}\sum_{k=1}^{N} \delta(\mathbf{x} - \mathbf{X}_k(t))\delta(\boldsymbol{\omega}, \boldsymbol{\omega}_k(t))$, we can formally derive the limit equation satisfied by $f = \lim f^N$ as the number of particles tends to $+\infty$:

$$
\begin{aligned}
&\partial_t f + \boldsymbol{\omega} \cdot \nabla_{\mathbf{x}} f + \nabla_{\boldsymbol{\omega}} \cdot ((\mathbf{F}_a - \mathbf{F}_r)\, f) = 0\,, \\
&\mathbf{F}_{a,r}(x, \boldsymbol{\omega}, t) = \nu_{a,r}(\mathrm{Id} - \boldsymbol{\omega} \otimes \boldsymbol{\omega})\boldsymbol{\xi}_{a,r}, \\
&\boldsymbol{\xi}_{a,r}(x, \omega, t) = \frac{\int K_{a,r}(\mathbf{y} - \mathbf{x})(\mathbf{y} - \mathbf{x})\varrho(\mathbf{y}, t)d\mathbf{y}}{\int K_{a,r}(\mathbf{y} - \mathbf{x})\varrho(\mathbf{y}, t)d\mathbf{y}}\,, \\
&\nu_r = \nu_r\left(\frac{\int K_r(\mathbf{y} - \mathbf{x})\varrho(\mathbf{y}, t)d\mathbf{y}}{\alpha \int K_r(\mathbf{y} - \mathbf{x})d\mathbf{y}}\right)\,,
\end{aligned}
\tag{5}
$$

where $\varrho(\mathbf{x}, \mathbf{v}, t) = \int f(\mathbf{x}, \mathbf{v}, t)d\mathbf{v}$ is the density and $K_{a,r}$ are the characteristic functions of the discs of radius R_a and R_r.

Hydrodynamic Scaling

To determine now the large time and space dynamics, we perform an hydrodynamic scaling. Let us introduce the new time and space variables: $\widetilde{\mathbf{x}} = \eta\mathbf{x}$, $\widetilde{t} = \eta t$, with $\eta << 1$. With this rescaling, the repulsive terms become local: $\nu_r^\eta(\mathbf{x}) = \nu_r(\varrho(\mathbf{x})) + o(\eta)$, $\xi_r^\eta(\mathbf{x}, \boldsymbol{\omega}, t) = \frac{R_r^2}{4}\nabla_{\mathbf{x}}\varrho^\eta(\mathbf{x}, t)/\varrho^\eta(\mathbf{x}, t) + o(\eta)$. As regards the attractive term, we suppose that it remains non local as η tends to 0 and weaker than the repulsive force: the scaled attractive kernel K_a^η and the scaled interaction frequency ν_a^η satisfy $K_a^\eta(\mathbf{z}) = K_a(\eta\mathbf{z}), \nu_a^\eta = \eta^2\nu_a$. Under all these model assumptions, the limit distribution function f we obtain as η tends to 0 in the new variables satisfies the system

$$
\begin{aligned}
&\partial_t f + \boldsymbol{\omega} \cdot \nabla_{\mathbf{x}} f + \nabla_{\boldsymbol{\omega}} \cdot ((\mathbf{F}_a - \mathbf{F}_r)\, f) = 0\,, \\
&\mathbf{F}_a(x, \omega, t) = \nu_a(\mathrm{Id} - \boldsymbol{\omega} \otimes \boldsymbol{\omega})\boldsymbol{\xi}_a,\ \boldsymbol{\xi}_a(x, t) = \left(\frac{\int K_a\left(|\mathbf{y} - \mathbf{x}|\right)(\mathbf{y} - \mathbf{x})\varrho(\mathbf{y}, t)d\mathbf{y}}{\int K_a\left(|\mathbf{y} - \mathbf{x}|\right)\varrho(\mathbf{y}, t)d\mathbf{y}}\right), \\
&\mathbf{F}_r(\mathbf{x}, \boldsymbol{\omega}, t) = \frac{R_r^2}{4}(\mathrm{Id} - \boldsymbol{\omega} \otimes \boldsymbol{\omega})\nabla_{\mathbf{x}} p(\varrho(\mathbf{x}, t))\,.
\end{aligned}
$$

Macroscopic Model

The last step of our derivation of models is to find the equation satisfied by the two first moments of the distribution function f: the density $\varrho = \int f d\boldsymbol{\omega}$

and the momentum $\varrho\mathbf{\Omega} = \int f\boldsymbol{\omega}d\boldsymbol{\omega}$. Supposing that f is regular enough and tends quickly enough to zero at infinity, then it can be checked that ϱ and $\varrho\mathbf{\Omega}$ satisfy

$$\partial_t \varrho + \nabla_{\mathbf{x}} \cdot \varrho\mathbf{\Omega} = 0, \tag{6}$$

$$\partial_t \varrho\mathbf{\Omega} + \nabla_{\mathbf{x}} \cdot \left(\int f\boldsymbol{\omega} \otimes \boldsymbol{\omega} d\boldsymbol{\omega} \right) = \int (Id - \boldsymbol{\omega} \otimes \boldsymbol{\omega}) f d\boldsymbol{\omega} \left(\nu_a \boldsymbol{\xi}_a - \frac{R_r^2}{4} \nabla_{\mathbf{x}} p(\varrho) \right). \tag{7}$$

where $\boldsymbol{\xi}_a$ is always given by (5). Unfortunately this system is not closed. We have to make new assumptions to express in term of the two first moments the quantities where f still appears. Here we assume that f is a monokinetic distribution: $f(\mathbf{x}, \boldsymbol{\omega}, t) = \varrho(\mathbf{x}, t)\delta(\boldsymbol{\omega}, \mathbf{\Omega}(x, t))$, with $|\mathbf{\Omega}(x, t)| = 1$. Finally, we obtain the following macroscopic system

$$\partial_t \varrho + \nabla_{\mathbf{x}} \cdot \varrho\mathbf{\Omega} = 0 \,, \tag{8}$$

$$\partial_t \left(\varrho\mathbf{\Omega} \right) + \nabla_{\mathbf{x}} \cdot \left(\varrho\mathbf{\Omega} \otimes \mathbf{\Omega} \right) = \varrho(Id - \mathbf{\Omega} \otimes \mathbf{\Omega})(\nu_a \boldsymbol{\xi}_a - \frac{R_r^2}{4} \nabla_{\mathbf{x}} p(\varrho)) \,. \tag{9}$$

2 Study of the Dilute-Congested Transition

These macroscopic equations (8)–(9) combine the congestion constraint embodied by p and the speed constraint embodied by the projection operator $(Id - \mathbf{\Omega} \otimes \mathbf{\Omega})$. It leads to two difficulties: the singularity of the pressure p and the non-conservativity of the (9). The first point has already been tackled in a one dimensional traffic jam model [2]. The goal of the following study is the treatment of the conjunction of the two in a 2 dimensional case. Since attractive and repulsive forces operate at different scales, we consider thereafter the attraction term as a source term and focus on the case $\nu_a = 0$ and $R_r << 1$.

2.1 Asymptotic Model

So as to study the singularity of the pressure, the principle is to enhance it by changing p into εp, $\varepsilon << 1$ ($R_r^{\varepsilon 2} = \varepsilon R_r^2$). By this way, the pressure term becomes negligible unless the density is near the maximal one. Let us denote by $(\varrho^\varepsilon, \mathbf{\Omega}^\varepsilon)$ the solution of the ε-system

$$\partial_t \varrho^\varepsilon + \nabla_x \cdot \varrho^\varepsilon \mathbf{\Omega}^\varepsilon = 0 \,, \tag{10}$$

$$\partial_t \mathbf{\Omega}^\varepsilon + \mathbf{\Omega}^\varepsilon \cdot \nabla_{\mathbf{x}} \mathbf{\Omega}^\varepsilon + \frac{R_r^2}{4} (Id - \mathbf{\Omega} \otimes \mathbf{\Omega}) \varepsilon \nabla_{\mathbf{x}} p(\varrho^\varepsilon) = 0 \,, \tag{11}$$

If $(\varrho^\varepsilon, \mathbf{\Omega}^\varepsilon)$ is a sequence of solutions converging to a solution $(\varrho, \mathbf{\Omega})$ when ε tends to zero, then the limit $\bar{p}(x) = \lim_{\varepsilon \to 0} \varepsilon p(\varrho^\varepsilon(x, t))$ is equal to zero unless ϱ^ε tends to ϱ^*. We assume that $\bar{p}$ is always finite. Thus, two interacting phases

with different dynamics appear at the limit: the phase of maximal density $\varrho = \varrho^*$, called congested phase, and the phase of density lower than ϱ^*, called the dilute phase. The limit $(\varrho, \mathbf{\Omega})$ fulfills the system

$$\partial_t \varrho + \nabla_{\mathbf{x}} \cdot \varrho \mathbf{\Omega} = 0 \,, \tag{12}$$

$$\partial_t \mathbf{\Omega} + \mathbf{\Omega} \cdot \nabla_{\mathbf{x}} \mathbf{\Omega} + \frac{R_r^2}{4} (Id - \mathbf{\Omega} \otimes \mathbf{\Omega}) \nabla_{\mathbf{x}} \bar{p} = 0 \,, \tag{13}$$

$$(\varrho^* - \varrho)\bar{p} = 0 \,, \tag{14}$$

where the last equality expresses the dichotomy $\varrho = \varrho^*$ or $\bar{p} = 0$.

In the dilute phase, where the density is lower than ϱ^*, we get a pressureless gaz dynamic model. Let us now investigate the system in the congested phase.

2.2 In the Congested Phase

In the congested phase $\varrho = \varrho^*$, the limit of (10)–(11) leads to an incompressible Euler system with speed constraint

$$\varrho = \varrho^*, \quad \nabla_{\mathbf{x}} \cdot \mathbf{\Omega} = 0 \,, \tag{15}$$

$$\partial_t \mathbf{\Omega} + \mathbf{\Omega} \cdot \nabla_{\mathbf{x}} \mathbf{\Omega} + \frac{R_r^2}{4} (Id - \mathbf{\Omega} \otimes \mathbf{\Omega}) \nabla_{\mathbf{x}} \bar{p} = 0 \,, \tag{16}$$

where the pressure $\bar{p}$ is the Lagrange multiplier of the incompressibility constraint.

The only incompressibility constraint (15) coupled with the speed constraint provide us enlightening ideas of the structure of clusters. Indeed, we can prove that if the vector field $\mathbf{\Omega}$ on the sphere ($|\mathbf{\Omega}| = 1$) satisfies the incompressibility constraint (15), then $\mathbf{\Omega}$ is constant on straight lines and orthogonal to these lines. Concerning the pressure $\bar{p}$, it satisfies an elliptic equation on the congested domain (easily obtained by taking the divergence of the momentum equation (16)).

As a result of these two last remarks, the only knowledge of the velocity $\mathbf{\Omega}$ and the pressure $\bar{p}$ on the border of the congested domain would enable us to find out the whole solution inside the congested zone. So given the interface dynamics, the whole problem could be solved.

2.3 The Interface Dynamics

So as to study the interface dynamics, we consider that our problem at the interface reduces to a one dimensional problem in the normal direction to this interface. Let us focus on the Riemann problem: the initial condition is a discontinuity between two constant states on both sides of the interface. The strategy is here to come back to the finite ε-system (10)–(11) and to extract the limit solutions of the Riemann problem as ε tends to zero with a left or a right state converging to ϱ^*.

By introducing θ with respect to the x_1 axis and assuming that the problem is uniform with respect to the x_2 axis, the non-conservative system (10)–(11) can be put for $\theta \in]0, \pi[$ in the form of the following conservative one

$$\partial_t \varrho + \partial_{x_1}(\varrho \cos(\theta)) = 0 \,, \tag{17}$$

$$\partial_t \Psi(\cos(\theta)) + \partial_{x_1}\left(\phi(\cos(\theta)) + \frac{R_r^2}{4}\varepsilon p(\varrho)\right) = 0 \,, \tag{18}$$

where $\Psi(u) = \frac{1}{2}\log\left(\frac{1+u}{1-u}\right)$ and $\phi(u) = \log\left(\frac{1}{\sqrt{1-u^2}}\right)$. The new conserved variables are ϱ and $\Psi(\cos(\theta))$.

This system is strictly hyperbolic and its associated fields are in the limit $\varepsilon = 0$ genuinely nonlinear. Therefore, classical results [3] provide us the entropic solution of the Riemann problem. For a non-congested left state $(\varrho_\ell, \theta_\ell)$ and a congested right state $(\varrho^*, \theta_r, \bar{p}_r)$, the two main possibilities are given by:

- In case of separating velocities $\cos(\theta_\ell) < \cos(\theta_r)$, vacuum appears between two contact discontinuities and there is an instantaneous declustering (the pressure becomes zero inside the congested domain);
- In case of incoming velocities $\cos(\theta_\ell) < \cos(\theta_r)$, the limit solution consists of one shock and a pressure jump in the congested domain. The new pressure is computable since it is the solution of an explicit non-linear equation.

The detailed study provides us also the interface dynamics in other cases ($\cos(\theta_\ell) = \cos(\theta_r)$, $\varrho^\ell = \varrho^r = \varrho^*$, etc.). It will be displayed in future papers.

3 Conclusion

In this paper, new tools for congestion modeling have been presented in the context of sheep herds modeling. We hope that it could be usefully adapted to supply chains modeling. The study of the congested/non-congested transition will be the ground of further challenging simulations taking into account both constraints (constant speed and maximal density).

References

1. Ringhofer, C., Armbruster, D., Degond, P.: Continuum Models for Interacting Machines. World Scientific, Singapore (2005)
2. Delitala M., Rascle, M., Berthelin, F., Degond, P.: A model for the formation and the evolution of traffic jams. Arch. Rationa. Mech. Anal., **187**(2), 185–220 (2008)
3. LeVeque, R.J.: Numerical Methods for Conservation Laws. Birkhäuser, Basel (1992)
4. Motsch, S., Degond, P.: Continuum Limit of Self-Driven Particles with Orientation Interaction. preprint.
5. Pillot, M.H.: Etude et modélisation des déplacements collectifs spontanés chez le mouton mérinos d'Arles (ovis aries). Master thesis (2006)

Computing the Value of Transshipment Flexibility in Distribution Networks

M. Laumanns

Institute for Operations Research, ETH Zurich, 8092 Zurich, Switzerland,
laumanns@ifor.math.ethz.ch

Summary. In inventory/distribution systems, lateral stock transshipments might lead to cost savings by effectively sharing inventory on the same echelon level. An approach is presented to quantify the value of this additional flexibility by determining the corresponding optimal control policies, and the resulting cost, in supply networks with transshipments under uncertain demand via stochastic dynamic programming. The minimum guaranteed cost reduction of the discounted expected cost compared to the base case without transshipment is proposed as the value of this flexibility option.

1 Introduction

In supply chain optimization, one typically distinguishes between decisions regarding the design and decisions regarding the operation of a system. Supply chain design concerns structural aspects, such as the choice and configuration of components, for example the nodes and links in a network. Supply chain operation, on the other hand, relates to the processes that take place on a given structure. This involves repeated decisions and actions over time, mainly to control the flow of material and to deploy the existing resources. Since structural design decision have long ranging effects, they are taken with a correspondingly long planning horizon on a strategic planning level, while supply chain control belongs to the shorter, operational planning level.

For both supply chain design and control, a multitude of methods and models exist in the literature [5] that mainly address each topic in isolation. But despite their differences with respect to time horizon, type of decisions to be taken and existing models, both aspects are obviously interrelated. Any design decision places constraints on how a system can be operated, and the value of a particular design alternative depends on how well the resulting structure can be utilized in operation. Thus, it is desirable from a conceptual standpoint to treat design and control simultaneously.

One approach for a simultaneous treatment of supply chain design and operation would be to formulate a bi-level optimization problem, where the

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*,
Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_85,

different design alternatives are represented by the upper level variables and the control problem constitutes the lower level. Bi-level optimization problems are normally non-convex and non-differentiable and thus hard even for simple instances where both the upper and lower level problems are linear programs [4]. Even more difficulties can be expected when the lower level problem, as assumed here, is a stochastic control problem, hence an infinite-dimensional optimization problem.

As a first step towards addressing design and control decisions in supply chains simultaneously, we assume here a scenario where we only have a few design alternatives that can essentially be enumerated. The focus is therefore on evaluating each design alternative with respect to the operational costs they incur. This cost is given by operating the supply chain optimally under each setting, i.e., by applying an optimal control policy that is of course constrained by the chosen design.

We demonstrate the approach for valuating design options for the particular case of inventory transshipments in distribution networks. Inventory transshipments are lateral movements of goods on the same echelon level that can be used to effectively 'share' inventory that is spatially distributed in a supply chain [1, 2]. Inventory transshipments usually involve additional transportation and handling cost, but have the potential to reduce the total inventory level, and hence the bound capital, within the network. We show, on a simple example, how these potential savings arise and how they can be quantified.

2 Inventory Control Model

We assume that the inventory system can be described as a state-based discrete-time controlled dynamical system with uncertainty,

$$\mathbf{x}_{k+1} = \mathbf{A}\mathbf{x}_k + \mathbf{B}\mathbf{u}_k + \mathbf{C}\mathbf{d}_k \tag{1}$$

where $\mathbf{x}_k \in X \subseteq \mathbb{R}^n$ denotes the state of the system, $\mathbf{u}_k \in U(x_k) \subseteq \mathbb{R}^{n_u}$ the control input, $\mathbf{d}_k \in \mathcal{D} \subseteq \mathbb{R}^{n_d}$ the (uncertain) disturbances, and k is the time index. In this case, the state variables represent the inventory levels at the different nodes in the inventory system, and the control inputs describe the controllable material flow (inventory replenishment orders and inventory shipments). The disturbance is given by the external customer demand that must be served by certain designated nodes. The matrices $\mathbf{A} \in \mathbb{R}^{n\times n}, \mathbf{B} \in \mathbb{R}^{n\times n_u}$, and $\mathbf{C} \in \mathbb{R}^{n\times n_d}$ express the connectivity structure between the nodes.

Consider the example shown in Fig. 1. Here, the customer demand d_1 is served from inventory x_1 and demand d_2 is served from x_2. New material can be ordered via the control variables u_1 and u_2, each with a time delay of 1 time step, which requires the auxiliary state variables x_3 and x_4. The control variables u_3 and u_4 represent the (optional) inventory transshipments between the inventories x_1 and x_2. The dynamics with the transshipment option is given by

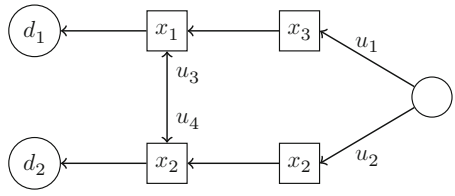


Fig. 1. Network structure of the inventory system

$$\mathbf{x}_{k+1} = \underbrace{\begin{bmatrix} 1&0&1&0 \\ 0&1&0&1 \\ 0&0&0&0 \\ 0&0&0&0 \end{bmatrix}}_{\mathbf{A}} \mathbf{x}_k + \underbrace{\begin{bmatrix} 0&0&1&-1 \\ 0&0&-1&1 \\ 1&0&0&0 \\ 0&1&0&0 \end{bmatrix}}_{\mathbf{B}} \mathbf{u}_k + \underbrace{\begin{bmatrix} -1&0 \\ 0&-1 \\ 0&0 \\ 0&0 \end{bmatrix}}_{\mathbf{C}} \mathbf{d}_k.$$

Without the transshipment option the control input is only two-dimensional (the components u_3 and u_4 are not present), and the dynamics is given by simply deleting the last two columns of $\mathbf{B}$. In the following we assume that the disturbance has finite support. For the example let $\lceil_k$ be drawn uniformly from $\mathcal{D} = \{(8,0)^T, (4,4)^T, (0,8)^T\}$.

The task is to solve the stochastic control problem optimally for both cases with and without the transshipment option. We assume linear inventory holding costs, which can be modeled as linear state costs $\mathbf{p}^T\mathbf{x}$, and linear transshipment and ordering costs $\mathbf{q}^T\mathbf{u}$. The objective is to minimize the average cost per time step. Thus we have to compute explicit state-feedback control policies, that is, control laws as functions of the current state of the system. In our setting it is sufficient to consider stationary policies [3]. Therefore let $\pi : \mathcal{X} \to \mathcal{U}$ be a stationary control policy and Π the set of all feasible stationary policies where $\pi(\mathbf{x}) \in U(x)$ for all $\mathbf{x} \in X$.

The objective is now to minimize the expected average cost per time, when starting from a given initial state x_0,

$$J_\pi(\mathbf{x}_0) := \lim_{K\to\infty} \frac{1}{K} E\left[\sum_{k=0}^{K-1} \mathbf{p}^T\mathbf{x}_k + \mathbf{q}^T\pi(\mathbf{u}_k)\right]$$

over the set of feasible policies Π, where the expectation is taken with respect to the random sequence (d_k) and $\mathbf{x}_{k+1} = \mathbf{A}\mathbf{x}_k + \mathbf{B}\pi(\mathbf{x}_k) + \mathbf{C}\mathbf{d}_k$.

The common approach for solving this stochastic infinite-time control problem is to use value iteration or policy iteration [3], usually on a discrete or discretized state space. Here, we slightly depart from this standard procedure and work with a continuous state space, by formulating the value iteration as a parametric optimization problem in the control variable vector $\mathbf{u}$ as

$$J_{k+1}(\mathbf{x}) = \min_{\mathbf{u}\in U(x)} \mathbf{p}^T\mathbf{x} + \mathbf{q}^T\mathbf{u} + E[J_k(\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u} + \mathbf{C}\mathbf{d})] \tag{2}$$

$$\text{s.t. } \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u} + \mathbf{C}\mathbf{d} \in X_k \ \forall \mathbf{d} \in \mathcal{D} \tag{3}$$

with initial values $J_0 \equiv 0$ and $X_0 = \mathcal{X}$. The sets X_k represent the sets of feasible states, i.e., the states where a feasible control action exists such that the next state is guaranteed to be feasible as well.

Assuming now that the control constraints $U(x)$ are polyhedral and can thus be specified as a system of linear inequalities

$$\mathbf{F}\mathbf{x} + \mathbf{G}\mathbf{u} \leq \mathbf{g},$$

it can be seen that the sets of feasible states X_k will remain polyhedral under the iteration and the value function will be a piecewise affine convex function. Consequently, the above problem will remain a parametric linear program in all iterations.

Using a solver for multi-parametric linear programs [6] we can now execute the value iteration (2) for the example until the sequence of value functions divided by the iteration counter, $\{J_k(x)/k\}$, converges to the limit cost $J^*(x)$. We can then extract the optimal policy π^* as the optimizer u^* in the last iteration, which will be a piecewise affine function of x, hence a state-feedback control policy.

For our example we assume linear state cost $\mathbf{p} = (1, 1, 0, 0)^T$, control cost $\mathbf{q} = (0, 0, 0, 0)^T$ and constraints

$$\begin{aligned} 0 \leq x_i \leq 40, i \in \{1, 2, 3, 4\}, \\ 0 \leq u_i \leq 8, i \in \{1, 2, 3, 4\}, \\ u_3 \leq x_2, \ \ u_4 \leq x_1. \end{aligned}$$

The resulting optimal policy with transshipment option is

$$\pi^*_{\mathrm{T}}(\mathbf{x}) = \begin{pmatrix} \max\{12 - (x_1 + x_2 + x_3 + x_4)/2, 0\} \\ \max\{12 - (x_1 + x_2 + x_3 + x_4)/2, 0\} \\ \max\{8 - x_1 - x_3, 0\} \\ \max\{8 - x_2 - x_4, 0\} \end{pmatrix}.$$

Without the transshipment option, the resulting optimal policy is

$$\pi^*_{\mathrm{N}}(\mathbf{x}) = \begin{pmatrix} \max\{16 - x_1 - x_3, 0\} \\ \max\{16 - x_2 - x_4, 0\} \end{pmatrix}.$$

3 Option and Policy Evaluation

Next, the optimal average cost per time step has to be computed for both settings. This can be done by the following steps:

Table 1. Recurrent states, control actions, state- and control cost and stationary probability distribution for the transshipment case under the optimal policy π_T^*

$\mathbf{x}$	$\pi(\mathbf{x})$	State cost $\mathbf{q}^T\mathbf{x}$	Control cost $\mathbf{p}^T\pi(\mathbf{x})$	$\mu_{\pi_T^*}$
(0,8,4,4)	(4,4,4,0)	8	0	$\frac{1}{3}$
(8,0,4,4)	(4,4,0,4)	8	0	$\frac{1}{3}$
(4,4,4,4)	(4,4,0,0)	8	0	$\frac{1}{3}$

1. Determine the stationary probability measure of the stochastic dynamical system induced by the optimal policy
2. Use the stationary measure to integrate the state and control cost over the state space.

For determining the stationary probability measure we note that there are only have a finite number of possible transitions in each state. This property, together with the special structure of the optimal policies, leads to only a finite number of recurrent states, which can easily be determined by straightforward construction of the transition graph from a given initial state. Consequently, the chosen policy induces a Markov chain on the set of recurrent states, whose stationary probability distribution can be computed by standard methods.

For the case with transshipment, there are only three recurrent states. Table 1 lists the set of recurrent states X_T together with the control action, the state- and control cost, and the stationary probability distribution $\mu_{\pi_T^*}$ on X_T under the policy π_T^*. The expected average cost per time step is given by

$$\sum_{\mathbf{x}\in X_T} \mu(\mathbf{x}) \left(\mathbf{q}^T\mathbf{x} + \mathbf{p}^T\pi(\mathbf{x})\right) = 8.$$

For the case without transshipment we obtain the set X_N of recurrent states nine elements

$$(0,16,8,0),(8,8,8,0),(4,12,8,0)$$
$$(8,8,0,8),(16,0,0,8),(12,4,0,8),$$
$$(4,12,4,4),(12,4,4,4),(8,8,4,4),$$

Again the stationary distribution is uniform. The optimal average cost per time step without transshipment is therefore

$$\sum_{x\in X_N} \frac{1}{9}((1,1,0,0)^T\mathbf{x} + (0,0)^T\pi(\mathbf{x})) = 16.$$

As the result, the *value* of the transshipment option in terms of its reduction of the expected average cost is $16 - 8 = 8$ units per time step.

It is also possible to study the sensitivity of this value for the chosen policy by considering hypothetical transportation and capacity reservation costs. If we assume that we have to pay c units per time step for providing the capacity

plus a linear transportation cost of r per unit to be transported, the value of the transshipment option reduces to

$$16 - \left(8 + c + \frac{8}{3}r\right) = 8 - c - \frac{8}{3}r$$

This way, the range of acceptable parameter combinations for the transshipment option can be easily characterized.

4 Conclusion

We have presented a way to valuate different supply chain configurations, or design options, with respect to their effect on the operational cost. The approach requires to compute the optimal operational (control) policy for each alternative design, and then to determine the resulting optimal average cost, by stochastic dynamic programming. For solving the stochastic control problems, an explicit discretization of the state space was avoided and replaced by an implicit enumeration via a sequence of parametric linear programs. It remains to be seen whether the approach is also practicable for larger problems than the small example considered here, as general parametric linear programs cannot be solved efficiently.

An important next step towards integrated supply chain design and control would be to find ways to incorporate the upper level decisions into the lower-level control problem and thus to formulate the bi-level problem as a true simultaneous optimization problem. The approach presented here is only applicable when the number of alternatives is small such that it searched exhaustively or as an evaluation method in iterative schemes, for example when using local search methods on the upper level.

References

1. Axsäter, S.: Manage. Sci. **36**, 1329–1338 (1990)
2. Axsäter, S.: Manage. Sci. **49**, 1168–1179 (2003)
3. Bertsekas, D.P.: Dynamic Programming and Optimal Control, Volume II. Athena, Belmont (2007)
4. Colson, B., Marcotte, P., Savard, G.: Ann. Oper. Res. **153**, 235–256 (2007)
5. de Kok, A.G., Graves, S.C.: Supply Chain management: Design, Coordination and Operation. Handbooks in Operations Research and Management Science. Elsevier, Amsterdam (2003)
6. Kvasnica, M., Grieder, P., Baotić, M.: Multi-Parametric Toolbox (MPT). Available via `http://control.ee.ethz.ch/mpt/` (2004)

Validated Methods: Applications to Modeling, Analysis, and Design of Systems in Medicine and Engineering

Andreas Rauh[1] and Ekaterina Auer[2]

[1] Chair of Mechatronics, University of Rostock, D-18059 Rostock, Germany
andreas.rauh@uni-rostock.de
[2] Faculty of Engineering, INKO, University of Duisburg-Essen, D-47048 Duisburg, Germany auer@inf.uni-due.de

During the last decades, computer assisted modeling and analysis of different industrial processes have increased in importance. Computers help to reduce the design and development time for new products and to substitute low cost virtual tests for expensive experiments on real life prototypes. However, the results are often unreliable due to errors that are generated either by the underlying computer arithmetic or by inaccuracy resulting from idealization of the mathematical model of the considered process. In this minisymposium, we focus on validated methods as a means to solve such problems.

A method is called *validated* if it guarantees the correctness of its output. In this context, intervals and Taylor models are widely used approaches to verifying results obtained on a computer. For example, the former provides a (multidimensional) box described in terms of floating point arithmetic which is guaranteed to contain the exact result. Besides, validated methods are able to allow for uncertainty in parameters, which helps to generate more realistic mathematical models or to take into account measurement errors.

The goal of this minisymposium is to make such techniques known to a broader circle of researchers and industry representatives. For this purpose, we outline their potential by presenting selected applications in medicine and engineering.

We begin the minisymposium by introducing validated methods and corresponding software libraries. Nathalie Revol, INRIA, Université de Lyon, France presented an overview focused on comparative advantages of different verified techniques. She highlighted slight but important differences in the definitions of basic interval concepts adopted by the modern libraries and gives insight on the current strivings for interval standardization.

After giving a general outline, we turn to the actual topic of the minisymposium – application of validated techniques to real life problems. Mathematical models describing static or dynamic processes are the basis of all

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*,
Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_86,

applications. These models are given by sets of algebraic equations, ordinary differential equations (ODEs), or differential-algebraic equations (DAEs). In all cases, there is an uncertainty in system parameters with a priori known bounds. In each application presented in this minisymposium, we use validated techniques to provide essential information about the parameter dependency, robustness, and safety of technical and medical systems in terms of guaranteed bounds for the quantities of interest. In contrast to non-validated techniques relying in most cases on grid-based or stochastic procedures for uncertainty quantification, interval methods, Taylor model approaches, and other validated techniques allow us to verify the worst-case influence of bounded uncertainties on mathematical system models almost as a by-product. In that sense they supplement traditional numerical techniques for modeling, analysis, and design of real life systems.

Andreas Rauh from the University of Rostock, Germany, discusses verification techniques for sensitivity analysis and design of controllers. He describes interval-based approaches for the analysis of reachability and observability of states of (nonlinear) dynamical systems with uncertainties and the validated simulation of sets of DAEs.

Mark A. Stadtherr, Department of Chemical and Biomolecular Engineering, University of Notre Dame, USA, analyzes the impact of infections within a population using epidemiological models with uncertainties by a method based on interval Taylor series to represent dependency on time and Taylor models to account for uncertainties in parameters and initial conditions.

Furthermore, Mareile Freihold, Institute of Measurement, Control, and Microtechnology, University of Ulm, Germany, discusses possibilities to reduce overestimation with the help of physical constraints. The potential of the approach is demonstrated for the validated ODE solver VALENCIA-IVP and an uncertain model of human blood cell dynamics.

Martin Tändl, Faculty of Engineering, University of Duisburg-Essen, Germany, presents applications of MOBILE, a software environment for modeling and simulation of multibody systems, to biomechanics. He accurately reconstructs bone motion from marker trajectories obtained in gait lab experiments using kinematical loops and functional human skeleton features. His approach is verified in SMARTMOBILE by Ekaterina Auer, Faculty of Engineering, University of Duisburg-Essen, Germany. SMARTMOBILE is a version of MOBILE which verifies kinematics and dynamics of various mechanical systems including closed loop ones and can additionally compute their sensitivity to parameters.

Finally, Michel Kieffer, Laboratoire des Signaux et Systèmes, Université Paris-Sud, France, focuses on applications of guaranteed computation in robotics with respect to robot localization and tracking, simultaneous localization and map building, and path planning under consideration of obstacles.

Verification Techniques for Sensitivity Analysis and Design of Controllers for Nonlinear Dynamical Systems with Uncertainties

Andreas Rauh[1], Johanna Minisini[2], and Eberhard P. Hofer[2]

[1] Chair of Mechatronics, University of Rostock, Germany, andreas.rauh@uni-rostock.de
[2] Institute of Measurement, Control, and Microtechnology, University of Ulm, Germany, johanna.minisini@uni-ulm.de, eberhard.hofer@uni-ulm.de

Summary. Controllers for nonlinear dynamical systems are often based on properties such as differential flatness or exact input-output as well as input-to-state linearizability. However, these approaches are limited to specific classes of system models. To generalize design procedures and to account for parameter uncertainties as well as modeling errors, an interval arithmetic approach for validated simulation of both ordinary differential equations and differential-algebraic equations is extended to the synthesis and sensitivity analysis of open-loop and closed-loop controllers. Furthermore, interval arithmetic routines for evaluation of criteria for reachability and observability of states are implemented using automatic differentiation.

1 Modeling, Analysis, and Design of Control Systems

In this paper, modeling, analysis, and design of open-loop as well as closed-loop controllers for nonlinear dynamical systems described by sets of ordinary differential equations (ODEs)

$$\dot{x}(t) = f(x(t), p(t), u(t), t) \quad \text{with} \quad x \in \mathbb{R}^{n_x},\ p \in \mathbb{R}^{n_p},\ u \in \mathbb{R}^{n_u} \tag{1}$$

are discussed. For that purpose, two different scenarios are distinguished.

First, the sensitivity of the controlled systems' trajectories $x(t)$ is analyzed with respect to uncertainties of the initial conditions $x(t_0)$ and the parameters $p(t)$. In this case, either open-loop control laws $u(t)$ or closed-loop control laws $u(x(t))$ are assumed to be given. The numerical solution approach is based on calculating guaranteed enclosures of all reachable states using validated ODE solvers such as ValEncIA-IVP. In Sect. 2, an overview of ValEncIA-IVP which has been developed to solve initial value problems (IVPs) for ODEs is given. Furthermore, extensions for computing partial derivatives of the states $x(t)$ with respect to (uncertain) parameters p using libraries for automatic and algorithmic differentiation (AD) are described.

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*, Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_87,

Second, basic steps of a validated IVP solver for differential-algebraic equations (DAEs) are highlighted in Sect. 3. In addition to sensitivity analysis, open-loop control sequences $u(t)$ are determined such that a specific output variable matches a predefined time response. Finally, relations to differential geometric criteria for analysis of reachability and observability of states and their potential for inclusion in ValEncIA-IVP are pointed out in Sect. 4.

2 Validated Sensitivity Analysis Using ValEncIA-IVP

In the following, ODEs $\dot{x}(t) = f(x(t), p, t)$ are considered which describe both open-loop and closed-loop systems. The vector p consists of all time-invariant system and controller parameters. The differential sensitivities $s_i(t)$ of the solution $x(t)$ with respect to the parameters p are defined by

$$\dot{s}_i(t) = \frac{\partial f(x(t), p, t)}{\partial x} \cdot s_i(t) + \frac{\partial f(x(t), p, t)}{\partial p_i} \quad \text{for all} \quad i = 1, \ldots, n_p. \tag{2}$$

The new state vectors $s_i(t)$ in (2) are given by

$$s_i(t) := \frac{\partial x(t)}{\partial p_i} \in \mathbb{R}^{n_x} \quad \text{with} \quad s_i(t_0) = \frac{\partial x(t_0, p)}{\partial p_i}. \tag{3}$$

If the initial states $x(t_0) \in [\underline{x}(t_0)\,;\,\overline{x}(t_0)]$ are independent of $p \in \left[\underline{p}\,;\,\overline{p}\right]$, the equality $s_i(t_0) = 0$ holds. In ValEncIA-IVP, the ODEs (2) do not need to be derived symbolically, since all required partial derivatives w.r.t. x and p are computed using AD provided by FADBAD++ (`www.fadbad.com`) [1,2].

As for the case of solving an IVP for the ODEs $\dot{x}(t) = f(x(t), p, t)$ using ValEncIA-IVP, guaranteed state enclosures

$$[x(t)] := x_{app}(t) + [R_x(t)] \tag{4}$$

are determined in a first stage. In (4), the approximate solution $x_{app}(t)$ for the IVP is determined numerically using a non-validated ODE solver. The guaranteed error bounds $[R_x(t)]$ are calculated iteratively, see e.g. [4]. In a second stage, after convergence of the iteration for the interval bounds $[x(t)]$, suitable approximate solutions $s_{i,app}(t)$ and additional enclosures

$$[s_i(t)] := s_{i,app}(t) + [R_{s,i}(t)] \quad \text{with} \quad s_{i,app}(t) \in \mathbb{R}^{n_x} \quad \text{and} \quad i = 1, \ldots, n_p \tag{5}$$

are determined for the sensitivities. For both exactly known and uncertain values of p and $x(t_0)$, the intervals $[s_i(t)]$ are determined such that the partial derivatives of all reachable states w.r.t. all possible p_i are included. For time-varying parameters $p(t)$, the sensitivities $s_i(t)$ are computed w.r.t. time-invariant variables $\epsilon_i \approx 0$ after substituting $p(t) + \epsilon$ with $\epsilon \in \mathbb{R}^{n_p}$ for $p(t)$.

The calculation of the differential sensitivities $s_i(t)$ using a validated ODE solver provides useful information for the design of controllers. Guaranteed

bounds of sensitivities of the state variables can be obtained for uncertain parameters $p \in [p]$ using a single evaluation of the state equations even for non-monotonic relations between the parameters p and the state variables x. Techniques for the reduction of overestimation are available which combine consistency tests and exponential state enclosures to avoid growth of the diameters of the state enclosures especially for asymptotically stable systems [5].

3 Application of ValEncIA-IVP to Sets of DAEs

In previous work, ValEncIA-IVP has been extended to determine guaranteed state enclosures also for DAEs [5]. In the following, semi-explicit DAEs

$$\dot{x}(t) = f\left(x(t), y(t), t\right) \qquad \text{with} \quad f : D \mapsto \mathbb{R}^{n_x} \tag{6}$$

$$0 = g\left(x(t), y(t), t\right) \qquad \text{with} \quad g : D \mapsto \mathbb{R}^{n_y}, D \subset \mathbb{R}^{n_x} \times \mathbb{R}^{x_y} \times \mathbb{R}^1 \tag{7}$$

and the consistent initial conditions $x\left(0\right)$ and $y\left(0\right)$ are evaluated. These DAEs may further depend upon uncertain parameters p. To simplify the notation in the Sects. 3 and 4, the dependency upon p is not explicitly denoted. However, all presented criteria are also applicable to $p \in \left[\underline{p}\,;\,\overline{p}\right]$ with $\underline{p} < \overline{p}$.

For validated DAE solvers, there are two important applications. First, as for validated simulation of ODEs, the influence of uncertainties is analyzed by calculating guaranteed state enclosures. Second, open-loop control strategies are determined such that the system's output signal matches a predefined time-response. This task is often referred to as the inverse control problem. E.g., for nonlinear exactly input-to-state linearizable sets of ODEs, this task can be solved symbolically by expressing $u\left(t\right)$ (as one component of $y\left(t\right)$ in (6),(7)) in terms of the state variables of the exactly linearized system. However, numerical design approaches based on interval analysis are more flexible since uncertainties and robustness requirements can be taken into account directly. For that purpose, sets of ODEs and DAEs are extended by time-dependent algebraic constraints to specify the desired output. The corresponding solution provides both the desired control and an enclosure of all reachable states.

3.1 Solving DAE Systems Using Interval Arithmetic

The approach for solving DAEs using ValEncIA-IVP is based on substituting state enclosures $x_i\left(t\right) \in [x_i\left(t\right)]$ and $y_i\left(t\right) \in [y_i\left(t\right)]$ defined by

$$\begin{aligned}
[x_i\left(t\right)] &:= x_{app,i}\left(t_k\right) + \left(t - t_k\right) \cdot \dot{x}_{app,i}\left(t_k\right) + [R_{x,i}\left(t_k\right)] + \left(t - t_k\right) \cdot [\dot{R}_{x,i}\left(t\right)] \\
&\text{for} \quad i = 1, \ldots, n_x \quad \text{and} \quad t \in [t_k\,;\,t_{k+1}]\ ,\ t_0 \le t \le T \\
[y_i\left(t\right)] &:= y_{app,i}\left(t_k\right) + \left(t - t_k\right) \cdot \dot{y}_{app,i}\left(t_k\right) + [R_{y,i}\left(t\right)] \quad \text{for} \quad i = 1, \ldots, n_y
\end{aligned} \tag{8}$$

for the differential and algebraic state variables $x(t)$ and $y(t)$, resp. In (8), t_k and t_{k+1} are two subsequent points of time for which guaranteed state enclosures are determined. For $t = t_0$, the conditions $[x(t_0)] = x_{app}(t_0) + [R_x(t_0)]$ and $[y(t_0)] = y_{app}(t_0) + [R_y(t_0)]$ have to be fulfilled.

However, the information provided by (6) and (7) is not always sufficient to solve for $[\dot{R}_x(t)]$ and $[R_y(t)]$ using validated methods such as the Krawczyk iteration. E.g., the equations (10) are singular for small step sizes $t_{k+1} \to t_k$. Then, *additional* hidden constraints are necessary to restrict the set of feasible solutions and to verify the consistency of initial conditions $x(t_0)$ and $y(t_0)$. For that purpose, those constraints $g_i(x)$ are considered, which do not depend explicitly upon y. Differentiation w.r.t. time leads to

$$\frac{d^j g_i(x)}{dt^j} = \left(\frac{\partial L_f^{j-1} g_i(x)}{\partial x}\right)^T \cdot f(x,y) = L_f^j g_i(x) = 0\ ,\ L_f^0 g_i(x) = g_i(x)\,. \quad (9)$$

The Lie derivatives $L_f^j g_i(x)$ are computed by FADBAD++ providing AD and automatic calculation of Taylor coefficients up to the smallest order $j > 0$ for which $L_f^j g_i(x)$ depends upon at least one component of y (i.e., up to the differentiation index of DAEs). The computation of Lie derivatives and Lie brackets in the Sects. 3 and 4 is inspired by [6]. To our knowledge, the author of [6] has not used this approach in an interval arithmetic framework to account for uncertainties in ODEs and DAEs and to design robust controllers.

A suitable extension for solving sets of DAEs in VALENCIA-IVP is to treat the intervals $[R_y(t)]$ as *constant interval parameters* in a first stage and to solve the non-algebraic equations for $[\dot{R}_x(t)]$. In a second stage, the consistency of the solution has to be proven with the help of the algebraic equations and their time derivatives by showing that feasible solutions are guaranteed to be contained in the interior of $[R_y(t)]$.

3.2 Example

To demonstrate the use of the hidden constraints (9), the pendulum example from [3] which is rewritten as a set of first order differential equations

$$\begin{aligned} \dot{x}_1 &= x_3 & \dot{x}_3 &= -x_1 y & g(x) &= x_1^2 + x_2^2 - 1 = 0 \\ \dot{x}_2 &= x_4 & \dot{x}_4 &= -x_2 y + 1 \end{aligned} \quad (10)$$

is considered. For this system, symbolic evaluation of (9) gives

$$\frac{dg(x)}{dt} = 2(x_1\dot{x}_1 + x_2\dot{x}_2) = 2(x_1 x_3 + x_2 x_4) \stackrel{!}{=} 0 \quad \text{and}$$
$$\frac{d^2 g(x)}{dt^2} = 2(\dot{x}_1 x_3 + x_1\dot{x}_3 + \dot{x}_2 x_4 + x_2\dot{x}_4) = 2\left(x_3^2 - x_1^2 y + x_4^2 - x_2^2 y + x_2\right) \stackrel{!}{=} 0$$

for the first and second derivatives with respect to time.

These expressions can be used to verify the results obtained by AD which show that $x(t_0) = \begin{bmatrix}1 & 0 & 0 & 1\end{bmatrix}^T$ and $y(t_0) = 1$ represent consistent initial values, while certainly no consistent initial conditions are included in $x(t_0) \in \begin{bmatrix}[-0.5\ ;\ 0.5] & [-0.5\ ;\ 0.5] & [-0.5\ ;\ 0.5] & [-1.0\ ;\ 1.0]\end{bmatrix}^T$ for $y(t_0) = 1$.

4 Relations to Exact Feedback Linearization

In the remainder of this paper, nonlinear input-affine dynamical systems

$$\dot{x}(t) = f(x(t)) + g(x(t)) \cdot u(t) \quad \text{with output equations} \quad y(t) = h(x(t)) \tag{11}$$

are considered. Non-input-affine systems $\dot{x} = f(x, u)$ can be transformed using artificial control inputs $\tilde{u}$ according to $\dot{x} = f(x, u)$, $\dot{u} = \tilde{u}$, $y = h(x)$. The goal is to convert the input-affine dynamical system into a set of linear ODEs

$$\dot{z}(t) = A \cdot z(t) + B \cdot v(t) \quad \text{with} \quad y(t) = C \cdot z(t) \tag{12}$$

using a coordinate transformation $z = \tau(x) : D \subset \mathbb{R}^{n_x} \mapsto \mathbb{R}^{n_x}$ and a control law $u = r(x) + V(x) \cdot v$ for exact input-to-state linearization. Let δ_i be the relative degree of the output y_i, $i = 1, \ldots, m$, i.e., the smallest order of the derivative $d^{\delta_i} y / dt^{\delta_i}$ which explicitly depends on u. Then, the state transformation

$$z = \tau(x) = \begin{bmatrix}\tau_1^1(x) \ \ldots \ \tau_1^{\delta_1}(x) \ \tau_2^1(x) \ \ldots\end{bmatrix}^T \tag{13}$$

can be computed using the Lie derivatives $\tau_i^{r_i} = L_f^{r_i - 1} h_i(x)$, $r_i = 1, \ldots, \delta_i$.

The feedback control law $u = r(x) + V(x) \cdot v$ is given by

$$r(x) = -D^{-1}(x)\varphi(x) \quad \text{and} \quad V(x) = D^{-1}(x) \tag{14}$$

with

$$\varphi(x) = \left[L_f^{\delta_1} h_1(x) \ L_f^{\delta_2} h_2(x) \ \ldots \ L_f^{\delta_m} h_m(x)\right]^T_{x(t) = \tau^{-1}(z)} \tag{15}$$

and the decoupling matrix

$$D(x) = \begin{bmatrix} L_{g_1} L_f^{\delta_1 - 1} h_1(x) & L_{g_2} L_f^{\delta_1 - 1} h_1(x) & \cdots & L_{g_m} L_f^{\delta_1 - 1} h_1(x) \\ L_{g_1} L_f^{\delta_2 - 1} h_2(x) & L_{g_2} L_f^{\delta_2 - 1} h_2(x) & \cdots & L_{g_m} L_f^{\delta_2 - 1} h_2(x) \\ \vdots & \vdots & \vdots & \vdots \\ L_{g_1} L_f^{\delta_m - 1} h_m(x) & L_{g_2} L_f^{\delta_m - 1} h_m(x) & \cdots & L_{g_m} L_f^{\delta_m - 1} h_m(x) \end{bmatrix}. \tag{16}$$

Linear feedback controllers can be designed for (12) if $\text{rank}\{D(x)\} = n_x$ and $\delta = \delta_1 + \ldots + \delta_m = n_x$ hold for the trajectories of all desired states $x(t)$, all $p \in [p]$, and all $x(t_0) \in [x(t_0)]$. This is verified by evaluation of $D(x)$ for interval boxes containing all $x(t)$. All derivatives in (15) and (16) are computed

by FADBAD++ as in the preceding section. If the sensitivity analysis, see Sect. 2, is successful and if $D(x)$ is regular for all desired states, reachability and observability have to be guaranteed to implement robust controllers. The corresponding criteria for nonlinear systems are generalizations of Kalman's criteria for state controllability and observability of linear systems.

With the help of $P_0(x) = g(x)$, $P_1(x) = [f(x), g(x)]$, and $P_k(x) = [f(x), P_{k-1}(x)]$, $k = 2, \ldots, n_x - 1$, i.e., the Lie brackets of $f(x)$ and $g(x)$ which are defined by $[f(x), g(x)] = \frac{\partial g(x)}{\partial x} f(x) - \frac{\partial f(x)}{\partial x} g(x)$, the state-dependent reachability matrix $P(x) = \left[P_0(x)\ P_1(x) \ldots P_{n_x-1}(x)\right]$ can be determined. The observability matrix is defined accordingly by

$$Q(x) = \left[\left(\frac{\partial h(x)}{\partial x} \right)^T \left(\frac{\partial L_f h(x)}{\partial x} \right)^T \ldots \left(\frac{\partial L_f^{n_x-1} h(x)}{\partial x} \right)^T \right]^T .$$

Using a validated LU-decomposition of interval matrices, the rank of $D(x) \in \mathbb{R}^{n_x \times n_x}$, $P(x) \in \mathbb{R}^{n_x \times n_x n_u}$, and $Q(x) \in \mathbb{R}^{n_x n_y \times n_x}$ is determined. Routines for trajectory planning have to make sure that these matrices have full rank n_x.

5 Outlook on Future Research

Further general purpose strategies will be developed to make use of the results obtained by interval evaluation of the criteria in Sect. 4 to generate reference signals for controllers such that reachability and observability are guaranteed. Additional extensions will be the proof of asymptotic stability of closed-loop controllers if the relative degree δ is smaller than n_x. In this case, instabilities of the internal dynamics have to be detected and avoided in a guaranteed way. Finally, combinations with feedforward control strategies (e.g. control sequences determined by the presented DAE approach) and other techniques for the design of nonlinear controllers will be investigated.

References

1. Bendsten, C., Stauning, O.: FADBAD, a Flexible C++ Package for Automatic Differentiation Using the Forward and Backward Methods. Technical Report 1996-x5-94, Technical University of Denmark, Lyngby, 1996
2. Bendsten, C., Stauning, O.: TADIFF, a Flexible C++ Package for Automatic Differentiation Using Taylor Series. Technical Report 1997-x5-94, Technical University of Denmark, Lyngby, 1997
3. Nedialkov, N.S.: Interval Tools for ODEs and DAEs. In CD-Proc. of the 12th GAMM-IMACS International Symposium on Scientific Computing, Computer Arithmetic, and Validated Numerics SCAN 2006, Duisburg, Germany, IEEE Computer Society (2007)

4. Rauh, A., Auer, E., Hofer, E.P.: ValEncIA-IVP: A Comparison with Other Initial Value Problem Solvers. In CD-Proc. of the 12th GAMM-IMACS International Symposium on Scientific Computing, Computer Arithmetic, and Validated Numerics SCAN 2006, Duisburg, Germany, IEEE Computer Society (2007)
5. Rauh, A., Auer, E., Minisini, J., Hofer, E.P.: Extensions of ValEncIA-IVP for Reduction of Overestimation, for Simulation of Differential Algebraic Systems, and for Dynamical Optimization. PAMM **7**(1), 1023001–1023002 (2007)
6. Röbenack, K.: On the Efficient Computation of Higher Order Maps $ad_f^k g(x)$ Using Taylor Arithmetic and the Campbell-Baker-Hausdorff Formula. In Zinober, A., Owens, D. (eds.) Nonlinear and Adaptive Control. Lecture Notes in Control and Information Science, vol. 281, pp. 327–336. Springer, Berlin (2002)

Verified Solution of Nonlinear Dynamic Models in Epidemiology

Joshua A. Enszer and Mark A. Stadtherr

Department of Chemical and Biomolecular Engineering
University of Notre Dame, Notre Dame, IN 46556, USA, jenszerl@nd.edu, markst@nd.edu

Summary. Epidemiological models can be used to study the impact of an infectious disease within a population. These models often involve parameters that are not known with certainty. A method is described for bounding the disease trajectories that are possible for given bounds on uncertain parameters and/or initial states. The method is based on the use of an interval Taylor series to represent dependence on time and the use of Taylor models to represent dependence on uncertain quantities. The use of this method in epidemiology is demonstrated using the SIRS model.

1 Introduction

Ordinary differential equations (ODEs) are the basis for many mathematical models in the sciences, including population models used in epidemiology. Specifically, the Kermack-McKendrick model [5] was one of the first developed to simulate the spread of infectious diseases such as bubonic plague and cholera. This model, and other compartmental models in epidemiology, partition the population into classes and describe the rate of population change in each class. Our focus here is on continuous epidemiological models that are systems of ODEs and formulated as initial value problems (IVPs). Thus, the model is integrated over time, starting with specified initial values for the different population classes.

Of interest here is the verified (i.e., mathematically and computationally guaranteed) solution of such systems of ODEs, especially systems that involve uncertainty in initial conditions or model parameters. Accounting for uncertainties is particularly important in the context of epidemiological models, since in many, if not most, cases, initial populations and model parameters (e.g., rate constants) may not be known exactly. We will assume that, for such uncertain quantities, only upper and lower bounds are available. That is, uncertain quantities will be represented by intervals. Since this implies that there are infinitely many possible values for the uncertain quantities, the underlying ODE system will have infinitely many possible solutions. To solve such a system, we seek rigorous, verified bounds on the possible trajectories.

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*,
Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_88,

For determining rigorous bounds on the solution of an ODE system, with or without uncertainties, the use of interval methods (also called validated or verified methods) is a natural approach, as computations with intervals, as opposed to real numbers, can provide both mathematically and computationally guaranteed enclosures. Excellent reviews of interval methods for IVPs are available in the literature [9,11]. There are several available software packages that treat interval-valued initial conditions, including AWA [7], VNODE [10], and COSY VI [1]. These packages can also deal indirectly with interval-valued parameters. An approach that deals directly with interval-valued parameters (and initial states) has recently been described by Lin and Stadtherr [6], who implemented this approach in a solver called VSPODE (Verifying Solver for Parametric ODEs). Both COSY VI and VSPODE use Taylor models [8], though in different ways, to deal with the uncertain quantities (parameters and initial values). In this paper, we propose the use of Taylor-model methods, specifically VSPODE, for propagating uncertainties through nonlinear ODE models in population epidemiology.

2 Problem Statement

We consider epidemiological models that can be represented as systems of ODEs, for which an IVP must be solved. In general mathematical form, this problem may be written

$$y'(t) = f(y, \theta), \qquad y(t_0) = y_0 \in [y_0], \qquad \theta \in [\theta], \tag{1}$$

where $t \in [t_0; t_m]$ for some $t_m > t_0$. Here y is the n-dimensional vector of state variables with initial value y_0, and θ is a p-dimensional vector of time-invariant parameters. The vectors $[y_0]$ and $[\theta]$ are intervals that enclose uncertainties in the initial states and parameters, respectively. If the ODE system is nonautonomous, or if the parameters have a known dependence on time, then such a model can be put into the form of (1) by the introduction of one or more new state variables. Our specific goal is to obtain a guaranteed enclosure of the state variables y at all times of interest.

3 Background

3.1 Interval Analysis

The real interval vector $[x] = [\underline{x}; \overline{x}]$ is an enclosure of the real vector $x = [x_1, \ldots, x_n]^{\mathrm{T}}$, $n \geq 1$. The real vectors $\underline{x} = [\underline{x}_1, \ldots, \underline{x}_n]^{\mathrm{T}}$ and $\overline{x} = [\overline{x}_1, \ldots, \overline{x}_n]^{\mathrm{T}}$ provide the lower and upper bounds, respectively, on the components of x. That is, $\underline{x}_i \leq x_i \leq \overline{x}_i$ or $x_i \in [\underline{x}_i; \overline{x}_i]$. Basic arithmetic operations are defined on interval scalars according to $[x] \circ [y] = \{x \circ y \mid x \in [x], y \in [y]\}, \circ \in$

$\{+, -, \times, \div\}$, with division in the case of $[y]$ containing zero allowed only in extensions of interval arithmetic [4]. Addition and multiplication are commutative and associative but only subdistributive. Interval versions of the elementary functions can also be defined.

For a real function $f(x)$, an interval extension $f^{\mathrm{I}}([x])$ encloses the range of $f(x)$ for $x \in [x]$. That is, $f^{\mathrm{I}}([x]) \supseteq \{f(x) \mid x \in [x]\}$. When $f(x)$ can be written as a series of arithmetic operations and elementary functions, an interval extension can be obtained by substituting $[x]$ into $f(x)$ and evaluating using interval arithmetic. In this case, $f^{\mathrm{I}}([x]) = f([x])$, which is referred to as the natural interval extension. Computing the interval extension in this way may result in overestimation of the function range due to the "dependency" problem. While a variable may take on any value within its interval, it must take on the *same* value each time it occurs in an expression. However, this type of dependency is not recognized when the natural interval extension is computed. Another source of overestimation that may arise in the use of interval methods is the "wrapping" effect. This occurs when an interval is used to enclose (wrap) a set of results that is not an interval. If this overestimation is propagated from step to step in an integration procedure for ODEs it can quickly lead to the loss of a meaningful enclosure. Both of these sources of overestimation can be addressed through the use of Taylor models.

3.2 Taylor Models

Makino and Berz [8] have described a remainder differential algebra (RDA) approach for bounding function ranges and control of the dependency problem of interval arithmetic. In this method, a function is represented using a model consisting of a Taylor polynomial and an interval remainder bound. Such a model is called a Taylor model.

One way of forming a Taylor model of a function is by using the Taylor theorem. Consider a real function $f(x)$ that is $(q+1)$ times partially differentiable on $[x]$ and let $x_0 \in [x]$. The Taylor theorem states that for each $x \in [x]$, there exists a real ζ with $0 < \zeta < 1$ such that

$$f(x) = p_f(x - x_0) + r_f(x - x_0, \zeta), \tag{2}$$

where p_f is a q-th order polynomial (truncated Taylor series) in $(x - x_0)$ and r_f is a remainder, which can be quantitatively bounded over $0 < \zeta < 1$ and $x \in [x]$ using interval arithmetic or other methods to obtain an interval remainder bound $[r_f]$. A q-th order Taylor model $T_f = p_f + [r_f]$ for $f(x)$ over $[x]$ then consists of the polynomial p_f and the interval remainder bound $[r_f]$ and is denoted by $T_f = (p_f, [r_f])$. Note that $f \in T_f$ for $x \in [x]$ and so T_f encloses the range of f over $[x]$.

In practice, it is more useful to compute Taylor models of functions by performing Taylor model operations. Arithmetic operations with Taylor models can be done using RDA operations [8], which include addition, multiplication,

reciprocal, and intrinsic functions. Using these, it is possible to start with simple functions such as the constant function $f(x) = k$, for which $T_f = (k, [0; 0])$, and the identity function $f(x_i) = x_i$, for which $T_f = (x_{i0} + (x_i - x_{i0}), [0; 0])$, and then to compute Taylor models for very complicated functions. Therefore, it is possible to compute a Taylor model for any function representable in a computer environment by simple operator overloading through RDA operations. It has been shown that, compared to other rigorous bounding methods, the Taylor model often yields sharper bounds for modest to complicated functional dependencies [8, 12].

4 Solution Procedure

In this section we briefly summarize the method used by VSPODE for solving the problem described in Sect. 2. A fully detailed description of this method has been given by Lin and Stadtherr [6].

At each integration step j in VSPODE, there are two phases in the computation. In the first phase, existence and uniqueness of the solution are proven over a time step $t \in [t_j; t_{j+1}]$ using the Picard-Lindelöf operator and the Banach fixed point theorem, and a rough enclosure $[\widetilde{y}_j]$ of the solution over this time step is computed. This is done using a high-order interval Taylor series (ITS) with respect to time, in a procedure similar to that used in VNODE [10].

In the second phase, a tighter enclosure $[y_{j+1}] \subseteq [\widetilde{y}_j]$ of $y_{j+1} = y(t_{j+1})$ is computed. This is done by using an ITS approach to compute $T_{y_{j+1}}(y_0, \theta)$, a Taylor model of y_{j+1} in terms of the initial values y_0 and parameters θ, and then bounding this Taylor model over $y_0 \in [y_0]$ and $\theta \in [\theta]$. To control the wrapping effect, the state enclosures are propagated using a new type of Taylor model consisting of a polynomial and a *parallelepiped* (as opposed to an interval) remainder bound. In performance comparisons [6], VSPODE provided tighter enclosures on the state variables than VNODE, and required significantly less computation time.

5 Example: SIRS Model

The basic SIRS model [2] is similar to the constant-population model first studied by Kermack and McKendrick [5]. There are three population classes, namely Susceptible, Infected and Recovered, with populations s, i and r, respectively. Unlike the original Kermack–McKendrick model, however, immunity is not permanent, so Recovered individuals can become Susceptible again. Assuming a constant total population $n = s + i + r$, the model equations are

$$\frac{ds}{dt} = -\beta si + \gamma r = -\beta si + \gamma(n - s - i) \tag{3}$$

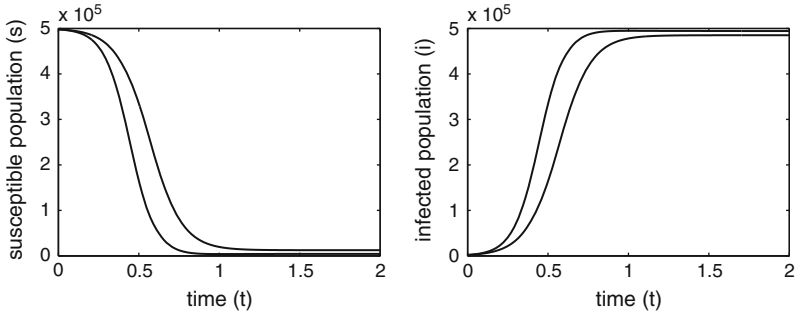


Fig. 1. VSPODE enclosure of Susceptible and Infected population trajectories of simple SIRS model

$$\frac{di}{dt} = \beta si - \nu i. \tag{4}$$

For this simple example, we have chosen a total population of $n = 500{,}000$ individuals (indv), with an initial Infected population of $i_0 = 2{,}000$ indv and initial Susceptible population of $s_0 = 498{,}000$ indv. We also set the susceptibility rate constant to be $\gamma = 50$ yr^{-1}. Uncertain values are assumed for the recovery rate constant, $\nu \in [0.125; 0.250]$ yr^{-1}, and for the infection probability, $\beta \in [2; 2.5] \times 10^{-5}$ yr^{-1}indv^{-1}.

VSPODE was applied to determine a verified enclosure of all possible solutions to this model for $t = 0$ to $t = 10$ yr. The results, out to $t = 2$ yr, are shown for $s(t)$ and $i(t)$ in Fig. 1. The curves shown in these figures are upper and lower bounds, which are mathematically and computationally guaranteed, on the possible trajectories of the Susceptible and Infected populations.

Since interval methods have a reputation of often producing only very loose bounds, we checked the tightness of the VSPODE bounds by comparison to the results of a Monte Carlo simulation with 100,000 trials. For each trial, real values of ν and β were selected at random from within their specified interval bounds. Bounds obtained from Monte Carlo analysis are not guaranteed and in general will yield an inner estimate of the true bounds (the guaranteed VSPODE bounds represent an outer estimate). Results for the Infected population are shown in Table 1, which provide a direct numerical comparison of the bounds obtained from VSPODE and from Monte Carlo analysis. The true bounds on the trajectories will be between the VSPODE bounds (outer estimate) and the Monte Carlo bounds (inner estimate). The closeness of these two sets of bounds demonstrates that the method used in VSPODE is capable of determining verified bounds that are in fact very tight. For the final time of $t = 10$ yr, the VSPODE bounds converge to a solution of $s \in [4{,}373; 12{,}501]$ and $i \in [485{,}073; 494{,}389]$. This numerical result can be compared to exact interval bounds obtained from the analytical steady-state solution, $s_s = \nu/\beta = [5{,}000; 12{,}500]$, and

Table 1. Numerical comparison of VSPODE enclosure and Monte Carlo simulation (MC) for Infected population i in simple SIRS model

t	$\underline{i}$ (VSPODE)	$\underline{i}$ (MC)	$\overline{i}$ (MC)	$\overline{i}$ (VSPODE)
0.2	13,612	13,728	22,751	22,761
0.4	80,385	82,448	180,048	180,898
0.6	282,559	286,360	430,625	434,961
0.8	440,344	441,604	487,722	490,957
1.0	477,942	478,411	493,240	494,764
1.2	483,989	484,123	493,713	494,565
1.4	484,904	484,945	493,753	494,436
1.6	485,045	485,062	493,756	494,400
1.8	485,068	485,079	493,757	494,391
2.0	485,072	485,080	493,757	494,389

$i_s = (\gamma n - s_s)/(\nu + \gamma) = [485{,}074; 493{,}766]$. The method employed by VSPODE accurately and tightly bounds the true solution. Several additional examples of the use of this technique will be presented elsewhere [3].

References

1. Berz, M., Makino, K.: Reliab. Comput. **4**, 361–369 (1998)
2. Edelstein-Keshet, L.: Mathematical Models in Biology, SIAM, Philadelphia (2005)
3. Enszer, J.A., Stadtherr, M.A.: Int. J. Appl. Math. Comput. Sci. **19**, 501–512 (2009)
4. Hansen, E.R., Walster, G.W.: Global Optimization Using Interval Analysis. Marcel Dekker, New York (2004)
5. Kermack, W.O., McKendrick, A.G.: Proc. R. Soc. Lond. A **115**, 700–721 (1927)
6. Lin, Y., Stadtherr, M.A.: Appl. Num. Math. **57**, 1145–1162 (2007)
7. Lohner, R.J.: In: Computational Ordinary Differential Equations, pp. 425–435. Clarendon, Oxford, UK (1992)
8. Makino, K., Berz, M.: In: Berz, M., Bishof, C., Corliss, G., Griewank, A. (eds.) Computational Differentiation: Techniques, Applicati.ons, and Tools, pp. 63–74. SIAM, Philadelphia (1996)
9. Nedialkov, N.S., Jackson, K.R., Corliss, G.F.: Appl. Math. Comput. **105**, 21–68 (1999)
10. Nedialkov, N.S., Jackson, K.R., Pryce, J.D.: Reliab. Comput. **7**, 449–465 (2001)
11. Neher, M., Jackson, K.R., Nedialkov, N.S.: SIAM J. Numer. Anal. **45**, 236–262 (2007)
12. Neumaier, A.: Reliab. Comput. **9**, 43–79 (2003)

Physically Motivated Constraints for Efficient Interval Simulations Applied to the Analysis of Uncertain Models of Blood Cell Dynamics

Mareile Freihold[1], Andreas Rauh[2], and Eberhard P. Hofer[1]

[1] Institute of Measurement, Control, and Microtechnology, University of Ulm, Germany m.freihold@web.de, eberhard.hofer@uni-ulm.de
[2] Chair of Mechatronics, University of Rostock, Germany andreas.rauh@uni-rostock.de

Summary. Interval arithmetic techniques such as ValEncIA-IVP allow one to calculate guaranteed enclosures of all reachable states of dynamical systems under consideration of bounded uncertainties of both initial conditions and system parameters. Considering the fact that in naive implementations of interval algorithms, overestimation might lead to unnecessarily conservative results, suitable consistency tests are essential for obtaining tightest possible enclosures. In this contribution physically motivated constraints are derived to implement a consistency test for a high-dimensional model of granulopoiesis in human blood cell dynamics.

1 Introduction

ValEncIA-IVP is an interval arithmetic solver which calculates guaranteed enclosures of all reachable states for initial value problems (IVPs) for sets of ordinary differential equations (ODEs). Through the use of interval arithmetic it allows for uncertainty in the initial conditions and system parameters [6]. Using only naive implementations, the resulting state enclosure might be too conservative. In order to obtain the tightest possible enclosures, overestimation has to be detected and reduced. A tighter enclosure of the exact solution is achieved by defining physically motivated constraints which allow to identify regions in the state space that are physically meaningless. The constraints are implemented in a Branch and Bound consistency test [1, 2] that confines the state enclosure with the aid of the constraints to physically meaningful areas. The goal is to enclose the exact solution as precisely as possible.

The applicability of the Branch and Bound algorithm is demonstrated on a dynamic system with uncertainties, a high-dimensional biomathematical model of blood cell growth according to Fliedner and Steinbach [4]. The constraints are based on decoupling of cell compartments, such that internal exchange of cells does not influence the input and output variables explicitly.

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*,
Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_89,

2 Human Blood Cell Dynamics: Granulopoiesis

In this paper, a biomathematical model describing the growth of human blood cells is analyzed. Since initial conditions and parameters of such mathematical models are subject to significant uncertainties, interval arithmetic can be applied to determine the worst-case influence of selected parameters on the state variables which correspond to the concentrations of cells in different compartments of the model [3]. Fliedner and Steinbach came up with a model consisting of a set of nonlinear coupled ODEs, see Fig. 1, expressing granulopoiesis in terms of cell and information flow [3]. Granulopoiesis is categorized into seven compartments representing the proliferation stages of granulocytes. All blood cells originate from the *stem cell compartment* (S) from which the cell grows until it reaches the circulating blood, represented by the *function compartment* (F). The stages in between are the *compartment bone marrow* (CBM) and the *compartment blood* (CBL), then the *precursor* (P), the *mature* (M), and the *reserve cells* (R). The state variables

$$\begin{aligned} x_1 &= S & x_{33} &= R \\ x_2 &= CBM_1\,, \ldots,\ x_{11} = CBM_{10} & x_{34} &= F \\ x_{12} &= CBL_1\,, \ldots,\ x_{21} = CBL_{10} & x_{35} &= RegI \\ x_{22} &= P_1\,, \ldots,\ x_{31} = P_{10} & x_{36} &= RegII \\ x_{32} &= M \end{aligned} \tag{1}$$

together with the abbreviations

$$\begin{aligned} u_1 &= \gamma_1 \exp(-\nu_1 x_1) + \gamma_2 \exp\big(-\nu_2 \cdot (x_2 + \ldots + x_{11} + x_{22} + \ldots + x_{33})\big) + \gamma_3 \\ u_2 &= \gamma_4 - \gamma_5 \exp(-\nu_3 x_{35}) & u_6 &= \gamma_6 - \gamma_y \exp(-\nu_4 x_{35}) \\ u_3 &= 2(1-\rho) u_1 x_1 & u_7 &= \lambda_p x_{31} \\ u_4 &= \beta & u_8 &= \gamma_8 - \gamma_9 \exp(-\nu_5 x_{36}) \\ u_5 &= \lambda_c x_{11} & u_9 &= u_8 x_{33} \\ u_{10} &= \gamma_{10} \exp\big(-\nu_6 \cdot \big(g_1 x_1 + g_2 \cdot (x_2 + \ldots + x_{11}) + g_3 \cdot (x_{22} + \ldots + x_{34})\big)\big) \\ u_{11} &= \gamma_{11} \exp(-\nu_7 x_{34}) \end{aligned} \tag{2}$$

lead to a state-space model, which is described by the coupled nonlinear ODEs

$$\begin{aligned} \dot{x}_1 &= (2\rho - 1) u_1 x_1 \\ \dot{x}_2 &= u_3 - \lambda_c x_2 + u_2 x_2 - u_4 x_2 + \Phi x_{12} \\ \dot{x}_i &= \lambda_c x_{i-1} - \lambda_c x_i + u_2 x_i - u_4 x_i + \Phi x_{i+10} && \text{for} \quad i = 3, \ldots, 11 \\ \dot{x}_i &= u_4 x_{i-10} - \Phi x_i && \text{for} \quad i = 12, \ldots, 21 \\ \dot{x}_{22} &= u_5 + u_6 x_{22} - \lambda_p x_{22} \\ \dot{x}_i &= \lambda_p x_{i-1} + u_6 x_i - \lambda_p x_i && \text{for} \quad i = 23, \ldots, 31 \\ \dot{x}_{32} &= u_7 - \lambda_M x_{32}, \quad \dot{x}_{33} = \lambda_M x_{32} - u_8 x_{33}, \quad \dot{x}_{34} = u_9 - \lambda_F x_{34} \\ \dot{x}_{35} &= u_{10} - \lambda_{RegI} x_{35}, \quad \dot{x}_{36} = u_{11} - \lambda_{RegII} x_{36}. \end{aligned} \tag{3}$$

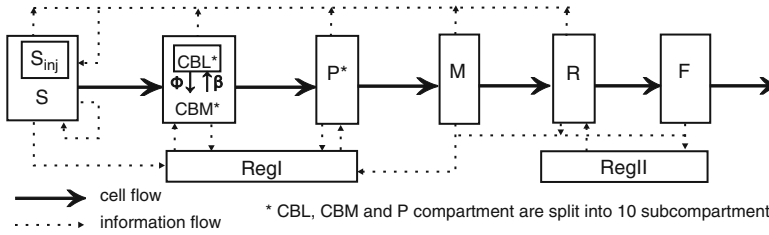


Fig. 1. The model of granulopoiesis according to Fliedner and Steinbach

In order to quantify the influence of parameter uncertainties on the state variables x_i as precisely as possible, suitable constraints are introduced which allow to detect and to reduce overestimation that is caused by interval simulations. In this application, the two compartments CBL and CBM are coupled directly. Constraints H_V are introduced that are independent of the parameters ϕ and β, which express the exchange of cells between CBL and CBM.

Physical constraints are calculated in two mathematically different ways in the ODE solver ValEncIA-IVP, see Sect. 3. The first method is to directly evaluate the constraints using a guaranteed state enclosure for $x = [x_1, \ldots, x_{36}]^T$ according to

$$H_{H,l} := x_{36+l} = x_{l+1} + x_{l+11} \tag{4}$$

with $l = 1, \ldots, 10$, by substituting the interval enclosures $[x_{l+1}]$, $[x_{l+11}]$ for the state variables. This constraint is denoted by H_H. The second way is to calculate the constraint by integration of the corresponding time derivatives. The ValEncIA-IVP constraint H_V is the solution of the additional ODEs

$$\dot{H}_{V,l} := \begin{cases} \dot{x}_{37} = u_3 - \lambda_c x_2 + u_2 x_2 & \text{for} \quad l = 1 \\ \dot{x}_{36+l} = \lambda_c x_{l+1} - \lambda_c x_{l+2} + u_2 x_{l+2} & \text{for} \quad l = 2, \ldots, 10. \end{cases} \tag{5}$$

The initial conditions $H_{V,l}(0) = H_{H,l}(0)$ for $l = 1, \ldots, 10$ are determined for the given enclosure $[x(0)]$ of the initial states $x(0)$ using equation (4). Considering these two compartments, up to $m = 10$ constraints can be identified for the blood cell system. Since the expressions (4) and (5) are two different mathematical formulations for the same physical quantities, however evaluated with different types of overestimation, they can be applied to detect and to reduce overestimation by the procedure described in Sects. 3 and 4.

3 ValEncIA-IVP

ValEncIA-IVP is a validated solver for IVPs for ODEs [6]. It has the ability to work with interval variables which result from propagation of uncertain

initial conditions and uncertain parameters. VALENCIA-IVP calculates guaranteed state enclosures

$$[x_{\text{encl}}(t)] := x_{\text{app}}(t) + [R(t)] \quad \text{for} \quad t \in [t_0; t_f] \quad \text{with} \quad x(t_0) \in [\underline{x}_0 \; ; \; \overline{x}_0] \tag{6}$$

via a two-stage approach. First, a suitable approximate solution $x_{\text{app}}(t)$ is computed using arbitrary non-validated ODE solvers, e.g. relying on explicit/implicit Euler methods or Runge–Kutta methods. Then, validated error bounds $[R(t)]$ are determined using an iteration procedure which can be derived using Banach's fixed-point theorem. A detailed derivation and proof of this two stage approach is found in [6].

Overestimation which is contained in the guaranteed state enclosure $[x_{\text{encl}}(t)]$ is to be reduced by consistency tests. The quantification of the influence of uncertainties and parameter variations on the dynamics of the blood cell system is better, the tighter the exact solution is enclosed.

4 Branch and Bound Algorithm

The constraints that have been identified for the blood cell model in equations (4) and (5) are used to exclude subdomains of the state enclosure computed by VALENCIA-IVP that are physically meaningless since they result from overestimation. A modified Branch and Bound algorithm has been implemented in this work as a consistency test. The Branch and Bound algorithm identifies intervals that are conform with the VALENCIA-IVP constraint H_V, which can be viewed as an optimization function. For each subdomain, the constraint H_H is re-evaluated using subintervals of the state enclosures to test if they are either within, partially outside, or completely outside the allowed range which is given by the guaranteed enclosure $[H_V]$. These intervals are distinguished as `true`, `undecided`, or `false` branches. The branches that are `undecided` or `true` are written in a list $\mathcal{L}$, so that at any time the list $\mathcal{L}$ is a validated solution of the problem [2].

In the case of multiple constraints $m > 1$, a branch is discarded if at least one constraint leads to a `false` interval. It is classified as an `undecided` interval if the constraints corresponds to a mixture of `true` intervals and at least one `undecided` interval, but no `false` ones.

An optimal time to apply the Branch and Bound algorithm has to be defined to detect and reduce as much overestimation as possible. The area expressed in terms of the constraints that is caused by overestimation is called reduction area

$$\mathcal{RA} := \text{diam}\{[H_H]\} - \text{diam}\,\{[H_H] \cap [H_V]\} \overset{!}{\geq} 0\,. \tag{7}$$

The goal is to calculate an optimal point of time $t^* \in [t_0; t_f]$ such that the reduction area $\mathcal{RA}$ is maximized. To simplify the optimization problem, the first local maximum of $\mathcal{RA}$ is used for t^*.

Several different subdivision strategies have been implemented in ValEncIA-IVP, they can be found in detail in [5]. The subdivision occurs at the midpoint for a single component x_{j^*}, $j^* \in \{1, \ldots, n\}$, of the state vector x, while all other components are identical to the original box. In the following application, the component j^* is determined according to

$$j^* = \underset{i=1,\ldots,n}{\arg\max} \left(\max_{l=1,\ldots,m} \left\{ \operatorname{abs}\left\{ \left. \frac{\partial H_l}{\partial x_i} \right|_{x=[x]} \right\} \cdot \operatorname{diam}\{[x_i]\} \right\} \right). \tag{8}$$

5 Simulation Results for the Blood Cell Model

For the dynamic system of granulopoiesis, simulations of both the nominal and uncertain system model are investigated. The nominal values for the initial states and parameters can be found in [4]. In order to identify overestimation efficiently, the constraints have to be chosen such that they are sensitive toward the parameters that introduce overestimation. The following simulation investigates uncertainties of $\pm 10\%$ for both parameters ϕ and β on which the constraint H_H depends through x_{l+1} and x_{l+11}, $l = 1, \ldots, 10$.

The optimal time t^* is calculated to $t^* = 6.0\,\mathrm{h}$ at which the subdivision strategy is evaluated for 1,000 subdivisions. In Fig. 2 the state enclosure of $x_2(t)$ is shown for both nominal and uncertain parameters. Especially for uncertain parameters, the consistency test significantly reduces overestimation at $t = t^*$. This leads to tighter interval bounds if the simulation is continued for $t > t^*$. In Fig. 3 it is depicted that the dependency of the number of blood cells in the function compartment with respect to uncertainties in the parameters ϕ and β is negligible in the time horizon under consideration. Therefore, techniques to simplify the system model or to reduce its order can be studied in future work to derive a mathematical description with identical input/output behavior which can be evaluated with less computational effort.

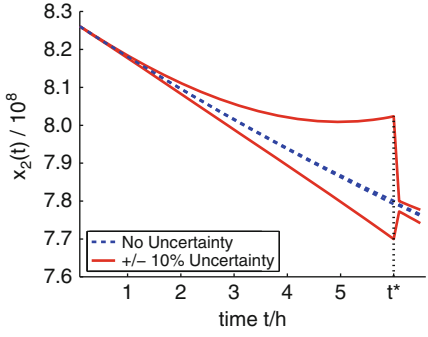


Fig. 2. Interval enclosure of $x_2(t)$, cell count of the CBM_1 compartment

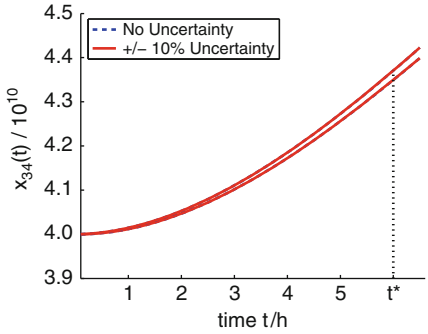


Fig. 3. Interval enclosure of $x_{34}(t)$, cell count of the F compartment

6 Conclusion

In the preceding sections, newly implemented physically motivated techniques for detection and reduction of overestimation in interval simulations of dynamic systems with uncertainties were introduced, analyzed, and discussed. ValEncIA-IVP is extended with a Branch and Bound algorithm which is based on the computationally efficient evaluation of physical constraints. Future work will deal with an automatic identification of suitable physically motivated constraints from the dynamical system model. The goal is to maximize the sensitivity of the constraints toward variables that introduce overestimation. This step will help to improve routines which are capable of accurately enclosing state variables of uncertain dynamical systems. This is the prerequisite for a software environment in which parameters of mathematical system models are identified. A further application is the quantification of approximation errors that arise in the reduction of the order of dynamical systems in a guaranteed way.

References

1. Clausen, J.: Branch and Bound Algorithms: Principles and Examples. `citeseer.ist.psu.edu/683497.html` (1999)
2. de Figueiredo, L.H., van Iwaarden, R., Stolfi, J.: Fast Interval Branch-and-Bound Methods for Unconstrained Global Optimization with Affine Arithmetic. Technical Report IC-97-08, Institute of Computing, Univ. of Campinas, Brazil, June 1997
3. Hofer, E.P., Fan, Y., Tibken, B.: Extraction of Rules for Model Based Estimation of Granulocytopoiesis. In: Frik, M. (ed.) 5th German-Japanese Seminar Nonlinear Problems in Dynamical Systems - Theory and Applications, pp. 58–68. Daun, Vulkaneifel (1991)
4. Hofer, E.P., Tibken, B., Fliedner, T.M.: Modern Control Theory as a Tool to Describe the Biomathematical Model of Granulocytopoiesis. In: Möller, D.P.H.,

Richter, O. (eds.) Analyse dynamischer Systeme in Medizin, Biologie, Ökologie, vol. 275, pp. 33–39. Springer, Berlin (1991)
5. Freihold, M., Hofer, E.P.: Derivation of Physically Motivated Constraints for Efficient Interval Simulations Applied to the Analysis of Uncertain Dynamical Systems. Int. J. Appl. Math. Comput. Sci. AMCS, vol. 19, No. 3, pp. 485–499, 2009
6. Rauh, A., Auer, E., Hofer, E.P.: ValEncIA-IVP: A Comparison with Other Initial Value Problem Solvers. In: CD-Proc. of the 12th GAMM-IMACS International Symposium on Scientific Computing, Computer Arithmetic, and Validated Numerics SCAN 2006, Duisburg, Germany, IEEE Computer Society (2007)

Application of M⟦⟧BILE for Accurate Bone Motion Reconstruction Using Motion-Measurements and MRI Measurements

M. Tändl, T. Stark, and A. Kecskeméthy

University of Duisburg-Essen, Lotharstraße 1, 47057 Duisburg, Germany,
laumanns@ifor.math.ethz.ch

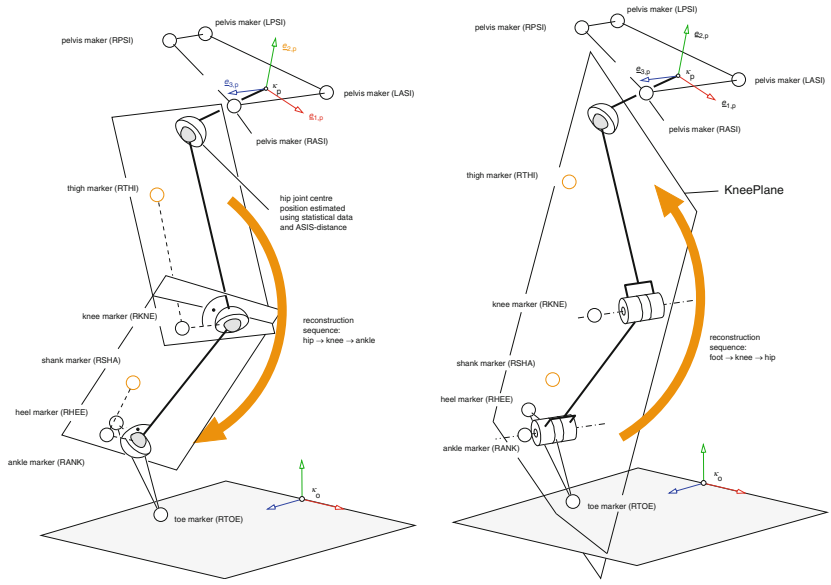
Summary. Accurate reconstruction of bone motion using marker tracking is still an open issue in biomechanics. In this paper a novel approach for gait motion reconstruction is presented that is based on the analysis of kinematical loops and the reconstruction of functional skeleton features from segmented MRI data. The method uses an alternative path for concatenating relative motion, starting at the feet and closing at the hip joints. The discrepancies between predicted and geometrically identified functional data, such as knee axis and hip joint centers, gives rise to a cost function, which is the basis for model addaptation. Computations are performed with the object-oriented library M⟦⟧BILE.

1 Motivation

The Vicon motion capture systems (www.vicon.com) is a widely used tool for the determination of motions of body segments upon the motion of skin-mounted markers. The "Plug-in-Gait marker model" is the marker setup used in the current version of the Vicon analysis software Nexus. Figure 1(a) shows the right leg with the markers used for this leg motion reconstruction model. For this model, a pelvis-fixed frame $\mathcal{K}_\mathrm{p}$ is defined using the hip markers RASI, LASI, RPSI, LPSI, such that the z-axis is aligned with the line connecting the markers RASI and LASI, and the centroid of the RPSI and LPSI markers lies in the xz-plane. For the pelvis, the position of hip joint centers is estimated using the Newington-Gage model [2] making use of the inter-ASI distance. The motion of the other leg segments is reconstructed starting at the hip joints, advancing down to the foot, sequentially computing new segment orientations using joint centers already determined and markers fixed to the next body part, according to the algorithm described in [8].

In the application of this procedure, accuracy problems can occur manifesting in wrong hip joint center positions and large (axial) rotations of tibia and femur. These errors result from inaccurate marker placement, skin motion at knee and ankle or soft tissue artifacts of the thigh marker (RTHI). These

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*,
Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_90,



a) VICON "Plug-in-Gait" model b) suggested kinematic model

Fig. 1. Kinematic structure of the Plug-in-Gait model and the proposed model

errors could be reduced by careful marker placement and performing new measurements with improved marker positions, but this is hardly possible in routine measurements with patients.

Approaches for reducing these errors after the measurement are presented in [1], and in [5], where the latter reduces model bone length variations by identifying the unknown constant offsets between the joint centres delivered by the motion capture system (prediction) and the anatomical joint centres.

In the approach presented in this paper the axes of the ankle joint and the knee joint are computed starting at the foot markers, and proceeding upward to the knee, using foot- and knee markers, which are subject to less (or at lest more predictable) soft tissue motion. It is assumed that the ankle joints $\mathrm{R}_{\mathrm{ankle}}$ and the knee joints $\mathrm{R}_{\mathrm{knee}}$ can be represented by revolute joints. Another assumption is that the line connecting the heel and the toe marker (unit vector $\underline{u}_{\mathrm{foot}}$) is perpendicular to both the ankle joint axis and the knee joint axis (Fig. 2a). An approximation of the hip joint center position is computed at the end of the procedure in a way that the relative motion of one femur point with respect to the pelvis is minimized. The segment motion is determined with respect to an inertially fixed frame $\mathcal{K}_0$, usually coinciding with the reference frame of the motion capture system. For describing vectors using cartesian coordinate frames, the notation ${}^{k}_{i}\underline{b}_{j}$ is used, where k denotes the frame of decomposition, i denotes the frame with respect to which the motion is measured, and j denotes the target frame. For motion measured

with respect to $\mathcal{K}_0$, the index $i = 0$ is omitted. Likewise, for decompositions in the target frame $k = j$, the index k is omitted. Hence $\underline{b}_1$ is equivalent ${}^1_0\underline{b}_1$.

2 Simulation Environment

The core component of the integrated simulation environment MobileBody© [7] is the mechanical model of the musculoskeletal system. Its implementation using object oriented programming makes it easy to combine it with image processing code or visualization libraries. Furthermore, by using the multibody simulation library M⛶BILE [4], model components (e.g. joints, muscles) can be easily replaced with more complex and realistic implementations.

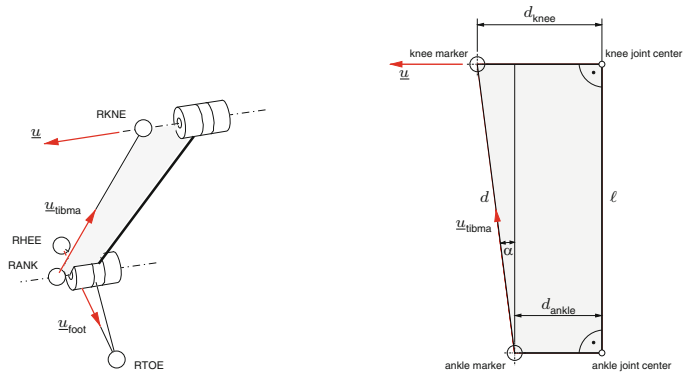
3 Segment Motion Estimation Procedure

Having measured the anthropometric distances d_{knee} and d_{ankle} between knee joint centre and ankle joint centre, and the corresponding marker (see Fig. 2b), the following simple procedure is used to determine estimates of tibia- and femur-fixed coordinate frames. By computing the distance between the ankle- and knee marker, the angle α between the connecting line of the tibia markers and the tibia joint centers is computed using the formula

$$\alpha = \arcsin \frac{d_{\text{knee}} - d_{\text{ankle}}}{d}. \tag{1}$$

With the unit direction vector $\underline{u}_{\text{tibma}}$ of the line connecting the tibia markers, the axis direction of knee- and ankle joint becomes

$${}^0\underline{u} = \cos\alpha \frac{{}^0\underline{u}_{\text{foot}} \times {}^0\underline{u}_{\text{tibma}}}{\| {}^0\underline{u}_{\text{foot}} \times {}^0\underline{u}_{\text{tibma}} \|} + \sin\alpha \, {}^0\underline{u}_{\text{tibma}} \quad . \tag{2}$$



a) Direction of foot and tibia markers b) Estimation of tibia joint centers

Fig. 2. Reconstruction of the knee joint axis

Next, with ${}^0\underline{u}$ and the position ${}^0\underline{r}_{\text{knema}}$ of the knee marker, the location of the knee joint center is calculated as

$$ {}^0\underline{r}_{\text{knee}} = {}^0\underline{r}_{\text{knema}} - d_{\text{knee}}\,{}^0\underline{u} \tag{3} $$

and a coordinate system can be aligned with the femur by setting

$$ \begin{aligned} {}^0\underline{e}_{3,\text{f}} &= {}^0\underline{u} \\ {}^0\underline{e}_{1,\text{f}} &= \frac{({}^0\underline{r}_{\text{tibma}} - {}^0\underline{r}_{\text{knee}}) \times {}^0\underline{u}}{\| ({}^0\underline{r}_{\text{tibma}} - {}^0\underline{r}_{\text{knee}}) \times {}^0\underline{u} \|} \\ {}^0\underline{e}_{2,\text{f}} &= {}^0\underline{u} \times {}^0\underline{e}_{1,\text{f}} \end{aligned} \quad . $$

This leads to the rotation matrix

$$ {}^0\mathbf{R}_{\text{f}} = \left[{}^0\underline{e}_{1,\text{f}},\, {}^0\underline{e}_{2,\text{f}},\, {}^0\underline{e}_{3,\text{f}} \right] \tag{4} $$

corresponding to the femur frame $\mathcal{K}_{\text{f}}$ displayed in Fig. 3. For estimation of the hip joint centre, let the position of the femur head ($\equiv$ hip joint centre) be given relatively to the femur frame $\mathcal{K}_{\text{f}}$ by the vector

$$ {}^{\text{f}}_{\text{f}}\underline{r}_{\text{hj}} = \begin{bmatrix} x_1 \\ x_2 \\ 0 \end{bmatrix} \tag{5} $$

which lies in the thigh-shank plane Fig. 3. Relatively to the pelvis frame, the coordinates of the femur-fixed point H representing the hip joint centre are

$$ {}^{\text{p}}_{\text{p}}\underline{r}_{\text{hj},i} = \mathbf{R}_i\, {}^{\text{f}}_{\text{p}}\underline{r}_{\text{f},i} + \mathbf{R}_{12,i}\,\underline{x} \tag{6} $$

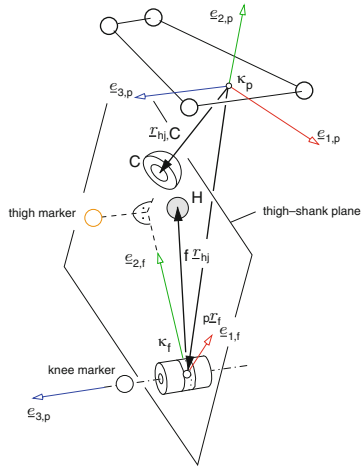


Fig. 3. Estimation of the hip joint centre location in the thigh-shank plane

where $\mathbf{R}_i = {}^{\mathrm{p}}\mathbf{R}_{\mathrm{f},i}$ is the rotation matrix between the estimated pelvis frame $\mathcal{K}_\mathrm{p}$ and the estimated femur frame $\mathcal{K}_\mathrm{f}$ at time t_i. Likewise, $\mathbf{R}_{12,i}$ is the matrix of the first two columns of $\mathbf{R}_i$ and $\underline{x} = [\, x_1 \,,\, x_2 \,]^{\mathrm{T}}$. The position of the femur frame relative to the pelvis frame at time t_i is given by ${}^{\mathrm{f}}_{\mathrm{p}}\underline{r}_{\mathrm{f},i}$. According to the "centre transformation technique" [3], the coordinates x_1 and x_2 are chosen such that the squared norm of errors

$$\mathrm{f}(\underline{x}) = \sum_{i=1}^{m} ({}^{\mathrm{p}}_{\mathrm{p}}\underline{r}_{\mathrm{hj},i} - {}^{\mathrm{p}}_{\mathrm{p}}\underline{r}_{\mathrm{hj,C}})^2 \tag{7}$$

between the points H_i of this trajectory and their centroid C (visualized as "cup" in Fig. 3)

$${}^{\mathrm{p}}\underline{r}_{\mathrm{hj,C}} = \frac{1}{m} \sum_{i=1}^{m} (\mathbf{R}_i {}^{\mathrm{f}}_{\mathrm{p}}\underline{r}_{\mathrm{f},i} + \mathbf{R}_{12,i}\, \underline{x}) = \underline{a} + \mathbf{B}\, \underline{x} \tag{8}$$

is minimized over all m measured poses summing over $i = 1, \ldots, m$. By summing up and factoring out $\underline{x}$ one obtains

$${}^{\mathrm{p}}\underline{r}_{\mathrm{hj,C}} = \underline{a} + \mathbf{B}\, \underline{x} \tag{9}$$

with $\underline{a}$ and $\mathbf{B}$ being constant. The first order conditions for the optimum become

$$\frac{\partial \mathrm{f}}{\partial \underline{x}}(\underline{x}) = \sum_{i=1}^{m} (\mathbf{R}_{12,i} - \mathbf{B})^{\mathrm{T}} \big((\mathbf{R}_i {}^{\mathrm{f}}_{\mathrm{p}}\underline{r}_{\mathrm{f},i} - \underline{a}) + (\mathbf{R}_{12,i} - \mathbf{B})\, \underline{x} \big) = \underline{a}_1 + \mathbf{B}_1\, \underline{x} = \underline{0} \tag{10}$$

with shortcuts $\underline{a}_1$ and $\mathbf{B}_1$, leading to the optimal point

$$\underline{x}^\star = -\mathbf{B}_1^{-1} \underline{a}_1. \tag{11}$$

When choosing the position of the hip joint relatively to the femur frame as

$${}^{\mathrm{f}}_{\mathrm{f}}\underline{r}^\star_{\mathrm{hj}} = \begin{bmatrix} x_1^\star \\ x_2^\star \\ 0 \end{bmatrix}, \tag{12}$$

the oscillation of the hip joint center with respect to the pelvis frame is minimized. This vector and the positions of the hip joint relatively to the pelvis frame are used to define the segment lengths of an open-chain kinematic model of the leg such as that described in [5]. The joint coordinates of the hip joint are obtained from the relative rotation matrix ${}^{\mathrm{p}}\mathbf{R}_{\mathrm{f},i}$. Analogously, the segment lengths of the shank and the joint coordinates in knee and ankle are computed.

4 Discussion

The described model displays a simplified kinematic model of the knee- and ankle joint, neglecting any axial rotation in these joints. This restricts its application to gait motion with low knee flexion angles, since larger external/internal rotation (up to 37°) in the knee are possible for large flexion angles according to a survey in [6]. In standard gait analysis, the model avoids huge, unrealistic internal/external rotations of the tibia and the femur, without requiring more markers than when using the "Plug-In-Gait model". On the other hand, the actual rotations in knee and ankle are not reflected in this model, which may lead to an inaccurate identification of the hip joint center if the distance between knee marker and knee joint center is not known precisely or in the case of valgus/varus deformities. If the distance between the hip joints is known from X-Ray or MRI images, this information can be exploited to improve the estimation of the hip joint centre in transversal direction, which is the focus of ongoing work.

Measurements are currently being produced and will be published in future.

References

1. Cerveri, P., Pedotti, A., Ferrigno, G.: Kinematical models to reduce the effect of skin artifacts on marker-based human motion estimation. J. Biomech. **38**(11), 2228–2236 (2005)
2. Davis, R.B. III, Õunpuu, S., Tyburski, D., Gage, J.R.: A gait analysis data collection and reduction technique. Hum. Motion Sci. **10**(5), 575–587 (1991)
3. Ehrig, R.M., Taylor, W.R., Duda, G.N., Heller, M.O.: A survey of formal methods for determining the centre of rotation of ball joints. J. Biomech. **39**(15), 2798–2809 (2006)
4. Kecskeméthy, A., Hiller, M.: An object-oriented approach for an effective formulation of multibody dynamics. Comput. Methods Appl. Mech. Eng. **115**, 287–314 (1994)
5. Kecskeméthy, A., Stolz, M., Strobach, D., Saraph, V., Steinwender, G., Zwick, B.: Improvements in measure-based simulation of the human lower extremity. In: Proceedings of the IASTED Conference on Biomechanics, pp. 155–160. Rhodes, Greece, June 30–July 2 2003
6. Piazza, S.J., Cavanagh, P.R.: Measurement of the screw-home motion of the knee is sensitive to errors in axis alignment. J. Biomech. **33**(8), 1029–1034 (2000)
7. Raab, D., Stark, T., Erol, N.E., Löer, F., Tändl, M., Straßmann, T., Kecskeméthy, A.: An integrated simulation environment for human gait analysis and evaluation. In: Proceedings of the 10th International Symposium Biomaterials: Fundamentals and Clinical Applications, Essen, Germany, 2008
8. Vicon Motion Systems Limited: Plug-in-Gait Marker Placement – Documentation, 2007

Toward Verified Modelling and Simulation of Closed Loop Systems in SmartMOBILE

E. Auer

Faculty of Engineering, NKO, University of Duisburg-Essen, D-47048 Duisburg, Germany, auer@inf.uni-due.de

Summary. Software for modelling and simulation (MSS) of mechanical systems helps to reduce production costs for industry. Usually, such software relies on (possibly erroneous) finite precision arithmetic and does not take into account uncertainty in the input data. The program SmartMOBILE enhances the existing MSS MOBILE with verified techniques to provide a guarantee that the obtained results are correct and measure the influence of data uncertainty. In this paper, particular attention is paid to the current strivings toward verified modelling and simulation of closed loop systems.

1 Introduction

Three major phases during modelling and simulation process are analysis, implementation (verification) and simulation (validation) [5]. The first step is to *analyze* the real world problem and to design a formal model of the system under consideration. The second is to *implement* this model. Usual tasks at this stage include code verification (finding logical and programming errors in the code) and numerical verification (minimizing numerical errors in results). The final step is *validation* during which model fidelity is established.

Modern numerical modelling and simulation software (MSS) such as MOBILE [2] automatizes parts of this cycle and in this way accelerates the development process for a product reducing production costs. However, some of the tasks are not covered. For example, MSS usually relies on floating point arithmetic, which either shifts the task of result verification into the validation stage of the cycle or leaves that question unanswered.

Development of methods for numerical result verification is the research area of the whole branch of numerics also known as "interval", "validated" or "verified" arithmetic. Such methods not only provide a guarantee that the obtained results are correct but also propagate initial data uncertainty almost as their by-product. A recently developed tool SmartMOBILE [1] interfaces libraries for result verification with MOBILE to be able to cover more tasks

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*, Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_91,

from the modelling and simulation cycle. At the stage of analysis, it offers techniques helping to take into account model errors and to perform uncertainty analysis in general. At the implementation stage, it provides result verification for kinematics and dynamics of various mechanical systems. Finally, the use of algorithmic differentiation and newly developed methods for sensitivity analysis identifies critical parameters and in this way makes the stage of validation easier.

In this paper, we focus on the uses of SmartMOBILE for accurate (and, in some cases, verified) simulation of kinematics and dynamics of closed loop systems. We begin by describing main features of SmartMOBILE which include free choice of underlying arithmetic. Further, we demonstrate options for result verification for models of closed loop systems. Such simulations are especially difficult since they have differential-algebraic equations (DAEs) as their basis. Finally, we recapitulate the main results.

2 Main Features of SmartMOBILE

SmartMOBILE is an object-oriented software for verification of mechanical systems based on MOBILE which employs floating point numerics. Models in both tools are executable C++ programs built of the supplied classes for transmission elements such as rigid links, for scalar or spatial objects such as reference frames and for solvers such as those for differential equations.

SmartMOBILE is one of the first integrated environments providing result verification for kinematical and dynamic simulations of mechanical systems. The advantage of this environment is its flexibility due to its template structure: the user can choose the kind of (non)verified arithmetics according to his task. Intervals and Taylor model arithmetics are currently available in SmartMOBILE. However, advanced users are not limited to them and are free to plug in their own implementations.

Although switching from MOBILE to SmartMOBILE is easy, users are assisted by converters in this process. However, automatically generated elements might require a heuristic improvement by the user, if they contain code fragments transformation of which cannot be algorithmized, for example, non-verified equation solvers.

For most kinematical problems, it is sufficient to use the supplied basic data types (e.g. intervals) as parameters to all the template classes for a particular model. The main idea for dynamic and special kinematical tasks such as finding of system equilibria is to use pairs basic data type/corresponding solver. Here, the basic data type should be constructed in such as way as to allow us to apply the given solver. That is, it should automatically deliver, for example, all the derivatives the solver requires.

3 Options for Simulation of Closed Loop Systems in SmartMOBILE

There are several options in MOBILE to simulate kinematics and dynamics of closed loop systems. We single out an aspect of this complicated process which is important for the following considerations.

Mathematical models behind this type of problems are systems of DAEs. Since the index of such systems is usually equal to three, common IVP solvers for DAEs such as Dassl cannot handle them as they are. This fact is one of the reasons why the original system of DAEs is automatically transformed into an equivalent system of ODEs in MOBILE. Two transmission elements are developed for this purpose. `MoExplicitConstraintSolver` handles systems with one or two constraints of a certain form where explicit solution of the corresponding algebraic system is possible. `MoImplicitConstraintSolver` uses Newton's method to obtain solutions to arbitrary (nonlinear) systems of algebraic equations.

As reported in [1], it is possible to verify the kinematics and dynamics of closed loop systems in SmartMOBILE by using `TMoExplicitConstraintSolver`. As for the second element, there exists a version of it called `MoIImplicitConstraintSolver` using Newton-Gauss-Seidel or Krawczyk methods to verify kinematics of systems modelled with it. The implementation of the latter element for dynamics seems impracticable since all iterations of a verified zero-finding method would have to be taken into the algorithmic differentiation graph for computing derivatives, which still cannot be handled satisfactorily by the software.

An alternative is to solve the DAE system directly. Unfortunately, verified solution of IVPs to DAEs is a very new research area. One tool available to us is an extension of ValEncIA-IVP which is still under development. However, the first results [4] are promising. Since this solver requires an approximation of the DAE solution in its first stage, an accurate solver for this purpose should be integrated into SmartMOBILE.

4 Equations of Motion for a Spatial Four Bar Mechanism with Result Verification

Four bar mechanisms are simplest closed loop systems relevant for real life applications. In this section, we consider the one shown in Fig. 1, left side. This closed system consists of two revolute joints `R1` and `R2`, a double-revolute joint modelled by two joints `R3` and `R4`, a spherical joint `S1`, and four rigid links `base`, `link_1`, `link_2`, and `coupler` between them. To model this task, the loop is dissected at the body `coupler` (cf. Fig. 1, right side). The closure condition `core` is the equality of the corresponding displacements and rotations for the reference frames `K7` and `K10`. Usually, `core` is an instance of the measurement object `MoChord3DPose`.

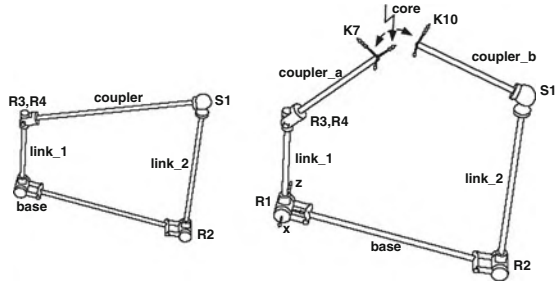


Fig. 1. The iconic model of a spatial four bar mechanism

For this type of closure conditions, we employ `TMoIImplicitConstraintSolver` in SmartMOBILE. The task is to find the mass matrix and the force for this system with result verification. However, this example shows more than just the possibility of verification. Using it, we can compare the method of obtaining derivatives of a function by algorithmic differentiation to the one based on physical considerations.

If we compute the Jacobian of the goal function in the interval version by using the force-based method supplied by MOBILE, the enclosure is equal to

$$\begin{pmatrix} [\pm 10^{-2}] & [-1.1;-0.9] & [\pm 10^{-4}] & [\pm 10^{-3}] & -[1.1] & [0;0] \\ [-0.8;1.8] & [\pm 10^{-3}] & [-0.7;-0.2] & [-1.3;-0.5] & [\pm 10^{-3}] & [0;0] \\ [0.5;3.1] & [0;0] & [0.9;1.1] & [0.2;1.0] & [\pm 10^{-3}] & [0;0] \\ [\pm 10^{-3}] & [0.2;0.6] & [\pm 10^{-3}] & [\pm 10^{-3}] & [0.0;1.1] & [0.3;1.4] \\ [\pm 10^{-3}] & [0.8;1.0] & [\pm 10^{-3}] & [\pm 10^{-3}] & [-1.4;-0.3] & [0.0;1.1] \\ [0.3;1.4] & [\pm 10^{-3}] & [-1.4;-0.3] & [0.3;1.4] & [\pm 10^{-3}] & [\pm 10^{-3}] \end{pmatrix}.$$

Here, the numbers are rounded up to the first digit after the decimal point. For the same parameter values, the enclosure of the Jacobian obtained with algorithmic differentiation is much tighter:

$$\begin{pmatrix} [0.0] & -[1.0] & [0.0] & [0.0] & -[1.1] & [0.0] \\ [1.5] & [0.0] & -[0.5] & -[0.5] & [0.0] & [0.0] \\ [1.0] & [0.0] & [1.0] & [1.0] & [0.0] & [0.0] \\ [0.0] & [0.5] & [0.0] & [0.0] & [0.0] & [1.0] \\ [0.0] & [0.9] & [0.0] & [0.0] & -[1.0] & [0.0] \\ [1.0] & [0.0] & -[1.0] & [1.0] & [0.0] & [0.0] \end{pmatrix}.$$

The notation [*number*] means that an enclosure of a *number* with a diameter of at most 10^{-12} is obtained. Since this Jacobian is important not only for the zero-finding method but also for correct computation of velocities and accelerations inside the implicit solver, it is crucial to obtain its tight enclosure.

It was possible to enclose the mass matrix and the force in the spatial four bar mechanism, used later to obtain equations of motion, in tight intervals

[1.03043; 1.03043] and [−0.05104; −0.05104], respectively (rounded, the diameter is at most 10^{-12}). However, these are results for the case in which all parameters are chosen to be point intervals. In general, this model is prone to overestimation. This is confirmed by the relatively narrow search intervals which both Krawczyk and Newton-Gauss-Seidel methods require to be able to compute zeros.

5 An Accurate Direct DAE Solving Method in SmartMOBILE

In this Section, we show how the DAE system underlying closed loop models can be solved directly in SmartMOBILE. For this purpose, the integrator `TMoDAETSIntegrator` has been implemented recently. It is based on the solver DAETS which computes accurate floating point solutions to IVPs for DAEs [3]. One advantage of `DAETS` is that it solves the problem as it is without the user having to transform it into an ODE problem or eliminating higher order derivatives, regardless of the problem's index.

A new element `TMoMechanicalSystemDAE` (Martin Tändl) was developed to provide equations of motion in the form $g(q, q', t) = 0$ required by DAETS. Note that the mass matrix does not have to be inverted for this representation.

The four bar mechanism we consider as an example is simpler than that from Sect. 4. Now the system consists of two simple pendulums modelled with two revolute joints and two rigid links with masses. They are connected by the third rigid link. The instance of the measurement class `TMoChordPointPointQuadratic` helps to formulate closure conditions for the loop.

Note that the DAE-based system cannot be solved in MOBILE because it uses Dassl for DAE solving. Besides, the solver DAETS can be employed only in SmartMOBILE because this MSS version, as opposed to the usual one, supplies the necessary derivatives.

We simulated the above system in SmartMOBILE with the help of the usual ODE-based floating point method using the explicit solver and the Adams integrator, the accurate DAE-based method with `TMoDAETSIntegrator`, and the verified ODE-based method using the explicit solver and `TMoValencia-Integrator`. The solutions for the consistent initial conditions supplied by `TMoDEATSIntegrator` are shown in Fig. 2. The trajectories coincide as expected. Both of the non-verified solutions lie inside the obtained verified bounds. For this simple example, it is not possible to decide if the DAE-based method is more accurate than the ODE-based one, although the solver `DAETS` is reported to be so in more complicated cases [3].

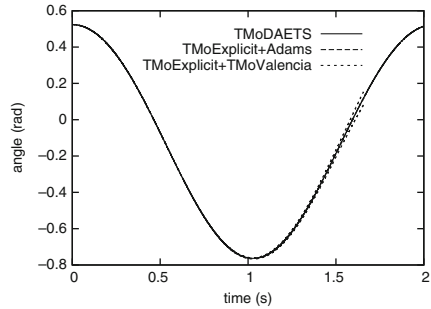


Fig. 2. The first angle of a four bar mechanism

6 Conclusions

In this paper, we presented the tool SmartMOBILE for guaranteed modelling and simulation of kinematics and dynamic of mechanical systems. With its help, the behavior of different classes of systems can be obtained with the guarantee of correctness, the option which is not given by tools based on floating point arithmetics. SmartMOBILE is flexible and allows the user to choose the kind of underlying arithmetics according to the task at hand.

A recent development concerned modelling and simulation of closed loop systems. New types of them were verified and a different kind of modelling was made possible.

Our short term task is the verification of the DAE-based approach to modelling and simulation of closed loop systems in SmartMOBILE using Valencia-IVP for DAEs as the basis. Almost all auxiliary components are already prepared for this purpose. First, `TMoMechanicalSystemDAE` generates the equations of motion in the required form. Next, a solver to compute the approximate solution is given by `TMoDAETSIntegrator`. Further, a verified zero-finding routine as required by ValEncIA-IVP for DAEs is already present in SmartMOBILE. Finally, a program for testing and computing consistent initial values is at the final stage of implementation. With these preparations, only the main DAE solving algorithm will have to be transferred.

References

1. Auer, E., Luther, W.: Proceedings of ICINCO **RA-1**, (2007)
2. Kecskeméthy, A., Hiller, M.: Comp. Meth. App. Mech. Eng. **115**, 287–314 (1994)
3. Nedialkov, N.S., Pryce, J.D.: J. Num. Anal. Ind. App. Math. **1**(1), 1–30 (2007)
4. Rauh, A., Auer, E., Minisini, J., Hofer, E.P.: Proc. App. Math. and Mech. **7**(1):1023001–1023002 (2007)
5. Schlesinger, S.: Simulation **32**:3, 103–104 (1979)

Reliably Safe Path Planning Using Interval Analysis

R. Pepy, M. Kieffer, and E. Walter

Laboratoire des Signaux et Systèmes, CNRS – SUPELEC – Univ Paris-Sud, 3 rue Joliot-Curie, 91192 Gif-sur-Yvette, France, pepy@lss.supelec.fr, kieffer@lss.supelec.fr, walter@lss.supelec.fr

Summary. This paper is devoted to path planning when the safety of the system considered has to be guaranteed in the presence of bounded uncertainty affecting its model. A new path planner addresses this problem by combining Rapidly-exploring Random Trees (RRT) and a representation of uncertain states by vectors of intervals. The resulting path planner is used for nonholonomic path planning in robotics.

1 Introduction

Robotics is a challenging and important field of application for validated numerical methods. Problems are often solved in this context by using local and random search algorithms, with no guarantee as to their results. A definite advantage of interval analysis is that it provides tools to obtain approximate but guaranteed results that take into account the uncertainty in the data and errors resulting from the finite nature of the representation of real numbers on a computer, see, e.g., [6].

This paper focuses on reliably safe path planning [9] using interval analysis. In robotics, our aim is to find a path for driving a vehicle (or, more generally, a robot) from an initial (potentially uncertain) state or *configuration* to a final desired configuration, despite the presence of uncertainty related to the model of the vehicle, to badly charted obstacles in the environment, etc. The control input and the corresponding paths (succession of states) that achieve this goal while eliminating the risk of collision are said to be *safe*.

Path planners involving Rapidly-exploring Random Trees (RRT) [10, 12] are widely used, since they allow an efficient exploration of configuration space. To the best of our knowledge, however, they do not provide any robustness to uncertainty. When taken into account, configuration uncertainty is usually described probabilistically, e.g., by a multivariate Gaussian probability density function [3, 8, 16]. The main drawback of path planners based on such a description is that the reliability of the path obtained may be guaranteed at best up to a given confidence level.

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*, Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_92,

Path planning may be facilitated by the presence of *relocalization zones*, in which the configurations become much more accurately known [2,4,13]. The price to be paid is the preparation of these relocalization zones, which are not a prerequisite here.

The problem is formalized in Sect. 2. A reliably safe path planner, Box-RRT, is then presented in Sect. 3. Starting from some uncertain initial configuration (represented by a vector of intervals or *box* [6,14]), Box-RRT aims at driving the vehicle to a final configuration set. Provided that the assumptions on the error bounds are not violated, if a path is found using this new path planner, it will be *guaranteed* to be safe. Section 4 applies Box-RRT to path planning for non-holonomic vehicles. Some conclusions are drawn in Sect. 5.

2 Reliably Safe Path Planning

Consider a vehicle described by the state equation

$$\frac{ds(t)}{dt} = f(s(t), u(t), w(t)), \tag{1}$$

where $s(t) \in \mathbb{S} \subset \mathbb{R}^n$ is the configuration of the vehicle, u belongs to $\mathcal{U}_{[u]}^{\Delta t}$, the set of piecewise-constant input functions over intervals of the form $[k\Delta t, (k+1)\Delta t[$, and bounded in $[u]$, with $\Delta t > 0$, and w belongs to $\mathcal{W}_{[w]}$, the set of random perturbation functions with values in $[w]$.

The configuration space $\mathbb{S}$ is partitioned into $\mathbb{S}_{\text{free}}$, to which the configuration is allowed to belong, and $\mathbb{S}_{\text{obs}} = \mathbb{S} \setminus \mathbb{S}_{\text{free}}$, to which it is not. At time $t = 0$, $s(0) \in [s_{\text{init}}] \subset \mathbb{S}_{\text{free}}$. The vehicle has to be driven to a given box of goal configurations $[s_{\text{goal}}] \subset \mathbb{S}_{\text{free}}$. Safe path planning amounts to determining $K > 0$ and a control input $u \in \mathcal{U}_{[u]}^{\Delta t}$ such that $\forall s \in [s_{\text{init}}]$ and $\forall w \in \mathcal{W}_{[w]}$, one has $s(K\Delta t) \in [s_{\text{goal}}]$ and $\forall t \in [0, K\Delta t]$, $s(t) \in \mathbb{S}_{\text{free}}$, where $s(t)$ is the solution of (1).

The main difficulty is that for a given $u \in \mathcal{U}_{[u]}^{\Delta t}$, the values of $s(t)$ consistent with all $s \in [s_{\text{init}}]$, $w \in \mathcal{W}_{[w]}$, and (1) belong to a set $\mathbb{S}_t$, the shape of which may be quite complex.

3 Box-RRT

To cope with an uncertain initial configuration and bounded state pertubations, the classical RRT path planner [10,12] has to be adapted to deal with sets. Dealing with general sets of $\mathbb{R}^n$ would be exceedingly difficult, even for the simplest uncertain state equations, so here, boxes will be used to wrap uncertain configurations sets $\mathbb{S}_t$ at each instant of time t. Boxes are quite simple sets, which may provide a very coarse description of complex-shaped sets. Using more accurate wrappers (ellipsoids, zonotopes, or union of interval vectors) may increase the number of problems to which solutions may be found, see [15].

Algorithm 5 Box-RRT($[s_{\text{init}}] \subset \mathbb{S}_{\text{free}}$, $[s_{\text{goal}}] \subset \mathbb{S}_{\text{free}}$, $\Delta t \in \mathbb{R}^+$, $\overline{K} \in \mathbb{N}$)

1: G.init($[s_{\text{init}}]$)
2: $i \leftarrow 0$
3: **repeat**
4: $[s_{\text{rand}}] \leftarrow$ random_box($\mathbb{S}_{\text{free}}$)
5: $[s_{\text{new}}] \leftarrow$ Box-RRT_extend(G, $[s_{\text{rand}}]$, Δt)
6: **until** $i{+}{+} > \overline{K}$ **or** ($[s_{\text{new}}] \neq \emptyset$ **and** $[s_{\text{new}}] \subset [s_{\text{goal}}]$)
7: **return** G

Algorithm 6 Box-RRT_extend(G, $[s_{\text{rand}}]$, Δt)

1: $[s_{\text{near}}] \leftarrow$ nearest_neighbor(G, $[s_{\text{rand}}]$)
2: $u \leftarrow$ select_input($[s_{\text{rand}}]$, $[s_{\text{near}}]$)
3: $[s_{\text{new}}] \leftarrow$ prediction($[s_{\text{near}}]$, u, Δt)
4: **if** collision_free_path($[s_{\text{near}}]$,$[s_{\text{new}}]$,u,Δt) **then**
5: G.add_guaranteed_node($[s_{\text{new}}]$)
6: G.add_guaranteed_edge($[s_{\text{near}}]$, $[s_{\text{new}}]$, u)
7: **return** $[s_{\text{new}}]$
8: **end if**
9: **return** $\emptyset$

The principle of Box-RRT is given in Algorithms 5 and 6. Box-RRT aims at generating iteratively a graph G consisting of nodes associated with boxes in configuration space. At each iteration, a box $[s_{\text{rand}}] \subset \mathbb{S}_{\text{free}}$ is chosen at random. The node $[s_{\text{near}}]$ of G that is the closest to $[s_{\text{rand}}]$ according to some metric d, here the Hausdorff distance [1], is then selected by the function `nearest_neighbor`. Assume that $[s_{\text{near}}]$ is associated with time $k\Delta t$. A control input $u_k \in [u]$ is chosen (for instance at random, in a finite set of options). A box $[s_{\text{new}}] = [s_{k+1}]$ containing all possible configurations at time $(k+1)\Delta t$ if the configuration was in $[s_{\text{near}}]$ at time $k\Delta t$ for a constant input u_k over $[k\Delta t, (k+1)\Delta t]$ and a perturbation w that can take any value in $\mathcal{W}_{[w]}$ is computed by the set `prediction` function involving guaranteed numerical integration for uncertain systems, see, e.g., [7] and the references therein. Finally, the collision test that guarantees the reliability of every path between $[s_{\text{near}}]$ and $[s_{\text{new}}]$ implemented in `collision_free_path` requires all possible state trajectories between $[s_{\text{near}}]$ and $[s_{\text{new}}]$ to be wrapped in a box. This is again performed using guaranteed numerical integration. Once all configurations along trajectories between $left[s_{\text{near}}]$ and $[s_{\text{new}}]$ have been proved to lie in $\mathbb{S}_{\text{free}}$, $[s_{\text{new}}]$ is deemed safe, added to G, and connected to $[s_{\text{near}}]$.

A new random box is chosen to start the next iteration of the algorithm. A path is found when $[s_{\text{new}}] \subset [s_{\text{goal}}]$. The algorithm also stops when the number of nodes generated reaches its limit $\overline{K}$.

4 Application in Robotics

The proposed Box-RRT algorithm is now applied to path planning for nonholonomic vehicles in a structured 2D environment, where obstacles are described by polygons.

4.1 Model of the Vehicle and Specific Difficulties

A model based on the classical *simple car* model [11] evolving in a 2D environment is considered. This model incorporates nonholonomic constraints and is given by

$$\begin{cases} \dot{x} = v(1+w_v)\cos\theta \\ \dot{y} = v(1+w_v)\sin\theta \\ \dot{\theta} = \frac{v(1+w_v)}{L}\tan(\delta(1+w_\delta)) \end{cases}, \tag{2}$$

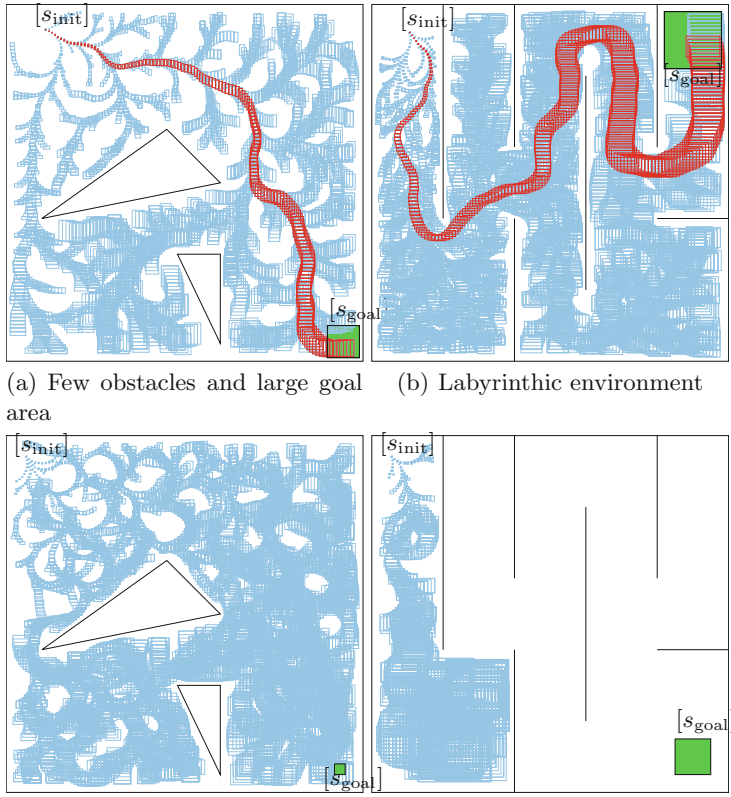
where the state vector $s = (x, y, \theta)^{\mathrm{T}}$ specifies the position (x, y) and orientation θ of a frame $\mathcal{V}$ attached to the vehicle with respect to a world frame $\mathcal{W}$ attached to the environment. The control input vector is $u = (v, \delta)^{\mathrm{T}}$, with v the longitudinal speed and $\delta \in [-\delta_{\max}, \delta_{\max}]$ the steering angle. Here, u is assumed to belong to a set $\mathbb{U}$ with finite cardinality. L is the distance between the front and rear wheels. The noise components $w_v \in [-v_{\mathrm{err}}, v_{\mathrm{err}}]$ and $w_\delta \in [-\delta_{\mathrm{err}}, \delta_{\mathrm{err}}]$ account for the slipping of the vehicle and for steering imprecision.

One of the difficulties of path planning in this context is the characterization of $\mathbb{S}_{\mathrm{free}}$, which may be quite complex. In [5], $\mathbb{S}_{\mathrm{free}}$ is characterized first or constructed iteratively. Here, $\mathbb{S}_{\mathrm{free}}$ is not determined explicitly: only the constraints of the environment are used to determine whether a set of paths is safe. For more details, see [15].

4.2 Results

Figures 1a–d present some results obtained with the Box-RRT algorithm; walls and obstacles to be avoided are represented by polygons. In Figs. 1a, b, no model error is considered ($v_{\mathrm{err}} = 0$ and $\delta_{\mathrm{err}} = 0$). The width of each component of $[s_{\mathrm{init}}]$ is 20 cm for the x and y components and 0.1 rad for θ. A box $[s_{\mathrm{goal}}]$, with size 10 m × 10 m ×2π rad, has to be reached in Fig. 1a. Its size is 15 m × 15 m ×2π rad in Fig. 1b. In both cases, a safe path is found, which is robust to uncertainty in the initial configuration.

If the size of $[s_{\mathrm{goal}}]$ is reduced, a path may no longer be found (see Fig. 1c), even if it may still exist. Since only prediction is used, and considering the form of the dynamical equation describing the motion of simple car, the size of the box describing the uncertain state always grows along the path. Thus, as soon as the size of $[s]$ at the end of a path exceeds that of $[s_{\mathrm{goal}}]$, there is no chance to reach $[s_{\mathrm{goal}}]$ from this box.



(a) Few obstacles and large goal area

(b) Labyrinthic environment

(c) Box-RRT fails to find a path (the goal area is too small)

(d) Box-RRT fails to find a path (skidding errors are too large)

Fig. 1. Some path-planning problems considered with Box-RRT

The same problem appears when the skidding error is too large. This problem is illustrated in Fig. 1d, where the size of $[s_{\text{init}}]$ is $10\,\text{cm} \times 10\,\text{cm} \times [1, 1.05]\,\text{rad}$, the size of $[s_{\text{goal}}]$ is $10\,\text{m} \times 10\,\text{m} \times 2\pi\,\text{rad}$, $v_{\text{err}} = 10^{-2}$ and $\delta_{\text{err}} = 10^{-3}$. Uncertainty then becomes exceedingly large and the vehicle no longer passes through the corridor. Thus, this problem cannot be solved using the present version of Box-RRT, unless some exteroceptive measurements are used at some points along the path to reduce uncertainty and the planning is restarted from time to time.

5 Conclusions

This paper has presented Box-RRT, an algorithm based on RRT able to perform robust and reliable path planning tasks for uncertain models of systems. Uncertain quantities are assumed to belong to sets. Albeit rather preliminary,

Box-RRT shows the potential of interval analysis to provide paths and prove that they are safe. Current work is on extending this approach to reachability analysis, see [15].

References

1. Berger, M.: Geometry I and II. Springer, Berlin (1987)
2. Bouilly, B., Simeon, T., Alami, R.: A numerical technique for planning motion strategies of a mobile robot in presence of uncertainty. In: Proceedings of the IEEE International Conference on Robotics and Automation, pp. 1327–1332, Nagoya, Japan, (1995)
3. Gonzalez, J.P., Stentz, A.: Planning with uncertainty in position: An optimal and efficient planner. In: Proceedings of the IEEE International Conference on Intelligent Robots and Systems, Edmonton, Canada, August 2005
4. Gonzalez, J.P., Stentz, A.: Planning with uncertainty in position using high-resolution maps. In: Proceedings of the IEEE International Conference on Robotics and Automation, pages 1015–1022, Roma, Italia, April 2007.
5. Jaulin, L.: Path planning using intervals and graphs. Reliable Comput. **7**(1), 1–15 (2001)
6. Jaulin, L., Kieffer, M., Didrit, O., Walter, E.: Applied Interval Analysis. Springer, London (2001)
7. Kieffer, M., Walter, E.: Guaranteed nonlinear state estimation for continuous–time dynamical models from discrete–time measurements. In: Proceedings of the 5th IFAC Symposium on Robust Control Design, 2006.
8. Lambert, A., Gruyer, D.: Safe path planning in an uncertain-configuration space. In: Proceedings of the IEEE International Conference on Robotics and Automation, pp. 4185–4190. Taipei, Taiwan, September 2003.
9. Latombe, J.C.: Robot Motion Planning. Kluwer Academic, Boston, MA, (1991)
10. LaValle, S.M.: Rapidly-exploring Random Trees: A new tool for path planning. Technical report, Computer Science Dept., Iowa State University, November 1998.
11. LaValle, S.M.: Planning Algorithms. Cambridge University Press, Cambridge (2006)
12. LaValle, S.M., Kuffner, J.J.: Rapidly-Exploring Random Trees: Progress and Prospects. In: Donald, B.R., Lynch, K.M., Rus, D. (eds.) Algorithmic and Computational Robotics: New Directions, pp. 293–308. A K Peters, Wellesley, MA (2001)
13. Lazanas, A., Latombe, J.C.: Motion planning with uncertainty: a landmark approach. Artif. intell. **76**(1–2), 287–317 (1995)
14. Moore, R.E.: Methods and Applications of Interval Analysis. SIAM, Philadelphia, PA (1979)
15. Pepy, R., Kieffer, M., Walter, E.: Reliable robust path planning. Int. J. Appl. Math. Comp. Sci. vol 19 413–424 (2009)
16. Pepy, R., Lambert, A.: Safe path planning in an uncertain-configuration space using RRT. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 5376–5381. Beijing, China (2006)

Minisymposium *Models and Methods for Viscous Jets, Break-up and Drop Forming*

Nicole Marheineke

Technical University Kaiserslautern, Department of Mathematics, P.O.Box 3049, D-67653 Kaiserslautern, Germany, nicole@mathematik.uni-kl.de

The understanding of the dynamics of viscous jets, break-up behaviour and drop formation is of interest in many industrial applications, including the drawing, tapering and spinning of polymer and glass fibres ([1, 4–7] and references within), sewing [2, 8, 9], pellets manufacturing [3, 10, 11], fuelling processes etc. In spinning or prilling processes for example, fluids of various properties are ejected from an orifice to form thin jets/fibres that might be subjected to surface tension, gravity, rotation and aerodynamic forces. These grow and break up into drops/filaments due to the growth of surface tension driven instabilities.

For industry the quality of the resulting goods (e.g. yarn, non-woven material, glass wool, pellets) counts. Manufacturing materials with desired specific properties requires the control of the production processes by choosing appropriate design parameters in the process. This task proposes a wide range of interesting challenges to mathematicians, natural scientists and engineers. The optimal control of the free boundary flows is based on the modelling and simulation of the motion and shape of the viscous jets due to the acting forces. Moreover, instabilities causing shape changes and rupture as well as transitions to other material behaviour (e.g. crystallisation) must be analysed.

Up to now, various aspects have been considered in numerous theoretical, numerical and experimental studies. In the minisymposium, Nicole Marheineke, from TU Kaiserslautern, Jamal Uddin, from University Birmingham, and Neil M. Ribe, from Institut de Physique du Globe, Paris, discuss string and rod models for viscous jets, their stability, applicability and validity in comparison to experimental data. Asymptotic one-dimensional models are derived via slender body theory, starting from cross-sectional averaging of the three-dimensional balance laws under systematic regular expansions and/or certain assumptions on geometry, velocity and stress profiles. String and rod models differ from each other in that a string model describes mass and linear momentum while a rod model consists additionally of a fully coupled equation for the angular momentum. The coupling of twist with the motion of the jet axis is very important for the coiling of a fibre falling on a rigid substrate,

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*, Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_93,

but not so relevant for the spinning of a fibre whose end is free. Finally, the asymptotic techniques can be also applied to formulate two-dimensional models for thin viscous films and sheets whose consideration by Thomas Goetz, from TU Kaiserslautern, follows similar optimisation aspects.

References

1. Caroselli, R.F.: In: Mark, H.F., Atlas, S.M., Cernia, E. (eds.) Man-Made Fibers, Science and Technology, vol. 3, pp. 361–389. Interscience, New York (1999)
2. Chiu-Webster, S., Lister, J.R.: J. Fluid Mech. **569**, 89–111 (2006)
3. Decent, S.P., King, A.C., Wallwork, I.M.: J. Eng. Math. **42**, 265–282 (2002)
4. Dewynne, J.N., Howell, P.D., Wilmott, P.: Quart. J. Mech. Appl. Math. **47**, 541–555 (1994)
5. Entov, V.M., Yarin, A.L.: J. Fluid Mech. **140**, 91–111 (1984)
6. Panda, S., Marheineke, N., Wegener, R.: Math. Meth. Appl. Sc. **31**, 1153–1173 (2008)
7. Pearson, J.R.A.: Mechanics of Polymer Processing. Elsevier, London (1985)
8. Ribe, N.M., Habibi, M., Bonn, D.: Phys. Fluids **18**, 084102 (2006)
9. Ribe, N.M., Lister, J.R., Chiu-Webster, S.: Phys. Fluids **18**, 124105 (2006)
10. Wallwork, I.M., Decent, S.P., King, A.C., Schulkes, R.M.S.M.: J. Fluid Mech. **459**, 43–65 (2002)
11. Wong, D.C.Y., Simmons, M.J.H., Decent, S.P., Parau, E.I., King, A.C.: Int. J. Multiphase Flow **30**, 499–520 (2004)

General String Theory for Dynamic Curved Viscida with Surface Tension

Nicole Marheineke[1] and Raimund Wegener[2]

[1] Technical University Kaiserslautern, Department of Mathematics, P.O.Box 3049, D-67653 Kaiserslautern, Germany, nicole@mathematik.uni-kl.de
[2] Fraunhofer-Institute for Industrial Mathematics (ITWM), Fraunhofer-Platz 1, D-67663 Kaiserslautern, Germany, wegener@itwm.fhg.de

Summary. This work deals with the asymptotic derivation and numerical investigation of a model for the dynamics of curved inertial viscous fibres under surface tension, as they occur in rotational spinning processes. The resulting string model accounts for the inner viscous transport and places no restriction on either motion or shape of the fibre centre-line. The boundary conditions for the free end of the fibre yield a description for its temporal evolution, depending on the ratio of viscous and surface tension (capillary number). The behaviour of the fibre is studied numerically as function of the effects of viscosity, gravity, rotation and surface tension.

1 Introduction

In rotational spinning processes of highly viscous fibres, the unrestricted motion and shape of a non-stationary centre-line is an important feature, see Fig. 1. In this work, we derive and investigate an asymptotic string model that is suitable for simulating the spinning of slender curved inertial viscous Newtonian fibres subjected to surface tension. Accordingly, we extend the slender body theory of [5] by including surface tension and deducing asymptotically appropriate boundary conditions for the free fibre end. For details we refer to [4].

2 Asymptotic Derivation

2.1 Three-Dimensional Free Boundary Value Problem

Let the fibre flow domain at time $t \in \mathbb{R}^+$ be $\Omega(t) \subset \mathbb{R}^3$ and its boundary $\partial\Omega(t) = \Gamma_{fr}(t) \cup \Gamma_{in}$ with $\Gamma_{fr}(t) \cap \Gamma_{in} = \emptyset$, where $\Gamma_{fr}(t)$ and Γ_{in} prescribe the time-dependent free surface and the time-independent planar inlet (spinning nozzle), respectively. Non-dimensionalising the underlying Navier–Stokes

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*, Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_94,

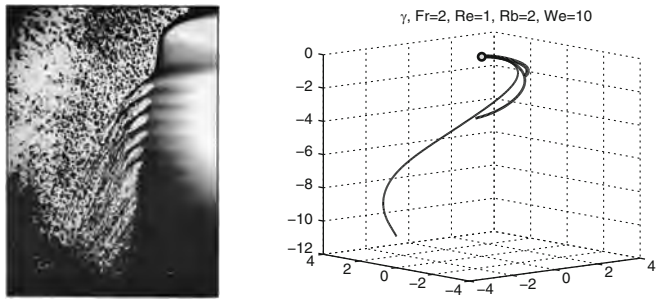


Fig. 1. *Left*: rotational device in glass wool production processes (photo by industrial partner). *Right*: centre-line of an extruded viscous fibre at different times

equations with the fluid density ρ, the mean velocity V at the nozzle and a typical length ℓ of the spun fibre, the small ratio between the nozzle width and the length ℓ is the slenderness parameter ϵ. Due to the scaling with V, the dimensionless inflow velocity profile $\mathbf{v}_{in}$ at the nozzle satisfies

$$|\Gamma_{in}|^{1/2} = \epsilon \ll 1, \qquad \int_{\Gamma_{in}} \mathbf{v}_{in} \cdot \boldsymbol{\tau_0} \, \mathrm{d}\mathcal{A} = \int_{\Gamma_{in}} \mathrm{d}\mathcal{A} = \epsilon^2,$$

where $|\Gamma_{in}| = \int_{\Gamma_{in}} \mathrm{d}\mathcal{A}$ and $\boldsymbol{\tau_0}$ is the inner normal vector of Γ_{in}. The model for the boundary value problem (BVP) reads

$$\begin{aligned}
\nabla_{\mathbf{r}} \cdot \mathbf{v}(\mathbf{r},t) &= 0 \\
\partial_t \mathbf{v}(\mathbf{r},t) + \nabla_{\mathbf{r}} \cdot (\mathbf{v} \otimes \mathbf{v})(\mathbf{r},t) &= \nabla_{\mathbf{r}} \cdot \mathbf{S}^T(\mathbf{r},t) + \mathbf{f}(\mathbf{r},t) && \mathbf{r} \in \Omega(t) \\
\mathbf{S} &= -p\,\mathbf{I} + \frac{1}{\mathrm{Re}}(\nabla_{\mathbf{r}}\mathbf{v} + (\nabla_{\mathbf{r}}\mathbf{v})^T) \\
(\mathbf{v} \cdot \mathbf{n})(\mathbf{r},t) = w(\mathbf{r},t), \qquad (\mathbf{S} \cdot \mathbf{n})(\mathbf{r},t) &= -\frac{\epsilon}{\mathrm{We}}(H\mathbf{n})(\mathbf{r},t) && \mathbf{r} \in \Gamma_{fr}(t) \\
\mathbf{v}(\mathbf{r},t) &= \mathbf{v}_{in}(\mathbf{r}) && \mathbf{r} \in \Gamma_{in}
\end{aligned}$$

with Reynolds $\mathrm{Re} = \ell\rho V/\mu$ and Weber number $\mathrm{We} = (\epsilon\ell/2)\rho V^2/\sigma$, dynamic viscosity μ and coefficient of surface tension σ. Apart from the unknown field variables for velocity $\mathbf{v}$ and hydrodynamic pressure p, the BVP determines the geometry $\Omega(t)$ specified by the outer normal vectors $\mathbf{n}$ and the scalar speed w of $\Gamma_{fr}(t)$. By choosing inhomogeneous dynamic boundary conditions for the stress tensor $\mathbf{S}$ the effects of surface tension are incorporated with mean curvature H deduced from the geometry. Body forces $\mathbf{f}$ complete the BVP.

The slenderness parameter ϵ enters the problem via the inflow domain. For the asymptotic reduction, we follow the concept of [5] and formulate the BVP in scaled curvilinear coordinates. These coordinates can be understood as generalisation of cylindrical ones along an arbitrary curve which is identified as fibre centre-line. Scaling leads to inflow conditions independent of the slenderness parameter ϵ, the corresponding balance laws are ϵ-dependent and form the basis for the rest of the asymptotic derivation.

Concretely, we define a bijective, time-dependent scaled curvilinear coordinate transformation $\check{\mathbf{r}}(\cdot,t)\colon \hat{\Omega}(t) \subset \mathbb{R}^3 \mapsto \Omega(t) \subset \mathbb{R}^3$

$$\check{\mathbf{r}}(\mathbf{x},t) = \boldsymbol{\gamma}(s,t) + \epsilon\, x_1 \boldsymbol{\eta_1}(s,t) + \epsilon\, x_2 \boldsymbol{\eta_2}(s,t) \quad \text{with} \quad s = x_3$$

with respect to the slenderness parameter ϵ and the arc-length parameterised fibre centre-line $\boldsymbol{\gamma}$. The normal vectors $\boldsymbol{\eta_1}$, $\boldsymbol{\eta_2}$ together with the tangent $\boldsymbol{\tau} = \partial_s \boldsymbol{\gamma}$ constitute a local orthonormal basis along the curve. We specify the flow domain $\Omega(t)$ in terms of the associated domain $\hat{\Omega}(t)$ in coordinates, i.e. $\Omega(t) = \check{\mathbf{r}}(\hat{\Omega}(t),t)$. Therefore, we assume that $\hat{\Omega}(t)$ is given by the fibre length $L(t)$ and a 2π-periodic radius function $R(\cdot,t) : [0,2\pi) \times [0,L(t)) \mapsto \mathbb{R}^+$ such that

$$\hat{\Omega}(t) = \{\, \mathbf{x} = (x_1,x_2,s) \in \mathbb{R}^3 \,|\, (x_1,x_2) \in \mathcal{A}(s,t), s \in [0,L(t))\}$$

with cross-sections

$$\mathcal{A}(s,t) = \{(x_1,x_2) \in \mathbb{R}^2 \,|\, x_1 = \varrho\cos(\psi), x_2 = \varrho\sin(\psi), \\ \varrho \in [0,R(\psi,s,t)],\ \psi \in [0,2\pi)\}.$$

Then, the fibre domain is described by L, $\boldsymbol{\gamma}$, R.

In addition to the geometrical quantities, we introduce the following characteristic quantities related to $\check{\mathbf{r}}$: coordinate transformation matrix $\mathbf{F} = \nabla_{\mathbf{x}}\check{\mathbf{r}}$, functional determinant $J = \det(\mathbf{F})$, inverse matrix $\mathbf{G} = \mathbf{F}^{-1}$ and coordinate velocity $\mathbf{q} = \partial_t \check{\mathbf{r}}$. Then, the governing equations of the free BVP in curvilinear coordinates $\mathbf{x} \in \hat{\Omega}(t)$ read

$$\begin{aligned} \partial_t J(\mathbf{x},t) + \nabla_{\mathbf{x}} \cdot (J\mathbf{u})(\mathbf{x},t) &= 0 \\ \partial_t (J\mathbf{v})(\mathbf{x},t) + \nabla_{\mathbf{x}} \cdot (\mathbf{u} \otimes J\mathbf{v})(\mathbf{x},t) &= \nabla_{\mathbf{x}} \cdot \mathbf{T}^T(\mathbf{x},t) + (J\mathbf{f})(\mathbf{x},t) \\ \mathbf{u} &= (\mathbf{v} - \mathbf{q}) \cdot \mathbf{G} \\ \mathbf{T} &= J\mathbf{S} \cdot \mathbf{G}. \end{aligned}$$

The physical and geometrical properties of the observables are preserved under the above transformation. The intrinsic velocity $\mathbf{u}$ describes the rate of convective transport of the unknowns in the coordinates, whereas the momentum-associated velocity $\mathbf{v}$ represents one of these transported quantities. Their relation is expressed in the coupling condition. See [4] for more details.

2.2 Asymptotic Analysis

The derivation of an asymptotic 1d model from the 3d free BVP is based on the cross-sectional averaging of the balance laws. Thereby, regular power expansions in the slenderness parameter to zeroth and first order, e.g. $\mathbf{v}_\epsilon = \mathbf{v}^{(0)} + \epsilon\mathbf{v}^{(1)} + \mathcal{O}(\epsilon^2)$, yield the necessary cross-sectional profile properties of the unknowns. We abbreviate $\langle f \rangle_{\mathcal{A}_\epsilon(s,t)} = \int_{\mathcal{A}_\epsilon(s,t)} f(x_1,x_2,s,t)\,\mathrm{d}x_1\mathrm{d}x_2$ and

$\langle f\rangle_{\partial\mathcal{A}_\epsilon(s,t)} = \int_{\partial\mathcal{A}_\epsilon(s,t)} f/\sqrt{n_1^2+n_2^2}\,\mathrm{d}l$ for any integrable function f on $\hat{\Omega}(t)$ and components n_i of the normal vector $\mathbf{n}$. The cross-sectionally averaged balance laws are

$$\begin{aligned}
&\partial_t\langle J_\epsilon\rangle_{\mathcal{A}_\epsilon(s,t)} + \partial_s\langle J_\epsilon u_{3,\epsilon}\rangle_{\mathcal{A}_\epsilon(s,t)} = 0\\
&\partial_t\langle J_\epsilon \mathbf{v}_\epsilon\rangle_{\mathcal{A}_\epsilon(s,t)} + \partial_s\langle J_\epsilon u_{3,\epsilon}\,\mathbf{v}_\epsilon\rangle_{\mathcal{A}_\epsilon(s,t)}\\
&\qquad = \partial_s\langle \mathbf{T}_\epsilon\cdot\mathbf{e_3}\rangle_{\mathcal{A}_\epsilon(s,t)} - \frac{\epsilon}{\mathrm{We}}\langle J_\epsilon H_\epsilon \mathbf{G}_\epsilon\cdot\mathbf{n}_\epsilon\rangle_{\partial\mathcal{A}_\epsilon(s,t)} + \langle J_\epsilon \mathbf{f}_\epsilon\rangle_{\mathcal{A}_\epsilon(s,t)}.
\end{aligned}$$

From the BVP we derive at zeroth and first order [4]

$$\begin{aligned}
&\mathbf{u}^{(-1)} = \mathbf{0},\ u_3^{(0)} = u_3^{(0)}(s,t),\ \langle\mathbf{v}^{(0)}\rangle_{\mathcal{A}_0} = u_3^{(0)}\partial_s\boldsymbol{\gamma}^{(0)} + \partial_t\boldsymbol{\gamma}^{(0)}\\
&\mathbf{T}^{(0)} = \mathbf{T}^{(1)} = \mathbf{0}, \qquad \langle\mathbf{T}^{(2)}\cdot\mathbf{e_3}\rangle_{\mathcal{A}_0} = \left(\frac{3}{\mathrm{Re}}|\mathcal{A}_0|\partial_s u_3^{(0)} - \frac{\sqrt{\pi}}{2\mathrm{We}}\sqrt{|\mathcal{A}_0|}\right)\partial_s\boldsymbol{\gamma}^{(0)}\\
&\langle JH\mathbf{G}\cdot\mathbf{n}\rangle_{\partial\mathcal{A}}^{(0)} = \mathbf{0}, \qquad \langle JH\mathbf{G}\cdot\mathbf{n}\rangle_{\partial\mathcal{A}}^{(1)} = -\sqrt{\pi}\partial_s(\sqrt{|\mathcal{A}_0|}\partial_s\boldsymbol{\gamma}^{(0)}).
\end{aligned}$$

The evaluation of the boundary integrals is based on the intuitive assumption of circular cross-sections at leading order which stands in accordance to the fact that the cross-sectional shape tends to a circle under surface tension. Abbreviating $u = u_3^{(0)}$, $A = |\mathcal{A}_0|$ and dropping the superscripts of leading order, we obtain the following asymptotic result.

Theorem 1 (Fibre String Model). *The spinning of a slender curved inertial viscous fibre subjected to surface tension is modelled by*

$$\begin{aligned}
\partial_t A + \partial_s(uA) &= 0\\
\partial_t(A\mathbf{v}) + \partial_s(uA\mathbf{v}) &= \partial_s\left((\frac{3}{\mathrm{Re}}A\partial_s u + \frac{\sqrt{\pi}}{2\mathrm{We}}\sqrt{A})\partial_s\boldsymbol{\gamma}\right) + A\mathbf{f}\\
\partial_t\boldsymbol{\gamma} + u\partial_s\boldsymbol{\gamma} &= \mathbf{v}
\end{aligned}$$

$$\frac{\mathrm{d}L(t)}{\mathrm{d}t} = u(L(t),t), \quad L(0) = 0, \qquad A\partial_s u(L(t),t) = \frac{\sqrt{\pi}}{6}\frac{\mathrm{Re}}{\mathrm{We}}\sqrt{A}(L(t),t)$$

$$A(0,t) = 1, \quad u(0,t) = 1, \quad \boldsymbol{\gamma}(0,t) = \boldsymbol{\gamma}_0, \quad \partial_s\boldsymbol{\gamma}(0,t) = \boldsymbol{\tau}_0, \quad \|\partial_s\boldsymbol{\gamma}\| = 1.$$

In rotational spinning processes the body densities $\mathbf{f}$ *arise due to gravity and rotation with Froude* Fr *and Rossby number* Rb

$$\mathbf{f} = \mathrm{Fr}^{-2}\mathbf{e_g} - 2\mathrm{Rb}^{-1}(\mathbf{e}_\omega\times\mathbf{v}) - \mathrm{Rb}^{-2}(\mathbf{e}_\omega\times(\mathbf{e}_\omega\times\boldsymbol{\gamma})).$$

The boundary conditions for the free end of the fibre come from a global balance of the volume-averaged 3d and the line-averaged 1d balance laws at leading order. We see that the viscous stresses balance the surface tension. Moreover, combining the boundary conditions with the conservation of mass yields a simple differential equation describing the evolution of the area at the fibre end, $\mathrm{d}A(L(t),t)/\mathrm{d}t = -(A\partial_s u)(L(t),t) = -\sqrt{\pi}/(6\mathrm{Ca})\sqrt{A}(L(t),t)$, $A(0,0) = 1$. Its unique solution is

$$A(L(t),t) = \begin{cases} (1-t/t_c)^2, & t \le t_c, \\ 0, & t > t_c \end{cases}, \quad t_c = \frac{12}{\sqrt{\pi}}\mathrm{Ca}$$

with capillary number $\mathrm{Ca} = \mathrm{Re}/\mathrm{We}$, since $A \ge 0$. Our model ceases to be valid when $A(L(t),t) = 0$ due to the definition of our fibre domain. Accordingly, the numerical simulations are performed for $t < t_c$.

3 Numerical Investigation

Our model places no restrictions on either the motion or shape of the fibre centre-line, and account for both the inner viscous transport and surface tension. Thus, it includes most of those previously derived for nearly straight (e.g. [2]) and curved centre-lines [1, 5] and can be regarded as a generalised theory of viscous strings. Moreover, our balance laws coincide with those of [3], considering $\partial_s u = \partial_s v_\tau - \kappa v_\eta$ with $v_\tau = \mathbf{v}\cdot\boldsymbol{\tau}$, $v_\eta = \mathbf{v}\cdot\boldsymbol{\eta}$, $\boldsymbol{\tau} = \partial_s\boldsymbol{\gamma}$, $\boldsymbol{\eta} = \partial_{ss}\boldsymbol{\gamma}/\kappa$ and $\kappa = \|\partial_{ss}\boldsymbol{\gamma}\|$. Supplemented with an additional angular momentum equation that is decoupled from the string part, the approach of [3] describes a rod model. A rod model with fully coupled equations for mass, linear and angular momentum is derived in [6], choosing a non-stationary material centre-line as reference curve in contrast to our geometrical one. However, twist is not relevant for rotational spinning.

For the simulation of a rotational spinning process (Fig. 1), we consider a situation in which a fibre is ejected horizontally from a spinning nozzle located on the curved face of a cylindrical drum with unit dimensionless radius that rotates about its vertical symmetry axis. This implies $\mathbf{e_\omega} = -\mathbf{e_g} = \mathbf{e_3}$ and $\boldsymbol{\gamma_0} = \boldsymbol{\tau_0} = \mathbf{e_1}$ in the rotating reference frame. Figure 2 shows the temporal evolution of the cross-section A, the intrinsic velocity u and the projection of the fibre centre-line onto the $\mathbf{e}_1$–$\mathbf{e}_2$-plane for different parameter combinations. Our results for high Reynolds number (small viscosity) agree well with those of [1, 7]: the smaller the Rossby number, i.e. the faster the rotation, the more pronounced is the curling of the fibre in the $\mathbf{e}_1$–$\mathbf{e}_2$-plane. The curling is also influenced by the surface tension. Moreover, decreasing the Weber number, i.e. increasing the surface tension, accelerates the thinning of the fibre end. The critical time t_c when the fibre end shrinks to a point is observed in the simulations. In contrast to the stationary centre-line approach of [1], our model allows the numerical simulation of a non-stationary fibre centre-line which is a significant feature in rotational glass spinning processes ($\mathrm{Rb}, \mathrm{Re} \to 0$) when the fibre centre-line is displaced during recoiling after drop detachment (cf. "dynamic break-up mode 4" [7]), for visualisation of this effect see Figs. 1, 2 and [4, 5].

Acknowledgement

This work has been supported by the Rheinland-Pfalz Excellence Center for Mathematical and Computational Modelling.

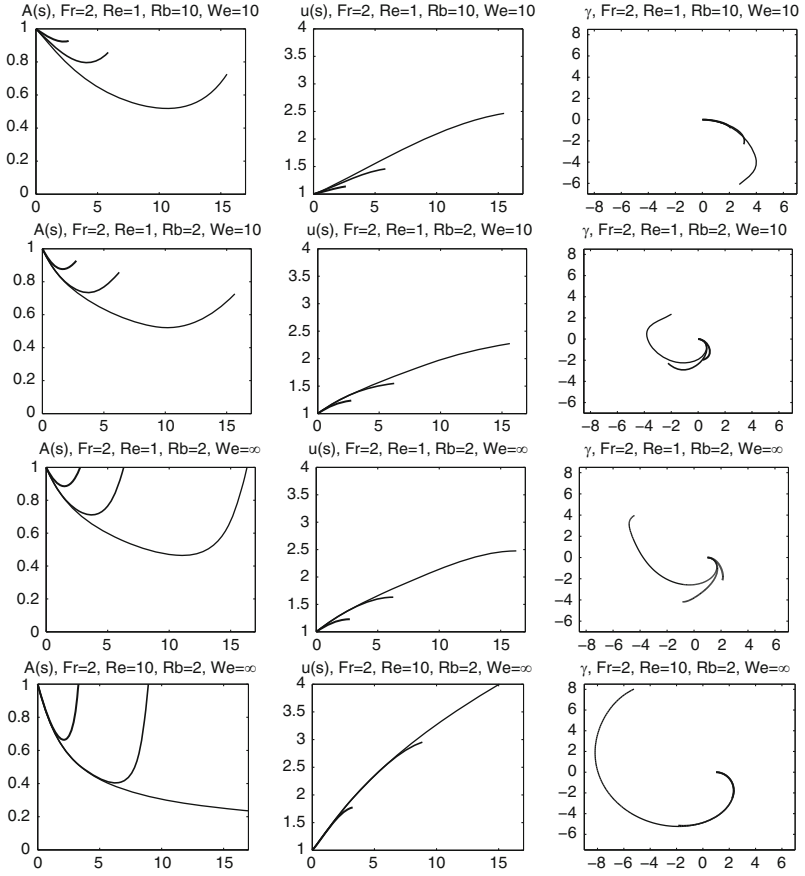


Fig. 2. *Left to right*: $A(s)$, $u(s)$, $\boldsymbol{\gamma}$ in $\mathbf{e_1}$–$\mathbf{e_2}$-plane at $t \in \{2.5, 5, 10\}$. *Top to bottom*: $(\mathrm{Re}, \mathrm{Rb}, \mathrm{We}) \in \{(1, 10, 10), (1, 2, 10), (1, 2, \infty), (10, 2, \infty)\}$ for fixed $\mathrm{Fr} = 2$

References

1. Decent, S.P., King, A.C., Wallwork, I.M.: J. Eng. Math. **42**, 265–282 (2002)
2. Dewynne, J.N., Ockendon, J.R., Wilmot, P.: J. Fluid Mech. **244**, 323–338 (1992)
3. Entov, V.M., Yarin, A.L.: J. Fluid Mech. **140**, 91–111 (1984)
4. Marheineke, N., Wegener, R.: J. Fluid Mech. **622**, 345–369 (2009)
5. Panda, S., Marheineke, N., Wegener, R.: Math. Meth. Appl. Sc. **31**, 1153–1173 (2008)
6. Ribe, N.M., Habibi, M., Bonn, D.: Phys. Fluids **18**, 084102 (2006)
7. Wong, D.C.Y., Simmons, M.J.H., Decent, S.P., Parau, E.I., King, A.C.: Int. J. Multiphase Flow **30**, 499–520 (2004)

Instability of Non-Newtonian Liquid Jets Curved by Gravity

J. Uddin and S.P. Decent

School of Mathematics, University of Birmingham, Birmingham B15 2TT, UK

Summary. A slender jet model is used to describe the instability of a curved liquid jet falling under gravity. The fluid is modelled as an inelastic non-Newtonian fluid obeying the Carreau model. A linear instability analysis is performed to examine the behaviour of the most unstable wavenumber and growth rate of instabilities.

1 Curved Liquid Jets

The breakup of a liquid jet into droplets is ubiquitous in many industrial and engineering applications (see [2] for a review). After emerging from an orifice or nozzle a liquid jet may be distorted by the influence of gravity or wind with the result being a liquid jet with a curved centreline. In such cases the interplay between the centreline of the jet and dynamics within the jet can affect the resulting instability and drop formation.

2 Governing Equations

Let us consider a liquid jet of density ρ emerging from a nozzle of radius a with speed U and falling under the influence of gravity. Upon leaving the nozzle the jet moves solely in the x-y plane and the centreline of the jet in the Cartesian plane can be described by $(X(s,t), Y(s,t))$ where s is the arc length along the jet and t is time (see Fig. 1). Any analysis of the jet in Cartesian coordinates leads to algebraic equations which are tedious and opaque. It is more amenable to describe the dynamics of the liquid jet in a coordinate system which has one coordinate vector along the axis of the jet with the remaining unit vectors as plane polar coordinates in any cross section of the jet. A similar coordinate system has been used by [3] but our system is based on those derived by [7] to investigate spiralling liquid jets in prilling. The derivation of the resulting system of unit vectors in this curved coordinate system can be found elsewhere (see for example [7]) and so we

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*,
Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_95,

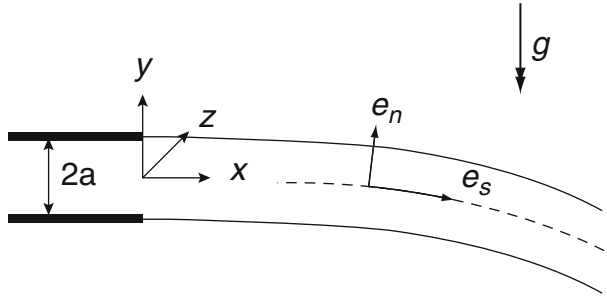


Fig. 1. A sketch of a liquid jet curved by the action of gravity. The centreline of the jet is shown as a *dashed line* and can be written an $(X(s,t), Y(s,t))$

will not repeat the derivation here. It suffices for our purposes to state that the unit vectors along the axis, along the radial direction and the azimuthal direction are denoted by $\boldsymbol{e_s}$, $\boldsymbol{e_n}$ and $\boldsymbol{e_\phi}$ respectively. They have the associated scale functions given by $h_s = 1 + n(X_s Y_{ss} - Y_s X_{ss})$, $h_n = 1$ and $h_\phi = n$.

We make our equations dimensionless by scaling velocity components with U and radial length scales with a. The jet curves over a length scale s_0 given by the ratio $s_0 = U^2/g$ where g is the acceleration due to gravity. Hence, we make lengths along the X and Y axis dimensionless with regards to U^2/g and time we scale with U/g. The pressure within the liquid jet is scaled with ρU^2. We assume that the aspect ratio $a/s_0 = \epsilon$ is small (i.e., $\epsilon \ll 1$ whence we have slender jets) and thus we are able to utilise a lubrication type approximation in the foregoing analysis. In order to have $\epsilon \ll 1$ we require $U^2/a \gg g$. The velocity components in the axial, radial and azimuthal directions are given by u, v and w respectively.

In the present study we take the emerging liquid to be non-Newtonian, and in particular, we examine the case of a shear thinning liquid jet which has a constitutive equation governed by the Carreau model

$$\mu = \tilde{\mu}_0 \tilde{\mu} = \tilde{\mu}_0 (1 - \xi)[1 + (h\dot{\gamma})^2]^{\frac{f-1}{2}} + \tilde{\mu}_0 \xi. \tag{1}$$

In this case, μ is the shear rate dependent viscosity, $\dot{\gamma} = \sqrt{\boldsymbol{\mathcal{E}} : \boldsymbol{\mathcal{E}}/2}$ is the second invariant of the deformation tensor $\boldsymbol{\mathcal{E}} = \nabla \mathbf{u} + (\nabla \mathbf{u})^T$, $\tilde{\mu}_0$ is the zero-shear rate viscosity, h is a constant, $\xi \tilde{\mu}_0$ is the viscosity in the limit of infinite shear and finally f is the flow index number such that $0 \leq f \leq 1$.

The equations governing the flow within the jet are given by

$$\boldsymbol{\nabla} \cdot \mathbf{u} = 0 \qquad \text{and} \qquad \rho \frac{D\mathbf{u}}{Dt} = -\nabla p + \nabla \cdot (\mu \boldsymbol{\mathcal{E}}) = -\nabla p + \mu \nabla^2 \mathbf{u} + \nabla \mu \cdot \boldsymbol{\mathcal{E}}, \tag{2}$$

where $\boldsymbol{u} = (u, v, w)$ is the velocity vector, ρ is the density of the liquid, t is time and p is the pressure.

These equations are supplemented by a number of boundary conditions at the liquid jet interface. Firstly, we have a kinematic condition

$$\frac{D}{Dt}(n - R(s,t)) = 0, \tag{3}$$

where $R(s,t)$ is the position of the free surface. We also have normal and tangential stress conditions at the liquid jet interface such that

$$\mu \boldsymbol{n} \cdot \boldsymbol{\mathcal{E}} \cdot \boldsymbol{n} = \sigma \cdot \kappa \quad \text{and} \quad \boldsymbol{t}_i \cdot \boldsymbol{\mathcal{E}} \cdot \boldsymbol{n} = 0, \tag{4}$$

where $\boldsymbol{n}$ and $\boldsymbol{t}_i$ for $i = 1$ and $i = 2$ are the unit normal and tangential vectors (note that the free surface has two tangential vectors) and σ is the surface tension of the liquid-gas interface. Furthermore, we have the additional constraint upon the centreline equation given by the arc length condition

$$X_s^2 + Y_s^2 = 1. \tag{5}$$

Expanding our variables using a slender jet model, with ϵ as our small term, the resulting leading order one dimensional equations are given in [6].

3 Steady State Solution

In order to determine steady trajectories of the curved liquid jet we now consider the steady state of this set of equations. This is given by the solution to the following set of equations (see [6])

$$u_0 u_{0s} = -\frac{1}{We}\left(\frac{1}{R_0}\right)_s - \frac{Y_{0s}}{\mathcal{F}^2}. \tag{6}$$

$$\frac{u_{0s}}{2} R_0 + u_0 R_{0s} = 0. \tag{7}$$

$$X_{0s}^2 + Y_{0s}^2 = 1. \tag{8}$$

$$\left({u_0}^2 - \frac{1}{R_0 We}\right)(X_{0s} Y_{0ss} - X_{0ss} Y_{0s}) - \frac{X_{0s}}{\mathcal{F}^2} = 0. \tag{9}$$

where the Weber number $We = \frac{\rho U^2 a}{\sigma}$ and the Froude number $\mathcal{F} = \frac{U}{\sqrt{s_0 g}}$. In the limiting case of zero surface tension (i.e., infinite Weber number) the centreline of the jet corresponds to the motion of a projectile released from the nozzle and falling under the influence of gravity. For finite values of the Weber number, our nonlinear ordinary differential equations must be solved using some suitable numerical method (like Runge–Kutta for example). The initial conditions we use are $u_0 = R_0 = X_{0s} = 1$ and $Y_{0s} = X_0 = Y_0 = 0$ at $s = 0$. Steady trajectories for different Froude numbers are shown in Fig. 2.

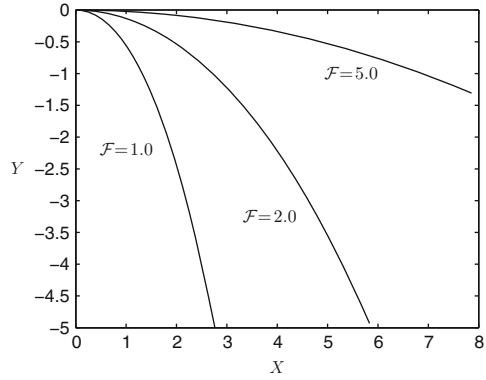


Fig. 2. Steady state trajectories for different Froude numbers. The centreline is seen to curve less when the Froude number is increased. ($We = 20.0$)

4 Linear Stability Analysis

In this section we consider the linear stability of disturbances about our leading order steady state solution obtained in the previous section. We should note that if the centreline of the falling jet is assumed to curve over a length-scale of $s = O(1)$ and perturbations along the jet are of order a (which is comparable to ϵ when $s = O(1)$) then travelling wave modes of the form $\exp(ik\bar{s} + \bar{t})$ must be considered, where $\bar{s} = s/\epsilon$ and $\bar{t} = t/\epsilon$, in order to have $k = k(s) = O(1)$ and $\lambda = \lambda(s) = O(1)$, which are the wavenumber and growth rate of disturbances along the jet respectively. We therefore have a multiple scales formulation with perturbations having wavelength of $O(\epsilon)$ as required. We now add small disturbances to the steady state solutions (obtained in the previous section) having the form $\delta \exp(ik\bar{s} + \lambda\bar{t})$ where δ is some small dimensionless constant. In this case the symbols with a subscript denote steady state solutions and k is a real wavenumber with λ being complex, so that $\lambda = \lambda_r + i\lambda_i$ where λ_r is the growth rate of disturbances and λ_i is the wavenumber of disturbances along the jet. The eigenvalue relationship between the growth rate and wavenumber of disturbances is given by the following equation

$$(\lambda + iku_0) - \frac{k^2 R_0}{2We(\lambda + ku_0)} \left(\frac{1}{R_0^2} - k^2 \right) + \frac{k^2 \hat{\mu}}{\bar{\mathcal{R}e}} = 0, \tag{10}$$

where $\hat{\mu} = (1 - \xi)[1 + (\sqrt{3}hu_{0s})^2]^{\frac{f-1}{2}} + \xi$ and $\bar{\mathcal{R}e} = \frac{\rho U a}{\bar{\mu}_0}$ is the Reynolds number. We now write $\lambda = \lambda_r + i\lambda_i$ and after substitution into (10) we find $\lambda_i = -ku_0$ (so that disturbances are simply convected with the liquid jet) and

$$\lambda_r = -\frac{k^2 \hat{\mu}}{2\bar{\mathcal{R}e}} + \frac{k}{2} \left[\frac{k^2 \hat{\mu}^2}{\bar{\mathcal{R}e}^2} + \frac{2}{R_0 We}(1 - (kR_0)^2) \right]^{\frac{1}{2}} \tag{11}$$

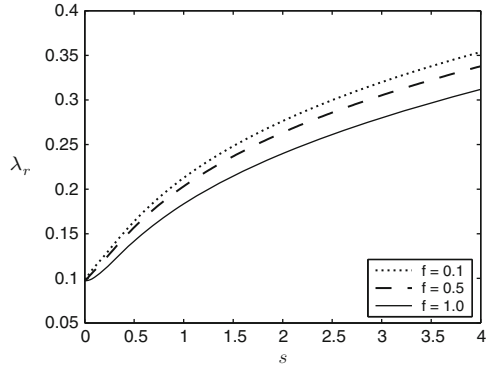


Fig. 3. The growth rate λ_r of the most unstable wavenumber $k^\star$ plotted against the arc length s along the jet for different flow index numbers f. ($\mathcal{F} = 0.5$, $Oh = 0.30$, $We = 9.0$, $h = 1.0$ and $\xi = 0.1$)

For unstable disturbances we require that λ_r be positive (otherwise disturbances decay along the jet). This necessitates that $0 < kR_0 < 1$, which for a straight jet where the radius along the jet remains a constant agrees with the classical result of [5]. Equation (11) may be differentiated to find the most unstable wavenumber $k = k^\star$ for which the growth rate λ_r attains a maximum, this is found to be given by

$$k^\star(s) = \frac{1}{(2R_0^3)^{\frac{1}{4}}} \frac{1}{[\hat{\mu}Oh + \sqrt{2R_0}]^{\frac{1}{2}}}, \tag{12}$$

where the Ohnesorge number $Oh = \sqrt{We}/\bar{\mathcal{R}}e$ is a measure of the relative importance of viscous effects to surface tension. Since both R_0 and $\hat{\mu}$ are determined from the steady state solutions, and both vary along the arc length s of the jet, the most unstable wavenumber $k^\star$ will also be a function of s.

5 Results and Conclusion

We find that the most unstable wavenumber increases along the jet and that larger values of $k^\star$ are seen to correspond to smaller values of the flow index number f (i.e., as we increase the shear thinning properties of the liquid we get larger values of $k^\star$). This pattern is also true for the growth rates of the most unstable mode as shown in Fig. 3. In the context of linear theory this shows that increasing shear thinning properties of a liquid jet lead to smaller droplets (since the most unstable wavenumber $k^\star$ is related to disturbance wavelengths Λ by the relationship $k^\star = 2\pi/\Lambda$) and shorter breakup lengths. The behaviour of the most unstable mode for changes in the Froude number for a highly shear thinning liquid ($f = 0.1$) is shown in Fig. 4. In this case increasing the importance of gravity (which corresponds to reducing the Froude number) leads to higher values of the most unstable mode.

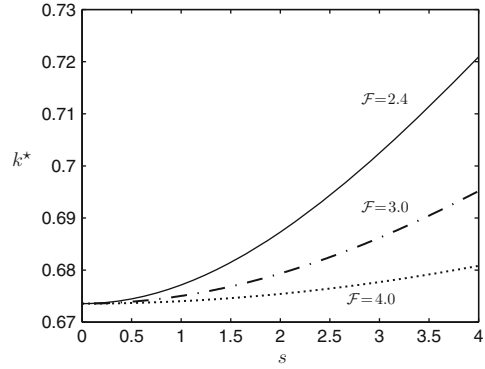


Fig. 4. The most unstable wavenumber $k^\star$ plotted against the arc length s along the jet for different Froude numbers $\mathcal{F}$. ($f = 0.3$, $Oh = 0.144$, $We = 12.0$, $h = 1.0$ and $\xi = 0.2$)

References

1. Decent, S.P., King, A.C., Simmons, M.J.H., Părău, E.I., Wallwork, I.M., Gurney, C.J., Uddin, J.: Appl. Math. Mod. **33**, 4283–4302 (2009)
2. Eggers, J.: Rev. Mod. Physics **69**(3), 865–929 (1997)
3. Entov, V.M., Yarin, A.L.: J. Fluid Mech. **140**, 91–113 (1984)
4. Părău, E.I., Decent, S.P., Simmons, M.J.H., Wong, D.C.Y., King, A.C.: J. Eng. Maths. **57**, 159–179 (2007)
5. Rayleigh, W.S.: Proc. Lond. Math. Soc. **10**, 4 (1878)
6. Uddin, J.: Ph.D. Thesis, University of Birmingham (2007)
7. Wallwork, I.M.: Ph.D. Thesis, University of Birmingham, Birmingham (2002)

Simulation and Optimization of Film Casting Processes

T. Götz[1] and K. Selvanayagam[2]

[1] Dept. of Mathematics, University of Kaiserslautern, D–67653 Kaiserslautern, Germany, goetz@mathematik.uni-kl.de
[2] Dept. of Mathematics, IIT Madras, Chennai 600 036, India

Summary. We present an optimal control approach for the isothermal film casting process with free surfaces described by averaged Navier–Stokes equations. We control the thickness of the film at the take-up point using the shape of the nozzle and the initial thickness. The control goal consists in finding an even thickness profile.

1 Introduction

Polymer films for video and magnetic tapes are produced by film casting. The molten polymer emerging from a flat die is first stretched a short distance between the die and a temperature controlled roll. The film shows a lateral neck-in as well as an inhomogeneous decrease of the thickness. The formation of edge beads surrounding a central area of constant thickness is generally called the dog bone defect or edge bead defect. In this paper we develop a mathematical model to predict the shape of the die which minimizes the edge bead defect.

2 Governing Equations

We consider the stationary, isothermal three-dimensional Newtonian model for the film casting process derived earlier by Demay and co-workers [2, 3, 8] or in [1]. The geometry of the film casting process is shown in Figure 1.

The polymer is pressed through the die (located in the yz-plane) with a velocity u_0 and wrapped up with velocity $u_L > u_0$ by a spindle at $x = L$. The die has a width of W_0 in the y-direction and a thickness of e_0 in the z-direction. For typical film casting processes, the thickness of the film at the nozzle is small compared to both the length and the width of the film i.e. $e_0/W_0 \ll 1$ and $e_0/L \ll 1$.

Averaging the mass and momentum equations describing the polymer flow over the z-direction, see [2, 8], leads to

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*, Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_96,

$$\nabla \cdot (eU) = 0 \tag{1a}$$

$$(U \cdot \nabla)U = \frac{1}{\mathrm{Re}} \left(\Delta U + 3\nabla \left(\nabla \cdot U\right)\right). \tag{1b}$$

Here $U = (u, v)$ denotes the velocity field in the x- and y-directions and e denotes the thickness of the film in the z-direction. The Reynolds number $\mathrm{Re} = \frac{Lu_L}{\nu}$ is based on the length of the film, the take-up velocity and the viscosity of the fluid. Using the notations of Fig. 1, the system (1) has to be solved inside the two-dimensional film domain $\Omega = \{(x, y) : \ 0 < x < L \ , \ -W(x) < y < W(x)\}$. Note that the width $W(x)$ of the film is a free boundary and not known a priori. The boundary of the domain consists of the extrusion line $\gamma_1 = \{0\} \times (-W(0), W(0))$, the take-up line $\gamma_2 = \{L\} \times (-W(L), W(L))$ and the lateral boundaries $\gamma_3 = (0, L) \times \{-W(x)\}$ and $\gamma_4 = (0, L) \times \{W(x)\}$. At the inflow and outflow flow boundary, we prescribe Dirichlet data

$$(u,\, v,\, e) = (u_0,\, 0,\, e_0) \quad \text{at } \gamma_1, \tag{1c}$$

$$(u,\, v) = (u_L,\, 0) \qquad \text{at } \gamma_2. \tag{1d}$$

The ratio $D = u_L/u_0 > 1$ between the winding and the extrusion velocity is called draw ratio. Due to the hyperbolic nature of (1a), there is no boundary condition for the thickness on γ_2. The treatment of the lateral boundaries γ_3, γ_4 is more sophisticated, since they are free boundaries. Their location is not known in advance and evolves with the width $W = W(x)$ of the film. The dynamic and kinematic conditions at the free boundary read as

$$\sigma \cdot n = 0 \quad \text{at } \gamma_3,\, \gamma_4, \tag{1e}$$

$$u\partial_x W - v = 0 \quad \text{at } \gamma_3,\, \gamma_4. \tag{1f}$$

Here n denotes the unit outer normal to γ_i, $i = 3, 4$. The stress tensor σ is given by $\sigma = (\nabla U) + (\nabla U)^T + 2\,(\mathrm{div}\, U)\, I = \begin{pmatrix} 4\partial_x u + 2\partial_y v & \partial_y u + \partial_x v \\ \partial_y u + \partial_x v & 2\partial_x u + 4\partial_y v \end{pmatrix}$, where I is the 2×2 identity matrix.

The following typical parameters are used throughout the paper: stretching distance $L = 0.4\,\mathrm{m}$, film width $W_0 = 1\,\mathrm{m}$, draw ratio $D = 10$ and Reynolds number $Re = 3$.

3 Optimal Control

The model (1) is capable to predict the final thickness $e(L, y)$ of the film. This thickness profile depends on the geometry e_0 of the nozzle as well as the draw-ratio D. Using a rectangular nozzle, i.e. a uniform initial thickness e_0, one obtains the well-known effect of edge beads, see Fig. 2. In this case the final film is thinner in the middle than at the lateral surfaces; an undesired

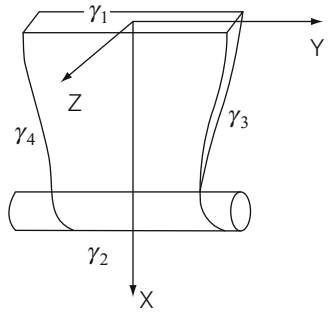


Fig. 1. Sketch of the considered geometry for the film casting process

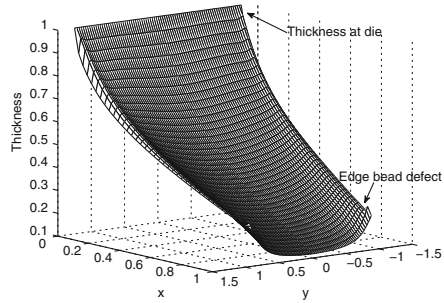


Fig. 2. Thickness profile of the film casting process with edge bead defect

result. In contrast to that, industrial applications aim to produce films with a uniform thickness profile at the take-up roll.

The parameters that can be modified are the initial thickness profile e_0 and the velocity at the die u_0 as well as the velocity of the take-up roll u_L. However, for simplicity, we focus on controlling the initial thickness e_0 and the initial velocity u_0 of the film.

To model the requirement of an even film thickness at the take-up roll, we consider the following tracking-type cost functional

$$J(z, \phi) = \|e(L, y) - e_d\|^2 + \alpha \|\phi\|^2 \tag{2}$$

where e_d is the desired thickness, $z = (u, v, e)$ are the state variables and $\phi = (e_0, u_0)$ are the control variables of the problem. The question of minimizing our cost functional $J(z, \phi)$ belongs to the class of constrained optimization problems [5,6], where the cost functional (2) is minimized with respect to the constraint given by the state system (1),

$$\text{minimize } J(z, \phi) \text{ with respect to } \phi \text{ subject to (1).} \tag{3}$$

In the sequel, we will formally introduce the Lagrangian for the problem (3).

Let Z denote the space of the state variables and C be the set of admissible controls, i.e. admissible nozzle shapes e_0 and possible input velocity

profiles u_0. To shorten the notation, we write the state system (1) together with its boundary conditions shortly as $P(z, \phi) = 0$, where $P : Z \times C \to W^*$ is called the state operator. Using a set $\xi = (\xi_u, \xi_v, \xi_e) \in W$ of Lagrangian multipliers, we introduce the Lagrangian $L : Z \times C \times W \to \mathbb{R}$ by

$$L(z, \phi, \xi) = J(z, \phi) + \langle P(z, \phi), \xi \rangle_{W^*,W} \ . \tag{4}$$

Here $\langle p, \xi \rangle_{W^*,W} \in \mathbb{R}$ denotes the duality pairing between $p \in W^*$ and $\xi \in W$.

Now, as a standard result from nonlinear optimization, the Karush–Kuhn–Tucker (KKT) system is a necessary first-order optimality condition. Assuming enough regularity, the Lagrangian is Fréchet-differentiable and the first-order optimality condition reads as $DL(z, \phi, \xi) = 0$ or componentwise

$$P(z, \phi) = 0 \quad \text{in } W^*, \tag{5a}$$
$$\partial_z P^*(\xi)[z, \phi] + \partial_z J(z, \phi) = 0 \quad \text{in } Z^*, \tag{5b}$$
$$\partial_\phi P^*(\xi)[z, \phi] + \partial_\phi J(z, \phi) = 0 \quad \text{in } C^*. \tag{5c}$$

In the system (5), we can easily identify the state (5a), adjoint (5b) and gradient (5c) in operator form.

4 Numerical Simulations

The KKT-system (5) corresponding to the first-order optimality conditions for the minimization problem (3) is a system of coupled, nonlinear PDEs. Hence, we will apply an iterative algorithm to solve them [7].

1. Given an initial controls e_0^0 and u_0^0. Set $k = 0$.
2. Solve the state equations (5a), i.e. (1) with its boundary conditions to obtain the new state variables z^{k+1}.
3. Given the state z^{k+1} corresponding to the controls e_0^k and u_0^k, solve the adjoint problem (5b) to obtain ξ^{k+1}.
4. Given ξ^{k+1}, update the control by

$$e_0^{k+1}(y) = e_0^k(y) - \xi_e^{k+1}(0, y)\, u_0^k(y) \tag{6a}$$
$$u_0^{k+1}(y) = u_0^k(y) - \xi_e^{k+1}(0, y)\, e_0^k(y) + \frac{4}{Re} \frac{\partial \xi_u^{k+1}}{\partial x}(0, y) \tag{6b}$$

5. Calculate the cost functional $J^{k+1} = J(z^{k+1}, \phi^{k+1})$.
6. If *Iteration has converged*
   then *Stop*
   else set $k = k + 1$ and go to step 2.

Since the boundaries γ_3 and γ_4 are free surfaces, it is difficult to implement the free boundary condition $\sigma \cdot n = 0$. To overcome the free surface, we transform the domain into a square domain by mapping the coordinates (x, y)

to $(x, \tilde{y})$ where $\tilde{y}(x) = \frac{y}{W(x)}$. Then, the new coordinates belong to a square domain $(x, \tilde{y}) \in [0, L] \times [-1, 1]$.

For the numerical simulations we use standard finite differences on a uniform grid with equal mesh widths in the x- and $\tilde{y}$-direction resp. The same grid is used for the state as well as for the adjoint equation. For the hyperbolic continuity equations we apply upwind methods. In the momentum equations the nonlinear terms are handled by iteration and centered differences are used to discretize the derivatives.

5 Simulation Results

Figure 2 on page 605 shows the thickness of the film for a given, constant initial thickness e_0. Figure 3 plots the transversal velocity $v(x, \cdot)$ at different lateral cuts $y = y_i$. Negative velocities imply that the fluid moves towards the centerline $y = 0$; this yields the neck-in of the film. This neck-in is also clearly visible in Fig. 4 showing the evolution of the width of the film.

Finally, we investigate the result of the optimization problem (3). The aim was to find an initial velocity profile u_0 and the shape of the nozzle, i.e. an initial thickness e_0 of the film, such that we obtain a uniform thickness e_d at the position of the spindel. Figure 5 shows the initial velocity profile u_0 before and after optimization. It is obvious that the velocity at the edge of the film

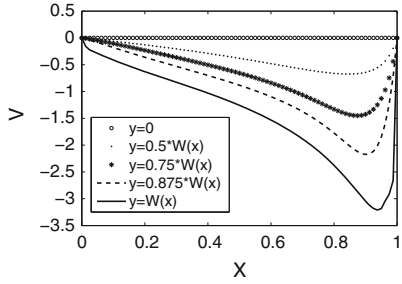


Fig. 3. Transversal velocity v

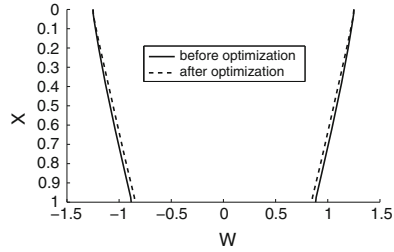


Fig. 4. Film width w

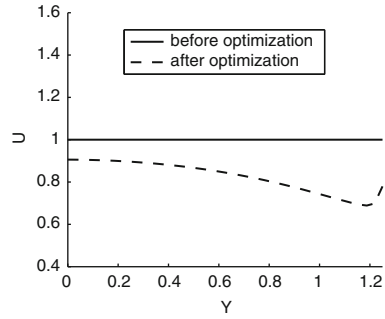


Fig. 5. Initial velocity profile u_0 before and after optimization

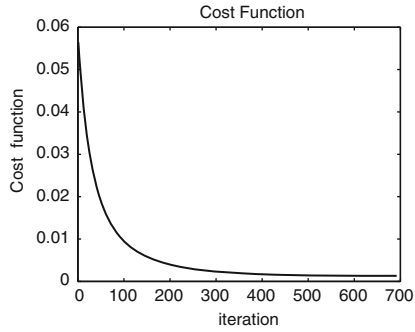


Fig. 6. Decrease of the cost functional (2) vs. iteration number

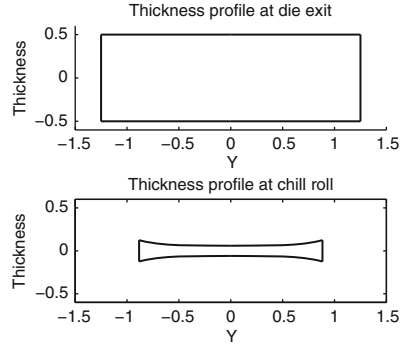


Fig. 7. Nozzle shape and film thickness at take-up before optimization

should be less compared to the center part of the film, so that the flux in the edge part will be less and it results in stopping the accumulation of fluid in the edge and removes the edge bead.

Figures 7 and 8 show a comparison of the un-optimized (left) and optimized situation (right). The initial thickness corresponding to the shape of the nozzle is shown in the upper part and the final film thickness is given below. In

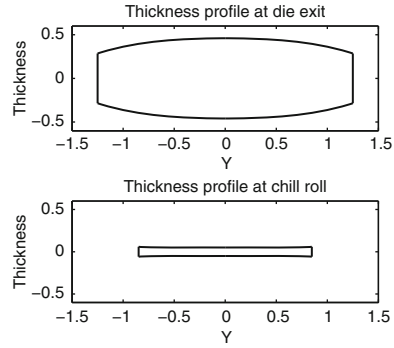


Fig. 8. Nozzle shape and film thickness at take-up after optimization

the optimized situation the close-to ellipsoidal shape of the nozzle with the modified initial velocity u_0 given in Fig. 5 counterbalances the edge-bead effect resulting in a uniform film thickness.

At the take-up point we obtain a constant film thickness of $e_d = 0.1$ corresponding to the draw ratio $D = 10$.

Figure 6 shows the decrease of the cost functional versus the iteration number. A rigorous justification of the observed convergence to the minimum based on space-mapping techniques is left open for future research. A more detailed discussion of the presented problem and the optimal control approach to its solution can be found in [4].

References

1. Barq, P., Haudin, J.M., Agassant, J.F.: Intern. Polymer Process. **8**(4), 334–349 (1992)
2. d'Halewyu, S., Agassant, J.F., Demay, Y.: Polymer Eng. Sci. **30**, 335–340 (1990)
3. Fortin, A., Carrier, P., Demay, Y.: Int. J. Numer. Meth. Fluids **20**, 31–57 (1995)
4. Götz, T., Selvanayagam, K.: e.a. Optimal Control of Film Casting Processes, Intern. J. Num. Meth. Fluids, **59**(10), 1111–1124 (2009)
5. Gunzburger, M.D., Hou, L.S., Manservisi, S., Yan, Y.: Int. J. Comp. Fluid Dyn. **11**, 181–191 (1998)
6. Hinze, M., Ziegenbalg, S.: ZAMM, **87**(6), 430–448 (2007)
7. Kelley, C.T.: Iterative methods for optimization. SIAM, Philadelphia (1999)
8. Silagy, D., Demay, Y., Agassant, J.F.: Int. J. Numer. Meth. Fluids **30**, 1–18 (1999)

On the Effect of an Atmosphere of Nitrogen on the Evaporation of Sessile Droplets of Water

S.K. Wilson[1], K. Sefiane[2], S. David[2], G.J. Dunn[1], and B.R. Duffy[1]

[1] Department of Mathematics, University of Strathclyde, Livingstone Tower 26 Richmond Street, Glasgow G1 1XH, United Kingdom
s.k.wilson@strath.ac.uk, b.r.duffy@strath.ac.uk

[2] School of Engineering and Electronics, The University of Edinburgh, The King's Buildings, Mayfield Road, Edinburgh EH9 3JL, United Kingdom
ksefiane@ed.ac.uk

Summary. The effect of an atmosphere of nitrogen on the evaporation of pinned sessile droplets of water is investigated both experimentally and theoretically.

1 Introduction

When liquid droplets are deposited on a solid substrate in an unsaturated atmosphere they will experience some degree of evaporation. This apparently simple phenomenon is encountered in everyday life as well as in a wide range of physical and biological processes. During the last decade renewed interest in droplet evaporation has been sparked by new developments in applications such as cooling technologies, desalination, painting, DNA synthesis and patterning technologies.

Many studies of the evaporation of sessile droplets have been undertaken, notably those by Picknett and Bexon [6], Bourges-Monnier and Shanahan [1], Deegan [3], Hu and Larson [5], and Popov [7]. The standard theoretical model (hereafter referred to as the "basic model") assumes that the rate-limiting mechanism for evaporation is the diffusive relaxation of the locally saturated vapour at the free surface of the droplet. The basic model decouples the concentration of vapour in the atmosphere from the temperature of the droplet and the substrate, and hence does not account for the effect of the thermal properties of the droplet and the substrate on the evaporation rate. Recently Dunn et al. [4] developed an improved mathematical model for the evaporation of a thin droplet on a thin substrate taking into account the temperature dependence of the saturation concentration of vapour at the free surface of the droplet, and found that its predictions are in reasonable agreement with the experimental results of David et al. [2]. The purpose of the present paper is to build on the progress made by David et al. [2] and Dunn et al. [4] by

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*,
Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_97,

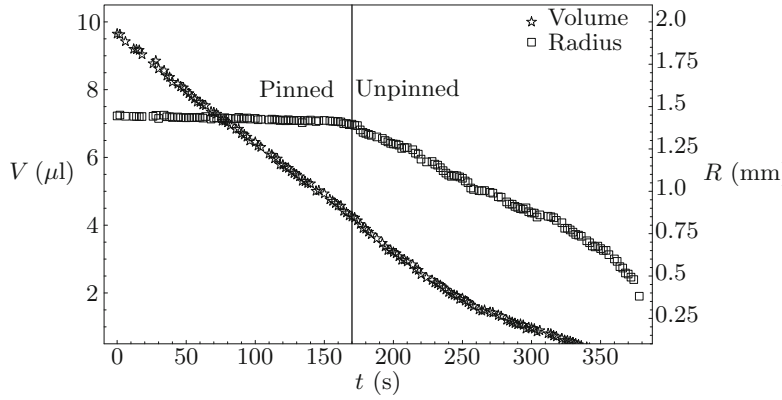


Fig. 1. Typical examples of the experimentally measured evolutions in time of the volume (*left hand axis*) and the base radius (*right hand axis*) of a droplet of water on an aluminium substrate evaporating into an atmosphere of nitrogen

investigating the effect of an atmosphere of nitrogen on the evaporation of pinned sessile droplets of water both experimentally and theoretically.

2 Experimental Procedure

The essence of the experiment consisted of depositing a liquid droplet of controlled volume on a substrate and allowing it to evaporate spontaneously. All of the experiments reported here were realised with droplets of pure deionised water resting on four different substrates chosen for their wide range of thermal conductivities, namely aluminium (Al), titanium (Ti), Macor and PTFE. The substrates had dimensions of $10\times10\times1$ mm (length $\times$ width $\times$ thickness), and their thermal conductivities are given by David et al. [2, Table 2]. In order to contain the ambient gas and to vary the atmospheric pressure, the droplet and the substrate were placed in a "low pressure" chamber. The chamber was cylindrical in shape (105 mm diameter and 95 mm height) with two observation windows and was connected to a gas supply and a vacuum pump. The experimental setup used a DSA100$^{\mathrm{TM}}$ Droplet Shape Analysis (DSA) system from KRÜSS GmbH to monitor the evolutions in time of the volume, contact angle, height and base radius of the droplet. Typical examples of the experimentally measured evolutions of the volume and the base radius are shown in Fig. 1. All of the experiments were carried out in a laboratory in which the room temperature was controlled at 295 K with an air-conditioning unit with a precision of ±1 K. Before each experiment, air was removed from the chamber and replaced with the chosen ambient gas. The pressure of the gas was varied in the range 40–1,000 mbar. Various ambient gases were used, but, for brevity, only results for an atmosphere of nitrogen are reported here.

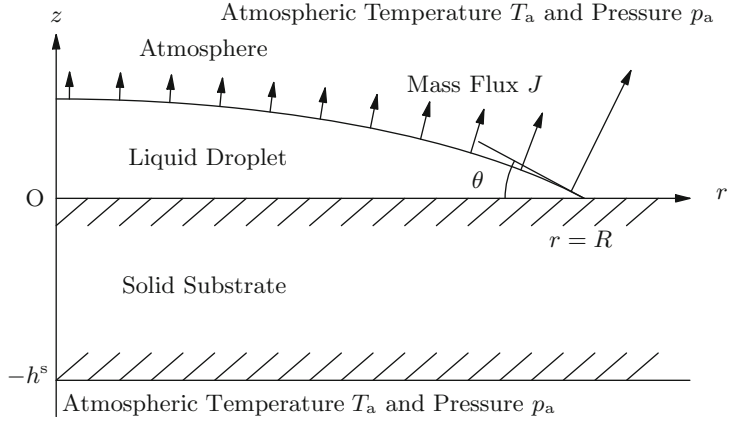


Fig. 2. Geometry of the mathematical model

3 Mathematical Model

The mathematical model used in the present work represents the quasi-steady diffusion-limited evaporation of an axisymmetric droplet of Newtonian fluid with constant viscosity, density ρ, surface tension γ, and thermal conductivity k resting on a horizontal substrate of constant thickness h^s with constant thermal conductivity k^s. Referred to cylindrical polar coordinates (r, θ, z) with origin on the substrate at the centre of the droplet with the z axis vertically upwards, the shape of the free surface of the droplet at time t is denoted by $z = h(r, t)$, the upper surface of the substrate by $z = 0$, and the lower surface of the substrate by $z = -h^s$, as shown in Fig. 2.

For a sufficiently small droplet whose base radius R is much less than the capillary length $\sqrt{\gamma/\rho g}$, the droplet shape can be approximated as a simple quasi-static spherical cap, and hence the relation between the volume $V = V(t)$ and the contact angle $\theta = \theta(t)$ is given by

$$V = \frac{\pi h_m (3R^2 + h_m^2)}{6}, \tag{1}$$

where $h_m = h_m(t) = h(0, t) = R \tan(\theta/2)$ is the maximum height of the droplet. The total evaporation rate is given by

$$-\frac{dV}{dt} = \frac{2\pi}{\rho} \int_0^R J \sqrt{1 + \left(\frac{\partial h}{\partial r}\right)^2}\, r\, dr, \tag{2}$$

where $J = J(r, t)\,(\geq 0)$ is the local evaporative mass flux from the droplet.

The atmosphere in the chamber surrounding the droplet and the substrate is assumed to be at constant atmospheric temperature T_a and atmospheric pressure p_a. The temperatures of the droplet and the substrate, denoted by

$T = T(r,z,t)$ and $T^{\rm s} = T^{\rm s}(r,z,t)$, respectively, satisfy Laplace's equation $\nabla^2 T = \nabla^2 T^{\rm s} = 0$. The mass flux from the droplet satisfies the local energy balance $\mathcal{L}J = -k\nabla T \cdot \mathbf{n}$ on $z = h$ for $r < R$, where $\mathcal{L}$ is the latent heat of vaporisation and $\mathbf{n}$ is the unit outward normal to the free surface of the droplet. We assume that the temperature and the heat flux are continuous between the droplet and the substrate, $T = T^{\rm s}$ and $-k\partial T/\partial z = -k^{\rm s}\partial T^{\rm s}/\partial z$ on $z = 0$ for $r < R$, and that the temperature is continuous between the substrate and the atmosphere, $T^{\rm s} = T_{\rm a}$ on $z = 0$ for $r > R$ and on $z = -h^{\rm s}$.

Assuming that transport of vapour in the atmosphere is quasi-static and is solely by diffusion, the concentration of vapour, denoted by $c = c(r,z,t)$, satisfies Laplace's equation $\nabla^2 c = 0$. At the free surface of the droplet we assume that the atmosphere is saturated with vapour and hence $c = c_{\rm sat}(T)$ on $z = h$ for $r < R$, where the saturation value of the concentration $c_{\rm sat} = c_{\rm sat}(T)$ is an increasing function of temperature, approximated quartically in $T_{\rm a} - T$ by

$$c_{\rm sat}(T) = \sum_{i=0}^{4} \alpha_i (T_{\rm a} - T)^i, \tag{3}$$

where the coefficients α_i for $i = 0,\ldots,4$ were chosen to fit experimental data given by Raznjevic [8], leading to $\alpha_0 = 1.93 \times 10^{-2}$, $\alpha_1 = 1.11 \times 10^{-3}$, $\alpha_2 = 2.78 \times 10^{-5}$, $\alpha_3 = 3.78 \times 10^{-7}$ and $\alpha_4 = 2.59 \times 10^{-9}$ in units of kg m^{-3} K^{-i}. Note that while a simple linear approximation is sufficient for situations with a relatively small evaporative cooling of a few degrees K, such as those considered by David et al. [2] and Dunn et al. [4], the quartic approximation (3) is necessary for situations with a larger evaporative cooling of up to 20 K, such as those considered in the present work. On the dry part of the substrate there is no mass flux, $\partial c/\partial z = 0$ on $z = 0$ for $r > R$, and, since the chamber is much larger than the droplets used in the experiments, far from the droplet the concentration of vapour approaches its far-field value of zero, $c \to 0$ as $(r^2 + z^2)^{1/2} \to \infty$. Once c is known the local evaporative mass flux from the droplet is given by $J = -D\nabla c \cdot \mathbf{n}$ on $z = h$ for $r < R$, where D is the coefficient of diffusion of vapour in the atmosphere. A standard result from the theory of gases is that D is inversely proportional to pressure, and hence we write $D = D_{\rm ref} p_{\rm ref}/p_{\rm a}$, where $D_{\rm ref}$ denotes the appropriate reference value of D at the reference pressure $p_{\rm ref} = 1$ atm. Note that the diffusion coefficient is the only parameter in the model that depends on either the nature of the ambient gas or its pressure $p_{\rm a}$.

In the special case $c_{\rm sat} \equiv c_{\rm sat}(T_{\rm a})$, corresponding to $\alpha_i = 0$ for $i = 1,\ldots,4$ in (3), the saturation concentration is constant and we recover the basic model in which the problem for the concentration of vapour in the atmosphere is decoupled from the problem for the temperature of the droplet and the substrate.

The model was solved numerically using the MATLAB-based finite element package COMSOL Multiphysics. The value of $D_{\rm ref}$ used to obtain the present numerical results was fitted by comparing the experimental results for evaporation on an aluminium substrate with the corresponding theoretical

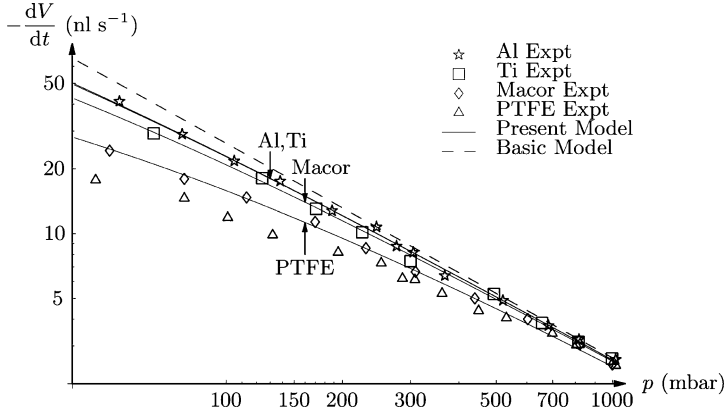


Fig. 3. Experimentally measured evaporation rates of droplets of water in an atmosphere of nitrogen on various substrates for different atmospheric pressures, together with the corresponding theoretical predictions of the mathematical model and the basic model

predictions. Specifically, for an atmosphere of nitrogen the fitted value of $D_{\mathrm{ref}} = 2.15 \times 10^{-5}$ m^2 s^{-1} differs by less than 15% from the value of $D_{\mathrm{ref}} = 2.47 \times 10^{-5}$ m^2 s^{-1} given by Reid et al. [9], i.e. the difference is comparable with the uncertainty in the theoretical value.

4 Results

As the typical experimentally measured evolutions in time of the volume and the base radius of a droplet shown in Fig. 1 illustrate, typically the evaporation process can be divided into two stages. In the first stage, the droplet is pinned and so the base radius is constant while the volume decreases approximately linearly with time. In the second stage, the droplet depins and so the base radius and the volume decrease until complete evaporation. The experimental and theoretical results presented here are for the first stage only, for which the basic model and previous experimental studies (such as, for example, that by David et al. [2]) indicate that the evaporation rate is proportional to the perimeter of the base of the droplet.

Figure 3 shows both the experimental results and the corresponding theoretical predictions of the mathematical model for all four substrates studied for an atmosphere of nitrogen, and shows that the theoretical predictions of the mathematical model using the value of the diffusion coefficient fitted for an aluminium substrate are in reasonable agreement with the experimental results for the other three substrates studied. Figure 3 also shows the corresponding prediction of the basic model, which are independent of the thermal properties of the droplet and the substrate, which consistently over-predicts

the evaporation rate. In particular, Fig. 3 shows that for all four substrates studied reducing the atmospheric pressure increases the evaporation rate. Furthermore, Fig. 3 also shows that droplets on substrates with higher thermal conductivities evaporate more quickly than those on substrates with lower thermal conductivities. Close inspection of Fig. 3 reveals that the agreement between theory and experiment is poorest for the substrate with the lowest thermal conductivity (namely PTFE).

5 Conclusions

An investigation into the effect of an atmosphere of nitrogen on the evaporation of pinned sessile droplets of water has been described. The experimental work investigated the evaporation rates of sessile droplets at reduced pressure using four different substrates with a wide range of thermal conductivities. Reducing the atmospheric pressure increases the diffusion coefficient of water vapour in the atmosphere and hence increases the evaporation rate. A mathematical model that takes into account the effect of the atmospheric pressure and the nature of the ambient gas on the diffusion of water vapour in the atmosphere was developed, and its predictions were found to be in encouraging agreement with the experimental results.

The present work was supported by the United Kingdom Engineering and Physical Sciences Research Council (EPSRC) via joint research grants GR/S59444 (Edinburgh) and GR/S59451 (Strathclyde).

References

1. Bourgès-Monnier, C., Shanahan, M. E. R.: Langmuir **11**, 2820 (1995)
2. David, S., Sefiane, K., Tadrist, L.: Colloids Surfaces A Physiochem. Eng. Asp. **298**, 108 (2007)
3. Deegan, R.: Phys. Rev. E **61**, 475 (1998)
4. Dunn, G.J., Wilson, S.K., Duffy, B.R., David, S., Sefiane, K.: Colloids Surfaces A Physiochem. Eng. Asp. **323**, 50 (2008)
5. Hu, H., Larson, R.G.: J. Phys. Chem. B **106**, 1334 (2002)
6. Picknett, R.G., Bexon, R.: J. Coll. Int. Sci. **61**, 336 (1977)
7. Popov, Y.O.: Phys. Rev. E **71**, 036313 (2005)
8. Raznjevic, K.: Handbook of Thermodynamic Tables. Begell House, New York (1995)
9. Reid, R.C., Prausnitz, J.M., Poling, B.E.: The Properties of Gases and Liquids, 4th edn. McGraw-Hill, New York (1987)

Similarity Solutions for Unsteady Rivulets

Y.M. Yatim, S.K. Wilson, B.R. Duffy, and R. Hunt

Department of Mathematics, University of Strathclyde, Livingstone Tower, 26 Richmond Street, Glasgow G1 1XH, United Kingdom
yazariah.mohd-yatim@strath.ac.uk, s.k.wilson@strath.ac.uk, b.r.duffy@strath.ac.uk

Summary. Similarity solutions representing unsteady gravity-driven flow of thin slender non-uniform rivulets down an inclined plane are described.

1 Introduction

Rivulets occur in a wide range of geophysical, biological and industrial contexts, ranging from the flow of lava to the flow of complex fluids in industrial devices such as condensers and heat exchangers; as a consequence there have been many studies of both steady and unsteady flows of thin rivulets.

Smith [6] obtained a similarity solution describing steady gravity-driven flow of a slender non-uniform rivulet of a Newtonian fluid down an inclined plane when surface-tension effects are negligible, and several papers concerning other steady rivulet flows have been written in the spirit of Smith's [6] analysis, including those by Duffy and Moffatt [1], Wilson and Duffy [8] and Wilson, Duffy and Hunt [9].

A similarity solution representing unsteady heat conduction from an instantaneous point source when the diffusivity is temperature dependent was obtained by Zel'dovich and Kompaneets [10] (and independently by Pattle [5] in another context). Smith [7] adapted this solution to the spreading of a thin drop of constant volume over a horizontal plane, and Huppert [2] generalized it to the case when the fluid is supplied from a time-dependent source. Huppert [3] obtained a similarity solution describing unsteady two-dimensional flow of a fluid film down an inclined plane, and Lister [4], in an extensive study of unsteady thin-film flow down an inclined plane from a point or line source, obtained similarity solutions valid for short times and long times. As part of his study, Lister [4] gave the asymptotic form of the appropriate thin-film equation for flow from a point source at large times (his equation (2.10b)); it is solutions of this equation that we investigate in detail in the present paper, in which we obtain similarity solutions for unsteady gravity-driven flow of a thin slender rivulet of Newtonian fluid down an inclined plane.

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*, Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_98,

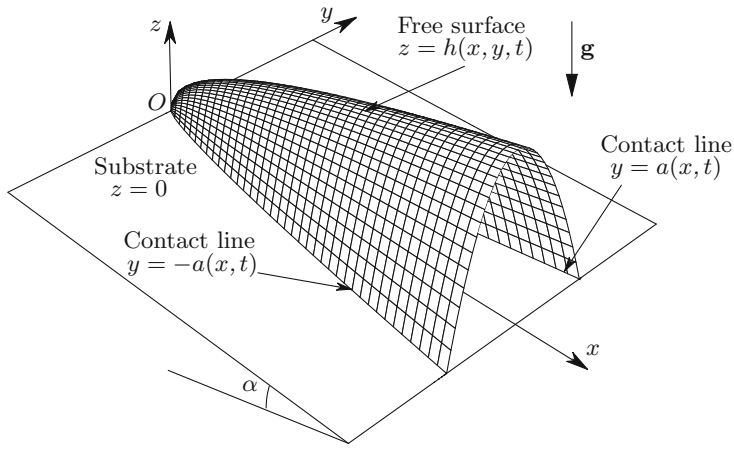


Fig. 1. Sketch of the geometry of the problem

2 Problem Formulation

Consider unsteady flow of a thin rivulet of Newtonian fluid with constant density ρ and viscosity μ driven by gravity g down a plane inclined at an angle α $(0 < \alpha < \pi)$ to the horizontal. Cartesian coordinates $Oxyz$ with the x axis down the line of greatest slope and the z axis normal to the substrate $z = 0$ are adopted, and we denote the (unknown) free surface of the rivulet by $z = h(x, y, t)$, where t denotes time. We restrict attention to flows that are symmetric about $y = 0$, with (unknown) semi-width $a = a(x, t)$. The geometry of the problem is sketched in Fig. 1.

With the familiar lubrication approximation the velocity and pressure of the fluid are determined in terms of h, which, for the case in which the rivulet is slender, satisfies the partial differential equation

$$3\mu h_t = \rho g \cos\alpha \left[h^3 h_y\right]_y - \rho g \sin\alpha \left[h^3\right]_x , \tag{1}$$

which is Lister's [4] (2.10b). The conditions of zero height and of zero mass flux at the contact lines $y = \pm a$ lead to

$$h = 0 \quad \text{at} \quad y = \pm a, \qquad h^3 h_y \to 0 \quad \text{as} \quad y \to \pm a. \tag{2}$$

We seek an unsteady similarity solution to (1) in the form

$$h = h_0 |x|^p |t|^q H(\eta), \qquad y = y_0 |x|^r |t|^s \eta, \tag{3}$$

where p, q, r, s, h_0, y_0 and the dimensionless function $H = H(\eta)$ (≥ 0) are to be determined. With (3), the terms in (1) balance provided that $p = \frac{1}{2}$, $q = -\frac{1}{2}$, $r = \frac{3}{4}$ and $s = -\frac{1}{4}$, and if we therefore write (3) in the form

$$h = \left(\frac{\mu|x|}{\rho g \sin\alpha |t|}\right)^{\frac{1}{2}} H(\eta), \qquad y = \left(\frac{4\mu\cos^2\alpha |x|^3}{9\rho g \sin^3\alpha |t|}\right)^{\frac{1}{4}} \eta, \tag{4}$$

then (1) reduces to a second-order ordinary differential equation for H, namely

$$S_t\left[\frac{1}{2}\eta H' - H\right] = S_g\left[H^3 H'\right]' + S_x\left[\frac{1}{2}\eta\left(H^3\right)' - H^3\right], \tag{5}$$

where a dash denotes differentiation with respect to η, and we have introduced the notation $S_t = \mathrm{sgn}(t) = \pm 1$, $S_g = \mathrm{sgn}(\cos\alpha) = \pm 1$ and $S_x = \mathrm{sgn}(x) = \pm 1$. For a symmetric rivulet, regular at $y = 0$, appropriate boundary conditions are

$$H = H_0, \quad H' = 0 \quad \text{at} \quad \eta = 0, \tag{6}$$

where the parameter H_0 (> 0) is to be determined. The position where $H = 0$ is denoted $\eta = \eta_0$ (corresponding to the contact-line position $y = a$), so that

$$H = 0 \quad \text{at} \quad \eta = \eta_0, \qquad H^3 H' \to 0 \quad \text{as} \quad \eta \to \eta_0. \tag{7}$$

At any time t the rivulet widens or narrows according to $|x|^{3/4}$ and thickens or thins according to $|x|^{1/2}$, and at any station x the rivulet widens or narrows according to $|t|^{-1/4}$ and thickens or thins according to $|t|^{-1/2}$. The 'nose' of the rivulet remains fixed at the origin O for all t.

3 Solution of the System (5)–(7)

A closed-form solution of the ordinary differential equation (5) is not available, and so it must, in general, be solved numerically for H subject to the boundary conditions (6) and (7), where H_0 and η_0 are (positive) parameters to be determined. Without loss of generality, we may choose $S_t = 1$ (i.e. $t \geq 0$) when solving (5). (The case $S_t = -1$, i.e. $t < 0$, has a different physical interpretation from the case $S_t = 1$; this will be discussed briefly in Sect. 4.) Thus, in principle, with two choices for each of S_g and S_x, there are four cases to consider; however, it turns out that the system (5)–(7) has solutions in only one of these cases, namely the one with $S_g = -1$ and $S_x = 1$, and so from now on we shall consider only that case.

Equation (5) with $S_t = 1$, $S_g = -1$ and $S_x = 1$ was solved numerically for H subject to (6) for a given value of H_0 by means of a shooting technique, the value of η_0 being determined as the point where $H = 0$, by (7a). It was found that there is a solution for all $H_0 > 0$ except in a narrow 'window' near $H_0 = 1$ given by $H_{01} < H_0 < H_{02}$, where $H_{01} \simeq 0.9995$ and $H_{02} \simeq 1.1059$, in which there is no solution.

The relation between H_0 and η_0 is not monotonic: for any given value of H_0 outside the interval $H_{01} < H_0 < H_{02}$ there is a corresponding unique value of η_0, but for any given η_0 there can be zero, one, two or three solutions, depending on the value of η_0.

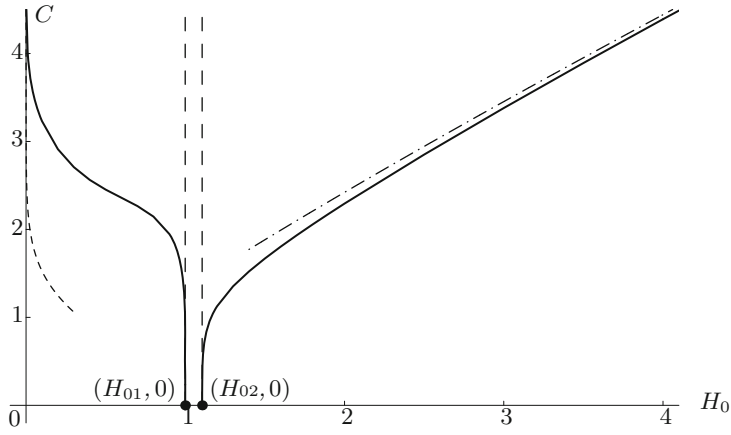


Fig. 2. Plot of C as a function of H_0 (*full line*), together with the leading order asymptotic solutions (11) in the limit $H_0 \to 0^+$ (*dashed line*) and (12) in the limit $H_0 \to \infty$ (*dashed-dotted line*); here $H_{01} \simeq 0.9995$ and $H_{02} \simeq 1.1059$

It is found that H satisfies

$$H = H_0 + \frac{1 - H_0^2}{2H_0^2}\eta^2 + O\left(\eta^4\right) \tag{8}$$

as $\eta \to 0$, and either

$$H \sim \left[\frac{3}{2}\eta_0\left(\eta_0 - \eta\right)\right]^{\frac{1}{3}} \tag{9}$$

or

$$H \sim C\left(\eta_0 - \eta\right)^{\frac{1}{4}} \tag{10}$$

as $\eta \to \eta_0^-$, where C is a positive constant. In the limit $H_0 \to 0^+$ the solution comprises an outer region in which $H = \left(3\eta^2/5\right)^{1/3}$ at leading order, together with two boundary layers, one near $\eta = 0$ of width $O(H_0^{3/2})$ in which $H = O(H_0)$, and another near the contact line $\eta = \eta_0 \to \infty$ of width $O(\eta_0^{-1/3})$ in which $H = O(\eta_0^{2/3})$; in particular, it is found that $\eta_0 \sim (-K \log H_0)^{3/4} \to \infty$ as $H_0 \to 0^+$, where $K \simeq 0.8498$. In the limit $H_0 \to \infty$ we have $H = O(H_0)$ and $\eta_0 = O(H_0^{1/2}) \to \infty$.

Thus far we have obtained a one-parameter family of solutions of (5), (6) and (7a), parameterised by H_0, and with η_0 determined in terms of H_0. However, this does not yet answer the problem of determining all physically sensible solutions of (1) of the form (3); to do this we must also impose condition (7b), or equivalently the condition $C = 0$. Figure 2 shows a plot of C as a function of H_0, together with appropriate leading order asymptotic forms of C, given by

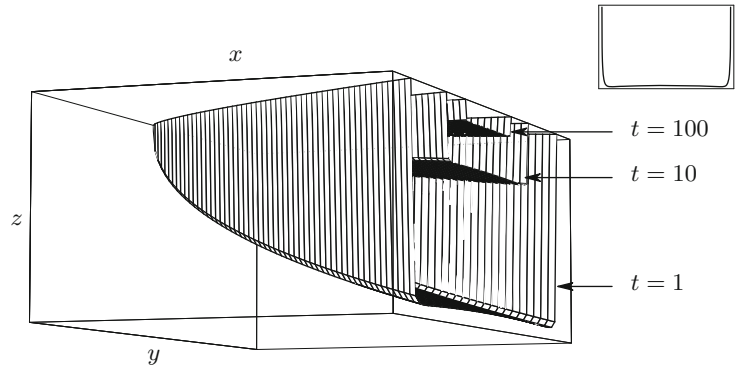


Fig. 3. Three-dimensional plot of the free surface $z = h$ of a pendent rivulet represented by the similarity solution (4) in which H satisfies (5)–(7) with $S_t = 1$, $S_g = -1$, $S_x = 1$ and $H_0 = H_{01}$, at times $t = 1$, 10 and 100. The *inset* shows the cross-sectional profile H

$$C \sim \left(\tfrac{6}{5}\right)^{\frac{1}{4}} (-K \log H_0)^{\frac{9}{16}} \simeq 0.9551\,(-\log H_0)^{\frac{9}{16}} \to \infty \tag{11}$$

in the limit $H_0 \to 0^+$, and

$$C \sim 1.3205 H_0^{\frac{7}{8}} \to \infty \tag{12}$$

in the limit $H_0 \to \infty$. Figure 2 shows that $C = 0$ at $H_0 = H_{01}$ and at $H_0 = H_{02}$, and is non-zero for all other values of H_0. We thus arrive at our main conclusion: there are similarity solutions of the type sought only for $H_0 = H_{01} \simeq 0.9995$ and $H_0 = H_{02} \simeq 1.1059$.

4 Conclusions

We have obtained similarity solutions of the form (4) (where $H(\eta)$ satisfies the system (5)–(7)), describing the free-surface profiles of thin slender rivulets undergoing unsteady gravity-driven flow down an inclined plane.

The choice $S_t = 1$, $S_g = -1$, $S_x = 1$ corresponds to pendent rivulets in $x > 0$ with $t > 0$. Figures 3 and 4 show three-dimensional plots of the free-surface profile $z = h$ (suitably non-dimensionalised) in the two cases for which solutions exist, namely $H_0 = H_{01} \simeq 0.9995$ and $H_0 = H_{02} \simeq 1.1059$, respectively, at various times. As shown in the insets, for $H_0 = H_{01}$ the cross-sectional profile is (barely discernibly) 'double-humped', and for $H_0 = H_{02}$ it is 'single-humped', in agreement with (8). At any time t (> 0) the rivulets widen according to $x^{3/4}$ and thicken according to $x^{1/2}$, and at any station x (> 0) they narrow according to $t^{-1/4}$ and thin according to $t^{-1/2}$.

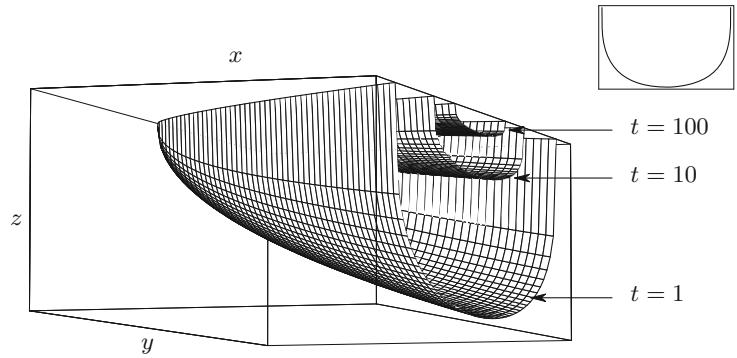


Fig. 4. As in Fig. 3, but with $H_0 = H_{02}$

The alternative choice $S_t = -1$, $S_g = 1$, $S_x = -1$ leads to exactly the same mathematical problem as that discussed above, but the physical interpretation is rather different, since it corresponds to sessile rivulets in $x < 0$, with $t < 0$. At any time t (< 0) the rivulets become thinner and narrower with increasing x (< 0), but at any x (< 0) they become wider and thicker as time elapses. In fact, in this interpretation the solutions exhibit a finite-time singularity, at $t = 0$, with h becoming infinite everywhere then.

There remains the question of whether rivulet solutions of the above types could occur in practice. In particular, the stability of the solutions is, of course, crucial, but it seems that even a restricted linear stability analysis is likely to be a formidable task.

Acknowledgement

The first author (YMY) wishes to thank the Ministry of Higher Education, Malaysia and University of Science, Malaysia for financial support via an Academic Staff Training Fellowship.

References

1. Duffy, B.R., Moffatt, H.K.: Euro. J. Appl. Math. **8**, 37–47 (1997)
2. Huppert, H.E.: J. Fluid Mech. **121**, 43–58 (1982)
3. Huppert, H.E.: Nature **300**, 427–429 (1982)
4. Lister, J.R.: J. Fluid Mech. **242**, 631–653 (1992)
5. Pattle, R.E.: Q. Jl Mech. Appl. Math. **12**, 407–409 (1959).
6. Smith, P.C.: J. Fluid Mech. **58**, 275–288 (1973)
7. Smith, S.H.: J. Appl. Math. Phys. **20**, 556–560 (1969)
8. Wilson, S.K., Duffy, B.R.: J. Engng. Math. **42**, 359–372 (2002)
9. Wilson, S.K., Duffy, B.R., Hunt, R.: Q. Jl Mech. Appl. Math. **55**, 385–408 (2002)
10. Zel'dovich, Ya.B., Kompaneets, A.S.: Izv. Akad. Nauk SSSR 61–71 (1950)

Depinning of 2d and 3d Droplets Blocked by a Hydrophobic Defect

P. Beltrame[1,2], P. Hänggi[1], E. Knobloch[3], and U. Thiele[4]

[1] Institut für Physik, Universität Augsburg, D-86135 Augsburg, Germany
Philippe.Beltrame@avignon.inra.fr,
Peter.Hanggi@physik.uni-augsburg.de
[2] Dept. de Physique, Universite d'Avignon, F-84000 Avignon, France
[3] Department of Physics, University of California, Berkeley CA 94720, USA
knobloch@berkeley.edu
[4] Department of Mathematical Sciences, Loughborough University, Loughborough, Leicestershire, LE11 3TU, UK u.thiele@lboro.ac.uk

Summary. On non-ideal real substrates the onset of droplet motion under lateral driving is strongly influenced by substrate defects. A finite driving force is necessary to overcome the pinning influence of microscale heterogeneities. The dynamics of depinning two- and three-dimensional droplets is studied using a long-wave evolution equation for the film thickness profile in the case of a localized hydrophobic wettability defect. It is found that the nature of the depinning transition explains the experimentally observed stick-slip motion.

1 Introduction

Both steady states and the evolution in time of droplets or liquid films on solid substrates in the limit of small contact angles and surface slopes are described well by the thin film or lubrication equation [13, 19]

$$\partial_t h \;=\; -\nabla \cdot [m(h)\nabla p(h) + \mu(h)\mathbf{e_x}]\,, \tag{1}$$

where $h(x, y, t)$ is the thickness profile, $m(h)$ is the mobility, and $\mu(h)$ represents a lateral driving force.

For a droplet on an incline this force might be due to gravity. In other situations similar forces arise as the result of rotation (centrifugal force), or gradients in wettability, although temperature and electrical field gradients can also introduce lateral forces into the problem. The pressure $p(h)$ may contain several terms, e.g., curvature pressure (capillary) or a thickness-dependent disjoining pressure $\Pi(h)$ modeling the effective molecular interactions between substrate and film surface (wettability), for example, due to van der Waals interactions [5, 12]. Other contributions may arise from electrostatic fields [16, 18, 38], thermal effects [2, 20, 37] or hydrostatics [4, 6, 9].

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*,
Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_99,

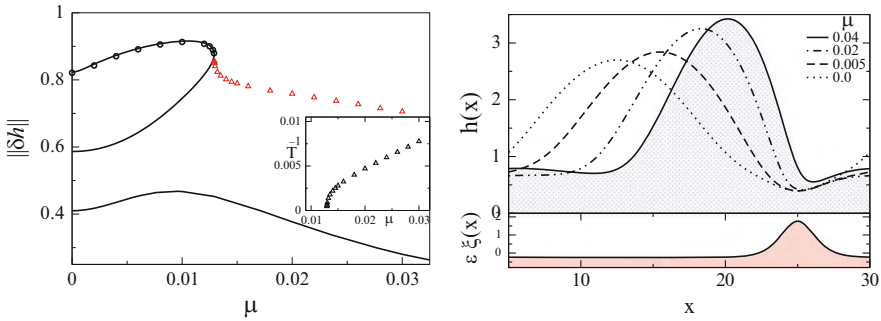


Fig. 1. (**a**) Bifurcation diagram for a droplet that depins via a sniper bifurcation. The localized hydrophobic defect with $s = 6$, $\epsilon = 0.4$ and $b = 0.1$ pins the droplet until depinning occurs at the lateral driving force $\mu_c \approx 0.014$. For details see main text. (**b**) Thickness profiles of pinned droplets for $\mu < \mu_c$ for $\epsilon = 1$. The lower panel gives the profile of the heterogeneity $\xi(x)$. The domain size is $L = 25$ and the liquid volume is $V = 37.5$

The system behavior is well studied for smooth homogeneous substrates. Without lateral driving force, an unstable film may structure via a long-wave instability resulting in patterns of holes, drops or labyrinths. The emerging structure sizes depend on the mean film thickness and the type of destabilizing influence [2,23,26,27,29]. One finds a similar situation for systems involving lateral driving forces such as gravity for a film on an incline [11,21]. The lateral driving gives rise to phenomena like transverse front instabilities [8,11,28,31]. A few studies focus on films and drops on heterogeneous substrates without lateral driving [3,14,30]. However, relatively little is known about the interplay of lateral driving and substrate heterogeneities. This is an important problem as such heterogeneities may cause stick-slip motion [25] or roughening [10,24] of moving contact lines, and are thought to be responsible for contact angle hysteresis [5,15,22].

Recently, [32, 33] studied the problem employing a dynamical systems approach based on thin film theory. In particular they use a wettability (disjoining pressure) that depends on the location on the substrate. In this way localized hydrophilic or hydrophobic substrate defects can be modeled. Constructing an idealized periodically heterogeneous substrate one can then employ tools of dynamical systems theory to study steady droplet constellations under small driving and the dynamics of the depinning process at a larger lateral driving. The dynamical approach has the advantage over static variational methods [17] in that it allows one to investigate the evolution of droplet shapes and stability as a function of the driving and the dynamics of the depinning process itself. In the following we will present a selection of results on the depinning of 2d droplets [33] and add material for the depinning of 3d droplets pinned by a line defect. As an example we use a hydrophobic defect that blocks the droplet at its front. For the case of a hydrophilic defect we refer to [1,32,33].

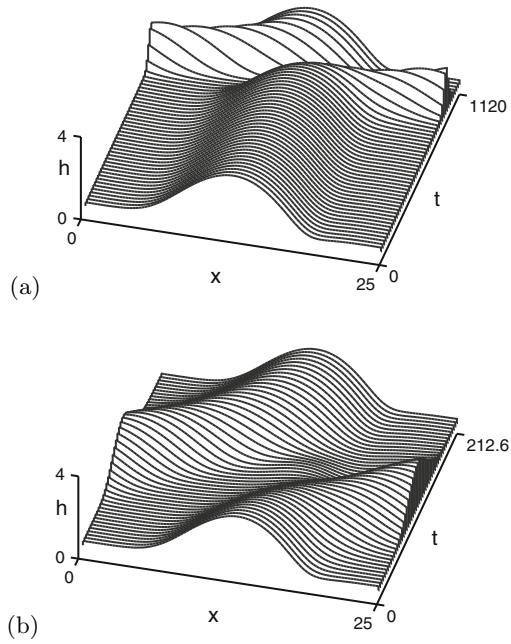


Fig. 2. The droplet dynamics beyond depinning is shown as space-time plot for one period in space and time (**a**) close to depinning at $\mu = 0.013$ with a temporal period of $T = 1119.9$, and (**b**) far from depinning at $\mu = 0.02$ ($T = 212.6$). The remaining parameters are as in Fig. 1a

2 Model

To model a wettability defect we let the short-range part of the disjoining pressure Π depend on the coordinate x (using expressions as in [33]). The resulting film evolution equation in dimensionless form for 3d droplets is

$$\partial_t h \;=\; -\nabla \cdot \left\{ h^3 \left[\nabla \left(\Delta h + \Pi(h,x) \right) + \mu \mathbf{e_x} \right] \right\}, \tag{2}$$

where

$$\Pi(h,x) \;=\; \frac{b}{h^3} - \left[1 + \epsilon \xi(x) \right] e^{-h} \tag{3}$$

is the dimensionless disjoining pressure (for details see [33, 35]). We chose $\xi(x) \;=\; \{2\,\mathrm{cn}[2K(k)x/L, k]\}^2 - \Delta$ with $K(k)$ being the complete elliptic integral of the first kind and Δ is either zero or chosen in such a way that the average of $\xi(x)$ is zero. L is the system size. For $k = 0$ the profile is sinusoidal whereas for $k \to 1$ one obtains for $\epsilon > 0$ localized hydrophobic defects. We further introduce the logarithmic measure $s \equiv -\log(1-k)$.

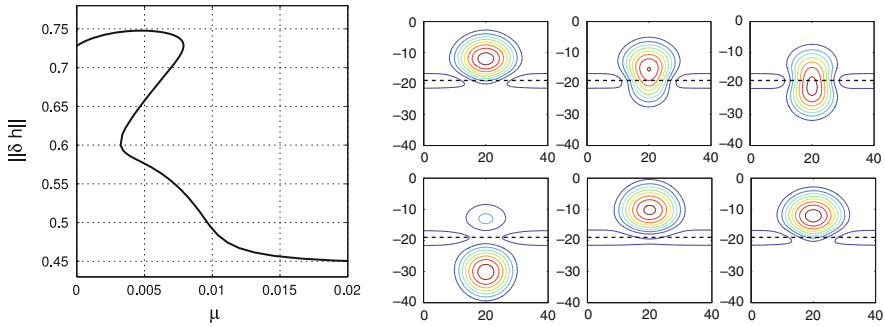


Fig. 3. (**a**) Bifurcation diagram for a 3d droplet for increasing lateral driving force μ and $\epsilon = 0.3$. The steady droplet solutions are characterized by their L^2 norm $||\delta h||$. (**b**) Shown are snapshots of height contour lines characterizing the droplet dynamics beyond their depinning via a sniper bifurcation for $\mu = 8.0 \times 10^{-3}$. From *top left* to *bottom right* droplets are shown at times $t = 0, 1160, 1240, 1340, 1680, 2930$. The position of the line defect is indicated by a *straight horizontal line*. The domain size is 40×40, $V = 1600$, and the distance between the contour lines $\Delta h = 0.4$

3 Results

Under lateral driving ($\mu > 0$) no droplet remains at rest on a homogeneous substrate ($\epsilon = 0$), i.e., a substrate without lateral variation or defect. All such droplets will (in the lubrication limit) slide with a constant velocity that is determined by the driving force, the properties of the liquid, and the wettability [34, 36]. The situation is very different on a heterogeneous substrate. There drops are pinned for small driving, and depin at a critical driving μ_c. For larger driving the droplets slide with a profile that is modulated when passing a defect.

Figure 1a presents the corresponding bifurcation diagram. It gives as solid line the L^2 norm for steady droplet solutions obtained by continuation [7], selected steady solutions as obtained by time integration (circles) and the time-averaged L^2 norm for the unsteady sliding droplet solutions beyond depinning (triangles). The inverse time period for the latter is given as inset. Its $(\mu - \mu_c)^{1/2}$ dependence indicates that the bifurcation represents a Saddle Node Infinite PERiod (SNIPER) bifurcation. Selected steady profiles before depinning ($\mu < \mu_c$) are shown in Fig. 1b. The time evolution beyond depinning is represented in the form of space-time plots for a typical stick-slip motion close to the sniper bifurcation (Fig. 2a) and for a larger force where droplet motion is more continuous (Fig. 2b).

Recently a path-following algorithm has been developed for three-dimensional droplets [1]. It has been shown that all main results on depinning obtained for 2d droplets hold as well for 3d droplets. Figure 3a shows the bifurcation diagram for a single droplet on a square domain blocked by a

hydrophobic stripe-like defect. It strongly resembles the corresponding result for the 2d case (Fig. 1a). Time simulations show that depinning occurs via a SNIPER bifurcation in this case as well. An example of a time series for a stick-slip motion of a single droplet is given in Fig. 3b as a series of snapshots. Note that the times at which the snapshots are taken are not equidistant. It takes the droplet about 1200 time units to slowly let an advancing 'protrusion' creep over the defect (snapshot 1–2). Then within 500 units it depins and slides to the next defect (snapshot 2–5), where it needs another 1300 units to reach again the same state as in snapshot 1 (snapshot 5–6). All together for the chosen value of μ the ratio of stick phase to slip phase is about 5:1. The ratio becomes larger if one approaches the bifurcation point.

4 Conclusion

We have reviewed recent work on the depinning dynamics of the depinning of two- and three-dimensional droplets under lateral driving. Here, we have focused on one type of defect (hydrophobic, localized) and one type of depinning transition (SNIPER). For results on hydrophilic defects and for larger drops see [1, 32, 33].

We have found that the depinning behavior is very similar for 2d and 3d droplets: Droplets are pinned up to a critical driving strength μ_c where they depin via a SNIPER bifurcation characterized by a square-root dependence of the inverse time scale of depinning on the distance $\mu - \mu_c$. Slightly above the bifurcation the unsteady motion resembles the stick-slip motion observed in experiment: The advancing motion is extremely slow when the drop 'creeps' over a hydrophobic defect, and very fast once the drop breaks away from the defect and slides to the next one. The difference in time scales for the stick- and the slip-phase can be many orders of magnitude.

Note that at very large driving depinning might as well occur via a Hopf instead of a SNIPER bifurcation [32]. It is thought that then the depinning is actually caused by the flow in the wetting layer, an effect that will for realistic forces not be observed for partially wetting nano- or micro-droplets on an incline with wettability defects. However, for dielectric liquids a thick wetting layer of 100 nm to 1 μm stabilized by van der Waals interaction can coexist with micro-droplets generated by an electric field [16, 18], and both depinning mechanisms should be observable using gravity as the driving force (see appendix of [33]).

We acknowledge support by the EU [MRTN-CT-2004005728 PATTERNS] and the DFG [SFB 486, project B13].

References

1. Beltrame, P., Thiele, U.: SIAM J. Applied Dynamical Systems, (2010) (at press)
2. Bestehorn, M., Pototsky, A., Thiele, U.: Eur. Phys. J. B **33**, 457–467 (2003)
3. Brinkmann, M., Lipowsky, R.: J. Appl. Phys. **92**, 4296–4306 (2002)
4. Burgess, J.M., Juel, A., McCormick, W.D., Swift, J.B., Swinney, H.L.: Phys. Rev. Lett. **86**, 1203–1206 (2001)
5. de Gennes, P.-G.: Rev. Mod. Phys. **57**, 827–863 (1985)
6. Deissler, R.J., Oron, A.: Phys. Rev. Lett. **68**, 2948–2951 (1992)
7. Doedel, E.J., Champneys, A.R., Fairgrieve, T.F., Kuznetsov, Y.A., Sandstede, B., Wang, X.J.: AUTO97: Continuation and bifurcation software for ordinary differential equations. Concordia University, Montreal (1997)
8. Eres, M.H., Schwartz, L.W., Roy, R.V.: Phys. Fluids **12**, 1278–1295 (2000)
9. Fermigier, M., Limat, L., Wesfreid, J.E., Boudinet, P., Quilliet, C.: J. Fluid Mech. **236**, 349–383 (1992)
10. Golestanian, R., Raphaël, E.: Europhys. Lett. **55**, 228–234 (2001)
11. Huppert, H.E.: Nature **300**, 427–429 (1982)
12. Israelachvili, J.N.: Intermolecular and Surface Forces. Academic, London (1992)
13. Kalliadasis, S., Thiele, U. (eds.) Thin Films of Soft Matter. Springer, Wien (2007)
14. Konnur, R., Kargupta, K., Sharma, A.: Phys. Rev. Lett. **84**, 931–934 (2000)
15. Leger, L., Joanny, J.F.: Rep. Prog. Phys. **55**, 431–486 (1992)
16. Lin, Z., Kerle, T., Baker, S.M., Hoagland, D.A., Schäffer, E., Steiner, U., Russell, T.P.: J. Chem. Phys. **114**, 2377–2381 (2001)
17. Marmur, A.: Colloid Surf. A-Physicochem. Eng. Asp. **116**, 55–61 (1996)
18. Merkt, D., Pototsky, A., Bestehorn, M., Thiele, U.: Phys. Fluids **17**, 064104 (2005)
19. Oron, A., Davis, S.H., Bankoff, S.G.: Rev. Mod. Phys. **69**, 931–980 (1997)
20. Oron, A., Rosenau, P.: J. Fluid Mech. **273**, 361–374 (1994)
21. Podgorski, T., Flesselles, J.-M., Limat, L.: Phys. Rev. Lett. **87**, 036102 (2001)
22. Quéré, D., Azzopardi, M.J., Delattre, L.: Langmuir **14**, 2213–2216 (1998)
23. Reiter, G.: Phys. Rev. Lett. **68**, 75–78 (1992)
24. Robbins, M.O., Joanny, J.F.: Europhys. Lett. **3**, 729–735 (1987)
25. Schäffer, E., Wong, P.Z.: Phys. Rev. Lett. **80**, 3069–3072 (1998)
26. Seemann, R., Herminghaus, S., Neto, C., Schlagowski, S., Podzimek, D., Konrad, R., Mantz, H., Jacobs, K.: J. Phys.-Condes. Matter **17**, S267–S290 (2005)
27. Sharma, A., Khanna, R.: Phys. Rev. Lett. **81**, 3463–3466 (1998)
28. Spaid, M.A., Homsy, G.M.: Phys. Fluids **8**, 460–478 (1996)
29. Thiele, U.: Eur. Phys. J. E **12**, 409–416 (2003)
30. Thiele, U., Brusch, L., Bestehorn, M., Bär, M.: Eur. Phys. J. E **11**, 255–271 (2003)
31. Thiele, U., Knobloch, E.: Phys. Fluids **15**, 892–907 (2003)
32. Thiele, U., Knobloch, E.: Phys. Rev. Lett. **97**, 204501 (2006)
33. Thiele, U., Knobloch, E.: New J. Phys. **8**(313), 1–37 (2006)
34. Thiele, U., Neuffer, K., Bestehorn, M., Pomeau, Y., Velarde, M.G.: Colloid Surf. A **206**, 87–104 (2002)
35. Thiele, U., Velarde, M.G., Neuffer, K.: Phys. Rev. Lett. **87**, 016104 (2001)

36. Thiele, U., Velarde, M.G., Neuffer, K., Bestehorn, M., Pomeau, Y.: Phys. Rev. E **64**, 061601 (2001)
37. VanHook, S.J., Schatz, M.F., Swift, J.B., McCormick, W.D., Swinney, H.L.: J. Fluid Mech. **345**, 45–78 (1997)
38. Verma, R., Sharma, A., Kargupta, K., Bhaumik, J.: Langmuir **21**, 3710–3721 (2005)

Minisymposium *ECMIMIM: Concepts of Mathematical Modelling in the Curriculum of Mathematics in Industry*

A. Noack

Technische Universität Dresden, 01062 Dresden, Germany
Antje.Noack@tu-dresden.de

Mathematics as a scientific field in conjunction with its large variety of applications turns out to be one of today's key skills. Products and processes of different kind are more and more developed and designed by means of mathematical modelling starting from the initial concept up to manufacturing and service. Hence, interaction with industry in the widest sense is a desirable aim for European studies in applied mathematics.

ECMI was founded with the aim to strengthen the collaboration between mathematical departments at European universities and the European industry and to promote the use of mathematical models in industry. The advantage is obvious for both sides: industry provides interesting real-life problems to train the student's skills in mathematical modelling and programming and on the other hand benefits from the expertise of the mathematicians in modern mathematical modelling methods and numerical solving techniques. As a feedback these collaborations generate demands for new methods, give impulse to new areas of research, and the industrial partners become aware what mathematics can do for them and, perhaps, start creating new tasks.

Interested in an appropriate education of students who in the future intend to work as mathematicians in industry, in the late 80s ECMI created a two-year postgraduate programme in Industrial Mathematics. The programme prescribes a number of core courses forming the basic knowledge every student of industrial mathematics should have and thus obligatory for them. Further it includes modelling activities and a study abroad at another ECMI university for at least one term. The modelling activities comprise a modelling seminar and the International ECMI Modelling Week. The remaining courses of the study can be chosen freely among a offered list of specialization courses. The programme finishes with a project placed in industry which is completed by a report written in English on the level of a master thesis. Graduates of this programme receive a certificate.

Up to now the implementation of the ECMI study programme varied considerably from ECMI node to ECMI node. Some universities implemented it

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*,
Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_100,

as part of their two years master study, some as additional programme during the diploma study, others after the masters degree, e.g. as initial part of the Ph.D. study. At present the ECMI educational system certifies too few students, i.e. too few accomplish the complete programme. Reasons are several: the extra work load, too hard common core requirements, not enough awareness of the certificate in the European world of science and industry, and therefore not enough advantage for the student.

The aim of the new EU project "ECMI Model Master in Industrial Mathematics", a common project of 10 ECMI partners under the leadership of the University Carlos III of Madrid, is to establish an innovative model of a European master programme in industrial mathematics. The model curriculum shall prescribe the general structure of the master programmes on the basis of the existing ECMI study programme. All partners of the project implement their own local master programmes during the next years in such a way that they fit the general model curriculum but let special strengths of each university flow in. Via dissemination activities the idea of this common model master will be spread all over Europe to encourage other universities, in particular other ECMI nodes, to join this programme and establish corresponding master programmes.

The partners expect that such a master programme will much more attract students than the ECMI study programme could do sofar since it does not imply additional work load and is finished by an official degree with some international attribute. The compatibility of the local programmes under the cover of the model curriculum allows a fluent exchange of students. All partners can benefit from common expert knowledge by interchange of lecturers and a common e-course concept.

One essential part of the master curriculum is mathematical modelling. Hence, an important point of discussion forms the question about the educational concepts that are favorite to equip the students with

- Skills in the development of mathematical models and in analyzing them.
- Knowledge of numerical methods.
- Training in advanced programming and simulations.
- Experience in tackling real-life problems coming from industry.
- Experience in team work, in the communication with engineers and the presentation of results for mathematicians and people from industry.

The aim of the minisymposium was to discuss such concepts of mathematical modelling. Topics of the minisymposium were the concept of modelling competence, the importance of teaching continuous modelling during the study, in particular via e-courses, and the cooperation of universities and industry by forming common study groups.

Why Teach Mathematical Modelling?

G. Brandell

Centre for Mathematical Sciences, Lund University, SE 221 00 Lund, Sweden,
Gerd.Brandell@math.lth.se

Summary. Research on the teaching and learning of mathematical modelling has attracted an increasing interest during the last 20 years. One concept of special significance in this field of research is the *modelling competence*, related to and intertwined with other mathematical sub-competences. In this paper some results from the research on the teaching and learning of mathematical modelling are presented and discussed and an example of an introductory course in mathematical modelling for engineers is presented.

1 Introduction

The overall goal of mathematics in primary, secondary and tertiary education is that the students become able to use mathematics in a variety of situations. Everybody needs to understand mathematical applications as a citizen as well as in the private life. For many, the use of mathematics is essential also in the working life. Phil Davis [7] reflected in 1991 on the use and limitations of mathematical descriptions at the ICTMA 4 (Fourth International Conference on the Teaching of Mathematical Modelling and Applications):

> Each age has preferred modes of prediction. Each mode opens up characteristic possibilities and creates realities. Mathematical modelling is today's high prestige way of predicting. It is the expression of our age, and it is likely to be around for a long while. However, we must watch it. We must watch it because mathematical descriptions tend to drive out all others. (p 1)

Today technology makes it easier to use advanced mathematics and therefore it is even more necessary to be able to exert a critical view on the results. Such a critical view may develop through an understanding of the entire process of applying mathematics to an extra-mathematical situation.

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*,
Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_101,

2 Modelling and the Modelling Competence

In any application of mathematics a mathematical model is involved. There are many definitions of a *mathematical model* in the literature. Here we follow Blomhøj and Jensen [2]. In short a mathematical model consists of the following: an extra-mathematical (real-world) domain, *the perceived reality*, a *mathematical system* describing some aspects of the perceived reality, and mathematical *model results* that may be applied to the real situation. See Fig. 1. During the modelling process the real system is delimited to a *domain of inquiry*, then reduced to a well defined *system* and translated into mathematical symbols and relations. Empirical data and mathematical theory give structure and content to the mathematical model. The results appear as solutions to the equations and relations in the mathematical system and are interpreted as insight into the reality or actions possible to carry out.

Mathematical ability is described in terms of competences by Niss [12]. The ability to carry out the modelling process is viewed as one specific mathematical competence among eight sub-competencies that together form what is called mathematical competence. The mathematical modelling competence is the ability to perform the processes that are involved in the construction and investigation of mathematical models [1]. According to Niss it includes among other things the following:

- Identify relevant questions, variables, relations and assumptions in a given real world situation.
- Simplify the real system and leave out factors of supposedly minor influence.
- Translate into mathematics.

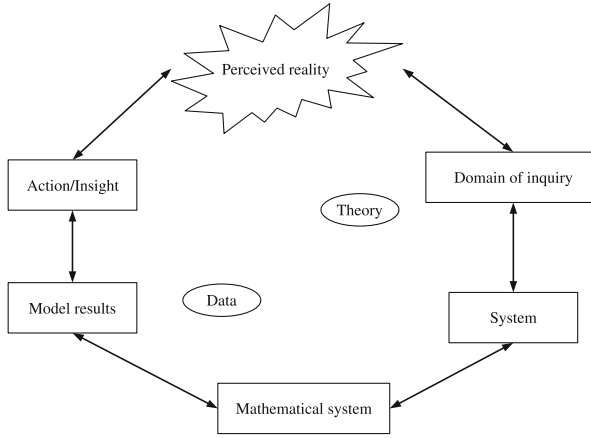


Fig. 1. The modelling process according to Blomhøj et al. [2]

- Work within the mathematical domain.
- Interpret and validate the solution of the model.
- Analyse and compare given mathematical models.
- Check properties and scope of a given model.

3 The Role and Place of Mathematical Modelling in Mathematics Education

Researchers in mathematics education have discussed reasons for modelling and applications in the mathematical curriculum at secondary and tertiary level and a number of aims have been put forward [5,10]. According to Blum [3] there are four types of arguments for including mathematical modelling in education. The *pragmatic* argument refers to the usefulness in extra-mathematical situations. A *formative* argument is that modelling develops general qualifications, such as translation between real-world and a description of reality. The *cultural* argument refers to knowledge of the role of mathematics in society. Finally the *psychological* argument is that modelling may be helpful for learning mathematics. However it is still open to discussion and research to which extent each of these arguments is valid in specific contexts.

Traditionally mathematical modelling is not included or made explicit in the mathematics school curriculum. Applications in other school subjects are polished and neat (physics, chemistry) and give no information of the modelling cycle. However, transfer of learning and knowledge to a new context is problematic and a classical educational issue both in practice and in research [9]. In general, transfer is only possible when the new context is quite similar to the old one. An important result from research in mathematics education seems to be that teaching and learning must address each specific mathematical competence that students should develop. If we wish students to reach some competence in mathematical modelling, then we have to teach mathematical modelling and make it explicit.

Different levels of understanding of the modelling process are required at various levels and for various groups of students. Students at higher levels ought to develop perspectives on mathematical modelling according to their future specialisation. Engineers use models constructed by others, modify existing mathematical models, and in some cases develop new mathematical models in cooperation with other specialists. Therefore – I claim – engineering students ought to develop a mathematical modelling competence. The same argument is relevant for students specialising in applied mathematics. The students need to experience that application of mathematics requires understanding and knowledge of the pertinent subject area, the relevant mathematics and the modelling process.

4 Sense-Making and Classroom Norms

One aspect of the modelling competence relevant already in primary education is the evaluation of a result in view of a real situation. A few examples of problems for which this competence is needed are the following:

- Carl has five friends and George has six friends. Carl and George decide to give a party together. They invite all their friends. All friends are present. How many friends are there at the party?
- Alice and Bruce go to the same school. Alice lives at a distance of 17 km from the school and Bruce at 8 km. How far do Alice and Bruce live from each other?
- John's best time to run 100 m is 17 s. How long will it take for him to run 1 km?

These problems have been used in several research studies, replicating an original study in the Netherlands and Northern Ireland. In the original investigation 90–95% of the pupils tested (upper elementary and lower secondary level) gave the answers 11 friends, 9 or 25 km and 170 s [8]. So why do students seem to avoid making sense in the mathematics classroom even if it would be possible for them to judge their result in view of a well-known reality?

The norms and values of a mathematics classroom influence the students' actions. According to Brousseau [6], a *didactical contract* shapes a student's beliefs and strategies in a teaching and learning situation. The teacher is responsible for making the learning meaningful for the student, and the student creates meaning for her- or himself by giving the right answers, following the instructions and so on. The implicit rules for mathematical problem solving implies that problems are not authentic, that any problem presented is solvable with one exact answer and that violations of your knowledge about everyday world may be ignored. An attitude is created that (school) mathematics forms a universe of its own – mathematics does not exist in a context where it is allowed or helpful to use common sense. Students bring this attitude with them to tertiary level. These attitudes and beliefs are probably most efficiently influenced and changed when students experience the full modelling cycle in an meaningful context.

5 Introductory Course in Mathematical Modelling

Traditionally the mathematical modelling process is not present in engineering programmes or enters at a late stage of the education, e.g. in a final examination project. However at some universities a different strategy is applied. Students get an early introduction into mathematical modelling, building on those mathematical tools at hand.

Courses in mathematical modelling may be designed in various ways. Two alternatives are to put more emphasis on mathematics through modelling

or on modelling through mathematics. It is easier to be explicit about the modelling process if the students work with well-known mathematics, as in the second alternative. However, also in the latter case, working with mathematical models has a potential to deepen the understanding of mathematical concepts already known to the students.

At two Swedish universities introductory courses in mathematical modelling for engineering students have been given according to the ideas presented above. At Luleå University of Technology a course has been given as part of the basic mathematics course of the engineering master programme since the 1980s. At Lund University a similar course is included in the first year of the engineering mathematics master programme since its start in 2002.

The purpose of the course in Lund is that the students change their attitudes towards the usefulness of mathematics and learn about the modelling process. More specifically the goals are to let the students

- Become aware of the meaning of mathematical modelling.
- Increase their self reliance when it comes to use mathematics in various contexts.
- Acquire modelling competence at an introductory level.
- Learn to handle computer support (Matlab, LaTeX).
- Acquire communication competence specifically about mathematics and modelling [11].

The course is given during seven weeks and corresponds to 4.5 ECTS-points. Students work with independent project work in co-operative groups of 3–4 persons. The projects are given by the teachers as open-ended problems in non-mathematical language and students have full responsibility for their model. All projects admit several valid models in reach of the students' capacity. Three projects are given with increasing complexity and several colleagues participate as supervisors. The students report on the projects in written reports and oral presentations. A peer review system is used to let the students learn from the experience of other students. However, the peer reviews are examined and supplemented by the supervisor. Only a few lectures are given. The students' evaluation is overwhelmingly positive. Generally the quality of the model improves through the sequence and the great majority of students raise the level of the presentation.

6 Issues in Research About Mathematical Modelling

Three phases may be discerned in the development of research about applications and modelling in mathematics education since the 1960s [13]. During the first phase (roughly 1965–1975) the main focus was on developing the arguments for including modelling in school curricula and educational programmes at universities. During the second phase (roughly 1975–1990) the development of courses in mathematical modelling and course materials were

brought into the focus. Much activity was reported from e.g. UK, US, Denmark and the Netherlands. During the last decades these strands have been further developed and complemented by empirical studies that give insight into the results of the teaching and learning of mathematical modelling. For those interested in the on-going research in this area there are a number of international conferences and publications. See a bibliography in [4].

7 Conclusion

In this article I have shown that mathematical modelling and applications has been a theme within research in mathematics education for decades, addressing among other issues the modelling competence and the various aims of modelling. However, it is still rare to include courses in mathematical modelling in programs for engineering students. In the article such courses are described that successfully meet the goal of teaching modelling as a specific competence and influence the students' beliefs about the usefulness of mathematics in applications.

References

1. Blomhøj, M., Jensen, T.H.: Teach. Math. Appl. **22**, 123–139 (2003)
2. Blomhøj, M., Jensen, T.H.: In: Blum, W., Galbraith, P.L., Henn, H.-W., Niss, M. (eds.) Modelling and Applications in Mathematics Education. The 14th ICMI Study,pp. 45–56. Springer, New York (2007)
3. Blum, W.: In: Niss, M., Blum, W., Huntley, I. (eds.) Teaching of Mathematical Modelling and applications, pp. 10–19. Ellis Horwood, New York (1991)
4. Blum, W., Galbraith, P.L., Henn, H.-W., Niss, M.: Modelling and Applications in Mathematics Education. The 14th ICMI Study. Springer, New York (2007)
5. Blum, W., Niss, M.: ESM **22**, 37–68 (1991)
6. Brousseau, G.: Theory of didactical situations in mathematics. Kluwer, Dordrecht (1997)
7. Davis, P.: In: Niss, M., Blum, W., Huntley, I. (eds.) Teaching of Mathematical Modelling and applications, pp. 1–9. Ellis Horwood, New York (1991)
8. Greer, B.: J Math. Beh. **12**, 239–250 (1993)
9. Haskell, R.E.: Transfer of Learning: Cognition, Instruction, and Reasoning. Academic, San Diego (2001)
10. Kaiser, G., Sriraman, B.: ZDM **38**, 302–310 (2006)
11. Mathematical Modelling. Syllabus academic year 2008/2009. Retrived Oct 13, 2008 from www.ka.lth.se/kursplaner/08_09%20eng/FMA045.html
12. Niss, M.: In Gagatsis, A., Papastavidris, S. 3rd Mediterranean Conference on Mathematical Education, pp. 115–124. Hellenic Mathematical Society, Athens (2003)
13. Niss, M., Blum, W., Galbraith, P.: In: Blum, W., Galbraith, P.L., Henn, H.-W., Niss, M. (eds.) Modelling and Applications in Mathematics Education. The 14th ICMI Study, pp. 3–32. Springer, New York (2007)

Differential Equations in the ECMIMIM Curriculum

P. Miidla

Institute of Mathematics, University of Tartu, J. Liivi 2, 50409 Tartu, Estonia
peep.miidla@ut.ee

Summary. A model is an analog of the object under consideration which replaces this object in human cognition. The main field of activity of the specialists of industrial mathematics is mathematical modelling and differential equations of all types are the main tools of continuous modelling. Because of that various courses connected with differential equations have an important place in the study programs on industrial mathematics, although there are differences in the proportions of those courses in the mandatory part of master programs among ECMIMIM partner universities. We also notice changes in historical development of the ECMI philosophy and execution of the curriculum of master on mathematics in industry from the establishment of ECMI in 1986. An additional conclusion to be taken into account for common online courses is that these need not be very large and expensive in the sense of ECTS assessment, compact teaching tools concentrated on a few fixed topics and which give 2–3 ECTS points to learner would be better.

1 Introduction

During the studies on industrial mathematics students have to learn to pose their own questions about the world, to understand the questions given by other people, particularly by the specialists of various human activities and to use mathematics to answer those questions. In other words, this means to develop the skills of students to understand real situations and then establish mathematical models, i.e. to use mathematical modelling to illustrate, explain and predict the behaviour of the object or phenomenon under consideration. Industrial innovation is increasingly based on the results and techniques of scientific research and that research, in turn, is both underpinned and driven by mathematics [7].

In this paper, at first, the general meaning of model is explained. The short overview of partner institutions master programs and development of ECMI viewpoint to industrial mathematics curriculum shows that the courses of differential equations have a significant place there. It is important that the individualities of partner universities are valuated, although there is enough

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*,
Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_102,

similarity to go on together in ECMIMIM community. From the modelling point of view some requirements arise for designing web-tool on differential equations.

2 Model and Modelling

Model by general definition is an analog, prototype of the object under consideration which replaces this object in human cognition. There are many types and classes of models in use in all human action fields, also several classification bases for these models. A model in science is a physical, mathematical, or logical representation of a system of entities, phenomena, or processes, so mathematical models are subclass of scientificals.

The replacement process of the object with its analog, the process of constructing models, is called modelling. In the case of mathematical modelling this replacement is done using mathematical means and tools, this building schedule is clearly explained in book [5]. Although there is no point to be too precise in defining the term "mathematical modelling", there is quite clear understanding of the types of these. Some models are explicative explaining a phenomenon in terms of simpler, more basic processes. All useful models, whatever type, are predictive in the sense that they allow us to make quantitative predictions that can be used either to test and refine the model, should that be necessary, or for use in practice [4].

The act of creating a model, mathematical included, forces the modeler to think deeply about the settings and conditions which must be fulfilled in the model. Translating an imprecise, complex, multivariate real-world situation into a simpler, more clearly defined mathematical structure such as functions, equations or a system of rules for a simulation can yield several properties or benefits of object [1]. Specialists of mathematics in industry must be able to do these kinds of "cuttings", they must be ready to discover that in many concrete cases, mathematical modelling and simulation have revealed unexpected behaviour of the relevant system and this presupposes the good enough knowledge of mathematics to where this translating is realized.

The main modelling stages which might be covered by courses on differential equations are: identification of the problem to be convinced that differential equations are applicable; formulation of the problem, including the choice of significant parameters; choice of type of differential equation depending on essence of phenomenon under investigation; choice of methods of research of the model in different stages – qualitative and quantitative, i.e. numerical methods; final elaboration of model, included to make clear the computer resource needed; application of model, simulations and numerical experiments; interpretation of results and solutions. For real differential equations modelling the first and last stages are quite complicated to teach as they contain enough heuristics, here presence of industrial specialists is

necessary. Success can be achieved by supporting the education with sufficient examples, case studies and interactive training possibilities. Other stages would be taught and trained in known ways.

The importance of simulation with differential equations models should be mentioned separately. Simulation is the implementation of a model over time, the act of imitating the behavior of some situation or some process by means of something suitably analogous. Results of numerical simulations can be tested with experimental data, and discrepancies resolved by improving the mathematical model [7].

3 Overview of the Master Programs

From the point of view of the university, mathematics has been an independent discipline for a long time. But new ideas develop either independently within the discipline or under the stimulus of applications. Thus, strengthening the ties between mathematics and industry will stimulate the development of mathematical sciences. Teaching of differential equations is already today often based on the modelling perspective. Workbooks are designed with accompanying software packages for solving and investigating differential equations and connected results [3]. Less and less we find pure theoretical courses, particularly on the undergraduate and master level. Applicational output is underlined as a very significant aspect of theory.

Ideas of industrial mathematics have begun to develop since 1986 when the ECMI was established. The exact history is not written yet, but we can say that during the academic year of 1986–1987, representatives of ten universities belonging to the European Consortium for Mathematics in Industry (ECMI) designed a two year postgraduate programme and reported the results in accounts dated February 20, 1987 and March 20, 1987. As intended, an educational programme which includes exchange of students, exchange of teachers, central international courses and cooperation with industry became operational. The experiences from the first few years of this ECMI-educational system have been discussed in detail by its partners and together with certain new insights they have led to an agreement on small changes in the philosophy and execution of the Programme. The resulting description, dated August 8, 1990, has been the guideline for the Programme for a period of about five years in which the educational system of ECMI was consolidated and gradually extended. After a long discussion, the final result has been approved in the meeting of the Council of ECMI on July 8, 1995.

In this curriculum the common core was fixed and this was obligatory to be taught at all educational centres. This mandatory part contained the following courses: analytical methods for ordinary differential equations; analytical methods for partial differential equations; numerical methods for ordinary differential equations; numerical methods for partial differential equations; nonlinear optimization; linear systems theory; regression analysis; discrete

optimization. The whole curriculum was divided into two parts: technomathematics and economathematics, the course list above is from the first branch. Notice that five courses out of eight are connected with differential equations. This shows the attention paid to those.

By the beginning of the 21th century the position of ECMI had changed. In 2002 there were only five topics fixed to be presented in the common core of master program for mathematics in industry: modelling seminar, scientific computing, optimization and statistics, ordinary differential equations, partial differential equations. The goals of education were clearly formulated: to produce trained students who can formulate a mathematical model from a description given by a non-mathematical industrial or specialist; carry out relevant mathematical analysis of an established model; select and implement an appropriate numerical method; use modern computation and communication tools; realize simulation and numerical experiments; interpret and improve results in consultation with the user; integrate all parts of the problem solving process; communicate on non-academic level; formulate the obtained results and reports.

These goals have been followed by ECMIMIM partner universities, but also in several other universities in Europe as well, where the master and even undergraduate programmes of mathematics in industry are working perfectly. At the same time academic institutions kept freedom to express their individualities and local strengths. From the point of view of continuous modelling we can find different obligatory courses in study programs which touch differential equations more or less directly. Some examples: differential equations, ordinary differential equations, partial differential equations, equations of mathematical physics, numerical methods in general and separately for ordinary and partial differential equations, modelling seminar, case study seminar, linear systems etc. In all of these the accent on the possibilities of using differential equations to construct models of real world is of great importance. The topics are taught in a way that the students understand the essence and nature of differential equation models and outputs of modelling and simulations.

4 Designing of a Web-Tool on Differential Equations

Computer, the most modern tool, is a very good teacher of differential equation models bridging numerical and continuous solutions and because of this it is reasonable to offer e-courses in the education of industrial mathematicians. Web makes it possible to also increase synergy and efficiency of the corresponding resources. For learners, online learning knows no time zones, and location and distance are not an issue. In asynchronous online learning, students can access online materials at any time, while synchronous online learning allows for real time interaction between students and the instructor. Learners can use the Internet to access up-to-date and relevant learning

materials, and can communicate with experts in the field in which they are studying. Situated learning is facilitated, since learners can complete online courses while working on the job or in their own space, and can contextualize the learning [2].

Differential equation courses are classical in theoretical sense and one can find a lot of materials online: lecture notes, exercises, assessment files etc. Open web-based study sources are useful for everybody. Though ECMIMIM partners need their own web-tools and online study sites due to the aim to reach, with ECMIMIM project a new model curriculum oriented to best practical solutions, which is intended to deliver an innovative set of European master programmes in industrial mathematics to be implemented through double degree agreements among the participants. Because of this it is important for partner universities to have common knowledge management.

The main parts of web-tools are official documents, texts or lecture notes with theory, overview with examples of methods and algorithms to solve exercises, bases of modelling with differential equations physical laws, demos and case examples and assessment materials [6]. To understand mathematics, its logic and beauty it is important for students to listen to live presentations, live lectures and for that purpose the web tool might also include video materials and lectures. For teaching mathematics the blended learning, where individual work with an e-course is accompanied by classwork, lectures, seminars and tutorial lessons seems useful. In general, adequate learning support for a student must be provided. Additionally it would be useful to supply web tools with short software manuals with orientation to applications using differential equations and with links to other resources and communication means as well.

Often e-teaching materials are not interactive, but are read-only. One significant feature which has been included in the web tool used in ECMIMIM community to teach industrial mathematicians is interactive modelling environment. It is important that student can "play" with their own problem, can change parameters and conditions and then visualize the result. This helps to promote student's creativity and demonstrate the link between theoretical concepts and applications. Interactive online learning can better create challenging activities that enable learners to link new information to old and acquire really meaningful and applicable knowledge. However, it is not the computer per se that makes students learn, but the design of real-life models and simulations, and the students' interaction with those models and simulations. The computer is merely the vehicle that provides the processing capability and delivers the instruction to learners [2].

The whole field of differential equations is very wide and it is hard to imagine that this would be placed into one only web-environment. There is also no need for such a colossal tool in the common curriculum. Even ordinary differential equations separately form too large area to be drawn together into one e-course. Another reason to be sceptical is that the one and only comprehensive tool can be too rigid to satisfy all the study programs of the partners as well as students who maybe want to study only some selected

topic of differential equations. Therefore the content of the whole field might be divided into reasonable parts. A good measure for this is the study credit system. As a rule, for example, the ordinary differential equations are included into curriculum as one classical course with up to 7.5 ECTS points. If we start to design and construct e-courses it is reasonable to think carefully about the volumes of components and instead of 7.5 credits e-course maybe it is better to prepare several e-courses with less credit points. For example, design separately the part of linear differential equations. The construction process of very huge web-tools may take too much time, instead of this we can complete a smaller course with more reasonable time and it might be finished and started. The final splitting of the field and also the programs of the courses on differential equations is a topic for consultations between partner universities.

The web-tool on differential equations must support high-quality studies of high levels of interactivity centred on the student, but also the involvement of new target groups. The online developer must know the different approaches to learning in order to select the most appropriate instructional strategies. Learning strategies should be selected to motivate learners, facilitate deep processing, build the whole person, cater for individual differences, promote meaningful learning, encourage interaction, provide feedback, facilitate contextual learning, and provide support during the learning process [2].

One appropriate methodological basement of the web-tool design is ADDIE, (Analyze, Design, Develop, Implement, Evaluate), which is in use in several universities. The course materials and tools are created according to the needs of the target group of students and essentials of the discipline, i.e. differential equations, ordinary or partial. Different design stages bring along different necessary activities and approaches [2].

References

1. Abrams, J.P.: Mathematical Modeling: Teaching the Open-ended Application of Mathematics (2001)
2. Anderson, T., Elloumi, F. (eds.): Theory and Practice of Online Learning. Printed at Athabasca University, 2004. http://cde.athabascau.ca/online_book
3. Borrelli, R.L., Coleman, C.S.: Differential Equations: A Modeling Perspective, 2nd edn. Wiley, New York (2004)
4. Howison, S.: Practical Applied Mathematics. Modelling, Analysis, Approximation. Cambridge University Press, Cambridge (2005)
5. Maki, D., Thompson M.: Mathematical Modeling and Computer Simulation. Thomson Brooks/Cole, Belmont (2006)
6. Miidla, P.: Web-tool on Differential Equations. In: Bonilla, L.L.; Moscoso, M.; Platero, G.; Vega, J.M. (eds.) Progress in Industrial Mathematics at ECMI 2006: 14th European Conference for Mathematics in Industry; Madrid, Spain; 10–14 July, 2006. Springer, Berlin (Mathematics in Industry) (2008)
7. OECD: Global Science Forum, Report on Mathematics in Industry. July 2008. `http://www.oecd.org/dataoecd/47/1/41019441.pdf`

Minisymposium *Topics in Learning Applied and Industrial Mathematics*

A. Kværnø[1] and H.G. ter Morsche[2]

[1] Department of Mathematical Sciences, Norwegian University of Science and Technology, 7491 Trondheim, Norway Anne.Kvarno@math.ntnu.no
[2] Technische Universiteit Eindhoven, University of Technology, Department of Mathematics and Computer Science, Den Dolech 2, 5600 MB Eindhoven, The Netherlands H.G.terMorsche@tue.nl

During their meeting in February 2008, the ECMI Educational Committee decided to propose minisymposia focusing on educational aspects for ECMI 2008. The main objective for the ECMI Educational Committee is to promote education of industrial mathematicians by collaboration on the development of course curriculum in mathematical disciplines relevant for applications in industry and commerce, and by exchange of students and teachers between the ECMI partner universities. Most of the committee's activities are directed to students at the master level. However, the following citation from the PISA 2006 survey, emphasize the importance of motivating students at the high school and undergraduate level as well:

"In today's technology-based societies, understanding fundamental scientific concepts and theories and the ability to structure and solve scientific problems are more important than ever. Yet the percentage of students in some OECD countries who are studying science and technology in universities has dropped markedly over the past 15 years. The reasons for this are varied, but some research suggests that student attitudes towards science, may play an important role (OECD, 2006a)."

The minisymposium aims to cover different aspects of teaching of applied and industrial mathematics not only at a bachelor or master level, but also at a high school level. The common divisor of all these aspects is the art of mathematical modeling. The ultimate goal is to train students to apply and develop mathematical methods and have the computational skills to solve industrial and engineering problems.

The presentation of Martin Bracke, (Technische Universität Kaiserslautern) covers how mathematical modeling activities have been introduced for high school students, as well as for students in teachers education. K. Schmidt (Technical University of Denmark) presents how Maple has been introduced in the introductory math courses at DTU in order to e.g. be able to introduce students to realistic problems at an early stage. He also discusses how technology

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*,
Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_103,

changes the student's use of textbooks and other knowledge resources in different study activities. Usually, to learn mathematical modeling is a process of doing by trial and error. By means of a wide range of examples one tries to build up competencies in the design of mathematical models. In his presentation, K. van Overveld (Technische Universiteit Eindhoven) shows that an overall systematic approach will give better insight in the modeling process and could lead to better results.

Modelling Reality: Motivate Your Students!

M. Bracke

Department of Mathematics, University of Kaiserslautern, P.O. Box 3049, 67653 Kaiserslautern, Germany, bracke@mathematik.uni-kl.de

Summary. Many universities have established modelling activities like special lectures on modelling, seminars or project work for math students; the ECMI Modelling Week is a nice example on a European level. In the past years there is a strong (and growing) interest in the integration of modelling activities at high school level: From the results of the PISA studies we know that students as well as their teachers need to enhance their modelling literacy.

In this paper we first motivate our strong focus on real world problems when conducting modelling activities and briefly summarise the general framework of our projects. Then we introduce three examples of modelling tasks we have presented to different groups of students at high school level as well as at university level. Finally, there is a short conclusion of our experiences made during the last 20 years.

1 Introduction – Motivation: The Hidden Component in Mathematical Modelling

Many modelling projects at university as well as at high school level have been conducted by the Department of Mathematics of the University of Kaiserslautern during the last 20 years. All of them have shown a big benefit for the participating students: they recognise that they can use the mathematical tools which are taught in school (or university, respectively) to understand and solve real world problems – and this insight gives a lot of motivation to learn mathematics! The logical consequence is to look for ways to incorporate mathematical modelling of real world problems (and there is a big emphasis on **real world**, for several reasons!) into standard mathematical education in high schools and universities.

Of course the concept of mathematical modelling is by no means a new one! Nor is it the idea of letting students deal with problems where they have to use at least some techniques of modelling in order to obtain a mathematical question to be solved. At the latest from publication of the results of the PISA studies [2] various activities to introduce mathematical modelling into

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*, Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_104,

standard math education have been started. There are many good ideas but from our experience quite a lot of them miss a very important point: the problem in focus has to be realistic and the modelling process should be complete![1] In order to save time many teachers simplify the problem. This can be done in several ways, among them:

- All information which is needed to solve the problem – and only that information! – is provided from the beginning.
- Some steps of the modelling process have been done before handing the problem to the students; as a consequence the students already start with a mathematical problem or obtain such a mathematical formulation of the original question from the pre-modelled parts without having a choice.
- Data is chosen in a way which simplifies the computations – with a possible loss of realism.

In certain situations there are reasons for introducing some simplifications either way, but from our point of view the experience of a complete modelling process without simplifying the original problem is worth all the effort and time. Especially those students who have a poor interest in mathematics as they know it from school gain a lot of motivation from such an experience. Very often it shows the interdisciplinary character of mathematical modelling and the usefulness of mathematical tools which are taught in school but rarely applied in a real context. From the feedback we have got from many participants of modelling projects in schools we learned that especially the 'soft features' like *correspondence to reality*, *applicability of mathematics*, *non-uniqueness of solutions* or *allowance to follow wrong ideas which are to be improved later* make the difference.

Hence our conclusion is to offer modelling projects which show a high degree of realism. Most of these projects are done in a compact form (1.5 up to 2.5 full days). They start with the presentation of the projects, then the participants choose a project they like to work on and form teams with 4–5 students; at the end of the whole project all teams present their results to the others and the important point is that they have to find explanations which can be understood by the problem poser – hence these presentations usually do not contain a lot of deep mathematics! In between the students try to really understand the original problem, obtain missing information/data, set up different mathematical models and try to solve them using the tools they have learnt before. Almost all projects have in common that a computer is necessary to solve the mathematical problems the students are facing. In an ideal setting there is a supervisor for each of the modelling groups who simulates the behaviour of the problem poser (since usually we do not have those people with us all the time). The supervisor is supposed to answer the

[1] These observations have been made in German schools and universities – and even here we do not claim them to be representative – but might be similar in other countries.

questions of the students in the way the problem poser (i.e. an engineer, a biologist, someone from administration) would do. In many situations the answer is again a question to the students – hence it is really them who determine the model and the solution process. In [1] we explain our framework for doing modelling projects with students in detail.

2 Examples: Three Real World Problems of Different Type

In this section we introduce three different problems which all have a real world character. The idea of this section is not to discuss various models and solutions of these problems (there is by far not enough space to do this even for one of the projects) but to present them as we usually do for the students: using the language of the problem poser, i.e. a darts player, an alpinist or a biologist, we describe a problem to be solved. There is no mathematics in these descriptions, some terms and notions might be new for the reader and a lot of information is missing for sure. Let's start!

2.1 How to Play an Optimal Darts Game?

In some parts of the world – and many European countries can be considered to belong to them – the game of *darts* is known and many people have even played the game. In Fig. 1 a dartboard is shown. The line behind which the throwing player must stand is generally 2.37 m from the face of the dartboard measured horizontally; the centre of the board should be at a distance of 1.73 m from the floor. The dartboard is divided into 20 numbered sectors scoring from 1 to 20 points. Moreover, there are several rings with a different meaning for the scores: The outermost ring (*double* ring) doubles the score of the corresponding sector, hitting the next ring (*triple* ring) gives you three times the number of points of the corresponding sector and the centre (*bull*) which is again divided into two areas gives 25 and 50 (for the innermost circle) points.

The rules of the game – in the standard variant – are quite simple (cf. *wikipedia*): The sport of darts is usually contested between two players who take turns in throwing up to three darts. Starting from a set score, usually 501 or 301, a player wins by reducing his score to zero. The last dart in the leg must hit either a double or the inner portion of the bullseye, which is the double of the outer bull, and must reduce the score to exactly 0.

Now the question sounds simple but nevertheless it is hard to grasp for some people: How to play an optimal darts game, i.e. where should a player aim in order to win the game? For sure this depends on his or her abilities to hit certain points on the board, the corresponding abilities of the opponent and the actual score... but how does this dependence look like?

Fig. 1. Darts in a dart board (*left*) and different sectors of a dartboard (*right*)

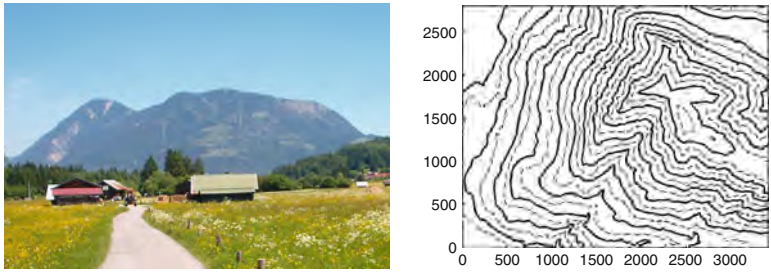


Fig. 2. *Wank* (1780m absolute altitude) near Garmisch-Partenkirchen, Germany

Most professional players aim at the triple 20, since three times 60 points would give the optimal result of 180 points per leg. But clearly, most amateur players are not able even to come close to 180! And this rises the question if the triple 20 is the optimal point to aim at – why not triple 19 or triple 14 or even the bull's eye (the central circle)?

2.2 The Optimal Way to the Top of a Hill

In Fig. 2 you can see a photo of mountain *Wank* near Garmisch-Partenkirchen together with an altitude profile (digital data based on a 1 m × 1 m grid is available). The simple question is: What is the best way in the sense of energy consumption to reach the top of the Wank?

Besides the high resolution contour data of the mountain there is some additional information regarding the relation between energy consumption when climbing a mountain and inclination of the path which was published by Italian physiologists (see Fig. 3) – but that's all the information given to the students!

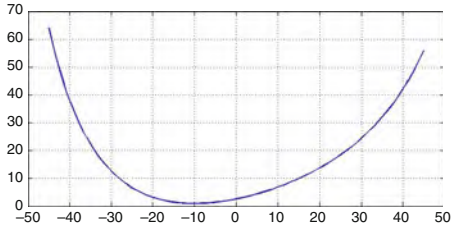


Fig. 3. Possible relation between inclination of a path (horiz. axis, angle in degree) and energy consumption of an alpinist (vert. axis, kJoule per meter height difference)

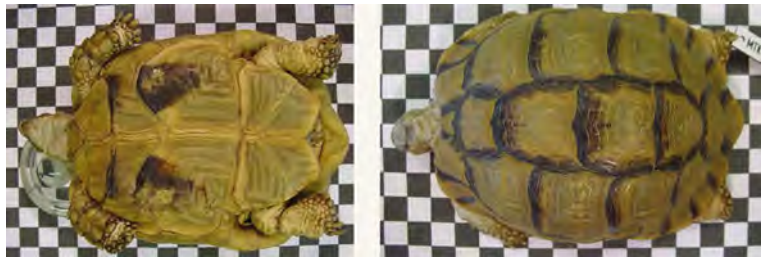


Fig. 4. Plastron & carapax of testudo kleinmanni

2.3 Design of an Identity Card for Turtles

How to identify a turtle using non-invasive methods (i.e. no transponder or other electronic devices being used)?[2] This question is motivated by CITES which is accepted by 173 countries (by October 2008). The aim of CITES is to ensure that international trade in specimens of wild animals and plants does not threaten their survival and hence it is necessary to be able to check the identity of animals! Fig. 4 shows photos of the plastron and carapax of an individual of testudo kleinmanni.

At first glance, there seem to be enough features in the images which allow for an identification of an individual. On the other hand, some animals clearly differ a lot while there are also some which are quite similar... But if you want to distinguish between 1,000,000 animals you clearly have to identify some appropriate features and determination of quantitative values has to be somewhat robust against various kinds of perturbations. Different models and solutions of student teams are discussed in [1].

2.4 It is Your Choice!

Having presented three different problems it is now your choice: which one would you like to think about or even to work on with your students? For the

[2] Original research project by biologist Dr. Carolin Bender (Kaiserslautern).

students it is very important to be able to choose a project. At the beginning of our activities we built the modelling teams based on various information about the abilities and interests of the participants. But over the years we have learnt that 'having the choice' has quite a big influence on the motivation of the students to work on a project for a longer period of time.

For all three problems various approaches exist – which will of course influence the solution and the final answer to our original question. But the very nice thing about modelling is that *the* solution does not exist (in the sense that there is only one model resulting in a certain answer)! The time you can spend for the whole process as well as the mathematical tools are going to have a very strong influence on the results.

3 Conclusions and Outlook

From our some twenty years experience of modelling with students, we can formulate two main consequences:

- The most important aspect concerning organization and implementation of modelling projects is learning by doing. Nobody would expect to become a good driver or even a pilot just by reading some books at this, quality is more or less proportional to the amount of practice (at least at the beginning...).
- Mathematical modelling should be integrated into teacher training including the learning by doing component, training of the supervisor role and learning how to find problems to the same extent. To achieve this one idea is to include student teachers in organization and implementation of modelling events in schools; we started to test this concept at TU Kaiserslautern some time ago and results are quite promising. A similar approach can be followed in advanced teacher trainings.

It is not our claim that the above consequences are completely new or particularly original but we have met many people thinking that having a collection of modelling problems at hand together with a good book on mathematical modelling is all they need to successfully do modelling projects with their students – and this is definitely not the case!

To conclude this paper we would like to appeal to all teachers interested in mathematical modelling of real world projects: just start modelling with your students – it is a lot of fun and it is worth one's while!

References

1. Bracke, M.: *Turtles in the Classroom – Mathematical Modelling in Modern High School Education*, Proceedings of the ICTMA-13, Bloomington (USA), (2007)
2. OECD: *First Results from PISA 2000/2003/2006*, (OECD Paris, 2001/2004/2007)

The Impact of CAS Use in Introductory Engineering Mathematics

K. Schmidt[1], P. Rattleff[2], and P.M. Hussmann[3]

[1] Dept. of Math., DTU, 2800 Lyngby, DK, karsten.schmidt@mat.dtu.dk
[2] Danish Univ. of Educ., 2400 Copenhagen NV, DK, rattleff@dpu.dk
[3] LearningLab, DTU, 2800 Lyngby, DK, pmh@llab.dtu.dk

Summary. In this paper we describe and discuss the way the usual challenges of CAS use has been met with in a first year engineering mathematics course which is based on a thorough and experienced implementation of the advanced CAS program Maple. The intensive use of Maple seems to cause a decrease in the students' preparation for the lessons and in their use of the classic textbooks. Not all of the students seem to fully utilize the experimental benefits of the software program and some of the students do not avoid pitfalls like doing exercises just by changing the variables in the provided Maple examples. In our conclusion we suggest an upgrading of the Maple Demos in a way that strengthens the interactive and conceptual potentials at the expense of the repetitive ones.

1 Introduction

Since 2001 the CAS tool, Maple, has been a fully integrated part of the first year introductory mathematics course at The Technical University of Denmark (DTU). Maple supports both symbolic and numeric computations, as well as visualizations. In a recent paper the authors of this paper has presented selected results from a broad survey study of the mathematics study habits of the students at DTU, see [1]. In the present paper we will focus on how the students experience the extensive use of Maple, and how Maple influences their working methods and use of the teaching resources. It is of special interest to describe how the attitudes and behavior of the students change in the course of the first study year, and to see if the hopes and expectations that normally are linked to the use of CAS are reflected in the students' evaluation of the study program. To embed our results in the right context our paper starts with a discussion on how DTU hitherto has met the usual challenges of implementing CAS use in a university study program.

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*,
Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_105,

2 Potentials and Pitfalls in the Use of CAS

A study in an early experiment in the use of Maple in introductory university mathematics [3] uses the following two archetypical statements from two classical papers on the subject to set up the potential benefits and dangers. In the first one the computer acts as a helper and in the second one as an obstacle on the road towards understanding: (1)"*The idea is for students to operate on a high conceptual level; in other words, they can concentrate on the operations that are intended to be the focus of the attention and leave the lower level operations to the computer.*" (2)"*Computers present particular problems to those who favour more work with deduction. Because of their ability to display example after example, computers encourage induction as a valid method of argument.*" (p. 4 in [3]).

Of course these positions do not necessarily have to be strictly opposite, a modern CAS supported education have to try to utilize the potentials and at the same time to find methods to avoid the pitfalls.

In a following paper one of the authors of [3] has divided the known pro et cons more precisely from a semiotic point of view [5]. The two above mentioned positions are here named *the lever potential* and *the particularity problem.* Two further dangers treated in the paper are of a special interest in our context: *The black box effect* refers to the risk that the student is satisfied when the CAS tool jumps from the problem to the solution without him or her controlling, or at least trying to understand, the intermediate steps. And *conflicting intentions* indicate that the contentment of the students might be motivated by primarily pragmatic reasons, e.g. minimizing the time spent while still passing the exam.

3 Trying to Utilize the Potentials

When the course Mathematics 1 (20 ECTS points, 600 students) was provided for the first time in the year 2000, the implementation of Maple was done through out economically, logistically and pedagogically, so that the software program from the beginning became an important part of teaching, as well as a tool for demonstrations at the lectures and as a key device offered for group exercises and for a mandatory (and credit giving) big project exercise which typically demands comprehensive calculations and visualisations. The philosophy (and hypotheses) at DTU is that a CAS tool with support for analytical as well as numerical computations ideally further the abilities and enjoyments of the future engineers in modelling the technical and scientific problems by using mathematics:

95% of every engineering model contains – and must be naturally based upon – kernels of precisely formulated physical, chemical, and constitutive laws and assumptions. The mathematical crux of these laws and assumptions can most effectively be fully understood, analyzed, modified, developed and unfolded

by applying analytical CAS tools like Maple to sufficiently small 'toy' versions of the model in question. In parallel the corresponding numerical tool (like Maple or Matlab) must be applied to solve and simulate solutions to the full scale real problems. None of the two types of attack can substitute the other. (p. 4 in [4]).

We will refer to this as *the toy model potential* and present an example of the DTU attempts to utilize the CAS potentials which seems to display the toy model potential on several levels: In a typical mathematics exercise in vector analysis there is given a vector field and a surface and with given formulas the students are able to calculate the flux of the field through the surface. It is an open question if this necessarily leads to a deeper understanding of the subject or to the ability later on to transfer the related methods to engineering tasks. By the support of a CAS tool the exercise can be lifted to a higher conceptual level (*the lever potential*): In a 2007 DTU homework exercise there are given two sunroofs, a hemisphere (cut from a sphere along the equator) and spherical cap (cut from a sphere along a small circle), respectively. The students are asked to calculate the total energy absorption in one day for each of the two sunroofs (p. 29 in [2]). The students are guided to model the light rays from the sun in the form of a parallel vector field and the energy absorption as a flux of the vector field through the surfaces. But then they have to think through the concept of flux, since it only seems relevant to calculate the flux through the time dependent illuminated parts of the roofs. By 2D and 3D visualizations they have to figure out and parametrize the related limiting geometric models. Following a students paper it was possible in the case of the hemisphere (placed on the equator) to express the absorption to a given time in a simple analytical way $E(t) = \pi \cdot (\sin(t) + 1)/2$ and thence also to calculate an exact result for the whole day absorption: $\pi + \pi^2/2$. In the more difficult case of the spherical cap it was still possible for the clever student to obtain a symbolic Maple output for the time dependent absorption. But this formula, including heavy square roots and several composed trigonometric and arcus functions, was not easy to cope with and certainly not of the type of "a beautiful answer" as in a classical standard exercise. Furthermore the student had to give up his attempt to force Maple to calculate an exact result for the whole day absorption, and therefore he exactly at this point experienced the limit of *the toy model* and turned over to numerical Maple methods, satisfied with displaying a decimal number as his final result.

4 Trying to Avoid the Pitfalls

At DTU the known dangers and problems are on the one hand been dealt with by current upgrading of that part of the teaching and curriculum, which makes visible the advantages of Maple: The number of experimental exercises have been increased, thematic exercises have been introduced, and new project based exercises with reference to recent research in diverse applications have

been developed. On the other hand the students and teachers at Mathematics 1 are carefully instructed in two rules of Maple use: (1) It is essential to teach the students to choose those Maple commands and Maple styles that support the present learning objective the best, and (2) Any Maple output must be provided with relevant and sufficient explanations and interpretations. One example: A Maple Demo concerning the subject of systems of linear equations exposes three different Maple commands: *LinearSolve, RowReducedEchelonForm* and *RowOperation*. The first one immediately presents the solution, without revealing anything about the used method (with the risk of *the black box effect*), the second one displays the result of a completed Gauss–Jordan elimination from where the solution has to be extracted, whereas the third one requires that the user himself performs the eliminations step by step only leaving elementary arithmetic to the computer, a method that we call "simulated paper and pencil calculations". When systems of linear equations is the subject being introduced, the student is expected to document through his Maple report that he fully has understood the basic concepts and methods in this field, while in other (later) cases, when this subject is subordinate, his finding of the most direct way to obtain the solution can be even laudable.

5 The Survey Study 2007–2008 at DTU

Our survey study in the academic year of 2007–2008 investigated how the Mathematics 1 students used the different study resources in relation to the different types of study activities. By *study resources* is understood primarily the three classic DTU textbooks[1], the Maple Demos[2], the Internet and diverse materials from the course homepage. By *study activities* is here meant one ordinary week of studies (that is: not working with project based exercises) consisting of two times a lecture followed by classroom teaching/group exercises supported by TAs and also non-scheduled activities: weekly quizzes and mandatory homework exercises (ten sets during the academic year).

The students were asked to fill out an online questionnaire three times during the academic year: in the weeks 3, 10 and 19 of the two semesters (a total of 26 weeks).

As initially the acts of reading the textbooks and listening to the teachers seemed fundamental in the way the students understood learning, a definite change seemed to happen during the academic year. While the part of the students who report that they have attended both group exercises is constant (89, 89 and 89%) during the three stages of the survey, and the amount of time spent on homework exercises is a bit increasing (305, 360, and 350 min.),

[1] Concerning Linear Algebra, One Variable Analysis and Multivariable Analysis.

[2] Exemplary Maple worksheets which by use of short explanations and examples provides an alternative introduction to the subject of the day, which the students can use in their own further work.

then the part of the students attending both lectures is a bit decreasing (94, 87, and 85%), but the part of the students who have been prepared for the classes is most definitely decreased (65, 48, and 30%).

The decrease in preparations seems closely connected with a change in the way the students experience the value of the textbooks for the learning process. A definite decrease among those who is reading contiguous text in the textbooks can be observed (23, 12, and 6%), while there is a lesser decrease among those who are skimming the subjects of the week in the textbooks (43, 44, and 38%).These numbers display the behavior during the group exercises, but they are typical for all the types of study activities.

While the use of the textbooks is decreasing during the academic year, the students use of Maple and especially their use of the Maple Demos is increasing. When asked to rank the study resources, the part of the students who gave the highest ranking to the textbooks was decreasing (33, 38, and 17%), while the ranking of the Maple Demos were increasing (7, 15, and 40%). These changes are also reflected in the students working methods. For instance is the part of the students who have used Maple shorter or longer time during the group exercises increasing (89, 90, and 95%), while the part who have used paper and pencil is decreasing (77, 75, and 51%). Correspondingly there is a decrease in that part of the students who respond that they have worked "with understanding concepts and proofs" during the group exercises (55, 42, and 30%), which might indicate *the particularity problem.*

A didactic evaluation of these changes of course depends on how Maple and the Maple Demos is actually used. The following are some typical answers which document that the attitudes of the students are very diverging:

Using the maple-demos is more fruitful as it leaves room for experiments through which you learn to understand how different components may affect a result.

I prefer the Maple Demo as it enables you to pick up promptly the tools to apply when doing math, thus preparing you to do your homework exercises and pass the course. Understanding the math itself comes second and, what is more, the books are very hard to get a grasp of.

I miss learning math. I think that the main focus is on how to use Maple. The understanding is lacking.

The first statement indicates a use of Maple according fully to the intended potentials. The next one is more ambiguous, although favourably disposed towards Maple it displays disquieting elements of *conflicting intentions.* The last one is definitely negative towards Maple, because *the black box effect* seems to block the understanding.

The students' actual use of the Maple Demos in relation to the three most important types of activity is shown in this diagram:

The diagram shows that a large part of the student *read* the Maple Demos in order to acquire a better understanding of the subject, while a lesser part use the experimental interactive potentials of the Maple Demos. The most typical use of the Demos is copy and pasting the Maple commands to their

Table 1. The use of Maple Demos

	Prep.(%)	Group ex.(%)	Home ex.(%)
To read through to understand	11	59	57
To play and experiment with	5	38	30
To copy commands from	9	78	72
To do exercises by changing variables	6	49	39

own worksheets. Finally a special variant of *the particularity problem* seems to be a disquieting important factor as a considerable part of the students report that they obtain some of their solutions just by changing some variables in the examples included in the Maple Demos.

6 Conclusions and Recommendations

The above mentioned survey study of the study habits of engineering students during the introductory DTU mathematics course shows the following facts about their attitudes to the use of the CAS tool Maple and their actual use of it:

1. The students do certainly not make up a homogeneous group. Maple supports that a mathematics course can be planned broadly with various co-ordinated teaching offers so that it is possible for the individual student to find his own learning styles.
2. There is a positive tendency towards a decrease in the importance of the classic textbooks during the course, while they seem to be in some respect compensated by the corpus of the Maple Demos.
3. Many students are utilizing the potentials, but some students do not avoid the pitfalls. We propose an upgrading of the corpus of Maple Demos in a way that strengthens the interactive and conceptual potentials on behalf of the repetitive ones.
4. A genuine evaluation of the impact of Maple on the changes in the mathematics learning processes would presuppose a throughout didactic study of the teaching materials and a parallel study of the actual benefits obtained by the students.
5. The changes of study habits during the mathematics course might be summarized in one single quote made by an unknown student. Referred in its entirety: “reading less, learning more.”

References

1. Rattleff P., Schmidt K., Hussmann P.M.: "Læser mindre og forstår mere" (Dansih). MONA 2009–3
2. www2.mat.dtu.dk/education/01005/MWS/INTEGRATOR6/IntegrationIflereVariableTekst2008.pdf
3. Solovej J., Winsløw C.: Maple på første års matematik (Danish). Report no. 14, Cent. for Educ. Develpm. in Univ. Sci., Aalbog.
4. Markvorsen S.: Math Education at a Crossroads. A report from a project at LearningLab DTU, Mat-Report No. 2006-19, (2006)
5. Winsløw C.: Semiotic and discursive variables in CAS-based didactical engineering. Educ. Stud. in Math. **52**, 271–288 (2003)

Minisymposium *Web Based Courses: Reaching a Distributed Audience*

Matti Heiliö[1] and Helle Rootzén[2]

[1] Lappeenranta University of Technology, Finland, matti.heilio@lut.fi
[2] Technical University of Denmark, Denmark, hero@imm.dtu.dk

In this minisymposium we discussed the challenge of web based solutions in organizing education in modelling and applied mathematics. The cutting edge knowledge in the art of mathematical technology is located in small nodes, research groups on applied mathematics, mathematical physics, scientific computing etc. Web-technologies are a viable media for innovative processes and knowledge transfer. Virtual educational environments enable novel solutions to training and education, they help to facilitate distributed processes and provide access to educational resources. Interactive cross-media allows easy time- and location-independent access and portability, flexible updates and the benefits of media technology, hypertext properties, animations etc. An evolution of educational methods, materials and means of delivery is taking place.

In this article we also draw attention to the possibility of creating added value for the knowledge repository of applied mathematics and computational methods via remote access educational modules. We suggest that we make a web portal for this. It will pool together and demonstrate the special knowledge and expertise in industrial math available in the network and create added value for European Masters Education in Industrial Mathematics.

This minisymposium wanted to give ideas on how to build up net based courses and environments. We discussed the challenge of distributed web based solutions in organizing education in applied mathematics and statistics. New ways for activating the students using the computer was presented and we discussed international and third world perspective on distance education.

Many pathfinder projects are underway. Time is mature for launching international collaboration. The minisymposium talks represented examples of ongoing e-learning projects which when brought together under a common portal could provide an important learning environment.

The first talk presented a versatile learning environment for design of experiments where the interactive system is intended to give the learner a feeling of the decisions in real industrial R&D situation. The second presentation described a net based master programme in applied statistics emphasizing the

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*,
Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_106,

modular approach of learning objects and special attention to the possibility of distant learners from business and working life. The third talk presented the benefits of providing access to demanding special courses in less favoured regions where lack of specialist academic skills is a severe handicap. Finally the possibility to use national academic potential more efficiently by pooling courses and teachers via network connecting several campuses was suggested in the last talk.

This article wants also to bring forward a vision of a European e-learning portal in applied mathematics. Such e-learning environment would be suitable for students in applied mathematics and engineering programmes in advanced BS and MS level. It would be designed also for persons who are already in their working life and are looking for continuing education and professional development.

For university use the courses would be especially suitable for those offering a specialization in Industrial Mathematics, Technomathematics etc. The courses would also be intended for continuing education of people working in industrial R&D. Some of the courses might represent more standard mathematical methods with high demand. The courses would be based on customised content for a special applications area that is active in the current European research and technology arena. The asset feature being the design and the usability for a certain target group of users.

The courses would create a learning environment for mathematical modelling, optimization and data analysis, business statistics and operations analysis. A natural base for such e-learning portal would be ECMI which represents a network of European universities and collaboration with industry in mathematical technology transfer, has a mission in European knowledge sharing and an educational programme in industrial mathematics.

The pooled expertise of the consortium provides a good foundation for a versatile, high quality and up-to-date content production representing forefront-knowledge in Europe. Some of the course topics would be based on current research and knowledge having relevance and demand in the industrial R&D and in educating industrial mathematicians or research scientists. The value of the product for the R&D community is the unique material, which is not yet available in abundance in commercial media.

The goal should be to create courses so that they can be used by students in several European countries and universities. A course addressed for a multi-campus audience is a viable possibility where the benefits of sharing expert knowledge are obvious. The e-course menu means pooling and sharing of expertise and it will strengthen the curriculum co-development in European Higher education.

Statlab: An Interactive Teaching Tool for DOE

M.A.A. Boon, A. Di Bucchianico, J.J.M. Rijpkema, and E.E.M. van Berkum

Eindhoven University of Technology, Department of Mathematics, P.O. Box 513, 5600 MB Eindhoven, The Netherlands, M.A.A.Boon@tue.nl, A.d.Bucchianico@tue.nl, J.J.M.Rijpkema@tue.nl, E.E.M.v.Berkum@tue.nl

Summary. An interactive web based teaching tool, Statlab, for Design of Experiments is presented. In this tool, the student is introduced to practical strategies for experimenting through virtual case studies. Statlab forces students to think about practical details since it hides options that students do not ask for. Engineering students as well as industrial participants in our courses consider Statlab as a stimulating learning environment. Statlab can be freely used through the web site www.win.tue.nl/statlab/.

1 Introduction

A good working knowledge of DOE (Design of Experiments) is essential for both industrial statisticians and engineers. It is therefore essential that statistics courses pay sufficient attention to this topic. However, a distinctive feature of DOE is that it is pro-active, unlike many other statistical techniques that are focused on extracting useful information after data has been collected. Hence, this requires a teaching approach that forces students to actively think about several aspects of setting up an experimental design, without steering the student too much. An additional feature required by us is that there should be room for the student to make mistakes and learning from them. In order to create such a teaching environment to be used in statistics courses at various departments of the Eindhoven University of Technology, a web based tool called Statlab has been developed. This tool adapts itself to the student, who is being led through one of several possible case studies. In this paper we describe the teaching philosophy behind this tool, as well as its technical implementation. The tool has been receiving positive reactions from students, who generally consider using Statlab as a stimulating teaching environment.

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*, Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_107,

2 Philosophy of Teaching DOE

Experimental design is a pro-active activity, unlike many other statistical methods. We feel that teaching experimental design should include conferring feeling for required choices in designing experiments. This should include learning from the consequences of omissions. Standard statistical packages are not suitable for these teaching tools, because they present many options to the student, who either accepts the default settings or simply chooses options that the software offers. We require a teaching environment that forces students to actively think about the construction of an experimental design, as well as analyzing data collected from such a design. In order to stay close to practice, the students should go through the following phases (see [4] for lots of useful advice and [1] for an example of an implementation in a physical experiment):

- Gather information about the goal of the experiments.
- Gather information about the experimental facilities.
- Construct an appropriate design.
- Analyse data collected from the chosen design.
- Formulate conclusions and recommendations.

Since statistical software often has an interface to a catalogue of experimental designs, we require that students are able to intelligently choose a design rather than have to construct designs themselves like fractions of factorial designs. We thus have the following requirements for our teaching environment:

- Ability to adapt to student behaviour.
- Ability to answer questions about the experiment.
- Hide options unless asked for.
- Force students to make choices.
- Create designs.
- Ability to simulate data from chosen design (including steepest ascent optimization, see Fig. 1).
- Ability to analyse simulated data (including determination of stationary points of response surfaces).

We implicitly assume that students have been introduced to the basics of DOE by lectures or self-study. The environment should help students to transfer their theoretical knowledge to practical situations, as well as develop a feeling for practical issues. These requirements, as well as our wish to teach large groups (over 100 students) led us to the choice of developing a dedicated software program for performing virtual experiments. The same conclusion has been reached by others (see e.g., [2] and [3]). We refer to [5] for a discussion on advantages and disadvantages of virtual experimentation environments, including a list of common pitfalls. In order to be flexible, the current version of Statlab is web based, but can also be run as a stand-alone application (see Sect. 3 for technical details).

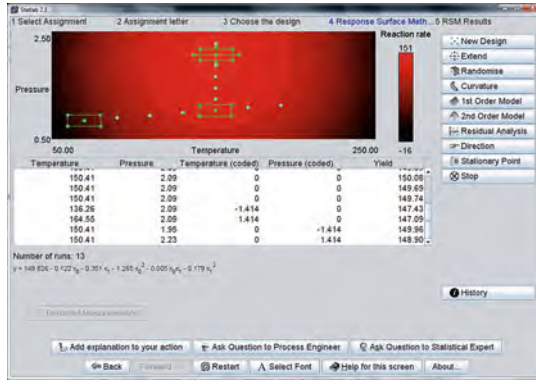


Fig. 1. A screenshot of a response optimization assignment

Statlab currently contains 15 case studies with an indication of its difficulty. The difficulty of a case study depends on a number of issues listed below (between parentheses we list the corresponding statistical notions):

- Restrictions on number of runs (fractions, replications).
- Restrictions on number of experiments under similar circumstances (blocks).
- Restrictions on experimental region (factor level settings, step size).
- Irregularities in the collected data (outliers).
- Knowledge of interactions.
- Curvature of response surface (centre points, lack-of-fit tests).
- Stationary points of the response surface that are not optima (saddle-point).

Of course, there are also issues like randomization that are always part of the case study. Each case study begins with an assignment letter that describes the problem in general engineering terms. We took this idea from the DOE case study in the German on-line statistics tool EMIL@A-stat (`www.emilea.de`). Students have to determine what kind of design is appropriate for this assignment. We currently offer the choice between screening designs, response surface designs and robust (Taguchi) designs.

Questions about the assignment may be asked to the process engineer through key words (see Fig. 2). The tool has a long list of synonyms. The process engineers knows how many experiments may be executed, sometimes has knowledge about interactions, knows whether experiments may be carried out under similar circumstances etc. She has no background in statistics, so she cannot (and is also not willing to) answer questions about randomization or centre points. A statistical expert is available if students wish to get a brief explanation about statistical concepts (again through key words entered by students). In order to teach students that time is limited in industry,

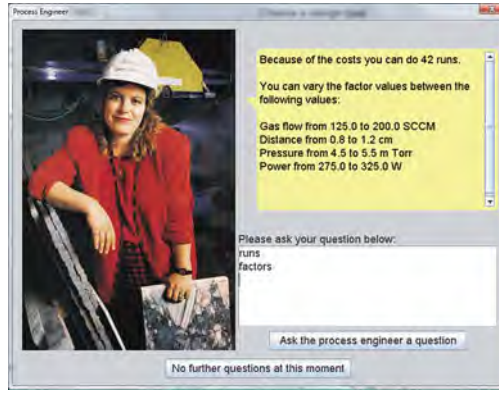


Fig. 2. The student should ask questions about the assignment to the process engineer

the process engineer display increasing levels of irritation when the student keep asking questions. The standard list of possible designs does not contain blocked designs or fractional designs. Students do not get a larger list of designs, unless they enter appropriate key words in the *Design options* field. The current version of Statlab has the following additional features:

- Adds a trend if the student forgets to randomize.
- Adds outliers to simulated data.
- Has a Design Wizard that creates and visualizes blocked fractional factorial designs (including the alias structure).
- Gives students feedback on their work by mentioning possible mistakes.
- Allows students to add explanation to their choices when needed.
- Implements an automatic grading system of student work that provides relevant feedback to the student.

Outliers should be noted by students and reported to the process engineer, who will ask the lab to investigate the suspicious observations. Currently the students get the answer that the observation was indeed wrong and obtain a corrected measurement. The feedback to students only gives possible mistakes, in order to avoid students to push some extra buttons without understanding their omissions. The Design Wizard is also separately available through `www.win.tue.nl/statlab/designApplet.html` and may support lectures on construction of fractional factorial designs.

3 Technical Implementation

Statlab is a freely accessible Java program. It is freely available through the URL `www.win.tue.nl/statlab`. The minimum required Java version is 1.4. The latest Java version can be downloaded freely from `java.com`. In order to

use Statlab during official examinations, the Java security settings must allow Statlab to save the results to the user's hard disk, and send an email to the teacher. When Statlab is started for the first time, the user is prompted to grant these permissions. But even when the Java security settings are correct, firewalls or virus scanners might prohibit Statlab to send email messages. This is why the opening page of Statlab contains a "Detect Java Security" button that checks whether the right Java version is installed and whether Statlab is allowed to save results to disk and send an email. Students should always run this security check on the system that they will use during the exams. By using Java Webstart, the tool can also be used off-line. When the tool is used on-line, automatic updates of the software will be installed without bothering the user. In this way we circumvented managing updates with users of our tool.

Currently, our tool supports two languages (Dutch and English), but the software architecture has been set up in such a way that we can support more languages. The tool automatically detects browser and language settings of the user.

4 Experiences

Throughout the years we have using our tool in various courses. These courses were mainly for Bachelor and Master's students of the Mathematics, Chemical Engineering, Industrial Engineering and Mechanical Engineering Departments, but we also used Statlab in industrial DOE courses. We used our tool both for instruction during lectures, as for official examinations. Using our tool during lectures usually gave rise to lively discussions about experimental design. Initially, students find it difficult to ask simple questions about designs. It is often an eye-opener that in practice the number of experiments that can be executed is restricted. A class demonstration together with the on-line student manual has proved to be sufficient for students to get acquainted with Statlab. Part of the case studies can be accesses freely so that students can practice. Case studies that we use for official examinations are password protected. We specifically ask questions about student experiences with Statlab in the standard course evaluation forms. Most students indicate that they feel that Statlab made statistics more attractive to them because it made them experience the practical sides of statistics. Furthermore, they indicated that using Statlab enhanced their understanding of applying DOE in practice. A minority adapts a minimalist approach by trying to work through the case studies using pre-defined lists of design options and questions to the process engineer.

Since at our university all students possess a personal notebook, it is the responsibility of the students to ensure that Statlab runs well on their notebooks. Over the years we had several hundreds of students using Statlab during official exams, but we only had very few cases where Statlab did not

work during an official examination. For such emergency cases, we always have one or two spare notebooks available. In order to prevent students from communicating via internet with each other during examinations, an ICT expert in our department developed a tool that prevents any communication between notebooks until the assignment has been submitted officially through the internet. Grading of examinations is easy since Statlab generates an easy-to-use, extensive grading report for each student where errors and omissions are marked in red. Instructors have the ability to overrule the grading results of Statlab.

5 Future Developments

Statlab is in continuous development. Although the number of case studies is large compared to other tools that we are aware of, we would like to have more case studies so that users with different backgrounds can choose case studies with context that appeal to them. We welcome feedback and ideas for new case studies from other instructors. There is a teacher manual that will be sent on request to persons that identify themselves are being involved in teaching. We plan to add more complexity in the initial and reporting phases of the assignment, to add functionality in communication with the virtual process engineer in order to imitate real-life situations more closely. Finally, we would like to add other types of designs like mixture designs and optimal designs.

Acknowledgement

The authors would like to thank the Dutch Ministry of Education for a grant to start the development of Statlab. We would also like to acknowledge the significant contributions to the initial development of Statlab by Leendert van Gastel and our former colleagues Mark van de Wiel and Michel Vollebregt.

References

1. Anderson-Cook, C.M., Dorai-Raj, S.: An active learning in-class demonstration of good experimental design. J. Stat. Educ. **9**(1) (2001)
2. Bulmer, M.: Virtual worlds for teaching statistics. CAL-laborate **11**, 1–4 (2004)
3. Darius, P.L., Portier, K.M., Schrevens, E.: Virtual experiments and their use in teaching experimental design. Int. Stat. Rev. **75**(3), 281–294 (2007)
4. Robinson, G.K., Robinson, D.G.: Practical Strategies for Experimenting. Wiley, New York (2000)
5. Schwarz, C.J.: Computer-aided statistical instruction-multi-mediocre techno-trash? Int. Stat. Rev. **75**(3), 348–354 (2007)

Statmaster and HEROS: Web-based Courses First and Second Generation

P.V. Larsen[1] and H. Rootzén[2]

[1] Department of Statistics, University of Southern Denmark, Denmark
[2] DTU Informatics, Technical University of Denmark, Denmark, hero@imm.dtu.dk

Summary. With the increasing focus on life-long learning, and with the convenience and accessibility of the Internet, the market for web-based courses has expanded vastly in recent times – in particular in connection with continuing education. However, teaching web-based courses presents various technical as well as pedagogical challenges. Some of these challenges are addressed, and means to dealing with them are suggested. A second generation of web-based courses is comprised of learning objects, which allows for tailoring courses for specialized groups of students, and accommodate individualized learning. The concept of learning objects and how they are used to form new courses are discussed.

1 Introduction

The first part of this paper, Sect. 2, considers some of the challenges in connection with teaching web-based courses; the second part, Sect. 3, is concerned with the evolution of web-based courses into a second generation of courses based on learning objects. The paper builds on practical experiences from Statmaster: a web-based master degree in applied statistics established in 2002 by four Danish research- and educational institutions: University of Southern Denmark, Technical University of Denmark, The Royal Veterinary and Agricultural University (now Faculty of Life Sciences at University of Copenhagen), and The Danish Institute of Agricultural Sciences (now Faculty of Agricultural Sciences at Aarhus University). The purpose of the Statmaster programme was to offer professionals, working with – or using results from – statistical analyses in their daily routine, a 2 1/2 year part-time master degree in applied statistics as a part of the Danish continuing education scheme. The material from some of the courses constituting the Statmaster programme has since been developed further into learning objects, which have been used to construct a new type of web-based courses that allow for individual students' different learning styles.

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*,
Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_108,

2 Teaching Web-Based Courses

A major strength of e-learning is its flexibility: students can work when and where they want, and at the pace that is most suitable for them. The flexibility is a particular advantage in part-time continuing education, as the students often have families and jobs to mind besides their studies. In the Statmaster programme, we extended the flexibility further by offering print-out versions of the course material, allowing students to study off-line when preferred, e.g. while commuting, and by offering instructor sessions at times convenient for the students for example in the evenings and/or at weekends.

The lack of face to face communication is one of the main weaknesses of e-learning – for some students this can be a barrier to taking active part in the course work. Establishing an atmosphere of a class-room or a student-community can break down the barrier and increase the students' motivation and their engagement in the course. There are various ways to stimulate such a sense of community. For example by including at the e-learning platform an informal discussion forum open for conferences not related to the course work – such as personal introductions of everyone, colloquial conversations, holiday greetings, etc. An additional (or alternative) option to is to arrange a number of assembly days during the course where students and instructors meet in person. A different way to strengthen interaction between students is through group work and group assignments. Our experience from Statmaster suggests that classes with lively colloquial activities in general also participate actively in the course work.

Diversity of the student-body is another potential challenge in continuing education: not only may the students have different professional and educational backgrounds, but also their interests and motivations regarding depth of understanding and complexity of applications may differ widely. To some extent, the flexibility and detached nature of e-learning can be used advantageously by incorporating multiple learning styles and supporting individualized learning. For some students it is sufficient having reading material and a discussion forum; but other students need more direct interaction with the instructor and through for example chat-room sessions, telephone- or video meetings, and/or group work. In addition, video lectures and slide presentations can be employed to provide overviews, and case studies to facilitate coherence and understanding of complex relationships. And, as the Statmaster programme showed us, a very popular motivating factor is to allow students to use their own data or problems in assignments and projects. The tailoring of courses for individual students or specialized groups of students can be taken even further by introducing learning objects as discussed in Sect. 3.

2.1 Technical Solution

The basic requirements of an e-learning platform are that it must be easy to access and easy to use, and it must contain an electronic forum where

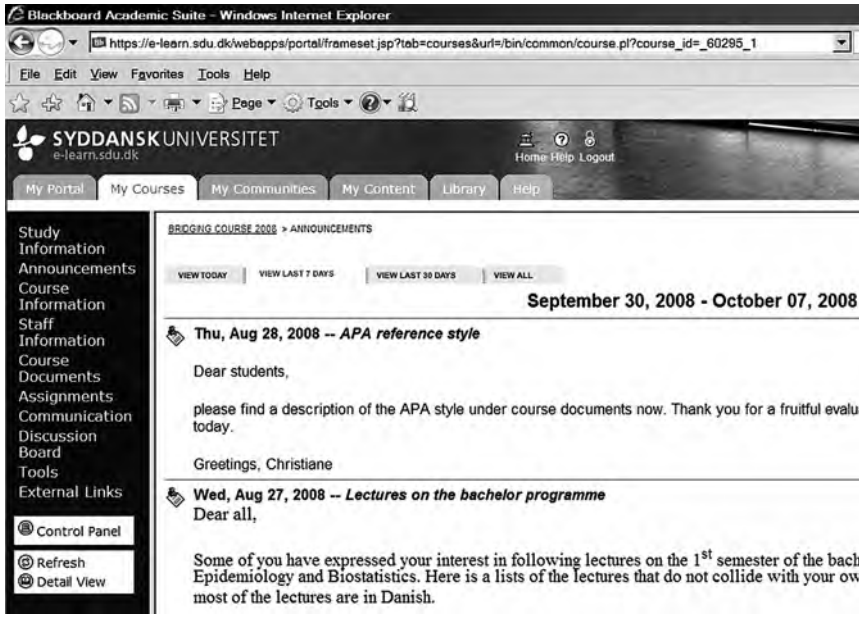


Fig. 1. Blackboard© E-learning platform

students and instructors can discuss the course work. Most e-learning platforms offer various useful additions to the bare basic requirements, such as notice boards, group sub-platforms, virtual meeting rooms, electronic drop-off boxes for assignments, video player, electronic multiple choice revision tests, etc. There exists several e-learning platforms answering to these needs, in Statmaster we used Blackboard© (http://www.blackboard.com) another example is JoomlaLMS© (http://www.joomlalms.com). An example of a course in Blackboard© is shown in Figure 1; here the students can choose between various options e.g. Discussion Board, Assignments or access to Course Documents via the left-hand menu.

3 Non-Linear Learning Using Learning Objects

Our experiences with Statmaster gave us ideas on developing a new concept for e-learning. Students nowadays – whether university students or other kinds – differ much more than before. Some are very good, and some have substantial difficulties even with basic concepts. Moreover, they have very different learning styles. Some learn best by first being exposed to theory and afterwards seeing examples, while others prefer the opposite order of presentation. Some prefer visual and graphical learning, some like to see theory written down in formulas, while others get most out of listening to oral presentations. Using "Leaning Objects" enables each student to design her own course, and ensures

that she gets course material at the right level and in the style that suits her best.

Learning objects represent a relatively new method of subdividing courses into smaller modules. According to David Wiley's contribution to the Educational Workshop in the 16th APAN Meeting (http://www.apan.net/meetings/busan03/materials/ws/education/articles/encyc-DavidWiley.pdf) a learning object is defined as follows: "Any digital resource that can be reused to support learning. The term "Learning Objects" generally applies to educational materials designed and created in small chunks for the purpose of maximizing the number of learning situations in which the resource can be utilized".

Traditional learning – e.g. reading a book or following a lecture – is linear. You start at page 1 or on the first slide and continue on – hopefully understanding a little bit more for every page or slide. To learn in this way is one example of a learning style among many. However, experience and theory suggest that different people learn in different ways. Another way of learning – which is not supported by reading a book – is by obtaining a lot of information before suddenly understanding the whole. For a discussion on different learning styles see [1]. Books and lectures can be good instruments for learning but should not stand alone. To combine them with a more "anarchistic" and non-linear kind of material can improve learning by individualizing it and making it more fun. With learning objects you can build your own course to suit your individual learning style, providing you with optimal learning.

Another reason for proposing a new generation of courses is that we are now faced with a generation of students who are used to exploiting the possibilities of the computer. Thus, a new type of education that will reflect a rethinking of content, form and duration is needed. In the future, education will be in the form of "voucher systems". As a student, you get a set of vouchers and use them to attend the specific chunks of a study programme you need – whenever and wherever it suits you. If the providers of education are to meet the requirements of such systems, the task of developing new courses and tailoring them to individual students has to be easy to manage.

Working with learning objects offers a wide range of flexibility for both the course providers and, as described, for the users. Creating new courses is much easier and more fun if there is a bank of learning objects where you can find most of the topics you need. It is a lot more manageable to make new learning objects, when you have new ideas on how to explain something better or just different than you did before, than rewriting your old book.

The course system we have developed for the second generation courses is called HEROS (Higher Education Re-usable Objects in Statistics). A course in HEROS consists of a collection of learning objects glued together using the e-learning authoring system "Lectora" – a tool made by Trivantis (http://www.trivantis.com). The heart of the system is a hyperbolic graph or a map (http://www.hypergraph.sourceforge.net) which gives an overview of the course, and makes it possible to navigate around in the system – see Figure 2.

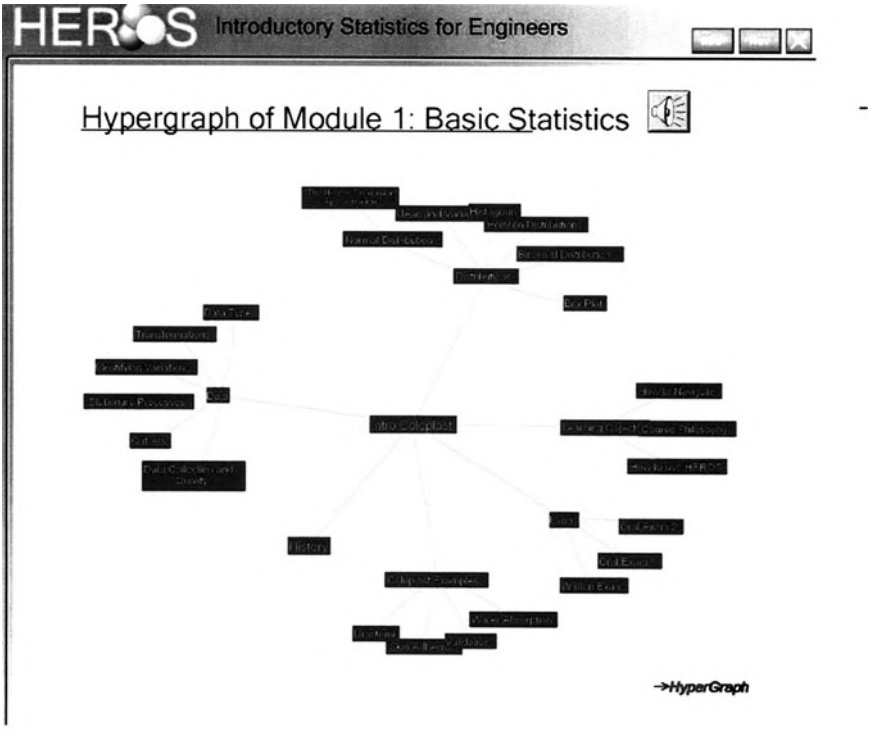


Fig. 2. HEROS: Course graph

The topics are arranged in different "clouds". One cloud consists of different introductions to the system, one is the company's motivations for the course, one is examples from the company, one is an exam, and further clouds consist of statistically oriented topics. When you click on a learning object you enter it and are able to play around e.g. listening to presentations, watching videos, reading, playing with applets and much more.

So far, we have created an introductory course in statistics using HEROS. The course has run twice as a continuing education course for engineers in a global company – the first run was for a Danish audience and the second run for an European audience. The course starts with an introduction – face to face when this is possible; when there are students from abroad we start with a video meeting instead. The students then play around with the system for one month. During that time they work in groups, receive supervision from coaches, and have e-mails from the coaches on things they have to do – all to keep them on track. The course ends with a multiple choice exam and an oral presentation – in person or on the web – of their group work.

4 Conclusion

Despite the lack of face to face communication, e-learning can accommodate many different learning styles by exploiting the diversity of web-based interaction. Indeed, second generation web-based courses using the HEROS system can be shaped to suit individual students with regard to their particular interests, abilities and learning styles.

Sharing our experiences and maybe sharing learning objects could help us making better courses – and make some of the challenges concerning e-learning more manageable.

References

1. Felder, R.M., Brent, R.: J. Engr. Educ. **94**(1), 57–72 (2005)

University Network of Virtual Education in Serbia

A. Tepavčević[1] and M. Heilio[2]

[1] Department of Mathematics and Informatics, University of Novi Sad, Serbia, etepavce@EUnet.yu
[2] Department of Information Technology, Lappeenranta University of Technology, Finland, matti.heilio@lut.fi

Summary. A network of virtual education is being established among four main state universities in Serbia, in the field of mathematics and natural sciences. It is being made following experiences of similar (more developed) networks in Finland and Spain, and using experiences of virtual education in Slovakia, under the framework of EU financed Tempus project SCM C024B06. The basic idea is to prepare electronic version of a certain number of courses (several of them in applied mathematics), so that students of each university can choose these as optional courses in their curriculum. It is planned to construct a web-site from which the courses can be accessed.

1 E-learning: Situation and Constraints

Like elsewhere in Europe, Universities in Serbia are adopting changes commonly known as 'Bologna suggestions', among which network education has a particular role. The challenge of new ways of studying, including e-learning and Web based courses, is more and more recognized by participants in high education. Still, there are many factors influencing these new trends: tradition, present Law of higher education, general situation in high education in the country etc.

In the sequel we analyze the situation, we present our investigation in the field and some results that are obtained in connecting our universities by new ways of teaching. We also stress the relevant connection of high education and industry, mostly in fundamental disciplines as they are applied; in our case it is applied mathematics.

There are four main state universities in the country, several private ones, all are presently under accreditation procedure. Universities consist of faculties, faculties of departments or institutes and chairs. Universities and faculties are legal entities, departments are not. Faculties have considerable independence within universities; connection among faculties is rather weak. Further,

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*,
Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_109,

fundamental disciplines are distributed over faculties. For instance, there are chairs for mathematics not only at faculties of science and mathematics, but also at faculties of technical engineering, agriculture etc., wherever mathematics is the subject. Similar is with computer science, physics, biology etc. Additional argument concerning independence and lack of connection among faculties: it is almost impossible to have students enrolling studies at one faculty and choosing a course at another (even at the same university).

Institutions involved in high education are well equipped with electronic and computational devices. Students are not. Our investigation (by particular questionnaires) has shown that less then 50% of students at two universities in central and southern Serbia do not use computers for learning purposes, they do not regularly (or not at all) have access to Internet. The situation at two northern universities (in particular in Novi Sad) is different. Students do use computers, have access to Internet, and communicate by e-mail with their teachers.

As we pointed out above, there are tendencies to improve teaching and learning methods using Internet and other electronic ways. However, traditional ways of education are still considered to be the most suitable. Either some teachers (or students) are not trained to use the relevant equipment, or they relay to the classical education, considering it unchangeable.

Considering applied mathematics at the university level, the situation is highly connected with our transition economy. Students are mostly interested in financial mathematics, since banks and other financial institutions do show interest for graduated mathematicians. Study programs for industrial mathematics are still not considered to be attractive. In one hand, there is no tradition in such education: the majority of mathematical courses usually use academic examples of 'applications' (this situation is changing presently). On the other hand, enterprises, companies, industry in general, they do not show much interest for young educated mathematicians, ready to solve practical problems. They are still more interested to pay for already developed programs, patents etc. It was so up to now, but also in this field, (slow) changes are visible.

2 Why Virtual Education?

In our opinion, Web based connection among faculties should be concentrated to special courses. Namely, on the higher years of studies (mostly at master level) there are courses with rather low number of students; we have in mind courses in applied mathematics, but also at other study programs in natural sciences. By the new rules, these are optional courses and it may happen that less than five students choose such a course at the beginning of the semester, at one faculty. This is in our opinion one reason for the network. Another is that for some special field (courses) there is no expert at each university in the country. Still the topic might be important.

Particular reason for e-learning approach is also connected to the profile of courses. Namely, in applied sciences special software is usually needed. It should be additionally developed and adopted by the teacher. Then the electronic (and Web) presentation offers convenient possibilities for students in all centers to access the teaching material and to use it in the interactive way. We have also in mind teachers in applied (fundamental) sciences which have well established connections with some company (e.g., industry – applied mathematics, or some medical institution – biology, physics etc.). Then their teaching material is specific and the best way to distribute it to students in other centers is over Internet.

Finally, it is obvious that virtual education in the above sense is not bounded to the region, not even to the particular country. Web supported course could be offered by the best expert at one university to students of any other university, anywhere in Europe.

3 Network of Faculties: Future and Present Status

What do we mean by a network of faculties? Our goal is not reached yet; we intend to develop it in time, step by step. Our vision is related to the existing situation in some other, more developed countries (e.g., Finland). Namely, faculties (of natural sciences and mathematics in our case) should be connected by an educational network which functions almost as a study program (or several study programs): a kind of virtual (part of the) faculty. Each course should be prepared electronically: teaching material, tutorials, exercises, using multimedia (video and text simultaneously) and a suitable interactive programs, together with regular (e.g., once a week) video conferences and permanent e-mail connections with the teacher and teaching assistants. The teacher is an expert from one faculty and students from all (four) centers can choose the course. The exam could be performed either at each faculty (distant examination) or at the faculty of the teacher; there are also other possibilities. Classical direct communication is also a part of education: students should have a possibility to meet the teacher and discuss problems they encounter as they learn.

Next we present our achievements up to now. We have mentioned the constraints. These are reasons for the difference of our network and the above 'ideal' situation. Precise data are given in the next section. We here explain the most important aspects. Four universities, through faculties in natural sciences and mathematics, participate in the network. Courses are chosen among applied fundamental disciplines, mostly from higher years of studies. In addition, most of these courses existed in the study program of only one of the universities. Now they are offered to students in all centers. Each course is prepared electronically and placed on the Web page of the network. Multimedia are present only as an introductory part of the course. Due to the regulation at our faculties (as mentioned above), it is not possible that the

same teacher leads the course and maintain the exam for students in different centers. Therefore, the formal responsibility for the course is distributed over faculties. In each of these, one teacher is responsible and leads (also signs) the final exam together with the teacher who offered the course. There are also hard copies of the teaching material, which are distributed to all students as text-books. Interactive ways for exercises for some of the courses are also possible [1–5].

Our plans for the future improvements of the network are connected with the overall changes in high education in our country. Namely, the more universities become centers for distribution of knowledge over faculties, the less constraints will we have in offering courses to the broad auditorium of students. In addition, we hope to get support from the industry: if companies show interest for particular knowledge in applied fundamental sciences, then relevant teachers as well as students in all centers will be motivated for such special courses. Obviously, the network is the best way for the preparation of some special course and its distribution over centers.

In addition, due to connection among universities in Europe, we plan to prepare some common special courses together with colleagues in other countries: each center will contribute to the course with the topic for which there is a specialized teacher. In our opinion this is also an advantage of Web supported study programs: several experts preparing a single course.

4 Results of the Tempus Projects

4.1 Mathematics Curricula for Technological Development and Accreditation of the Applied Mathematics Program

First Tempus project that influenced a lot the situation at Department of Mathematics and Informatics at University of Novi Sad was CD JEP 17017-2002 project "Mathematics Curricula for Technological Development". The main project objective was: "Post graduate curricula in Mathematics in industry and credit transfer system developed according to EU standards and University of Novi Sad inclusion to ECMI Educational System prepared."

Consortium consisted of University of Novi Sad (coordinating and beneficiary university), TU Dresden (contracting university) Germany, Lappeenranta University of Technology Finland, University of Milan Italy, Belgrade Stock Exchange, Commerce Chamber of Vojvodina, and 2 individual experts from Austria and Poland.

It was a 3 year project (Oct. 2003–Sept. 2006), first phase consisting of the preparation of curriculum and the second phase of introduction of the program at University of Novi Sad in coordination with ECMI.

The main idea of the new introduced program in Applied Mathematics was interdisciplinary, applicable mathematics within the model 3+2 (Bachelor + Master). Master thesis consisted of a real-world problem and intensive cooperation with other departments was planned.

On 12 April, 2008 the Quality Control Commission of Republic of Serbia approved the accreditation of Faculty of Science, University of Novi Sad in the first round: among other programs also the accreditation of the new program of Applied Mathematics was approved.

Approved curricula in Mathematics are within the model 3+2+3 (Bachelor + Master + Doctoral) with two programs in Mathematics on bachelor level: BSc in Mathematics and BSc in Applied Mathematics and also two on master level: Master in Mathematics and Master in Applied Mathematics.

4.2 Network education at Faculties of Science in Serbia

The main objective of this Tempus Structural and Complementary Measures SCM C036A06 Project was "to contribute to the actual changes in high education in Serbia, by improving specific contemporary aspects of teaching, communication and studying possibilities".

EU project partners of this project were University Mateja Bela, Faculty of Natural Sciences, Banska Bystrica, Slovakia (Grant holder), Lappeenranta Technical University, Finland and University of Oviedo, Spain.

Project partners from Serbia were University of Novi Sad (coordinating institution), University of Belgrade, University of Kragujevac and University of Niš.

Duration of the project was 1 year, from 15 June 2007–14 June 2008

Main project outcomes were:

1. Establishing Internet based network framework among faculties of sciences in Serbia.
2. Preparing joint courses for all universities in Serbia (about 10 in this pilot phase).
3. Preparing Internet based network framework for joint courses EU – Serbia.

Main activities were preparation of 11 joint courses and of teaching material and implementation of courses and also training visits of teachers from Serbia to Slovakia, Spain and Finland.

The aim of the project could be also defined as a transfer of good practice from EU universities:

From LUT Finland it was a transfer of good practise from Teaching Mathematical Modelling National Network Project.

University of Oviedo, Spain belongs to Group 9 of Spanish universities, which have joint Internet-based study programmes.

Faculty of Sciences and Mathematics, University Matej Bel Banska Bystrica has developed teaching and communication with students through Moodle.

Minimum requirements for such a network (we produced under the project) are:

Video cameras (with microphones and tripods), software for making internet presentations (Microsoft producer or Adobe premier,) and also at each university computer specialist trained to technically prepare courses.

Teachers should be trained for some basics methods on preparing lectures for Web, teachers and students should learn basic elements of Moodle (or other similar course management system) and such platform should exist at the university.

Methodology of preparing courses:

Introductory lectures for every course were prepared consisting of mixed video and power point presentations. For every lecture the following material was prepared: power point presentation and written teaching material (text books).

Students from one university can chose a course from another university. They can access the course material through Web. Communication teacher-students and students-students is done by Moodle as well as by e-mail, forums, chats etc.

Courses from higher years of studies are chosen because of smaller number of students at each center, lack of teachers for specialized courses and also since students of higher years are more familiar with communication through computer.

Still some open questions are left to be resolved in future:

1. Formal aspects (recognition of courses chosen at other university).
2. Distance knowledge evaluation.
3. Practical question (in Serbia): Does every student have a computer OR whether there are computers at the department at the students' disposal?

References

1. Tempus CD JEP 17017-2002 project "Mathematics Curricula for Technological Development" http://www.im.ns.ac.yu/Projects/Tempus/
2. Tempus SCM C036A06 Project "Network Education of Faculties of Science in Serbia" http://sites.im.ns.ac.yu/projects/Tempus/
3. Teaching and student knowledge verification methods. Study. Tempus, University of Novi Sad, Department of Mathematics and Informatics, 2005–2007
4. Welcome to Moodle! http://moodle.org/
5. Miidla, P.: Web-tool on Differential Equations, Mathematics in Industry Volume 12, Progress in Industrial Mathematics at ECMI 2006, pp. 746-750, Springer Berlin Heidelberg

Introducing eLearning in Industrial Mathematics in Tanzania and Rwanda

Verdiana Grace Masanja

Department of Mathematics, National University of Rwanda, Rwanda,
vmasanja@nur.ac.rw

Summary. Many efforts have been undertaken by African countries to promote the use of eLearning in Higher Education Institutions (HEIs), however, it is noted that the uptake of that little which is available is extremely poor. Although it is largely claimed by many that this dismal state is due to economic and technological circumstances, this presentation argues that most efforts have been invested in infrastructure improvement, increased band width provision, hardware and supporting software technologies acquisition and very minimum investment has been put into training and re-training of educators in eLearning delivery modes. This is the major contributor to poor utilisation of eLearning opportunities in most HEIs in Africa. Examples from Tanzania and Rwanda are presented giving good practice approaches for addressing the challenge of poor uptake of eLearning in HEIs mathematics education. Existing opportunities for Africa's eLearning

1 Political Will and Investment

At continental level, many efforts and resources have been invested in ICT infrastructure improvement, increased bandwidth and hardware supply [1]. In this regard, numerous regional efforts are being implemented by individual HEIs, regional networks such as the Association of Africa Universities (AAU), Inter University Council for East Africa (IUCEA), African Union etc in partnership with donor agencies [3]. Recent surveys on eLearning in Africa suggest that expertise and management skills of the practitioners are vital to the success of eLearning on the continent [7]. A survey presented at the eLearning Africa conference held in Accra, Ghana, on 28–30 May 2008, established that many respondents to the survey were unaware of how to manage eLearning programmes and, furthermore, did not feel that they were involved in the development of eLearning content [6]. Other respondents said the only use of eLearning was in accessing information from the Internet. Training and human capacity building should be emphasised alongside developing infrastructure, the survey concluded.

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*,
Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_110,

1.1 Networks

UbuntuNet Alliance

Numerous efforts have been initiated in Africa to use distance learning and ICTs to promote higher education in Africa. The AAU for example continues to struggle to address the issues of ICT policies, infrastructure and increased cost effective bandwidth for its member universities. In issues of pedagogy and research, national networks of research and education institutions of higher education are being formed and linked to one umbrella network called the UbuntuNet. The UbuntuNet is an alliance of National Research and Education Networks (NRENs) aiming to cover the whole of Africa. It is a new initiative since 2007 and already some African Countries have formed NRENs. These are Democratic Republic of Congo, Kenya, Malawi, Mozambique, Rwanda, Sudan, South Africa, Tanzania, Uganda and Zambia. Others in formation stage are Botswana, Burundi, Ethiopia, Lesotho, Namibia, Somalia, Swaziland and Zimbabwe.

The Alliance has been established to capitalise on the emergence of optical fibre and other terrestrial infrastructure opportunities and thus become the Research and Education Network (REN) backbone of Africa. Tertiary education and research institutions throughout the rest of the world are connected to the Internet using fast low-cost fibre. The UbuntuNet Alliance plans to link NRENs in Africa through Géant, the EU academic and research fibre network, to other academic and research fibre networks around the globe.

The Africa Virtual University (AVU)

AVU initiated in 1997 with World Bank funding is the largest network of Open Distance and e-Learning institutions in Africa. AVU is Established in more than 27 countries with 53 partner Institutions has the ability to work across borders and languages in Anglophone, Francophone and Lusophone Africa. AVU delivers degrees, Diploma Programmes, and short professional courses in Computer Science, Languages & Journalism. AVU has received heavy funding and has been working with existing HEIs in Africa in collaboration with External HEIs to deliver eLearning using modules developed external to Africa. AVU programmes remain un-integrated into ongoing programmes where they are offered. Lecturers, who use eLearning mode in AVU programmes, continue to use the conventional mode of delivery in their HEIs. Obviously the existence of AVU has not permeated the teaching and learning cultures of Africa's HEIs.

The NetTel@Africa

NetTel@Africa [5] is a transnational network for capacity building and knowledge sharing in the information communication technologies (ICT) and

telecommunications (telecom) policy, regulation, and applications whose aim is to build the capacities of policy makers, regulators, private sector operators, consumer advocates, and academic institutions. The NetTel@Africa runs a Postgraduate Training Programme in ICT Policy and Regulation, an International Peer-to-Peer Network, a Research Programme in ICT Policy and Regulation an International Community-to-Community ICT Application Network NetTel@Africa.

NetTel is operational in 18 African partner institutions and offers postgraduate degrees in ICT Policy and Regulation with the engagement of academics from five Universities from the USA and six African and American policy makers and regulators of ICT. The consortium of African universities mentioned above with the active support of the international and continental stakeholders offer the programmes. The programme structure, modules and systems of course offerings and evaluations were developed centrally and monitored by the Academic Board of the NetTel@Africa Network. The Academic Board is constituted by the Deans of the respective faculties of the member universities and the invited experts in the field. It is having a chairperson, vice-chairperson and the committees to monitor the quality issues.

1.2 Wireless Campuses Projects

Currently, most of the Local Area Networks (LANs) in Higher Education Institutions (HEIs) are fixed and most eLearning take place in computer laboratories, lecture halls, libraries, and other structures especially constructed for this work. Most institutions are running short of such facilities. Laboratory space for eLearning is so limited that many classes either do away with eLearning as components of their mode of instruction or just leave it to the individual students or participants to cover on their own outside the normal class hours. Some HEIs have introduced wireless LANs. These have helped do away with the problems listed above; these institutions have adopted to change quite easily. For example at Strathmore University in Kenya, wirelesses LANs with laptops have transformed every lecture hall into an eLearning place. Students of Strathmore are truly learning anytime and anywhere: in the cafeteria, at the recreational areas, in their hideouts, and wherever they spend their time on campus. They do their assignments even when eating in the cafeteria. Students do assignments anywhere even when eating and get immediate feedback [2]. This self-paced learning is restricted to the digital (wireless) campus for Africa this is far from being an ideal situation.

1.3 Mobile Phones Learning Projects

Mobile phone learning (mLearning) provides new avenues for distance and open learning. In Africa, most of our communities are still rural and usually

without basic infrastructure for eLearning. Fixed or land phones are non-existent or next to impossible to get working in such remote and rural communities. The beauty of wireless communication is that with Internet-enabled devices such as laptops with mobile telephone subscription, it is possible to take mLearning into most African rural communities.

At a workshop for Science, Mathematics and Technology teachers held in Ghana in September 2007, a facilitator, Mr. Fred Kofi de Heer-Menlah demonstrated that with a mobile-phone Internet subscription line running the GPRS/EDGE technology, it was possible to introduce the teachers to the Internet and eLearning. MLearning is the future of education in Africa, particularly for the less privileged HEIs and those with spread out campuses. The mobile phone technology is improving and wireless connection is a necessity in most institutions of learning today and in the future, we need to explore the use of wireless technologies to promote mLearning. Diverse eLearning practices are currently being used across the continent. Recent trends show that students in Africa are using mobile phones as part of eLearning. Some projects are being implemented on mLearning; for example the Maths for Girls (M4G) is a project in South Africa whose aim is to teach mathematics to girls using technologies that are not usually permitted in the classroom. In this project, South African female secondary school pupils have made extensive use of videos on cellular phones. In Nigeria, due to Internet connection constraints, mobile phones tutorials are a great help to part-time students at the Nnamdi Azikiwe University. Several projects have been recently launched, such as a Ghana-wide e-learning project for mathematics and science curriculum for primary and secondary school by Intel's World Ahead programme, which is testing similar efforts in Nigeria and South Africa. Although some successful and useful projects on mLearning are being implemented, there is a need to move into sustainable large-scale, long-term implementation of this eLearning mode.

2 eLearning in Applied Mathematics in Tanzania and Rwanda

Rwanda and Tanzania HEIs have been at the forefront of eLearning initiatives in Africa through AAU, IUCEA, hosting the AVU programmes and recently the NREN. HEIs in Tanzania and Rwanda are collaborating with HEIs in Finland in training postgraduate students (masters' level and PhD) and retraining academic staff into the field of industrial mathematics. The recent collaboration needs to reach-out to as many as possible cost effectively, necessitating the use of eLearning.

2.1 eLearning at the University of Dar es Salaam (UD), Tanzania

UD uses the conventional mode of delivery. She instituted an ICT Policy in 1995 and has managed to implement the eLearning platform since 1997

using WEBCT and Blackboard, which are eLearning proprietary software. UD has participated in AVU and has a well established virtual learning centre. UD implemented the eLearning system through the financial support from the Flemish University Council. The major problem that UD faced in the implementation of the project is the issue of software license. It is from this fact that the University of Western Cape (UWC) in South Africa initiated a KEWL (Knowledge Environment for Web-Based Learning) project for developing eLearning platform. Currently the UWC has started another project called KEWL – NextGen project under AVOIR (African Virtual Open Initiatives and Resources). AVOIR is a network of African universities working on Open Source applications. Their primary work at the moment is in developing a next-generation of the KEWL learning management software originally developed at the University of the Western Cape in South Africa. UD is a partners in this project. Currently, UD has replaced the commercial platform Blackboard with KEWL Next Generation Learning Management System (KNG LMS).

2.2 eLearning at the Open University of Tanzania (OUT)

OUT is a purely distance and Open learning University which offers academic degrees, diploma and certificate programmes in diverse fields. The University serves a broad spectrum of local and foreign communities by means of distance education which has made it possible to reach students in their communities. Educational delivery is attained through various means of communication such as broadcasting, telecasting, Information and Communication Technologies (ICT), correspondence enhanced face to face, seminars, contact programmes. In 2004, OUT received funding from SPIDER for the establishment of Tanzania's first eLearning centre, located in Dar es Salaam. With support for SIDA, and collaboration with the Open Polytechnic of New Zealand, OUT has made improvements in eLearning using the open source software Moodle. OUT staff got trained in Moodle administration and how to use Moodle to create and deliver courses. Because of Moodle's low cost, OUT is now able to deliver eLearning using this platform to reach a greater number of students more cost-effectively. The university conducts its operations through Provincial Centres and Study Centres. Currently there are 25 Provincial Centres and 69 Study Centres scattered all over the country. At each Regional centre there are study centres to service distance study students. Within each Province several institutions with adequate facilities have been identified to serve as study centre. For example Secondary schools, Colleges and Institutes.

Also OUT serves students residing in neighbouring countries of Uganda and further North (e.g. Sudan), Kenya, Rwanda and Burundi, Democratic Republic of Congo, Zambia and further South, Mozambique and Indian Ocean Islands (e.g. Seychelles, Comoros) and some students from other countries as far as Europe in Diplomatic Missions. Study centres serve as general points for

project work, interaction with other students, attending seminars and tutorials, practical work and demonstrations and of using reference materials. They also provide counselling and tutoring services for the Open University students as well as physical facilities such as classrooms, libraries and laboratories. Students from outside Tanzania are affiliated to one of the Provincial Study centre. Also OUT has a Centre at Egerton University in Kenya servicing students living in Kenya. While currently UD is the African academic coordinator of the NetTel@Africa network, OUT hosts the Tanzania NREN secretariat.

2.3 eLearning at the National University of Rwanda

The National University of Rwanda (NUR) started the NetTel Post Graduate Diploma in ICT Policy and Regulation in July 2005. NetTel Program at NUR are fully integrated in to the normal management process the financial administration of NetTel is also integrated with the NUR's department of finance. NUR has a unit managed by two staff that takes care of eLearning. NUR is the University earmarked by Rwanda to deliver PhDs and other Research graduate programmes to generate the national highly needed higher level cadres.

3 Challenges

Besides the infrastructure, power supply and hardware problems that have been widely pointed out, the main challenges are lack of experts in the field and the negative attitude towards eLearning by the systems, policies, teachers and learners [4]. There have been many opportunities that remain untapped. For example the Virtual Africa University (operating in 27 Countries of Africa) is running own programmes. UD, OUT and NUR hosted AVU. Mathematics departments did not take up the opportunities mainly because of lack of the know-how in eLearning by academic staff and importation of foreign programmes that could not fit the local needs. Existence of the NetTel project remains within the areas where funding collaboration is available. Mathematics departments are not tapping into this opportunity either. Most arguments for low or non-use of eLearning in Africa is attributed to infrastructure, hardware, cost of technology and attitudes towards eLearning quality and standards. Except for HEIs whose delivery mode is pure open and distance learning, and besides some small projects to retrain practicing teachers, Mathematics has not made use of eLearning efforts and opportunities such as those of AVU and NetTel programmes. Same goes for institutional efforts. Many programmes are in IT related courses, Business studies and to a small extent languages and Journalism. Mathematics has the potential to tap into some existing good practices with collaborations with partners

from experienced countries. Rwanda and Tanzania intend to use the existing infrastructure, facilities and networks to introduce eLearning in Applied Mathematics in collaboration with Finland Universities.

4 Way Forward

UD is running two Masters programmes in Mathematics, one of them is on Mathematical Modelling which is regional programme for the Eastern and Southern Africa region supported by NOMA (Norway). NUR is running Masters Degree in Applied Mathematics focusing on Inverse problems and remote sensing supported by Sida/ SAREC Sweden. Both UD and NUR are collaborating with Finland Universities – Lappeenranta University of Technology and Tampere University of Technology focusing on postgraduate training and academic staff re-training in Industrial Mathematics. The collaboration plans to re-train secondary school teachers in industrial mathematics content and pedagogy. This collaboration can work with the Open University of Tanzania and the NetTel network which is operating in both NUR and UD to introduce eLearning in Industrial Mathematics starting with Financial Mathematics and Mathematics for Policy and Decision Makers. The Open University of Tanzania can also play a major role in re-training of secondary school teachers. Since teachers are scattered in rural areas, the use of eLearning (with mLearning) could be explored.

References

1. Anderson, P.: What is Web 2.0? Ideas, technologies and implications for education. JISC Technology and Standards Watch (2007). http://www.jisc.ac.uk/media/documents/techwatch/tsw0701b.pdf
2. Arabasz, P., Baker, M.B.: Evolving Campus Support Models for E-Learning Courses (2003). http://www.educause.edu/ir/library/pdf/ERS0303/ekf0303.pdf
3. Association of African Universities (AAU), Core programme of activities 2001–2004: themes and sub-themes (2001). http://www.aau.org
4. Hunsinger, J.: How to determine your readiness for mobile e-learning. Information policy (2005). http://i-policy.typepad.com/informationpolicy/2005/04/how_to_determin.html
5. NetTel@Africa http://cvs.uwc.ac.za/kewl_documentation/manuals/Doucuments
6. Opoku, F.B., Scott, C.: A survey of people involved in e-learning in 42 African countries. eLearning Africa Conference, Accra, Ghana, 28–30 May, 2008
7. Sife, A.S., Lwoga, E.T., Sanga, C.: New technologies for teaching and learning: Challenges for higher learning institutions in developing countries (2008)

Part III

Contributed Papers

Management of Several Purifying Plants in the Same Area: A Multi-Objective Optimal Control Problem

L.J. Alvarez-Vázquez[1], N. García-Chan[2], and A. Martínez[1], and M.E. Vázquez-Méndez[2]

[1] Departamento Matemática Aplicada II, ETSI Telecomunicación, Universidad de Vigo, 36310 Vigo, Spain. lino@dma.uvigo.es, aurea@dma.uvigo.es
[2] Departamento Matemática Aplicada, EPS, Universidad de Santiago de Compostela, 27002 Lugo, Spain. netog_g@hotmail.com, ernesto@usc.es

Summary. In this paper we deal with a parabolic multi-objective optimal control problem related to the management of a wastewater treatment system. The problem is studied from a non-cooperative point of view (looking for a Nash equilibrium), and also from a cooperative point of view (looking for Pareto solutions "better" than the Nash equilibrium). Numerical results for a real world situation in the estuary of Vigo (NW Spain) are presented.

1 The Multi-Objective Optimal Control Problem

We consider a shallow water domain Ω located in an urban area with a wastewater treatment system consisting of several purifying plants. We assume that each of the plants is controlled by a different organization and we suppose that each of them has to take care of some sensitive areas, in such a way that a penalty is imposed on the plant if the water pollution levels in one of its associated zones is greater than a threshold level. In each plant there is a purification cost associated to the purification process, and the problem consists of finding the discharge strategy in each plant minimizing costs (purification cost and penalties) at every plant.

In this paper we assume N_E purifying plants discharging wastewater in points $P_1, \ldots, P_{N_E} \in \Omega$, take faecal coliform bacteria (FC) as indicator of the water quality and denote by $m_j(t)$ the mass flow rate of coliform discharged in P_j (with low and up bounds, respectively $0 < \underline{m}_j < \overline{m}_j$). If we define $M_j = \{m_j \in L^\infty(0,T) : \underline{m}_j \leq m_j(t) \leq \overline{m}_j, \text{ a.e. in } (0,T)\}$ and $M = \prod_{j=1}^{N_E} M_j$, then the problem can be formulated (see [1]) as the following multi-objective optimal control problem ($\mathcal{P}$): Find the discharge strategy $m(t) = (m_1(t), m_2(t), \ldots, m_{N_E}(t)) \in M$ which, for $j = 1, \ldots, N_E$, minimizes the functionals

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*, Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_111,

$$J_j(m) = \int_0^T f_j(m_j(t))\,dt + \sum_{i=1}^{n_j} \frac{1}{2\epsilon_i^j} \int_{A_i^j \times (0,T)} \left(\rho(x,t) - \sigma_i^j\right)_+^2 dxdt, \qquad (1)$$

where f_j represents the purification cost at j plant, $A_1^j, \ldots, A_{n_j}^j \subset \Omega$ are the sensitive areas associated to that plant, σ_i^j is the FC threshold in A_i^j, ϵ_i^j is a penalty parameter, $(.)_+$ denotes the *positive part function*, and $\rho(x,t)$ is the FC concentration given by:

$$\left.\begin{array}{ll} \dfrac{\partial \rho}{\partial t} + \mathbf{u}\cdot\nabla\rho - \beta\Delta\rho + \kappa\rho = \dfrac{1}{h}\displaystyle\sum_{j=1}^{N_E} m_j(t)\delta(x-P_j) & \text{in } \Omega\times(0,T), \\ \rho(x,0) = \rho_0(x) & \text{in } \Omega, \\ \dfrac{\partial \rho}{\partial \mathbf{n}} = 0 & \text{on } \partial\Omega\times(0,T). \end{array}\right\} \qquad (2)$$

In this system $\delta(x-P_j)$ denotes the *Dirac measure* at P_j, $\mathbf{n}$ is the unit normal outward vector and $h(x,t)$ (height of water), $\mathbf{u}(x,t)$ (depth-averaged horizontal velocity of water), $\rho_0(x)$ (initial FC concentration), β (viscosity coefficient collecting turbulent and dispersion effects) and κ (experimental coefficient related to the loss rate of FC) are known data.

2 A Non-Cooperative Study: Nash Equilibria

First we recall that each plant is controlled by a different organization which looks for its own discharge strategy ($m_j \in M_j$) in order to minimize its own objective functional J_j. So, we look for a whole discharge strategy (vector $m \in M$) accepted by all of the plant managers in the sense that none can change its strategy without increasing its cost functional, if the others do not change their strategies. This vector $m \in M$ is known as a Nash equilibrium:

Definition 1. *We say that $m = (m_1, \ldots, m_{N_E}) \in M$ is a Nash equilibrium of problem $(\mathcal{P})$ if it verifies that, for all $j = 1, \ldots, N_E$,*

$$J_j(m_1,\ldots,m_j,\ldots,m_{N_E}) = \min_{m_j^* \in M_j} J_j(m_1,\ldots,m_{j-1},m_j^*,m_{j+1},\ldots,m_{N_E}) \qquad (3)$$

Nash equilibria can be characterized by using classical optimal control theory of partial differential equations: For each $j = 1, \ldots, N_E$ we introduce the j-th *adjoint problem*:

$$\left.\begin{array}{ll} -\dfrac{\partial q_j}{\partial t} - \beta\Delta q_j - div(q_j\mathbf{u}) + \kappa q_j = \displaystyle\sum_{i=1}^{n_j} \frac{1}{\epsilon_i^j}\chi_{A_i^j}(\rho-\sigma_i^j)_+ & \text{in } \Omega\times(0,T), \\ q_j(x,T) = 0 & \text{in } \Omega, \\ \beta\dfrac{\partial q_j}{\partial \mathbf{n}} + q_j\,\mathbf{u}\cdot\mathbf{n} = 0 & \text{on } \partial\Omega\times(0,T), \end{array}\right\} \qquad (4)$$

where $\chi_{A_i^j}$ denotes the characteristic function of the set A_i^j, i.e. $\chi_{A_i^j}(x) = 1$ only if $x \in A_i^j$. Then we have the following very useful result (see [2]):

Theorem 1. *A vector $m = (m_1, \ldots, m_{N_E}) \in int(M)$ is a Nash equilibrium of the problem $(\mathcal{P})$ if and only if it verifies the optimality system given by:*

$$\left.\begin{array}{r} \textit{State system } (2), \\ \textit{Adjoint systems } (4), \ \textit{for } j = 1, \ldots, N_E. \\ f_j'(m_j) + \dfrac{1}{h(P_j,t)} q_j(P_j,t) = 0 \ \textit{in } (0,T), \ \textit{for } j = 1, \ldots, N_E. \end{array}\right\} \quad (5)$$

Then, to obtain a Nash equilibrium we introduce a time discretization: we take $N \in \mathbb{N}$, $\Delta t = \frac{T}{N}$, and $t_n = n\Delta t$, for $n = 0, \ldots, N$. We define $M^{\Delta t} = \prod_{j=1}^{N_E} [\underline{m}_j, \overline{m}_j]^N$, and consider the discrete control

$$m^{\Delta t} = (m_1(t^1), \ldots, m_1(t^N), \ldots, m_{N_E}(t^1), \ldots, m_{N_E}(t^N)) \in M^{\Delta t}.$$

The optimality system (5) is now approximated by:

$$\text{Find } \ m^{\Delta t} \in M^{\Delta t} \ \text{ verifying } \ F(m^{\Delta t}) = 0, \quad (6)$$

where the function $F : M^{\Delta t} \subset \mathbb{R}^{N \times N_E} \longrightarrow \mathbb{R}^{N \times N_E}$ is given by:

Algorithm 1. (Computation of $F(m^{\Delta t})$)

Initial inputs: Polygonal approximation Ω_h of Ω, admissible triangulation τ_h of Ω_h, and $m^{\Delta t} \in M^{\Delta t}$.

- Step 1.1: Numerical resolution of the state system:
  Taking $m^{\Delta t} \in M^{\Delta t}$ as data, we solve system (2) by using a characteristic-Galerkin method (see [3]) and obtain, for $n = 0, \ldots, N$, functions $\rho_h^n(x)$ verifying $\rho_h^n(x) \approx \rho(x, t^n)$ in Ω_h.
- Step 1.2: Numerical resolution of the adjoint systems:
  Taking approximations $\rho_h^n(x)$ as data, we solve systems (4) by using the previous characteristic-Galerkin method and obtain, for $n = N, \ldots, 0$ and $j = 1, \ldots, N_E$, functions $q_{jh}^n(x)$ verifying $q_{jh}^n(x) \approx q_j(x, t^n)$ in Ω_h.
- Step 1.3: Time discretization of the optimality condition:
  We compute $F(m^{\Delta t}) = ((f_j'(m_j(t^n)) + \dfrac{1}{h(P_j,t^n)} q_{jh}^n(P_j))_{n=1}^N)_{j=1}^{N_E}$

Finally, a discrete approximation of a Nash equilibrium is obtained from solving problem (6) by any standard numerical method for nonlinear systems.

3 A Cooperative Study: Pareto Solutions

Once we have already obtained a Nash equilibrium, we wonder if it is an optimal solution. That is, the Nash equilibrium is a discharge strategy (m) accepted by all plant managers because if one of them (j plant) changes its particular strategy (m_j), then its particular cost functional (J_j) necessarily increases. But now the question is: If all plant managers are ready to cooperate, can we obtain a *better* strategy which brings off a simultaneously decrease of all cost functionals? According to this we introduce the concept of Pareto solution:

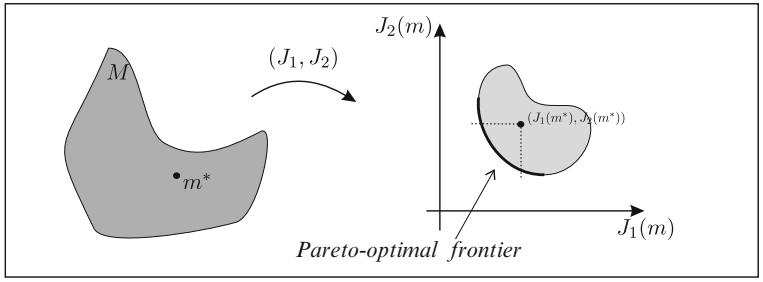


Fig. 1. Geometrical interpretation of Pareto solutions and Pareto-optimal frontier

Definition 2. *We say that* $m = (m_1, \ldots, m_{N_E}) \in M$ *is a Pareto solution of problem* $(\mathcal{P})$ *if there does not exist any* $m^* \in M$ *such that* $J_j(m^*) \leq J_j(m)$, *for all* $j = 1, 2, \ldots, N_E$, *and for at least one* $j \in \{1, 2, \ldots, N_E\}$, $J_j(m^*) < J_j(m)$. *If* $m \in M$ *is a Pareto solution, the objective vector* $(J_1(m), \ldots, J_{N_E}(m))$ *is called Pareto-optimal and the set of Pareto-optimal objective vectors is called Pareto-optimal frontier.*

Figure 1 shows the geometrical interpretation for two plants. An admissible set and its image are illustrated. The fat line is the Pareto-optimal frontier and, for a non Pareto solution $m^* \in M$, dashed lines bound objective vectors corresponding to strategies $m \in M$ better than m^*. Strategies $m \in M$ with image on the arch bounded by dashed lines are Pareto solutions better than m^*.

Pareto solutions can be characterized by means of the weighting method. For each vector $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_{N_E}) \in \mathbb{R}^{N_E}$ such that $\lambda_i \geq 0$, for all $i = 1, \ldots, N_E$, and $\sum_{i=1}^{N_E} \lambda_i = 1$, we introduce the *weighting problem:*

$$\text{minimize } J(m) = \sum_{j=1}^{N_E} \lambda_j J_j(m) \quad \text{subject to } m \in M. \tag{7}$$

We can prove the following very useful result (see [1]):

Theorem 2. *Let* $f_j \in C^1([\underline{m}_j, \overline{m}_j])$ *be strictly convex in* $[\underline{m}_j, \overline{m}_j]$, *for all* $j = 1, \ldots, N_E$. *For each vector* $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_{N_E}) \in R^{N_E}$, $\lambda \geq 0$ *and* $\sum_{k=1}^{N_E} \lambda_k = 1$, *the weighting problem (7) has only one solution. Moreover,* $m \in M$ *is a Pareto solution of problem* $(\mathcal{P})$ *if and only if there exists* $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_{N_E}) \in R^{N_E}$, $\lambda \geq 0$ *and* $\sum_{k=1}^{N_E} \lambda_k = 1$ *such that* m *is a solution of (7).*

From this result, Pareto solutions can be obtained by solving (7) for every weight vector λ. From a computational viewpoint, it is divided in two stages: **Stage 1.** We must fix the number $imax + 1$ of Pareto solutions we are interested in, and we have to choose their corresponding *weights* $\{\lambda^0, \lambda^1, \ldots, \lambda^{imax}\}$.

In this paper we use an algorithm generating the family of weight vectors by splitting the interval $[0,1]$ in a regular way, as given by Caballero et al. [4].
Stage 2. For each $i = 0, 1, \ldots, imax$, we have to solve the problem (7) taking $\lambda = \lambda^i$. In order to do it, we recall the time discretization introduced in section 2, and approach the problem (7) by the discrete problem:

$$\text{minimize } J^{\Delta t}(m^{\Delta t}) \quad \text{subject to } m^{\Delta t} \in M^{\Delta t}, \tag{8}$$

where

$$J^{\Delta t}(m^{\Delta t}) = \sum_{j=1}^{N_E} \lambda_j \Delta t \sum_{n=1}^{N} (f_j(m_j(t^n)) + \sum_{i=n_{j-1}+1}^{n_j} \frac{1}{2\epsilon_i} \int_{A_i} (\rho_h^n(x) - \sigma_i)_+^2 \, dx),$$

and, for $n = 1, \ldots, N$, $\rho_h^n(x)$ is the approximation of $\rho(x, t^n)$ obtained as described in Step 1.1 of algorithm 1. The gradient of $J^{\Delta t}$ at $m^{\Delta t}$ can be also approximated by a discretization of adjoint systems (4). To be exact,

$$\nabla J^{\Delta t}(m^{\Delta t}) \approx ((f_j'(m_j(t^n)) + \sum_{k=1}^{N_E} \lambda_k \frac{1}{h(P_j, t^n)} q_{kh}^n(P_j))_{n=1}^{N})_{j=1}^{N_E},$$

where, for $n = N, \ldots, 1$ and $k = 1, \ldots, N_E$, $q_{kh}^n(x)$ is the approximation of $q_k(x, t^n)$ obtained as described in Step 1.2 of algorithm 1. The discrete problem (8) can now be solved by any method for convex differentiable optimization.

4 Numerical Results

Problem $(\mathcal{P})$ has been solved in a realistic situation posed in the *ría* of Vigo (NW Spain). We have considered two sewage purifying plants, and two sensitive areas, each one associated to its corresponding plant. For the numerical simulation we considered a complete tidal cycle ($T = 12.4\,\text{h}$), chose $N = 120$, supposed $\rho_0 = 0$, and used the height/velocity obtained by solving the shallow water equations on this domain. Related to purification characteristics we have assumed that area associated to plant 1 is more sensitive than area associated to plant 2 ($\sigma_1^1 < \sigma_1^2$), we have taken the same purification cost function for both plants ($f_1 = f_2$) and also same penalty parameters ($\epsilon_1^1 = \epsilon_1^2$).

First we have looked for a Nash equilibrium in this situation and the result can be seen in Fig. 2a. Next, we have looked for Pareto solutions: Figure 3 shows the Pareto-optimal frontier. Cost for plant 1 is represented in the abscissa axis and cost for plant 2 is represented in the ordinate axis. An empty circle represents the cost associated to the Nash equilibrium given in Fig. 2a. As we can see, the Nash equilibrium is not a Pareto solution, and discharge strategies with cost inside the dashed lines are better that the discharge strategy given by the Nash equilibrium. Plant managers have to negotiate to choose one of them (for instance, a reasonable option is that giving a similar improvement – in cost reduction – for both plants). That discharge strategy, with cost pointed out in Fig. 3, is represented in Fig. 2b.

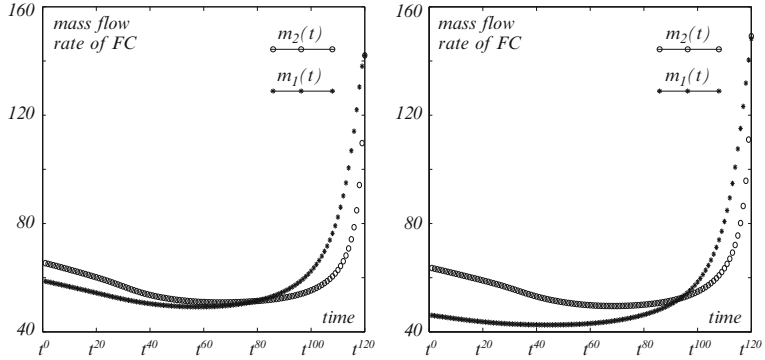


Fig. 2. Optimal discharge strategies

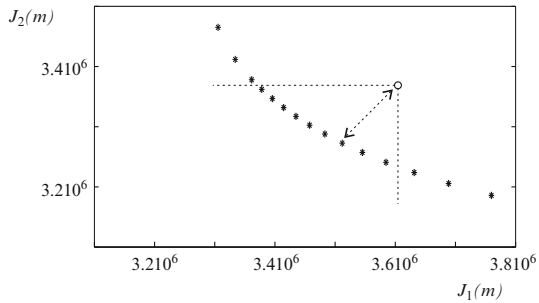


Fig. 3. Pareto-optimal frontier

Acknowledgement

Work partially supported by MCI of Spain (Project MTM2009-07749), Xunta de Galicia (Project Incite 09-291083PR) and CONACyT of Mexico (code 165721).

References

1. Alvarez-Vázquez, L.J., García-Chan, N., Martínez, A., Vázquez-Méndez, M.E.: Comput. Optim. Appl. **46**, 135–157 (2010)
2. García-Chan, N., Muñoz-Sola, R., Vázquez-Méndez, M.E.: ESAIM-control. Optim. Calc. Var. **15**, 117–138 (2009)
3. Alvarez-Vázquez, L.J., Martínez, A., Rodríguez, C., Vázquez-Méndez, M.E.: Appl. Math. Model. **25**, 1015–1024 (2001)
4. Caballero, R., Rey, L., Ruiz, F., González, M.: In: Multiple Criteria Decision Making. Lecture Notes in Econ. and Math. Systems, vol. 448, pp. 275–284. Springer, Heidelberg (1997)

Vector Space of Cooperative Games: Construction of Basis Related with Solutions Based on Marginal Contributions and Determination of Games with Predefined Allocations

R. Amer[1] and J.M. Giménez[2]

[1] Department of Applied Mathematics II, Technical University of Catalonia, ETSEIAT, Colom 11, 08222 Terrassa, Spain, rafel.amer@upc.edu
[2] Department of Applied Mathematics III, Technical University of Catalonia, EPSEM, Bases de Manresa 61, 08242 Manresa, Spain
jose.miguel.gimenez@upc.edu

Summary. Semivalues form a wide family of solutions for cooperative games that assign to each player a weighted sum of its marginal contributions to the coalitions. The Shapley value and the Banzhaf value belong to the family of semivalues. In this work, all semivalues that admit a basis related with the concept of potential are determined, obtaining an explicit expression for its games. Also, for each semivalue whose potential basis has been found, a method to construct all cooperative games with a predefined payoff vector is offered.

1 Introduction

Probabilistic values as solution concept for cooperative games were introduced in [11]. The payoff that assigns a probabilistic value to each player is a weighted sum of marginal contributions to the coalitions, where the weighting coefficients form a probabilistic distribution over the coalitions he/she is a member.

A type of probabilistic values is formed by the semivalues that were defined in [5]. In this case the weighting coefficients are independent of the players and they only depend on the coalition size. Semivalues represent a natural generalization of both the Shapley value [10] and the Banzhaf value [2, 9]. Many properties of these solutions can extend to the set of semivalues. For instance, the potential, which was introduced in [7] for the Shapley value. The potential of the Shapley value assigns to each game and all its restricted games a number recursively obtained, so that the marginal contribution of each player to the potential coincides with the payoff to the player by the

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*,
Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_112,

Shapley value. In a similar way, Dragan [3] defines a potential for the Banzhaf value, as well as a potential for every semivalue on cooperative games [4].

Indeed in [3], it is obtained for the Banzhaf value a series of new concepts and properties already known for the Shapley value. Among these concepts we find a particular basis for the vector space of cooperative games whose determination is directly related with the potential. This basis allows to solve an inverse problem for the Banzhaf value: find all games with a predefined payoff vector. For each vector space of cooperative games, this basis is know as potential basis.

The main purpose of this work consists in finding all games with a pre-established allocation for the greater possible number of semivalues. In a similar way to the Banzhaf value, the process passes through the potential basis, but now it has two levels: (1) determine the semivalues for which a potential basis can be obtained and (2) construct the games of the basis according to the weighting coefficients of each semivalue.

Our inverse problem for semivalues coincides with the resolution of a non-homogeneous system of linear equations. In a classical way, we obtain the solution as a sum of the general solution for the homogeneous system, the so-called *null space*, and a particular solution for the non-homogeneous system, *the short game*. In both cases, the potential basis plays an essential role, since its games are $\{0,1\}$-valued for the potential; these values easily allow to modulate the payoff vector according to predefined allocations.

2 Preliminaries

A *cooperative game* with transferable utility is a pair (N, v), where N is a finite set of *players* and $v : 2^N \to \mathbb{R}$ is the so-called *characteristic function*, which assigns to every *coalition* $S \subseteq N$ a real number $v(S)$, the *gain* or *worth* of coalition S, and satisfies the natural condition $v(\emptyset) = 0$. With G_N we denote the set of all cooperative games on N. For a given set of players N, we identify each game (N, v) with its characteristic function v.

With the usual operations, addition $(v_1+v_2)(S) = v_1(S)+v_2(S)$, and product $(\lambda v)(S) = \lambda v(S)$, $\lambda \in \mathbb{R}$, the set G_N has structure of real vector space. For every nonempty coalition T, the *unity game* 1_T is defined by $1_T(S) = 1$ if $S = T$ and 0 otherwise. The family of unity games $\{1_T \mid \emptyset \neq T \subseteq N\}$ forms a basis in G_N and the dimension of G_N as real vector space is $2^n - 1$.

A function $\psi : G_N \to \mathbb{R}^N$ is called a *solution* and it represents a method to measure the negotiation strength of the players in the game. The payoff vector space $\mathbb{R}^N$ is also called the allocation space. In order to calibrate the importance of each player $i \in N$ in a cooperative game (N, v), we can look at his/her marginal contribution to the coalitions, $v(S) - v(S \setminus \{i\})$. If these contributions are weighted by means of identical weights according to the coalition size, we obtain the solution concept known as *semivalue*, introduced and axiomatically characterized in [5].

The payoff to the players in a game $v \in G_N$ by a semivalue ψ is an average of marginal contributions of each player:

$$\psi_i[v] = \sum_{S \ni i} p_s^n [v(S) - v(S \backslash \{i\})] \quad \forall i \in N \ (s = |S|),$$

where the weighting coefficients $(p_s^n)_{s=1}^n$ verify $\sum_{s=1}^n \binom{n-1}{s-1} p_s^n = 1$ and $p_s^n \geq 0$ for $1 \leq s \leq n$. With $Sem(G_N)$ we denote the set of all semivalues on G_N.

Given a semivalue $\psi \in Sem(G_N)$, $|N| = n$, with weighting coefficients $(p_s^n)_{s=1}^n$, the recursively obtained numbers

$$p_s^m = p_s^{m+1} + p_{s+1}^{m+1} \quad 1 \leq s \leq m < n,$$

define a *induced semivalue* ψ^m (see [4]) on the space of cooperative games with m players. Adding the own semivalue, the so-called *family of induced semivalues by* ψ in spaces of cooperative games with less than or equal n players is formed by $\psi^m \in Sem(G_M)$ with $1 \leq m \leq n$.

If the initial semivalue on G_N is the Shapley value, $p_s^n = 1/[n\binom{n-1}{s-1}]$, the Banzhaf value, $p_s^n = 1/2^{n-1}$, or *binomial semivalues* as they are defined in [1], $p_s^n = \alpha^{s-1}(1-\alpha)^{n-s}$, $\alpha \in (0,1)$, then the induced semivalues are also of the same initial types.

Definition 1. *Let us suppose* $\psi \in Sem(G_N)$ *with weighting coefficients* $(p_s^n)_{s=1}^n$. *The potential of game* v *restricted to coalition* $T \subseteq N$, $T \neq \emptyset$, *according to semivalue* ψ *is defined by*

$$P_\psi(T, v) = \sum_{S \subseteq T} p_s^t v(S) \quad \forall T \subseteq N.$$

We find this definition in [4]. It generalizes the potential for the Shapley value introduced in [7] and also the potential for the Banzhaf value in [3]. The above definition verifies the condition of potential, i.e.,

$$P_\psi(T, v) - P_\psi(T \setminus \{i\}, v) = \psi_i^t[T, v] \quad \forall T \subseteq N, \ |T| \geq 2,$$

and, for $|T| = 1$: $P_\psi(\{i\}, v) = v(\{i\}) = \psi_i^1[\{i\}, v] \quad \forall i \in N$.

3 Potential Basis for Semivalues

A basis in the game space G_N is potential basis with respect to a solution concept that has a potential if the components of every game $v \in G_N$ in this basis agree with the potentials of game (N, v) and its restricted games (T, v), $T \subset N$, $T \neq \emptyset$. We find the potential basis for the Banzhaf value in [3]. Now, we rewrite this definition for any semivalue.

Definition 2. *Let ψ be a semivalue on G_N. A basis in G_N $\{v_S \in G_N \,|\, S \subseteq N, S \neq \emptyset\}$ is potential basis with respect to semivalue ψ iff:*

$$\forall v \in G_N, \quad v = \sum_{S \subseteq N} \alpha_S v_S \;\; \Rightarrow \;\; \alpha_S = P_\psi(S, v).$$

Lemma 1. *A basis in the game space G_N $\{v_S \in G_N \,|\, S \subseteq N, S \neq \emptyset\}$ is potential basis with respect to semivalue ψ on G_N if and only if for every $S \subseteq N$, $S \neq \emptyset$, $P_\psi(S, v_S) = 1$; $P_\psi(T, v_S) = 0 \quad \forall T \subseteq N, T \neq S$.*

Proposition 1. *Let us suppose $\psi \in Sem(G_N)$ with weighting coefficients $(p_s^n)_{s=1}^n$. Every game $v \in G_N$ can be recursively reconstructed from the potential P_ψ if and only if $p_n^n > 0$. Then, the recursive expression is:*

$$v(T) = \frac{1}{p_t^t}\Big[P_\psi(T, v) - \sum_{S \subset T} p_s^t v(S)\Big] \quad \forall T \subseteq N,\ 2 \leq |T| \leq n,$$

and $v(\{i\}) = P_\psi(\{i\}, v) \quad \forall i \in N$.

According to Lemma 1, the games in the so-called potential basis are characterized by the potentials of their restricted games. By means of Proposition 1 we can reconstruct these games from their potentials and obtain an explicit expression for them.

Proposition 2. *[6] Let ψ be a semivalue on G_N whose last weighting coefficient is $p_n^n > 0$. If P_ψ denote the potential of ψ, for every $S \subseteq N$, $S \neq \emptyset$, there exists a unique game $c_{\psi,S} \in G_N$ with $P_\psi(S, c_{\psi,S}) = 1$ and $P_\psi(T, c_{\psi,S}) = 0$ $\forall T \subseteq N$, $T \neq S$, which has like explicit expression:*

$$c_{\psi,S}(T) = \begin{cases} (-1)^{t-s} \displaystyle\sum_{h=0}^{t-s} \binom{t-s}{h} \frac{(-1)^h}{p_{t-h}^{t-h}} & \text{if } T \supseteq S, \\ 0 & \text{otherwise.} \end{cases}$$

Theorem 1. *[6] If ψ is a semivalue on G_N with last weighting coefficient $p_n^n > 0$, then the family of games $C_\psi = \{c_{\psi,S} \in G_N \mid S \subseteq N, S \neq \emptyset\}$ is potential basis in the vector space G_N with respect to the semivalue ψ.*

4 Inverse Problem for Semivalues

The potential for the games in a potential basis only takes values 0 and 1; it leads to simple allocations for these games, as we can see in the next Lemma.

Lemma 2. *Let us suppose $\psi \in Sem(G_N)$ with last weighting coefficient $p_n^n > 0$. If e_j, $1 \leq j \leq n$, are the unit vectors in the standard basis for $\mathbb{R}^n$, for the games of a potential basis $c_{\psi,S}$, $S \subseteq N$, $S \neq \emptyset$, we have:*

$$(a)\ \psi[N, c_{\psi,N}] = \sum_{j=1}^{n} e_j;$$
$$(b)\ \psi[N, c_{\psi,N\setminus\{j\}}] = -e_j \quad \forall j \in N;$$
$$(c)\ \psi[N, c_{\psi,S}] = 0 \quad \forall S \subset N,\ 1 \leq |S| \leq n-2.$$

Definition 3. *Let ψ be a semivalue on G_N. We call null space by ψ to the vector subspace of games in G_N that obtain payoff vector 0 according to semivalue ψ.*

$$NS(\psi) = \{v \in G_N \mid \psi[N, v] = 0\}.$$

Games in a null space are a solution for our inverse problem in a particular case. By means of a vector treatment, for semivalues with non-null last weighting coefficient, the next property shows the solution for the homogeneous inverse problem and, as a result, we can solve the general inverse problem.

Proposition 3. *Let us suppose $\psi \in Sem(G_N)$ with last weighting coefficient $p_n^n > 0$, then $\dim(NS(\psi)) = 2^n - n - 1$ and a basis for $NS(\psi)$ is formed by*

$$\Big\{\ c_{\psi,N} + \sum\nolimits_{j=1}^{n} c_{\psi,N\setminus\{j\}}\ ,\quad c_{\psi,S} \mid 1 \leq |S| \leq n-2\ \Big\}.$$

Corollary 1. *For a given semivalue $\psi \in Sem(G_N)$ with last weighting coefficient $p_n^n > 0$ and a given payoff vector $\eta = (\eta_1, \ldots, \eta_n) \in \mathbb{R}^N$, the solution for the equation*

$$\psi[N, v] = \eta, \tag{1}$$

has by expression:

$$v = \sum_{S\subseteq N,\, 1\leq|S|\leq n-2} \lambda_S\, c_{\psi,S} + \lambda_N \Big[c_{\psi,N} + \sum\nolimits_{j\in N} c_{\psi,N\setminus\{j\}}\Big] - \sum_{j\in N} \eta_j\, c_{\psi,N\setminus\{j\}},$$

where λ_N, λ_S, $1 \leq |S| \leq n-2$, are freedom degrees of the set of solutions; for every selection, the numbers λ_N, λ_S are the potentials of game v on N and games v restricted to S, $1 \leq |S| \leq n-2$, respectively.

Definition 4. *We call short game that verifies equation (1) to the particular solution obtained by imposing $\lambda_N = 0$ and $\lambda_S = 0$ for $1 \leq |S| \leq n-2$; we denote it by $\overline{v}_\psi$.*

$$\overline{v}_\psi = -\sum_{j\in N} \eta_j\, c_{\psi,N\setminus\{j\}}.$$

Game $\overline{v}_\psi$ is a linear combination of games $c_{\psi,N\setminus\{j\}}$, $j \in N$, that only take non-null values on the coalitions $N \setminus \{j\}$ and N. An explicit expression for the short game $\overline{v}_\psi$ is:

$$\overline{v}_\psi = \frac{1}{p_{n-1}^n + p_n^n}\Big[-\sum_{j\in N} \eta_j\, 1_{N\setminus\{j\}} + \frac{p_{n-1}^n}{p_n^n}\Big(\sum_{j\in N} \eta_j\Big)\, 1_N\Big].$$

Theorem 2. *For a given semivalue ψ defined on game space G_N with last weighting coefficient $p_n^n > 0$ and a given vector $\eta = (\eta_1, \ldots, \eta_n) \in \mathbb{R}^N$, the general solution of the non-homogeneous equation $\psi[N, v] = \eta$ is obtained as a sum of the general solution of the homogeneous equation $\psi[N, v] = 0$ and one particular solution of the non-homogeneous equation, i.e.,*

$$v = NS(\psi) + \overline{v}_\psi.$$

Acknowledgement. Research partially supported by Grant SGR 2005-00651 of the Catalonia Government (*Generalitat de Catalunya*) and Grant MTM 2006-06064 of the Education and Science Spanish Ministry and the European Regional Development Fund.

References

1. Amer, R., Giménez, J.M.: Modification of semivalues for games with coalition structures. Theor. Decis. **54**, 185–205 (2003)
2. Banzhaf, J.F.: Weighted voting doesn't work: a mathematical analysis. Rutgers Law Rev. **19**, 317–343 (1965)
3. Dragan, I.: New mathematical properties of the Banzhaf value. TR#300, Department of Mathematics, University of Texas at Arlington, 1–22 (1995)
4. Dragan, I.: Potential and consistency for semivalues of finite cooperative TU games. Int. J. Math. Game Theory Algebra **9**, 85–97 (1999)
5. Dubey, P., Neyman, A., Weber, R.J.: Value theory without efficiency. Math. Oper. Res. **6**, 122–128 (1981)
6. Giménez, J.M.: Contributions to the study of solutions for cooperative games (in Spanish). Ph.D. thesis, Technical University of Catalonia, Spain (2001)
7. Hart, S., Mas-Colell, A.: The potential of the Shapley value. In: Roth, A.E. (ed.) The Shapley Value: Essays in Honor of L.S. Shapley, pp. 127–137. Cambridge University Press, Cambridge (1988)
8. Owen, G.: Multilinear extensions of games. Manage. Sci. **18**, 64–79 (1972)
9. Owen, G.: Multilinear extensions and the Banzhaf value. Nav. Res. Log. Quarterly **22**, 741–750 (1975)
10. Shapley, L.S.: A value for n-person games. In: Kuhn, H.W., Tucker, A.W. (eds.) Contributions to the Theory of Games II, pp. 307–317. Princeton University Press, Princeton (1953)
11. Weber, R.J.: Probabilistic values for games. In: Roth, A.E. (ed.) The Shapley Value: Essays in Honor of L.S. Shapley, pp. 101–119. Cambridge University Press, Cambridge (1988)

Introduction of Measurement Rules on the Nodes of Oriented Structures by Using Concepts of Game Theory

R. Amer[1], J.M. Giménez[2] and A. Magaña[1]

[1] Department of Applied Mathematics II, Technical University of Catalonia, ETSEIAT, Colom 11, E-08222 Terrassa, Spain rafel.amer@upc.edu, antonio.magana@upc.edu

[2] Department of Applied Mathematics III, Technical University of Catalonia, EPSEM, Bases de Manresa 61, E-08242 Manresa, Spain jose.miguel.gimenez@upc.edu

Summary. An oriented structure can model the feasible coalitions according to the sequences of nodes obtained by means of its oriented edges. As an extension of the Shapley value for the classical cooperative games, a solution for games modified by oriented structures is considered in this work. In addition, if a game is symmetric, the allocations only depend on the geometry imposed by the oriented structure. Then, to obtain a measure among the nodes without a predefined game, an exogenous procedure based on a family of symmetric games is proposed.

1 Introduction

A central problem of Game Theory consists of distributing the total utility by using acceptable allocation rules. One of the most important solution concepts for cooperative games is the Shapley value [3], whose payoff to each player is a weighted sum of his/her marginal contributions to the coalitions. This solution verifies symmetry and efficiency – the sum of payoffs to all players equals the utility of the grand coalition N.

In a classical cooperative game, every coalition can form. Nevertheless, a directed graph on N can model the feasible coalitions according to the sequences of nodes obtained by means of its oriented edges. The cooperation is only possible when the players are related by means of the directed graph. This idea directly leads to our solution concept for a game modified by a directed graph: cooperation is possible when there exists accessibility. The introduced solution for games modified by directed graphs is an extension of the Shapley value. If a game is symmetric – the utilities only depend on the coalition size – the Shapley value assigns the same payoff to all players. Thus, according to our introduced solution, the allocations to the players in a symmetric game will only depend on the geometry imposed by the directed edges.

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*, Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_113,

2 Accessibility in a Digraph

A directed graph or *digraph* is a pair (N, D) where N is a finite set of *nodes* and D is a binary relation defined on N. The visual interpretation of pair $(i, j) \in N \times N$ corresponds to an oriented edge that links node i to node j. We consider digraphs without loops. Thus, the *complete digraph* is (N, D_N) with $D_N = N \times N \setminus \{(i, i) \,/\, i \in N\}$. Fixed N, we identify each digraph (N, D) with the binary relation D. In this way, all digraphs on N are the subsets $D \subseteq D_N$.

A *cooperative game* with transferable utility is a pair (N, v), where N is a finite set of *players* and $v : 2^N \to \mathbb{R}$ is the so-called *characteristic function*, which assigns to every *coalition* $S \subseteq N$ a real number $v(S)$, the *worth* of coalition S, and satisfies the natural condition $v(\emptyset) = 0$. With G_N we denote the set of all cooperative games on N. For a given set of players N, we identify each game (N, v) with its characteristic function v. A cooperative game $v \in G_N$ is *superadditive* if $v(S_1 \cup S_2) \geq v(S_1) + v(S_2)$ for every coalitions $S_1, S_2 \subset N$ with $S_1 \cap S_2 = \emptyset$.

Following the formal development for games in generalized characteristic function form in Nowak and Radzik [2], for each nonempty subset $S \subseteq N$, we denote by $H(S)$ the set of all orders of the elements in S. The elements $T \in H(S)$, $\emptyset \neq S \subseteq N$, will be called *ordered coalitions*. A game in *generalized characteristic function form* is a pair (N, v) where N is a finite set of players and v is a function that assigns to every $T \in H(S)$, $\emptyset \neq S \subseteq N$, a real number $v(T)$ with $v(\emptyset) = 0$.

For a nonempty ordered coalition $T = (i_1, i_2, \ldots, i_s) \in H(S)$, we say that i_1 and i_s are, respectively, the *first* and the *last* element in T. As well, i_{j+1} is the *consecutive* element of i_j in T, for $1 \leq j \leq s-1$ (or i_j is the *previous* element of i_{j+1}). And we will say that a subset of consecutive elements in T, $Q = (i_p, i_{p+1}, \ldots, i_{p+u})$ with $1 \leq p \leq p+u \leq s$, is a *consecutive subcoalition* of T.

Definition 1. *Given a digraph D defined on N, a consecutive subcoalition $Q = (i_p, i_{p+1}, \ldots, i_{p+u})$ of T is a connected consecutive subcoalition according to the digraph D if, and only if, $u = 0$ or $(i_j, i_{j+1}) \in D$ for $j = p, \ldots, p+u-1$. If, in addition, (i) $p = 1$ or $(i_{p-1}, i_p) \notin D$ and (ii) $p+u = s$ or $(i_{p+u}, i_{p+u+1}) \notin D$, we say that Q is a maximal connected consecutive subcoalition according to D.*

Note that the individuals $T = (i_p)$, $1 \leq p \leq s$, are connected consecutive subcoalitions.

Definition 2. *Let v and D be a cooperative game and a digraph respectively defined on N. The game v modified by digraph D is the game in generalized characteristic function form defined by*

$$v_D(T) = \sum_{Q \in T/D} v(Q') \quad \forall T \in H(S),\ \forall S \subseteq N,\ S \neq \emptyset,$$

where T/D denotes the set of maximal connected consecutive subcoalitions of T according to digraph D, and Q' denotes the (non-ordered) coalition in N formed with the elements of the ordered subcoalition Q.

Example 1. Let v be the symmetric game defined on $N = \{1, 2, 3, 4\}$ where $v(\{i\}) = 1$ for all $i \in N$, $v(S) = 3$ if $|S| = 2$, $v(S) = 6$ if $|S| = 3$ and $v(N) = 9$. Let us consider the digraph $D = \{(1, 2), (1, 3), (2, 1), (2, 3), (4, 2)\}$ on N. The game v modified by digraph D is defined on the set of all ordered coalitions $H(S)$ with $\emptyset \neq S \subseteq N$. For instance,

$$\begin{aligned} v_D(2, 3, 4) &= v(\{2, 3\}) + v(\{4\}) = 4, \\ v_D(1, 2, 3, 4) &= v(\{1, 2, 3\}) + v(\{4\}) = 7, \\ v_D(1, 2, 4, 3) &= v(\{1, 2\}) + v(\{4\}) + v(\{3\}) = 5, \\ v_D(4, 2, 1, 3) &= v(\{1, 2, 3, 4\}) = 9. \end{aligned}$$

Definition 3. *Fixed N, let v and D be a cooperative game and a digraph defined on N. The accessibility of node i according to v and D is*

$$\alpha_i[v; D] = \frac{1}{n!} \sum_{T \in H(N)} [v_D(T|_i, i) - v_D(T|_i)],$$

where $T|_i$ denotes the consecutive subcoalition of T with the same first element as in T and whose last element is the previous one to element i in T, and $(T|_i, i)$ is the consecutive subcoalition obtained from $T|_i$ adding element i at its end.

Example 2. We return to the game and the digraph introduced in Example 1. In order to obtain the accessibility, we consider the 4! ordered coalitions in $H(N)$ and compute the marginal contributions of each node.

$$\alpha_1[v; D] = \frac{1}{4!}\big\{18v(\{1\}) + 4[v(\{1, 2\}) - v(\{2\})] + 2[v(\{1, 2, 4\}) - v(\{2, 4\})]\big\} = \frac{4}{3}$$

For the remaining nodes: $\alpha_2[v; D] = 3/2$, $\alpha_3[v; D] = 7/4$, $\alpha_4[v; D] = 1$.

The symmetric game v has been modified according to the non-symmetric digraph D. The concept of accessibility gathers this circumstance and offers diverse allocations to the different nodes. In addition, we can see that the sum of allocations to the nodes is $67/12 = 5.5833$ and this value does not agree with the utility that the grand coalition could obtain. The restrictions to the cooperation imposed by digraph D have reduced the sum of allocations to the nodes with respect to the global utility.

Proposition 1. *The notion of accessibility according to a game $v \in G_N$ and a digraph $D \subseteq D_N$ verifies:*

(a) Linearity. For all v, $w \in G_N$ and all λ, $\mu \in \mathbb{R}$, $\alpha[\lambda v + \mu w; D] = \lambda\, \alpha[v; D] + \mu\, \alpha[w; D]$ $\forall D \subseteq D_N$.

(b) *Dummy player. If i is a dummy in game v_D ($v_D(T|_i, i) = v_D(T|_i) + v(\{i\})\,\forall T \in H(N)$), then* $\alpha_i[v; D] = v(\{i\})$.
(c) *Average efficiency. The sum of the accessibilities coincides with the average utility obtained by all ordered coalitions in $H(N)$:* $\sum_{i\in N} \alpha_i[v; D] = \frac{1}{n!} \sum_{T\in H(N)} v_D(T)$.
(d) *For the complete digraph D_N, the accessibility of every node i equals the Shapley value of player i in game v:* $\alpha[v; D_N] = Sh[v]$.
(e) *If a node i is inaccessible in the digraph D, then* $\alpha_i[v; D] = v(\{i\})$.
(f) *The accessibility of a node i does not vary by addition of an oriented edge leaving i:* $\alpha_i[v; D \cup (i, j)] = \alpha_i[v; D]$.
(g) *If game v is superadditive, the accessibility of a node i does not decrease by addition of an oriented edge arriving at i:* $\alpha_i[v; D] \leq \alpha_i[v; D \cup (j, i)]$.

An *oriented path* is a pair (N, P) where N is a finite set of nodes and the binary relation is

$$P = \{(i_1, i_2), (i_2, i_3), \ldots, (i_{l-1}, i_l)\} \text{ with } i_j \neq i_k \text{ for } j \neq k.$$

Proposition 2. *If $v \in G_N$ is a convex game,*

$$v(S_1) + v(S_2) \leq v(S_1 \cup S_2) + v(S_1 \cap S_2) \quad \forall S_1, S_2 \subseteq N,$$

the accessibility of the last node of an oriented path does not decrease by the addition of nodes previous to the first node in the oriented path.

Proposition 3. *Let v be a cooperative game defined on N:*

(a) *For every oriented path P on N, if an oriented edge with the opposite direction is added, then the accessibility of the subsequent nodes do not vary.*
(b) *For every pair of oriented paths P, P' on N with last node i, if $P \cap P'$ also is an oriented path with last node i, then*

$$\alpha_i[v; P \cup P'] = \alpha_i[v; P] + \alpha_i[v; P'] - \alpha_i[v; P \cap P'].$$

(c) *Reduction to oriented paths. The accessibility of node i in a digraph $D \subseteq D_N$ equals its accessibility in the digraph D_i formed by the union of all the oriented paths in D with last node i,*

$$\alpha_i[v; D] = \alpha_i[v; D_i].$$

3 Oriented Structures Without Cooperative Games

We want to study the importance of each node in the system of directed connections of a digraph D without a predefined game.

3.1 Classical Solution

The importance of each node in an oriented structure is proportional to the sum of the importances of all nodes that link to it. The importances are the unknowns of a system of linear equations and the solutions are the components of an eigenvector of matrix $A = (a_{ij})$,

$$a_{ij} = \begin{cases} 1 & \text{if } (j,i) \in D, \\ 0 & \text{otherwise.} \end{cases}$$

According to the idea due to Wei [4] and Kendall [1], the ranking system is based on the eigenvector of matrix A whose components are all positive.

3.2 Exogenous Procedure Based on Game Theory

For every coalition $S \subseteq N$, the *unanimity game* u_S is defined by $u_S(T) = 1$ if $T \supseteq S$ and 0 otherwise.

Definition 4. *From the unanimity games, we construct the so-called test games:*

$$\overline{v}_1 = \sum_{S:\,|S|=2} u_S; \qquad \overline{v}_2 = \sum_{S:\,|S|\geq 2} u_S.$$

Test games $\overline{v}_1$ and $\overline{v}_2$ are symmetric and convex. Game $\overline{v}_1$ – the pairs game – orders the nodes according to each paired comparison. Game $\overline{v}_2$ – the conferences game – orders the nodes according to its relevance in each coalition formed by two or more elements (*conference*).

We obtain rankings for the nodes of an oriented structure based on the accessibility by using symmetric and convex test games as $\overline{v}_1$ and $\overline{v}_2$.

Example 3. Let $D = \{(1,2),(1,5),(2,4),(2,5),(3,2),(4,1),(5,3),(5,4)\}$ be a digraph on the set of nodes $N = \{1,2,3,4,5\}$ without no cooperative game defined on N.

The ranking for the nodes according to the accessibility is

$$\alpha[\overline{v}_1; D] = \frac{1}{120}\,(\ 42,\ 65,\ 42,\ 77,\ 71\ ),$$
$$\alpha[\overline{v}_2; D] = \frac{1}{120}\,(\ 80,\ 96,\ 80,\ 126,\ 108\ ).$$

The classical solution for the ranking requires the matrix

$$A = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 \end{pmatrix}.$$

To compare with $eig^{+}(A)$, we normalize both accessibilities:

$$nor\{\alpha[\overline{v}_1; D]\} = (\ 0.3069,\ 0.4750,\ 0.3069,\ 0.5627,\ 0.5189\ ),$$
$$nor\{\alpha[\overline{v}_2; D]\} = (\ 0.3594,\ 0.4312,\ 0.3594,\ 0.5660,\ 0.4851\ ),$$
$$eig^{+}(A) = (\ 0.3656,\ 0.4267,\ 0.3132,\ 0.5814,\ 0.4981\ ).$$

Both normalized accessibilities are quite close to the *classic* solution, particularly the one that offers test game $\overline{v}_2$.

Acknowledgments

Research partially supported by Grant SGR 2005-00651 of the Catalonia Government (*Generalitat de Catalunya*) and Grant MTM 2006-06064 of the Education and Science Spanish Ministry and the European Regional Development Fund.

References

1. Kendall, M.G.: Further contributions to the theory of paired comparisons. Biometrics **11**, 43–62 (1955)
2. Nowak, A., Radzik, T.: The Shapley value for n-person games in generalized characteristic function form. Games Econ. Behav. **6**, 150–161 (1994)
3. Shapley, L.S.: A value for n-person games. In Kuhn, H.W., Tucker, A.W. (eds.), Contributions to the Theory of Games II, pp. 307–317. Princeton University Press, Princeton (1953)
4. Wei, T.H.: The Algebraic Foundations of Ranking Theory. Cambridge University Press, Cambridge (1952)

Quasicontinuum Method at Finite Temperature Applied to the Study of Nanovoids Evolution in Fcc Crystals

C. Arévalo[1], Y. Kulkarni[2], M.P. Ariza[1], M. Ortiz[3], J. Knap[4], and J. Marian[4]

[1] Escuela Superior de Ingenieros, Universidad de Sevilla, Sevilla, Spain, carevalo@us.es, mpariza@us.es
[2] University of California, San Diego, CA, USA, ykulkarni@uh.edu
[3] Division of Engineering and Applied Science, California Institute of Technology, Pasadena, CA, USA, ortiz@aero.caltech.edu
[4] Lawrence Livermore National Laboratory, Livermore, CA, USA

Summary. Breaking tensile test of ductile materials starts with the formation, in the test material central area, of a choking followed by the nucleation of several cavities at nanoscopic scale. Nanovoids growth and coalescence give rise to a crack which propagates towards the surface in the perpendicular direction to the applied charge. This work is focused in the study of the evolution of these nanovoids for face centered cubic (fcc) crystals. The Quasicontinuum (QC) method at finite temperature has been performed to carry out such an analysis.

1 Introduction

In metals, at room temperature, fracture process entails plastic deformation that the material undergoes before fracture occurs [2]. The failure of many ductile materials occurs in stages that initiate after necking begins. First, small nanovoids form inside the material. Next, deformation continues and the nanovoids enlarge to form a crack. The crack continues to grow and it spreads laterally towards the edges of the specimen. Finally, crack propagation is rapid along a surface that makes about a 45° angle with the tensile stress axis. The new fracture surface has a very irregular appearance [1, 4].

It is, therefore, of great interest for the study of metals behaviour under stress, to understand the growth and the evolution of these, initially nanoscopic size, cavities when stress is applied. This is the aim of this work. The Quasicontinuum (QC) method at finite temperature has been modified to carry out such study. This method is framed inside the multiscale modelling techniques and it is based on a mixed approximation of the system, continuum and atomistic.

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*,
Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_114,

2 QC Method at Finite Temperature

The static theory of the QC model, at zero temperature, was originally developed by Tadmor et al. [13, 15] to solve quasi-static deformation of single crystals in two dimensions. Subsequently, Knap and Ortiz developed the static version of the code in three dimensions [6]. Kulkarni and Ortiz [7], developed a non-equilibrium finite temperature extension of the QC method using a variational formulation based on the maximum entropy principle.

The key idea of QC method is the use of a full atomistic description of the material near the region of interest (nanovoid in this study) while a coarse-grained finite element model is used as we move away from this region and the displacement field becomes slow varying on the scale of the lattice (heterogeneity where the lattice is not highly distorted). In the atomistic area, as a first approach, Lennard-Jones potential has been considered for the interaction between atoms.

A crystal with N atoms is considered in reference configuration occupying a subset $\mathcal{L}$ of a simple Bravais lattice. Denoting the basis vectors by $\mathbf{a_i}$, the reference coordinates of the atoms are:

$$\mathbf{X}(\mathbf{l}) = \sum_{i=1}^{3} l^i \mathbf{a_i}, \mathbf{l} \in \mathcal{L} \subset \mathbb{Z}^3$$

where $\mathbf{l}$ are the lattice coordinates associated with individual atoms. $\mathbf{q}(X)$ is defined as the array of atomic positions in the deformed configuration, where X denotes the configuration space of the ensemble. Total potential energy in the deformed configuration ($\Phi(\mathbf{q})$) is the sum of the atomic interaction energy and the possibility of an external potential. Then, the problem of determining the equilibrium configurations of the system (the main objective of this method) is a problem of seeking the local minima of the energy functional consistent with the boundary conditions.

For systems with a very large number of atoms, this minimization problem presents a significant computational difficulty. To solve this, there are three key components of the QC framework that impart the method its capabilities: constrained minimization problem, lattice summation rules and adaptive refinement.

The essence of the QC theory lies in replacing the former minimum equation, by an approximate minimization problem over a suitable chosen subspace X_h of X. X_h is constructed by selecting a reduced set $\mathcal{L}_h$ of N_h representatives atoms or nodes. Minimization problem is defined now as a minimization over only the representative atoms.

$$\min_{\mathbf{q}_h \in X_h} \Phi(\mathbf{q}_h)$$

Sampling the behavior of the crystal over clusters of atoms around the representative atoms, the equations of equilibrium are reduced to the form:

$$\mathbf{f}_h(\mathbf{l}_h) \approx \sum_{\mathbf{l}'_h \in L_h} n_h(\mathbf{l}'_h) \left[\sum_{\mathbf{l}'_h \in \mathcal{L}_h} \frac{\partial \Phi}{\partial \mathbf{q}(\mathbf{l})} \varphi(\mathbf{l}|\mathbf{l_h}) \right]$$

where $\varphi(\mathbf{l}|\mathbf{l_h})$ denotes the continuous and piecewise linear shape function and $n_h(\mathbf{l}_h)$ are the clusters weights associated with the representative atoms $\mathbf{l}_h$.

The third key component is the use of mesh adaptation in order to tailor the computational mesh to the structure of the deformation field. It has been adopted, as empirical adaptation indicator, the measure of the displacement field variation ε over a simplex K.

$$\varepsilon(K) = \sqrt{|II_{E^d}|} h(K)$$

where II_{E^d} denotes the second invariant of the Lagrangian strain tensor for simplex K and $h(K)$ is the size of K. The element K is deemed acceptable if when you divide $\varepsilon(K)$ by the smallest Burgers vector of the crystal, the results is lower than a tolerance (TOL) value, where TOL is less than 1. The value of TOL involves a compromise between conflicting demands on accuracy and computational efficiency.

The formulation of the QC method at finite temperature is based on the principle of maximum entropy [8]. This principle provides a way to analyze the available information in order to determinate an unique probability distribution function. The principle states that the least biased function maximizes the entropy of the system subject to all the imposed constraints. As in statistical mechanics, the basic idea is to account for the energy contained in the thermal oscillations of the atoms to obtain effective macroscopic thermodynamic potentials while circumventing the treatment of all the atomic degrees of freedom. This goal is achieved by constructing a probability distribution function for the system by way of a mean field approximation. The task of determining the metastable configurations of the crystal, when it is in thermal equilibrium at a uniform temperature T, may be enunciated as follows,

$$\min_{\overline{\mathbf{q}} \in X} \min_{w \in \mathbb{R}^3} \Phi(\overline{\mathbf{q}}, T, w), \quad \Phi(\overline{\mathbf{q}}, T, w) = F(\overline{\mathbf{q}}, T, w) + \Phi^{ext}(\overline{\mathbf{q}})$$

$F(\overline{\mathbf{q}}, T, w)$ is the Helmholtz free energy of the crystal and w approximates the averages of the local frequencies and establishes a link between the energetics of the microscopic dynamics and the effective macroscopic energy of the system. Using the framework of the static theory of QC, described above, the reduced equilibrium equations are of the form:

$$\sum_{\mathbf{l}'_h \in \mathcal{L}_h} n_h(\mathbf{l}'_h) \left[\sum_{\mathbf{l} \in \mathcal{C}(\mathbf{l}'_h)} \frac{\partial \Phi}{\partial \overline{\mathbf{q}}(\mathbf{l})} \varphi(\mathbf{l}|\mathbf{l_h}) \right] = 0$$

$$\sum_{\mathbf{l}'_h \in \mathcal{L}_h} n_h(\mathbf{l}'_h) \left[\sum_{\mathbf{l} \in \mathcal{C}(\mathbf{l}'_h)} \frac{\partial \Phi}{\partial\, w(\mathbf{l})} \varphi(\mathbf{l}|\mathbf{l_h}) \right] = 0$$

where $\mathcal{C}(\mathbf{l}_h)$ represents a cluster of lattice sites within a sphere of radius $r(\mathbf{l}_h)$ centered at the node $\mathbf{l}_h$.

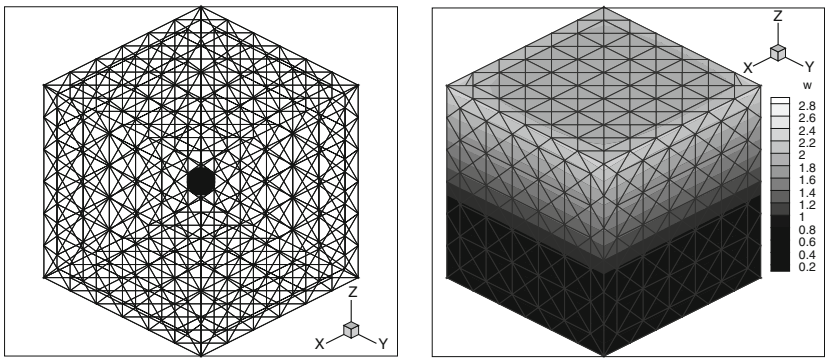
3 Results

The main parameters needed in this study to define a nanovoid deformation problem at finite temperature were based on the study made by Knap and Marian [11,12] for nanovoid in Al with the static method.

Recent experiments data suggest that the material response to a strong shock is essentially volumetric [10]. We therefore drive the void expansion by prescribing pure dilatational displacements over the extension boundary of the computational cell. If ε is the normal axial strain imparted on the sample, we increase ε steadily from $\varepsilon = 0$ to $\varepsilon = 3\%$ by 0.1% increments. At each loading step, a new stable equilibrium configuration is obtained using the Polak-Ribiere variant of the conjugate gradient algorithm [14].

Computational cell is a cube of size $162a_0 \times 162a_0 \times 162a_0$ ($a_0 = 0.5312$) nm corresponding to a size of 86 nm. Solid Argon is used as a test material since it can be modelled using the Lennard-Jones pair potential. The cell is oriented along cubic lattice directions. A 4.25 nm radius void is initially created in the centre of the cell with full atomistic resolution being provided ab initio within an $8a_0 \times 8a_0 \times 8a_0$ region surrounding the void. This system contains 1.5×10^6 atoms reduced to 4,530 nodes in this model. The temperature in the system is 42 K, corresponding to half the melting temperature of solid Ar. As boundary conditions, free surfaces are considered.

For our nanovoid study at finite temperature, the QC simulation was run in two parts: We first allow thermal expansion of the crystal at the specified temperature (42 K) under isothermal conditions. Then, we subject this relaxed crystal to the tri-axial deformation described before. Figure 1 shows (a) the initial simulation mesh with the void in the centre of the box and (b) the initial thermal expansion in z-direction.



(a) (b)

Fig. 1. (**a**) Initial triangulation with the void in the centre of the simulation box. (**b**) Thermal expansion in z-direction (w)

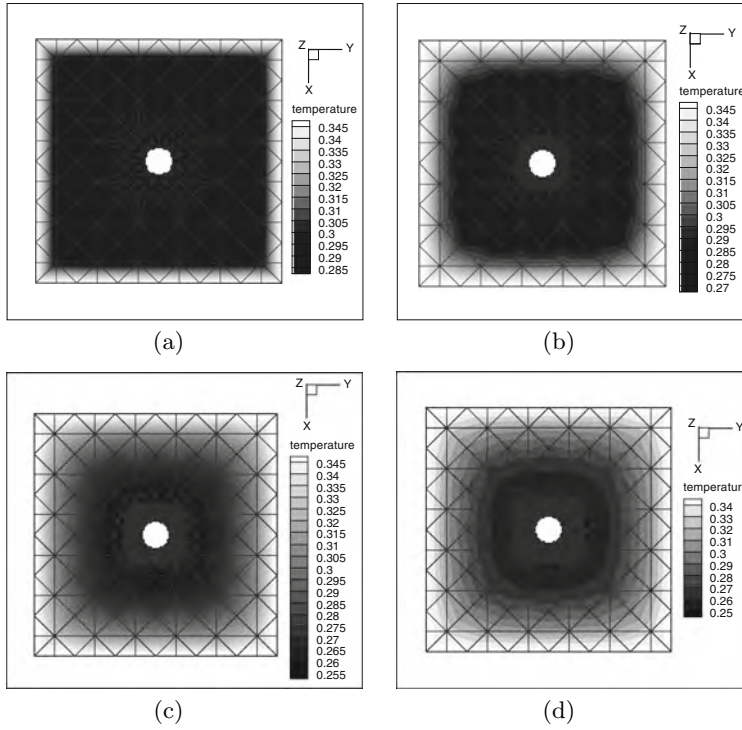


Fig. 2. Snapshots at several deformation steps 0, 1, 2 and 3% respectively, showing the temperature profile of the cross-section Z = 0

Figure 2 represents several snapshots showing the temperature profile of the system for four different deformation steps: 0, 1, 2 and 3% deformation respectively. The unit of temperature is the melting point of the material, $T_m = 83\,\mathrm{K}$. Several aspects must be taken into account: The initial mesh comprises atomistic region only in the vicinity of the nanovoid. When deformation increases, remeshing is clearly observed in the region around the nanovoid in order to capture the microscopic evolution. As the nanovoid expands, the local temperature decreases. At 3% strain, the T drops by 25% around the void.

4 Conclusions and Further Work

In this study, fundamental aspects of the multiscale modelling technique QC have been shown. This technique has been applied to the study of growth and behaviour of nanovoids in order to understand ductile fracture mechanism at finite temperature. Results for Solid Ar at 42 K have been presented, showing temperature profile.

Simulations are now in progress to consider 10.5% deformation to study elastic and plastic regimes. In order to reliable identify the defects and the dislocations in the crystal, centrosymmetry deviation parameter [5] will be used in our calculations. Studies considering bigger domain and higher temperatures will be carried out for fcc aluminium crystal using Ercolessi-Adams interatomic potential [3, 9]. Then we will be able to make a comparison to results obtained at zero temperature with the static model by Marian and co-workers [11].

Acknowledgment

This work has been performed under funding of project P06-TEP-010514 given by Consejería de Innovación, Ciencia y Empresa, Junta de Andalucía, Spain.

References

1. Bai, Y., Dodd, B.: Adiabatic Shear Localization. Pergamon, New York (1992)
2. Callister, W.D.: Materials Science and Engineering: An Introduction, 7th edn. Wiley, New York (2006)
3. Ercolessi, F., Adams, J.B.: Europhys. Lett. **26**, 583–588 (1994)
4. Hai, Q., Enoki, M., Hiraoka, K., Kishi, T.: Eng. Fract. Mech. **72**, 1624–1633 (2005)
5. Kelchner, C.L., Plimpton, S.J., Hamilton, J.C.: Phys. Rev. B **58**, 11085–11088 (1998)
6. Knap, J., Ortiz, M.: J. Mech. Phys. Solids **49**, 1899–1923 (2001)
7. Kulkarni, Y.: Coarse-graining of atomistic description at finite temperature. Ph.D. Thesis, CALTECH, 159 pp. (2007)
8. Kulkarni, Y., Knap, J., Ortiz, M.: J. Mech. Phys. Solids **56**, 1417–1449 (2008)
9. Liu, X.Y., Ercolessi, F., Adams, J.B.: Model. Simul. Mater. Sci. Eng. **12**, 665–670 (2004)
10. Loveridge-Smith, A., et al.: Phys. Rev. Lett. **86**, 2349–2352 (2001)
11. Marian, J., Knap, J., Ortiz, M.: Phys. Rev. Lett. **93**, 165503-1–165503-4 (2004)
12. Marian, J., Knap, J., Ortiz, M.: Acta Mater. **53**, 2893–2900 (2005)
13. Miller, R., Ortiz, M., Phillips, R., Shenoy, V.B., Tadmor, E.B.: Eng. Fract. Mech. **61**, 427–444 (1998)
14. Press, W., Vetterling, W., Teutolsky, S., Flannery, B.: Numerical Recipes C++, 2nd edn. Cambridge University Press, Cambridge (2002)
15. Tadmor, E.B., Ortiz, M., Phillips, R.: Philos. Mag. **73**, 1529–1563 (1996)

Second-Order Asymptotic Expansion for an Eigenvalue Set in Domain with Small Iris

A. Bendali, A. Tizaoui, S. Tordeux, and J.P. Vila

Institut de Mathématique de Toulouse, INSA de Toulouse, 135 avenue de Rangueil, 31077 Toulouse, France, Abderrahmane.Bendali@insa-toulouse.fr, atizaoui@insa-toulouse.fr, sebastien.tordeux@insa-toulouse.fr, Jean-Paul.Vila@insa-toulouse.fr

Summary. We derive the second-order asymptotic expansion of an eigenvalue problem for the Laplace eigenfunction with Dirichlet boundary conditions set in a domain corresponding to two cavities linked by a small iris. Several convergence rates are obtained and illustrated by numerical experiments.

1 Introduction and Motivation

In a turbo engine, the temperature of the combustion chamber can reach 2,000°. In order to protect the structure, small holes are perforated throw the wall linking the combustion chamber to the casing and fresh air is injected. These small holes give rise to disturbance of the acoustic resonance frequencies and modes of the combustion chamber. This has often a negative impact on the combustion but a positive impact on the noise generated by the engine. As a result, a sharp numerical modeling of the effects of these holes has to be performed in order to fulfil two contradictory requirements: ensure a correct functioning of the engine and prevent it from emitting a two high level of noises.

Unfortunately, a direct numerical approach is nowadays technically not feasible due to two main reasons:

- A refined mesh cannot be avoided due to the small characteristic length of the holes.
- The mesh generation of a perforated structure is a hard task, especially when there are a large number of small holes.

This contribution presents a part of a bigger project which aims in providing efficient numerical procedure to take into account the small holes. The desired methods should fulfil the following requirements:

- Mesh refinement are avoided in the neighborhood of the holes.
- Only quantities that can be easily computed have to be used in the numerical procedure.

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*, Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_115,

Two natural approaches can be considered. The first one consists in replacing the effect of the wall by an equivalent transmission condition based on a surface homogenization technique, see for example [2]. The second approach consists in replacing each hole by an equivalent source whose intensity is obtained from a multiscale analysis.

Direct numerical simulations (see for example [4, 7]) do not clearly determine which of these approaches has to be preferred. In this study, we investigate the performance of the equivalent point source method to deal with this kind of problem.

The acted full-wave problem is too complicated to be considered at this stage of the study. Rather a 2-D "toy" model is used to disjoint the main features of the dominant modes and their associated eigenfrequencies, specially their asymptotic behavior related to the size of a characteristic hole. The asymptotic expansion involves a located boundary layer the vicinity of the hole which is dealt with the method of matched asymptotic expansion, see for example [3, 6].

2 Governing Equations

Let $\Omega_{\rm int}$ (the combustion chamber) and $\Omega_{\rm ext}$ (the casing chamber) be two open subsets of $\mathbb{R}^2$ with

$$\Omega_{\rm int} \cap \Omega_{\rm ext} = \emptyset \quad \text{and} \quad \exists a > 0 : \Big(\{0\}\times]-a; a[\Big) \in \partial\Omega_{\rm int} \cap \partial\Omega_{\rm ext}. \tag{1}$$

We consider the domain Ω^δ consisting of Ω_{ext} and $\Omega_{\rm int}$ linked by an iris of width δ

$$\Omega^\delta := \Omega_{\rm int} \cup \Omega_{\rm ext} \cup \Big(\{0\}\times]-\frac{\delta}{2}; \frac{\delta}{2}[\Big) \subset \mathbb{R}^2 \tag{2}$$

which goes to

$$\Omega := \Omega_{\rm int} \cup \Omega_{\rm ext} \subset \mathbb{R}^2, \tag{3}$$

if δ tends to 0.
The eigenvalue problem has the following statement

$$\begin{cases} \text{Find } u^\delta \in \Omega^\delta \to \mathbb{R},\ u^\delta \neq 0; \ \text{ and } \lambda^\delta \in \mathbb{R} \text{ satisfying} \\ -\Delta u^\delta(x,y) = \lambda^\delta u^\delta(x,y) \text{ in } \Omega^\delta, \\ u^\delta(x,y) = 0 \text{ on } \partial\Omega^\delta, \end{cases} \tag{4}$$

$$\begin{cases} \text{Find } u \in \Omega \to \mathbb{R},\ u \neq 0 \text{ and } \lambda \in \mathbb{R} \text{ satisfying} \\ -\Delta u(x,y) = \lambda u(x,y) \text{ in } \Omega, \\ u(x,y) = 0 \text{ on } \partial\Omega. \end{cases} \tag{5}$$

These problems have both a countable set of eigenmodes and associated eigenvalues (Fig. 1):

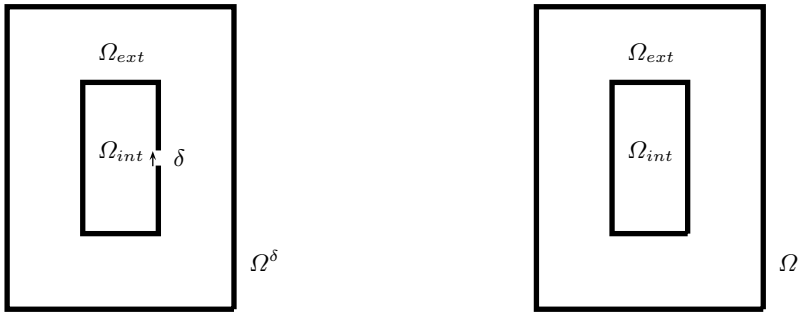


Fig. 1. Geometry of the domain of propagation

- $(u_n^\delta, \lambda_n^\delta)_{n\geq 0}$ (resp. $(u_n, \lambda_n)_{n\geq 0}$) chosen in such a way that $(u_n^\delta)_{n\geq 0}$ (resp. $(u_n)_{n\geq 0}$) is an orthogonal basis of $L^2(\Omega^\delta)$ and $H^1(\Omega^\delta)$ (resp. $L^2(\Omega)$ and $H^1(\Omega)$)

$$\lambda_0 \leq \lambda_1 \leq \lambda_2 \leq \ldots \quad \text{and} \quad \lim_{n\to+\infty} \lambda_n = +\infty, \tag{6}$$

$$\lambda_0^\delta \leq \lambda_1^\delta \leq \lambda_2^\delta \leq \ldots \quad \text{and} \quad \lim_{n\to+\infty} \lambda_n^\delta = +\infty. \tag{7}$$

Derived questions are in order:

- Does the eigenvalue λ_n^δ converge to λ_n?
- Is it possible to obtain an asymptotic expansion of λ_n^δ relatively to δ?
- With this asymptotic expansion, is it possible to construct a numerical method to compute an approximation of λ_n^δ at a small computational cost?

For simplicity, we assume that

The eigenvalues $(\lambda_n)_{n\in\mathbb{N}}$ of the limit problem are all distinct. **(A)**

3 Matching of Asymptotic Expansions

In this study we determine a second-order asymptotic expansion of the eigenvalue λ_n^δ defined by (4) with respect to the size of the iris:

$$\lambda_n^\delta = \lambda_n^0 + \delta\lambda_n^1 + \delta^2\lambda_n^2 + \underset{\delta\to 0}{o}(\delta^2). \tag{8}$$

The obtention of (8) is achieved in parallel to the derivation of the asymptotic expansion of the eigenvector u_n^δ. The method of matching of asymptotic expansions is used to deal with the boundary layer which arises in this kind of problems. We give here a quick overview of the technique.

An Asymptotic Domain Decomposition. The matching of asymptotic expansions consists in expanding the eigenvector u_n^δ with respect to δ

in different scalings. To take care of the propagative phenomenon and of the boundary layer having respectively as a characteristic length $O(1)$ and $O(\delta)$, we will write with the two scalings (x, y) outside a near zone of the iris and $(x/\delta, y/\delta)$ in the vicinity of the iris of asymptotic expansions of the eigenvector u_n^δ.

The Far-Field Expansion. The far-field expansion is defined as the asymptotic expansion of $u_n^\delta(x, y)$ where here (x, y) is varying in the limit domain Ω with no holes.

Looking for a second-order asymptotic expansion, we assume that the following asymptotic expansion

$$u_n^\delta(x, y) = u_n^0(x, y) + \delta u_n^1(x, y) + \delta^2 u_n^2(x, y) + \underset{\delta\to 0}{o}(\delta^2) \tag{9}$$

is holding.

When restricted to a subset of Ω excluding a small neighborhood of the iris, the far-field expansion should provide a good approximation of u_n^δ.

The Near-Field Expansion. The near-field expansion is defined as the asymptotic expansion of $u_n^\delta(x/\delta, y/\delta)$.
Denoting by $(X, Y) = (x/\delta, y/\delta)$ the so called test variables, we assume that $u_n^\delta(\delta X, \delta Y)$ has the following the second-order asymptotic expansion

$$u_n^\delta(\delta X, \delta Y) = \Pi_n^0(X, Y) + \delta \Pi_n^1(X, Y) + \delta^2 \Pi_n^2(X, Y) + \underset{\delta\to 0}{o}(\delta^2). \tag{10}$$

When used in a small neighborhood of iris, the near-field expansion should provide a good approximation of u_n^δ.

The Matching. The matching zone consists in an intermediate zone between the far-field zone and near-field zone. In the matching zone, both the far and near field asymptotic expansions can be used to approximate the same function u_n^δ, and they expressed exactly the matching procedure.

Asymptotic Expansion of λ_n^δ. Under hypothesis $(\mathbf{A})$, the second-order asymptotic expansion (8) of λ_n^δ is given by

$$\lambda_n^0 = \lambda_n,\ \lambda_n^1 = 0 \text{ and } \begin{cases} \lambda_n^2 = -\dfrac{\pi}{16}\dfrac{|\partial_x u_n^0|_{\Omega_{\rm int}}(0,0)|^2}{\|u^0\|_0^2}, \text{ if } u_n^0|_{\Omega_{\rm ext}} = 0, \\ \lambda_n^2 = -\dfrac{\pi}{16}\dfrac{|\partial_x u_n^0|_{\Omega_{\rm ext}}(0,0)|^2}{\|u^0\|_0^2}, \text{ if } u_n^0|_{\Omega_{\rm int}} = 0, \end{cases} \tag{11}$$

(see [1] for the details).

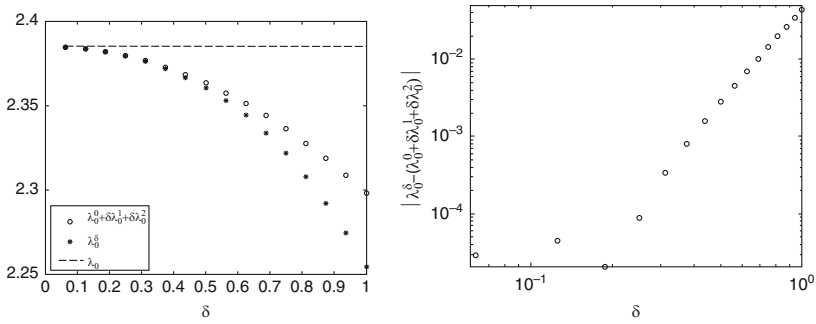


Fig. 2. The *left plot* describes the first eigenvalue of (4) with respect to δ, obtained respectively by the present algorithm (M.A.E.), the direct problem, and limit problem, using a FE of degree 2 in each triangle. In the *right*, we report the error resulting from the second-order asymptotic expansion using the direct numerical procedure as the reference solution

4 Error Estimates

In the previous section, the second-order asymptotic expansions of λ_n^δ has been obtained by means of a formal procedure based on the technique of matching asymptotic expansion. There is no evidence that $\lambda_n^0 + \delta\lambda_n^1 + \delta^2\lambda_n^2$ yields an approximation of the eigenvalue λ_n^δ.

The following theorem provides a theoretical basis for this approximation procedure (for the proof see [1]).

Theorem 1. *Let λ_n^δ (resp. λ_n) the n^{th} eigenvalue of Ω^δ (resp. Ω). Under the hypothesis (A), we have*

$$\left|\lambda_n^\delta - (\lambda_n + \delta^2\lambda_n^2)\right| \leq C\ \delta^3\ |\ln(\delta)|. \tag{12}$$

with λ_n and λ_n^2 given by (11).

5 Numerical Experiments

This section is devoted to numerical validation of the previous second-order asymptotic expansion of λ^δ. The geometry is chosen so that an explicit expression for λ_n^δ is available. The eigenvalue λ_n^δ is computed by using a high order finite element method on a refined triangular mesh. The effective implementation has been done using the GETFEM library [5]. Let Ω^δ be the domain reported in Fig. 1 with Ω_{ext} and Ω_{int} given by (Figs. 2 and 3)

$$\Omega_{\text{int}} =]-2, 0[\times]-2, 1.5[\quad \text{and} \quad \Omega_{\text{ext}} =]0, 2.5[\times]-2.5, 1[. \tag{13}$$

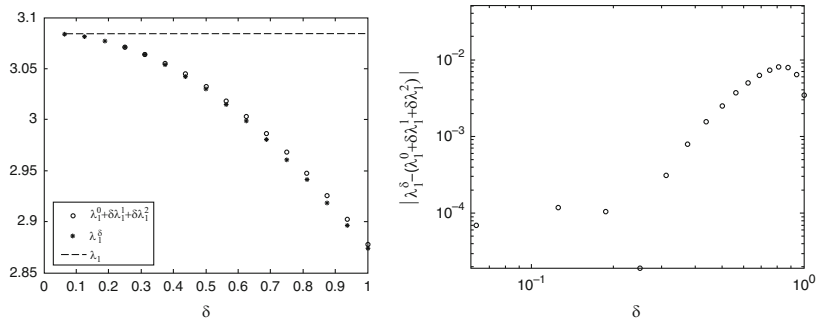


Fig. 3. The *left plot* presents the second eigenvalue of (4) with respect to δ, obtained respectively by the asymptotic expansion, the direct procedure and the problem obtained by neglecting the effect of the iris using the FE in order 2 in each triangle. The *right plot* depicts the error computed in the same way in Fig. 2

6 Conclusion

We have obtained a second-order asymptotic expansion of an eigenvalue problem on a domain consisting of two cavities linked by a small iris (see problem (4)). This expansion yields very sharp approximation of the eigenmodes that can be practically use without any mesh refinement the small iris. The theoretical results are in good agreement with numerical tests.

References

1. Bendali, A., Tizaoui, A., Tordeux, S., Vila, J.P.: Matching of asymptotic expansion for an eigenvalue problem with two cavities linked by a narrow hole. Research Report, INSA of Toulouse (2008)
2. Bonnet-Ben Dhia, A.S., Drissi, D., Gmati, N.: Simulation of muffler's transmission losses by a homogenized finite element method. J. Comput. Acoust. **12**(3), 447–474 (2004)
3. Dyke, M.V.: Perturbation Methods in Fluid Mechanics, Annotated edn. The Parabolic Press, Stanford, xiv+271, 76.41 (1975)
4. Garcia, J.M., Mendez, S., Staffelbach, G., Vermorel, O., Poinsot, T.: Growth of rounding errors and the repetitivity of large eddy simulations. AIAA J. **46**(7), 1773–1781 (2008)
5. http://home.gna.org/getfem/
6. Ilin, A.M.: Matching of Asymptotic Expansions of Solutions of Boundary Value Problems. Translations of Mathematical Monographs, vol. 102. American Mathematical Society, Providence, RI (1992)
7. Mendez, S., Nicoud, F.: Large-eddy simulation of a bi-periodic turbulent flow with effusion. J. Fluid Mech. **598**, 27–65 (2008)

Mathematical Modelling of Fuel Cells

P. Berg

University of Ontario Institute of Technology, 2000 Simcoe Street N., Oshawa, ON, Canada L1H 7K4, peter.berg@uoit.ca

Summary. Fuel cells are electrochemical energy conversion devices that hold great promise to enable a move towards a low-carbon energy economy. However, several technological and scientific obstacles impede their commercialisation, namely, weight, cost, durability and power density issues. This paper discusses what role mathematical modelling plays in fuel cell R&D, and briefly describes four selected topics, as presented in this minisymposium.

1 Fuel Cells

For climate change and energy security reasons, there is a growing need to diversify the energy supply mix of our economy and boost the efficiency of energy conversion devices.

Fuel cells (FCs), which convert chemical energy of fuels efficiently into electrical energy, have the potential to play a key role in this endeavour. These devices circumvent the widely used Carnot cycle by making use of electrochemical reactions which allow for better control of the reaction rate. Because they are not subject to the thermodynamic limit of the Carnot cycle, they exhibit higher efficiency [5].

There are several types of FCs defined by what fuels they convert and how. Among the most popular to-date are (1) proton exchange membrane (PEM) and (2) solid oxide (SO) fuel cells. In both types, hydrogen (H_2), the fuel, reacts with oxygen (O_2) to produce water, separated into two half-cell reactions at anode (a) and cathode (c)

$$\text{PEMFC}: \quad H_2 \rightarrow 2H^+ + 2e^- \,(\text{a}), \quad O_2 + 4H^+ + 4e^- \rightarrow 2H_2O \,(\text{c}) \qquad (1)$$

$$\text{SOFC}: \quad H_2 + O^{2-} \rightarrow H_2O + 2e^- \,(\text{a}), \quad O_2 + 4e^- \rightarrow 2O^{2-} \,(\text{c}) \qquad (2)$$

Ions, namely protons in PEMFC and oxide ions in SOFC, flow between the anode and the cathode in the respective electrolyte; electrons flow through an external circuit, driven by a potential gradient, that can power a load.

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*, Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_116,

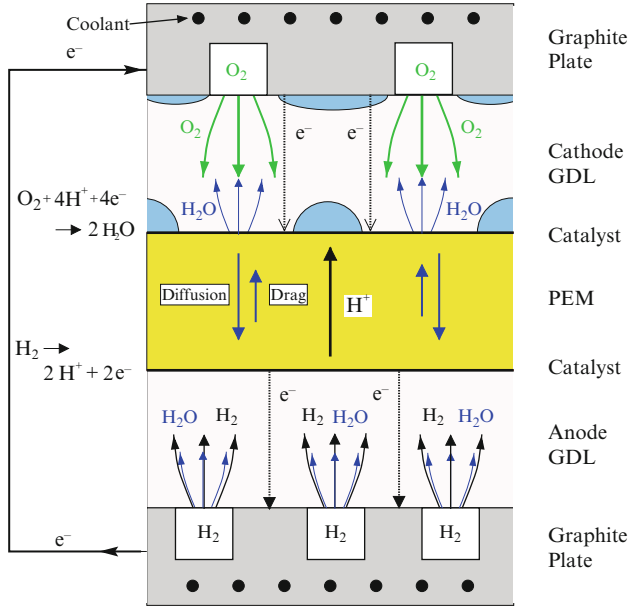


Fig. 1. Cross section of a PEM fuel cell membrane electrode assembly (courtesy of Prof. Keith Promislow, Michigan State University, USA)

Figure 1 shows a cross section of a PEMFC membrane electrode assembly (MEA) perpendicular to the gas flow through the channels. It consists of current collector plates, gas channels, gas diffusion layers (GDLs) which supply the gases to the catalyst layers (CLs) where the reactions take place, and the electrolyte, a proton exchange membrane.

Each domain has its own specific purpose related to mass, charge and heat transfer and, therefore, its own specific scientific and technological challenges [5]. Fundamental materials science is at the core of FC advances, particularly designing, modelling and optimising complex porous media.

2 Mathematical Modelling of Fuel Cells: Key Issues

Fundamental fuel cell research aims at improving critical characteristics, such as costs, power density and durability, related to their commercialisation [2]. Since only a few in situ FC measurement techniques are available, mathematical modelling can greatly help in understanding transport processes which are difficult to observe in detail experimentally. The ultimate goal is to lead the design process through the application of predictive models, resulting in optimised performance.

Some of the most common features and challenges in fuel cell modelling are [5]:

- *Aspect ratios, multiple scales.* While the length of gas channels is typically on the order of centimetres, a catalyst layer can be as thin as 5 µm. This poses challenges when defining numerical grids but, at the same time, it also allows for simplified fuel cell models.
  Moreover, mass transport processes take place at a variety of scales (nm to cm). Processes at the micro or nano level need to be scaled up to derive predictive macroscopic models.
- *Complex porous media.* The GDL, CL and PEM are porous media which consist of two or three phases and they each have a different, yet characteristic pore size distribution. To establish effective macroscopic transport parameters for each domain, homogenisation techniques need to be employed (see Sect. 2.4).
- *Multi-component, two-phase flow (PEMFC).* The gas flow at both anode and cathode consists of more than one species. While SOFCs operate at above 500°C, PEMFCs typically run at 80°C and a pressure of about 2 atm. Since water is produced, this results in two-phase flow in the CL, GDL and channels. The corresponding models are very stiff, owing to the source/sink terms that describe evaporation/condensation. Consequently, numerical convergence often becomes a challenge and the results are quite sensitive to certain boundary conditions.
  In addition, capillary pressure functions are used to describe the flow of liquid water, yet they are difficult to measure experimentally (see Sect. 2.4). Also, the GDL can likely not be modelled as an isotropic domain with uniform characteristics.
- *PEM.* The membrane is a dynamic porous medium whose morphology changes with its water content. PEM water and proton transport is not fully understood, including their related interface phenomena.
- *Optimisation.* Maximising the reaction rate and platinum utilisation in catalyst layers while minimising their degradation, are critical goals. For given materials, these three aspects depend mainly on the choice of the geometry and morphology of the layer (see Sect. 2.3).

2.1 3-D Simulation of a Rolls-Royce SOFC Stack

The first talk, presented by Dr. Ben Haberman [3] (Imperial College, UK), dealt with a 3-D simulation of a Rolls-Royce SOFC stack, containing several unit cells connected in series. The novelty of the design lies in the cathode gas channel which is just bulk air flow, a simple and cost effective solution that also serves as the coolant.

Two key issues which are addressed in this work, are the convective cooling (radiative losses are neglected) provided by the bulk air flow through the stack and the electrical coupling of the unit cells. This simulation tool aids

Rolls-Royce engineers in the design process, resulting in improved cooling and reduced Ohmic losses.

It is vital for R&D staff to have fast solvers at their disposal, something that commercial software often fails to deliver. Dr. Haberman accomplished an order of magnitude decrease in computational (CPU) time by developing his own code, employing parallel computing. The latter is based on domain splitting for which SOFC stacks are well suited, owing to their modular structure, namely unit cells which, in turn, consist of several domains (channels, diffusion layers, catalyst layers, electrolyte).

2.2 Simplified Models for Fuel Cell Stacks

Dr. Andrei Kulikovsky [4] (Research Center Jülich, Germany) pointed out that the lifetime of a unit cell in a stack can be lower than 5,000 h while a single, isolated unit cell can easily operate past 5,000 h. The reason must lie in the impact that thermal and electrical coupling in stacks have on degradation. In order to investigate each, two simplified models are analysed.

Assuming that the cell cooling is mainly provided by the channel flow, a simple heat transfer model for PEMFC unit cells is derived that allows for analytical steady-state solutions for the channel temperature profile, using perturbation methods. It reveals that a balance of (1) water cross-over, (2) liquid water evaporation in the GDL and/or channel, and (3) reaction heat can result in a constant temperature down the channel. In addition, an eigenvalue problem was presented to gauge whether the steady-state solutions are stable. It was concluded that direct methanol fuel cells (DMFCs) should be stable at high temperature.

The second modelling effort addressed how electrically resistive spots in a unit cell, where the through-plane current density might even drop to zero, affect adjacent unit cells through electrical coupling and, hence, impact the stack performance. The conductivity plays a key role in determining how far the perturbation spreads, i.e. how many neighbouring cells are affected. Since Ohm's law results in a Laplace equation for the electrical potential, the voltage distribution within a thin unit cell can be approximated by its boundary values. In principle, this provides an in situ method during fuel cell operation.

2.3 PEMFC Catalyst Layers: Porous Structure and Water Accumulation

Water management and platinum utilisation in PEMFC catalyst layers are critical issues in understanding and optimising these domains, and are driven by their morphology. Dr. Michael Eikerling [6] (Simon Fraser University, Canada) focussed on water flow and reaction kinetics at several scales (nano, micro, macro), trying to assess what role the membrane plays in CLs and how to define the active catalyst surface area.

There is experimental evidence for a bi-modal gas pore distribution in CLs which gives rise to two prevailing values for the capillary pressure within the domain, determined by the local pore size. It was shown that for two-phase (liquid, vapour) flow within the domain, there exists the potential for bi-stability, meaning non-unique steady-state solutions for the same operating conditions. If these steady-state solutions are unstable, this might result in voltage fluctuations of a unit cell for a prescribed, fixed current density. Some experimental results indicate that such fluctuations exist.

This type of research which is currently related to random CL structures, might lead to the design of more rational, ordered morphologies. The goal is to make design and manufacturing serve the functionality of the layer, rather than having functionality limited by the manufacturing process.

2.4 Two-Phase Behaviour in PEMFC Gas Diffusion Layers

In the last talk, Dr. Jürgen Becker [1] (Fraunhofer ITWM, Germany) introduced the novel concept of virtual material design. In particular, his group simulates two-phase flow in GDLs by creating a 3-D model whose fibre morphology is based on synchrotron tomography images, as shown in Fig. 2a. The goal is to extract two-phase mass transport parameters (permeability, diffusivity and capillary pressure) numerically, since they are difficult to measure experimentally.

The first step is to compute the liquid water distribution in the GDL. In this process, pore sizes need to be determined numerically by use of an efficient algorithm that defines a distance for each pore grid point ($350 \times 350 \times 100$ volume pixels, or *voxels*, corresponding to about $1\,\mu m^3$ each) to the pore wall. The capillary pressure distribution, based on the resulting pore size distribution and using the Young-Laplace equation, will determine the liquid water saturation, thereby yielding the much desired capillary pressure function $P = P(s)$ when simulating drainage and imbibition (see Fig. 2c).

The second step is to solve for the fluid flow of the gas phase only, given the microscopic, steady liquid water distribution. Macroscopic transport parameters (permeability, K, and diffusivity, D, in Fig. 2b, d) as functions of saturation are then derived via up-scaling. The resulting functional dependencies, $K(s)$ and $D(s)$, exhibit power laws which can be compared to those in the literature to-date, e.g. the Bruggemann correction.

Specifically, for the permeability the Stokes equation is solved at the micro level and the up-scaling takes place by use of Darcy's law. For the diffusivity, the Laplace equation is solved and the up-scaling uses Fick's law.

3 Outlook

With the increase in computational power, further advances in fundamental modelling of fuel cell processes can be expected, particularly at the micro

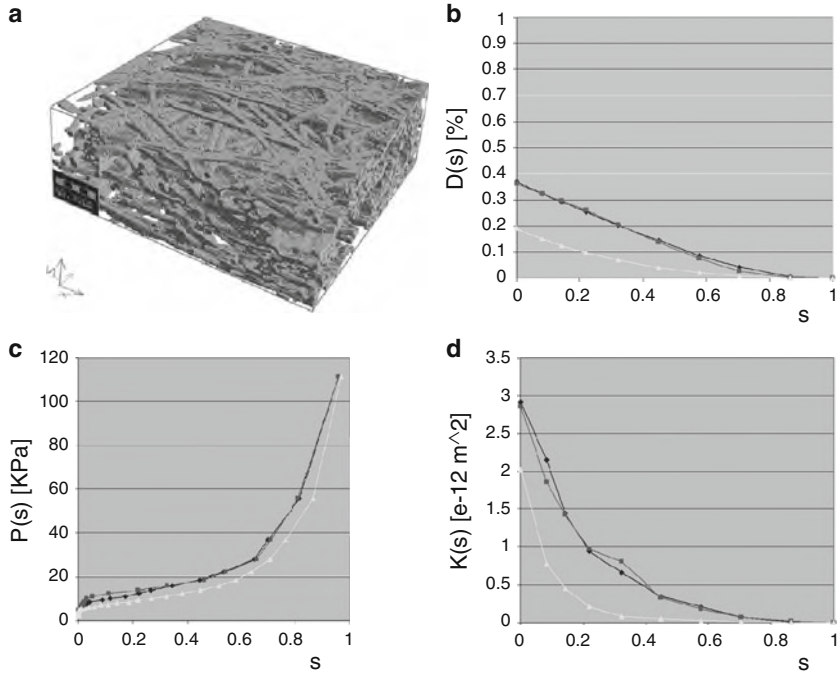


Fig. 2. Simulation of two-phase transport phenomena in a PEM fuel cell GDL: (**a**) computational domain, (**b**) diffusivity (two in-plane, one through-plane), (**c**) capillary pressure (two drainage, one equilibrium curve), and (**d**) permeability (two in-plane, one through-plane). Courtesy of Dr. Jürgen Becker, Fraunhofer ITWM, Germany [1]

and nano scale. Currently, there exists a push towards developing predictive capabilities that can actively guide the design process, thereby replacing the status quo of "passive" research which follows the design process. A key role in advancing PEM fuel cells will be the replacement and/or optimisation of platinum as a catalyst, but it remains an open question whether a breakthrough in this area will emerge from computational research.

References

1. Schulz, P., Becker, J., Wiegmann, A., Mukherjee, P., Wang, C.-Y.: J. Electrochem. Soc. **154**, B419–B426 (2007)
2. Eikerling, M., Kornyshev, A., Kulikovsky, A.: Fuel Cell Rev. **1**, 15–24 (2005)
3. Haberman, B., Young, J.: J. Fuel Cell Sci. Technol. **5**, 011006 (2008)
4. Kulikovsky, A.: J. Electrochem. Soc. **154**, B817–B822 (2007)
5. Li, X.: Principles of Fuel Cells. Taylor & Francis, New York (2006)
6. Liu, L., Eikerling, M.: Electrochim. Acta **53**, 4435–4446 (2008)

Meshless Solution of Singular Potential Flows in Strong Formulation

Francisco Bernal and Manuel Kindelan

G. Millán Institute for Modeling, Simulation, and Industrial Mathematics, Universidad Carlos III de Madrid, 28911 Legans, Madrid, Spain, bernal@maths.ox.ac.uk, kinde@ing.uc3m.es

Summary. In Computational Fluid Dynamics (CFD) it is critical that the numerical solution preserves the total mass of incompressible flows, without introducing spurious sources or sinks. Weak formulations such as the Finite Element Method (FEM) are often preferred, because they implicitly enforce the harmonicity of the approximation. However, these methods typically possess algebraic convergence only and require that a mesh be generated over the computational domain, which may be an expensive task in the event of an expanding fluid.

For that reason, meshless radial basis function (RBF) collocation methods are an appealing alternative to FEM in CFD. We show how to modify the basic setting so that problems involving boundary singularities can also be successfully tackled with RBF collocation. Focussing on an engineering problem (injection molding) we show that RBF collocation can outperform FEM on both simple and non-trivial domains.

1 Introduction

The motivation for this paper is the simulation of plastic injection molding (PIM), a process of industrial interest whereby molten polymer is fed into a thin cavity through an injection machine in order to manufacture plastic parts. The flow is driven by the pressure field $p(x, y)$ modeled by the Hele-Shaw approximation [4], under which the average velocity field may be regarded as two-dimensional and incompressible. In order to simulate the evolution of the flow into the cavity, the following algorithm is used [5]:

1. Solve p at the filled portion of the plan view of the mold at time t, $\Omega(t)$.
2. Compute the velocity field along the front.
3. Update the position of the front according to the velocity at time $t + \triangle t$.
4. Go back to 1 until the mold is entirely flooded.

In this paper we will be concerned only about the first stage of the algorithm, i.e. the elliptic PDE which yields p. Under certain assumptions [1]

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*, Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_117,

the velocity field $\mathbf{v}(x, y)$ can be regarded as potential, $\mathbf{v} \propto \nabla p$, and the instantaneous pressure $p(x, y)$ obeys the following Laplace PDE:

$$\nabla^2 p = 0 \text{ if } \mathbf{x} \in \Omega \tag{1}$$

$$p = p_I \text{ if } \mathbf{x} \in \partial\Omega_I, \qquad p = 0 \text{ if } \mathbf{x} \in \partial\Omega_F, \qquad \frac{\partial p}{\partial n} = 0 \text{ if } \mathbf{x} \in \partial\Omega_W \tag{2}$$

where we have dropped the argument t, and $\partial\Omega_I$, $\partial\Omega_F$, and $\partial\Omega_W$ are the portions of the boundary corresponding to the injection gate(s), the fluid front, and the mold walls, respectively. The pressure exerted by the injection machine, p_I, is assumed constant along the injection segment. Notice that a similar singularity to that of the Motz's problem arises [10], with the BCs changing abruptly from $\frac{\partial p}{\partial n} = 0$ to $p = p_I$ at both ends of the injection segment $\partial\Omega_I$. As a result, the pressure surface is nonsmooth at those points and the straightforward application of RBF collocation fails, as will be explained next.

2 RBF Collocation

In order to illustrate the mechanics of RBF collocation (also known as Kansa's method [6, 7]), we will solve the above PDE on the semicircular domain depicted in Fig. 1. The idea is to discretize it into $N - N_B$ collocation nodes $\mathbf{x}_k$ scattered over Ω and N_B nodes along $\partial\Omega$, and to find an approximate solution in the form

$$p(\mathbf{x}) = \sum_{k=1}^{N} \alpha_k \phi_k(\mathbf{x}) + \sum_{k=N+1}^{N+N_B} \alpha_k \phi_k(\mathbf{x}) \tag{3}$$

where $\phi_k(\mathbf{x}) = \phi(\| \mathbf{x} - \mathbf{x}_k \|)$ is the chosen RBF. There are N_B more RBF centers than nodes, because on the N_B boundary nodes we will want to enforce *both* the BC and the PDE in order to improve results (the so-called PDEBC strategy, discussed in [3]). Such points $\mathbf{x}_j$, $j = N + 1, \ldots, N + N_B$ are placed *outside* Ω and not collocated on. Therefore the number of RBFs must be enlarged to match that of collocation equations – which render the coefficients $\{\alpha_k\}$. They are:

$$\sum_{k=1}^{N+N_B} \alpha_k \nabla^2 \phi_k(\mathbf{x}_i) = 0 \text{ if } \mathbf{x}_i \in \Omega \cup \partial\Omega,\ i \le N \tag{4}$$

$$\sum_{k=1}^{N+N_B} \alpha_k \phi_k(\mathbf{x}_i) = 0 \text{ if } \| \mathbf{x}_i \| = 0.6,\ i \le N \tag{5}$$

$$\sum_{k=1}^{N+N_B} \alpha_k \frac{\partial \phi_k}{\partial n}(\mathbf{x}_i) = 0 \text{ if } x_i = 0 \text{ and } |y_i| > 0.2,\ i \le N \tag{6}$$

$$\sum_{k=1}^{N+N_B} \alpha_k \phi_k(\mathbf{x}_i) = 1 \text{ if } x_i = 0 \text{ and } |y_i| \leq 0.2,\ i \leq N \tag{7}$$

As RBF we have chosen the multiquadric (MQ),

$$\phi_k(\mathbf{x}) = \sqrt{\| \mathbf{x} - \mathbf{x}_k \|^2 + c^2} \tag{8}$$

As $c \to \infty$, the approximation error steadily drops until numerical blow-up occurs – for the MQ becomes more and more flatter, causing a large condition number κ and numerical instability.

The meshless approximation for $N = 281$, $N_B = 56$, and $c = 0.1$ is shown on the left side of Fig. 1 (top row), where ϵ is the error. Gibbs' oscillations are noticeable at both ends $(X_1, Y_1) = (0, 0.2)$ and $(X_2, Y_2) = (0, -0.2)$ of the injection gate, where the pressure field is nonsmooth. The right column of Fig. 1 shows the residual to the PDE, R (minus the Laplacian in this problem). It is apparent that the amplitude of the oscillations has grown by orders of magnitude. In a potential-flow problem such as this, R is the amount to which mass (or area) is locally not conserved, and represents non-physical mass sources or sinks, depending on the sign. The crucial drawback for fluid-mechanical simulations is that in the general case – as well as it happens here – oscillations will not cancel out, resulting in a net integral of R over the domain, thus preventing the simulation of the flow evolution from conserving mass at each iteration. We remark that this kind of BVP and the associated difficulties are representative of a wide class of fluid-mechanical problems. For instance, it also appears in the context of seepage flow [8], or as the first stage of a linearization procedure for solving the full nonlinear Hele-Shaw equation [2].

3 Enrichment of the RBF Interpolant with Singularity-Absorbing Terms

The reason why Kansa's method fails to provide a good approximation is that the solution exhibits nonsmooth features which do not belong in the interpolation space spanned by translates of an infinitely smooth RBF. Therefore, we enrich the interpolant with analytical functions which effectively remove the singularity for the RBFs, as was done in [9]. Consider polar coordinates (r, θ) centered at each of the singularities and seek $f(r, \theta)$ such that

$$\frac{\partial^2 f}{\partial r^2} + \frac{1}{r}\frac{\partial f}{\partial r} + \frac{1}{r^2}\frac{\partial^2 f}{\partial \theta^2} = 0 \tag{9}$$

$$\frac{\partial f}{\partial \theta}(r, \theta = 0) = 0 \qquad f(r, \theta = \pi) = 0 \tag{10}$$

Because the BCs are missing, (9)–(10) is not well posed and its solution is not unique:

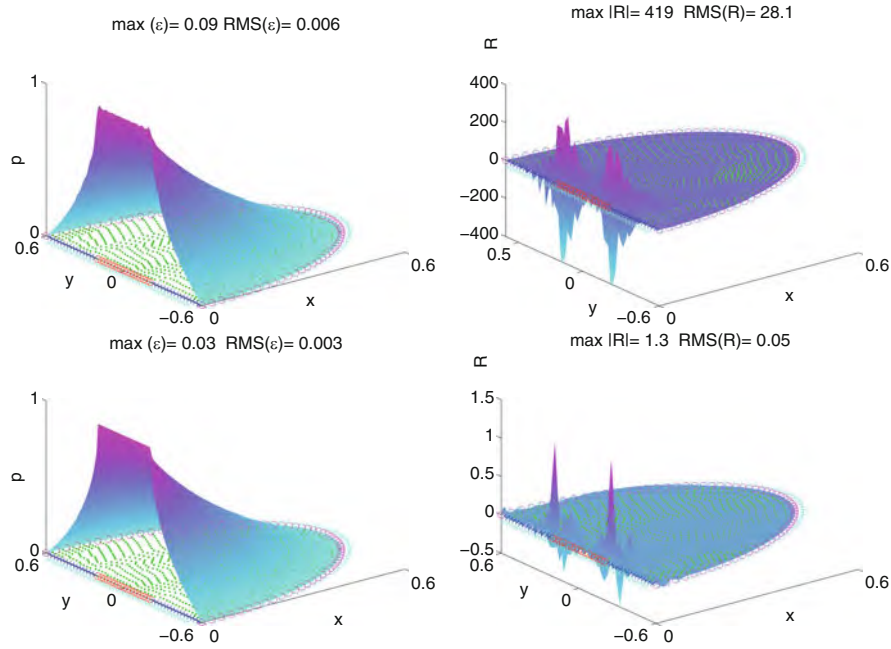


Fig. 1. Conservation of mass in the RBF interpolant: only MQs (*top row*) and enriched (*bottom row*). In both cases $c = 0.1$ and $\kappa = \mathcal{O}(10^{11})$

$$f_k(r,\theta) = r^{(2k-1)/2} \cos\left[\left(\frac{2k-1}{2}\right)\theta\right] \qquad k \geq 1 \tag{11}$$

From this set we pick $f_1(r\theta)$ which reproduces the derivative discontinuity at the origin. Therefore, the enriched interpolant is

$$p = \sum_{k=1}^{N+N_B} \alpha_k \phi_k(\mathbf{x}) + \alpha_{N+N_B+1} \sqrt{r^{(1)}} \cos\frac{\theta^{(1)}}{2} + \alpha_{N+N_B+2} \sqrt{r^{(2)}} \cos\frac{\theta^{(2)}}{2} \tag{12}$$

where $(r^{(i)}, \theta^{(i)})$, $i = 1, 2$ are the polar coordinates centered at either singularity. In order to keep a square system, two further equations are needed. We follow [9] in requiring

$$\sum_{j=1}^{N+N_B} \alpha_j \sqrt{r_j^{(i)}} \cos\frac{\theta_j^{(i)}}{2} = 0, \qquad i = 1, 2 \tag{13}$$

4 Numerical Example on a Nonconvex Domain

The 'elbow' domain is shown on the left side of Fig. 2. It is inscribed in the square $[-1, 1] \times [-1, 1]$ with the centers of the upper and lower circular arcs at

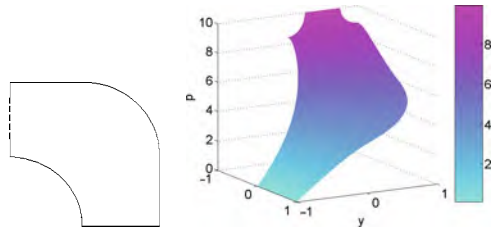


Fig. 2. 'Elbow' domain (*left*). Notice the injection gate (*dashed line*) and the front (*thick line*). *Right*: MQ solution ($c = \sqrt{40/N_{MQ}}$)

Table 1. Comparison of vanilla and enriched RBF interpolant for the elbow domain

	RMS(ϵ)		RMS(R)		$\triangle\Phi$	
$c/ < h >$	n=0	n=1	n=0	n=1	n=0	n=1
3.16	0.15	0.0012	10.21	0.193	−1.285	−0.0195
4.47	0.21	0.0025	14.93	0.216	−1.549	−0.0243
5.48	0.24	0.0034	17.55	0.247	−1.670	−0.0302
6.32	0.14	0.0015	8.62	0.123	−0.881	−0.0176
7.07	0.15	0.0014	8.67	0.105	−0.947	−0.0169
7.75	0.04	0.0014	1.73	0.106	−0.357	−0.0180
8.37	0.03	0.0016	0.74	0.043	−0.322	−0.0007
8.94	0.04	0.0018	0.52	0.027	−0.310	0.0009
9.49	0.06	0.0017	0.11	0.013	−0.259	−0.0003
10.00	0.06	0.0015	0.02	0.009	−0.254	−0.0012
10.95	0.06	0.0013	0.016	0.005	−0.257	−0.0026
11.40	0.06	0.0013	0.018	0.004	−0.262	−0.0033
11.83	0.06	0.0013	0.019	0.004	−0.265	−0.0044
12.25	0.06	0.0016	0.022	0.004	−0.274	−0.0054

$(0, 0)$ and $(-1, -1)$ respectively. The left straight side of the domain therefore has length 1 with the injection gate centered on it and having length 1/2. The injection pressure is $p_I = 10$. The adaptive finite element mesh holding the reference solution is made up of $N = 25{,}034$ vertices. The RBF point set consists of 599 evenly distributed nodes (including the outlying ones), plus $2 \times n$, $n = \{0, 1\}$ enriching functions.

For increasing values of $c/ < h >$ (where $< h >$ is the average distance between collocation nodes), the results yielded by the vanilla RBF method ($n = 0$) are compared with those obtained with the enriched interpolant $n = 1$ (Table 1). Let us define the flow balance as

$$\triangle\Phi = \int_{\partial\Omega_F} \frac{\partial p}{\partial n} dl - \int_{\partial\Omega_I} \frac{\partial p}{\partial n} dl \tag{14}$$

It is apparent that all the estimators RMS(ϵ), RMS(R), and $\triangle\Phi$, drop by at least one order of magnitude from $n = 0$ to $n = 1$. For comparison purposes,

the reference FEM inflow and outflow were -4.599 and 4.585, respectively, thus yielding $\triangle\Phi_{FEM} = -0.014$. On the other hand, if the collocation nodes in the RBF point set were the vertices of a mesh, the resulting value of the flow balance for that FEM approximation would be 1.322. FEM requires a much larger number of vertices for obtaining similar results to those of the enriched RBF scheme, as well as the generation of a mesh. Therefore, the enriched version of RBF collocation outperforms FEM in this problem.

References

1. Bernal, F., Kindelan, M.: RBF meshless modeling of non-Newtonian Hele-Shaw flow. Eng. Anal. Bound. Elem. **31**, 863–874 (2007)
2. Bernal, F., Kindelan, M.: A meshless solution to the p-Laplace equation. In: Ferreira, A.J.M., Kansa, E.J., Fasshauer, G.E., Leitao, V. (eds.) Progress on Meshless Methods 17–35. Springer, Berlin (2009)
3. Fedoseyev, A.I., Friedman, M.J., Kansa, E.J.: Continuation for nonlinear elliptic partial differential equations discretized by the multiquadric method. Int. J. Bifur. Chaos **10**, 481–492 (2000)
4. Hele-Shaw, H.S.: Proceed. Royal Inst. **16**, 49–64 (1899)
5. Hieber, C.A., Shen, S.F.: A finite-element/finite-difference simulation of the injection-molding filling process. J. Non-Newtonian Fluid Mech. **7**, 1–32 (1980)
6. Kansa, E.J.: Multiquadrics – a scattered data approximation scheme with applications to computational fluid-dynamics. I. Surface approximations and partial derivative estimates. Comput. Math. Appl. **19**, 127–145 (1990)
7. Kansa, E.J.: Multiquadrics – a scattered data approximation scheme with applications to computational fluid-dynamics. II. Solutions to parabolic, hyperbolic and elliptic partial differential equations. Comput. Math. Appl. **19**, 147–161 (1990)
8. Li, J., Cheng, A.H.D., Chen, C.S.: A comparison of efficiency and error convergence of multiquadric collocation method and finite element method. Eng. Anal. Bound. Elem. **27**, 251–257 (2003)
9. Platte, R., Driscoll, T.A.: Computing eigenmodes of elliptic operators using radial basis functions. Comput. Math. Appl. **48**, 561–576 (2004)
10. Wait, R., Mitchell, A.R.: Corner singularities in elliptic problems by finite element methods. J. Comp. Phys. **8**, 45–52 (1971)

Estimation of a Piecewise Constant Function Using Reparameterized Level-Set Functions

Inga Berre[1,2], Martha Lien[1,2], and Trond Mannseth[1,2]

[1] Department of Mathematics, University of Bergen, Johs. Brunsgt. 12, N-5008 Bergen, Norway, Inga.Berre@math.uib.no, Martha.Lien@uni.no, Trond.Mannseth@uni.no

[2] CIPR – Centre for Integrated Petroleum Research, University of Bergen, Realfagbygget, Allégaten 41, N-5007 Bergen, Norway

Summary. In the last decade, the use of level-set functions has gained increasing popularity in solving inverse problems involving the identification of a piecewise constant function. Normally, a fine-scale representation of the level-set functions is used, yielding a high number of degrees of freedom in the estimation. In contrast, we focus on reparameterization of the level-set functions on a coarse scale. The number of coefficients in the discretized function is then reduced, providing necessary regularization for solving ill-posed problems. A coarse representation is also advantageous to reduce the computational work in solving the estimation problem.

1 Level-Set Representation of a Piecewise Constant Parameter Function

The identification of a piecewise constant parameter function is an inverse problem arising in various applications. Examples are image segmentation and identification of electric or fluid conductivity in reservoirs. Three features of the function can potentially be unknown: the number of regions of different constant value, the geometry of the regions, and the constant values of the parameter function.

For representing piecewise constant functions in a manner suitable for identification, the level set approach [6] has become popular as it provides a flexible tool to represent region boundaries. The first approach related to inverse problems is due to Santosa [7]. Recent reviews are given by Tai and Chan [8], Burger and Osher [2], and Dorn and Lesselier [3]. The main idea is that the boundary between two regions can be represented implicitly as the zero level set of a function – the level-set function. If we consider a domain Ω consisting of two regions Ω_1 and Ω_2, the level-set function is defined to have the following properties:

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*, Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_118,

$$\begin{aligned}\phi(\mathbf{x}) &> 0 \quad \text{for } \mathbf{x} \in \Omega_1;\\ \phi(\mathbf{x}) &< 0 \quad \text{for } \mathbf{x} \in \Omega_2;\\ \phi(\mathbf{x}) &= 0 \quad \text{for } \mathbf{x} \in \partial\Omega_1 \cap \partial\Omega_2.\end{aligned}$$

Hence, the boundary between the regions Ω_1 and Ω_2 is incorporated as the zero level-set of $\phi(\mathbf{x})$. A parameter function $p(\mathbf{x})$ taking different constant values c_1 in Ω_1 and c_2 in Ω_2 can now be written as

$$p(\mathbf{x}) = c_1 H(\phi(\mathbf{x})) + c_2[1 - H(\phi(\mathbf{x}))],$$

where H is the Heaviside function. Commonly, the level-set functions are initialized as signed distance functions. For representation of a partitioning with more than two regions, Vese and Chan [9] provide an extension of the above idea, which enables representation of up to 2^l regions with l level-set functions.

2 Coarse-Scale Level-Set Representation

Usually, level-set functions are represented with one degree of freedom for each grid cell of the computational grid. We apply an approach based on a coarse-scale representation, where each level-set function is reparameterized by a few coefficients only. This approach has several advantages: the reduced number of coefficients makes sensitivity computations less demanding; the need for regularization is diminished since the possible variations in the region boundaries are more limited; and, we can achieve convergence in a low number of iterations. The latter is particularly important in solving inverse problems requiring computationally demanding forward computations and sensitivity calculations. A prime example is the inverse problem of fluid conductivity estimation for oil reservoirs [1, 5].

A reparameterization of the level-set function is written

$$\phi(\mathbf{x}) = \sum_{i=1}^{n} a_i \theta_i(\mathbf{x}),$$

where $\{\theta_i\}$ denotes the set of basis functions and $\{a_i\}$ denotes the set of coefficients in the discretization. A coarse representation of the level-set functions enables fast identification of coarse-scale features of the parameter function with a low number of estimated coefficients. In addition, the approach provides regularization as the boundaries are confined to certain shapes based on the chosen representation. For fine-scale solutions of inverse problems, the coarse representation can enhance convergence by serving as a preconditioner for more fine-scale updates. Another strategy is to apply a sequential estimation, where the detail in the representation of the level-set functions is successively refined. The number of degrees of freedom in the estimation is then gradually

increased, and we can achieve estimates more in correspondence with the information content of the data.

When it comes to the choice of basis in the representation of the level-set functions, various alternatives are possible. However, once a set of basis functions is chosen, the reparameterized functions are restricted with respect to which shapes they can take. In the following, we discuss two different choices: a piecewise constant representation and a continuous representation.

A piecewise constant representation of the level-set functions [1, 5] is well suited in combination with adaptive multiscale estimation [4], which provides a rough identification of the number and geometry of the regions of different constant value for the parameter function. In Berre et al. [1], reparameterization by characteristic basis functions in combination with a narrow-band approach is proposed. The representation of the region boundaries is gradually refined during estimation. In Lien et al. [5], a sequential approach is developed, where the number of regions of constant parameter value is sought found as part of the estimation. Through successive estimations with increasing resolution in the representation of $p(\mathbf{x})$ identification of rather general parameter functions can be achieved.

The choice of characteristic basis functions yields piecewise constant level-set functions. This leads to updates of the boundaries between the constant states of $p(\mathbf{x})$, where the jumps are determined by the spatial support of the basis functions for $\phi(\mathbf{x})$, regardless of the resolution of the computational grid for the forward problem. The two plots to the left in Fig. 1 illustrate this situation in 1-D with $n = 2$, supp $\boldsymbol{\theta}_1 = [0, 1/2)$, and supp $\boldsymbol{\theta}_2 = [1/2, 1]$.

A continuous representation of the level-set functions enables more gradual updates of the region boundaries, depending on the resolution of the computational grid for the forward problem. Hence, the number of possible updates of the region boundaries is greatly enhanced. The two plots to the right in Fig. 1 illustrate this situation in 1-D with $n = 2$ and $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ being the standard linear basis on $[0, 1]$.

As a simple example of a continuous reparameterization of $\phi(\mathbf{x})$ with a low number of coefficients in 2-D, we consider a bilinear basis. The computational domain is partitioned into a coarse quadrilateral grid of rectangular elements. A bilinear reparameterization of a level-set function on a reference element $D_r = [0, 1] \times [0, 1]$ is given by

$$\phi(x_1, x_2) = a_1(x_1 - 1)(x_2 - 1) - a_2(x_1 - 1)x_2 - a_3x_1(x_2 - 1) + a_4x_1x_2.$$

By increasing the resolution of the quadrilateral grid, the boundaries can have more complex shapes.

A numerical example illustrating the potential of applying a coarse-scale reparameterization of the level-set functions is given below; see Figs. 2 and 3. We consider the inverse problem of fluid conductivity estimation for an oil reservoir based on a level-set representation with bilinear basis functions. The forward problem describes horizontal, two-phase, immiscible, and incompressible fluid flow in porous media. The available data for the inversion consist

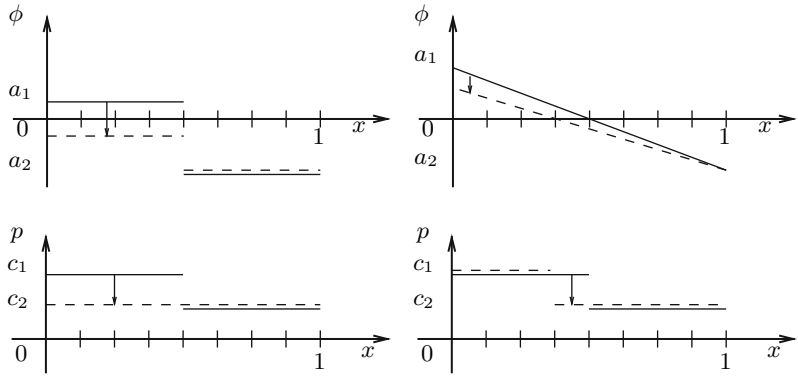


Fig. 1. Coarse representation of ϕ with two coefficients, a_1 and a_2, by characteristic (*top left*) and linear (*top right*) basis functions. Optimizing with respect to a_1 cause ϕ to change sign in some area. The resulting updates of the parameter function are shown in the *lower figures*. The initial states are drawn by *solid lines* and the final states by *dashed lines*

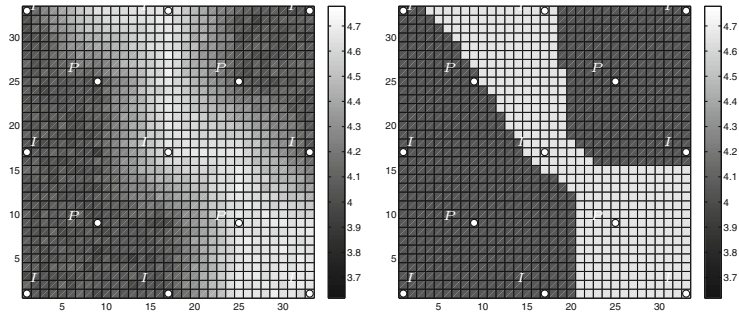


Fig. 2. Reference fluid conductivity and final estimation result

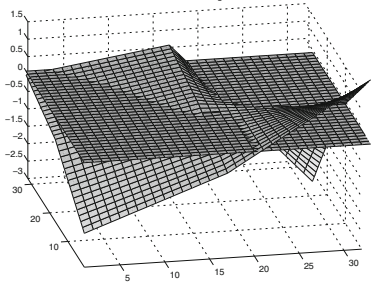


Fig. 3. Level-set function at convergence together with the surface $z = 0$

of time series of pressure data $\mathbf{d}$ logged in the injection wells. There are nine injection (I) and four production (P) wells; hence, the data are very sparsely distributed. The data are obtained from a forward simulation with the reference fluid conductivity, where normally distributed errors are added to the calculated pressures.

To illustrate the regularizing effect of the coarse level-set representation, we minimize a weighted least squares objective function with no additional regularizing terms:

$$J(p) = [\mathbf{m}(p) - \mathbf{d}]^T \mathbf{C}^{-1} [\mathbf{m}(p) - \mathbf{d}].$$

Here $\mathbf{m}(p)$ denote the calculated pressures given a parameter function $p(\mathbf{x})$, and $\mathbf{C}$ denotes the (diagonal) covariance matrix of the measurement errors in the data.

The reference fluid conductivity describes a channel with areas of low conductivity on both sides where the coarse-scale features are contaminated with fine-scale conductivity variation; see Fig. 2 (left). The initial guess for $p(\mathbf{x})$ was a constant value for the whole reservoir. The estimation was started with a coarse level-set representation on a grid of only one element before we continued with a quadrilateral grid of four elements. With our coarse level-set approach, we are able to recover the main trends in the conductivity distribution. The data are not reconciled by the coarse estimate though the objective function is greatly reduced. Hence, in this case, the coarse-scale estimate can serve as a good starting point for more fine-scale estimations.

References

1. Berre, I., Lien, M., Mannseth, T.: A level-set corrector to an adaptive multiscale permeability prediction. Comput. Geosci. **11**(1), 27–42 (2007)
2. Burger, M., Osher, S.: A survey on level set methods for inverse problems and optimal design. Eur. J. Appl. Math. **16**(2), 263–301 (2005)
3. Dorn, O., Lesselier, D.: Level set methods for inverse scattering. Inverse Probl. **22**, 67–131 (2006)
4. Grimstad, A.-A., Mannseth, T., Nævdal, G., Urkedal, H.: Adaptive multiscale permeability estimation. Comput. Geosci. **7**(1), 1–25 (2003)
5. Lien, M., Berre, I., Mannseth, T.: Combined adaptive multiscale and level-set parameter estimation. Multiscale Model. Simul. **4**(4), 1349–1372 (2005)
6. Osher, S., Sethian, J.A.: Fronts propagating with curvature-dependent speed: algorithms based on Hamilton-Jacobi formulations. J. Comput. Phys. **79**(1), 12–49 (1988)
7. Santosa, F.: A level-set approach for inverse problems involving obstacles. ESAIM Contrôle Optim. Calc. Var. **1**, 17–33 (1995/96)
8. Tai, X.-C., Chan, T.F.: A survey on multiple level set methods with applications for identifying piecewise constant functions. Int. J. Numer. Anal. Model. **1**(1), 25–47 (2004)
9. Vese, L.A., Chan, T.F.: A multiphase level set framework for image segmentation via the Mumford and Shah model. Int. J. Comput. Vision. **50**(3), 271–293(23) (2002)

On the Trajectory of Rockets in the Atmosphere

L.M.B.C. Campos and P.J.S. Gil

Centro de Ciências e Tecnologias Aeronáuticas e Espaciais (CCTAE),
luis.campos@ist.utl.pt
Instituto Superior Técnico, 1049-001 Lisboa, Portugal, p.gil@dem.ist.utl.pt

Summary. The trajectory of rockets in the atmosphere, e.g. a multi-stage rocket launcher, are specified by ordinary differential equations which involve (1) the rocket mass, varying with time due to the burning of the propellant; (2) the aerodynamic forces, (lift and drag) which are non-linear functions of velocity; and (3) the aerodynamic forces are proportional to the atmospheric mass density which varies by several orders of magnitude from the ground level to the near vacuum of space. Thus the differential equations of rocket trajectories in the atmosphere are non-linear with variable coefficients depending on time and altitude. Three methods of calculating trajectories are presented, which can be combined to determine the maximum orbital capability of a rocket, in terms of payload and velocity.

1 Introduction

The calculation of rocket trajectories in the atmosphere is considered taking into account the following effects: (1) the time variable mass, associated with propellant consumption; (2) the effect of thrust vectoring at an angle to the rocket axis; (3) the lift and drag for flight at an angle-of-attack; (4) the proportionality of the aerodynamic forces on the square of the velocity and on the atmospheric mass density; and (5) the dependence of the atmospheric mass density on altitude. This combination of effects has not been considered in some of the literature on rocket trajectories (cf. [1–8]). The equations of motion in an earth-fixed Cartesian frame, are solved for a gravity turn by three methods: direct expansion of the coordinates as a Taylor series of time (Sect. 2); use of the atmospheric mass density as direct (Sect. 3) or inverse (Sect. 4) altitude variable. The combination of there methods can be used to address both single and two point boundary value problem (Sect. 5).

2 Iterative Use of Initial Conditions in Taylor Series

The equations of motion of a rocket take the form:

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*,
Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_119,

$$[m_0 - c(t-t_0)]\begin{bmatrix}\ddot{x}\\ \ddot{z}+g\end{bmatrix} = |\dot{x}^2+\dot{z}^2|^{-1/2}\begin{bmatrix}T_{11} & T_{12}\\ T_{21} & T_{22}\end{bmatrix}\begin{bmatrix}\dot{x}\\ \dot{z}\end{bmatrix} - \tfrac{1}{2}S\rho_0\exp[(z-z_0)/\ell]|\dot{x}^2+\dot{z}^2|^{1/2}\begin{bmatrix}F_{11} & F_{12}\\ F_{21} & F_{22}\end{bmatrix}\begin{bmatrix}\dot{x}\\ \dot{z}\end{bmatrix} \quad (1)$$

in an earth-fixed reference frame where (a) the variation of mass with time, first equation in (2), corresponds to a constant fuel rate; (b) the acceleration of gravity is assumed to be uniform, and the altitude coordinate z is taken opposite to it; (c) the thrust matrix, first equation in (3) is involves the thrust T, angle-of-attack α and angle ϵ of the thrust with the rocket axis:

$$m(t) = m_0 - c(t-t_0) \qquad \rho(z) = \rho_0\exp[(z-z_0)/\ell], \quad (2)$$

$$T_{ij} = \begin{bmatrix}\cos(\alpha+\epsilon) & -\sin(\alpha+\epsilon)\\ \sin(\alpha+\epsilon) & \cos(\alpha+\epsilon)\end{bmatrix} \qquad F_{ij} = \begin{bmatrix}\cos\alpha & \sin\alpha\\ \sin\alpha & -\cos\alpha\end{bmatrix}\begin{bmatrix}C_D & C_L\\ C_L & -C_D\end{bmatrix}, \quad (3)$$

(d) the aerodynamic matrix, second equation in (3) involves the angle-of-attack α and lift C_L and drag C_D coefficients; (e) the aerodynamic forces are proportional to the square of the velocity and to the atmospheric mass density; and (f) the latter decays exponentially on the scale height ℓ for an isothermal atmosphere (second equation in (2)).

The equations of motion (1) and (2) are to be solved subject to initial conditions specifying the coordinates and velocity components (4) at initial time

$$t = t_0: \qquad \{x(t_0), z(t_0)\} \equiv \{x_0, z_0\}, \quad \{\dot{x}(t_0), \dot{z}(t_0)\} \equiv \{u_0, w_0\}. \quad (4)$$

Instead of $\{u_0, w_0\}$ in (4), the initial speed v_0 and flight path angle γ_0 could be specified:

$$v_0 = \sqrt{u_0^2+w_0^2}, \qquad \tan\gamma_0 = w_0/u_0; \quad (5)$$

the flight path angle γ_0 is indeterminate for zero initial velocity $v_0 = 0$ (5), whereas the initial conditions, second equation in (4) are always determinate. Substituting the initial conditions (4) in the equations of motion (1) specifies

$$m_0 v_0\begin{bmatrix}\ddot{x}_0\\ \ddot{z}_0+g\end{bmatrix} = \begin{bmatrix}T_{11} & T_{12}\\ T_{21} & T_{22}\end{bmatrix}\begin{bmatrix}u_0\\ w_0\end{bmatrix} - \tfrac{1}{2}\rho_0 S v_0^2\begin{bmatrix}F_{11} & F_{12}\\ F_{21} & F_{22}\end{bmatrix}\begin{bmatrix}u_0\\ w_0\end{bmatrix}, \quad (6)$$

the initial accelerations $(\ddot{x}_0, \ddot{z}_0)$.

Differentiating the equations of motion (1) and (2) with regard to time and setting $t = t_0$ specifies the $O(t^3)$ coefficients

$$m_0\begin{bmatrix}\dddot{x}_0\\ \dddot{z}_0\end{bmatrix} = c\begin{bmatrix}\ddot{x}_0\\ \ddot{z}_0+g\end{bmatrix} + v_0^{-3}(w_0\ddot{x}-u_0\ddot{z})\begin{bmatrix}T_{11} & T_{12}\\ T_{21} & T_{22}\end{bmatrix}\begin{bmatrix}w_0\\ -u_0\end{bmatrix} - \tfrac{1}{2}S\rho_0 v_0^{-1}\begin{bmatrix}F_{11} & F_{12}\\ F_{21} & F_{22}\end{bmatrix}\begin{bmatrix}2u_0^2\ddot{x}_0 + w_0^2\ddot{x}_0 + u_0w_0\ddot{z}_0 - v_0^2u_0w_0/\ell\\ 2w_0^2\ddot{z}_0 + u_0^2\ddot{z}_0 + u_0w_0\ddot{x}_0 - v_0^2w_0^2/\ell\end{bmatrix}, \quad (7)$$

in the Taylor series expansion

$$\begin{bmatrix}x(t)\\ z(t)\end{bmatrix} = \begin{bmatrix}x_0\\ z_0\end{bmatrix} + \theta\begin{bmatrix}u_0\\ w_0\end{bmatrix} + \frac{\theta^2}{2}\begin{bmatrix}\ddot{x}_0\\ \ddot{z}_0\end{bmatrix} + \frac{\theta^3}{6}\begin{bmatrix}\dddot{x}_0\\ \dddot{z}_0\end{bmatrix} + O(\theta^4), \quad (8)$$

of coordinates of the trajectory versus time, with $\theta \equiv t - t_0$. Thus cubic approximation can be applied step-by-step along the trajectory, allowing the thrust $T(t)$ and its angle with the rocket axis $\epsilon(t)$ and angle-of-attack $\alpha(t)$ to vary at each step, as well as the scale height $\ell(z)$ in the mass density.

3 Mass Fraction of Burned Fuel as the Time Variable

The equations of motion (1) may be put in a dimensionless form (9) using as dimensionless time variable τ, the mass of burned fuel as a fraction of lift-off mass, and making both altitude z and downrange distance x dimensionless, respectively ζ and χ, dividing by the atmospheric scale height, viz.:

$$\tau \equiv c(t - t_0)/m_0, \quad \zeta \equiv (z - z_0)/\ell, \quad \chi \equiv (x - x_0)/\ell : \tag{9}$$

$$(1 - \tau)\begin{bmatrix} \chi'' \\ \zeta'' + a \end{bmatrix} = \frac{1}{\sqrt{\chi'^2 + \zeta'^2}} [b_{ij}] \begin{bmatrix} \chi' \\ \zeta' \end{bmatrix} - e^{-\zeta}\sqrt{\chi'^2 + \zeta'^2}\,[f_{ij}] \begin{bmatrix} \chi' \\ \zeta' \end{bmatrix}, \tag{10}$$

where $[x_{ij}] \equiv \left[\begin{smallmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{smallmatrix}\right], x_{ij} = \{b_{ij}, f_{ij}\}$, and: (1) the gravity parameter a and (2) the thrust parameter b appears in the matrix b_{ij} (11).

$$a \equiv \frac{m_0}{c^2 \ell}, \quad b \equiv \frac{m_0 T}{c^2 \ell}, \quad b_{ij} \equiv \frac{m_0}{c^2 \ell} T_{ij} = b \begin{bmatrix} \cos(\alpha + \epsilon) & -\sin(\alpha + \epsilon) \\ \sin(\alpha + \epsilon) & \cos(\alpha + \epsilon) \end{bmatrix} \tag{11}$$

(3) the drag parameter f appears in the aerodynamic matrix F_{ij} (12):

$$m_* = \rho_0 S \ell, \; f \equiv \frac{c_0 \rho_0 S \ell}{2 m_0}, \; f_{ij} = \frac{\rho_0 S \ell}{2 m_0} F_{ij} = \frac{m_*}{2 m_0} \begin{bmatrix} \cos\alpha & \sin\alpha \\ \sin\alpha & -\cos\alpha \end{bmatrix} \begin{bmatrix} C_D & C_L \\ C_L & -C_D \end{bmatrix}, \tag{12}$$

involving both the drag C_D and lift C_L coefficients, and the reference mass m_*. Using as dimensionless altitude variable η the atmospheric mass density at the initial altitude divided by that at arbitrary altitude, the exponential non-linearity in the equations of motion (10) is replaced by algebraic non-linearities of degree up to three in (13):

$$\begin{aligned} \eta \equiv e^{\zeta} = \rho(0)/\rho(z) : \quad & (1 - \tau)\eta \begin{bmatrix} \chi'' \eta \\ \eta'' \eta^2 + a\eta^2 \end{bmatrix} \\ = & \left(\eta'^2 + \eta^2 \chi'^2\right)^{-1/2} \eta^3 [b_{ij}] \begin{bmatrix} \chi' \\ \eta' \end{bmatrix} - \left(\chi'^2 \eta^2 + \eta'^2\right)^{1/2} [f_{ij}] \begin{bmatrix} \chi' \\ \eta' \end{bmatrix}. \end{aligned} \tag{13}$$

Note that if $x - x_0 = \text{const}$, or $\chi' = 0$, then (13) reduces to the case of a vertical climb. The solution of (13) is taken in (14) for down-range and altitude, respectively η and χ:

$$v_* \equiv \frac{c\ell}{m_0} : \quad \eta(\tau) = \frac{w_0}{v_*}\tau + \sum_{n=2}^{\infty} Z_n \tau^n, \quad \chi(\tau) = \frac{u_0}{v_*}\tau + \sum_{n=2}^{\infty} X_n \tau^n, \tag{14}$$

since (a) the initial position in (4) corresponds to $\chi = 0$ in (9) and $\eta = 1$ in (13) and (b) the initial velocities in (4) are divided by v_* defined in (14)

in the coefficient of τ, as shown in for $\dot{z}$ and likewise for $\dot{x}$. The remaining coefficients (X_n, Z_n) in (14) for $n = 2, 3, \ldots$ are obtained by substitution in the equation of motion (13) and equating to zero the coefficients of successive powers $\tau^2, \tau^3 \ldots$

This process is illustrated by substituting the trajectory (14) in the equation of motion (13) up to order τ^3, viz. the first and second equations becomes

$$(1-\tau)(1+w_0\tau/v_*)^2(2X_2+6X_3\tau)$$
$$= \left\{G^{-1/2}(1+w_0\tau/v_*)^3\left[b_{11}\ b_{12}\right] - G^{1/2}\left[f_{11}\ f_{12}\right]\right\}\begin{bmatrix} u_0/v_*+2X_2\tau \\ w_0/v_*+2Z_2\tau\end{bmatrix}, \quad (15)$$

$$(1-\tau)(1+w_0\tau/v_*)^2(2Z_2+6Z_3\tau)$$
$$-(1+w_0\tau/v_*)(w_0/v_*+2Z_2\tau)^2 + a(1+w_0\tau/v_*)^3$$
$$= \left\{G^{-1/2}(1+w_0\tau/v_*)^3\left[b_{21}\ b_{22}\right] - G^{1/2}\left[f_{21}\ f_{22}\right]\right\}\begin{bmatrix} u_0/v_*+2X_2\tau \\ w_0/v_*+2Z_2\tau\end{bmatrix}, \quad (16)$$

where

$$G \equiv (\eta')^2 + \eta^2(\chi')^2 = (w_0/v_*+2Z_2\tau)^2 + (1+w_0\tau/v_*)^2(u_0/v_*+2X_2\tau)^2$$
$$= (v_0/v_*)^2 + 2\left[(w_0/v_*)(u_0/v_*)^2 + (X_2u_0+Z_2w_0)/v_*\right]\tau \quad (17)$$

using v_0 from (5). Equating powers of τ^0 in (15) and (16) yields:

$$2X_2 = b_{12} + b_{11}u_0/w_0 - w_0(f_{11}u_0+f_{12}w_0)/v_*^2, \quad (18)$$
$$2Z_2 = -a + b_{22} + b_{21}u_0/w_0 - w_0[f_{21}u_0+(f_{22}-1)w_0]/v_*^2, \quad (19)$$

and equating powers of τ yields the next pair of coefficients (omitted for brevity). The solution (14) converges best for small τ, i.e. short time or mass of fuel burned a small fraction of initial mass.

4 Atmospheric Mass Densities as Altitude Variable

A faster convergence close to burn-out conditions is obtained by (a) using the residual mass fraction μ in (20) as dimensionless independent variable instead of time; (b) replacing altitude as dependent variable by ζ from (9) the ratio of atmospheric densities at arbitrary and initial altitudes ξ in (20); and (c) using as the other dependent variable the same downrange distance normalized to the scale height χ. The equation of motion (10) becomes:

$$\mu = -c(t-t_0)/m_0, \quad \xi = e^{-\zeta} = \rho(z)/\rho_0 : \qquad \mu\begin{bmatrix}\chi'' \chi \\ -\xi''\xi + \xi'^2 + a\xi^2\end{bmatrix}$$
$$= \left(\xi'^2 + \xi^2\chi'^2\right)^{-1/2}\left[b_{ij}\right]\begin{bmatrix}\chi' \\ \eta'\end{bmatrix} - \xi\left(\xi'^2+\xi^2+\chi'^2\right)^{1/2}\left[f_{ij}\right]\begin{bmatrix}\chi' \\ \eta'\end{bmatrix}. \quad (20)$$

The solution is sought as power series similar to (14) viz.:

$$\chi(\mu) = \sum_{n=0}^{\infty} P_n \mu^{n+p}, \qquad \xi(\mu) = \sum_{n=0}^{\infty} Q_n \mu^{n+q}, \tag{21}$$

where the initial conditions $t = t_0$ imply (a) for the coordinates in (4), then $\mu = 1$ and $\xi = 1$ in the first and second equations of (20), respectively, and $\chi = 0$ in the third equation of (9), viz.:

$$0 = \sum_{n=0}^{\infty} P_n, \qquad 0 = \sum_{n=0}^{\infty} Q_n; \tag{22}$$

(b) for the initial velocities (16), then $\mu = 1$ in the first equation of (20), implies $\chi' = u_0 m_0 / c_0 \ell = u_0 / v_*$ in and also $\xi' = w_0 / v_*$, viz:

$$u_0 = v_* \sum_{n=0}^{\infty} P_n (n+p). \qquad w_o = v_* \sum_{n=0}^{\infty} Q_n (n+q). \tag{23}$$

If the coefficients (P_n, Q_n) are known for $n = 0, 1, \ldots, N$, then (22) and (23) can be used to determine them for $n = N+1, N+2$.

The trajectory is specified by (21), where the coefficients (P_n, Q_n) and the exponents (p, q) are to be determined. The exponents (p, q) can be determined from the lowest powers in the trajectory (21), viz. χ and ξ in (24):

$$\chi \sim P_0 \mu^p, \qquad \xi \sim Q_0 \mu^q, \qquad H \equiv \xi'^2 + \xi^2 \chi'^2, \tag{24}$$

which imply that H in (24) is given to lowest order by (25)

$$p \geq 1: \qquad H = (Q_0 \mu^{q-1})^2 \left[1 + (p P_0 \mu^{p-1})^2\right] \sim (Q_0 \mu^{q-1})^2 \tag{25}$$

if $p \geq 1$. Taking also the lowest powers in (20) leads to

$$\begin{bmatrix} P_0 Q_0 p(p-1) \mu^{q+p-2} \\ Q_0^2 \mu^{2q-2} (q + a\mu^3) \end{bmatrix} = Q_0 \left\{ \mu^{q+1} \left[b_{ij}\right] - Q_0 \mu^{2q-2} \left[f_{ij}\right] \right\} \begin{bmatrix} P_0 p \mu^{p-1} \\ Q_0 q \mu^{q-1} \end{bmatrix}; \tag{26}$$

on the l.h.s. the term μ^{2q+1} is of higher order than μ^{2q-1} and can be omitted, and on the r.h.s. μ^{q+1} is of higher order than μ^{2q-2} for $q+1 > 2q-2$ or $q \leq 2$, and can also be omitted, simplifying (26) to

$$\begin{bmatrix} P_0 p(p-1) \mu^{q+p-2} \\ Q_0 q \mu^{2q-2} \end{bmatrix} = -Q_0 \mu^{2q-2} \left[f_{ij}\right] \begin{bmatrix} P_0 p \mu^{p-1} \\ Q_0 q \mu^{q-1} \end{bmatrix}. \tag{27}$$

This can be satisfied by choosing the exponents $p = q = 1$

$$p = 1 = q: \qquad \begin{bmatrix} 0 \\ Q_0 \end{bmatrix} = -Q_0 \left[f_{ij}\right] \begin{bmatrix} P_0 \\ Q_0 \end{bmatrix}, \tag{28}$$

leading to the system of equations (28), which can be solved

$$\{f_{12}, f_{11}\} = \{P_0, -Q_0\}(f_{11} f_{22} - f_{12} f_{21}), \tag{29}$$

to determine the lowest order coefficients in (21).

The leading terms of the trajectory (21) are determined by (29), viz.:

$$\chi(\mu) = \mu(P_0 + P_1\mu + \ldots), \qquad \xi(\mu) = \mu(Q_0 + Q_1\mu + \ldots) \tag{30}$$

and substituting in the equation of motion (20) up to order μ^2, specifies the next coefficients, viz.

$$\mu \begin{bmatrix} 2P_1Q_0\mu \\ -2Q_0Q_1\mu + Q_0^2 \end{bmatrix} = \frac{(Q_0\mu)^2}{\sqrt{H}} \left[b_{ij}\right] \begin{bmatrix} P_0 \\ Q_0 \end{bmatrix} - \sqrt{H}\mu(Q_0 + Q_1\mu) \left[f_{ij}\right] \begin{bmatrix} P_0 + 2P_1\mu \\ Q_0 + 2Q_1\mu \end{bmatrix}, \tag{31}$$

which involves (24), calculated to order μ

$$H \equiv (Q_0 + 2Q_1\mu)^2 + (Q_0P_1\mu)^2 = Q_0(Q_0 + 4Q_1\mu) \tag{32}$$

Equating to zero the coefficients of μ in (20), the system (29) is regained, and equating to zero the coefficients of μ^2 leads to the system of equations:

$$\begin{bmatrix} 2P_1 \\ -2Q_1 \end{bmatrix} = \left[b_{ij}\right] \begin{bmatrix} P_0 \\ Q_0 \end{bmatrix} - 2Q_0 \left[f_{ij}\right] \begin{bmatrix} P_1 \\ Q_1 \end{bmatrix} - (Q_0 + 3Q_1) \left[f_{ij}\right] \begin{bmatrix} P_0 \\ Q_0 \end{bmatrix}, \tag{33}$$

which determines the coefficients (P_1, Q_1). The next two pairs of coefficients follow from (22) and (23), viz.:

$$P_0 + P_1 + P_2 + P_3 = 0 = Q_0 + Q_1 + Q_2 + Q_3, \tag{34}$$

$$\{u_0, w_0\}/v_* = \{P_0, Q_0\} + 2\{P_1, Q_1\} + 3\{P_2, Q_2\} + 4\{P_3, Q_3\}; \tag{35}$$

this specifies the trajectory (21) up to order four:

$$\left\{(x - x_0)/\ell, e^{[-(z-z_0)/\ell]}\right\} = \sum_{n=0}^{\infty} \{P_n, Q_n\}[1 - c(t - t_0)/m_0]^{n+1} \tag{36}$$

with exponents $p = q = 1$, and coefficients (29) and (33)–(35).

5 Discussion

All of the preceding three methods can be used to show the effect on a rocket trajectory of a non-zero angle-of-attack and non-zero thrust angle; this specifies the burn-out altitude and speed for a powered first stage, and the downrange and peak altitude for a ballistic stage; they can be combined for a multi-stage rockets. The methods also apply to (1) the single-point boundary-value problem (SPBVP) of finding the trajectory of a rocket given the initial conditions and (2) the two-point boundary-value problem (TPBVP) of achieving a given orbital velocity from a given launch condition. The TPBVP is the most important, but may have no solution, e.g. if the rocket does not have the required orbital capability. The three methods indicated can be used to solve the TPBVP and to determine the maximum orbital capability of a rocket. The methods of calculation of rocket trajectory are usually numerical, but may use analytical solutions for initialization.

References

1. Connor, M.A.V.: J. Spacecraft **3**, 1308–1311 (1966)
2. Culler, G.J., Fried, B.D.: J. Appl. Phys. **28**, 672–676 (1957)
3. Di Sotto, E., Teofilatto, P.: J. Guid. Control Dyn. **25**, 538–545 (2002)
4. Miele, A.: Flight Mechanics, 2 vols. Addison-Wesley, Reading, MA (1961)
5. Miele, A.: J. Opt. Theor. Appl. **2**, 260–273 (1968)
6. Miele, A., Pritchard, R.E., Damoulakis, J.N.: J. Opt. Theor. Appl. **5**, 235–283 (1970)
7. Rutherford, D.E.: Classical Mechanics, 2nd edn. Oliver & Boyd, Edinburgh (1967)
8. Thomson, W.T.: Space Dynamics, 2nd edn. Dover, New York (1986)

On Aircraft Response and Control During a Wake Encounter

L.M.B.C. Campos[1,2] and J.M.G.Marques[2,3]

[1] Instituto Superior Técnico (IST), 1049-001 Lisboa, Portugal
[2] Centro de Ciências e Tecnologias Aeronaúticas e Espaciais (CCTAE), 1049-001 Lisboa, Portugal, luis.campos@ist.utl.pt
[3] Universidade Lusófona de Humanidades e Tecnologias (ULHT), 1749-024 Lisboa, Portugal, jmgmarques@ist.utl.pt

Summary. In this paper the exact analytical solution of the airplane response to a wake encounter appears as a power series of a damping factor, whose coefficients are exponential integrals of time. It is shown that in the absence of control action, the roll response tends to an asymptotic bank angle.

1 Introduction

The separation between aircraft due to wake effects determines aircraft spacing at take-off and landing, and hence runway and airport capacity. This is the motivation for the current research on the topic [1], which can be grouped under four broad headings: (a) measurement, calculation and prediction of the evolution of wake vortices in the atmosphere, including wind and ground effects [2]; (b) determination of the effect of a wake vortex encounter in terms of aircraft response, control, load factor and other dynamical aspects [3]; (c) identification of possible measures to reduce or mitigate the effect of the wake vortex encounter [4]; and (d) classification of the consequences of the wake vortex encounter into hazard classes [5]. The response to wakes is affected by the application of controls and damping effects, which are considered in the present paper, as an extension of an earlier analytical theory [6].

The roll motion of the airplane will consist of (a) a free response to an arbitrary initial condition as regards bank angle and roll rate; (b) plus a forced response to the wake encounter and aileron deflection. Both responses will be affected by the aerodynamic damping, especially for longer times. This is illustrated for an example of identical leading and following aircraft of the same type (Boeing 757), considering the wake vortex effects alone and in combination with aileron control deflection.

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*,
Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_120,

2 Roll Equation with Damping and Controls

In this model the roll dynamics equation, with one degree-of-freedom [6] in the form:

$$\ddot{\phi} - \frac{1}{2}\frac{\rho U_2 S_2}{W_2/g}\left(\frac{b_2}{r_2}\right)^2 C_{\dot{\phi}}\,\dot{\phi} = \\ = \frac{\rho S_2 b_2}{W_2/g}\left(\frac{U_2}{r_2}\right)^2 C_\delta\,\delta(t) - \frac{2h}{1+\lambda}\frac{C_{L_\alpha}}{2\pi}\frac{W_1}{W_2}\frac{S_2}{S_1}\frac{U_2}{U_1}\left(\frac{a}{r_2}\right)^2 \frac{c_{r_1} g}{\eta t}\exp\left(-\frac{a^2}{2\eta t}\right), \tag{1}$$

where ρ is the atmospheric mass density and g the acceleration of gravity. The weight W, velocity U and wing area S have subscript "1" for the leading aircraft (W_1, U_1, S_1) that creates the wake impinging on the following aircraft, for which the subscript "2" is used (W_2, U_2, S_2). The following aircraft has wing span b_2, taper ratio λ (ratio of tip to root chord), and radius of gyration r_2 in roll; the leading aircraft has wing root chord c_{r1}. Its trailing wind tip vortices have core radius a, and decay with time t with total kinematic viscosity η. For the following aircraft (with subscript "2" omitted) $C_{L\alpha}$ is the slope of the lift coefficient, C_δ the aileron rolling moment coefficient and the corresponding damping coefficient; the dimensionless encounter parameter h defined in [6] is of order unity.

The case of an aileron control law which compensates the wake vortex encounter is the only situation in which there is no aircraft roll response, because the forced response to the ailerons (b) exactly balances the response to the wake vortex (c), leaving only the free response (a), which is zero if there are no initial perturbation. The three terms of the response (a,b,c) are calculated next in turn, starting with the free response $\phi_1(t)$, which is the solution of the roll equation (1) without forcing terms on the r.h.s., viz.:

$$\ddot{\phi}_1 + \bar{\mu}\dot{\phi}_1 = 0, \tag{2a}$$

where the overall damping coefficient is specified by:

$$\bar{\mu} \equiv -\frac{1}{2}\frac{\rho U_2 S_2}{W_2/g}\left(\frac{b_2}{r_2}\right)^2 C_{\dot{\phi}}, \tag{2b}$$

and the damping time by $1/\bar{\mu}$. The damping increases with (a) the ratio of span to gyration radius squared; (b) the roll damping coefficient $C_{\dot{\phi}}$; and (c) the air density (lower altitude), airspeed and wing area divided by the mass.

It follows that the free response is given by:

$$\phi_1(t) = \phi_0 + (\dot{\phi}_0/\bar{\mu})[1 - e^{-\bar{\mu}t}], \tag{3}$$

for arbitrary initial bank angle ϕ_0 and roll rate $\dot{\phi}_0$.

The forced response to the ailerons $\phi_2(t)$ is even simpler, since it is a particular solution of the roll dynamics equation (1), omitting the last term on the r.h.s. side representing wake vortex effects:

$$\ddot{\phi}_2 + \bar{\mu}\dot{\phi}_2 = \nu, \tag{4}$$

where the aileron deflection was taken to be maximum:

$$\nu \equiv \frac{\rho S_2 b_2}{W_2/g} \left(\frac{U_2}{r_2} \right)^2 C_\delta \, \delta_{\max}. \tag{5}$$

The forcing term increases with (a) the air density times span and wing area (which is the mass of a parallelepiped of air, with base area equal to the wing area and height equal to the span) divided by the aircraft mass, specifying a relative density; (b) the square of airspeed divided by the radius of gyration; (c) the aileron rolling moment coefficient; and (d) the aileron deflection taken at maximum value for fastest response. The forced response to constant aileron deflection is a bank angle varying linearly with time:

$$\phi_2(t) = \frac{\nu}{\bar{\mu}} t \equiv -2 \frac{U_2}{b_2} \frac{C_\delta}{C_{\dot{\phi}}} \delta_{\max} t. \tag{6}$$

3 Response Forced by Wake Encounter

Taking into account the dependence of the induced rolling moment on time leads to a less simple response $\phi_3(t)$, specified by a particular integral of the roll dynamics equation (1), without the first term on the r.h.s.:

$$\ddot{\phi}_3 + \bar{\mu} \dot{\phi}_3 = -\bar{\xi} t^{-1} \exp(-a^2/2\eta t), \tag{7a}$$

where the vortex wake effect is specified by:

$$\bar{\xi} \equiv \frac{2h}{1+\lambda} \frac{C_{L_\alpha}}{2\pi} \frac{W_1/S_1}{W_2/S_2} \frac{U_2}{U_1} \left(\frac{a}{r_2} \right)^2 \frac{c_{r_1} g}{\eta}, \tag{7b}$$

and increases for (a) larger encounter factor h and smaller taper ratio λ; (b) larger ratio of wing loading of the leading aircraft to the wing loading of the following aircraft; (c) larger ratio airspeed of following aircraft (catches wake sooner) to the airspeed of the leading aircraft (leaves stronger wake for lower airspeed); (d) larger square ratio of vortex core radius to radius of gyration; (e) larger root chord of leading aircraft; and (f) smaller viscosity leading to slower vortex decay.

It is convenient to introduce a dimensionless time divided by the time of peak vorticity [6]:

$$\tau \equiv t/t_* = 2\eta t/a^2, \qquad \Phi(\tau) = \phi_3(t), \tag{8a,b}$$

so that the roll response forced by the wake vortex satisfies (7a) in the form:

$$\ddot{\Phi} + \mu \dot{\Phi} = -(\xi/\tau) \; e^{-1/\tau}, \tag{9}$$

where the dimensionless aerodynamic damping and vortex effect are given respectively by:

$$\mu \equiv \frac{\bar{\mu} a^2}{2\eta} = -\frac{1}{4} \left(\frac{b_2}{r_2} \right)^2 \frac{\rho S_2 a}{W_2/g} \frac{U_2 a}{\eta} C_{\dot{\phi}}, \tag{10a}$$

$$\xi \equiv \frac{\bar{\xi} a^2}{2\eta} = \frac{2h}{1+\lambda} \frac{C_{L_\alpha}}{2\pi} \frac{W_1/S_1}{W_2/S_2} \frac{U_2}{U_1} \left(\frac{a}{r_2} \right)^2 \frac{c_{r_1} a^2 g}{2\eta^2}. \tag{10b}$$

The forced solution of (9) is found by the method of variation of parameters, i.e. as the free solution (3) with non-constant coefficients, leads the forced response:

$$\Phi(\tau) = -\xi \int e^{-\mu\tau} d\tau \int \tau^{-1} e^{-1/\tau} e^{\mu\tau} d\tau, \tag{11}$$

to the wake vortex.

4 Time Evolution of the Forced Response

The total roll response is the sum of the free response with the forced responses to the ailerons and the wake vortex:

$$\phi(t) = \phi_1(t) + \phi_2(t) + \phi_3(t). \tag{12}$$

Assuming that the initial bank angle and the roll rate are zero there is no free response $\phi_1(t) = 0$, and the total forced response

$$\phi_0 = 0 = \dot{\phi}_0 : \qquad \phi(t) = \phi_2(t) + \Phi(2\eta t/a^2) \tag{13}$$

consists of (a) the response to the ailerons and (b) the response to the wake vortex, which in the absence of damping is expressed in terms of the exponential integral [7]:

$$T = 1/\tau : \quad E_n(T) \equiv \int_T^\infty T^{-1-n} e^{-T} dT = \int_0^{1/\tau} \tau^{n-1} e^{-1/\tau} d\tau = E_n(1/\tau), \tag{14}$$

of order zero $n = 0$ by (11) with $\mu = 0$:

$$\mu = 0 : \quad \Phi_0(\tau) = -\xi \int d\tau \int d\tau \; e^{-1/\tau} \tau^{-1} = -\xi \int E_0(1/\tau) \, d\tau, \tag{15}$$

in agreement with [6].

Since the dimensionless roll rate in the absence of damping is specified by an exponential integral of order zero:

$$-\xi^{-1} \dot{\Phi}_0(\tau) = E_0(1/\tau) = \int \tau^{-1} e^{-1/\tau} d\tau, \tag{16}$$

the comparison with the dimensionless roll rate in the presence of damping (11):

$$-\xi^{-1}\dot{\Phi}(\tau) = e^{-\mu\tau}\int \tau^{-1}e^{-1/\tau}e^{\mu\tau}d\tau, \tag{17}$$

leads to the solution:

$$\dot{\Phi}(\tau) = -\xi e^{-\mu\tau}\left\{E_0(1/\tau) + \sum_{n=1}^{\infty}\frac{\mu^n}{n!}E_n(1/\tau)\right\}, \tag{18}$$

which specifies the dimensionless roll response:

$$\Phi(\tau) = -\xi\sum_{n=0}^{\infty}\frac{\mu^n}{n!}\int e^{-\mu\tau}E_n(1/\tau)\ d\tau, \tag{19}$$

as a series of powers of the damping, with exponential integrals of order n as coefficients (14). If the damping is weak, only the first terms of the series are needed, e.g. the first two for $\mu^2 \ll 1$.

5 Numerical Results

As an example, the case of two Boeing 757 flying one behind the other is considered. The data needed is available from open sources [8]. Table 1 has shown the bank angle response for the same aircraft, replacing the actual roll damping $\bar{\mu} = 0.45$, by larger and smaller hypothetical values, up to more than the double. Table 2 shows the sum of the first $N+1$ terms of the series, viz.:

$$\phi_3^N(t) \equiv \Phi^N(\tau) = -\xi\sum_{n=0}^{N}\frac{\mu^n}{n!}\int e^{-\mu\tau}E_n(1/\tau)\ d\tau, \tag{20}$$

for several values of N.

Table 1. Effect of increasing or decreasing roll damping

Roll damping	Asymptotic bank angle (°)	Time to within 1% (s)
μ	ϕ_∞	$1-\phi_3/\phi_\infty < 0.01$
0.25	50.13	$t \geq 5.71$
0.375	21.7	$t \geq 3.95$
0.5	14.89	$t \geq 3.34$
0.75	5.24	$t \geq 2.08$
1	2.99	$t \geq 1.66$

Table 2. Asymptotic bank angle reached during wake encounter

Number of terms of series	Peak bank angle (°)	Relative error
N+1	$\phi_\infty^N \equiv \phi_3^{(n)}(\infty)$	$1 - \phi_\infty^N/\phi_\infty$
1	12.7237	1.76×10^{-2}
2	14.6418	1.74×10^{-3}
3	14.8657	1.45×10^{-4}
4	14.8875	1.61×10^{-5}
5	14.8893	8.04×10^{-6}
∞	14.8895	0

6 Conclusion and Further Work

The first two terms $N = 1$ of the series solution (20) to $\mathcal{O}(\mu)$ gives an error of less than 2% in the bank angle response. This can be confirmed from Table 2, which indicates the asymptotic bank angle $\phi_\infty^N \equiv \phi_3^{(N)}(\infty)$ calculated with $N + 1$ terms of the series (20), and shows rapid convergence for $N \geq 2$.

The value $\phi_\infty = 14.89°$ is above the threshold $\phi_{max} = 10°$ for which airline procedures regarding passenger comfort require a go-round on approach to land. Since the asymptotic bank angle does not include the effect of aileron deflection, it is clear that roll control could be used to keep the bank angle below the threshold, which will be the object of future work.

References

1. Schwarz, C.W., Klaus-Uwe, H.: Full-Flight Simulator study for wake vortex hazard area investigation. Aerosp. Sci. Technol. **10**, 136–143 (2006)
2. Hallock, J.N., Burnham, D.C.: Decay characteristics of wake vortex from jet transport aircraft. In: 35th Aerospace Sciences Meeting & Exhibit, Reno, NV, 1997
3. Gerz, T., Holzpfel, F., Darracq, D.: Commercial aircraft wake vortices. Progr. Aerosp. Sci. **38**, 181–208 (2002)
4. McMillan, O.J., Nielsen, J.N., Schwind, R.G., Dillenius, M.F.: Rolling moments in a trailing-vortex flow field. AIAA J. Aircraft **15**, 280–286 (1978)
5. Hohne, G., Fuhrmann, M., Luckner, R.: Critical wake vortex encounter scenarios. Aerosp. Sci. Technol. **8**, 689–701 (2004)
6. Campos, L.M.B.C., Marques, J.M.G.: On wake vortex response for all combinations of five classes of aircraft. Aeronaut. J. **108**, 295–310 (2004)
7. Abramowitz, M., Stegun, I.: Tables of Mathematical Functions. Dever, New York (1965)
8. Jackson, P.: Jane's All-the-World's Aircraft 2006–2007. MacDonald and Jane's, London (2006)

On Alternative Safety Metrics for the Probability of the Collision Between Aircraft

L.M.B.C. Campos[1,2] and J.M.G. Marques[2,3]

[1] Instituto Superior Técnico (IST), 1049-001 Lisboa, Portugal
[2] Centro de Ciências e Tecnologias Aeronaúticas e Espaciais (CCTAE), 1049-001 Lisboa, Portugal, luis.campos@ist.utl.pt
[3] Universidade Lusófona de Humanidades e Tecnologias (ULHT), 1749-024 Lisboa, Portugal, jmgmarques@ist.utl.pt

Summary. This paper examines several safety metrics: (1) the one-dimensional cumulative probability of coincidence P_1 along the line joining the two aircraft; (2) the three-dimensional probability of coincidence P_3 over all space; and (3) a two dimensional probability of coincidence, defined in the present paper P_2.

1 Introduction

The safety of air traffic is based on separation rules [1], and when they fail to be observed, on conflict resolution measures [2] to avoid a collision. The chosen separation can be reduced to increase air traffic capacity [3] if the risk of collision remains below the threshold set by International Civil Aviation Organization (ICAO) Target Level of Safety (TLS), which is an example of a safety metric [4]. The pioneering work on the calculation of collision probabilities [5] is based on the penetration of a safety volume around an aircraft, by another aircraft. It can be shown that in the particular but important case of air corridors, the probability of coincidence is an upper bound for the probability of collision [6].

The value of the probability of collision depends on the statistical distribution of aircraft position errors [7]. It can be argued that the Gaussian is suited to frequent events like small flight path deviations; collisions are due to large deviations [5,7], which are rare events [8] modeled by Laplacian or generalized error distributions [9].

The simplified model of collision probability used makes six assumptions. The first assumption is that the three dimensional position error is decomposed into horizontal along track and across track errors and vertical error; in this way only a one-dimensional collision problem needs to be considered at a time. The second assumption is that the aircraft are treated as a mass points located at centers of mass. It can be shown [6] that the aircraft size

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*,
Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_121,

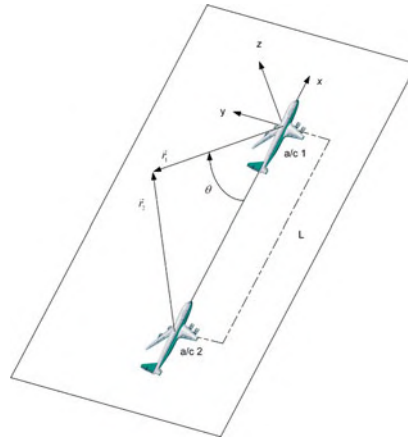


Fig. 1. Aircraft flying along the same flight path at a minimum separation distance

affects collision probability if it is comparable to the r.m.s. position error. The third assumption is that the aircraft are assumed to move in unbounded space. The fourth assumption is that the position errors are specified by a Gaussian probability distribution. It is known from flight data records and radar tracks that the Gaussian underestimates the probability of large flight path deviations and collision [3], and whereas the Laplace distribution is an improvement [5]; a more accurate representation is provided by the generalized error distribution [9]. These probability distributions can be extended to a combined Gamma and generalized error distribution [10] to model the whole range of flight path deviations from small to large [11]. These distributions can be introduced [12] as a correction to the Gaussian. The fifth assumption, allows the calculation of the probabilities of collision as a function of position if aircraft dynamics do not appear explicitly; since aircraft dynamics would limit the possible displacements, the probabilities of collision calculated in this way are upper bounds. The sixth assumption concerns the geometry considered.

In the present paper two simple but important cases are considered, both with aircraft flying at the same speed, on: (case I) the same flight path at a given distance; (case II) on parallel flight paths at given distance.

2 Comparison of Probabilities of Coincidence in Several Dimensions

The calculation of maximum coincidence probabilities is made for the case of two aircraft flying on the same straight flight path at a distance L. A coincidence occurs if the first aircraft deviates by $\mathbf{r}_1$ and the second by $\mathbf{r}_2$ such that $\mathbf{r}_2 = \mathbf{r}_1 + L\mathbf{e}_x$ in Fig. 1. The coincidence occurs on a plane through the flight path, and thus it is possible to introduce polar coordinates in this

plane, with origin at aircraft one, and axis along the flight path:

$$\mathbf{r}_1 = (-r\cos\theta, r\sin\theta)\,, \qquad \mathbf{r}_2 = (L - r\cos\theta, r\sin\theta)\,. \qquad (1a,b)$$

The probability distributions are initially assumed to be Gaussian:

$$P(\mathbf{r}_i) = \left[1/\left(\sigma_1\sqrt{2\pi}\right)\right] \exp\left[-\left(\|\mathbf{r}_i\|/\sigma_i\right)^2/2\right],\ i = 1,2 \qquad (2)$$

with r.m.s. position errors respectively σ_1 and σ_2, which may or may not be equal. Exactly the same formulas (1,2) will apply (case II) for two aircraft on parallel flight paths at a distance L, using polar coordinates with origin on the first aircraft and axis perpendicular to the flight paths as shown in Fig. 2. Assuming that the position errors are statically independent, the probability of coincidence is the product of (2):

$$P(r,\theta) = P(\mathbf{r}_1)\,P(\mathbf{r}_2)\,, \qquad (3)$$

and depends only on (r,θ) in both cases I and II. From (3) can be defined a one-dimensional cumulative probability of coincidence, by integrating along the polar axis $\theta = 0$:

$$\bar{P} \equiv \int_{-\infty}^{+\infty} P(r,0)\,dr \equiv P_1, \qquad (4a)$$

viz.: (case I) the integration is along the flight path; (case II) the integration is along a line perpendicular to the flight paths. Substitution of (3) into (4a) leads [6] to the one-dimensional probability of collision:

$$P_1 = \left[1/\left(2\bar{\sigma}\sqrt{\pi}\right)\right]\ \exp\left\{-\left[L/\left(2\bar{\sigma}\right)\right]^2\right\}, \qquad (4b)$$

which involves the r.m.s. position error $\bar{\sigma}$ corresponding to:

$$\bar{\sigma} \equiv \sqrt{\left[\left(\sigma_1\right)^2 + \left(\sigma_2\right)^2\right]/2}. \qquad (4c)$$

The three-dimensional cumulative probability of coincidence involves an integration over all space in spherical coordinates:

$$\overline{\overline{P}} \equiv \int_0^{2\pi} d\varphi \int_0^{\pi} d\theta\,\sin\theta \int_0^{\infty} dr\,r^2\,P(r,\theta) \equiv P_3, \qquad (5a)$$

and leads via a broadly similar integration to:

$$P_3 = \frac{\sqrt{\pi}}{2}\frac{\bar{\sigma}}{f^2}\ \exp\left\{-\left(\frac{L}{2\bar{\sigma}}\right)^2\right\}, \qquad (5b)$$

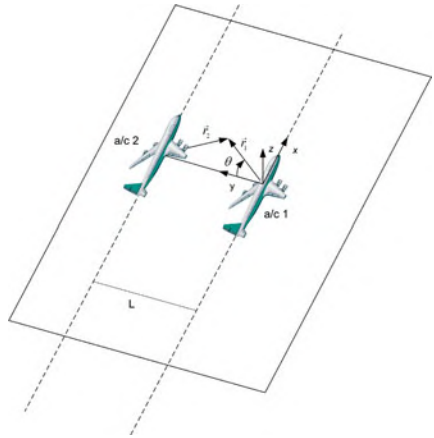


Fig. 2. Aircraft flying on parallel flight paths at the minimum lateral distance

involving $\bar{\sigma}$ in (4c) and the aircraft dissimilarity function:

$$f \equiv (\sigma_1/\sigma_2 + \sigma_2/\sigma_1)/2. \tag{5c}$$

The only case of dimensionless probability of coincidence is the two-dimensional case:

$$P_2 \equiv \int_0^{2\pi} d\theta \int_0^{\infty} dr\, r\, P(r,\theta) = \frac{2}{f} \exp\left\{-\left[L/(2\bar{\sigma})\right]^2\right\}. \tag{6}$$

which may be interpreted as the collision probability integrated over: (case I) a plane passing through the trajectory; (case II) a plane perpendicular to the trajectories passing through both aircraft.

3 Comparison with the ICAO TLS

The ICAO TLS specifies a probability of collision $S_1 = 5\times10^{-9}$ per hour flown, which can be converted in probability of collision per nautical mile S_1/V by dividing by the speed V in knots. Thus the ICAO TLS is directly comparable to the one-dimensional cumulative probability of coincidence $P_1 V \leq S_1$. To apply the three-dimensional cumulative probability of coincidence $P_3/V \leq S_3$ would need the introduction of another safety metric or modified ICAO TLS with the dimensions of hour flown.

Taking as reference case aircraft with identical r.m.s. position errors, the ratio to the minimum separation distance which typically meets [6] the current ICAO TLS, is indicated in the first line of Table 1. Taking in this table the geometric mean of (a) the least strict condition for dissimilar aircraft $\left(1.26\times10^{-12}\right)$ and (b) the intermediate condition for similar aircraft

Table 1. Two-dimensional probability of coincidence

$L/\bar{\sigma}$	10	11	12
P_2	6.95×10^{-12}	3.64×10^{-14}	1.16×10^{-16}
P_2/f	1.26×10^{-12}	6.62×10^{-15}	2.11×10^{-17}

Table 2. One-dimensional probability of coincidence

m	1	2	3	4
L_m	60 nm	5 nm	1,000 ft	2,000 ft
σ_m	5.61 nm	0.468 nm	93.6 ft	187 ft
P_{1m} (per nm)	2.01×10^{-14}	2.41×10^{-13}	7.33×10^{-12}	3.66×10^{-12}
$V_m < S_1/P_{1m}$	2.48×10^5 kt	2.07×10^4 kt	6.82×10^2 kt	1.36×10^3 kt

$\left(3.64 \times 10^{-14}\right)$ leads to $\sqrt{1.26 \times 10^{-12} \times 3.64 \times 10^{-14}} = 2.14 \times 10^{-13}$, which suggests:

$$P_2 \leq S_2 = 2 \times 10^{-13}, \tag{7}$$

as the alternative absolute ICAO TLS, which will be checked next.

In order to asses the implications of this choice of absolute safety standard, it is applied to the following four typical Air Traffic Management (ATM) cases: (a) lateral separation in transoceanic airspace $L_1 = 60$ nm; (b) lateral separation in controlled airspace $L_2 = 5$ nm; (c) Reduced Vertical Separation Minima (RVSM) $L_3 = 1{,}000$ ft in controlled airspace at lower flight levels (above FL 290); and (d) vertical separation $L_4 = 2{,}000$ ft elsewhere. For these four values L_m with $m = 1, 2, 3, 4$, the proposed absolute TLS (8) corresponds by (7) to $L_m/\sigma_m = 10.7$ and thus to a r.m.s. position error σ_m indicated in Table 2 together with the one-dimensional probability of coincidence (4b) per nautical mile, which satisfies the ICAO TLS:

$$S_1 = 5 \times 10^{-9} \text{ per hour} \tag{8}$$

for airspeeds up to V_m. Since V_m exceeds the speed capability of all current subsonic airliners, the absolute alternative ICAO TLS (8) is safe in all these conditions. The values of V_m suggest that the absolute ICAO TLS (7) is stricter than the original ICAO TLS (8) in the four cases considered. This is the price to be paid for having an ICAO TLS which is absolute, i.e. applies to all separation conditions, not just the four examples given.

The suggested alternative TLS is dimensionless for all probability distribution of aircraft deviations. The preceding examples using the Gaussian distribution can be extended to other distributions in future work.

4 Discussion

In conclusion the ICAO TLS of $S_1 = 5 \times 10^{-9}$ per hour is comparable to the one-dimensional probability coincidence $P_1 V \leq S_1$. For the maximum

probability of coincidence $P_0V^2 \leq S_0$ a modified TLS $S_0 = 5 \times 10^{-9}$ per hour squared is needed, if the same value is chosen. For the three-dimensional probability of coincidence P_3/V another modified TLS $S_3 = 5 \times 10^{-9}$ times hour would be needed. Of course, the value of S_0 and S_3 need not be numerically equal to S_1. Since the two-dimensional probability of coincidence $P_2 \leq S_2$ is dimensionless, the modified TLS $S_2 = 5 \times 10^{-9}$ would also be dimensionless. The numerical value of S_2 need not equal S_1 or S_0 or S_3. The procedure indicated has lead to a value (7) of S_2 consistent with S_1, justifying the following reasoning (a) the original ICAO TLS (8) has been applied to three of the most common ATM traffic situations and (b) for these situations it is comparable to the absolute level of safety (7). The latter is preferable to the former, because it is dimensionless, and thus independent of flight time or speed. Thus the Absolute Level of Safety (ALS) in (7) can be proposed as a more general dimensionless substitute to the original ICAO TLS in (8).

References

1. Rules of the Air and Air Traffic Services, 13th edn. International Civil Aviation Organization, Montreal, Canada (1996)
2. Tomlin, C., Pappas, J., Sastry, S.: Conflict resolution in air traffic management: a study in multi-agent hybrid systems. IEEE Trans. Autom. Control **43**, 509–521 (1998)
3. Anderson, E.W.: Principles of Navigation. Hollis and Carter, London (1966)
4. Kinnersly, S.R.: Objective measures of ATM system safety: safety metrics. In: CARE-INTEGRA Report to Eurocontrol (2000)
5. Reich, P.G.: Analysis of long-range air traffic systems: separation standards. J. Inst. Navig. **19**, 88–98, 169–186, 331–347 (1966)
6. Campos, L.M.B.C., Marques, J.M.G.: On safety metrics related to aircraft separation. J. Navig. **55**, 39–63 (2002)
7. European Studies of Vertical Separation above FL 290. Summary Report, Eurocontrol 88/20/10 (1988)
8. Reiss, R.D., Thomas, M.: Statistical Analysis of Extreme Values. Birkhauser, Basel (2001)
9. Johnson, N.L., Balakrishnan, N.: Continuous Univariate Probability Distributions. Wiley, New York (1995)
10. Campos, L.M.B.C., Marques, J.M.G.: On the combination of the Gamma and generelized error distribution with application to aircraft flight path deviations. Commun. Stat. **33**, 2307–2332 (2004)
11. Ballin, M.G., Wing, D.J., Hughes, M.F., Conway, S.R.: Airborne separation assurance and air traffic management: research of concepts and technology. AIAA Pap. 99-3989 (1999)
12. Campos, L.M.B.C., Marques, J.M.G.: On the probability of collision between climbing and descending aircraft. AIAA J. Aircraft **44**, 550–557 (2007)

Homogeneous Branched-Chain Explosions

M. Carretero[1], L.L. Bonilla[1], and J.B. Keller[2]

[1] G. Millán Institute, Fluid Dynamics, Nanoscience & Industrial Mathematics, Universidad Carlos III, 28911 Leganés, Spain, manuel.carretero@uc3m.es, bonilla@ing.uc3m.es

[2] Department of Mathematics, Stanford University, Stanford, CA 94305, USA keller@math.stanford.edu

Summary. A model of homogeneous explosions due to Kapila is analyzed by singular perturbation methods. The results are compared with those obtained by the method of self-adjusting multiple scales.

1 Introduction

A distinctive characteristic of combustion processes is the self-accelerating nature of their associated chemical reactions. In chain-branching processes [6], the accelerating factor is the autocatalytic character of the chain-branching reactions. Chain-branching explosions are observed for instance in hydrogen-oxygen mixtures when the initial temperature is above the so-called crossover temperature. Branched-chain explosions are examples of problems termed jump phenomena, that are characterized by large amplitude dynamic responses to small amplitude disturbances and typically involve different time scales: the system may evolve slowly during long time intervals which are separated by fast transition layers during which the system changes abruptly [3,5]. In [2], we introduced a method of self-adjusting time scales to describe homogeneous branched-chain explosions, whose main ingredient is a fast time scale which is a nonlinear function of one of the system variables. This method is not standard in that it requires two different solvability conditions depending on whether time is smaller or larger than the very large induction time. An approximate solution valid for all times was obtained by patching two different approximations at the induction time. Here we present an overview of an alternative singular perturbation method detailed in [1]. This method is based on an exact relation between the fuel concentration and a slowly varying combination of fuel and radicals. Therefore, it is possible to approximate the solutions of the explosion problem before and after the induction time, and match them to find an uniform approximation.

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*, Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_122,

2 Governing Equations

2.1 Two-Step Chemistry Model

A model of homogeneous explosions with competing branching and recombination processes due to Kapila [4] is based in the following two-step chemical reaction scheme:

$$\begin{aligned} \mathrm{F} + \mathrm{X} &\xrightarrow{\mathrm{B}} 2\mathrm{X} \quad \text{branching reaction,} \\ \mathrm{X} &\xrightarrow{\mathrm{R}} \mathrm{P} \quad \text{recombination reaction.} \end{aligned}$$

Where F, X and P are the reactant, radical and product of the chemistry description, respectively. The previous scheme does not have an initiation reaction, therefore we assume that there is a small amount of radical X from the beginning. All the heat is generated through the recombination reaction.

2.2 Non-Dimensional Equations

Manipulation of the kinetic rate equations for a homogeneous branched-chain explosion at constant pressure, similar to that presented in [4], leads to the following dimensionless problem [1,2].

$$\frac{dx}{dt} = \exp\left[\frac{\beta\theta}{1+\theta}\right] x\, f - \epsilon\, x, \tag{1}$$

$$\frac{df}{dt} = -\exp\left[\frac{\beta\theta}{1+\theta}\right] x\, f, \tag{2}$$

$$\frac{d\theta}{dt} = q\,\epsilon\, x, \tag{3}$$

to be solved with the initial conditions

$$x(0) = \nu, \quad f(0) = 1, \quad \theta(0) = 0. \tag{4}$$

where $f(t)$, $x(t)$ are the normalized concentration of fuel and of radical, respectively and θ is the temperature at time t. The nondimensional parameters satisfy, in realistic applications, $q = O(1)$, $\beta = O(1)$, and $0 < \nu \ll \epsilon \ll 1$, see details in [2].

3 Solution by Singular Perturbation Methods

Introducing the variable $y = x + f$ and the result of a linear combination of (1)–(4) that gives $\theta = q(1 + \nu - y)$, the problem reduces to:

$$\frac{df}{dt} = -(y - f)\, f\, e^{a(y)}, \tag{5}$$

$$\frac{dy}{dt} = -\epsilon\,(y - f), \tag{6}$$

$$a(y) = \frac{\beta q(1 + \nu - y)}{1 + q(1 + \nu - y)} \tag{7}$$

with $f(0) = 1, \quad y(0) = 1 + \nu$. The previous equations lead to the relation

$$f(y) = \exp\left(-\frac{1}{\epsilon}\int_y^{1+\nu} e^{a(y)}ds\right) \tag{8}$$

Using (8) in (6), we obtain the following integro-differential equation for $y(t)$:

$$\frac{dy}{dt} = -\epsilon\left[y - \exp\left(-\frac{1}{\epsilon}\int_y^{1+\nu} e^{a(y)}ds\right)\right]. \tag{9}$$

Its solution with $y(0) = 1 + \nu$ is given by

$$\epsilon\, t = -\int_y^{1+\nu} \frac{ds}{s - \exp\left(-\frac{1}{\epsilon}\int_s^{1+\nu} e^{a(r)}dr\right)}. \tag{10}$$

$y(t)$ can be obtained evaluating numerically the integral in (10), however, the asymptotic forms of y, both for t large, and for ϵ small and t finite, can be determined analytically.

3.1 The Outer Expansion

The outer expansion (see [1]) is given by $y(t) \sim (1+v)e^{-\epsilon(t-t_o)}, t \gg 1, \epsilon \ll 1$, where t_o is the induction time defined by

$$\epsilon t_o = -\int_{y_\infty}^{1+\nu}\left[\frac{1}{s - f(s)} - \frac{(s - y_\infty)^{-1}}{1 - f'(y_\infty)}\right]ds, \tag{11}$$

and the stationary state $y_\infty = \exp\left(-\frac{1}{\epsilon}\int_{y_\infty}^{1+\nu} e^{a(s)}ds\right)$ is reached when $t \to \infty$. The integral (11) for t_o can be evaluated for ϵ small:

$$t_o \sim t_{o,0} = \frac{\ln \nu^{-1}}{1+\nu}, \quad \epsilon \ll 1. \tag{12}$$

3.2 The Inner Expansion

For ϵ small and $t = O(1)$, the inner expansion (see [1]) is

$$y(t) = 1 + \nu + \epsilon \ln\left[\frac{1+\nu}{1 + \nu e^{(1+\nu)t}}\right] + O(\epsilon^2). \tag{13}$$

3.3 Matching and the Composite Expansion

For $t \gg 1$, the inner expansion becomes $y(t) \sim 1+\nu-\epsilon(1+\nu)(t-t_o), \epsilon \ll 1$. For $\epsilon(t-t_o) \ll 1$ the outer expansion becomes $y(t) \sim 1+\nu-\epsilon(1+\nu)(t-t_o), \quad 0 \leq (t-t_o) \ll 1$. Thus the composite expansion for $\epsilon \ll 1$ is

$$y_c(t) \sim 1+\nu-\epsilon \ln\left(\frac{1+e^{(1+\nu)(t-t_o)}}{1+\nu}\right) + H(t-t_o)\left[(1+\nu)e^{-\epsilon(t-t_o)} - \{1+\nu-\epsilon(1+\nu)(t-t_o)\}\right]. \quad (14)$$

We have used the Heaviside function H because the outer expansion only works for $\epsilon(t-t_o) > 0$. Finally the approximations for f, x and θ are

$$f_c(t) = \exp\left(-\frac{1}{\epsilon}\int_{y_c(t)}^{1+\nu} e^{a(s)}ds\right), \; x_c(t) = y_c(t) - f_c(t), \; \theta_c(t) = q[1+\nu-y_c(t)]. \quad (15)$$

4 Results

4.1 Results with the Approximation of the Induction Time

As can be seen in Fig. 1 the approximate solutions given by (14)–(15) capture rather well the behavior of the solution of the model equations (1)–(4). The differences are due to the fact that the induction time given by (12) is not a good approximation to the real value.

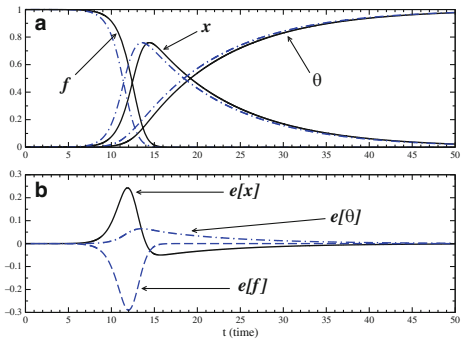


Fig. 1. (**a**) Time evolution of x, f, and θ, for $\epsilon = 0.1$, $\nu = 10^{-5}$, $\beta = 5$, $q = 1$, obtained by numerical integration of (1)–(4) (*solid lines*), and using (15) (*dot-dashed lines*) with the induction time given by (12). (**b**) Errors

4.2 Improving the Induction Time

Our analysis points up that the accuracy of the approximations (14)–(15) are deeply related with the value of the induction time. Figure 2 shows that we obtain more satisfactory results because we calculate a better value of t_o integrating numerically the integral (11). The differences are now of order ϵ^2.

4.3 Comparison with the Method of Self-Adjusting Multiple Scales

With the ideas of the our singular perturbation method we have improved the method of self-adjusting multiple scales by matching two terms of the expansion valid before the induction time to one term of the expansion for later times (see details in [1]). Of course, the quality of the composite expansion depends on the accuracy with which we calculate the induction time. Figure 3 shows the good agreement of the approximations with f, x and θ. It is very similar to

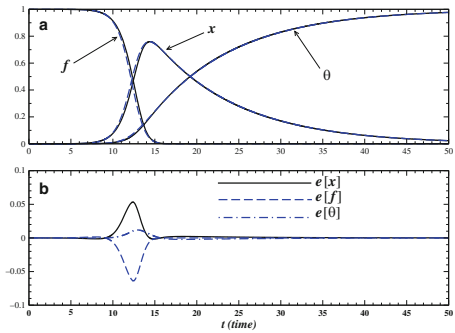


Fig. 2. (**a**) Same as in Fig. 1 but now the induction time is given by numerical evaluation of (11). (**b**) Errors

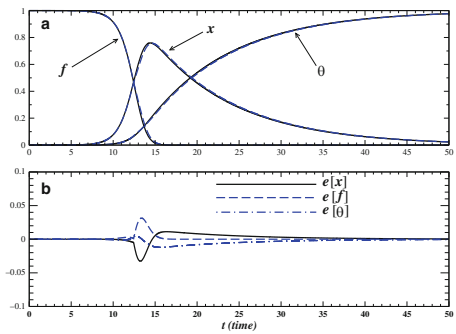


Fig. 3. (**a**) Same as in Fig. 1 but now the approximations (*dashed lines*) are obtained improving the method of self-adjusting time scales with the ideas from the boundary layer method. (**b**) Errors

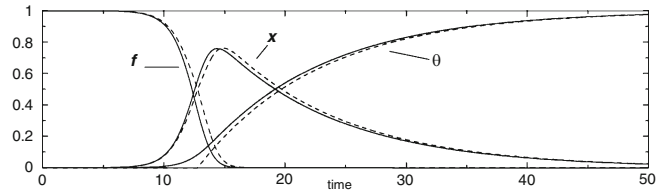


Fig. 4. Same as in Fig. 1 but now the *dashed lines* are the leading-order approximations given by the method of self-adjusting time scales, see [2]

the results shown in Fig. 2. If we compare Fig. 4 with Fig. 1, we observe, that to leading order, the method of self-adjusting scales gives approximated x and f that are somewhat better than f_c and x_c, but this method uses patching of two different asymptotic expansions at the induction time. This patching implies that the approximation of θ is a worse approximation than θ_c.

5 Conclusions and Further Work

We have found a composite of two matched asymptotic expansions and the approximations for the radical, the fuel and the temperature providing very good agreement with the numerical solution. We have investigated the influence of the induction time to obtain better agreement with the numerical solution of the model. With the ideas of the boundary layer method we have improved the method of multiple self-adjusting time scales described in [2], finding another composite expansion, although the greater simplicity of the boundary layer method makes it preferable. Future work will try to apply our methods to other jump phenomena problems.

References

1. Bonilla, L.L., Carretero, M., Keller, J.B.: SIAM J. Appl. Math. **68**(3), 619–628 (2007)
2. Bonilla, L.L., Sánchez, A.L., Carretero, M., SIAM J. Appl. Math. **61**, 528–550 (2000)
3. Haberman, R.: SIAM J. Appl. Math. **37**, 69–106 (1979)
4. Kapila, A.K.: J. Eng. Math. **12**, 221–235 (1978)
5. Reiss, E.L.: SIAM J. Appl. Math. **39**, 440–455 (1980)
6. Semenov, N.N.: Some Problems in Chemical Kinetics and Reactivity, vol. 2. Princeton University Press, Princeton, NJ (1959)

Wind Simulation Refinement: Some New Challenges for Particle Methods

C. Chauvin[1], F. Bernardin[2], M. Bossy[1], and A. Rousseau[3]

[1] INRIA, Sophia Antipolis, France
Claire.Chauvin@inria.fr, Mireille.Bossy@inria.fr
[2] CETE de Lyon, LRPC, Clermont-Ferrand, France
Frederic.Bernardin@developpement-durable.gouv.fr
[3] INRIA, Laboratoire Jean Kuntzmann, Grenoble, France
Antoine.Rousseau@inria.fr

Summary. We present two new challenges related to the stochastic downscaling method (SDM) that we applied to wind simulation refinement in Bernardin et al. (Stoch. Environ. Res. Risk Assess. 23:851–859, 2009). After setting the framework, we introduce the boundary forcing issue, and propose a numerical scheme adapted to *Particle in Cell* methods. Then we turn to the uniform density constraint raised by SDM and propose some new methods that rely on optimization algorithms.

1 The Stochastic Downscaling Method

We are interested in the behaviour of an *incompressible fluid* in a domain $\mathcal{D}$ of $\mathbb{R}^3$; $\mathcal{D}$ is such that the mass density ρ is supposed constant. We decompose the unknown functions as the sum of a large-scale component and a turbulent one. Rather than solving the Reynolds Averaged Navier Stokes (RANS) equations on the mean velocity $\langle U\rangle$ and pressure $\langle\mathcal{P}\rangle$, we consider some stochastic differential equations (SDEs) that describe the stochastic dynamics of a fluid particle with state variables $(X_t,\, \mathcal{U}_t)_{t\geqslant 0}$:

$$dX_t = \mathcal{U}_t dt, \tag{1a}$$

$$\begin{aligned} d\mathcal{U}_t = &-\frac{1}{\rho}\nabla_x\langle\mathcal{P}\rangle(t,X_t)dt - \left(\frac{1}{2}+\frac{3}{4}C_0\right)\langle\omega\rangle(t,X_t)\left(\mathcal{U}_t - \langle U\rangle(t,X_t)\right)dt \\ &+ \sqrt{C_0\varepsilon(t,X_t)}dW_t \\ &- \sum_{0\leq s\leq t} 2\,\mathcal{U}_{s-}\,\mathbb{1}_{\{X_s\in\partial\mathcal{D}\}} + \sum_{0\leq s\leq t} 2\,V_{\text{ext}}(s,X_s)\,\mathbb{1}_{\{X_s\in\partial\mathcal{D}\}}, \end{aligned} \tag{1b}$$

where ε is the turbulent kinetic energy dissipation rate, $\langle\omega\rangle$ the turbulent frequency, and $(W_t)_{t\geq 0}$ is a three dimensional Brownian motion. The foundation

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*,
Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_123,

of such a model can be found in [1] and was inspired from [7]. The last two terms of (1b) model a Dirichlet condition (see [3]):

$$\langle U\rangle(t,x) = V_{\text{ext}}(t,x),\ \ x \in \partial\mathcal{D}, \tag{2}$$

V_{ext} denoting a known external velocity field (provided e.g. by measures, large scale simulations, or statistics). In the general RANS equations, $\langle\mathcal{P}\rangle$ is recovered thanks to the following Poisson equation:

$$-\frac{1}{\rho}\Delta_x\langle\mathcal{P}\rangle = \sum_{i,j=1}^{3}\left(\partial_{x_j}\langle U^{(i)}\rangle\ \partial_{x_i}\langle U^{(j)}\rangle + \partial^2_{x_ix_j}\langle u^{(i)}u^{(j)}\rangle\right), \tag{3}$$

which requires the knowledge of the second order moments of the velocity; this can be done thanks to turbulent closures, see [5] for a review of these models.

Assume that there exists a Lagrangian density f_L, such that at every time t the measure $f_L(t;x,V)dxdV$ is the law of the random process $(X_t,\mathcal{U}_t)$ solution of (1); a fluid particle satisfying (1) and (3) also satisfies (at least formerly), for almost all $x \in \mathcal{D}$

$$\int_{\mathbb{R}^3} f_L(t;x,V)\ dV = \rho, \tag{4a}$$

$$\nabla_x \cdot \langle U\rangle(t,x) = 0. \tag{4b}$$

The method that we define, called the *Stochastic Downscaling Method (SDM)*, is of a totally new type: its consists in simulating a solution of (1), (2), (4) with a given V_{ext}.

2 Numerical Description of SDM

2.1 The Stochastic Particle Method

The time is discretized with a sequence $t_k = k\Delta t, k = 0,\dots,K,\ \Delta t = T/K$. The stochastic dynamics is approximated at time t_k by the discrete random variables $(X_k^n,\mathcal{U}_k^n, 1 \le n \le N)$ associated to N fluid particles dropped inside $\mathcal{D}$. The statistics on these variables are defined using a local approximation, as in the *Particle in Cell* method (see [8]). More precisely, in the *Nearest Grid Point* method, a partition of $\mathcal{D}$ into N_c cells is defined: $\mathcal{D} = \cup_{i=1}^{N_c}\mathcal{C}_i$, associating N_i particles to each cell $\mathcal{C}_i$. A statistics $Q(U)$ is defined on each cell $\mathcal{C}_i$ by

$$\langle Q(U)\rangle_k(x) = \frac{1}{N_i}\sum_{n=1}^{N_i} Q\left(\mathcal{U}_k^n\right), \qquad x \in \mathcal{C}_i. \tag{5}$$

Notice that the method we construct is not hybrid. In particular, inside $\mathcal{D}$ the pressure gradient term $-\frac{1}{\rho}\nabla_x\langle\mathcal{P}\rangle$ is not computed by mean of a PDE solver.

Moreover, the computation of the right-hand-side of (3) is far too costly since it requires a very fine cell subdivision. Instead, we proceed to a fractional step algorithm inspired from Pope (see [7]): at each step, we solve (1b) without the term $-\frac{1}{\rho}\nabla_x\langle\mathcal{P}\rangle$. We simulate the pressure effect by solving the constraints (4), more adapted to a particle method [4].

2.2 Two New Numerical Challenges

In this paper, we focus our work on two issues: first, the confinement scheme required by (2). To the best of our knowledge, the case of (inhomogeneous) imposed boundary conditions in the framework of stochastic particle methods has not been formerly studied in the literature. Second, we focus on the transportation problem raised by (4a) (see [4] for some first studies in the SDM context).

Solving the Boundary Condition (2)

The external velocity V_{ext} is imposed at the boundaries of $\mathcal{D}$. The guidance is modelled by the two last terms of (1b). For robustness considerations (see [6]), we introduce the exponential version of the explicit Euler scheme. Hereafter, we sketch the main steps of the algorithm. After a prediction step, the dynamics of the outgoing particles is treated by the following reflection scheme:

At time t_k, for each particle n:

1. *Prediction.* Predict the position $\widetilde{X}_k^n = X_{k-1}^n + \Delta t\ \mathcal{U}_{k-1}^n$ and the velocity $\widetilde{\mathcal{U}}_k^n$ using an exponential scheme [10]:

$$d\mathcal{U}_t^n = -\left(\frac{1}{2} + \frac{3}{4}C_0\right)\langle\omega\rangle_{k-1}\left(\mathcal{U}_t^n - \langle U\rangle_{k-1}\right)dt + \sqrt{C_0\varepsilon_{k-1}}dW_t, \quad t \in [t_{k-1}, t_k], \tag{6}$$

   where $\langle U\rangle_{k-1}$, $\langle\omega\rangle_{k-1}$ and ε_{k-1} are evaluated in the cell containing X_{k-1}^n. If $\widetilde{X}_k^n \in \mathcal{D}$, then set $X_k^n = \widetilde{X}_k^n$ and $\mathcal{U}_k^n = \widetilde{\mathcal{U}}_k^n$.

2. *Reflection.* When $\widetilde{X}_k^n \notin \mathcal{D}$; let $\delta_{out} = \lambda\Delta t$ be the boundary hitting time, and $x_{out} = X_{k-1}^n + \delta_{out}\mathcal{U}_{k-1}^n$ be the hitting position, then the reflected position reads

$$X_k^n = x_{out} + (\Delta t - \delta_{out})\left(2V_{\text{ext}}(t_{k-1}, x_{out}) - \mathcal{U}_{k-1}^n\right). \tag{7}$$

   The reflected velocity is constructed by two successive steps. First, we simulate Equation (6) between t_{k-1} and t_{out^-} with an exponential scheme to obtain the velocity $\mathcal{U}_{t_{out^-}}$. Then, in order to match the boundary conditions, a *jump* is imposed to the velocity at $t = t_{out}$, leading to $\mathcal{U}_{t_{out^+}} = 2V_{\text{ext}}(t_{k-1}, x_{out}) - \mathcal{U}_{t_{out^-}}$. The second advancement is done between t_{out^+} and t_k.

3. *Kill-Build Procedure*. It may happen that the reflected position (7) does not belong to $\mathcal{D}$. In this case, the particle is *killed, and created* in a boundary cell with incoming velocity V_{ext}.

The linear equation (6) is exactly solved in step 1; the same holds for the two velocity advancements in step 2, knowing the boundary hitting time t_{out}, and the velocity jump.

Solving the Constant Mass Density Constraint (4a)

We come now to the second difficulty of this paper. The condition (4a) implies that the number of particles per cell has to be constant: for each cell C_i, $N_i = N_{pc}$, and thus the total number of particles is $N = N_c N_{pc}$. After steps 1–3 above, this condition may not hold anymore. Let us denote x_i the particle locations at the end of step 3. When $N_i < N_{pc}$, locations are randomly created in $\mathcal{C}_i$, and the set $\{q_j\}_{1\leq j\leq N}$ is constructed by taking N_{pc} particles per cell.

At this point, the constant mass density problem can be interpreted (at least formally) as an *optimal transport problem* (see [1, 4]): defining the cost $p_{ij} = \|x_i - q_j\|_{L_2}^2$ of transporting a particle from x_i to q_j, the problem consists of finding an element σ of the set of permutations $\mathcal{S}_N$ of $\{1, \ldots, N\}$ which minimizes the overall transport cost:

$$(P) \quad \text{Find } \sigma^* \in \mathcal{S}_N \text{ such that } D^* := \sum_{i=1}^{N} p_{i\sigma^*(i)} = \min_{\sigma\in\mathcal{S}_N} \sum_{i=1}^{N} p_{i\sigma(i)}. \tag{8}$$

This so-called *Assignment Problem* has been tackled by D. Bertsekas in [2], introducing the *Auction Algorithm*, where the optimality condition (8) is ε-relaxed:

$$D^* \leqslant \sum_{i=1}^{N} p_{i\sigma^*(i)} \leqslant D^* + N\varepsilon. \tag{9}$$

The overall cost of the final assignment is within $N\varepsilon$ of being optimal. Numerical tests (see [4]) have shown that in our specific configuration, the optimal solution is obtained when $\varepsilon \backsimeq \frac{C}{N}$, with a complexity of order N^2. Such a computational cost involves a very slow execution of SDM, since we need a large number of particles N for the Monte Carlo method to converge.

Hereafter, in the SDM framework, we present our strategies to reduce the number of objects involved in the Auction Algorithm.

3 Benchmarks

In order to decrease the number of particles involved in the Auction Algorithm, we consider the supernumerary particles and possibly a set of particles coming from tanks, defined in each cell. Let be the sets $\mathbb{X}$, containing the particles to be transported, and $\mathbb{Q}$, the final locations, constructed as follows:

Table 1. Comparison of several transportation algorithms: Auction Algorithm with Tank (AAT), for several tank sizes, and Triangular Transport (TT)

		Run time (s)	D	c_{max}	c_{move}
AAT	$\alpha = 1$	467,183	8.8	1.5	4,846
AAT	$\alpha = 0.01$	1,646	66	1.7	3,012
AAT	$\alpha = 0$	557	85	2.3	2,414
TT		0.38	110	1.4	9,843

Initialization: $\mathbb{X} = \mathbb{Q} = \varnothing$, *and the tank size* $N_{tank} = \alpha N_{pc} \in \mathbb{N}$, $0 \leq \alpha \leq 1$.
For all $\mathcal{C}_i$*:*

If $N_i > N_{pc}$*: add* $N_i - N_{pc}$ *particles to* $\mathbb{X}$*, and add* N_{tank} *other particles of* $\mathcal{C}_i$ *to* $\mathbb{X}$ *and* $\mathbb{Q}$*.*

If $N_i < N_{pc}$*: create* $N_{pc} - N_i$ *particles in* $\mathcal{C}_i$*, and add them to* $\mathbb{Q}$*. If* $N_{pc} - N_i < N_{tank}$ *then add* $N_{tank} - (N_{pc} - N_i)$ *other particles to* $\mathbb{X}$ *and* $\mathbb{Q}$*.*

If $N_i = N_{pc}$ *then add* N_{tank} *particles to* $\mathbb{X}$ *and* $\mathbb{Q}$*.*

The Auction Algorithm with Tank (AAT) is applied to $(\mathbb{X}, \mathbb{Q})$, and by then the particles of $\mathbb{X}$ are assigned to the final locations of $\mathbb{Q}$, leading to the global transport cost $D = \sum_{i=1}^{|\mathbb{X}|} p_{i\sigma^*(i)}$.

In a previous work [9], the triangular transport procedure (TT) was presented as a competitive method for the uniformization of the mass density: in the case of dimension one, the transport cost is known to be optimal, with a very cheap complexity of $\mathcal{O}(n \log n)$.

Table 1 compares the AAT procedure with several tank sizes to the TT procedure. The initial locations $\{x_i\}_{1 \leq i \leq N}$ are randomly created inside $\mathcal{D}$, and $\mathcal{D}$ is partitioned into $6 \times 6 \times 6$ cells, with $N_{pc} = 800$ particles per cell. The four columns correspond to the mean of the following quantities: the computational time on a work station (run time (s)), the transport cost D, the largest number c_{max} of cells crossed by the particles during their transportation (expected to be close to 1), and finally the number c_{move} of particles which have leaved their initial cell. The variable c_{max} plays a crucial role in SDM: particles transport physical information, and hence we look for an optimization procedure that preserves the physics inside each cell.

When $N_{tank} = N_{pc}$ (AAT with $\alpha = 1$, full tank), the Auction Algorithm is applied to the N particles: an optimality condition can be written (see (9)). This test is taken as a reference in terms of transport cost D and c_{max}. Nevertheless the computational time is far too large, and unsuitable for SDM since the mass density uniformization has to be done at *every* time step. Setting $\alpha = 0$ in AAT (empty tank) consists in transporting the supernumerary particles towards the cells that miss particles. With 1% particles in the tank, we obtain a satisfying trade-off between computational and transport costs (see Table 1). Although this procedure does not lead to an optimal transport cost, the number c_{max} of crossed cells is surprisingly small; this is precisely

what matters in our application. The introduction of a better-adapted metric to define $p_{i\sigma^*(i)}$ for our application is still an open problem. Meanwhile, our preferred method remains TT as it both minimizes the computational cost and the number of crossed cells.

4 Conclusion

We have introduced a new numerical scheme which ensures that the Dirichlet condition (2) is satisfied. Then, we have presented a new adaptation of the Auction Algorithm, that improves the resolution of the optimal transport problem in the context of SDM: the computational cost is reduced, involving few particles in the process, with a satisfying transport cost.

References

1. Bernardin, F., Bossy, M., Chauvin, C., Drobinski, P., Rousseau, A., Salameh, T.: Stochastic downscaling method: application to wind refinment. Stoch. Environ. Res. Risk Assess. **23**(6) 851–859 (2009)
2. Bertsekas, D.P.: Auction algorithms. In: Floudas, C.A., Pardalos, P.M. (eds.) Encyclopedia of Optimization, vol. 1, pp. 73–77. Kluwer, London (2001)
3. Bossy, M., Jabir, J.-F.: Confined Langevin processes and mean no-permeability condition. Preprint (2008)
4. Chauvin, C., Hirstoaga, S., Kabelikova, P., Rousseau, A., Bernardin, F., Bossy, M.: Solving the uniform density constraint in a downscaling stochastic model. ESAIM Proc. **24**, 97–110 (2007)
5. Mohammadi, B., Pironneau, O.: Analysis of the k-Epsilon Turbulence Model. Masson, Paris, (1994)
6. Mora, C.M.: Weak exponential schemes for stochastic differential equations with additive noise. IMA J. Numer. Anal. **25**(3), 486–506 (2005)
7. Pope, S.B.: P.D.F. methods for turbulent reactive flows. Prog. Energy Comb. Sci. **11**, 119–192 (1985)
8. Raviart, P.-A.: An analysis of particle methods. In: Numerical Methods in Fluid Dynamics (Como, 1983), vol. 1127, pp. 243–324. Springer, Berlin (1985)
9. Rousseau, A., Bernardin, F., Bossy, M., Drobinski, P., Salameh, T.: Stochastic particle method applied to local wind simulation. In: IEEE International Conference on Clean Electrical Power, pp. 526–528. Springer, Berlin (2007)
10. Talay, D.: Probabilistic numerical methods for partial differential equations: elements of analysis. In: Talay, D., Tubaro, L. (eds.) Probabilistic Models for Nonlinear Partial Differential Equations, vol. 1627, pp. 148–196. Springer, Berlin (1996)

Parallel Numerical Algorithm for Simulation of Counter Propagation of Two Laser Beams

R. Čiegis[1], I. Laukaitytė[1], and V. Trofimov[2]

[1] Vilnius Gediminas Technical University, Saulėtekio al. 11, Vilnius LT-10223, Lithuania, rc@fm.vgtu.lt
[2] M.V. Lomonosov Moscow State University, Vorob'evy gory, Moscow 119992, Russia, vatro@cs.msu.ru

Summary. ParSol library is applied to implement the finite difference scheme used to solve numerically a system of PDEs describing a nonlinear interaction of two counter-propagating laser waves. Results of computational experiments are presented.

1 Problem Formulation

The interaction of counter propagating laser beams is of great practical interest. We mention only few applications, including development of optical switches with short response time and creation of optical processors. Then the new types of optical bistability are important and this problem is actively studied (see [6, 7] and references given in these papers).

Since in practical applications the greatest interest is given to nonlinear response of medium, we shall consider the Kerr nonlinearity. In the domain $D(z, X) = (0 \le z \le L_z) \times D(X),\ D(X) = \{0 \le x_k \le L_x,\ \ k = 1, 2\}$ dimensionless equations and boundary conditions describing a nonlinear interaction of two counter propagating laser beams are given by the system of equations

$$\frac{\partial A^+}{\partial t} + \frac{\partial A^+}{\partial z} + i\sum_{k=1}^{2} D_k \frac{\partial^2 A^+}{\partial x_k^2} + i\gamma\big(0.5|A^+|^2 + |A^-|^2\big)A^+ = 0, \tag{1}$$

$$\frac{\partial A^-}{\partial t} - \frac{\partial A^-}{\partial z} + i\sum_{k=1}^{2} D_k \frac{\partial^2 A^-}{\partial x_k^2} + i\gamma\big(0.5|A^-|^2 + |A^+|^2\big)A^- = 0, \tag{2}$$

and the boundary conditions

$$A^+(t, z = 0, x_1, x_2) = A_0(t) \exp\left(-\sum_{k=1}^{2} \frac{(x_k - x_{ck})^{m_k}}{r_{pk}} \right),$$

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*, Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_124,

$$A^-(t, z = L_z, x_1, x_2) = A^+(z = L_z, x_1, x_2, t)R_0 \tag{3}$$

$$\times \Big(1 - \exp\Big(-\sum_{k=1}^{2} \frac{(x_k - x_{mk})^{q_k}}{R_{ak}}\Big)\Big) \exp\Big(i \sum_{k=1}^{2} \frac{(x_k - x_{mk})^2}{R_{mk}}\Big). \tag{4}$$

Here $A^{\pm}$ are complex amplitudes of counter propagating pulses, γ characterizes the nonlinear interaction of laser pulses, x_{ck} are coordinates of the beam center, r_{pk} are radius of input beam on the transverse coordinates and $A_0(t)$ is a temporal dependence of input laser pulses. In the boundary conditions, R_0 is the reflection coefficient of the mirror, R_{ak} are the radius of the hole along the transverse coordinates, x_{mk} are coordinates of the hole center, R_{mk} characterize curvature of the mirror.

At the initial time moment the amplitudes of laser pulses are equal to zero $A^{\pm}(0, z, x_1, x_2) = 0$, $(z, x_1, x_2) \in D$. Boundary conditions along transverse coordinates are equal to zero.

It is well known that the solution of the given system satisfies some important invariants [1,9]. In this paper we consider only two main invariants. Let us introduce new local space coordinates $\eta^{\pm} = z \pm (t - t_0)$ and define new functions $a^{\pm}(t, \eta^{\mp}, x_1, x_2) = A^{\pm}(t, z, x_1, x_2)$. It follows from (1) and (2) that these functions satisfy the following system of equations

$$\frac{\partial a^{\pm}}{\partial t} + i \sum_{k=1}^{2} D_k \frac{\partial^2 a^{\pm}}{\partial x_k^2} + i\gamma\big(0.5|a^{\pm}|^2 + |a^{\mp}|^2\big)a^{\pm} = 0. \tag{5}$$

Multiplying differential equations (5) by $(a^{\pm})^*$ and integrating over $(t_0 - h/2, t_0 + h/2) \times D(X)$ we prove that

$$\|a^{\pm}(\eta^{\mp}, t_0 + h/2)\|^2 = \|a^{\pm}(\eta^{\mp}, t_0 - h/2)\|^2.$$

Taking $t_0 = \bar{t} - h/2$ and denoting $z^{\pm} = z \pm h/2$, we get that the full energy of each laser pulse is conserved during propagation along the directions of characteristics:

$$\|A^+(z^+, \bar{t})\|^2 = \|A^+(z^-, \bar{t} - h)\|^2, \quad \|A^-(z^-, \bar{t})\|^2 = \|A^-(z^+, \bar{t} - h)\|^2. \tag{6}$$

Here the L_2 norm is defined as $\|A(z,t)\|^2 = \int_0^{L_x} \int_0^{L_x} |A|^2 \, dx_1 dx_2$.

Multiplying differential equations (5) by $\frac{\partial (a^{\pm})^*}{\partial t}$ and integrating over $(t_0 - h/2, t_0 + h/2) \times D(X)$ we prove that

$$I_2(t_0 + h/2) := \sum_{k=1}^{2} D_k \Big(\Big\| \frac{\partial a^+(\eta^-, t_0 + h/2)}{\partial x_k} \Big\|^2 + \Big\| \frac{\partial a^-(\eta^+, t_0 + h/2)}{\partial x_k} \Big\|^2 \Big)$$

$$- \gamma \int_0^{L_x} \int_0^{L_x} \Big(\frac{1}{4} |a^+(\eta^-, t_0 + h/2, X)|^4 + \frac{1}{4} |a^-(\eta^+, t_0 + h/2, X)|^4$$

$$+ |a^+(\eta^-, t_0 + h/2, X)|^2 |a^-(\eta^+, t_0 + h/2, X)|^2 \Big) \, dx_1 dx_2 = I_2(t_0 - \frac{h}{2}).$$

Taking $t_0 = \bar{t} - h/2$ and denoting $z^{\pm} = z \pm h/2$, we get the second invariant

$$\sum_{k=1}^{2} D_k\Big(\Big\|\frac{\partial A^+(z^+,\bar{t})}{\partial x_k}\Big\|^2+\Big\|\frac{\partial A^-(z^-,\bar{t})}{\partial x_k}\Big\|^2\Big)-\gamma\int_0^{L_x}\int_0^{L_x}\Big(\frac{1}{4}|A^+(z^+,\bar{t},X)|^4$$
$$+\frac{1}{4}|A^-(z^-,\bar{t},X)|^4+|A^+(z^+,\bar{t},X)|^2\,|A^-(z^-,\bar{t},X)|^2\Big)\,dx_1dx_2$$
$$=\sum_{k=1}^{2} D_k\Big(\Big\|\frac{\partial A^+(z^-,\bar{t}-h)}{\partial x_k}\Big\|^2+\Big\|\frac{\partial A^-(z^+,\bar{t}-h)}{\partial x_k}\Big\|^2\Big)$$
$$-\gamma\int_0^{L_x}\int_0^{L_x}\Big(\frac{1}{4}|A^+(z^-,\bar{t}-h,X)|^4+\frac{1}{4}|A^-(z^+,\bar{t}-h,X)|^4$$
$$+|A^+(z^-,\bar{t}-h,X)|^2\,|A^-(z^+,\bar{t}-h,X)|^2\Big)\,dx_1dx_2. \tag{7}$$

These two invariants (6) and (7) describe very important features of the solution and therefore it is important to guarantee that the discrete analogous are satisfied for the numerical solution. In many cases this helps to prove the existence and convergence of the discrete solution. Conservative discrete schemes for problems of nonlinear optics are investigated in many papers, see e.g. [4,5,8,9], where a comparison of conservative and non-conservative discrete schemes is done for the nonlinear Schrödinger problem and systems of such equations.

In this paper we develop a conservative finite difference scheme, solution of which satisfies both discrete invariants. The given mathematical model depends on three space coordinates (z, x_1, x_2) thus the computational complexity is much larger than in the case of 2D models used previously. We propose a parallel version of the numerical algorithm and implement it using ParSol tool of parallel numerical objects [2,3]. Some results of computational experiments are presented.

2 Finite Difference Scheme

In the domain $[0,T]\times D$ we introduce a uniform grid $\Omega=\Omega_t\times\Omega_z\times\Omega_x$, where

$$\Omega_t=\{t^n=nh_t,\ \ n=0,\dots,N\},\quad \Omega_z=\{z_j=jh_z,\ \ j=0,\dots,J\},$$
$$\Omega_x=\{(x_{1l},x_{2m}),\ \ x_{km}=mh_x,\ \ k=1,2,\ \ m=0,\dots,M\}.$$

In order to approximate the transport part of the differential equations by using the finite differences along the characteristics $z\pm t$ we take $h_t=h_z$. Let us denote discrete functions, defined on the grid Ω by $E^{\pm,n}_{j,lm}=E^{\pm}(z_j,x_{1l},x_{2m},t^n)$. We also will use the following operators:

$$\bar{E}^+=\frac{E^{+,n}_j+E^{+,n-1}_{j-1}}{2},\ \bar{E}^-=\frac{E^{-,n}_{j-1}+E^{-,n-1}_j}{2},\ \beta(E,W)=\gamma(\frac{1}{2}|E|^2+|W|^2),$$
$$\mathcal{D}E_{j,kl}=D_1\frac{E_{j,l+1,m}-2E_{j,lm}+E_{j,l-1,m}}{h_x^2}+D_2\frac{E_{j,l,m+1}-2E_{j,lm}+E_{j,l,m-1}}{h_x^2}.$$

Then the system of differential equations is approximated by the following finite difference scheme

$$\frac{E_j^{+,n} - E_{j-1}^{+,n-1}}{h_t} + i\mathcal{D}\bar{E}_j^{+} + i\beta(\bar{E}_j^{+}, \bar{E}_j^{-})\bar{E}_j^{+} = 0, \tag{8}$$

$$\frac{E_{j-1}^{-,n} - E_j^{-,n-1}}{h_t} + i\mathcal{D}\bar{E}_j^{-} + i\beta(\bar{E}_j^{-}, \bar{E}_j^{+})\bar{E}_j^{-} = 0$$

with corresponding boundary and initial conditions.

We shall prove that this scheme is conservative, i.e. two discrete invariants are satisfied for its solution. Let us define the scalar product and the L_2 norm of the discrete functions as $(U, V) = \sum_{l=1}^{M-1}\sum_{m=1}^{M-1} U_{lm}V_{lm}^{*}h_x^2$, $\|U\|^2 = (U, U)$. Taking scalar products of equations (8) by $\bar{E}^{+}$ and $\bar{E}^{-}$ respectively and considering the real parts of the equations, we get that the discrete analogs of the invariants (6) are satisfied

$$\|E_j^{+,n}\|^2 = \|E_{j-1}^{+,n-1}\|^2, \quad \|E_{j-1}^{-,n}\|^2 = \|E_j^{-,n-1}\|^2, \quad j = 1, \ldots, J.$$

Taking scalar products of (8) by $(E_j^{+,n} - E_{j-1}^{+,n-1})$ and $(E_{j-1}^{-,n} - E_j^{-,n-1})$ respectively, adding the obtained equalities and considering the imaginary part of the equation, we get the discrete analog of the second invariant (7):

$$\begin{aligned}
&(\mathcal{D}E_j^{+,n}, E_j^{+,n}) + (\mathcal{D}E_{j-1}^{-,n}, E_{j-1}^{-,n}) - \gamma\Big(\frac{1}{4}\big(|E_j^{+,n}|^2, |E_j^{+,n}|^2\big) \\
&\qquad + \frac{1}{4}\big(|E_{j-1}^{-,n}|^2, |E_{j-1}^{-,n}|^2\big) + \big(|E_j^{+,n}|^2, |E_{j-1}^{-,n}|^2\big)\Big) \\
&= (\mathcal{D}E_{j-1}^{+,n-1}, E_{j-1}^{+,n-1}) + (\mathcal{D}E_j^{-,n-1}, E_j^{-,n-1}) - \gamma\Big(\frac{1}{4}\big(|E_{j-1}^{+,n-1}|^2, |E_{j-1}^{+,n-1}|^2\big) \\
&\qquad + \frac{1}{4}\big(|E_j^{-,n-1}|^2, |E_j^{-,n-1}|^2\big) + \big(|E_{j-1}^{+,n-1}|^2, |E_j^{-,n-1}|^2\big)\Big).
\end{aligned}$$

3 Parallel Algorithm

The finite difference scheme (8) uses the structured grid and the complexity of computations at each node of the grid is approximately the same (it depends on the number of iterations used to solve a nonlinear discrete problem for each z_j). The parallelization of such algorithms can be done by using domain decomposition paradigm and ParSol is exactly targeted for such algorithms. In this paper we apply the 1D block domain decomposition algorithm, decomposing the grid only in z direction. Such a strategy enables us to use a sequential version of the FFT algorithm for solution of the 2D linear systems with respect to (x_1, x_2) coordinates.

This parallel algorithm is generated semi-automatically by ParSol. The parallel vectors, which are used to store discrete solutions $E^{\pm}$, are created by specifying three main attributes:

(a) The dimension of the parallel vector is 3D.
(b) The topology of processors is 1D and only z coordinate is distributed.
(c) The 1D grid stencil is defined by the points (z_{j-1}, z_j, z_{j+1}).

Thus in order to implement the computational algorithm the kth processor $(k = 0, 1, \ldots, p)$ defines its subgrid as well its ghost points $\Omega(k)$, where $\Omega(k) = \{(z_j, x_{1l}, x_{2m}),\ z_j \in \Omega_z(k),\ (x_{1l}, x_{2m}) \in \Omega_x\},\ \Omega_z(k) = \{z_j : \widetilde{j}_L(k) \le j \le \widetilde{j}_R(k)\},\ \widetilde{j}_L(k) = \max(j_L(k) - 1, 0),\ \widetilde{j}_R = \min(j_R(k) + 1, J)$. At each time step t^n and for each $j = 1, 2, \ldots, J$ the processors must exchange some information for ghost points values. Since the computations move along the characteristics $z \pm t$ only a half of the full data on ghost points is required to be exchanged. The kth processor (a) sends to $(k+1)$th processor vector $E^{+,n}_{j_R,\cdot}$ and receives from him vector $E^{-,n}_{\widetilde{j}_R,\cdot}$, and sends to $(k-1)$th processor vector $E^{-,n}_{j_L,\cdot}$ and receives from him vector $E^{+,\cdot}_{\widetilde{j}_L,\cdot}$. Obviously, if $k = 0$ or $k = (p-1)$, then a part of communications is not done. In ParSol, such an optimized communication algorithm is obtained by defining temporal reduced stencils for vectors E^+ and E^-, they contain ghost points only in the required directions but not in both.

Let assume that p processors are used. The total complexity of the parallel algorithm is given by

$$T_p = \gamma \max_{0 \le k < p} K(k)\big(\lceil (J+1)/p \rceil + 1\big)(M+1)^2 \log M + 2\big(\alpha + \beta(M+1)^2\big),$$

where γ estimates the CPU time required to implement one basic operation of the algorithm, α is the message startup time and β is the time required to send one element of data, K is the averaged number of iterations done at one time step. We assume that communication between neighbour processors is done in parallel.

The parallel code was tested on the cluster of PCs at Vilnius Gediminas Technical University. It consists of 16 Intel Quad Core nodes interconnected via Gigabit Smart Switch (http://vilkas.vgtu.lt). Some results of computational experiments are presented in Table 1. Here coefficients of the algorithmic speed up $S_p = T_1/T_p$ and efficiency $E_p = S_p/p$ are presented. $p_1 \times n$ denotes that p_1 nodes with n tasks per node are used, thus the total number of processors is $p = p_1 n$. The size of the discrete problem is $M = 123$ and $J = 640$.

More applications of the developed parallelization tool ParSol are described in [2,3].

Table 1. Results of computational experiments on Vilkas cluster

	2×1	1×2	4×1	2×2	1×4	8×1	4×2	2×4
S_p	1.85	1.82	3.20	3.23	3.18	5.13	5.12	4.94
E_p	0.92	0.91	0.80	0.81	0.79	0.64	0.64	0.62

Acknowledgments

R. Čiegis and I. Laukaitytė were supported by the Lithuanian State Science and Studies Foundation within the project on B-03/2008 "Global optimization of complex systems using high performance computing and GRID technologies."

References

1. Ablowitz, J.M., Prinari, B., Trubatch, A.D.: Discrete and Continuous Nonlinear Schrodinger Systems. Cambridge University Press, Cambridge (2004)
2. Čiegis, R., Jakušev, A., Krylovas, A., Suboč, O.: Parallel algorithms for solution of nonlinear diffusion problems in image smoothing. Math. Model. Anal. **10**(2), 155–172 (2005)
3. Čiegis, R., Jakušev, A., Starikovičius, V.: Parallel tool for solution of multiphase flow problems. In: Wyrzykowski, R., Dongarra, J., Meyer, N., Wasniewski, J. (eds.) Lecture Notes in Computer Science, vol. 3911, pp. 312–319. Sixth International Conference on Parallel Processing and Applied Mathematics, Poznan, Poland, September 10–14, 2005. Springer, Berlin (2006)
4. Čiegis, R., Štikonienė, O.: Semiimplicit schemes for nonlinear Schrödinger type equations. In: Amann, H., Galdi, G., Pileckas, K., Solonikov, V. (eds.) Proceedings of the 6th International Conference NSEC-6, Palanga, Lithuania, 1997. Navier-Stokes Equations and Related Nonlinear Problems, VSP/TEV, Utrect/Vilnius, pp. 53–68 (1998)
5. Ismail, M., Taha, T.: A linearly implicit conservative scheme for the coupled nonlinear Schrödinger equation. Math. Comput. Simul. **74**, 302–311 (2007)
6. Nikitenko, K.Yu., Trofimov, V.A.: Optical bistability based on nonlinear oblique reflection of light beams from a screen with an aperture on its axis. Quantum Electron. **29**(2), 147–150 (1999)
7. Osuch, K., Pura, B., Petykiewicz, J., Wierzbicki, M., Wrzesinski, Z.: The optical bistability of polarisation in B_5NH_4 crystal caused by the optical Kerr effect. Opt. Mater. **27**(1), 39–43 (2004)
8. Sanz-Serna, J.M., Verwer, J.G.: Conservation and nonconservation schemes for the solution of the nonlinear Schrödinger equation. IMA J. Numer. Anal. **6**, 25–42 (1986)
9. Tereshin, E.B., Trofimov, V.A.: Conservative finite difference scheme for the problem of propagation of a femtosecond pulse in a photonic crystal with combined nonlinearity. Comput. Math. Math. Phys. **46**(12), 2154–2165 (2006)

Modelling Burglaries in Streets

John P. Curtis, Frank T. Smith, and Xiang Ye

Department of Mathematics, UCL, Gower Street, London WC1E 6BT, UK
j.p.curtis@talk21.com, frank@math.ucl.ac.uk, frankye211@hotmail.com

Summary. Basic modelling of the evolution of burglary probabilities along a street or system of streets is described. The central cases investigated are for a single street, a system of streets in series, a system of streets in parallel, and the effects of security.

1 Introduction

Interest in understanding and modelling certain aspects of crime has been growing significantly in recent times. Studies include those in [1–6]. The present work arose from a workshop in 2007 and is motivated by links established with the Jill Dando Institute at UCL which in turn is linked closely with the London Metropolitan Police in research terms.

The modelling here is described in Sect. 2 for a single street, then extended in Sect. 3 to two or more streets in series and to junctions. This is followed by Sect. 4 which is on special interactions concerned with streets in an in-parallel configuration. Section 5 addresses the modelling of security or watchfulness effects. The investigation is concentrated on the discrete versions of models rather than continuum ones. Also the methodology was tested for analytical cases and compared with limit continuum calculations. Section 6 provides further comments.

2 Basic Model for One Street

This is a simple first model for a rectilinear street consisting of K houses or other dwellings. The reasoning here is that the probability p_k of a burglary at the k^{th} house of the street evolves according to the differences between the probability at that house and those at its nearest neighbours. The integer k runs from 1 to K, and in addition superscripts $i, i+1$ are to be used, standing for the values at the current time level and at the next time level in turn.

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*,
Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_125,

The typical time step in mind is one day say. Thus the normalised evolution equation has the form

$$p_k^{(i+1)} = p_k^{(i)} + \mu(p_{k+1}^{(i)} - p_k^{(i)}) + \mu'(p_{k-1}^{(i)} - p_k^{(i)}). \tag{1}$$

The coefficients μ, μ' are taken as constants and are expected to be positive, such that if the probability at $k+1$ (or at $k-1$) is higher than that at k then the latter is likely to increase in proportion during the next time step. For definiteness we then suppose the coefficients μ, μ' to be equal, effectively yielding a diffusivity of crime through the relation

$$p_k^{(i+1)} = p_k^{(i)} + \mu(p_{k+1}^{(i)} - 2p_k^{(i)} + p_{k-1}^{(i)}). \tag{2}$$

The case of unequal coefficients is also of interest however and is mentioned in the final section of the paper. Next we note that (2) is of course the diffusion equation in discretised form, controlling heat conduction for example; that is, the continuum limit of (2) for many houses is

$$\frac{\partial p}{\partial t} = \kappa \frac{\partial^2 p}{\partial x^2} \tag{3}$$

where the coefficient $\kappa = \mu(\Delta x)^2/(\Delta t)$, subject to the appropriate limiting process involving the effective step lengths $\Delta x, \Delta t$ in space x and time t respectively. Development in space here corresponds to movement from one house to the next. Most of our concern in this article is with the form (2).

Initial conditions (set at $i = 0$ for all k) and boundary conditions (usually at $k = 1, K$ for all positive i) are imposed appropriately on (2). In addition a refined model of the single street case is currently under investigation in which the burglar is modelled as an agent. In brief, the probabilities evolve as above (and below) in the absence of a burglary. Concerning the probability of a burglary on day i (as a first step only one burglary is envisaged), the agent proceeds down the street from house 1 to 2, $\ldots k \ldots$ and commits a crime with a certain probability which is a function of security etc, as considered in Sect. 6. The crime itself acts to produce new initial conditions for the evolution in (2). An object-oriented C++ code for this is under development.

Figure 1 shows a sample solution of (2), in which the diffusive effect is clear as the discrete time i increases. The probabilities of crime are initially zero in all the houses, save House 6, where we model a crime having just occurred by raising the probability to the value of 0.5. This value may be artificially high, but it serves to illustrate the diffusion of probability of crime in a manner directly analogous to thermal diffusion. It can be seen how the probability of crime in neighbouring houses rises rapidly in close proximity to the burgled house and more slowly and steadily further away. It is worth remarking that, in the present discrete setting, if the initialised p values are set as nonzero only towards the middle of the street with p imposed as zero at the two ends then p remains zero near those ends for a finite time.

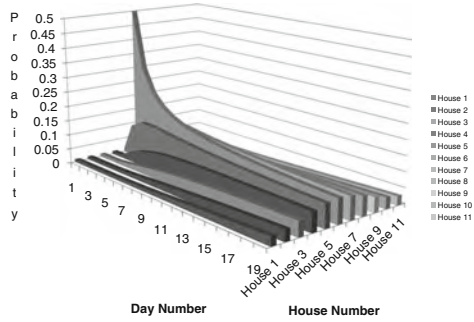


Fig. 1. The evolution of probability of crime in a street of 11 houses is shown

3 Two or More Streets in Series or with Junctions

The first extension now to two streets in series, say street P (with K_p houses) leading in to street Q (with K_q houses), has the corresponding probabilities being p, q. Here p evolves in the form (2) with a coefficient now written μ_p and q evolves in a similar form, namely

$$q_k^{(i+1)} = q_k^{(i)} + \mu_q(q_{k+1}^{(i)} - 2q_k^{(i)} + q_{k-1}^{(i)}), \tag{4}$$

where k now runs from 1 to K_p in (2) but from 1 to K_q in (4). The junction of the streets is at $k = K_p$ in P street, $k = 1$ in Q street. The un-joined ends are taken to have zero probability conditions,

$$p_1^{(i)} = 0, q_{Kq}^{(i)} = 0, \text{ for all } (i), \tag{5}$$

whereas a modelled interaction condition applies at the junction. This is

$$(q_2^{(i)} - q_1^{(i+1)}) = \alpha_3(p_{Kp}^{(i+1)} - p_{Kp-1}^{(i)}). \tag{6}$$

Here α_3 is a prescribed constant, reflecting proportionality between the probability rates on either side of the junction, with p_{Kp}, q_1 being taken to be identical purely for convenience in these first studies. For certain situations, e.g. involving a change in housing type, it can be argued that α_3 should be the ratio of the μ values but other situations such as at crossroads suggest a wider range of α_3 values can hold; again, taking continuity of p, q is open to some debate. The junction condition allows a type of leakage to occur from one street to the other.

Examples have been calculated for particular values of the coefficients involved and will be in [6]. For the scenario of all the $q's$ initially being zero and the $p's$ likewise except at a finite number of houses in the middle, we observe that Q street remains at zero probability until the diffusion in P street makes the p values near the junction become nonzero. Further, even after such a delay, an extreme value for the leakage coefficient in (6) can make the response in Q street grow dramatically.

A similar description holds (second) for more streets, say $P, Q, R, S, \ldots$, in series: P leading in to Q, then Q in to R and so on. The evolution equation in each street is as in (2), (4), while the interaction conditions at each junction are (6) between p, q, then q, r, etc. Examples are in [6].

A similar account can also be made of junctions in which several streets join together, for instance with street P leading in to both Q and R at the same junction, or with a crossroads. Here again interaction of the kind in (6) seems reasonable as a first step.

4 Streets in Parallel

When two streets P, Q are in effect in parallel it is considered that there is not necessarily any interaction of the above type directly between them; however interaction may occur at a finite number of special positions, corresponding to an alleyway or parkland for instance offering possibly enhanced access. At such a position k the probability p_k at the house in P street is taken to be influenced positively by the probability q_k at the corresponding house in Q street and likewise for the effect of P street on Q street. In consequence the evolution system has the form

$$p_k^{(i+1)} = p_k^{(i)} + \mu_p(p_{k+1}^{(i)} - 2p_k^{(i)} + p_{k-1}^{(i)}) + \beta_p(q_k^{(i)} - \epsilon p_k^{(i)}), \tag{7}$$

$$q_k^{(i+1)} = q_k^{(i)} + \mu_q(q_{k+1}^{(i)} - 2q_k^{(i)} + q_{k-1}^{(i)}) + \beta_q(p_k^{(i)} - \epsilon q_k^{(i)}), \tag{8}$$

at such k values, where the access factors β_p , β_q are constants which are generally positive.

The constant factor ϵ plays an interesting role. If ϵ is unity then the evolution remains as before dependent on various differences between probabilities, whereas if ϵ is zero then there is a stronger linkage between the two streets. In fact exponential growth in time arises whenever ϵ is between zero and unity. This can be seen readily if there is no significant diffusion in the system, so that the μ values are negligibly small, since then (7) and (8) give the property that at large times

$$p_k^{(i)}, q_k^{(i)} \text{ grow as } b^i, \text{ with } b - 1 = [-\epsilon(\beta_p + \beta_q) + \{\epsilon(\beta_p - \beta_q)^2 + 4\beta_p\beta_q\}^{\frac{1}{2}}]/2, \tag{9}$$

at the access positions k. Hence exponential growth is encountered ($b > 1$) except when ϵ is unity; the unit case is associated with all previous interactions addressed, by the way. Even if there is significant diffusion essentially the same conclusion is reached, by virtue of a single or double summation over all k values, depending on the end conditions, although the exponent is then altered from that in (9).

Numerical studies support the result (9) and its extension to three or more streets in parallel when they interact in the special access way. An example

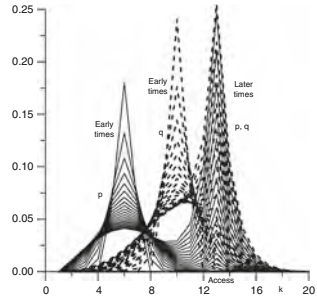


Fig. 2. For two in-parallel streets

is presented in Fig. 2. This and other cases examined allow for special access occurring at a number of locations; thesis studies are being conducted by XY. It might be argued that the accesses for the current in-parallel situations should merely be treated as crossroads as in the previous section, but against that these accesses are believed to have a special status.

5 Effects of Security

Extending the model to allow for security or watchfulness s_k in the same simple-minded way we first have the system

$$p_k^{(i+1)} = p_k^{(i)} + \mu_p(p_{k+1}^{(i)} - 2p_k^{(i)} + p_{k-1}^{(i)}) - a_3 s_k^{(i+1)}, \tag{10}$$

$$s_k^{(i+1)} = s_k^{(i)} + b_1 p_k^{(i)} - b_2, \tag{11}$$

where a_3, b_1, b_2 are non-negative constants. The main idea is that security put in at the typical k^{th} house reduces the crime probability there, as reflected in (10), while any increase in crime probability tends to induce further security as represented in (11) but subject to security being reduced if there is negligible probability.

An example is presented in Fig. 3. It shows crime probability decreasing when security is imposed, along with neighbouring probability levels which may decrease or increase relatively depending on the parameters present. We note that under some circumstances the system can produce apparently unrealistic negative values for the p_k and /or the s_k at certain positions although this feature can be prevented by sensible adjustments of the model, one of which is discussed in the next section.

6 Final Comments

- According to the modelling the crime probabilities mostly diffuse with time for the single street and multiple street cases, with the diffusion rates depending sensitively on the parameters involved. A possible exception

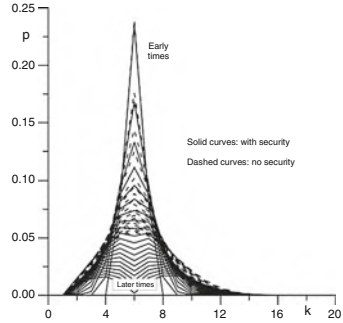


Fig. 3. Showing effects of security

is for the in-parallel scenarios described in Sect. 3 which admit temporal growth of the probabilities.

- In the multiple street cases the junctions between streets can play a key role, as is well known in other network settings. We see this aspect in all the models of Sects. 3 and 4.
- For the most part linear effects have been addressed so far. An exception should probably be made in future studies for the setting of Sect. 5 concerning security effects. The relations in (10) and (11) could be better treated as giving growth or decay rates, which would make the interactions nonlinear in addition to ensuring that the probabilities and security levels remain sensibly non-negative.
- In the conference itself there were several interesting points made by audience members at the talk associated with this article. Most had already been accommodated in the research, including (first) the continuum version (3) and its solution properties and (second) the possibility of unequal coefficients in (1) which provokes a directional preference corresponding to a convective effect proportional to $(\mu\text{-}\mu')$ in (2). Other comments are covered by this text or will be covered in later research including [6]. There are indeed many further issues to be studied.

Acknowledgments

We thank especially Shane Johnson, Steve Bishop, John Ockendon and the audience for helpful comments and interest. JPC acknowledges gratefully a Royal Society Industry Fellowship.

References

1. Bernasco, W.: Them again?: Same-offender involvement in repeat and near repeat burglaries. Eur. J. Criminol. **5**, 411–432 (2008)

2. Johnson, S.D.: Repeat burglary victimisation: a tale of two theories. J. Exp. Criminol. **4**, 215–240 (2008)
3. Kleemans, E.R.: In: Farrell, G., Pease, K. (eds.) Crime Prevention Studies, vol. 12. CRC, Monsey (2001)
4. Pease, K.: The Home Office: Police Research Group: Crime Detection and Prevention Series Paper 90 (1998)
5. Short, M.B., D'Orsogna, M.R., Pasour, V.B., Tita, G.E., Brantingham, P.J., Bertozzi, A.L., Chayes, L.: A statistical model of criminal behavior. Math. Models Methods Appl. Sci. Special Issue on Traffic, Crowds, and Swarms **18**, 1249–1267 (2008)
6. Ye, X.: Ph.D. Thesis, University College London (2011) (in preparation)

Approximate Numerical Solutions of Autonomous Second-Order Matrix Models Using Cubic Matrix Splines

E. Defez[1], M.M. Tung[1], J. Ibañez[2], and A. Hervás[1]

[1] Instituto de Matemática Multidisciplinar, Universidad Politécnica de Valencia, Camino de Vera s/n, 46022 Valencia, España, edefez@imm.upv.es, mtung@imm.upv.es, ahervas@mat.upv.es

[2] Instituto Universitario de Aplicaciones de las Tecnologías Información y Comunicaciones Avanzadas, Universidad Politécnica de Valencia, Camino de Vera s/n, 46022 Valencia, España, jjibanez@dsic.upv.es

Summary. This work discusses the construction of approximate solutions for the initial matrix problem $Y''(x) = f(Y(x), Y'(x))$ using cubic matrix splines.

1 Introduction and Review

Autonomous matrix initial value problems are frequently encountered in diverse fields of physics and engineering, see e.g. [1] and references therein. Usually, initial matrix problems of the type $Y''(x) = f(Y(x), Y'(x))$ can be recast as an extended first order matrix problem [2]. This standard approach, however, comes with an increase of additional computational cost due to the higher dimensionality of the problem. Cubic splines were used in the scalar case to obtain approximations for first-order differential equations [3], who have the advantage to be of class C^1 in a given approximation interval. Furthermore, scalar splines are easy to program on the computer and the associated approximation errors are only of the fourth order of the chosen step size in the iteration algorithm. This method has also been used in the resolution of linear matrix problems [4], first-order matrix differential equations [5], and for the particular cases of second-order problems in [6].

In the present work, we extend this scheme to the resolution of second-order initial matrix problems $Y''(x) = f(Y(x), Y'(x))$ without recurring to any additional increase in dimensionality of the problem. Note that after the presentation of this work, the more general case was accepted for publication in [7]. Throughout, we will adopt the common notation for norms and cubic matrix splines as in [5]. The paper is organized as follows: Sect. 2 develops the proposed method; in Sect. 3 an algorithm is given; and finally, in Sect. 4, an example is provided.

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*, Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_126,

2 Construction of the Method

The autonomous initial-value problem under consideration is given by the following system of equations

$$Y''(x) = f(Y(x), Y'(x)) \ , \ Y(a) = Y_0 \ , \ Y'(a) \ = \ Y_1, \ a \leq x \leq b, \qquad (1)$$

where $Y_0, Y_1, Y(t) \in \mathbb{C}^{r\times s}$ and $f : \mathbb{C}^{r\times s} \times \mathbb{C}^{r\times s} \longmapsto \mathbb{C}^{r\times s}$, $f \in C^0(T)$, with $T \ = \ \{(Y, Z);\ Y, Z \in \mathbb{C}^{r\times s}\}$. To guarantee the existence and uniqueness of the continuously differentiable solution $Y(x)$ of the system (1), we impose the following Lipschitz conditions on function f so that [8, p. 99]

$$\|f(Y_1, Y) - f(Y_2, Y)\| \leq L_1 \|Y_1 - Y_2\|, \|f(Z, Z_1) - f(Z, Z_2)\| \leq L_2 \|Z_1 - Z_2\|, \qquad (2)$$

for $Y_1, Y_2, Z_1, Z_2, Y, Z \in \mathbb{C}^{r\times s}$. We also split the interval $[a, b]$ into subintervals according to

$$\Delta_{[a,b]} = \{a = x_0 \ < \ x_1 \ < \ \ldots \ < \ x_n = b\}, \quad x_k = a + kh, \quad k = 0, 1, \ldots, n, \qquad (3)$$

where $h = (b-a)/n$ is the step size for any positive integer n. In each of these subintervals $[a+kh, a+(k+1)h]$, our objective is to find cubic matrix splines approximating the solution of (1). For the first interval $[a, a+h]$, we simply assume that the spline is of the form

$$S_{\big|_{[a,a+h]}}(x) \ = \ Y(a) \ + \ Y'(a)(x-a) \ + \ \frac{1}{2!}Y''(a)(x-a)^2 \ + \ \frac{1}{3!}A_0(x-a)^3, \qquad (4)$$

where the matrix $A_0 \in \mathbb{C}^{r\times s}$ is an unknown to be determined. Given this definition of the initial spline (4), it is straightforward to check that for $k = 0, 1$, one gets

$$S^{(k)}_{\big|_{[a,a+h]}}(a) = Y^{(k)}(a) \text{ and } S''_{\big|_{[a,a+h]}}(a) = Y''(a) = f(S_{\big|_{[a,a+h]}}(a), S'_{\big|_{[a,a+h]}}(a)), \qquad (5)$$

and hence (4) satisfies (1) at $x = a$. For a full determination of the spline in subinterval $[a, a+h]$, it is still necessary to solve for A_0. By requiring that (4) is a solution of problem (1) at $x = a + h$, we obtain

$$S''_{\big|_{[a,a+h]}}(a+h) \ = \ f\left(S_{\big|_{[a,a+h]}}(a+h), S'_{\big|_{[a,a+h]}}(a+h)\right), \qquad (6)$$

and thus find the matrix equation for the remaining unknown matrix

$$A_0 = \frac{1}{h}\left[f\left(Y(a)+Y'(a)h \ + \ \frac{Y''(a)h^2}{2} + \frac{A_0h^3}{6}, Y'(a) + Y''(a)h + \frac{A_0h^2}{2}\right) - Y''(a)\right]. \qquad (7)$$

With the uniqueness of solution A_0 given by (7), the matrix spline of the first subinterval $[a, a+h]$ is finally fully determined.

Now, we move on to the next subinterval $[a+h, a+2h]$, where the cubic matrix spline takes the following form

$$S_{\big|_{[a+h,a+2h]}}(x) = S_{\big|_{[a,a+h]}}(a+h)+S'_{\big|_{[a,a+h]}}(a+h)(x-(a+h)) + \frac{1}{2!}S''_{\big|_{[a,a+h]}}(a+h)(x-(a+h))^2+\frac{1}{3!}A_1(x-(a+h))^3. \quad (8)$$

Hence, it becomes obvious that $S(x)$ is of class $C^2([a,a+h]\cup[a+h,a+2h])$, and again all of the coefficients of the spline $S_{\big|_{[a+h,a+2h]}}(x)$ are determined with exception of matrix $A_1 \in \mathbb{C}^{r\times s}$. Following the same procedure as before for the first subinterval, we consider a spline of the form (8) which fulfills the differential equation (1) at point $x = a + h$. Then, it is possible to find the expression of A_1, imposing that the differential equation (1) also holds at $x = a + 2h$:

$$S''_{\big|_{[a+h,a+2h]}}(a+2h) \;=\; f\left(S_{\big|_{[a+h,a+2h]}}(a+2h), S'_{\big|_{[a+h,a+2h]}}(a+2h)\right).$$

It is now straightforward to identify the matrix equation for the only unknown quantity A_1:

$$A_1 = \frac{1}{h}\left[f\left(S_{\big|_{[a,a+h]}}(a+h) + S'_{\big|_{[a,a+h]}}(a+h)h + \frac{1}{2}S''_{\big|_{[a,a+h]}}(a+h)h^2+\frac{1}{6}A_1h^3, S'_{\big|_{[a,a+h]}}(a+h) + S''_{\big|_{[a,a+h]}}(a+h)h + \frac{1}{2}A_1h^2\right) - S''_{\big|_{[a,a+h]}}(a+h)\right]. \quad (9)$$

Assuming again that the matrix equation (9) has only the solution A_1, the spline of subinterval $[a + h, a + 2h]$ is fully determined. From the explanation of the previous steps, it should be clear how to generalize this iteration process for all subsequent subintervals up to the last interval of the partition. Without any loss of generality, let us consider the cubic matrix spline in an arbitrary subinterval $[a + (k - 1)h, a + kh]$. Then, for the subsequent subinterval $[a + kh, a + (k + 1)h]$, we can define the corresponding spline in analogy to the steps before:

$$S_{\big|_{[a+kh,a+(k+1)h]}}(x) = \beta_k(x) + \frac{1}{3!}A_k(x-(a+kh))^3, \quad (10)$$

where $\beta_k(x) \;=\; \sum_{l=0}^{2}\frac{1}{l!}S^{(l)}_{\big|_{[a+(k-1)h,a+kh]}}(a+kh)(x-(a+kh))^l$. With definition (10), the cubic matrix spline is of class $C^2\left(\bigcup_{j=0}^{k}[a+jh,a+(j+1)h]\right)$ and satisfies the differential equation (1) at point $x = a+kh$. Additionally, we require that $S(x)$ is a solution of the differential equation (1) at point $x = a+(k+1)h$, so that the equation for the unknown matrix A_k is

$$A_k = \frac{1}{h}\left[f\left(\beta_k(a+(k+1)h) + \frac{1}{6}A_kh^3, \beta'_k(a+(k+1)h) + \frac{1}{2}A_kh^2\right) - \beta''_k(a+(k+1)h)\right]. \quad (11)$$

Note that (11) is a general result for A_k and reduces to (7) and (9) for $k = 0$ and $k = 1$, respectively. It remains to be shown that (11) yields a unique solution A_k of the problem. Existence and uniqueness can be demonstrated by a fixed-point argument (see [5] for the first-order case). For this purpose, we consider the matrix function $g : \mathbb{C}^{r\times s} \to \mathbb{C}^{r\times s}$ for a fixed step size h defined by

$$g(T) = \frac{1}{h}\left[f\left(\beta_k(a+(k+1)h) + \frac{1}{6}Th^3, \beta_k'(a+(k+1)h) + \frac{1}{2}Th^2\right) - \beta_k''(a+(k+1)h)\right]. \tag{12}$$

Obviously (11) will only be valid if and only if $A_k = g(A_k)$, and as a consequence A_k is the fixed-point solution of function $g(T)$. Applying Lipschitz's conditions (2)–(12) yields

$$\|g(T_1) - g(T_2)\| \leq \left(\frac{L_1h^2}{6} + \frac{L_2h}{2}\right)\|T_1 - T_2\|.$$

For $h < \left(\sqrt{24L_1 + 9L_2^2} - 3L_2\right)/2L_1$, it follows $\left(L_1h^2/6 + L_2h/2\right) < 1$ and hence $g(T)$ is a contractive matrix function. Therefore (11) has unique solutions A_k for $k = 0, 1, \ldots, n-1$. This completes the uniqueness proof and the cubic matrix spline is completely determined. In summary, the following result can been established (see also [3]):

Theorem 1. *Let L_1, L_2 be Lipschitz constants defined by (2). If step size $h < \left(\sqrt{24L_1 + 9L_2^2} - 3L_2\right)/2L_1$ is chosen, then there exists a matrix-cubic spline $S(x)$ for each subinterval $[a+kh, a+(k+1)h]$, $k = 0, 1, \ldots, n-1$. If $f \in C^1(T)$, then $\|Y(x) - S(x)\|$ is, at least, of global order $O(h^2)$ $\forall x \in [a, b]$, where $Y(x)$ is the theoretical solution of* (1).

3 Algorithm

The following algorithm implements the approximate solution of system (1) in the interval $[a, b]$ with an error, at least, of global order $O(h^2)$.

- **Step 1.** Let L_1, L_2 be Lipschitz constants defined by (2). Take

$$n > \frac{2(b-a)L_1}{\sqrt{24L_1 + 9L_2^2} - 3L_2}, \quad h = (b-a)/n, \tag{13}$$

  and the partition $\Delta_{[a,\,b]}$ given for partition (3).
- **Step 2.** For $k = 0$, solve the matrix equation (7). Compute $S_{\big|_{[a,a+h]}}(x)$ as defined in (4).
- **Step 3.** For $k = 1, \ldots, n-1$, solve the matrix equation (11). Compute $S_{\big|_{[a+kh,a+(k+1)h]}}(x)$ as defined in (10).

Only for some exceptional cases (7) and (11) can be solved analytically [9]. Otherwise they can be tackled with standard iterative methods (see e.g. [10]) using $T_{l+1}^q = g(T_l^q)$, where T_0^q is an arbitrary matrix in $\mathbb{C}^{r\times s}$ and $q = 0, 1, \ldots, n-1$. Note that matrix function $g(T)$ is given by (12).

4 Example: Incomplete Second-Order Differential System

It is well known that the initial problem

$$Y''(t) + A^2 Y(t) = 0, Y(0) = Y_0, Y'(0) = Y_1 \ , \tag{14}$$

has the exact solution $Y(t) = \cos(At)Y_0 + (A)^{-1}\sin(At)Y_1$. A major disadvantage of this formal solution is the non-trivial computation of $\cos(At)$ and $\sin(At)$. Our proposed method avoids these difficulties. In this example we choose the parameters $A = \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix}, Y_0 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, Y_1 = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, t \in [0, 1]$, so that the exact solution of (14) is $Y(t) = \cos\left[\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} t\right] = \begin{pmatrix} \cos(t) & -t\sin(t) \\ 0 & \cos(t) \end{pmatrix}$. In this case, we have $L_1 \approx 2.82843$ and $L_2 = 0$, and according to Theorem 1 we need to take $h < 1.45647$. As in [6], we choose $h = 0.1$. To obtain solutions for the algebraic equations which arise from the algorithm, we employ *Mathematica*. Table 1 displays the corresponding numerical estimates (rounded to the fourth relevant digit). The maximum error for each subinterval is indicated in the third column of the table. In Fig. 1, we list the maximum error margins of the incomplete second-order differential system (14) for the interval $[0, 1]$ with step sizes $h = 0.1$, $h = 0.01$ and $h = 0.001$, respectively.

Table 1. Approximation for the incomplete second-order differential system (14) in the interval $[0, 1]$ with step size $h = 0.1$

Interval	Approximation	Error
$[0, 0.1]$	$\begin{pmatrix} 1.-0.5x^2+0.0083x^3 & -1.x^2+0.0333x^3 \\ 0 & 1.-0.5x^2+0.0083x^3 \end{pmatrix}$	1.7601×10^{-5}
$[0.1, 0.2]$	$\begin{pmatrix} 0.999983+0.0005x-0.5050x^2+0.0249x^3 & -0.00006+0.0020x-1.0198x^2+0.0993x^3 \\ 0 & 0.99998+0.0005x-0.5050x^2+0.0249x^3 \end{pmatrix}$	6.9897×10^{-5}
$[0.2, 0.3]$	$\begin{pmatrix} 0.999853+0.0025x-0.5148x^2+0.0412x^3 & -0.0006+0.0097x-1.0585x^2+0.1638x^3 \\ 0 & 0.9999+0.0025x-0.5148x^2+0.0412x^3 \end{pmatrix}$	1.5537×10^{-4}
$[0.3, 0.4]$	$\begin{pmatrix} 0.9994+0.0067x-0.5291x^2+0.0571x^3 & -0.0023+0.0265x-1.1144x^2+0.2258x^3 \\ 0 & 0.9994+0.0067x-0.5291x^2+0.0571x^3 \end{pmatrix}$	2.7154×10^{-4}
$[0.4, 0.5]$	$\begin{pmatrix} 0.9984+0.0141x-0.5475x^2+0.0724x^3 & -0.0060+0.0546x-1.1848x^2+0.2845x^3 \\ 0 & 0.9984+0.0141x-0.5475x^2+0.0724x^3 \end{pmatrix}$	4.1500×10^{-4}
$[0.5, 0.6]$	$\begin{pmatrix} 0.9966+0.0251x-0.5694x^2+0.0870x^3 & -0.0128+0.0954x-1.2663x^2+0.3389x^3 \\ 0 & 0.9966+0.0251x-0.5694x^2+0.0870x^3 \end{pmatrix}$	5.8146×10^{-4}
$[0.6, 0.7]$	$\begin{pmatrix} 0.9937+0.03989x-0.5941x^2+0.1007x^3 & -0.0235+0.1486x-1.3550x^2+0.3881x^3 \\ 0 & 0.9937+0.03989x-0.5941x^2+0.1007x^3 \end{pmatrix}$	7.6589×10^{-4}
$[0.7, 0.8]$	$\begin{pmatrix} 0.9893+0.0586x-0.6208x^2+0.1135x^3 & -0.0383+0.2124x-1.4461x^2+0.4315x^3 \\ 0 & 0.9893+0.0586x-0.6208x^2+0.1135x^3 \end{pmatrix}$	9.6259×10^{-4}
$[0.8, 0.9]$	$\begin{pmatrix} 0.9833+0.0809x-0.6486x^2+0.1251x^3 & -0.0572+0.2831x-1.5345x^2+0.4683x^3 \\ 0 & 0.9833+0.0809x-0.6486x^2+0.1251x^3 \end{pmatrix}$	1.1653×10^{-3}
$[0.9, 1]$	$\begin{pmatrix} 0.9758+0.1060x-0.6766x^2+0.1354x^3 & -0.0788+0.3551x-1.6145x^2+0.4980x^3 \\ 0 & 0.9758+0.1060x-0.6766x^2+0.1354x^3 \end{pmatrix}$	1.3674×10^{-3}

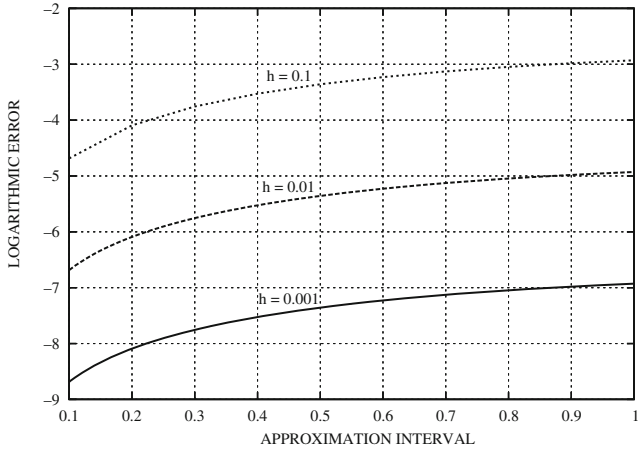


Fig. 1. Maximum error margins for the incomplete second-order differential system (14) in the interval $[0, 1]$ with step size $h = 0.1$, $h = 0.01$ and $h = 0.001$

Acknowledgment

This work has been partially supported by the Generalitat Valenciana GV/2007/009 and PAID-06-07/3283 of the Universidad Politécnica de Valencia.

References

1. Zhang, J.F.: Mech. Syst. Signal Process. **16**(1), 61–67 (2002)
2. Coddington, E.A., Levinson, N.: Theory of Ordinary Differential Equations. McGraw-Hill, New York (1955)
3. Loscalzo, F.R., Talbot, T.D.: SIAM J. Numer. Anal. **4**(3), 433–445 (1967)
4. Defez, E., Soler, L., Hervás, A., Santamaría, C.: Comput. Math. Appl. **50**, 693–699 (2005)
5. Defez, E., Soler, L., Hervás, A., Tung, M.M.: Comput. Modelling **46**(5–6), 657–669 (2007)
6. Tung, M.M., Soler, L., Defez, E., Hervás, A.: Cubic-matrix splines and second-order matrix model. In: Bonilla, L.L., Moscoso, M.A., Platero, G., Vega, J.M. (eds.) Progress in Industrial Mathematics at ECMI 2006. Mathematics in Industry, vol. 12, pp. 897–901. Springer, Berlin (2007)
7. Tung, M.M., Defez, E., Sastre, J.: Comput. Math. Appl. **56**, 2561–2571 (2008)
8. Flett, T.M.: Differential Analysis. Cambridge University Press, Cambridge (1980)
9. Lancaster, P.: Explicit solutions of linear matrix equations. SIAM Rev. **12**, 544–566 (1970)
10. Ortega, J.M., Rheinboldt, W.C.: Iterative Solution of Nonlinear Equations in Several Variables. Academic, New York (1972)

The Mathematical Model of the Pan-Tilt Unit Used in Noise Measurements in Urban Traffic

O.A. Detesan, M. Arghir, and G. Solea

Department of Applied Mechanics and Computer Programming, Technical University of Cluj-Napoca, Cluj-Napoca, B-dul Muncii 103-105, Romania, Ovidiu.Detesan@mep.utcluj.no, Mariana.Arghir@mep.utcluj.no, marylandprod@gmail.com

Summary. One of the most important aspects in urban agglomerations is to find the right way to monitor the sound polluters, such as the surface traffic. The paper presents the geometric and kinematic model of the Pan-Tilt Unit (PTU) used as an orientation device of a digital camcorder, in urban noise measurements. Using programs written in Matlab, the authors find the symbolic equations of the mathematical model, useful in the motion control of the PTU, with the purpose of orienting the camera according to the environmental requirements.

1 Introduction

The environmental noise becomes a worldwide problem in the last years. It is estimated that over 250 million European people live or work in areas where the surrounding noise has an unacceptable level [1]. No matter the source of this noise (road traffic, airports, building sites, industrial plants), the efforts to diminish or eliminate these sound polluters are increasing.

One of the methods used to measure the sound level in connection to the urban traffic is to record the flux of vehicles in well established points of the artery, in the principal moments of the day, using a camera. The recorded sounds and images can be online or offline analyzed, resulting key data about the traffic and the noise level it generates.

2 The Pan-Tilt Unit and the Camera

The camera orientation device, Pan-Tilt Unit PTU-46-17.5 [2] is supplied by Directed Perception, Inc., California and it is suitable for the following applications: computer vision and robotics, security and surveillance, industrial automation, tracking, laser ranging, teleconference and distance learning, antenna support, photo, video and special effects. It has the following general features: precise on-the-fly control of position, speed and acceleration,

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*, Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_127,

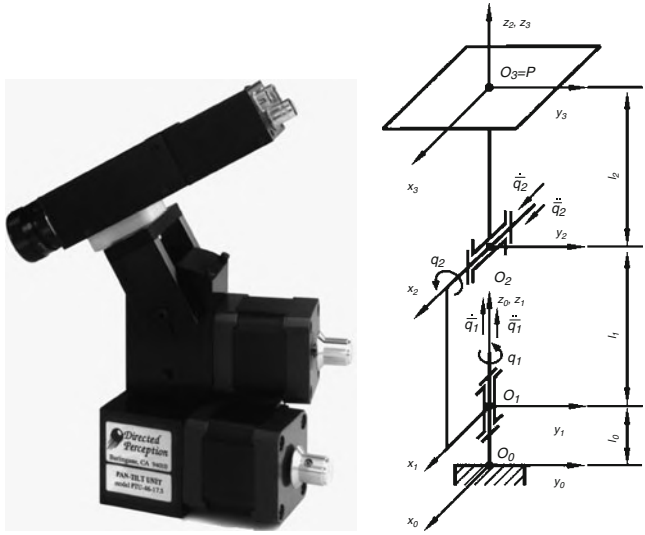


Fig. 1. The Pan-Tilt Unit PTU-46-17.5 [2] and its kinematic diagram

small form factor, self calibration upon reset. The Pan-Tilt characteristics (for the model PTU-46-17.5) are: speeds to 300°/s, resolution to 0.0051428°, load capacity to 1.81 kg, tilt range (approx): minimum 31° up and 47° down with option of 80° down, pan range (approx): ±159° with option of ±180°.

3 The *Robot_Symbolic* Application

The *Robot_Symbolic* software package [3] is a generalized application for symbolic modelling of robots and manipulators of any structure and number of degrees of freedom. It is written in Matlab, exploiting its symbolic computation libraries and it consists of the following main modules: *Robot_definition*, *Robot_geometry*, *Robot_kinematics* and *Robot_dynamics*. They represent a Matlab implementation of the following methods and algorithms [4]: the rotation matrices method, the iterative method of kinematics and the Newton-Euler formulation.

4 The Mathematical Model of the Pan-Tilt Unit

The kinematic diagram of the Pan-Tilt Unit is represented in Fig. 1, along with the following notations:

l_0, l_1, l_2 – the geometrical parameters of the unit
q_1, q_2 – the generalized coordinates from the joints of the unit
$\dot{q}_1$, $\dot{q}_2$ – the generalized velocities
$\ddot{q}_1$, $\ddot{q}_2$ – the generalized accelerations

The Pan-Tilt Unit can be regarded as a manipulator with two degrees of freedom. The first step in using the *Robot_Symbolic* application is defining the mechanical structure of the unit. The name of the device (PTU), the number and the type of joints (2R), the position and orientation of the attached frames are defined, by calling the *Robot_definition* function from the General Model menu. The program generates the position vectors and the rotation matrices relating two adjacent frames $\{i\}$ and $\{i-1\}$.

4.1 The Geometrical Model

The equations of the geometric model are determined by calling the *Robot-geometry* function, yielding (1) and (2) which express the position and the orientation of the frame attached to the mobile platform of the unit, with respect to the base frame 0.

$$\bar{p}_3 = \begin{bmatrix} l_2 sq_1 sq_2 \\ -l_2 cq_1 sq_2 \\ l_0 + l_1 + l_2 cq_2 \end{bmatrix} \tag{1}$$

$${}^0_3[R] = \begin{bmatrix} cq_1 & -sq_1 cq_2 & sq_1 sq_2 \\ sq_1 & cq_1 cq_2 & -cq_1 sq_2 \\ 0 & sq_2 & cq_2 \end{bmatrix} \tag{2}$$

The Euler angles, as generated by the *Robot_Symbolic* application and considering the set of rotation angles about mobile axes ($\alpha_z - \beta_x - \gamma_z$), are the following:

$$\alpha_z = q_1;\ \beta_x = q_2;\ \gamma_z = \pi/2. \tag{3}$$

Therefore, (1) and (2) or (1) and (3) express the geometric model of the PTU, namely the position and orientation of the mobile platform with respect to the generalized coordinates q_1, q_2.

4.2 The Kinematic Model

The equations of the kinematic model of the PTU are generated using the *Robot_kinematics* function, obtaining the linear and angular operational velocities (4) and (5), also the linear and angular operational accelerations (6) and (7).

$${}^0\bar{v}_3 = l_2 \cdot \begin{bmatrix} cq_1 sq_2 \dot{q}_1 + sq_1 cq_2 \dot{q}_2 \\ sq_1 sq_2 \dot{q}_1 - cq_1 cq_2 \dot{q}_2 \\ -sq_2 \dot{q}_2 \end{bmatrix} \tag{4}$$

$${}^0\bar{\omega}_3 = \begin{bmatrix} cq_1 \dot{q}_2 \\ sq_1 \dot{q}_2 \\ \dot{q}_1 \end{bmatrix} \tag{5}$$

$$ {}^0\dot{\bar{v}}_3 = l_2 \cdot \begin{bmatrix} cq_1 sq_2 \ddot{q}_1 + 2cq_1 cq_2 \dot{q}_1 \dot{q}_2 + sq_1 cq_2 \ddot{q}_2 - \left(\dot{q}_1^2 + \dot{q}_2^2\right) sq_1 sq_2 \\ sq_1 sq_2 \ddot{q}_1 + 2sq_1 cq_2 \dot{q}_1 \dot{q}_2 - cq_1 cq_2 \ddot{q}_2 + \left(\dot{q}_1^2 + \dot{q}_2^2\right) cq_1 sq_2 \\ -sq_2 \ddot{q}_2 - cq_2 \dot{q}_2^2 \end{bmatrix} \tag{6} $$

$$ {}^0\dot{\bar{\omega}}_3 = \begin{bmatrix} cq_1 \ddot{q}_2 - sq_1 \dot{q}_1 \dot{q}_2 \\ sq_1 \ddot{q}_2 + cq_1 \dot{q}_1 \dot{q}_2 \\ \ddot{q}_1 \end{bmatrix}. \tag{7} $$

Equations (4)–(7) express the kinematic model of the PTU. They characterize the motion of the mobile platform with respect to the fixed frame $\{0\}$.

5 The Simulation of the Pan-Tilt Unit Behaviour

In order to simulate the Pan-Tilt Unit, the symbolic data of the geometric and kinematic models are loaded (file *PTU_kin.mat*). The geometric elements of the device are defined numerically as follows:

$$ \begin{aligned} \mathrm{l}_0 &= 45.72 \text{ mm} \\ \mathrm{l}_1 &= 46.23 \text{ mm} \\ \mathrm{l}_2 &= 39.12 \text{ mm} \end{aligned} \tag{8} $$

To establish a sequence of orientations of the unit, a number $ncfg = 5$ distinct configurations were considered, defined by relative displacements in the two joints (pan and tilt). To each configuration a moment in time (expressed in seconds) was associated, as shown in Table 1.

The generalized variables were cubically spline interpolated and numerically derived, resulting the generalized velocities and accelerations. The symbolic data of the mathematical model was furthermore processed by numerical substitutions with data from the technical documentation of the device [2], resulting vectors representing the position of the characteristic point of the mobile platform on each three Cartesian axes, the module of the operational velocities (linear and angular) and the module of the operational accelerations (linear and angular). The following results were graphically represented:

- The generalized coordinates *qi* and velocities *qidot* (Fig. 2)

Table 1. Distinct configurations by relative displacements in joints

Nr. cfg. $i = \overline{1, ncfg}$	Time $t(i)$ (s)	Pan (degree)	Tilt (degree)
1	0	0	0
2	30	−15	−10
3	50	45	25
4	80	−30	5
5	100	15	−10

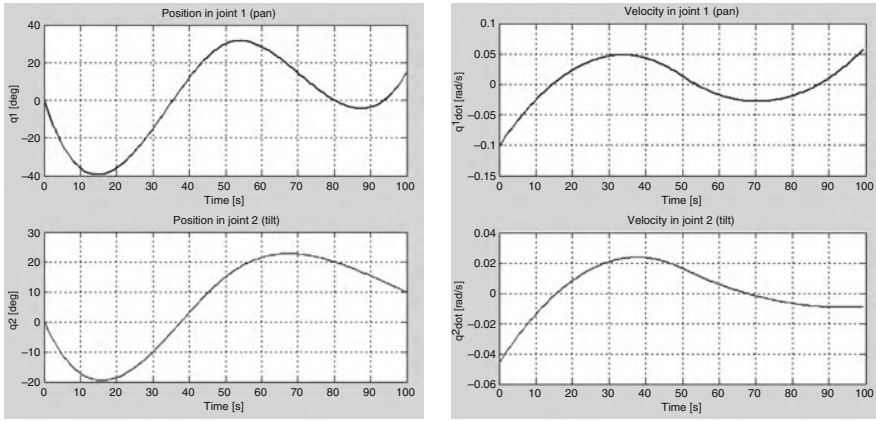


Fig. 2. The generalized coordinates and velocities

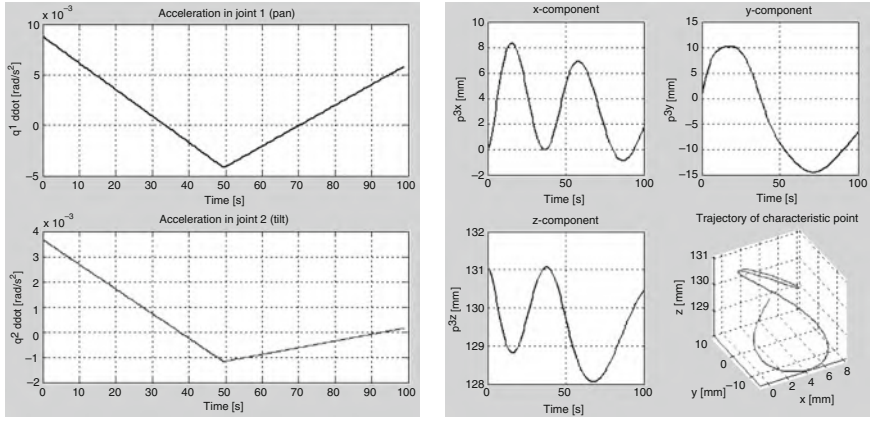


Fig. 3. The generalized accelerations and the position of the characteristic point

- The generalized accelerations *qiddot* and the position of the characteristic point of the platform, on each of the three axes (*p3x, p3y, p3z*) with respect to the time and in the 3D space (Fig. 3)
- The operational velocities and accelerations, linear and angular (Fig. 4)

6 Conclusions

The paper presents the mathematical model of the Pan-Tilt Unit PTU-46-17.5 used as an orientation device for a Sony HDD Handycam DCR-SR30 with the purpose of measuring the urban traffic noise. Using the *Robot_Symbolic* application, the authors determine the geometric and kinematic models of the PTU, necessary to simulate the behaviour of the PTU-camera assembly.

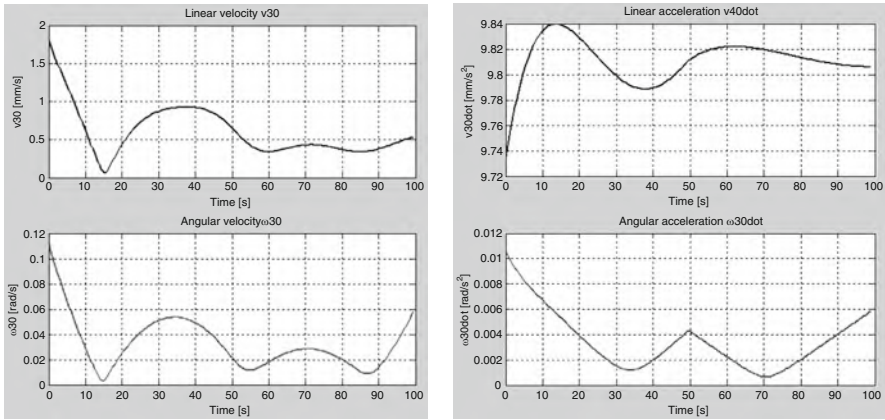


Fig. 4. The operational velocities and accelerations

Given the geometrical elements of the PTU, a sequence of five orientations of the device was imposed, associated with a time vector, whose numerical values were substituted into the symbolic equations of the models, generating numerical and graphical results about the geometric and kinematic behaviour of the PTU.

Acknowledgements

The paper was developed with the support of the project "Modern Concepts According to the Specific European Rules Regarding the Surface Transport Noise Greening," led by The Technical University of Cluj-Napoca, Romania.

References

1. Arghir, M., Ispas, V., Stoian, I., Blaga, F., Borzan, C.: The Surface Transport Greening in Urban Agglomerations. EDP, Bucharest (2008)
2. Computer-Controlled Pan-Tilt Unit. Models PTU-46-17.5 and PTU-46-70, Technical Specifications, Directed Perception, Inc., http://www.DPerception.com
3. Detesan, O.A.: Research Regarding the Modeling, Simulation and Building of Minirobots. Ph.D. Thesis, Technical University of Cluj-Napoca (2007)
4. Negrean, I.: Kinematics and Dynamics of Robots – Modelling, Experiment, Accuracy. EDP, Bucharest (1999)

Spread of Epidemics and Rumours with Mobile Agents

M. Draief[1] and A. Ganesh[2]

[1] Department of Electrical and Electronic Engineering, Imperial College London Exhibition Road, London SW7 2AZ, UK, m.draief@imperial.ac.uk
[2] Department of Mathematics, University of Bristol, University Walk, Clifton, Bristol BS8 1TW, UK, a.ganesh@bristol.ac.uk

Summary. We propose a simple model of infection that enables to study the coincidence time of two random walkers on an arbitrary graph. By studying the coincidence time of a susceptible and an infected individual both moving in the graph we obtain estimates of the infection probability. The main result of this paper is to pinpoint the impact of the network topology on the infection probability.

1 Introduction

In much of the literature on mathematical epidemiology, the members of the population are assumed to occupy fixed locations and the probability of infection passing between a pair of them in a fixed time interval is taken to be some function of the distance between them. Mean-field models are a special case in which this function is a constant [5]. In this work, we consider a different model in which the agents are mobile and can only infect each other if they are in sufficiently close proximity. The model is motivated both by certain kinds of biological epidemics, whose transmission may be dominated by sites at which individuals gather in close proximity (e.g. workplaces or public transport for a disease like SARS, cattle markets for foot-and-mouth disease, etc.) and by malware spreading between wireless devices via Bluetooth connections, for example.

To our knowledge the first attempts to model virus spreading in mobile networks relies on the use of a non-rigorous mean-field approximations that incorporate the mobility patterns of users. In [10], the authors derive a threshold for the persistence of the epidemic by computing the average number of neighbours of a given node. Using a similar approach but with different mobility patterns, Nekovee et al. [11,13] explore the evolution of the number of devices that are infected in terms of the contact rate between users. A related line of work studying the dissemination of information in opportunistic networks [3] focuses on the following analogous problem: Suppose that all

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*, Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_128,

individuals are interested in a piece of information that is initially held by one user. The information is transmitted between users who happen to be close to each other. As in the case of static networks [12], one may be interested in the time it takes for the rumour to be known to all users. To this end we need to understand how information is transmitted between an informed and an ignorant user. Our work gives some insight on the impact of the network structure on the likelihood of successfully transmitting the rumour.

2 Models and Results

We consider a simple mathematical model of the spread of infection as follows. There is a finite, connected, undirected graph $G = (V, E)$ on which the individuals perform independent random walks: they stay at each vertex for an exponentially distributed time with unit mean, and then move to a neighbour of that vertex chosen uniformly at random. The infection can pass from an infected to a susceptible individual only if they are both at the same vertex, and the probability of its being passed over a time interval of length τ is $1 - \exp(-\beta\tau)$, where $\beta > 0$ is a parameter called the infection rate. We shall consider a single infected and a single susceptible individual and ask what the probability is that the susceptible individual becomes infected by time t. This probability has been studied in the case of a complete graph in [6]. Here, we extend their results to a much wider class of graphs.

It is simplistic to consider just a single infective and a single susceptible individual. Nevertheless, insights gained from this setting are relevant in the "sparse" case, where the number of both infected and susceptible individuals is small and inter-contact times are fairly large.

We now describe the model precisely. Let $X_t, Y_t \in V$ denote the positions of the susceptible and infected individuals respectively at time t. We model $(X_t, t \geq 0)$ and $(Y_t, t \geq 0)$ as independent continuous-time Markov chains (CTMCs) on the finite state space V, with the same transition rate. We define the coincidence time up to time t, denoted $\tau(t)$, as the total time up to t during which both walkers are at the same vertex, i.e.,

$$\tau(t) = \int_0^t \mathbf{1}_{(X_s = Y_s)} ds. \tag{1}$$

Let $\gamma(t)$ denote the probability that the initial susceptible becomes infected by time t. Then, conditional on $\tau(t)$, we have

$$\gamma(t) = 1 - \exp(-\beta\tau(t)), \tag{2}$$

where $\beta > 0$ is the infection rate. We are interested in estimating the coincidence time $\tau(t)$ and the infection probability $\gamma(t)$ for different families of graphs. Observe that the Markov chains X_t, Y_t have invariant distribution π given by

$$\pi_x = \frac{\text{degree}(x)}{\sum_{v \in V} \text{degree}(v)} \tag{3}$$

and that they are reversible, i.e., $\pi_x q_{xy} = \pi_y q_{yx}$ for all $x, y \in V$. We consider the case when these chains are started independently in the stationary distribution and provide estimates on the coincidence time and the infection probability, for arbitrary graphs. A direct computation yields

Theorem 1. *Suppose X_0 and Y_0 are chosen independently according to the invariant distribution π. Then, we have*

$$\mathbb{E}[\tau(t)] = \sum_{v \in V} \pi_v^2 t, \quad \text{and} \quad \mathbb{E}[\gamma(t)] \leq 1 - \exp\Big(-\beta t \sum_{v \in V} \pi_v^2\Big).$$

3 Examples of Graphs

We present models of networks of interest to which we are going to apply the result of Theorem 1.

3.1 Complete Graphs

Consider the complete graph on n nodes, namely the graph in which there is an edge between every pair of nodes, $\text{degree}(v) = n - 1$ and $\pi_v = 1/n$ for all $v \in V$, so we have by Theorem 1 that $\mathbb{E}[\tau(t)] = t/n$. This result should be intuitive by symmetry. Theorem 1 also gives us an upper bound on the infection probability, $\mathbb{E}[\gamma(t)] \leq 1 - \exp(-\beta t/n)$. Roughly speaking, this says that it takes time of order n/β for the susceptible individual to become infected; for $t \ll n/\beta$, the probability of being infected is vanishingly small.

3.2 Regular Graphs

A graph $G = (V, E)$ is said to be r-regular if $\text{degree}(v) = r$ for all $v \in V$, so that $\pi_v = 1/n$ for all $v \in V$ if G is for any $r \geq 2$. Hence, if G is connected, we have the same estimates for $\tau(t)$ and $\gamma(t)$ as for the complete graph, which is a special case corresponding to $r = n - 1$.

The next examples we consider will be families of random graphs widely used in practice to model networks.

3.3 Erdős-Rényi Random Graphs

The Erdős-Rényi graph $G(n, p)$ is defined as a random graph on n nodes, wherein each edge is present with probability p, independent of all other edges. Let p to be a function of n chosen so that $np > c \log n$ for some constant $c > 1$ ensuring that the graph is almost surely connected. In this model, the node degrees concentrate around the mean value np, and have exponentially decaying tails away from this value. Thus, while Erdős-Rényi graphs are not exactly regular, they exhibit considerable homogeneity in node degrees.

3.4 Power Law Random Graphs

In contrast to the above graph models, many real-world networks exhibit considerable heterogeneity in node degrees, and have empirical degree distributions whose tails decay polynomially; see, e.g., [1, 8]. This observation has led to the development of generative models for graphs with power-law tails [1, 2] as well as random-graph models possessing this property [4]. For definiteness, we work with the model proposed in [4], but we believe that similar results will hold for the other models as well.

In the model of [4], each node v is associated with a positive weight w_v, and edges are present independently with probabilities related to the weights by

$$\mathbb{P}((u,v) \in E) = \frac{w_u w_v}{W} \text{ where } W = \sum_{x \in V} w_x. \tag{4}$$

We assume that $W \geq w_{\max}^2$, so that the above defines a probability. It can be verified that $\mathbb{E}[\text{degree}(v)] = w_v$ and so this model is also referred to as the expected degree model. If the weights are chosen to have a power-law distribution, then so will the node degrees. The following 3-parameter model for the ordered weight sequence is proposed in [4], parametrised by the mean degree d, the maximum degree m, and the exponent $\gamma > 2$ of the weight distribution:

$$w_i = m\Big(1 + \frac{i}{i_0}\Big)^{-\frac{1}{\gamma-1}}, \quad i = 0, 1, \ldots, n-1, \tag{5}$$

where

$$i_0 = n\Big(\frac{d(\gamma-2)}{m(\gamma-1)}\Big)^{\gamma-1}. \tag{6}$$

Note that $W = \sum_{i=0}^{n-1} w_i \sim nd$, for n large.

We consider a sequence of such graphs indexed by n. The maximum expected degree m and the average expected degree d may, and indeed typically will, depend on n. In models of real networks, we can typically expect d to remain bounded or to grow slowly with n, say logarithmically, while m grows more quickly, say as some fractional power of n. In this paper, we only assume the following:

$$d \geq \delta > 0, \quad d = o(m), \quad m \leq \sqrt{nd}, \quad \frac{m}{d} = o\Big(n^{\frac{1}{\gamma-1}}\Big). \tag{7}$$

Here, δ is a constant that does not depend on n. In other words, the average expected degree is uniformly bounded away from zero. The third assumption simply restates the requirement that $w_0^2 \leq W$, so that (4) defines valid probabilities. The last assumption ensures that i_0, defined in (6), tends to infinity. We now describe our results about these models.

Theorem 2. *Consider a sequence of graphs $G = (V, E)$ indexed by $n = |V|$. On each graph, consider two independent random walks with initial condition*

X_0, Y_0 chosen independently from the invariant distribution π for the random walk on that graph.

We have $\mathbb{E}[\tau(t)] = t/n$ for regular graphs, including the complete graph, on n nodes.

For Erdős-Rényi random graphs $G(n,p)$ conditioned to be connected, and having $np \geq c \log n$ for some $c > 1$, we have $\mathbb{E}[\tau(t)] \sim t/n$, as n tends to infinity.

Finally, consider a sequence of power law random graphs defined via (4) and (5), and satisfying the assumptions in (7). Then, we have the following:

$$\frac{n\mathbb{E}[\tau(t)]}{t} \sim \begin{cases} c_1, & \text{if } \gamma > 3, \\ c_2(\log m), & \text{if } \gamma = 3, \\ c_3(m.d)^{3-\gamma}, & \text{if } 2 < \gamma < 3, \end{cases}$$

where $c_1, c_2, c_3 > 0$ are constants that do not depend on n, m or d.

The proof is rather long involving the computation of moments of the degree distributions and using concentration results. For lack of space it is omitted (see [7]).

4 Conclusion and Further Work

In this work we have presented a simple model for the spread of epidemics where individuals are mobile. In this framework we were interested in the setting where there are two individuals one infected and one healthy both performing random walks on the network. Our preliminary investigation highlights the effect of the topology on the spread of an epidemic, motivated by networking phenomena such as worms and viruses, failures, and dissemination of information. Under this natural model, we provided an explicit relationship between the structure over which the walks are performed and the coincidence time of the two walkers. To this end we analysed both homogeneous (regular, complete and Erdös-Rényi graphs) and heterogeneous (power-law graphs) networks. We pinpointed the existence of a phase transition for the coincidence time in the case of power-law networks depending on the parameter of the power-law degree distribution. We also derived bounds on the probability of infection.

As a final remark, we propose some several interesting directions to pursue the work presented here. In our present model individuals are supposed to start their walks in stationary regime. This can be relaxed since the networks we study are expanders and thus random walks on such networks have nice mixing properties as illustrated in [9] through the computation of the isoperimetric constant of the underlying graphs. We also anticipate that similar results can be derived when considering k walkers as long as k is small with respect to n the number of sites in the network.

References

1. Barabási, A.-L., Albert, R.: Emergence of scaling in random networks. Science **286**, 509–512 (1999)
2. Bollobás, B., Riordan, O.: The diameter of a scale-free random graph. Combinatorica **4**, 5–34 (2004)
3. Chaintreau, A., Hui, P., Scott, J., Gass, R., Crowcroft, J., Diot, C.: Impact of human mobility on opportunistic forwarding algorithms. IEEE Trans. Mobile Comput. **6**(6), 606–620 (2007)
4. Chung, F., Lu, L.: The average distances in random graphs with given expected degrees. Internet Math. **1**, 91–114 (2003)
5. Daley, D.J., Gani, J.: Epidemic Modelling: An Introduction (Cambridge Studies in Mathematical Biology). Cambridge University Press, Cambridge (2001)
6. Datta, N., Dorlas, T.C.: Random walks on a complete graph: a model for infection. J. Appl. Prob. **41**, 1008–1021 (2004)
7. Draief, M., Ganesh, A.: Spread of epidemics and rumours with mobile agents. Preprint (2008)
8. Faloutsos, M., Faloutsos, P., Faloutsos, C.: On power-law relationships of the Internet topology. In: Proceedings ACM SIGCOMM, 1999
9. Ganesh, A., Massoulié, L., Towsley, D.: The effect of network topology on the spread of epidemics. In: Proceedings of the IEEE INFOCOM, 2005
10. Mickens, J.W., Noble, B.D.: Modeling epidemic spreading in mobile environments. In: Proceedings of the 4th ACM Workshop on Wireless Security, pp. 77–86, 2005
11. Nekovee, M.: Worm epidemics in wireless ad hoc networks. New J. Phys. **9**, 189 (2007)
12. Pittel, B.: On spreading a rumor. SIAM J. Appl. Math. **47**(1), 213–223 (1987)
13. Rhodes, C.J., Nekovee, M.: The opportunistic transmission of wireless worms between mobile devices. arXiv arXiv:0802.2685v1 (2008)

A Two-Layer Algebraic Turbulence Model for Compressible Flow in Turbomachinery Cascade

A. Dumitrache[1], H. Dumitrescu[1], and F. Frunzulica[1,2]

[1] Institute of Mathematical Statistics and Applied Mathematics, Calea 13 Septembrie no. 13, 050711 Bucharest, Romania,
alexandru.dumitrache@ima.ro, horia.dumitrescu@ima.ro
[2] "POLITEHNICA" University from Bucharest, Bucharest, Romania
ffrunzi@aero.pub.ro

Summary. The objective of this study is to examine the Baldwin-Lomax turbulence model in a finite volume solver to introduce a computer code for complex two-dimensional flows in turbomachinery. The turbulent model was tested with investigating the turbulent flow over a flat plate and other test cases. Three test cases are presented and the computed results are compared with experimental data. The calculated velocity profile agreed well with the experimental data in plate test case and the solver is validated in test case of flow over a semi NACA-0012 airfoil. The solver is used for flow through a multi-blade cascade of an axial compressor in design condition to show its capability of multi-block solution.

1 Introduction

The development of CFD methods has resulted in very useful analysis tools that are able to provide detailed information to enhance the understanding of complex flow physics at design and off-design conditions in compressor/turbine design [3]. The flow calculations have to be carried out on the basis of the averaged Navier-Stokes equations completed with transport equations for turbulence models. One of the groups of statistical turbulence models is the algebraic one or two-layer turbulence closure, but they require the determination of boundary layer parameters to calculate the eddy viscosity. In complex flow such as the flow through a turbine or compressor cascade, the calculation of e.g. shear layer thickness in a CFD code is difficult, because no realistic criterion can be used to define the edge of the boundary layer [2]. That is the specially the case when flow separation exists within the domain.

An algebraic model, which is not written in terms of the boundary layer quantities and is very robust in separated regions, is the standard Baldwin-Lomax (BL) model [1]. The model was modified by Granville [4] and used by

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*,
Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_129,

He [5] for pressure gradient effects. The objective of the present computational study was to examine the BL model in a finite volume solver to introduce a computer code for 2-D flows in turbomachinery studies based on Van Leers flux splitting methods with using of high order limiters.

2 Governing Equations

The integral form of the quasi-three dimensional unsteady Navier-Stokes equations over a moving finite area **A** is

$$\begin{array}{c} \frac{\partial}{\partial t}\int\int_{\triangle A} U dxdy + \oint_s [(F - Uu_g - V_x)\mathbf{n}_x + (G - Uv_g - V_y)\mathbf{n}_y] \\ = \int\int_{\triangle A} S dxdy \end{array} \tag{1}$$

where

$$\begin{aligned} U &= h\,[\rho \quad \rho u \quad \rho v r \quad \rho\rho e]^T, \\ F &= h\,[\rho \quad \rho uu + p \quad \rho uvr \quad (\rho e + p)u]^T, \\ G &= h\,[\rho \quad \rho uv \quad (\rho vv + p)r \quad (\rho e + p)u]^T, \\ S &= [0 \quad p\partial h/\partial x \quad 0 \quad 0]^T. \end{aligned}$$

and M^T denotes the transpose of a matrix M.

The quasi-three dimensional effects are introduced by allowing specified variations of r and h in the axial direction. Both u_g and v_g are the moving mesh grid velocities, to rotor blades and blade vibration [6, 7]. In present work, only the former is considered, thus u_g is zero and V_g is equal to the blade rotation velocity. V_x and V_y are the viscous terms:

$$V_x = h\,[0 \quad \tau_{xx} \quad r\tau_{xy} \quad -q_x + u\tau_{xx} + v\tau_{xy}]^T, \tag{2}$$

$$V_y = h[0 \quad \tau_{xy} \quad r\tau_{yy} \quad -q_y + u\tau_{xy} + v\tau_{yy}]^T \tag{3}$$

where

$$\tau_{xx} = \frac{2}{3}\mu\left(2\frac{\partial u}{\partial x} - \frac{\partial v}{\partial y}\right), \quad \tau_{yy} = \frac{2}{3}\mu\left(2\frac{\partial v}{\partial y} - \frac{\partial u}{\partial x}\right),$$

$$\tau_{xy} = \frac{2}{3}\mu\left(2\frac{\partial u}{\partial y} - \frac{\partial v}{\partial x}\right), \quad q_x = -k\frac{\partial T}{\partial x}, \quad q_y = -k\frac{\partial T}{\partial y}.$$

Flow calculations have to be carried out on the basis of the above averaged Navier-Stokes equations in conjunction with transport equations for BL turbulence closure.

2.1 Baldwin-Lomax Turbulence Model

The Baldwin-Lomax turbulence model is a relatively simple algebraic model that makes use of a two-layer diffusivity formulation

$$\mu_t = \begin{cases} \mu_{t_{inner}} & \text{if } y_n \leq y_{crossover} \\ \mu_{t_{outer}} & \text{if } y_n > y_{crossover} \end{cases}$$

where y_n is the normal distance to the wall and $y_{crossover}$ is the minimum value of the y_n at which $\mu_{t_{inner}} = \mu_{t_{outer}}$. In the inner layer, the eddy viscosity coefficient is defined as $\mu_{t_{inner}} = \rho l^2 \Omega$ where $l = k y_n [1 - \exp(y^+/A^+)]$ is the length scale of the turbulence in the inner region, k and A^+ are model constants, Ω is the magnitude of the vorticity,

$$\Omega = \left| \frac{\partial v}{\partial x} - \frac{\partial u}{\partial y} \right|$$

The wall factor is given by $y^+ = \dfrac{\sqrt{\rho_w \tau_w}}{\mu_w} y_n$ where, ρ_w, μ_w and τ_w are the density, molecular viscosity and laminar shear stress at the wall.

In the outer layer, the eddy viscosity is defined by $\mu_{t_{outer}} = K C_{cp} F_{wake} F_{kleb}$ where K and C_{cp} are the model constants, and the function F_{wake} is taken by

$$F_{wake} = \min \left(y_{max} F_{max}, C_{wk} y_{max} U_{max}^2 / F_{max} \right)$$

and

$$F_{kleb} = \left[1 + 5.5 \left(\frac{C_{kleb} y}{y_{max}} \right)^6 \right]^{-1}$$

Here, F_{max} is determined by the maximum value of the function $F = y_n \Omega [1 - \exp(-y^+/A^+]$ and y_{max} is the value of y_n at which this maximum occurs. Also, U_{max} is the maximum difference of the magnitude of the velocity in the profile.

The model constants are given by

$$k = 0.4, \quad A^+ = 26, \quad K = 0.0168$$

$$C_{cp} = 1.6, \quad C_{wk} = 1.0, \quad C_{kleb} = 0.8$$

Transition to turbulence can be modeled by setting a cut off value for the computed eddy diffusivity. The suggested criterion is $\mu_t = 0$, if $\mu_{max} < C_{mutm}$, $C_{mutm} = 14$. For use with multigrid, the turbulence viscosity is evaluated only on the fine mesh and frozen on all coaster meshes.

3 Numerical Schemes and Results

The Van-Leer's method is used for splitting the convective fluxes and central differencing is used for viscous terms. To improve the accuracy of convective terms the Universal Van Leers limiter is used in mid points of computation area and first order accuracy is used for boundary cells. The numerical

equations are used with explicit scheme because of its faster convergence history. The stability criteria, is based on characteristic values of Jacobian matrices of convective terms. The non-reflective boundary condition is used for downstream of computational domain to speed up the convergence process, especially in low Mach number flows. A multi-block algorithm is implemented for complicated geometries, to divide the area to multi-channels. Each channel is solved with using neighborhood channel boundary condition in jointed boundaries of channels to give the influence of adjacent zones in being solved area. All channels are solved alternatively form bottom to up and then time increases to new time step of solution. To minimize the convergence history the non-reflective boundary condition [9] is used for downstream of computational domain.

The validation of the implemented turbulence model is first done for the turbulent flow over a flat plate. The calculation are carried out with inviscid inlet boundary conditions located 100% of the plate length upstream of the leading edge with a Mach number of 0.2. The transition point is set to close to the leading edge. The Reynolds number for flow around flat plate is set to 10^6 based on plate length scale. The results give good agreement with experimental data of Wieghardt [10].

In the next test case, the subsonic viscous flow over a semi NACA-0012 airfoil is considered. The Reynolds number is set to 10^6 and the free upstream Mach number is set to 0.4. In the next test case, the transonic viscous flow over NACA-0012 airfoil with a 2° angle of attack is considered and the free stream Mach number is set to 0.82. Figure 1 illustrates the comparison of pressure coefficient distribution for upper and lower surface of airfoil with experimental results. The good agreement of computed results and the experimental data, gives the adequate assurance about the solver, which implements the Van Leers flux splitting scheme and the related high ordering limiter and the BL turbulence model in compressible flows in presence of pressure gradients and shock induced cases.

As the final case study, a two- dimensional cascade of an axial compressor with NACA-65(10) airfoil geometry for rotor and stator blades is considered. Figure 2 shows the pressure contours, of the stage when the rotor is located in face to face situation with respect to stator location. The Mach number is set to 0.2, because many instability phenomenas happens in low Mach numbers for transonic compressors, for example during starting process of a gas turbine engine. Figure 3 shows the pressure contours when the rotor is moved in rotation direction for a half pitch of blades row. The mesh points are clustered for boundary layer capturing. The mesh resolution is such set to have at least a numerical cell in viscous sub-layer of boundary layer in all test cases. This means the mesh resolution is adequate to capture turbulent behavior. The transition point is set at the near of leading edge of the blades in stage problem.

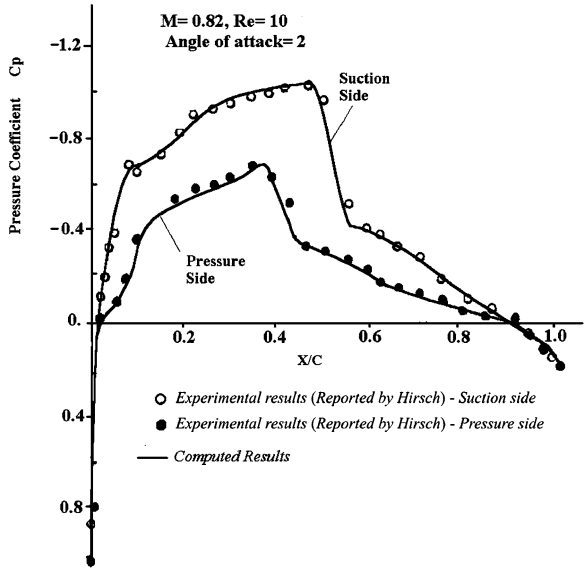


Fig. 1. Computed pressure coefficient over a NACA0012 airfoil at M = 0.82, Re = 10^6 and 2° angle of attack, compared with experimental results [8]

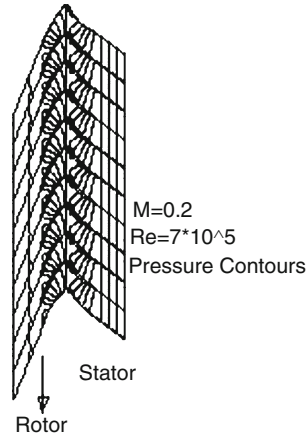


Fig. 2. Pressure contours for a 10-passage-stage of an axial compressor (no. of cells = 320,000)

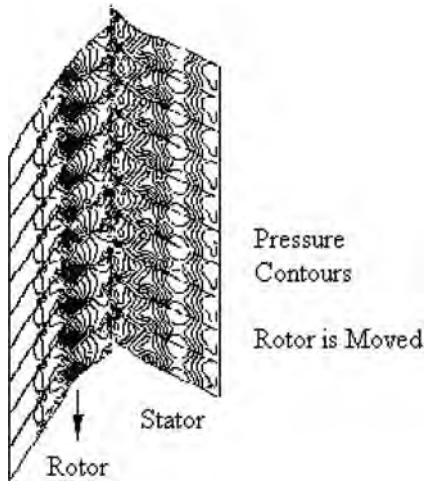


Fig. 3. Pressure contours for a 10-passage-stage of an axial compressor when the rotor is moved about half pitch of cascade

4 Conclusions

A computer code for numerical simulation of the 2-D inviscid and viscous flow in turbomachinery blade channels was presented. The turbulent model was tested with investigating the turbulent flow over a flat plate and other test cases. The calculated velocity profile agreed well with the experimental data in plate test case and the solver is validated in test case of flow over a semi NACA-0012 airfoil. The good agreement of pressure distribution with experimental data, in turbulent flow around NACA-0012 airfoil with angle of attack, gives the robustness of the code and the implemented schemes in transonic flows with the existence of adverse pressure gradients and the interaction of shock and boundary layer. Consequently, a solver with an algebraic turbulent modeling for compressible viscous-inviscid flows in complicated geometries is achieved, which may be useful in 2-D unsteady studies of turbomachinery investigations in off-design conditions.

References

1. Baldwin, B.S., Lomax, H.: AIAA Paper No. 78-257 (1978)
2. Bohn, D., Edmunds, R.: Proceedings of the International Gas Turbine and Aeroengine Congress and Exposition, Houston, 95-GT-90, 1995
3. Deardorff, J.W., J. Fluid Mech. **41**, 453–480 (1970)
4. Granville, P.S.: AIAA J. **25**, 1624–1627 (1987)
5. He, L., J. Comput. Phys. **36**, 55–70 (1980)
6. He, L.: J. Propul. Power **9**, 272–280 (1993)
7. He, L.: Int. J. Comput. Fluid Dyn. **3**, 217–231 (1994)

8. Hirsch, C.: Numerical Computation of Internal and External Flows: Fundamentals of Numerical Discretization, vol. 1. Wiley, New York (1988)
9. Rudy, D.H., Strikwerda, J.C.: J. Propul. Power **13**, 31–38 (1997)
10. Wieghardt, K., Tillmann, W.: Computation of turbulent boundary layers. In: Proceedings of AFOSR-IFP-Stanford Conference, 1968

Aerodynamic and Aeroacoustic Analysis of a HAWT in Yaw

H. Dumitrescu, A. Dumitrache, and V. Cardos

Institute of Mathematical Statistics and Applied Mathematics, Calea 13 Septembrie no. 13, 050711 Bucharest, Romania, horia.dumitrescu@ima.ro, alexandru.dumitrache@ima.ro, vladimir.cardos@ima.ro

Summary. Stall control and pitch control are the most commonly used methods of regulating power. However, through the opportunities presented by the flexible (or teetered) hub of a two-bladed teetered rotor one can also utilize yaw control to regulate power. This paper presents the aerodynamic and aeroacoustic results obtained from theoretical models for such a rotor when is yawed to the undisturbed flow. Some comparisons between calculated and measured noise spectra of a yaw controlled wind turbine show good agreement over all angles up to 60° of yaw.

1 Introduction

Small wind turbines, with rated power values in the 0.5–20 kW range, mainly utilize furling (or yawing) as their mechanism for power regulations. This is achieved by adjusting the capture area of the rotor disk relative to the dominant wind direction. A modified strip theory approach has been used to determine the effects of non-axial flow on the power performance and sound pressure level.

2 Aerodynamics Analysis

Blade element momentum (BEM) theory is the standard computational technique for the prediction of power curves of wind turbines; it is based on the 2-D aerodynamic characteristics of airfoil blade elements and some corrections accounting for 3-D wing aerodynamics. Before being able to calculate the force and moments on a blade, it is necessary to derive the velocity components of the air flow relative to any point on the blade and also the induced velocity components.

2.1 Velocity Components at the Blade

The following analysis has been applied to upwind horizontal axis wind turbine (HAWT) as shown in Fig. 1. The undisturbed flow can be expressed in the

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*, Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_130,

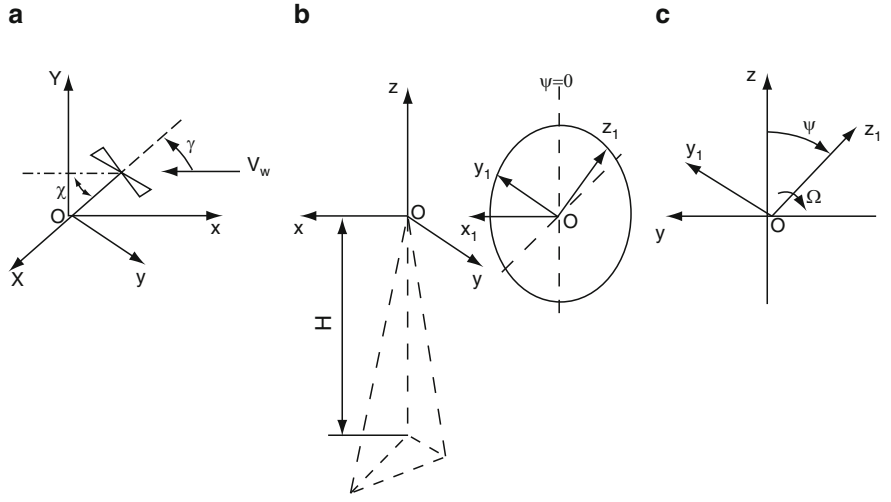


Fig. 1. Coordinate systems describing HAWT: (**a**) (X, Y, Z) – ground based coordinates; (**b**) (x, y, z) – rotor based coordinates; and (**c**) (x_1, y_1, z_1) – blade based coordinates

coordinate system, defined in Fig. 1a as

$$\overrightarrow{V}_w = V_w \cos\gamma\, \overrightarrow{i} - V_w \sin\gamma\, \overrightarrow{j}\,. \tag{1}$$

Transforming the rotor coordinate system in the blade coordinate system we have

$$\overrightarrow{V}_w = V_w \cos\gamma\, \overrightarrow{i_1} - V_w \sin\gamma\cos\psi\, \overrightarrow{j}_1 + V_w \sin\gamma\sin\psi\, \overrightarrow{k_1}\,. \tag{2}$$

The rotational motion of the blade will add a velocity component Ωr in the direction $\overrightarrow{j}_1$, so that the total velocity vector relative to the blade, $\overrightarrow{W}$ is

$$\overrightarrow{W} = V_w \cos\gamma\, \overrightarrow{i_1} + (\Omega r - V_w \sin\gamma\cos\psi)\, \overrightarrow{j_1} + V_w \sin\gamma\sin\psi\, \overrightarrow{k_1}\,. \tag{3}$$

If the blade is additionally allowed to flap through angle β about y_1, and the induced velocities are included, then the total velocity vector $\overrightarrow{W}$ transformed into the final coordinate system, $(\overrightarrow{i}_2, \overrightarrow{j}_2, \overrightarrow{k}_2)$, becomes

$$\begin{aligned}
\overrightarrow{W}_i &= [V_w \cos\gamma\cos\beta - v_a\cos\beta - (V_w \sin\gamma\sin\psi\sin\beta \\
&\quad - v_a \tan(\chi/2)\sin\psi\sin\beta)]\, \overrightarrow{i}_2 \\
\overrightarrow{W}_j &= [\Omega r\cos\beta + v_t - (V_w \sin\gamma\cos\psi - v_a\tan(\chi/2)\cos\psi]\, \overrightarrow{j}_2 \\
\overrightarrow{W}_k &= [V_w \sin\gamma\sin\psi\cos\beta - v_a\cos\beta - v_a\tan(\chi/2)\sin\psi\cos\beta \\
&\quad + V_w\cos\gamma\sin\beta - v_a\sin\beta]\, \overrightarrow{k}_2,
\end{aligned}$$

where χ is the wake skew angle. Further on we define the non-dimensional velocities: $a = v_a/(V_w\cos\gamma)$, $a' = v_t/(\Omega r\cos\beta)$, $\lambda = \Omega R\cos\beta/V_w$,

$X = \Omega r \cos\beta / V_w$, $C_i = W_i/(V_w \cos\gamma)$, $C_j = W_j/(V_w \cos\gamma)$ and $C_k = W_k/(V_w \cos\gamma)$, and substitute them into above equations to obtain

$$C_i = (1-a)\cos\beta - (\tan\gamma - a\tan(\chi/2))\sin\psi\sin\beta, \tag{4}$$

$$C_j = X/\cos\gamma\,[1 + a' - (\sin\gamma\cos\psi)/X + (a\cos\gamma\tan(\chi/2)\cos\psi)/X]\,, \tag{5}$$

$$C_k = \tan\gamma\sin\psi\cos\beta - a\tan(\chi/2)\sin\psi\sin\beta + (1-a)\sin\beta]. \tag{6}$$

2.2 Aerodynamic Loads

The basic idea of BEM models is to balance both the linear and angular momentum changes of the air masses flowing through the rotor disc with the axial force and torque generated on the rotor blades respectively. This balance is carried out in a detailed fashion, considering the flow through annular strips of width dr and the aerodynamic forces on blade elements of the same width; the forces are obtained from 2D wind tunnel data for the lift coefficient $C_L(\alpha)$ and the drag coefficient $C_D(\alpha)$. Both coefficients depend mainly on the angle of attack and Reynolds number. By using an average for an annular ring the elemental values of axial force and moment integrated around the ring are:

$$dF_b = \int_0^{2\pi} \frac{1}{2}\rho W^2 \sigma_r C_x \cos^2\beta\, r\, dr\, d\psi, \tag{7}$$

$$dM_b = \int_0^{2\pi} \frac{1}{2}\rho W^2 \sigma_r C_y r^2 \cos^2\beta\, dr\, d\psi, \tag{8}$$

and from the momentum theory we also have

$$dF_m = \int_0^{2\pi} 2\rho V_W^2 \cos^2\gamma\, a f\,(1-a)\cos^2\beta\, r\, dr\, d\psi, \tag{9}$$

$$dM_m = \int_0^{2\pi} 2\rho V_W \cos\gamma\, a' f\,(1-a)\,\Omega r\cos^4\beta\, r^2 dr d\psi, \tag{10}$$

where $C_x = C_L\cos\phi + C_D\sin\phi$, $C_y = C_L\sin\phi - C_D\sin\phi$, $\sigma_r = Bc/2\pi r\cos\beta$ is the solidity parameter, c is the chord length, B is the number of blades and f is the tip loss factor.

It has been assumed that any radial flow corresponding to (6) and the expansion of wake can be neglected and that (4) can be reduced to $C_i = (1-a)\cos\beta$. The flow angle ϕ is then determined by the components of velocity C_i and C_j

$$tg\phi = \frac{(1-a)\cos\gamma\cos\beta}{X\,(1+a') + \cos\psi\cos\gamma\left(a\tan\frac{\chi}{2} - \tan\gamma\right)}. \tag{11}$$

Now equating (7), (9) and (8), (10), by means of the flow angle ϕ, we obtain

$$\frac{a}{1-a} = \frac{\sigma_r \cos^2 \beta}{8\pi f} \int_0^{2\pi} \frac{C_x}{\sin^2 \phi} d\psi,$$

$$\frac{a'}{1+a'} = \frac{\sigma_r}{8\pi f} \int_0^{2\pi} \frac{C_y}{\sin\phi\cos\phi \left[1 - \dfrac{\tan\phi \tan\gamma \cos\psi}{(1-a)\cos\beta}\right]} d\psi.$$

The non-dimensionalized resultant velocity relative to a blade element is given by

$$\frac{W^2}{V_W^2 \cos^2 \gamma} = [(1-a)\cos\beta]^2 + \left[\frac{X(1+a')}{\cos\gamma} + \cos\psi \left(a \tan\frac{\chi}{2} - \tan\gamma\right)\right]^2 \quad (12)$$

Once the induction factors a and a' are known as a function of the radial variable r, the power coefficient for the complete blade can be calculated from $C_P = 8\lambda^2 \cos\gamma \int_0^1 \left(\frac{r}{R}\right)^3 a'(1-a)\, d\left(\frac{r}{R}\right)$. The elemental lift and drag forces per unit length are: $L = {}^1\!/_2 \rho W^2 c C_L$ and $D = {}^1\!/_2 \rho W^2 c C_D$, with W^2given by (12).

2.3 Results

A rotor of fixed pitch and two bladed configurations, one designed to operate at an optimum tip speed ratio of nine. Figure 2 shows the torque coefficients as a function of the tip speed ratio (λ) and yaw angle (γ), with zero coning angle ($\beta = 0$). A negative value of the torque (C_{MZ}) indicates that the operating turbine will veer in a direction which restores axial flow.

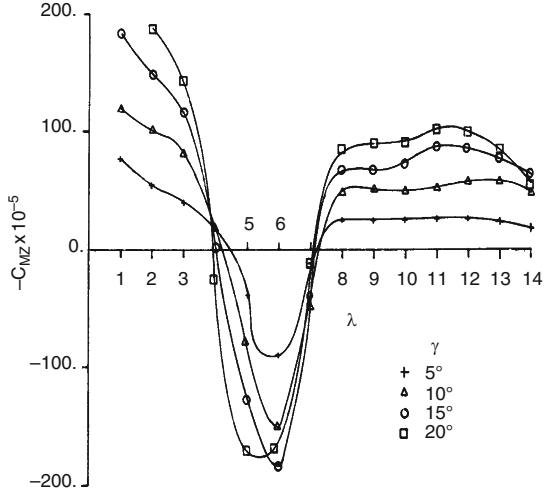


Fig. 2. Effect of yaw on the restoring torque

3 Aeroacoustic Analysis

The aerodynamic noise prediction method combines a model for predicting aerodynamic noise due to the effects of inflow turbulence upon on airfoil section, with prediction schemes for airfoil self-noise. The method is integrated into a program which averages the noise of the previous two bladed rotor over one revolution and gives as output a 1/3 octave A-weighted spectrum at a user selectable location. Output spectra are predicted at a downwind location for various yaw angles and comparisons with experiment are presented.

3.1 Inflow-Turbulence Noise (INT)

The adopted prediction model for turbulence inflow noise is based on the semi-empirical model of Amiet [1] derived for a single airfoil section under turbulent inflow and extended to the case of rotating blade by Lowson [4].This model can be applied for both high and low frequency, with smooth transition between the two regions:

$$L_{p,INF} = L^{H}_{p,INF} + 10\log_{10}\frac{K_c}{1+K_c},$$

where $L^{H}_{p,INF}$ is the sound pressure level for high frequency region and K_c is the low frequency correction [1].

3.2 Turbulent Boundary Layer-Trailing Edge Noise (TBLTE)

As its name implies TBLTE noise is caused by the flow of a turbulent boundary layer over the impedance discontinuity existing at the trailing edge (TE) of the airfoil. The effect of the edge is to radically increase the efficiency of the acoustic radiation of the turbulence, particularly at lower speeds. The noise is described as a function of local Mach number M, displacement thickness, δ^*, length of blade segment, ΔS, angle of attack, α and the distance of the source to observer position r. The total sound pressure level, in 1/3 octave band, is given by Brooks, Pope, Marcolini model as [2]:

$$L_{p,TBLTE} = 10log_{10}(10^{L_{P,\alpha}} + 10^{L_{P,s}} + 10^{L_{P,p}})$$

with $L_{P,\alpha}$ representing the effect of angle of attack, $L_{P,s}$, the contribution of the suction side and $L_{P,p}$ the contribution of the pressure side of the airfoil.

3.3 Blunt-Trailing Edge Noise (BTE)

The total sound pressure level caused by a blunt trailing edge is also modeled by a scaling law proposed by Brooks et al. [2]:

$$L_{p,BTE} = 10\log_{10}\left[\left(t^* M^{5.5}\Delta s\overline{D}_h\right)/r^2 + G_3\left(t^*/\delta^*_{avg}, \psi_{TE}\right)\right.$$
$$\left. + G_4\left(t^*/\delta^*_{avg}, \psi_{TE}, St^{'}/St^{'}_{peak}\right)\right],$$

with the Strouhal number based on the TE thickness t^*, $St' = ft^*/U$.

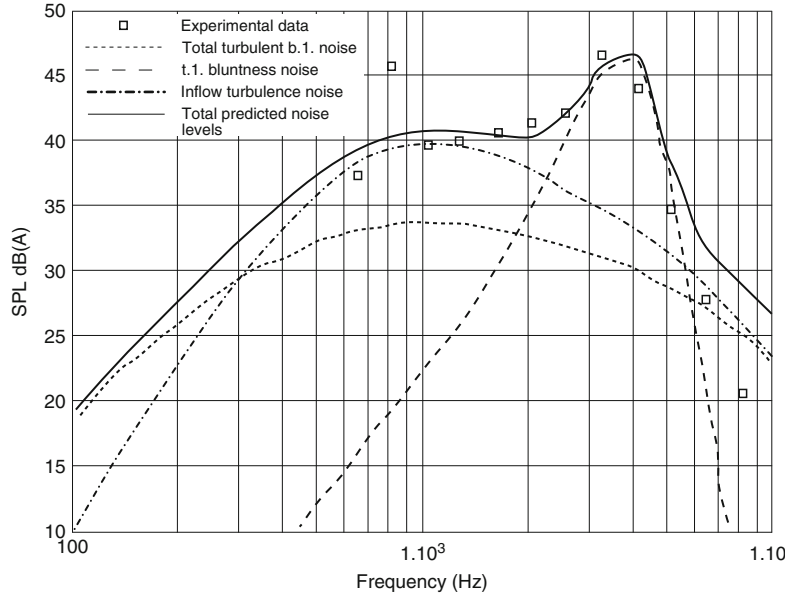


Fig. 3. Comparison between total predicted noise levels and experimental data

3.4 Results

The aerodynamic code is coupled with the aerodynamic noise prediction model. For HAWT's operating at certain yaw angle, the velocity field is accurately computed taking into account the skew angle of vortex cylinder wake. Also, for each airfoil, the boundary layer displacement thicknesses are calculated at both pressure side and suction side using the airfoil code XFOIL [3]. The acoustic analysis is based on a 10 kW HAWT for which some measurements have been performed in steady yaw state [5]. The noise from all sources (INT, TBLTE, BTE) are plotted together with experimental data in Fig. 3 for the yaw angles of 30° (not shown here for 0° and 60°). In all cases there is an overprediction of the noise levels at frequencies above 7,000 Hz, probably due to the incomplete input data.

4 Conclusions

An aerodynamic BEM model and aeroacoustic method were developed to predict the power output and noise from a HAWT rotor in yaw. The yawed rotor is less efficient than non-yawed rotor and so it is vital to assess the efficiency for purposes of energy production estimation and power control. The noise prediction model developed in conjunction with the aerodynamic model captures the key features of the noise produced by such a rotor. Some

comparisons between predicted and experimental data show good agreement over all yaw angles up to 60°. The model accurately predicts that the noise of the rotor is dominated by the tonal noise due to the TE bluntness and this form of noise is controllable by simply sharpening of TE.

References

1. Amiet, R.K.: J. Sound Vibr. **41**, 407–420 (1975)
2. Brooks, F.T., Pope, D.S., Marcolini, M.A., NASA **RP-1218**, 1–137 (1989)
3. Drela, M.: Lecture Notes in Engineering Low Reynolds Number Aerodynamics, vol. 54. Springer, Berlin (1989)
4. Lowson, M.V.: ETSU W/13/00284/Rep. Harwell, England, pp. 1–46 (1993)
5. Moroz, E.M.: Experimental and theoretical characterization of acoustic noise from a yaw controlled teetered rotor wind turbine. Master Thesis, UTEP (1995)

Quasi-Positive Continuous Darcy-Flux Finite-Volume Methods

Michael G. Edwards and Hongwen Zheng

Civil and Computational Engineering Centre, Swansea University, Swansea SA2 8PP, UK, m.g.edwards@swansea.ac.uk, h.zheng@swansea.ac.uk

Summary. A new family of flux-continuous finite-volume methods are presented for the full-tensor pressure equation with general discontinuous coefficients. Full pressure continuity that is built into the new methods leads to a quasi-positive formulation that minimises spurious oscillations in discrete pressure solutions for strongly anisotropic full-tensor fields.

1 Introduction

Approximation of the pressure equation resulting from Darcy's law requires that key physical constraints of continuity in normal flux and pressure be imposed at control-volume interfaces, across which strong discontinuities in permeability can occur.

In this paper a new family of flux-continuous, locally conservative, finite-volume schemes is presented for solving the general tensor pressure equation. The new schemes have full pressure continuity imposed across control-volume faces, in contrast to the earlier families of schemes with point-wise continuity in pressure and flux.

For strongly anisotropic full-tensor cases where M-matrix conditions are violated, the earlier flux-continuous schemes e.g. [1–5] cannot avoid decoupling of the solution [6], which leads to severe spurious oscillations in the discrete solution. The new schemes are shown to be quasi-positive and minimize spurious oscillations in discrete pressure solutions.

Section 2 gives the formulation of the single phase flow problem. The family of Full-Pressure Support (FPS) schemes is introduced in Sect. 3. M-matrix and Quasi-positive QM-matrices are presented in Sect. 4. Results are presented in Sect. 5. Conclusions follow in Sect. 6.

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*, Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_131,

2 Flow Equation and Problem Description

The pressure equation is formulated in a general curvilinear coordinate system defined with respect to a uniform dimensionless transform space with a (ξ, η) coordinate system. Choosing Ω to represent an arbitrary control volume comprised of surfaces that are tangential to constant (ξ, η) respectively, where $\partial\Omega$ is the boundary of Ω. Resolving the Darcy velocity $-\mathbf{K}\nabla\phi$ along the surface normals to (ξ, η) gives rise to the general tensor flux components $F = -\int(T_{11}\phi_\xi + T_{12}\phi_\eta)d\eta$ and $G = -\int(T_{12}\phi_\xi + T_{22}\phi_\eta)d\xi$ where general elliptic (Piola) tensor $\mathbf{T} =\mid J \mid \mathbf{J}^{-1}\mathbf{K}\mathbf{J}^{-T}$, elements are given in [1]. The tensor $\mathbf{K}$ can be discontinuous across internal boundaries of Ω. The closed integral of velocity divergence is written as

$$\int\int_{\Omega} \frac{(\partial_\xi \tilde{F} + \partial_\eta \tilde{G})}{J} J d\xi d\eta = \triangle_\xi F + \triangle_\eta G = m \tag{1}$$

where e.g. $\triangle_\xi F$ is the difference in net flux with respect to ξ and $\tilde{F} = T_{11}\phi_\xi + T_{12}\phi_\eta$, $\tilde{G} = T_{12}\phi_\xi + T_{22}\phi_\eta$. For incompressible flow pressure is specified at least at one domain point. Full tensors can arise from upscaling, unstructured grids and local orientation of the grid and permeability field.

3 Family of Flux-Continuous FPS Schemes

Here we introduce the continuous full pressure support (FPS) schemes. Cell-centred and cell-vertex formulations are developed. A cell-centred quadrilateral formulation is outlined (details in [6]). The support is shown in Fig. 1a, where flow and rock variables are assigned to the grid cells. A dual-cell (dashed) is introduced by connecting the primal nodes, partitioning the primal cells into sub-cells. Local flux continuity conditions are imposed over the subcell faces in the dual-cell to handle jumps in permeability. First interface pressures are introduced at the indicated positions (n, s, e, w) in Fig. 1b, these will be determined in terms of the primary cell centred pressures via four

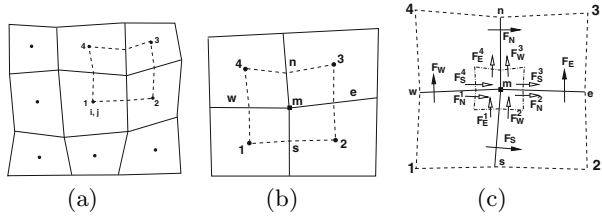


Fig. 1. (**a**) Nine-node support, cell-centered control-volume i, j, with dual-cell (*dashed line*) at $i + 1/2, j + 1/2$, (**b**) FPS dual cell and auxiliary pressure nodes n, s, e, w, m, and (**c**) fluxes in dual cell: *solid arrow* = primal-flux N, S, E, W, *hollow arrow* = auxiliary-flux

local flux continuity conditions Fig. 1c. Full sub-cell face pressure continuity is achieved by introducing a further interface pressure at the common corner m of the four subcells as indicated in Fig. 1b, i.e. at the dual-cell centre. This enables bilinear support in pressure to be introduced over each subcell so that *full pressure continuity* is achieved over the faces of each control-volume. The bilinear support retains a degree of freedom in position of flux continuity on a sub-face, leading to a new family of flux-continuous schemes with linear flux in the transverse direction and full pressure support. The additional degree of freedom at m is defined by imposing the discrete integral form of divergence (1) to hold over an auxiliary control-volume surrounding the dual-cell centre, as indicated in Fig. 1c. The four flux continuity conditions, imposed per dual-cell, together with the auxiliary discrete divergence condition lead to the local algebraic system

$$\begin{aligned} F_N &= -(T_{11}\phi_{\tilde{\xi}} + T_{12}\phi_{\tilde{\eta}})|_N^3 = -(T_{11}\phi_{\tilde{\xi}} + T_{12}\phi_{\tilde{\eta}})|_N^4, \\ F_S &= -(T_{11}\phi_{\tilde{\xi}} + T_{12}\phi_{\tilde{\eta}})|_S^1 = -(T_{11}\phi_{\tilde{\xi}} + T_{12}\phi_{\tilde{\eta}})|_S^2, \\ F_E &= -(T_{12}\phi_{\tilde{\xi}} + T_{22}\phi_{\tilde{\eta}})|_E^2 = -(T_{12}\phi_{\tilde{\xi}} + T_{22}\phi_{\tilde{\eta}})|_E^3, \\ F_W &= -(T_{12}\phi_{\tilde{\xi}} + T_{22}\phi_{\tilde{\eta}})|_W^1 = -(T_{12}\phi_{\tilde{\xi}} + T_{22}\phi_{\tilde{\eta}})|_W^4, \\ &- \textstyle\sum_{\partial\Omega_{AUX}} (\mathbf{K}\nabla\Phi)\cdot\hat{\mathbf{n}}\Delta s = 0 \end{aligned} \tag{2}$$

from which the interface pressures $(\phi_n, \phi_s, \phi_e, \phi_w, \phi_m)^T$ are eliminated, leading to fluxes expressed in terms of primal node pressures, which are assembled to form a divergence approximation over each primal cell.

4 M-Matrices and QM-Matrices

Conditional M-Matrix: *Any single η-parameter family of consistent locally conservative schemes on or within the 9-point stencil applied to a constant full-tensor field has an M-matrix if*

$$| T_{12} | \leq \eta(T_{11} + T_{22}) \leq min(T_{11}, T_{22}) \tag{3}$$

where η defines quadrature point [1, 6]. This theorem applies on a uniform quadrilateral grid of rectangles or parallelograms. Note: FPS schemes are exact for piece-wise linear and bilinear fields since the pressure basis functions are piecewise bilinear.

Triangular Grids A (cell-vertex scheme) M-matrix will be obtained on a triangle grid if $| T_{12} | < T_{11}$ and $| T_{21} | < T_{22}$. For a symmetric tensor $| T_{12} | < \min(T_{11}, T_{22})$ [5,6], which is consistent with the quadrilateral optimal support (4 below) and M-matrix condition.

4.1 Variable Support and 7-Point Schemes

If we choose a quadrature point with

$$\eta = | T_{12} | /(T_{11} + T_{22}) \tag{4}$$

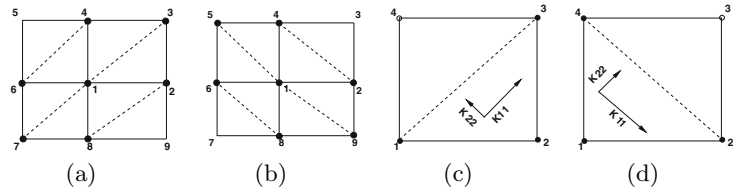


Fig. 2. (**a**) Positive T_{12} over all contributing dual-cells – right inclined 7 pt scheme, (**b**) negative T_{12} over all contributing dual-cells – left inclined 7 pt scheme, (**c**) positive T_{12} over a dual-cell, control-vol at 2 uses nodes 1,2,3, and (**d**) negative T_{12} over a dual-cell, control-vol at 1 uses nodes 4,1,2

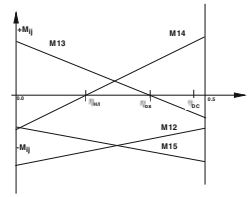


Fig. 3. Unique coefficients M12, M13, M14, M15 versus η (quadrature range) QM-matrix

then an M-matrix is obtained subject to a sufficient condition for ellipticity, i.e. if $\mid T_{12} \mid \leq \min(T_{11}, T_{22})$ for a triangle grid . This quadrature point reduces a 9-point scheme to a 7-point (triangle) scheme with upward support if tensor cross-term $T_{12} > 0$ over the supporting dual cells or downward support if $T_{12} < 0$. We refer to this as an *optimal support condition* with $\eta = \eta_{OS}$ (Fig. 2). However, in the general case approximate optimal support will be obtained via quadrature [6]. Optimal support may still be achieved by triangulation according to anisotropy angle [7] or by special case construction [8].

4.2 Quasi-Positive QM-Matrices

Definition. A Quasi-M-matrix (QM-matrix) is a matrix that has the minimum of only one unique positive off-diagonal coefficient that violates the M-matrix conditions, where the matrix would otherwise be an M-matrix. Here QM-matrices are considered for $\mid T_{12} \mid > \min(T_{11}, T_{22})$ (Fig. 3). The *optimal support condition* of (4) (η_{OS}) leads to an optimal QM-matrix scheme with 7-points (Fig. 2) [6]. An anisotropy favoring triangulation of a quadrilateral grid will also lead to a QM-matrix since the same optimal support is obtained.

Note on Decoupled Approximation. The quadrature point $\eta = 1/2$ is singular and results in a discretization that has a *checker board* solution that is strongly oscillatory and decoupled [6]. The earlier pointwise continuous TPS schemes have a limited quadrature range for highly anisotropic full tensors that remains close to $\eta = 1/2$ [6], leading to a decoupled approximation.

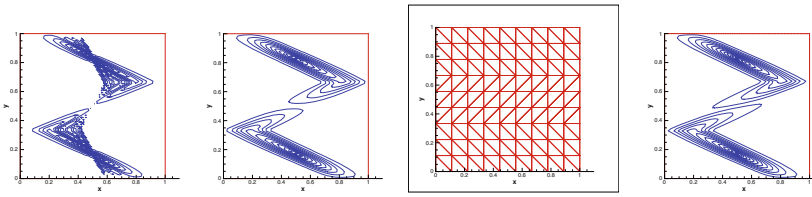


Fig. 4. Pressure contours – quadrilateral grid: (**a**) TPS scheme, (**b**) FPS OS scheme, (**c**) anisotropy favoring triangles, and (**d**) triangle grid FPS contours

5 Numerical Results

A comparison is now presented between the new full pressure support (FPS) formulation and the earlier pointwise triangular pressure support (TPS) formulation for a strong discontinuous full-tensor (zigzag) field.

The boundary conditions for the unit domain involve a source and sink placed at opposite mid-points of the first and last thirds of the domain together with zero pressure prescribed on all boundary walls. The permeability principal axes change direction in anisotropy at one third and two thirds the way across the domain (i.e. minus, plus, minus 25°). The discontinuous full-tensor permeability field is defined in sections varying from Sect. 1 $\mathbf{K} = [2{,}464.360020, -1{,}148.683643, -1{,}148.683643, 536.6399794]$ Sect. 2 with $\mathbf{K} = [2{,}464.360020, +1{,}148.683643, +1{,}148.683643, 536.6399794]$ and Sect. 3 with the tensor of Sect. 1, with principal anisotropy ratio of 3,000:1, violating the M-matrix conditions in each section.

Results are presented for TPS with $q = 1$ (Fig. 4a), there are very strong oscillations in the solution with violation of the maximum principle. The (FPS) quadrature of (4) leads to optimal support away from the discontinuities, according to local orientation of the full-tensor field. The FPS formulation yields oscillation free results (compare to TPS) with a QM-matrix solution (Fig. 4b). Solution resolution is seen to sharpen with η increasing. The cell-vertex scheme with triangulation favoring the anisotropy (Fig. 4c) also shows very good resolution (Fig. 4d), again leading to a QM-matrix optimal support scheme.

6 Conclusions

New families of locally conservative flux-continuous, finite-volume schemes are presented for solving the general tensor pressure equation on quadrilateral and triangular grids. The new schemes have full pressure continuity and flux continuity imposed across control-volume faces and are quasi-positive yielding solutions essentially free of spurious oscillations. An optimal support scheme is identified via quadrature which adapts the discretization accord-

ing to anisotropy. An optimal support scheme is also obtained by anisotropy favoring triangulation.

Acknowledgements

Supported by EPSRC (GR/S70968/01) and ExxonMobil Upstream Research Company.

References

1. Edwards, M.G., Rogers, C.F.: Finite volume discretization with imposed flux continuity for the general tensor pressure equation. Comput. Geosci. **2**, 259–290 (1998)
2. Edwards, M.G.: Unstructured, control-volume distributed, full-tensor finite volume schemes with flow based grids. Comput. Geosci. **6**, 433–452 (2002)
3. Pal, M., Edwards, M.G., Lamb, A.R.: Convergence study of a family of flux-continuous, finite-volume schemes for the general tensor pressure equation. Int. J. Numer. Methods Fluids **51**, 1177–1203 (2006)
4. Aavatsmark, I.: Introduction to multipoint flux approximation for quadrilateral grids. Comput. Geosci. **6**, 405–432 (2002)
5. Lee, S.H., Jenny, P., Tchelepi, H.A.: A finite-volume method with hexahedral multiblock grids for modeling flow in porous media. Comput. Geosci. **6**, 353–379 (2002)
6. Edwards, M.G., Zheng, H.: A quasi-positive family of continuous darcy-flux finite volume schemes with full pressure support. J. Comput. Phys. **227**, 9333–9364 (2008)
7. Pal, M., Edwards, M.G.: Quasi-monotonic continuous darcy-flux approximation for general 3-D grids of any element type. In: SPE Reservoir Simulation Symposium, Paper SPE 106486, 14 p., Houston, TX, 26–28 February 2007. DOI 10.2118/106486-MS
8. Aavatsmark, I., Eigestad, G.T., Mallison, B.T., Nordbotten, J.M.: A compact multipoint flux approximation method with improved robustness. Numer. Methods Part. Diff. Equ. **24**, 1329–1360 (2008) (http://www.interscience.wiley.com)

Are Copying and Innovation Enough?

T.S. Evans[1,2], A.D.K. Plato[2], and T.You[1]

[1] Theoretical Physics, Imperial College London, London SW7 2AZ, UK
T.Evans@imperial.ac.uk, tevong@gmail.com
[2] Institute for Mathematical Sciences, Imperial College London, London SW7 2PG, UK, Alexander.Plato@imperial.ac.uk

Summary. Exact analytic solutions and various numerical results for the rewiring of bipartite networks are discussed. An interpretation in terms of copying and innovation processes make this relevant in a wide variety of physical contexts.

1 Introduction

There are many situations where an 'individual' chooses only one of many 'artifacts' but where their choice depends in part on the current choices of the community. Names for new babies and registration rates of pedigree dogs often reflect current popular choices [10, 11]. The allele for a particular gene carried ('chosen') by an individual reflects current gene frequencies [8]. In Urn models the probabilities controlling the urn chosen by a ball can reflect earlier choices [9]. In all cases copying the state of a neighbour, as defined by a network of the individuals, is a common process because it can be implemented without any global information [7]. At the other extreme, an individual might pick an artifact at random.

2 The Basic Model

We first consider a non-growing bipartite network in which E 'individual' vertices are each attached by a single edge to one of N 'artifact' vertices. At each time step we choose to rewire the artifact end of one edge, the *departure* artifact chosen with probability Π_R. This is attached to an *arrival* artifact chosen with probability Π_A. Only after both choices are made is the graph rewired as shown in Fig. 1. The degree distribution of the artifacts when averaged over many runs of this model, $n(k,t)$, satisfies the following equation:-

$$\begin{aligned} n(k,t+1) = n(k,t) &+ n(k+1,t)\Pi_R(k+1,t)\left(1-\Pi_A(k+1,t)\right) \\ &- n(k,t)\Pi_R(k,t)\left(1-\Pi_A(k,t)\right) - n(k,t)\Pi_A(k,t)\left(1-\Pi_R(k,t)\right) \\ &+ n(k-1,t)\Pi_A(k-1,t)\left(1-\Pi_R(k-1,t)\right), \quad (E \geq k \geq 0), \end{aligned} \tag{1}$$

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*,
Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_132,

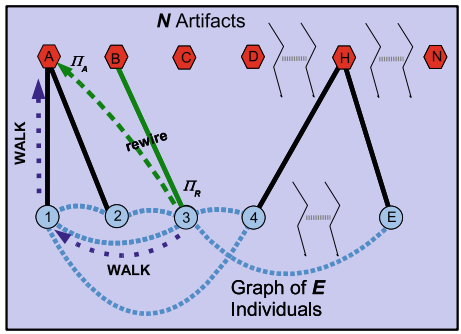


Fig. 1. The bipartite network of E individual vertices, each connected by a single edge (*solid lines*) to any one of N artifacts. The *dashed lines* below the individuals are a social network. In the event shown individual **3** updates their choice, making **B** the departure artifact. They do this by copying the choice of a friend, friend of a friend, etc., found by making a random walk on the social network. Here this produces **A** as the arrival artifact so edge **3B** is rewired to become edge **3A**

where $n(k) = \Pi_R(k) = \Pi_A(k) = 0$ for $k = -1, (E+1)$. If Π_R or Π_A have terms proportional to k^β then this equation is exact only when $\beta = 0$ or 1 [5]. We will use the most general Π_R and Π_A for which (1) is exact, namely

$$\Pi_R = \frac{k}{E}, \qquad \Pi_A = p_r \frac{1}{N} + p_p \frac{k}{E}, \qquad p_p + p_r = 1 \qquad (E \geq k \geq 0)\,. \tag{2}$$

This is equivalent to using a complete graph with self loops for the social network at this stage but these preferential attachment forms emerge naturally when using a random walk on a general network [7]. This choice for Π_A has two other special properties: one involves the scaling properties [5] and the second is that these exact equations can be solved analytically [3–6]. The generating function $G(z,t) = \sum_k z^k n(k,t)$ is decomposed into eigenmodes $G^{(m)}(z)$ through $G(z,t) = \sum_{m=0}^{E} c_m(\lambda_m)^t G^{(m)}(z)$. From (1) we find a second order linear differential equation for each of the eigenmodes with solution [5]

$$G^{(m)}(z) = (1-z)^m {}_2F_1(a+m, -E+m; 1-E-a(N-1); z)\,, \quad a = \frac{p_r}{p_p}\frac{E'}{N}\,,$$
$$\lambda_m = 1 - m(m-1)\frac{p_p}{EE'} - m\frac{p_r}{E}, \qquad 0 \leq m \leq E\,, \tag{3}$$

where $E' = E$. These solutions are well known in theoretical population genetics as those of the Moran model [8] and one may map the bipartite model directly onto a simple model of the genetics of a haploid population [5].

The equilibrium result for the degree distribution [3,5] is proportional to $\frac{\Gamma(k+a)}{\Gamma(k+1)}\frac{\Gamma(E+a(N-1)-1-k)}{\Gamma(E+1-k)}$. This has three typical regions. We have a condensate, where most of the edges are attached to one artifact $p(k=E) \sim O(N^0)$, for $p_r \ll (E+1-\langle k\rangle)^{-1}$. On the other hand when $p_r \gg (1+\langle k\rangle)^{-1}$ we get a

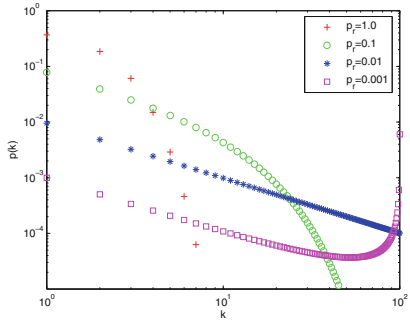


Fig. 2. The equilibrium degree probability distribution function $p(k) = n(k)/N$ for $N = E = 100$. Shown are (from top to bottom at low k) $p_r = 1$ (*crosses*), $10/E$ (*circles*), $1/E$ (*stars*) and $0.1/E$ (*squares*)

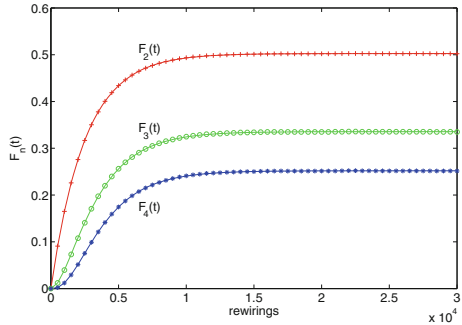


Fig. 3. Plots of various $F_n(t)$ for $E = N = 100$, $p_r = 0.01$. The *points* are averages over 10^5 runs while the *lines* are the exact theoretical results. From top to bottom we have: $F_2(t)$ (*crosses*), $F_3(t)$ (*circles*), $F_4(t)$ (*stars*)

peak at small k with an exponential fall off, a distribution which becomes an exact binomial at $p_r = 1$. In between we get a power law with an exponential cutoff, $p(k) \propto (k)^{-\gamma} \exp\{-\zeta k\}$ where $\gamma \approx (1 - \frac{p_r}{p_p}\langle k\rangle)$ and $\zeta \approx -\ln(1 - p_r)$. For many parameter values the power γ will be indistinguishable from one and this is a characteristic signal of an underlying copying mechanism seen in a diverse range of situations (e.g. see [1,12]; Fig. 2).

One of the best ways to study the evolution of the degree distribution [5,6] is through the *Homogeneity Measures*, F_n. This is the probability that n distinct edges chosen at random are connected to same artifact, and is given by $F_n(t) := (\Gamma(E + 1 - n)/\Gamma(E + 1))(d^n G(z,t)/dz^n)_{z=1}$. Further, each F_n depends only on the modes numbered 0 to n so they provide a practical way to fix the constants c_n in the mode expansion. Since $F_0 = E$ and $F_1 = 1$, we find $c_0 = 1$ and $c_1 = 0$ while equilibration occurs on a time scale of $\tau_2 = -1/\ln(\lambda_2)$ (Fig. 3).

3 Communities

Our first generalisation of the basic model is to consider two distinct communities of individuals, say E_x (E_y) of type X (Y). The individuals of type X can now copy the choices made by their own community X with probability p_{pxx}, a different rate which is used when an X copies the choice made by somebody in community Y, p_{pxy}. An X individual will then innovate with probability $(1 - p_{pxx} - p_{pxy})$. Another two independent copying probabilities can be set for the Y community. At each time step we choose to update the choice of a member of community X (Y) community with probability p_x $(1 - p_x)$. Complete solutions are not available but one can find exact solutions for the lowest order Homogeneity measures and eigenvalues using similar techniques to those discussed above. The unilluminating details are given in [6].

4 Complex Social Networks

An obvious generalisation is to use a complex network as the Individual's social network [6]. When copying, done with probability p_p, an individual does a random walk on the social network to choose another individual and finally to copy their choice of artifact, as shown in Fig. 1. The random walk is an entirely local process, no global knowledge of the social network is needed, so it is likely to be a good approximation of many processes found in the real world. It also produces an attachment probability which is, to a good approximation, proportional to the degree distribution [7]. The alternative process of innovation, followed with probability p_r, involves global knowledge through its normalisation N in (2). However when $N \gg E$ this can represent innovation of new artifacts as it is likely that the arrival artifact has never been chosen before. However this process could also be a first approximation for other unknown processes used for artifact choice.

Results shown in Fig. 4 show that the existence of hubs in the Scale Free social network enhances the condensate while large distances in the social networks, as with the lattices, suppress the condensate.

An interesting example is the case of $N = 2$ which is a Voter Model [13] with noise (innovation $p_r \neq 0$) added. One can then compare the probability that a neighbour has a different artifact (the interface density) $\rho(t)$, a local measure of the inhomogeneity, with our global measure $(1 - F_2(t))$. These coincide when the social network is a complete graph. However as we move from 3D to 1D lattices, keeping N, E and p_r constant, we see from Fig. 5 that both these local and global measures move away from the result for the complete graph but in opposite directions [6].

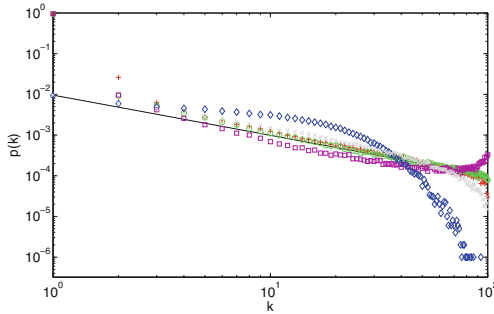


Fig. 4. The degree distributions $p(k)$ averaged over 10^4 runs for different social networks of average degree of 4: Erdős-Réyni (*pluses*), Exponential (random with $p(k) \propto \exp(-\zeta k)$, *circles*), Scale Free (random with $p(k) \propto k^{-3}$, *squares*), periodic lattices of two (*grey crosses*) and one (*diamonds*) dimension. The line is the analytic result where the social network is a complete graph with self loops. $N = E = 100$, $p_r = 1/E$

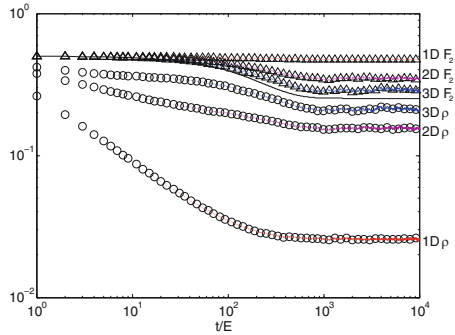


Fig. 5. Inhomogeneity measures for various lattices against t/E. The *black solid line* represents the analytic result $(1 - F_2(t))$ for $N = 2$, $p_r = 1/E$ and $E = 729$. Numerical results for $(1 - F_2(t))$ (*triangles*) and for the average probability that a neighbour has a different artifact, $\rho(t)$ (*circles*) shown for social networks which are lattices of different dimensions. Averaged over 1,000 runs

5 Different Update Methods

Another way we can change the model is to change the nature of the update. Suppose we first select the edge to be rewired and immediately remove it. Then, based on this network of $E' = (E - 1)$ edges, we choose the arrival artifact with probability $\Pi_A = (p_r/N) + (1 - p_r)k/E'$. The original master equation (1) is still valid and exact. Moreover it can still be solved exactly giving exactly the same form as before, (3), but with $E' = (E-1)$ not E. This gives very small differences of order $O(E^{-1})$ when compared to the original simultaneous update used initially.

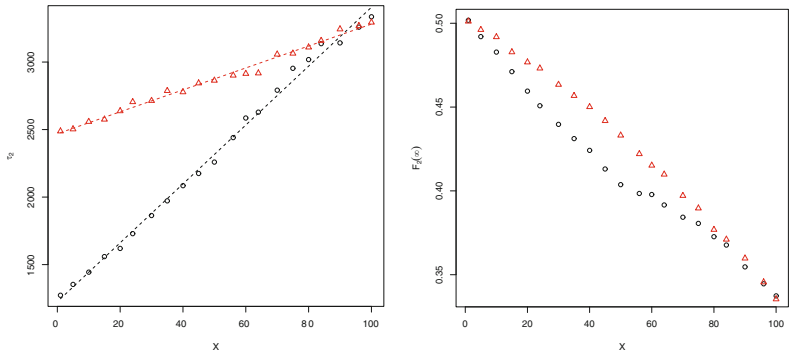


Fig. 6. $\tau_2 = -1/\ln(\lambda_2)$ (*left*) and $F_2(\infty)$ (*right*) obtained by fitting $A + B(\lambda_2)^t$ to the data for $F_2(t)$. For sequential ($m = 4$ circles, *lower lines*) and random ($m = 6$ triangles, *upper lines*) updates of X individuals at a time. $N = E = 100$, $p_r = 1/E = 0.01$ and averaged over 10^4 runs. The *dashed lines* represent the best linear fit with $\tau_2 \approx 1{,}230(20) + 21.8(3)X$ for $m = 4$ and $\tau_2 \approx 2{,}470(10) + 8.1(2)X$ for $m = 6$. Theoretical values are $\tau_2 \approx 2{,}512.1$ and $F_2(\infty) \approx 0.50251$ for $X = 1$ random update and $\tau_2 \approx 3{,}316.6$ and $F_2(\infty) \approx 0.33669$ for $X = 100$ either update

Instead we will consider the simultaneous rewiring of X edges in our bipartite graph at each step. We will choose the individuals, whose edges define the departure artifacts, in one of two ways: either sequentially or at random. The arrival artifacts will be chosen as before using Π_A of (2).

The opposite extreme from the single edge rewiring case we started with ($X = 1$) is the one where all the edges are rewired at the same time, $X = E$. This is the model used in [2,10,11] to model various data sets on cultural transmission. It is also the classic Fisher-Wright model of population genetics [8]. From this each homogeneity measure F_n and the n-th eigenvector λ_n may be calculated in terms of lower order results F_m ($m < n$). Non trivial information again comes first from $F_2(t) = F_2(\infty) + (\lambda_2)^t\,(F_2(0) - F_2(\infty))$ where

$$F_2(\infty) = \frac{p_p^2 + (1 - p_p^2)\langle k\rangle}{p_p^2 + (1 - p_p^2)E}\,, \qquad \lambda_2 = \frac{p_p^2(E-1)}{E}\,. \tag{4}$$

Comparing with the results for $X = 1$ we see that there are large differences in the equilibrium solution and in the rate at which this is approached (measured in terms of number of the rewirings made). For intermediate values of X we have not obtained any analytical results so for these numerical simulations are needed, as shown in Fig. 6.

References

1. Anghel, M., et al.: Phys. Rev. Lett. **92**, 058701 (2003)
2. Bentley, R.A., et al.: Evol. Hum. Behav. **28**, 151 (2007)
3. Evans, T.S.: Eur. Phys. J. B **56**, 65 (2007)
4. Evans, T.S.: arXiv:0711.0603
5. Evans, T.S., Plato, A.D.K.: Phys. Rev. E **75**, 056101 (2007)
6. Evans, T.S., Plato, A.D.K.: Netw. Heterogen. Media **3**, 221 (2008)
7. Evans, T.S., Saramäki, J.P.: Phys. Rev. E **72**, 026138 (2005)
8. Ewens, W.J.: Mathematical population genetics. I. Theoretical introduction, 2nd edn. Springer, New York (2004)
9. Godreche, C., Luck, J.M.: J. Phys.: Condens. Matter **14**, 1601 (2002)
10. Hahn, M.W., Bentley, R.A.: Proc. R. Soc. Lond. B **270**(Suppl. 1), S120 (2003)
11. Herzog, H.A., Bentley, R.A., Hahn, M.W.: Proc. R. Soc. Lond. B **271**(Suppl.), S353 (2004)
12. Laird, S., Jensen, H.J.: Europhys. Lett. **76**, 710 (2006)
13. Liggett, T.M.: Stochastic Interacting Systems: Contact, Voter and Exclusion Processes. Springer, New York (1999)

Pricing Options Under Stochastic Volatility with Fourier-Cosine Series Expansions

F. Fang[1] and C.W. Oosterlee[2]

[1] DIAM, Delft University of Technology, Delft, the Netherlands
wellstone_ff@hotmail.com
[2] CWI – Centrum Wiskunde & Informatica, Amsterdam, the Netherlands
c.w.oosterlee@cwi.nl

Summary. An option pricing method for European options based on the Fourier-cosine series, called the COS method, is presented. It can cover underlying asset processes for which the characteristic function is known, and in this paper, in particular, we consider stochastic volatility dynamics.

1 Introduction: The COS Method

Efficient numerical methods are required to rapidly price complex contracts and calibrate financial models. During calibration, i.e., when fitting model parameters of the stochastic asset processes to market data, we typically need to price European options at a single spot price, with many different strike prices, very quickly. Particular examples of where this is important would be processes with several parameters, like the Heston model [4] or the infinite activity Lévy processes, since there the pricing problem (for many strikes) is used inside an optimization method.

The integration methods are used for calibration purposes whenever the characteristic function of the asset price process is known analytically. State-of-the-art numerical integration techniques have in common that they rely on a transformation to the Fourier domain. The Carr-Madan method [1] is one of the best known examples of this class.

In this paper we will focus on *Fourier-cosine expansions* in the context of numerical integration as an alternative for methods based on the FFT. We will show that this method, called the COS method [2, 3], can improve the speed of pricing plain vanilla options.

The point-of-departure for pricing European options with numerical integration techniques is the risk-neutral valuation formula:

$$v(x,t_0) = e^{-r\Delta t}\mathbb{E}^{\mathbb{Q}}\left[v(y,T)|x\right] = e^{-r\Delta t}\int_{\mathbb{R}} v(y,T)f(y|x)dy, \tag{1}$$

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*,
Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_133,

where v denotes the option value, Δt is the difference between the maturity, T, and the initial date, t_0, and $\mathbb{E}^{\mathbb{Q}}[\cdot]$ is the expectation operator under risk-neutral measure $\mathbb{Q}$. x and y are state variables at time t_0 and T, respectively; $f(y|x)$ is the probability density of y given x, and r is the risk-neutral interest rate.

Since the density rapidly decays to zero as $y \to \pm\infty$ in (1), we truncate the infinite integration range without loosing significant accuracy to $[a, b] \subset \mathbb{R}$, and we obtain approximation v_1:

$$v_1(x, t_0) = e^{-r\Delta t} \int_a^b v(y, T) f(y|x) dy. \tag{2}$$

Since $f(y|x)$ is usually not known whereas the characteristic function is, we replace the density by its cosine expansion in y,

$$f(y|x) = {\sum}'_{k=0}^{+\infty} A_k(x) \cos\left(k\pi \frac{y-a}{b-a}\right) \tag{3}$$

with

$$A_k(x) := \frac{2}{b-a} \int_a^b f(y|x) \cos\left(k\pi \frac{y-a}{b-a}\right) dy, \tag{4}$$

so that

$$v_1(x, t_0) = e^{-r\Delta t} \int_a^b v(y, T) {\sum}'_{k=0}^{+\infty} A_k(x) \cos\left(k\pi \frac{y-a}{b-a}\right) dy. \tag{5}$$

$\sum'$ indicates that the first term in the summation is weighted by one-half.

We interchange the summation and integration, and insert the definition

$$V_k := \frac{2}{b-a} \int_a^b v(y, T) \cos\left(k\pi \frac{y-a}{b-a}\right) dy, \tag{6}$$

resulting in

$$v_1(x, t_0) = \frac{1}{2}(b-a) e^{-r\Delta t} \cdot {\sum}'_{k=0}^{+\infty} A_k(x) V_k. \tag{7}$$

The V_k are the cosine series coefficients of payoff function $v(y, T)$ in y.

We have analytic solutions for V_k for several contracts. As we assume the characteristic function of the log-asset price to be known, we represent the payoff as a function of the log-asset price, $x := \ln(S_0/K)$ and $y := \ln(S_T/K)$, with S_t the underlying price at time t and K the strike price. Focusing on a put option, we obtain

$$V_k^{put} = \frac{2}{b-a} K \left(-\chi_k(a, 0) + \psi_k(a, 0)\right). \tag{8}$$

where χ_k and ψ_k are given by

$$\chi_k(c,d) := \frac{1}{1+\left(\frac{k\pi}{b-a}\right)^2}\left[\cos\left(k\pi\frac{d-a}{b-a}\right)e^d - \cos\left(k\pi\frac{c-a}{b-a}\right)e^c + \frac{k\pi}{b-a}\sin\left(k\pi\frac{d-a}{b-a}\right)e^d - \frac{k\pi}{b-a}\sin\left(k\pi\frac{c-a}{b-a}\right)e^c\right] \quad (9)$$

and

$$\psi_k(c,d) := \begin{cases} \left[\sin\left(k\pi\frac{d-a}{b-a}\right) - \sin\left(k\pi\frac{c-a}{b-a}\right)\right]\frac{b-a}{k\pi} & k \neq 0, \\ (d-c) & k = 0. \end{cases} \quad (10)$$

For a call we find a similar expression.

Due to the rapid decay rate of the V_k, we further truncate the series summation in (7) to obtain approximation v_2:

$$v_2(x,t_0) = \frac{1}{2}(b-a)e^{-r\Delta t}\cdot{\sum}'_{k=0}^{N-1} A_k(x)V_k. \quad (11)$$

Coefficients $A_k(x)$, defined in (4), can be approximated by $F_k(x)$ defined as

$$F_k(x) := \frac{2}{(b-a)}\mathrm{Re}\left\{\phi\left(\frac{k\pi}{b-a};x\right)\cdot e^{-i\frac{ka\pi}{b-a}}\right\} \quad (12)$$

with $\phi(\omega;x)$ the characteristic function:

$$\phi(\omega;x) := \int_{\mathbb{R}} e^{i\omega y} f(y|x)dy.$$

This gives

$$v(x,t_0) \approx v_3(x,t_0) = e^{-r\Delta t}{\sum}'_{k=0}^{N-1}\mathrm{Re}\left\{\phi\left(\frac{k\pi}{b-a};x\right)e^{-ik\pi\frac{a}{b-a}}\right\}V_k, \quad (13)$$

where $\mathrm{Re}\{\cdot\}$ denotes taking the real part of the argument.

Equation (13) can be improved for the Lévy and the Heston models, so that options *for many strike prices* can be computed simultaneously. In the Heston model [4], the volatility, denoted by $\sqrt{u_t}$, is modeled by an additional stochastic differential equation,

$$\begin{aligned} dx_t &= \left(\mu - \tfrac{1}{2}u_t\right)dt + \sqrt{u_t}dW_{1t}, \\ du_t &= \lambda(\bar{u}-u_t)dt + \eta\sqrt{u_t}dW_{2t} \end{aligned} \quad (14)$$

where x_t denotes the log-asset price variable and u_t the variance of the asset price process. Parameters $\lambda \geq 0, \bar{u} \geq 0$ and $\eta \geq 0$ are the speed of mean reversion, the mean level of variance and the volatility of volatility, respectively. Furthermore, the Brownian motions W_{1t} and W_{2t} are assumed to be correlated with correlation coefficient ρ.

For the Heston model, we have $\phi(\omega;\mathbf{x},u_0)=\varphi_{hes}(\omega;u_0)\cdot e^{i\omega\mathbf{x}}$, with u_0 the volatility of the underlying at the initial time and $\varphi_{hes}(\omega;u_0):=\phi(\omega;0,u_0)$. The characteristic function of the log-asset price, $\varphi_{hes}(\omega;u_0)$, reads

$$\varphi_{hes}(\omega;u_0)=\exp\left(i\omega\mu\Delta t+\frac{u_0}{\eta^2}\left(\frac{1-e^{-D\Delta t}}{1-Ge^{-D\Delta t}}\right)(\lambda-i\rho\eta\omega-D)\right)\cdot \exp\left(\frac{\lambda\bar{u}}{\eta^2}\left(\Delta t(\lambda-i\rho\eta\omega-D)-2\log(\frac{1-Ge^{-D\Delta t}}{1-G})\right)\right),$$

with

$$D=\sqrt{(\lambda-i\rho\eta\omega)^2+(\omega^2+i\omega)\eta^2}\quad\text{and}\quad G=\frac{\lambda-i\rho\eta\omega-D}{\lambda-i\rho\eta\omega+D}.$$

Recalling the V_k-formula for a European options, like (8), we now define them as a vector multiplied by a scalar, $\mathbf{V}_k=U_k\mathbf{K}$, where

$$U_k=\begin{cases}\frac{2}{b-a}\left(-\chi_k(a,0)+\psi_k(a,0)\right) & \text{for a put}\\ \frac{2}{b-a}\left(\chi_k(0,b)-\psi_k(0,b)\right) & \text{for a call.}\end{cases}\tag{15}$$

We then find

$$v(\mathbf{x},t_0,u_0)\approx\mathbf{K}e^{-r\Delta t}\cdot\mathrm{Re}\left\{\sum\nolimits'_{k=0}^{N-1}\varphi_{hes}\left(\frac{k\pi}{b-a};u_0\right)U_k\cdot e^{ik\pi\frac{\mathbf{x}-a}{b-a}}\right\}.\tag{16}$$

This is the COS formula, pricing European options under Heston dynamics very efficiently. The convergence rate of the Fourier-cosine series depends on the properties of the functions on the interval $[a,b]$. From the error analysis in [2], we can summarize that, with a properly chosen truncation of the integration range, the overall error converges either exponentially for density functions, with nonzero derivatives, belonging to $\mathbb{C}^\infty([a,b]\subset\mathbb{R})$.

We define the truncation range by

$$[a,b]:=[c_1-12\sqrt{|c_2|},c_1+12\sqrt{|c_2|}].$$

in which the cumulants, c_n, are given by the derivatives, at zero, of $g(t)=\log(E(e^{t\cdot X}))$,

$$c_1=\mu T+(1-e^{-\lambda T})\frac{\bar{u}-u_0}{2\lambda}-\frac{1}{2}\bar{u}T,$$

$$c_2=\frac{1}{8\lambda^3}\big(\eta T\lambda e^{-\lambda T}(u_0-\bar{u})(8\lambda\rho-4\eta)+\lambda\rho\eta(1-e^{-\lambda T})(16\bar{u}-8u_0)+2\bar{u}\lambda T(-4\lambda\rho\eta+\eta^2+4\lambda^2)+\eta^2((\bar{u}-2u_0)e^{-2\lambda T}+\bar{u}(6e^{-\lambda T}-7)+2u_0)+8\lambda^2(u_0-\bar{u})(1-e^{-\lambda T})\big)$$

Cumulant c_2 may become negative for sets of Heston parameters that do not satisfy the Feller condition, i.e., $2\bar{u}\lambda>\eta^2$. We therefore use the absolute value of c_2.

The Greeks, like Delta, Gamma and also Vega can be obtained, basically at no cost, by differentiating the COS formula (16).

2 Numerical Results

We perform a numerical test on European options under the Heston process to evaluate the efficiency and accuracy of the COS method. We compare our results to the Carr-Madan method [1], in which the FFT has been used. Parameter N, in the experiments to follow, denotes for the COS method the number of terms in the Fourier-cosine expansion, and the number of grid points for the Carr-Madan method. Some experience is helpful when choosing the correct truncation range and damping factor in Carr-Madan's method. A suitable choice appears to be $\alpha = 0.75$ for the Heston experiments.

The computer used has an Intel Pentium 4 CPU, 2.80 GHz with cache size 1,024 KB; The code is written in Matlab 7-4.

2.1 The Heston Model

We choose the Heston model and price puts with the following parameters:

$$\begin{aligned} &S_0 = 100, K = 100, r = 0, q = 0, \lambda = 1.5768, \eta = 0.5751, \\ &\bar{u} = 0.0398, u_0 = 0.0175, \rho = -0.5711, T = 1. \end{aligned} \tag{17}$$

In this test, we compare the COS method with the Carr-Madan method. The option price reference values are obtained by the Carr-Madan method using $N = 2^{17}$ points, and the truncated Fourier domain is set to [0, 1,200] for the experiment with $T = 1$.

We mimic the calibration situation and price several strikes simultaneously. We choose $T = 1$ and 21 consecutive strikes, $K = 50, 55, 60, \ldots, 150$, see the results in Table 1. The maximum error over all strike prices is presented. Note the very different values of N, that the two methods require for satisfactory convergence. With $N = 160$, the COS method can price all options for 21 strikes highly accurately, within 3 ms. The COS method appears to be approximately a factor 20 faster than the Carr-Madan method for the same level of accuracy.

Table 1. Error convergence and cpu time for puts under the Heston model by the COS and Carr-Madan method, pricing 21 strikes, with $T = 1$, parameters as in (17)

	N	32	64	96	128	160
COS	cpu time (ms)	0.85	1.45	2.04	2.64	3.22
	max. abs. err.	1.43×10^{-1}	6.75×10^{-3}	4.52×10^{-4}	2.61×10^{-5}	4.40×10^{-6}
	N	512	1,024	2,048	4,096	8,192
Carr-Madan	cpu time (ms)	7.44	12.84	20.36	37.69	76.02
	max. error	4.70×10^{6}	6.69×10^{1}	2.61×10^{-1}	2.15×10^{-3}	2.08×10^{-7}

3 Conclusions and Discussion

In this paper we have discussed an option pricing method based on Fourier-cosine series expansions, the COS method, for European-style options. The method can be used as long as a characteristic function for the underlying price process is available. The COS method is based on the insight that the series coefficients of many density functions can be accurately retrieved from their characteristic functions. The computational complexity of the COS method is linear in the number of terms, N, chosen in the Fourier-cosine series expansion. Very fast computing times were reported here for the Heston model.

References

1. Carr, P.P., Madan D.B.: Option valuation using the fast Fourier transform. J. Comp. Finance **2**, 61–73 (1999)
2. Fang, F., Oosterlee, C.W.: A novel pricing method for European options based on Fourier-cosine series expansions. SIAM J. Sci. Comput. **31**, 826–848 (2008)
3. Fang, F., Oosterlee, C.W.: Pricing early-exercise and discrete barrier options by Fourier-cosine series expansions. Numer. Math. **114**, 27–62 (2009) (http://ta.twi.tudelft.nl/mf/users/oosterle/oosterlee/bermCOS.pdf)
4. Heston, S.: A closed-form solution for options with stochastic volatility with applications to bond and currency options. Rev. Financ. Stud. **6**, 327–343 (1993)

Topology and Motion Planning Algorithms in Robotics

M. Farber

Department of Mathematical Sciences, Durham University, Durham DH1 3LE, UK, Michael.Farber@durham.ac.uk

Summary. Methods of algebraic topology are used to analyze the structure of motion planning algorithms in robotics. Navigational complexity of a mechanical system is measured by a numerical invariant $\mathsf{TC}(X)$ depending on the homotopy type of the configuration space X. Computations of $\mathsf{TC}(X)$ use various topological tools including cohomology algebras and cohomology operations. This paper is a brief survey introducing these topics to the community of applied and industrial mathematicians. The method is illustrated by applications to problems of simultaneous control of multiple moving objects.

1 Introduction

Algorithmic motion planning in robotics is a well established discipline which provides a wide variety of general-purpose efficient algorithms as well as algorithms designed for a number of fairly involved special scenarios, see [13,14,16]. One considers a moving system S with k degrees of freedom and a two or three-dimensional workspace W. The geometry of S and W is given in advance which determines the configuration space of the system, X. The latter is a subset of $\mathbf{R}^k$ consisting of all admissible placements (or configurations) of the system S, each represented by a tuple of k real parameters. Being a subset of the Euclidean space $X \subset \mathbf{R}^k$, the configuration space X naturally inherits its topology.

A motion planning algorithm takes as input the present and the desired states of the system and produces as the output a continuous motion of the system from its current state to the desired state. The topology of the configuration space X of the system imposes important restrictions on the discontinuities of the robot motion viewed as a function of the input data, see [3, 4]. The complexity of motion planning algorithms is measured by a numerical invariant $\mathsf{TC}(X)$ which has at least three different appearances in robotics applications [6]. Firstly, it is the minimal number of domains of continuity of any motion planning algorithm for a system having X as its

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*, Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_134,

configuration space; this can also be understood as the minimal number of subprograms (operating only with continuous functions) in any motion planning algorithm for the system. Secondly, it is the minimal order of instability [4] which have motion planning algorithms in X. The third interpretation allows to measure $\mathsf{TC}(X)$ while dealing with random motion planning algorithms: $\mathsf{TC}(X)$ is the minimal integer n such that there exists an n valued random motion planning algorithm for the system [5].

The phenomenon described in this work may have a significant impact only in situations when the dimension of the configuration space X is large. Since we are mainly concerned with systems operating in three-dimensional space, the high-dimensionality happens when the system contains many independently moving parts. We study in detail several specific examples of this kind: these are problems of simultaneous control of multiple moving objects which are either totally independent or have to perform their tasks avoiding collisions with each other and with obstacles.

2 The Concept $\mathsf{TC}(X)$

Let X denote the configuration space of a mechanical system. States of the system are represented by points of X, and continuous motions of the system are represented by continuous paths $\gamma : [0, 1] \to X$. Here the point $A = \gamma(0)$ represents the initial state and $\gamma(1) = B$ represents the final state of the system. We assume that X is path connected, i.e. the system can be brought to an arbitrary state from any given state by a continuous motion. We are interested in algorithms producing such motions.

Denote by $PX = X^I$ the space of all continuous paths $\gamma : I = [0, 1] \to X$. The space PX is supplied with the compact - open topology [2]. Let

$$\pi : PX \to X \times X$$

be the map which assigns to a path γ the pair $(\gamma(0), \gamma(1)) \in X \times X$ of the initial–final configurations. The map π is a fibration in the sense of Serre.

Definition 1. *A motion planning algorithm is a section of fibration π. In other words it is a (not necessarily continuous) map $s : X \times X \to PX$ satisfying $\pi \circ s = 1_{X \times X}$.*

A motion planning algorithm $s : X \times X \to PX$ is *continuous* if the suggested route $s(A, B)$ of going from A to B depends continuously on the states A and B, for all $A, B \in X$. It is not difficult to see that a continuous motion planning algorithm in X exists if and only if X is contractible. This explains why motion planning algorithms in most real life applications are discontinuous. One wants to measure and minimize discontinuities of such algorithms.

Definition 2. *A motion planning algorithm $s : X \times X \to PX$ is called tame if $X \times X$ can be split into finitely many sets $X \times X = F_1 \cup F_2 \cup \cdots \cup F_k$ such that* (i) *the restriction $s|F_i : F_i \to PX$ is continuous for all $i = 1, \ldots, k$,* (ii) *$F_i \cap F_j = \emptyset$ for $i \neq j$ and* (iii) *each F_i is an ENR (see [2]).*

Definition 3. *The topological complexity of a tame motion planning algorithm $s : X \times X \to PX$ is defined as the minimal number k appearing in all possible decompositions of Definition 2.*

Definition 4. *The topological complexity $\mathsf{TC}(X)$ of a path-connected finite-dimensional polyhedron X is defined as the minimal topological complexity of tame motion planning algorithms for X.*

The topological complexity $\mathsf{TC}(X)$ can be also defined for more general path-connected spaces which are not necessary polyhedra. For such spaces one defines $\mathsf{TC}(X)$ to be the genus in the sense of A. Schwarz [15] of the path-fibration $\pi : PX \to X \times X$. More explicitly $\mathsf{TC}(X)$ is the minimal integer k such that $X \times X$ admits an open cover $X \times X = U_1 \cup U_2 \cup \cdots \cup U_k$ such that for each $i = 1, \ldots, k$ the projections $X \leftarrow U_i \to X$ on the first and the second factor are homotopic to each other.

The notion of Schwarz genus was used by Smale [17] and Vassiliev [18] to study complexity of algorithms for solving polynomial equations.

3 Upper and Lower Bounds for $\mathsf{TC}(X)$

In many instances calculation of $\mathsf{TC}(X)$ is based on a combination of upper and lower bounds; also the homotopy invariance of $\mathsf{TC}(X)$ plays a role [7].

Theorem 1. *If X is an r-connected polyhedron where $r \geq 0$ then*

$$\mathsf{TC}(X) < \frac{2\dim(X) + 1}{r + 1} + 1.$$

The simplest lower bound uses cohomology classes $u \in H^*(X \times X; R)$ satisfying

$$u|\Delta_X = 0 \in H^*(X; R');$$

such cohomology classes are called *zero-divisors*. Here R is a local coefficient system on $X \times X$, $\Delta_X \subset X \times X$ denotes the diagonal and R' denotes $R|\Delta_X$.

The following simple construction of zero-divisors is quite useful: for an abelian group R and a cohomology class $u \in H^j(X; R)$ one has the zero-divisor (Fig. 1)

$$\bar{u} \;=\; 1 \times u - u \times 1 \;\in\; H^j(X \times X; R).$$

Theorem 2. *One has $\mathsf{TC}(X) > k$ assuming that there exist k zero-divisors $u_i \in H^*(X \times X; R_i)$ where $i = 1, \ldots, k$, whose cup-product is nonzero.*

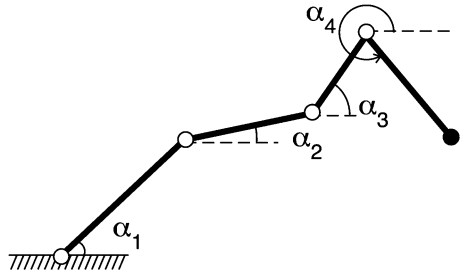


Fig. 1. Robot arm with $n = 4$ bars

Applying these results one easily finds that the topological complexity of spheres is given by $\mathsf{TC}(S^n) = 2$ for all odd n and $\mathsf{TC}(S^n) = 3$ for all even n.

If Σ_g denotes a compact orientable surface of genus g then $\mathsf{TC}(\Sigma_g) = 5$ for all $g \geq 2$ although $\mathsf{TC}(\Sigma_g) = 3$ for $g = 0, 1$.

The planar robot arm with n bars has as its configuration space the n-dimensional torus $T^n = S^1 \times S^1 \times \cdots \times S^1$. The topological complexity of T^n equals $\mathsf{TC}(T^n) = n + 1$. The configuration space of an n-bar robot arm in $\mathbf{R}^3$ is the product of n spheres $S^2 \times S^2 \times \ldots S^2$ whose topological complexity is $2n + 1$, see [3]. Explicit motion algorithms use the methods of navigation functions [7].

For an aspherical space X (i.e. assuming that $\pi_i(X) = 0$ for all $i \geq 2$) the topological complexity depends only on the fundamental group $\pi_1(X)$. Cohen and Pruidze [1] found explicitly the topological complexity of Eilenberg–MacLane spaces of right-angled Artin groups.

4 Simultaneous Control of Multiple Objects

Suppose that one controls simultaneously n systems $S_1, \ldots, S_n$. The total configuration space in the case of centralized control is the Cartesian product $X_1 \times X_2 \times \cdots \times X_n$ where X_i denotes the configuration space of S_i. For simplicity we will assume that the spaces X_i are all homeomorphic to each other $X_i \simeq X$; this assumption is valid if we control several systems having identical properties. From the product inequality [3] one obtains the inequality

$$\mathsf{TC}(X_1 \times X_2 \times \cdots \times X_n) \leq n \cdot [\mathsf{TC}(X) - 1] + 1.$$

Denote by $\mathrm{zcl}(X)$ the zero-divisors cup-length of X, i.e. the longest nontrivial cup-product of zero-divisors. Then

$$\mathsf{TC}(X_1 \times X_2 \times \cdots \times X_n) > n \cdot \mathrm{zcl}(X).$$

Comparing these two inequalities we see that the topological complexity of centralized control is bounded above and below by functions linear in n.

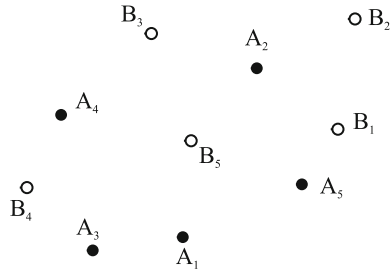


Fig. 2. Two (*black and white*) configurations of $n = 5$ moving objects in $\mathbf{R}^m$

Let us now consider algorithms of distributed control, i.e. when each of the controllable objects has its own motion planning algorithms and behaves independently of the others. The motion planning algorithm of the ith object is given by a splitting $F_1^i \cup F_2^i \cup \cdots \cup F_{s_i}^i = X \times X$ and by defining a continuous section $\sigma_j^i : F_j^i \to PX$ for $j = 1, \ldots, s_i$. Here clearly $s_i \geq \mathsf{TC}(X)$. The domains of continuity for the system of n objects are of the form $F_{r_1}^1 \times F_{r_2}^2 \times \cdots \times F_{r_n}^n$ where $1 \leq r_i \leq s_i$. We see that any distributed motion planning algorithm has at least $s_1 s_2 \ldots s_n \geq \mathsf{TC}(X)^n$ domains of continuity.

Corollary 1. *The topological complexity of centralized control of n identical objects growths as a linear function of n as $n \to \infty$. However, any distributed motion planning algorithm has exponential in n topological complexity (which is bounded below by $\mathsf{TC}(X)^n$).*

Thus, the centralized control has significant advantages compared to the distributed control. It is important to emphasize that specific motion planning algorithms for centralized control with complexity linear in n can be designed using the technique developed in the proof of the product inequality [3] (Fig. 2).

5 Collision Free Motion Planning

Next we consider the problem of controlling n objects moving in $\mathbf{R}^m$ without collisions. We will assume that the objects are represented by points $A \in \mathbf{R}^m$. An allowed configuration of n objects is labeled by n distinct points in $\mathbf{R}^m$, and the configuration space in this case is

$$F(\mathbf{R}^m, n) = \{(A_1, A_2, \ldots, A_n);\ A_i \in \mathbf{R}^m, A_i \neq A_j \text{ for } i \neq j\}.$$

A motion planning algorithm for this problem decides how to move one configuration to another so that no collisions occur in the process of motion. The topological complexity of $F(\mathbf{R}^m, n)$ was computed in [10] using the theory of hyperspace arrangements:

$$\mathsf{TC}(F(\mathbf{R}^m, n)) = \begin{cases} 2n-1 \text{ for any odd } m, \\ 2n-2 \text{ for any even } m. \end{cases}$$

The case when $m \geq 4$ even was resolved in a recent preprint [8]. It is a challenging problem to find explicit motion planning algorithms in $F(\mathbf{R}^m, n)$ having linear in n topological complexity. In [6] we suggested an algorithm which has complexity quadratic in n.

Paper [9] studies collision free motion planning algorithms for multiple moving objects in the presence of moving obstacles.

Consider now the case when the controlled objects move along a graph Γ. The study of configuration spaces $F(\Gamma, n)$ was initiated by Ghrist and Koditschek [11, 12]. The topological complexity of this configuration space satisfies

$$\mathsf{TC}(F(\Gamma, n)) \leq 2m(\Gamma) + 1,$$

see [5]. Here $m(\Gamma)$ denotes the number of essential vertices of Γ, i.e. vertices incident to at least 3 edges. We conclude that the complexity of collision free control of multiple objects moving along a graph is bounded above by a constant independent of n. This result contrasts greatly the computation of $\mathsf{TC}(F(\mathbf{R}^m, n))$ (see above) which is linear in n.

References

1. Cohen, D., Pruidze, G.: Bull. LMS **40**, 249–262 (2008)
2. Dold, A.: Lectures on Algebraic Topology. Springer, New York (1972)
3. Farber, M.: Discr. Comput. Geom. **29**, 211–221 (2003)
4. Farber, M.: Topol. Appl. **40**, 245–266 (2004)
5. Farber, M.: In: Erdmann, M., et al. (eds.) Algorithmic Foundations of Robotics IV, pp. 123–138 Springer, Berlin (2005)
6. Farber, M.: In: Biran, P., et al. (eds.) Morse Theoretic Methods in Nonlinear Analysis and in Symplectic Topology, pp. 185–230 Springer, Berlin (2006)
7. Farber, M.: Invitation to Topological Robotics. EMS, Zürich (2008)
8. Farber, M., Grant, M.: Preprint, arXiv:0806.4111
9. Farber, M., Grant, M., Yuzvinsky, S.: In: Farber, M., et al. (eds.) Topology and Robotics – Contemporary Mathematics No. 438, pp. 75–83. AMS, Providence, RI (2007)
10. Farber, M., Yuzvinsky, S.: Am. Math. Soc. Transl. **212**, 145–156 (2004)
11. Ghrist, R.: Knots, braids, and mapping class groups. AMS/IP Stud. Adv. Math. **24**, 29–40 (Am. Math. Soc., Providence, 2001)
12. Ghrist, R., Koditschek, D.: SIAM J. Control Optim. **40**, 1556–1575 (2002)
13. Halperin, D., Sharir, M.: In: Goldberg, K., et al. (eds.) The Algorithmic Foundations of Robotics, pp. 495–511, Boston, MA, 1995
14. Latombe, J.-C.: Robot Motion Planning. Kluwer, Norwell, MA (1991)
15. Schwarz, A.S.: Am. Math. Soc. Transl. **55**, 49–140 (1966)
16. Sharir, M.: In: Goodman, J.E., et al. (eds.) Handbook of Discrete and Computational Geometry. CRC, Boca Raton, FL (1997)
17. Smale, S.: J. Complexity **3**, 81–89 (1987)
18. Vassiliev, V.A.: Funct. Anal. Appl. **22**, 15–24 (1988)

Some Hints on Finding the Most Important Components in a System

Josep Freixas and Montserrat Pons

Department of Applied Mathematics, Technical University of Catalonia (UPC), Barcelona, Spain, josep.freixan@upc.edu, montserrat.pons@upc.edu

Summary. Some ideas for selecting the most important component(s) in a system, from a reliability point of view, are given. This is done using algebraic tools and taking only into account the structure of the system and the ranking of component reliabilities. Three measures of component importance are considered: Birnbaum, Improvement Potential and Risk Achievement Worth.

1 Introduction

The aim of this paper is to present some hints for finding the component(s) in a system that have the highest reliability importance (under some measure), when only the structure function of the system, and the ranking of component reliabilities (but possibly not their exact values) are known. To help identifying these (most important) components, we use some binary relations defined on the set of nodes of the system, i.e., binary relations that do not depend on the nature of the components, but only depend on the structure of the system. The reliability importance measures considered in the paper are: Birnbaum, Improvement potential and Risk achievement worth.

Some work has previously been done in similar contexts. In particular, the binary relation called the "criticality relation" was introduced in [1] as a tool to find an optimal component arrangement that maximizes system reliability when all components have identical reliability. Other relations between nodes have also been considered in the literature [2, 5]. These relations were used in [7–9] to compare Birnbaum measures of system components.

Two new pre-orderings between nodes (stronger than the criticality relation) were introduced in [3, 4], and it was proved there that they are useful to compare component reliability importance, not only for Birnbaum's but for other importance measures, when components reliabilities are not all equal.

We consider binary systems, i.e., systems in which there is a random Boolean variable associated to each node which takes the value 1 if the component placed on it is functioning and 0 otherwise. These random variables are

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*, Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_135,

assumed to be statistically independent. The structure function of the system is a boolean function of these random variables that takes the value 1 if the system is functioning and 0 otherwise. The systems we consider are coherent, i.e., systems with nondecreasing structure function and with all components being relevant. The reliability of a system, i.e., the probability of it being functioning, is given by the expectation of its structure function but it can also be expressed in terms of its path sets. This is why we usually denote a system by (N, π), where $N = \{1, 2, \ldots, n\}$ is the set of nodes and π is the collection of path sets. Recall that a path set is a subset of nodes such that if all of them are operative then the system is functioning.

From now on we assume that (N, π) is a coherent system with n components, $\mathbf{p} = (p_1, p_2, \ldots, p_n) \in (0, 1)^n$ is a vector of component reliabilities, and $h(\mathbf{p})$ is the reliability of the system, with $0 < h(\mathbf{p}) < 1$.

Along the paper we will consider three reliability measures of component importance. The formal definitions of these measures are given next.

Definition 1 (Reliability Importance Measures). *For a component $i \in N$ the following measures are considered:*

- $I_i^B(\mathbf{p}) = h(1_i, \mathbf{p}) - h(0_i, \mathbf{p}) = \dfrac{\partial h}{\partial p_i}(\mathbf{p})$ *(Birnbaum)*
- $I_i^{IP}(\mathbf{p}) = h(1_i, \mathbf{p}) - h(\mathbf{p})$ *(Improvement Potential)*
- $I_i^{RAW}(\mathbf{p}) = \dfrac{1 - h(0_i, \mathbf{p})}{1 - h(\mathbf{p})}$ *(Risk Achievement Worth)*

The organization of the paper is as follows. In the next section, three pre-orderings in the set of nodes are defined and their properties are given. The main results are included in Sect. 3, where we compare the importance of two components linked by the some of the pre-orderings, for each one of the reliability importance measures considered in the paper. The results in this section allow us to describe systems in which the most important components can be identified independently of the values of their component reliabilities and depending only on their ordering. Some final comments end the paper in Sect. 4.

2 The Algebraic Tools

We are going to introduce three binary relations between nodes. The first one of them is known as the desirability relation in game theory (see, e.g., [6,10]), or as the criticality relation in reliability theory (see [1]). The other two binary relations (the external and the internal domination relations) were introduced in [3]. The three considered binary relations are pre-orderings, i.e., they are reflexive and transitive.

Definition 2 (Pre-Orderings on the Set of Components).

- *The* **criticality relation**
  $i \succeq j$ *iff* $[S \cup \{j\} \in \pi \ \Rightarrow \ S \cup \{i\} \in \pi,$ *whenever* $S \subseteq N \backslash \{i, j\}]$.
  If $i \succeq j$ *we say that component* i *is at least as critical as component* j.
- *The* **external domination relation**.
  $i \vDash j$ *iff* $i = j$, *or* $[S \in \pi, j \in S, i \notin S \ \Rightarrow \ S \backslash \{j\} \in \pi\,]$.
  If $i \vDash j$, *we say that component* i *externally dominates* j.
- *The* **internal domination relation**.
  $i \trianglerighteq j$ *iff* $i = j$, *or* $[S \in \pi, i, j \in S \ \Rightarrow \ S \backslash \{j\} \in \pi\,]$.
  If $i \trianglerighteq j$, *we say that component* i *internally dominates* j.

Other binary relations defined from the former ones will also be used in this paper, and they are put together in the following definition.

Definition 3 (Strict and Equivalence Relations).

$i \succ j$ *iff* $i \succeq j$ *and* $j \not\succeq i$; $i \equiv j$ *iff* $i \succeq j$ *and* $j \succeq i$.
$i \vdash j$ *iff* $i \vDash j$ *and* $j \nvDash i$; $i \Subset j$ *iff* $i \vDash j$ *and* $j \vDash i$.
$i \rhd j$ *iff* $i \trianglerighteq j$ *and* $j \not\trianglerighteq i$; $i \bowtie j$ *iff* $i \trianglerighteq j$ *and* $j \trianglerighteq i$.

The following proposition states that the external and the internal domination relations extend to (imply) the criticality relation.

Proposition 1.

$i \vDash j \Rightarrow i \succeq j$ *and* $i \vdash j \Rightarrow i \succ j$
$i \trianglerighteq j \Rightarrow i \succeq j$ *and* $i \rhd j \Rightarrow i \succ j$.

In general the three pre-orderings are neither antisymmetric (i.e., they are not ordering relations) nor total (i.e., there exist incomparable elements). The following example illustrates most of these properties.

Example 1. Let $N = \{1, 2, 3, 4, 5, 6\}$ be the set of components of the system in Fig. 1. The minimal path sets are: $\{1, 2, 3\}, \{1, 4\}, \{5\}, \{6\}$.

In this system the relations between components are:

$6 \equiv 5 \succ 1 \succ 4 \succ 2 \equiv 3$; $1 \vdash 2 \Subset 3,\ 1 \vdash 4$; $6 \bowtie 5 \rhd 4 \rhd 3,\ 4 \rhd 2$.

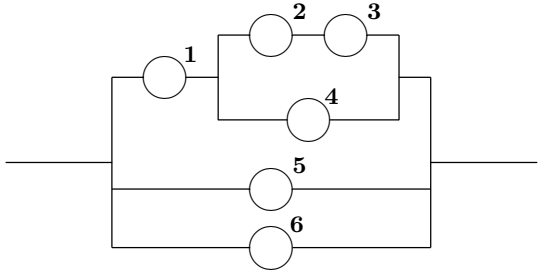


Fig. 1. System in Example 1

3 Ranking of Component Importance

The results in this section provide useful hints for comparing the importance (for the different measures) of two components linked by some of the considered binary relations. As a consequence, they allow us to identify the most important components in some systems independently of the values of their reliabilities and depending only on their ordering.

Criticality relation has been used to compare the Birnbaum measure between system components in [7–9], and the results are reproduced here for the sake of completeness. We proved the results for the Birnbaum measure based on the external or the internal domination relations, and also all statements referring to the other reliability importance measures in [3,4].

Theorem 1 (Birnbaum Measure). *Let i, j be different elements in N. Then,*

$$\begin{aligned}
&i \succeq j \text{ and } p_i = p_j \Rightarrow I_i^B(\mathbf{p}) \geq I_j^B(\mathbf{p})\\
&i \succ j \text{ and } p_i = p_j \Rightarrow I_i^B(\mathbf{p}) > I_j^B(\mathbf{p})\\
&i \vDash j \text{ and } p_i = p_j \Rightarrow I_i^B(\mathbf{p}) = I_j^B(\mathbf{p})\\
&i \vDash j \text{ and } p_i < p_j \Rightarrow I_i^B(\mathbf{p}) > I_j^B(\mathbf{p})\\
&i \trianglerighteq j \text{ and } p_i = p_j \Rightarrow I_i^B(\mathbf{p}) = I_j^B(\mathbf{p})\\
&i \trianglerighteq j \text{ and } p_i > p_j \Rightarrow I_i^B(\mathbf{p}) > I_j^B(\mathbf{p})
\end{aligned}$$

Theorem 2 (Improvement Potential). *Let i, j be different elements in N. Then,*

$$\begin{aligned}
&i \succeq j \text{ and } p_i \leq p_j \Rightarrow I_i^{IP}(\mathbf{p}) \geq I_j^{IP}(\mathbf{p})\\
&i \succ j \text{ and } p_i \leq p_j \Rightarrow I_i^{IP}(\mathbf{p}) > I_j^{IP}(\mathbf{p})\\
&i \succeq j, i \not\trianglerighteq j \text{ and } p_i < p_j \Rightarrow I_i^{IP}(\mathbf{p}) > I_j^{IP}(\mathbf{p})\\
&i \bowtie j \Rightarrow I_i^{IP}(\mathbf{p}) = I_j^{IP}(\mathbf{p})
\end{aligned}$$

Theorem 3 (Risk Achievement Worth). *Let i, j be different elements in N. Then,*

$$\begin{aligned}
&i \succeq j \text{ and } p_i \geq p_j \Rightarrow I_i^{RAW}(\mathbf{p}) \geq I_j^{RAW}(\mathbf{p})\\
&i \succ j \text{ and } p_i \geq p_j \Rightarrow I_i^{RAW}(\mathbf{p}) > I_j^{RAW}(\mathbf{p})\\
&i \succeq j, i \not\vDash j \text{ and } p_i > p_j \Rightarrow I_i^{RAW}(\mathbf{p}) > I_j^{RAW}(\mathbf{p})\\
&i \Subset j \Rightarrow I_i^{RAW}(\mathbf{p}) = I_j^{RAW}(\mathbf{p})
\end{aligned}$$

Example 1 revisited. Let's consider again the system in Example 1 to see how these results apply. Assume, for example, that $\mathbf{p} = (p_1, p_2, \ldots, p_6)$ is such that $p_6 < p_1 = p_4 = p_5 < p_3 < p_2$ and that we are looking for the most important components in this case. To alleviate the notation we will write I_i instead of $I_i(\mathbf{p})$ throughout the example, for the different measures considered.

For the Birnbaum measure (using Theorem 1): Taking into account that $5 \succ 1 \succ 4$ and $p_5 = p_1 = p_4$ we have $I_5^B > I_1^B > I_4^B$. The fact that $1 \vdash 2 \Subset 3$ and $p_1 < p_3 < p_2$ implies $I_1^B > I_3^B > I_2^B$. Finally, since $6 \bowtie 5$ and $p_5 > p_6$ then, $I_5^B > I_6^B$. Consequently, the most important component for this measure is 5.

For the Improvement Potential measure (using Theorem 2): Since $5 \succ 1 \succ 4$ and $p_5 = p_1 = p_4$ then $I_5^{IP} > I_1^{IP} > I_4^{IP}$. Moreover, $4 \succ 3$ and $p_4 < p_3$ implies $I_4^{IP} > I_3^{IP}$. The fact that $3 \succeq 2$, $3 \not\trianglerighteq 2$ and $p_3 < p_2$ ensures that $I_3^{IP} > I_2^{IP}$, and, finally, $6 \bowtie 5$ implies $I_5^{IP} = I_6^{IP}$. The most important components for this measure are 5 and 6.

For the Risk Achievement Worth measure (using Theorem 3): The fact that $5 \succ 1 \succ 4$ and $p_5 = p_1 = p_4$ implies $I_5^{RAW} > I_1^{RAW} > I_4^{RAW}$. Since $5 \succeq 6$, $5 \not\models 6$ and $p_5 > p_6$ it is $I_5^{RAW} > I_6^{RAW}$, and, finally, $2 \models\!\!\models 3$ implies $I_2^{RAW} = I_3^{RAW}$. The most important components for this measure belong to the set $\{2, 3, 5\}$. In this case we would need to know the values of p_2, p_3 and p_5 to decide which one of them is the most important for this measure.

4 Some Final Comments

In this section we summarize the main results in the paper. They illustrate the influence of the different pre-orderings on the ranking of components importance, for the three measures considered in it.

Under the assumption of identical reliabilities of components, Theorems 1–3 tell us that the ordering of the components by all the reliability importance measures considered coincides with the ranking given by the node criticality relation. Moreover, if two components are related by either the internal or the external domination relations, their Birnbaum measures coincide.

If component reliabilities are not the same, Theorem 1 states that if two components i, j are comparable by one of the domination relations, and their reliabilities p_i, p_j satisfy the adequate (in)equality, then we can determine the ordering between their Birnbaum measures without computing them. Similar results are shown in Theorems 2 and 3 using the node criticality relation and referring to the other considered measures. All these results can be useful in the case that the reliabilities of components could be modified (for redundancy or other reliability actions) to increase system reliability, and a component has to be selected according to some importance measure.

For nodes equivalent by the criticality relation, it is clear from Theorem 1 that their Birnbaum measures are the same if their reliabilities coincide. The same happens for the other two considered measures, as shown in Theorems 2 and 3. But if the component reliabilities are not equal then the lack of domination relation between the components is decisive. More precisely, $i \not\trianglerighteq j$ and $p_i < p_j$ imply $I_i^{IP}(\mathbf{p}) > I_j^{IP}(\mathbf{p})$, and, similarly, $i \not\models j$ and $p_i > p_j$ imply $I_i^{RAW}(\mathbf{p}) > I_j^{RAW}(\mathbf{p})$. These results are applicable, among others, in k-out-of-n systems, and they prove that in general k-out-of-n systems the ordering of component reliabilities completely determines the ordering of component importance for the three measures considered.

Acknowledgements

Research partially supported by Grant MTM 2006–06064 of "Ministerio Español de Educación y Ciencia, y el Fondo Europeo de Desarrollo Regional" and Grant SGRC 2005–00651 of "Generalitat de Catalunya."

References

1. Boland, P.J., Proschan, F., Tong, Y.L.: Optimal arrangements of components via pairwise rearrangements. Nav. Res. Log. **36**, 807–815 (1989)
2. Butler, D.A.: A complete importance ranking for components of binary coherent systems with extensions to multi-state systems. Nav. Res. Log. Q. **4**, 565–578 (1979)
3. Freixas, J., Pons, M.: Identifying optimal components in a reliability system. IEEE Trans. Reliab. **57**(1), 163–170 (2008)
4. Freixas, J., Pons, M.: The influence of the node criticality relation on some measures of components importance. Oper. Res. Lett. **36**, 557–560 (2008)
5. Hwang, F.K.: A hierarchy of importance indices. IEEE Trans. Reliab. **54**(1), 169–172 (2005)
6. Isbell, J.R.: A class of majority games. Q. J. Math. Oxford Ser. **7**(2), 183–187 (1956)
7. Meng, F.C.: Some further results on ranking the importance of system components. Reliab. Eng. Syst. Saf. **47**(2), 97–101 (1995)
8. Meng, F.C.: Comparing Birnbaum importance measure of system components. Probab. Eng. Inform. Sci. **18**, 237–245 (2004)
9. Mi, J.: A unified way of comparing the reliability of coherent systems. IEEE Trans. Reliab. **52**(1), 38–43 (2003)
10. Taylor, A.D., Zwicker, W.S.: Simple Games: Desirability Relations, Trading, and Pseudoweightings. Princeton University Press, Princeton, NJ (1999)

An Advanced Aeroelastic Model for Horizontal Axis Wind Turbines

F. Frunzulica[1,2], H. Dumitrescu[2], A. Dumitrache[2], and V. Cardos[2]

[1] "POLITEHNICA" University from Bucharest, Bucharest, Romania
ffrunzi@aero.pub.ro
[2] "Gheorghe Mihoc – Caius Iacob" Institute of Mathematical Statistics and Applied Mathematics, Bucharest, Romania, horia.dumitrescu@ima.ro, alexandru.dumitrache@ima.ro, vladimir.cardos@ima.ro

Summary. In this paper an advanced aeroelastic numerical tool for horizontal axis wind turbines is presented. The tool is created by non-linear coupling an unsteady aerodynamic model based on the lifting-line approximation with an elastodynamic model based on the beam approximation. The aero-to-elastic interface defines the loads exercised on the structure, whereas the elastic-to-aero interface transmits the rates of deformations. The aeroelastic model is evaluated through comparisons of its predictions with experimental data as well as with predictions obtained by simpler models.

1 Introduction

The design problem of horizontal axis wind turbines (HAWT) is related to two dominant model problems: the aerodynamic problem and the elastodynamic one. Their combination leads to the aeroelastic problem of a horizontal axis wind turbine. Input to this problem is the wind inflow conditions.

In this work, the authors present a non-linear advanced aeroelastic model based on a lifting line model as regards the aerodynamics, and a beam structural model adapted to this problem, useful to a number of design problems.

2 The Numerical Method

The key point of the approach adopted herein is based on the formulation of the aeroelastic problem as an appropriate coupling of two different problems: the aerodynamic and the elastodynamic.

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*,
Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_136,

2.1 The Aerodynamic Model

The response of an horizontal axis wind turbine to dynamic excitation is a special case of the aerodynamic performance problem. In this connection, a computational environment based on a lifting line method [2, 3], and a semiempirical dynamic stall model of Leishman and Beddoes [4] has been developed. This method is an unsteady model based on the vorticity filament approximation of the vorticity on blades.

For the classical prescribed wake-lifting line blade model, the unknown quantity is the spanwise bound circulation distribution. The relationship for the section bound circulation (Γ) can be shown to be related to the local velocity (U), section chord (c), and section lift coefficient (C_L) through the Kutta-Joukowski theorem,

$$\Gamma = \frac{1}{2}\, c\, C_L(\alpha)\; U \tag{1}$$

The local velocity and effective angle of attack (α) are functions of the tangential velocity (U_T), the local axial and azimuthal induced velocities (u, w), wind velocity (V_w), and blade pitch angle (θ_p) ,

$$U = \left[(U_T + w)^2 + (V_w - u)^2\right]^{1/2}, \alpha = -\theta_p + \tan^{-1}\left(\frac{V_w - u}{U_T + w}\right), U_T = \Omega\; r \tag{2}$$

The components of the induced velocity for a given wake geometry, at control point (i) and at any time $k\,\Delta t$, can be shown to be a function of the bound circulation distribution and individual vortex filament influence coefficients (GC) as

$$u_i^{(k)} = \frac{1}{4\pi R}\sum_{j=1}^{N} GC_{u,ij}^{(k)}\; \Gamma_j^{(k)}, w_i^{(k)} = \frac{1}{4\pi R}\sum_{j=1}^{N} GC_{w,ij}^{(k)}\; \Gamma_j^{(k)} \tag{3}$$

where N is the number of blade inflow solution stations [3]. The total lift force coefficient is given by sum of circulatory and noncirculatory components under attached flow conditions

$$C_L^{\beta}(t) = C_{LI}(t) + C_{LC}(t) \tag{4}$$

and by sum of nonlinear separation and vortex components under dynamic stall conditions [4]

$$C_L^{s}(t) = C_{LF}(t) + C_{LV}(t) \tag{5}$$

The nonlinear solution is based on the linearization of the relationships given above ((1), (2), (3), and (4) or (5)) to form a system of linear equations which can be corrected for the actual nonlinearities of the problem by a lagged iteration procedure. Since the induced velocities are also functions of the circulation distribution (3), the equation at the ith blade segment (1) can be reformulated as

$$\Gamma_i = \frac{1}{2}\, c\, a \left(U_T \overline{\theta} - \frac{1}{4\pi R} \sum_{j=1}^{N} G C_{u,ij} \Gamma_j \right) + \frac{1}{2} c\, a\, C_{f,i} \tag{6}$$

where a is the linear lift curve slope, $\bar{\theta} = -\theta_p - \alpha_0 + V_w/U_T$, α_0 is the zero lift offset angle, and C_f is the correction to the linearized equations for the nonlinearities of the actual problem,

$$C_f = U\ (C_L/a) - U_T\,(-\theta_p - \alpha_0 + (V_w - u)/U_T)$$

This equation can be written for each blade segment, and thus, a system of simultaneous linear equations results if the correction term (C_f) is assumed known. In matrix form, this can be expressed as

$$[A]\,[\Gamma^n] = [B] - \left[C\ \Gamma^{n-1}\right] \tag{7}$$

where $[A]$ is the matrix of influence coefficients and $\left[C\ \Gamma^{n-1}\right]$ is a correction vector calculated based on the circulation solution from the previous iteration. Once (7) has been solved, the circulation distribution on blades is known at the present time and the foregoing procedure may be repeated to obtain the solution at future times. The last phase of the computations consists of calculations of blade forces and the performance.

2.2 The Elastodynamic Model

The aspect ratio of wind turbine blades is, usually large and, therefore, beam theory can be used to describe, the elastodynamic behaviour of the blade. Let $O\,[X_e, Y_e, Z_e]$ denote the beam coordinate system, and it is assumed that the elastic axis is straight and coincides with axis Y_e. In this model three types of deformations are introduced: $\delta_x(y)$ – the bending deformation along X_e direction (flapwise bending), $\delta_z(y)$ – the bending deformation along Z_e direction (leadwise bending) and $\theta_y(y)$ – the spanwise torsional deformation.

The first step in structural computation is to calculate beam cross-sectional properties of thin-walled beam, multicell, nonhomogeneous, closed sections, within the framework of Bernoulli's bending theory and St. Venant torsion theory [5, 7]. The key idea is the approximation of the airfoil's shape by ne straight, homogeneous elements. The thickness of every element is considered constant and is evenly distributed across the two sides of its midline.

At the beginning, the element coordinates can be given with respect to any coordinate system, but after the calculation of the elastic centre coordinates, we switch to the elastic coordinate system.

The finite element technique. By using Lagrange equations the following linear equations of motion are obtained [1, 8]

$$\boldsymbol{M}\,\ddot{\boldsymbol{D}}_n + \boldsymbol{C}\,\dot{\boldsymbol{D}}_n + \boldsymbol{K}\,\boldsymbol{D}_n = \boldsymbol{R}_n^{ext} \tag{8}$$

where $\boldsymbol{M} = \rho \int_V \boldsymbol{N}^T \boldsymbol{N}\, dV$ is the mass matrix, $\boldsymbol{C} = \int_V \boldsymbol{N}^T \boldsymbol{C}^* \boldsymbol{N}_d\, dV$ is the structural damping matrix, $\boldsymbol{K} = \int_V \boldsymbol{N}_d^T \boldsymbol{E}\, \boldsymbol{N}_d\, dV$ is the stiffness matrix, $\boldsymbol{R}^{ext}$ is the load vector, $\boldsymbol{D}$ is the displacement vector, which contain the degrees of freedom, $\boldsymbol{N}_d$ – the derivative matrix of shape functions, and $\boldsymbol{N}$ – the matrix of shape functions (for the beam element the shape functions most commonly used are the third-degree polynomials and the first degree in the case of torsion) [8].

The time advancement of (8) with the appropriate initial conditions is performed with the specialized algorithm (Crank-Nicolson) method [3]. This is an unconditionally stable implicit one-step method, which is second order accurate in time.

2.3 Coupling Models

The solution of the aeroelastic problem requires the coupling of an aerodynamic and an elastodynamic model. In previous paragraphs a brief description of each part was done separately. In this paragraph the basic principles of the communication between the two parts will be discussed.

As regards the elastodynamic part, the load vector must be input. This vector is calculated by superimposing the gravitational forces on the aerodynamic loads. The quantities that have to be transferred from the aerodynamic part are, therefore, the aerodynamic forces that act on the blade.

The solution of (8) yields the vector $\boldsymbol{D}$ of the deformations, the vector $\dot{\boldsymbol{D}}$ of the deformation rates and the vector $\ddot{\boldsymbol{D}}$ of the accelerations at the nodes of the beam that simulates the blade.

The main modules can be summarized in the following flowchart Fig. 1.

The flowchart of the aeroelastic code has the following steps: (a) Initialize code; (b) Perform same pure aerodynamic steps for the calculation of the circulation distribution; (c) Time marching scheme. For every time step: (c.1) start time step; (c.2) calculate the circulation distribution; (c.3) calculate the force and the velocity distribution on the blades; (c.4) perform elastodynamic

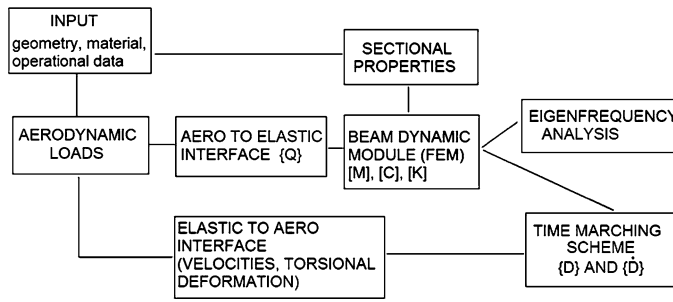


Fig. 1. The flowchart of the coupling between aerodynamic and elastodynamic models

Table 1. Comparison between experiment and numerical results

	Case 1		Case 2	
	Experiment	Aeroelastic code	Experiment	Aeroelastic code
U_∞ (m/s)	12.5	12.5	8.7	8.7
$t_{1,st}$ (s)	4.70	14.0	2.0	21.0
$t_{1,end}$ (s)	6.00	15.3	2.5	21.5
$t_{2,st}$ (s)	34.58	24.0	32.0	34.0
$t_{2,end}$ (s)	35.7	25.12	32.7	34.7
θ_1 (deg)	−1.164	−1.164	−0.07	−0.07
θ_2 (deg)	−3.19	−3.19	−3.716	−3.716

steps for a period of time equal to aerodynamic time step; (c.5) circulation calculation step; (c.6) go to next time step. The only communication between the two parts is in step (c.4) where the aerodynamic forces are imposed on the beam and the elastodynamic problem is solved.

3 Results

The results presented in the sequel concern the two cases of double pitch steps for the Tjaereborg HAWT, for which experimental data are available [6]. The parameters used for each case are (Table 1) : the inflow velocity – U_∞(m/s), the starting time of first pitch step $t_{1,st}$ (s), the ending time of first pitch step $t_{1,end}$ (s), the starting time of second pitch step $t_{2,st}$ (s), the ending time of second pitch step $t_{2,end}$ (s), the initial pitch angle θ_1 (deg), the pitch angle after the first pitch step θ_2 (deg). The numerical results are shown in Fig. 2.

4 Conclusions

A complete aeroelastic tool has been presented together with its self consistency tests and some results. In this stage, we cannot conclude on its accuracy. However, the experience suggests that this could be expected.

There are three points that must be underlined: (1) in some tests it appeared necessary to introduce artificial damping; (2) the coupling, within the context of approach described, must be non-linear and (3) the computational effort required to couple the aerodynamics with the structural part, is small compared to the whole.

Prospective work: we will make a most elaborate model based on the coupling of the aerodynamic model with a structural code based on the shell model and composite materials.

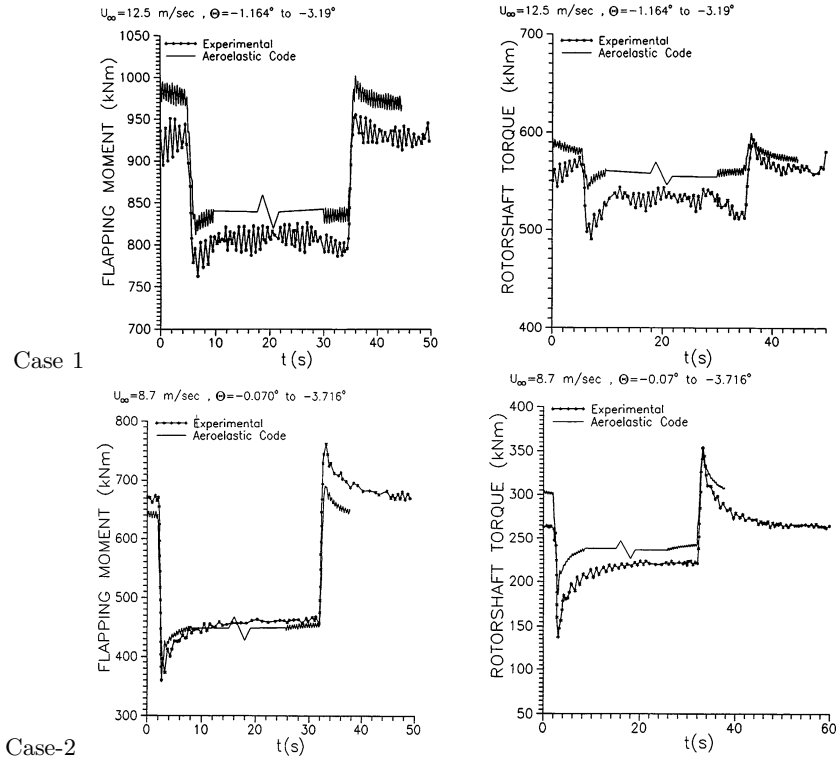


Fig. 2. Comparison between experiment and aeroelastic code

References

1. Bisplinghoff, R.L., Ashley, H., Halfman, R.: Aeroelasticity. Dover, New York (1996)
2. Dumitrescu, H., Cardos, V.: Int. J. Appl. Mech. Eng. **9**(4), 675–686 (2004)
3. Dumitrescu, H., Cardos, V., Frunzulica, F., Dumitrache, A.: Aerodynamics, Aeroelasticity and Aeroacoustics for Wind Turbines. Romanian Academy, Bucharest, Romania (2007)
4. Leishman, J.G., Beddoes, T.S.: J. Am. Helicopter Soc. **34**(3), 3–17 (1989)
5. Megson, T.H.G.: Aircraft Structures for Engineering Students. Edward Arnold, London (1977)
6. Oye, S.: Tjaereborg Wind Turbine: Fifth Dynamic Inflow Measurement. AFM VK-233, Department of Fluid Mechanics, Technical University of Denmark, Lyngby, Denmark (1992)
7. Timoshenko, S., Goodier, J.N.: Theory of Elasticity. McGraw-Hill, New York (1992)
8. Zienkiewiecs, O.C.: The Finite Element in Engineering Science. McGraw-Hill, London (1971)

On One Nonlinear Mathematical Model for Intensive Steel Quenching and Its Analytical Solution in Closed Form

Sh.E. Guseynov[1,2], J.S. Rimshans[1], and N.I. Kobasko[3]

[1] Institute of Mathematical Sciences and Information Technologies, University of Liepaja, Liepaja LV3401, Latvia, janis.rimsans@liepu.lv
[2] Transport and Telecommunication Institute, Riga LV1019 sh.e.guseinov@inbox.lv
[3] FASM, IQ Technologies, Inc., PO Box 1787, Ohio 44309-1787, USA nikobasko@yahoo.com

Summary. Evaluation of non-stationary nucleate boiling can be done on the basis of solving hyperbolic heat conductivity equation with the nonlinear boundary condition responsible for the process of nucleate boiling. Results of calculations can be used for modification of IQ-2 method of quenching, which is environment friendly, less expensive process, which significantly improves service life of steel parts.

1 Mathematical Statement of the Problem in General Form and Its Solution

As it is well known there is intensive IQ-2 quenching method, which is based on non-stationary, nucleate boiling (self-regulated thermal process) [1, 5]. At present time there are some questions, which are discussed between mathematicians: what kind of equations (namely, the parabolic equation or hyperbolic heat equations) should be used to get the best results of calculations conformed to experimental data? In [2–4] the following direct problem was considered: it is necessary to determine the function $u(x,t)$ that satisfies the hyperbolic heat transfer equation:

$$\frac{\partial u(x,t)}{\partial t} + \tau_r \frac{\partial^2 u(x,t)}{\partial t^2} = a^2 \frac{\partial^2 u(x,t)}{\partial x^2} + f(x,t), \quad (1)$$
$$0 < x < l < \infty,\ 0 < t \le T < \infty,$$

where $a^2 = \frac{k}{c \cdot \rho}, k, c, \rho, \tau_r \equiv const > 0$, the initial conditions:

$$u(x,t)|_{t=0} = u_0(x), \quad 0 \le x \le l, \quad (2)$$

$$\left.\frac{\partial u(x,t)}{\partial t}\right|_{t=0} = u_1(x), \quad 0 \le x \le l, \quad (3)$$

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*, Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_137,

the boundary conditions:

$$-k\frac{\partial u\left(x,t\right)}{\partial x}\bigg|_{x=0}+\beta^{m}\left\{u\left(x,t\right)|_{x=0}-\theta\left(t\right)\right\}^{m}=0, \tag{4}$$

$$\frac{\partial u\left(x,t\right)}{\partial x}\bigg|_{x=l}=0,\ \ 0\leq t\leq T, \tag{5}$$

where $m=\frac{10}{3}$, $\beta\equiv const>0$, $\theta\left(t\right)$ is are given function, and with corresponding consistency constrains.

It was shown by authors [2–4] that (1)–(5) can be reduced to the problem of finding the unknown function $h\left(\varphi\right)$ from the following nonlinear integral equation:

$$h\left(\varphi\right)=g\left(\varphi\right)\cdot\left\{1-\int\limits_{0}^{\varphi}G\left(\varphi,\psi\right)\cdot h\left(\psi\right)d\psi\right\}^{m}, \tag{6}$$

where the functions $h\left(\varphi\right)$, $g\left(\varphi\right)$, $G\left(\varphi,\psi\right)$ and the variables $\varphi=\frac{\beta^{m}a^{2}}{k\tau_{r}}t$, $\psi=\frac{\beta^{m}a^{2}}{k\tau_{r}}\tau$ have the following meaning if we consider the corresponding problem for the hyperbolic heat equation (1):

$$h\left(\varphi\right)\overset{def}{\equiv}k\cdot\beta^{-m}\cdot\vartheta_{2}\left(\frac{k\cdot\tau_{r}}{\beta^{m}\cdot a^{2}}\cdot\varphi\right),\ g\left(\varphi\right)\overset{def}{\equiv}\overline{g}\left(\frac{k\cdot\tau_{r}}{\beta^{m}\cdot a^{2}}\cdot\varphi\right), \tag{7}$$

$$G\left(\varphi,\psi\right)\overset{def}{\equiv}\overline{G}\left(0,0,\frac{k\cdot\tau_{r}}{\beta^{m}\cdot a^{2}}\cdot\left(\varphi-\psi\right)\right),\overline{g}\left(t\right)\overset{def}{\equiv}e^{-\frac{m-1}{2\tau_{r}}t}V^{m}\left(t\right), \tag{8}$$

$$\vartheta_{2}\left(t\right)\overset{def}{\equiv}\frac{\partial\vartheta\left(x,t\right)}{\partial x}\bigg|_{x=0},\vartheta\left(x,t\right)\overset{def}{\equiv}e^{\frac{t}{2\tau_{r}}}\left\{u\left(x,t\right)-\theta\left(t\right)\right\}, \tag{9}$$

$$V\left(t\right)\overset{def}{\equiv}\int\limits_{0}^{l}\frac{\partial G\left(x,\xi,t\right)}{\partial t}\bigg|_{x=0}\vartheta_{0}\left(\xi\right)d\xi+\int\limits_{0}^{l}G\left(x,\xi,t\right)|_{x=0}\vartheta_{1}\left(\xi\right)d\xi+$$

$$\int\limits_{0}^{t}d\tau\int\limits_{0}^{l}G\left(x,\xi,t-\tau\right)|_{x=0}F\left(\xi,\tau\right)d\xi\neq0, \tag{10}$$

$$F\left(x,t\right)\overset{def}{\equiv}e^{\frac{t}{2\tau_{r}}}\left\{\frac{1}{\tau_{r}}f\left(x,t\right)-\frac{1}{\tau_{r}}\theta'\left(t\right)-\theta''\left(t\right)\right\}, \tag{11}$$

$$\overline{G}\left(x,\xi,t\right)|_{\xi=0,x=0}\overset{def}{\equiv}\frac{4\tau_{r}}{V\left(t\right)}\left(\sum_{n=1}^{N}\frac{\sinh\left(t\frac{\sqrt{\left|\left(2a\pi n\sqrt{\tau_{r}}\right)^{2}-l^{2}\right|}}{2l\tau_{r}}\right)}{\sqrt{\left|\left(2a\pi n\sqrt{\tau_{r}}\right)^{2}-l^{2}\right|}}+\right.$$

$$\left.\sum_{n=N+1}^{\infty}\frac{\sin\left(t\frac{\sqrt{\left(2a\pi n\sqrt{\tau_{r}}\right)^{2}-l^{2}}}{2l\tau_{r}}\right)}{\sqrt{\left(2a\pi n\sqrt{\tau_{r}}\right)^{2}-l^{2}}}+\frac{2\tau_{r}}{l}\sinh\left(\frac{t}{2\tau_{r}}\right)\right). \tag{12}$$

Without loss of generality in the expression for the Green function $G(x,\xi,t)$ we have assumed the existence of such natural number N, that for all $n=\overline{1,N}$ the following inequality is valid: $\tau_r \le \left(\frac{l}{2a\pi n}\right)^2$, but for the natural numbers $n=N+1,N+2,...$ the following inequality is valid: $\tau_r > \left(\frac{l}{2a\pi n}\right)^2$.

Having designated the solution of the nonlinear integral equation (6) as $h_{exact}^{hyp.heat}(\varphi)$, the solution of the original problem (1)–(5) is the function:

$$u(x,t)=\theta(t)+e^{-\frac{t}{2\tau_r}}\left\{\int_0^l \frac{\partial G(x,\xi,t)}{\partial t}\left[u_0(\xi)-\theta(0)\right]d\xi+\right.$$

$$\int_0^l G(x,\xi,t)\left[\frac{u_0(\xi)}{2\tau_r}+u_1(\xi)-\frac{\theta(0)}{2\tau_r}-\theta'(0)\right]d\xi++$$

$$\int_0^t e^{\frac{\tau}{2\tau_r}}d\tau\int_0^l G(x,\xi,t-\tau)\left[\frac{f(\xi,\tau)}{\tau_r}-\frac{\theta'(\tau)}{\tau_r}-\theta''(\tau)\right]d\xi-$$

$$\left.\frac{a^2\beta^m}{\tau_r k}\int_0^t G(x,\xi,t-\tau)|_{\xi=0}\,h_{exact}^{hyp.heat}\left(\frac{a^2\beta^m}{\tau_r k}\tau\right)d\tau\right\}. \tag{13}$$

Thus, the last formula determines a desired solution of the general problem (1)–(5) if only we will solve the integral equation (6). For solving the nonlinear Volterra integral equation of the second kind (6) we will introduce the following definition:

Definition 1. *The functional $H(\varphi,h)$ $(\varphi>0)$ will be said to be Volterra functional if the value of H is the number that depends on parameter values and depends on values of function $h(\psi)$ in the half-open interval $0\le\psi<\varphi$.*

Now let us consider the functional equation:

$$h(\varphi)=H(\varphi,h). \tag{14}$$

It is obvious that our integral equation (6) belongs to such type functional equation, where

$$H(\varphi,h)\overset{def}{\equiv} g(\varphi)\left\{1-\int_0^{\varphi} G(\varphi,\psi)\,h(\psi)\,d\psi\right\}^m. \tag{15}$$

Let us assume that the above defined functional $H(\varphi,h)$ satisfies the following conditions:

A. If $h(\psi)\in C[0,\varphi)$ then the functional $H(\varphi,h)$ is well defined and $H(\varphi,h)$ is continuous function on variable φ;

B. If $h_1(\psi)$ and $h_2(\psi)$ are continuous functions on variable ψ and if $|h_1(\psi)| < M$, $|h_2(\psi)| < M$ for $0 \le \psi \le \varphi_0$ then the following inequality is valid:

$$|H(\varphi, h_1) - H(\varphi, h_2)| \le \int_0^{\varphi} K(\varphi, \psi; M, \varphi_0) |h_1(\psi) - h_2(\psi)| \, d\psi, \qquad (16)$$

for $\forall \varphi \in [0, \varphi_0]$, where $K(\varphi, \psi; M, \varphi_0) \overset{def}{=} \frac{K_1(M, \varphi_0)}{K_2(\varphi, \psi)}$, $K_1(M, \varphi_0)$ is some constant depending of M and φ_0, $K_2(\varphi, \psi) \overset{def}{=} (\phi - \psi)^{\alpha}, \alpha \in [0, 1)$.

Remark 1. By immediate verification we can make sure that these two propositions are valid for our functional $H(\varphi, h)$ defined by (6).

Theorem 1. *Let some functional (not merely our functional (15)) satisfies the above-enumerated conditions A and B. Then the functional equation (14) has unique solution in some segment* $0 \le \varphi \le \varphi_0$.

We will prove this theorem by a step-by-step method. Therefore let us consider the functional transformation $S(\varphi) = H(\varphi, h)$, which transforms the function $h(\varphi)$ to $S(\varphi)$.

Our main task is the following: we must determine such φ_0 so that the continuous function $h(\varphi)$ with the property $|h(\psi)| < M$, $0 \le \psi \le \varphi_0$ has been transformed to a function possessing the same property. We can reach this in the following way: we assume $\overline{H}(\varphi) \overset{def}{=} H(\varphi, h)|_{h=0}$. Evidently $\overline{H}(\varphi)$ is a continuous function (by the property A). We assume $L \overset{def}{=} \max\limits_{0 \le \varphi \le \varphi_0} |\overline{H}(\varphi)|$. Then for the function $h(\varphi)$, where $|h(\varphi)| < M$ and $M > L$, we have the following inequality: $|H(\varphi, h) - H(\varphi, h)|_{h=0}| \le M \int\limits_0^{\varphi} K(\varphi, \psi; M, \varphi_0) \, d\psi = MK_1(M, \varphi_0) \frac{\varphi^{1-\alpha}}{1-\alpha}$. From here $|S(\varphi)| = |H(\varphi, h)| \le L + MK_1(M, \varphi_0) \frac{\varphi^{1-\alpha}}{1-\alpha}$. It is clear that for $\forall M > L$ it is possible to find such φ_1 $(\varphi_1 < \varphi_0)$ that $L + M \cdot K_1(M, \varphi_0) \frac{\varphi_1^{1-\alpha}}{1-\alpha} = M$. From here we obtain the final value of unknown number φ_1 as:

$$\varphi_1 = \left(\frac{(M - L) \cdot (1 - \alpha)}{M \cdot K_1(M, \varphi_0)} \right)^{\frac{1}{1-\alpha}}. \qquad (17)$$

Thus, we establish the following fact: if $|h(\varphi)| < M$ for $0 \le \varphi \le \varphi_1$ then $|S(\varphi)| < M$ for $0 \le \varphi \le \varphi_1$.

Now let us take any function $h_0(\varphi)$ that $|h_0(\varphi)| < M$ in the segment $0 \le \varphi \le \varphi_1$. For the sake of definiteness we take, for example, $h_0(\varphi) = 0$. Then we can construct the successive approximations $h_1(\varphi), h_2(\varphi), \ldots, h_n(\varphi), \ldots$ by principle:

$$h_n(\varphi) = H(\varphi, h_{n-1}). \qquad (18)$$

Then owing to (17) it is clear that $|h_n(\varphi)| < M$ for $\forall n \in N$. From here we can write:

$$|h_n(\varphi) - h_{n-1}(\varphi)| = |H(\varphi, h_{n-1}) - H(\varphi, h_{n-2})| \leq$$
$$K_1(M, \varphi_0) \cdot \int_0^{\varphi} \frac{h_{n-1}(\psi) - h_{n-2}(\psi)}{K_2(\varphi, \psi)} d\psi. \tag{19}$$

Here it is significant that in the inequality (19) the constant $K_1(M, \varphi_0)$ is identical for all n $\in N$.

Thus, we receive the following estimations:

$$h_0(\varphi) = 0,$$
$$|h_1(\varphi) - h_0(\varphi)| = |H(\varphi, h)|_{h=0}| \leq L,$$
$$|h_2(\varphi) - h_1(\varphi)| \leq LK_1\varphi^{1-\alpha} \int_0^1 \frac{d\psi}{(1-\psi)^{\alpha}} = LK_1 \frac{\Gamma(1-\alpha)}{\Gamma(2-\alpha)}\varphi^{1-\alpha}, ...,$$
$$|h_n(\varphi) - h_{n-1}(\varphi)| \leq LK_1^{n-1} \frac{\Gamma^{n-1}(1-\alpha)}{\Gamma(1+(n-1)(1-\alpha))}\varphi^{(n-1)(1-\alpha)}. \tag{20}$$

Using the so-called Stirling's formula

$$\Gamma(\beta) = \sqrt{\frac{2\pi}{\beta}} \left(\frac{\beta}{e}\right)^{\beta} \left(1 + O\left(\frac{1}{\beta}\right)\right) > \sqrt{\frac{2\pi}{\beta}} \left(\frac{\beta}{e}\right)^{\beta},$$

from (20) we will obtain:

$$|h_n(\varphi) - h_{n-1}(\varphi)| \leq L\{K_1(M, \varphi_0)\Gamma(1-\alpha)\}^{n-1}$$
$$\left(\frac{e}{1+(n-1)(1-\alpha)}\right)^{(n-1)(1-\alpha)} \sqrt{\frac{1+(n-1)(1-\alpha)}{2\pi}} \varphi^{(n-1)\cdot(1-\alpha)}. \tag{21}$$

From (21) follows that the successive approximations $h_1(\varphi)$, $h_2(\varphi)$, ..., $h_n(\varphi)$ are uniformly convergent to some function $h(\varphi)$ in the open-interval $(0, \varphi_1)$, at which $|h(\varphi)| < M$.

On the other hand we have $|H(\varphi, h) - H(\varphi, h_n)| < \int_0^{\varphi} \frac{|h(\psi) - h_n(\psi)|}{K_2^n(\varphi, \psi)} d\psi$.

Consequently, $H(\varphi, h_n) \overset{n\to\infty}{\rightrightarrows} H(\varphi, h)$. From here we obtain the final result (14): $h(\varphi) = H(\varphi, h)$, $0 \leq \varphi \leq \varphi_1$. The theorem is proved.

Thus, we offer the following recurrence formula for solving of nonlinear integral equation (6):

$$h_0(\varphi) = 0, \; h_n(\varphi) = g(\varphi) \left\{1 - \int_0^{\varphi} G(\varphi, \psi) h_{n-1}(\psi) d\psi\right\}^m, \quad n = 1, 2, 3, ...$$

Moreover, having designated the exact solution of (6) as the function $h_{exact}(\varphi)$, then we have:

$$h_n(\varphi) \overset{n\to\infty}{\Rightarrow} g(\varphi)\left\{1-\int_0^{\varphi} G(\varphi,\psi)\cdot h_{exact}(\varphi)\,d\psi\right\}^m,$$

i.e. the functional sequence $h_n(\varphi)$ converges uniformly to the desired solution $h_{exact}(\varphi)$: $h_n(\varphi) \overset{n\to\infty}{\rightrightarrows} h_{exact}(\varphi)$.

Remark 2. In the proof of theorem we have used the constraint $0 \le \varphi \le \varphi_1$ only for proofing the fact $|h_n(\varphi)| < M\ \forall\varphi$. If the successive approximations $h_1(\varphi), h_2(\varphi), \ldots, h_n(\varphi), \ldots$ are uniformly bounded in some segment $[0,\varphi_2]$ or in the semi-infinite interval $[0,\infty)$ then these successive approximations $h_1(\varphi), h_2(\varphi), \ldots h_n(\varphi), \ldots$ will converge in all area and give us the solution of the functional equation (14).

By using the non-degenerate transformations

$$\tau = \bar{t},\ \xi = \frac{\bar{x}^{c_1}}{\bar{t}},\ u(\bar{x},\bar{t}) = \tau^{c_2}\cdot U(\xi,\tau),$$

where c_1 and c_2 are some constants, it can be shown that (1) has zero approximation

$$u_{zero}(x,t) = \frac{|a|\sqrt{\tau_r}}{x-|a|\sqrt{\frac{1}{\tau_r}}t} - \frac{|a|\sqrt{\tau_r}}{x+|a|\sqrt{\frac{1}{\tau_r}}t},$$

which describes a wave with a growing amplitude. This zero approximation is a solution of the wave equation

$$\frac{\partial^2 u_{zero}(x,t)}{\partial t^2} = \frac{a^2}{\tau_r}\cdot\frac{\partial^2 u_{zero}(x,t)}{\partial x^2},$$

i.e. the term $\frac{\partial u}{\partial t}$ in hyperbolic heat equation characterizes damping of the temperature waves.

In the future works it will be investigated some order approximate solutions for determining solutions character at large value of time.

References

1. French, H.J.: The Quenching of Steels. American Society for Steel Treating, Cleveland, OH (1930)
2. Guseynov, Sh.E.: Methods of the solution of some linear and nonlinear mathematical physics inverse problems. Doctoral Thesis, University of Latvia (2006)
3. Guseynov, Sh.E., Buikis, A.A.: WSEAS Trans. Math. **6**, 43–47 (2007)
4. Guseynov, Sh.E., Kobasko, N.I.: Proc. OTTOM-7 **2**, 22–27 (2006)
5. Kobasko, N.I.: Steel Quenching in Liquid Media under Pressure. Naukova Dumka, Kiev (1980)

Designing a Cover for a Tank

G. Gutiérrez[1], S. Merino[1], J. Martínez[1] and I. Ladrón de Guevara[2]

[1] Universidad de Málaga, Departamento de Matemática Aplicada, Spain
ggutierrez@uma.es, smerino@uma.es, jmartinez@uma.es
[2] Universidad de Málaga, Departamento de Expresión Gráfica, Diseño y Proyectos, Spain ilguevara@uma.es

Summary. Our research group is working on the design of a rigid cover over a tank or a cavity. Usually, swimming pools, wells or drainage channels are protected with high resistance plastic sailcloth and fences. Such security measurements block the access to dangerous areas. We have designed a roof allowing the use of the covered zones.

The system consists of a set of pieces made of high density polyethylene than can be joined as in a puzzle whose junctions are reinforced with bolts. The pieces have been designed using MicroStation and SolidWorks. The system is easily assembled and disassembled, in fact, it is not necessary to have skilled labour to do this.

1 Introduction

Usually (see [2], [5], [10]) we can consider the following steps in designing mechanical pieces:

1. Fixing the function.
2. To analyze the stress that these pieces are going to support.
3. Choosing the material.
4. Determining the shape and size of the pieces.
5. To analyze if the material and the shape are adequate to the required conditions and, if necessary, to introduce modifications to the design.

Moreover, we have to consider the marketing and the maintenance.

Nowadays, the software tools have made things easier for designers (see [6]). CAD packages have a lot of automated features available and they are really useful for steps 4 and 5 in the previous list. The main advantage is that we can model in such a way that the model easily adapts to future changes.

2 A Cover System for a Tank

In order to avoid accidents it is usual the protection of tanks and cavities with sailcloth and fences. In this way, the protected zones become isolated areas.

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*,
Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_138,

Our objective is the design of a solid roof allowing the passage of people over it. So there is no risk of accident and it is possible to make use of these zones.

One of the most important requirements is that the system must be easily assembled and disassembled. Bearing that in mind we have designed a system made up of a set of pieces that can be joined as in a puzzle.

Once we have fix our objective of designing a rigid cover for a tank, we present in Sect. 3 the analysis of the requirements (the stress that the pieces support), we show the pieces we have designed and how do the fit together in Sect. 4 and we analyze the warp of the pieces even in extreme conditions (Sect. 5)

3 Analysis of the Requirements

We have considered the worst load conditions of the *Spanish Building Technical Code* ([3]), using the appropriate security factors (sf). For this reason we will use the following load combinations:

$$\sum_{j\geq 1} \gamma_{G,j} \cdot G_{k,j} + \gamma_p \cdot P + A_d + \gamma_{Q,1} \cdot \psi_{1,1} \cdot Q_{k,1} + \sum_{i>1} \gamma_{Q,i} \cdot \psi_{2,i} \cdot Q_{k,i}$$

where

- G_k denotes permanent loads (In our case it is the total weight).
- P denotes permanent prestressed (it is not necessary because we use homogeneous material).
- A denotes occasional loads (In our case are not necessary).
- Q_k denotes variable loads (In our case is the used load).
- Finally, γ represents security factors and ψ represents simultaneity or synchronized coefficients.

Using the *Spanish Building Technical Code* over this equation we obtain the following values:

- Permanent weight:

$$10 \text{ kg/m}^2 \times 1.35 \text{ (sf for weight)}$$

- Uniform load in public zones:

$$\begin{aligned} 500 \text{ kg/m}^2 &\times 1.5 \text{ (sf most unfavourable)} \\ &\times 1.4 \text{ (sf for syncronized loads)} \end{aligned}$$

- Snow load:

$$\begin{aligned} 100 \text{ kg/m}^2 &\times 1.2 \text{ (sf for climatic zone)} \\ &\times 1.5 \text{ (sf most unfavourable)} \\ &\times 0.7 \text{ (sf for syncronized loads)} \end{aligned}$$

Using all these formulae, we obtain that the total load is 1189.5 kg/m^2 and so we have used 1,200 kg/m^2 as applied load.

4 Designing and Assembling the Pieces

In this section we show the pieces that we have designed using an adequate software. The pieces are designed to fit as in a puzzle. We also justify that high density polyethylene is the most adequate material for our system.

4.1 The Modeling Tools

The software tools have allowed us to build and test models through automated techniques. Our pieces have been designed using both MicroStation and SolidWorks. More specifically, we have used MicroStation for the design phase and SolidWorks to analyze the deformation that the pieces suffer when they bear a weight (see [7], [9]).

We have chosen these CAD platforms because they are extremely powerful and interoperable modeling tools. Secondly, the interface with the user is very flexible and intuitive.

In fact, the software allows us to create sketches and edit them. This is very useful when experimenting with changes to the sketch because we can see both the new and the old states of the sketch and decide on possible modifications to the design. In this way, we can also create different versions of a piece.

Moreover, when using the view tools we can visualize the prototypes in a very realistic way. The software offers many options to manipulate the view, for instance rotating the point of view or changing the projection plane.

We can also apply the material properties to the pieces. The materials define certain properties (density and resistance as example) that are associated to Finite Element Analysis (FEA) (see [1], [8], [11], [12], [13]). According to the material, we can see a textured display.

4.2 The Pieces and the Assembly

The pieces are 2 mm thick. The size of each one in approximately 1 m^2, its height is 0.3 m and the weight is less than 10 kg. The size and weight of the pieces make transport and storage easy.

Each piece is inserted into surrounding ones. The shape of the pieces allows a perfect fit between them. Moreover, the junctions are reinforced with bolts. In this way, we have a rigid roof allowing that the weight applied over a point in the surface can be transmitted to the edge of the tank (Figs. 1 and 2).

We have also designed special pieces for the borders of the tank. These pieces are fixed with bolts and they allow the roof to fit to the size of the tank.

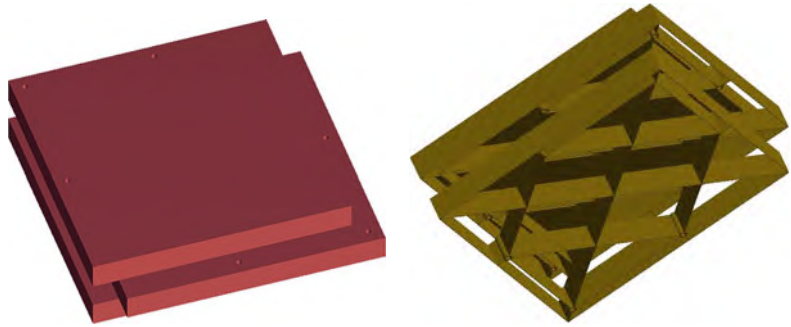


Fig. 1. Outside and inside perspectives

Fig. 2. Detail of inside perspective

In the case of big gaps we have considered the inclusion of fixing bars.

If the tank is full of liquid (for example, a swimming pool) the pieces float. In this case, by Archimedes's Principle, the resistance of the system increases. In addition, each piece is itself a tank that can be empty (if we want to increase the flotation) or may contain some liquid (if we want to increase the rigidity). The mass of liquid into the pieces is limited by the fact that it can become frozen if the environment temperature is cool enough (see [4]).

4.3 Choosing the Material

The pieces are made of high density polyethylene (HDPE). We have chosen HDPE because its properties are really adequate to our project: low density and very high mechanical and chemical resistance. In fact, the resistance to the more usual chemical products in a swimming pool is excellent. Moreover, HDPE is a suitable material for outside equipment (it can resist a big gap of temperatures, there is not a problem with the environmental humidity and it has an acceptable resistance to solar radiation). Of course, we have also taken into account the economical questions (HDPE is a very cheap plastic, there are not problems with the supplying, it is easy to shape in a factory) and the fact that it is non toxic, recyclable and easy to clean.

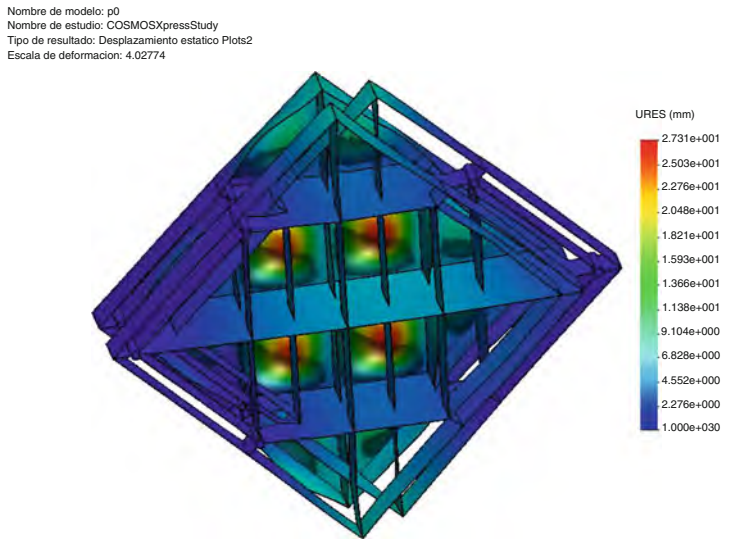


Fig. 3. Displacement results

5 Analysis of the Strain

In the picture we can see the effects of the applied loads over the pieces. It is obvious than even with a load of 1,200 kg/m^2, the deformation is imperceptible (Fig. 3).

6 Design Optimization

At the moment we are working in a variable design that let us to obtain the optimal shape. With this philosophy we impose the following conditions:

1. The thickness is variable from 1 to 4 mm.
2. The height is variable from 15 to 45 cm.
3. Deformations are limited to 5 mm.

and we will choose the minimum weight solution.

References

1. Akin, J.E.: Application and implementation of finite element methods. Computational Mathematics and Applications. Academic [Harcourt Brace Jovanovich Publishers], London (1982)
2. Argüelles, R.: Cálculo de Estructuras. Escuela Técnica Superior de Ingenieros de Montes, Madrid (1981)

3. BOE: Código Técnico de la Edificación. Obra Completa. Madrid (2006)
4. Bonilla, L.L., Moscoso, M., Platero, G., Vega, J.M. (eds.): Progress in industrial mathematics at ECMI 2006, volume 12 of Mathematics in Industry. Springer, Berlin (2008). Papers from the 13th European Conference on Mathematics in Industry held in Madrid, June 2006, The European Consortium for Mathematics in Industry (Berlin)
5. Félez, J., Martínez, M.L., Cabanellas, J.M., Carretero, A.: Fundamentos de Ingeniera Gráfica. Síntesis, Madrid (1996)
6. Ford, B., Rault, J.C., Thomasset, F. (eds.): Tools, methods and languages for scientific and engineering computation, vol. 1. Elsevier Science, Amsterdam (1984). Proceedings of the International Conference on Tools, Methods and Languages for Scientific and Engineering Computation held in Paris, France, 17–19 May 1983
7. Lombard, M.: SolidWorks 2007 Bible. Wiley, Indianapolis (2007)
8. Oñate, E.I.: Cálculo de Estructuras por el Método de Elementos Finitos. Análisis estático lineal. Centro Internacional de Métodos Numéricos en Ingeniera, Barcelona (1992)
9. Planchard, D., Planchard, M.: A Commands Guide For Solidworks 2008. Thomson Delmar Learning, Madrid (2008)
10. Prieto, M.: Fundamentos Geométricos del Diseño en Ingenieria. Aula Documental de Investigación, Madrid (1992)
11. Weaver, W. Jr., Johnston, P.R.: Finite elements for structural analysis, 1st edn. Prentice-Hall, Inc., Englewood Cliffs, New Jersey (1984)
12. Zienkiewicz, O.C.: The finite element method in engineering science. McGraw-Hill, London (1971). The second, expanded and revised, edition of ıt The finite element method in structural and continuum mechanics
13. Zienkiewicz, O.C., Taylor, R.L.: The Finite Element Method: The Basis, vol. 1, 5th edn. Butterworth-Heinemann, Oxford (2000)

An Advection-Dispersion Model for Spray Droplet Transport Including Interception by a Shelterbelt

S.A. Harper, R. McKibbin, and G.C. Wake

Institute of Information and Mathematical Sciences, Massey University, Auckland, NEW ZEALAND, s.harper@niwa.co.nz, R.McKibbin@massey.ac.nz, G.C.Wake@massey.ac.nz

Summary. This paper presents an overview of a simple advection-dispersion model for the transport of spray drift from orchard spraying, including droplet interception by a shelterbelt. Solutions to the model provide an estimate of the drift deposition profile on the ground.

1 Introduction

Advection-dispersion models are widely-used in analysis of particle transport, and often analytic solutions are possible. Whilst analytical models usually involve simplifying assumptions, they are always useful as an initial estimate, and also allow for efficient analysis of parameter variations. This paper presents an overview of an advection-dispersion model for the transport of airborne spray drift from orchard spraying, including the interception of spray droplets by a shelterbelt.

Studies have shown that shelterbelts can reduce spray drift by as much as 90 % [6]. While there has been some analysis of the droplet capture efficiency of a shelterbelt [5], there is little information available with which to predict the deposit downwind, particularly for a fully-sheltered orchard block.

The objective here is to develop a simple analytical model to capture the major features of spray drift transport, including droplet capture by a shelterbelt. We apply an advection-dispersion equation, based on the approach of [4] in modelling particle transport in a forest canopy.

2 Model Formulation

Drifting spray droplets are advected by the wind and dispersed by turbulence, all whilst falling under the influence of gravity and losing mass by evaporation. Within a shelterbelt some of the droplets may be intercepted by the foliage, and we refer to this as trapping.

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*, Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_139,

2.1 Droplet Trapping

Conceptually, the rate at which droplets are trapped will depend upon how many are within the shelterbelt. We model the trapping by

$$T = k_b R c, \tag{1}$$

where T is the rate of droplet mass removal by trapping per unit volume ($\mathrm{kg\,s^{-1}\,m^{-3}}$), $R = R(x, y, z)$ is a dimensionless function which is non-zero only within the shelterbelt, and $c = c(x, y, z, t)$ is the droplet mass concentration per unit volume ($\mathrm{kg\,m^{-3}}$). The proportionality constant k_b ($\mathrm{s^{-1}}$) is called the background trapping rate; it is defined as the fraction of droplets removed per unit time, and is related to the physical properties of the shelterbelt (see [3]). We treat k_b as being constant throughout the shelterbelt.

2.2 Advection-Dispersion Model Without Evaporation

It is more straightforward at first to consider the case where there is no evaporation. Our advection-dispersion model, including trapping within the shelterbelt, is then

$$\frac{\partial c}{\partial t} + u\frac{\partial c}{\partial x} - S\frac{\partial c}{\partial z} = D_L\frac{\partial^2 c}{\partial x^2} + D_T\frac{\partial^2 c}{\partial y^2} + D_V\frac{\partial^2 c}{\partial z^2} + q - T, \tag{2}$$

where u is the mean wind speed (the positive x-axis is aligned to point directly downwind), S is the downward droplet settling speed, and D_L, D_T and D_V are the dispersion coefficients alongwind, crosswind and vertically. These parameters are all assumed to be constant. The source is represented by q ($\mathrm{kg\ s^{-1}\ m^{-3}}$), and T is the trapping term as given by (1).

We solve the model for a cohort of droplets which are all of the same size. Assuming that mass Q is released instantaneously at time $t = 0$ from the point (X_0, Y_0, H), the source term q may be written

$$q = Q\delta(x - X_0)\,\delta(y - Y_0)\,\delta(z - H)\,\delta(t). \tag{3}$$

The distribution of droplet sizes produced by a sprayer would be simulated by superposing solutions to the model, each for a different droplet size. Other source types, such as a line release, can also be constructed from the results for the point release [4].

The initial and far-field boundary conditions are

$$c = 0 \text{ at } t = 0^-, \text{ and } c \to 0 \text{ as } x, y \to \pm\infty \text{ and } z \to +\infty. \tag{4}$$

The ground is approximately horizontal – it is assumed to be impervious to the droplets, so that they cannot disperse through it and the boundary condition on the ground is

$$\frac{\partial c}{\partial z} = 0 \text{ on } z = 0. \tag{5}$$

Needless to say, the actual dynamics of the spraying process and the airflow through and around the shelterbelt are very complicated. The model is intended to capture the main features of the droplet transport, yet be simple enough to enable an analytic solution; some points to note are:

- The mean wind speed is assumed to be uniform in speed and direction.
- Turbulence in the airflow is modelled as having some dominant length scales alongwind, crosswind and vertically. These would be typical mean values for the flow, since turbulence has a variety of scales.
- The dispersion coefficients are represented as the dominant turbulence length scales multiplied by the mean wind speed.

3 A Point Representation for Trapping

To solve the model (2) analytically, we introduce a mathematical simplification whereby the effect of continuous trapping in a small block is represented as occurring at a single point. As shown in Fig. 1, the shelterbelt is then discretised by dividing it into an $N \times L \times M$ array of blocks, each of the same size, with the trapping in each block represented as occurring at the point in its centre.

The blocks each have dimensions $\Delta x \times \Delta y \times \Delta z$ and are labelled by alongwind, crosswind and vertical indices $n = 1, \ldots, N$, $l = 1, \ldots, L$ and $m = 1, \ldots, M$ respectively. For each block, we concentrate the trapping to the point at its centre by defining

$$R_{nlm} = \Delta x \Delta y \Delta z \delta\left(x - X_n\right) \delta\left(y - Y_l\right) \delta\left(z - Z_m\right). \tag{6}$$

Next we introduce an effective trapping rate for the point; denoted by k, this is the background trapping rate scaled by the block size:

$$k = k_b \Delta x \Delta y \Delta z. \tag{7}$$

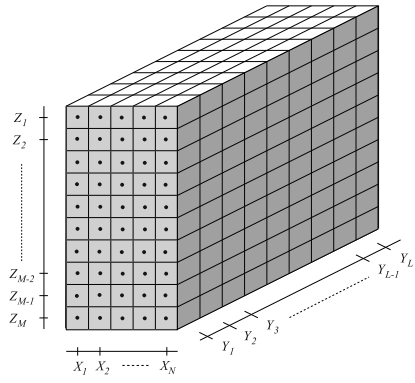


Fig. 1. A rectangular shelterbelt discretised using a 3-D array of trapping points

Note that k is the same for each trapping point. The total rate of droplet mass removal per unit volume for the discretised shelterbelt is the summed effect of all of the points, so that T in (1) and (2) becomes

$$T = \sum_{n=1}^{N} \sum_{l=1}^{L} \sum_{m=1}^{M} kc\left(X_n, Y_l, Z_m\right) \delta\left(x - X_n\right) \delta\left(y - Y_l\right) \delta\left(z - Z_m\right). \tag{8}$$

4 Model Solution

By taking Fourier transforms (in x and y) and applying a Green function, we obtain a solution to the model which is embedded in a convolution equation of the form

$$\begin{aligned} c\left(x, y, z, t\right) &= Qf\left(x, y, z, t; X_0, Y_0, H\right) \\ &\quad - \sum_{n=1}^{N} \sum_{l=1}^{L} \sum_{m=1}^{M} \int_0^t kc\left(X_n, Y_l, Z_m, \tau\right) f\left(x, y, z, t - \tau; X_n, Y_l, Z_m\right)\, d\tau\,. \end{aligned} \tag{9}$$

Due to its length, the expression for the function f is not included here but may be found in [2] and [3]. Of particular interest is the total amount trapped and the deposit on the ground; these quantities may be conveniently evaluated using Laplace transforms, by noting that

$$\int_0^\infty c\left(x, y, z, t\right)\, dt = \left[\int_0^\infty e^{-pt} c\left(x, y, z, t\right)\, dt\right]_{p=0} = \overline{c}\left(x, y, z, 0\right), \tag{10}$$

where $\overline{c}\left(x, y, z, p\right)$ is the Laplace transform of $c\left(x, y, z, t\right)$ with respect to t.

The total mass trapped by the discretised shelterbelt, M_{TT} (kg), is the integral of (8) with respect to space and time. Using (10) this becomes

$$M_{TT} = \sum_{n=1}^{N} \sum_{l=1}^{L} \sum_{m=1}^{M} k\overline{c}\left(X_n, Y_l, Z_m, 0\right). \tag{11}$$

Note that the concentration at each trapping point depends upon the concentration at all of the others. Thus, to evaluate (11) we must solve the system of simultaneous equations formed by the Laplace transform of (9) at each trapping point.

The density of deposit on the ground, M_D (kg m^{-2}), is the integral of the downward mass flux per unit area at $z = 0$ with respect to time:

$$M_D = S\overline{c}\left(x, y, 0, 0\right). \tag{12}$$

An example of the percentage reduction in the density of deposit caused by trapping is shown in Fig. 2. The parameters used are $u = 1\,\mathrm{m\,s^{-1}}$, $S =$

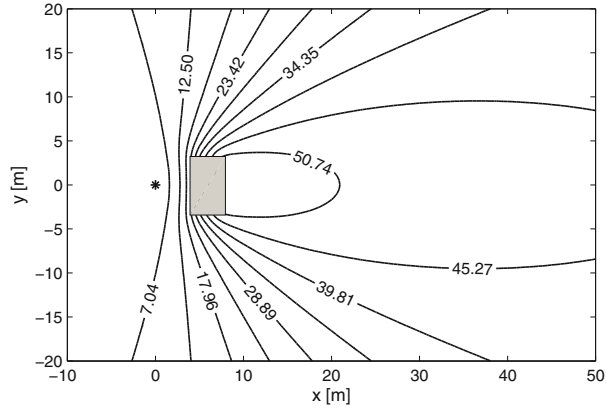


Fig. 2. An example of the percentage reduction in the density of deposit as a result of trapping. See the text for parameter values

$0.2\,\mathrm{m\,s^{-1}}$, $(D_L, D_T, D_V) = (2, 2, 1)\,\mathrm{m^2\,s^{-1}}$, $Q = 1\,\mathrm{kg}$, $(X_0, Y_0, H) = (0, 0, 4)$ and $k_b = 0.5\,\mathrm{s^{-1}}$. The shelterbelt, shown shaded in grey, is 4 m wide × 8 m long × 8 m high; it is divided into $2 \times 4 \times 40$ blocks, so there is a total of 320 trapping points. The strongest reduction is immediately downwind of the shelterbelt, and there is also some effect upwind and around the sides.

5 Advection-Dispersion Model with Evaporation

Droplets in the air evaporate at a rate proportional to the ambient temperature and relative humidity [1]. Here we give a brief discussion of the model with evaporation; for more detail see [3].

The droplets become lighter as they evaporate, so their settling speed decreases; this tends to increase the distances over which the droplets travel. It is more convenient to use the droplet number concentration, since all droplets count as the same in the number concentration no matter their mass; therefore evaporation only appears in the model via the non-constant settling speed. The model is

$$\frac{\partial C}{\partial t} + u\frac{\partial C}{\partial x} - S\,(t)\,\frac{\partial C}{\partial z} = D_L\frac{\partial^2 C}{\partial x^2} + D_T\frac{\partial^2 C}{\partial y^2} + D_V\frac{\partial^2 C}{\partial z^2} + q - T, \qquad (13)$$

where $C = C\,(x, y, z, t)$ is the number concentration ($\#\,\mathrm{m^{-3}}$), q is the same as in (3) and T is the same as in (8) but with c replaced by C. The initial and boundary conditions are the same as in the previous model. The non-constant settling speed combined with the boundary condition on the ground makes it very difficult to solve the model analytically; the solution we obtain is embedded in an integral equation which must be evaluated numerically.

The total mass trapped and the density of the deposit on the ground can be found using the numerically evaluated values for C:

$$M_{TT} = \int_0^{t_s} \sum_{n=1}^{N} \sum_{l=1}^{L} \sum_{m=1}^{M} m(t)\, kC(X_n, Y_l, Z_m, t)\ dt, \tag{14}$$

where $m(t)$ is the mass of an individual droplet and t_s is the time at which the droplets evaporate completely, and

$$M_D = \int_0^{t_s} S(t)\, m(t)\, C(x, y, 0, t)\ dt. \tag{15}$$

Expressions for $m(t)$, $S(t)$ and t_s are given in [3]. Though not ideal, as they require numerical evaluation, these expressions can still be used to observe the effects of parameter variations.

6 Summary

We have used an advection-dispersion model to simulate the transport of airborne drifting spray droplets, including a trapping term to represent droplet interception by a shelterbelt. In order to solve the model analytically we discretised the shelterbelt using a point representation for trapping. Without evaporation, the total amount trapped and the deposit on the ground could be explicitly evaluated from the model by using Laplace transforms; with evaporation, however, these quantities had to be evaluated numerically.

Acknowledgement

Thank you to Lincoln Ventures Ltd, NZ, for their support of this research.

References

1. Davies, C.N.: Evaporation of airborne droplets. In: Shaw, D.T. (ed.) Fundamentals of Aerosol Science, pp. 135–164. Wiley, New York (1978)
2. Harper, S.A.: Gaz. Aust. Math. Soc. **34**, 28–34 (2007)
3. Harper, S.A.: Mathematical models for the dispersal of aerosol droplets in an agricultural setting. PhD Thesis, Massey University, Auckland (2008)
4. McKibbin, R.: JSME Int. J. Ser. B. **49**, 583–589 (2006)
5. Raupach, M.R., Woods, N., Dorr, G., Leys, J.F., Cleugh, H.A.: Atmos. Environ. **35**, 3373–3383 (2001)
6. Ucar, T., Hall, F.R.: Pest Manag. Sci. **57**, 663–675 (2001)

Numerical Modelling of a Pulse Combustion Burner: Limiting Conditions of Stable Operation

P.A. van Heerbeek[1], M.B. van Gijzen[1], C. Vuik[1], and M.R. de la Fonteijne[2]

[1] Delft Institute of Applied Mathematics, Delft University of Technology, Mekelweg 4, 2628 CD Delft, The Netherlands, p.a.vanheerbeek@ziggo.nl, m.b.vangijzen@tudelft.nl, c.vuik@tudelft.nl

[2] DLF Sustainable, Delft, P.O. Box 1077, 2600 BB Delft, The Netherlands, marcel@dlfsustainable.nl, http://www.dlfsustainable.nl/

Summary. Numerical modelling of pulse combustors may give important guidelines on how design parameters should be chosen. This paper gives a mathematical analysis of a simple model for thermal pulse combustion and determines conditions under which this model can describe stable pulse operation.

1 Introduction

Compared to conventional combustion, pulse combustion has significant advantages in terms of thermal efficiency, energy savings and environmental impact. The high heat transfer rate makes it particularly attractive for applications such as heating, particle drying, waste incineration, etc. Areas for which industrial application of pulse combustion can be beneficial include heating, drying, calcinating, gasification, and waste incineration.

The operation of a pulse combustor is based on a coupling between intermittent (pulse) combustion and resonant acoustics in the burner system. Self-sustained pulse combustion and high-intensity sound waves result if the system's acoustics and the combustion process are in phase (i.e. if *Rayleigh's criterion* [2] is satisfied).

The pulse combustion characteristics are determined by complex interactions between physical and chemical processes, which depend on many parameters (e.g. fuel supply, mixing processes, reaction rates, tailpipe length). This severely complicates the design process.

Numerical modelling may give important guidelines on how the design parameters should be chosen in order to achieve an optimal performance of the pulse combustion process. To gain insight into the role of various design parameters, we study a simple model of a so-called thermal pulse combustor. By integrating the model equations in time it is possible to predict whether

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*, Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_140,

stable pulse operation for a given set of design parameter values is possible. Such an analysis, however, is very time consuming if many combinations of design parameter values have to be considered. In order to make a computationally less demanding analysis we perform a stability analysis on the model equations. We will show that the stability analysis provides insightful information by comparing it with the results of a time integration analysis.

2 Thermal Pulse Combustion: A Mathematical Model

Richards et al. [3] introduced a mathematical model that describes pulse combustion in a system with a continuous fuel supply, which they call *thermal pulse combustion*. Figure 1 gives a schematic representation of such a thermal pulse combustor. Thermal pulse combustion is different from ordinary pulse combustion, where fuel periodically enters the combustion chamber because of time-dependent pressure differences over valves. However, Richards et al. show that pulsating combustion can occur even in the case of a continuous fuel supply.

Richards et al. model this device with a simple lumped parameter model, taking the combustion chamber as a control volume. The amount of energy in the combustion chamber is changed by inflow of reactants, combustion, outflow of combustion products, and heat transfer to the chamber wall. The combustion process is modelled by a one-step Arrhenius law for a bimolecular reaction between fuel and oxidizer. The gases in the combustion chamber are assumed to be well-stirred. The tailpipe flow is modelled as a plug flow, i.e. with a uniform density over its volume and driven by the pressure difference over the tailpipe. Flow from the combustion chamber into the tailpipe is assumed to be isentropic. Wall friction of the tailpipe gases is taken into account. It is assumed that the gases are perfect, and that all mixtures of reactants and products have the same (constant) specific heats. By applying conservation of mass, energy and species to the control volume, and coupling this to the tailpipe dynamics by conservation of momentum, Richards et al. derived a system of four ordinary differential equations. It can be written as

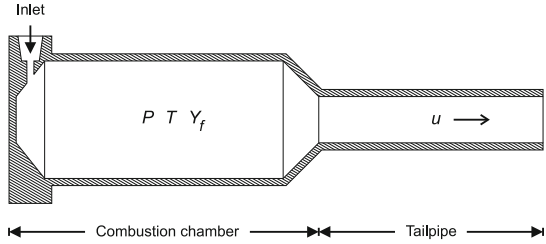


Fig. 1. Control volumes and variables of the thermal pulse combustor model

$$\frac{dP}{dt} = \gamma(A + B \cdot RR + CD - (C + GZ_e)\,T), \tag{1}$$

$$\frac{dT}{dt} = \gamma(A + B \cdot RR + CD)\frac{T}{P} - (A + \gamma C + (\gamma - 1)GZ_e)\frac{T^2}{P}, \tag{2}$$

$$\frac{du}{dt} = E(P_e - 1)\,\frac{T_e}{P_e} - F\,u\,|u|, \tag{3}$$

$$\frac{dY_f}{dt} = (A(Y_{f,i} - Y_f) - RR)\frac{T}{P}. \tag{4}$$

where RR, P_e, T_e and Z_e are functions of P, T, u and Y_f, and $A, \ldots, G$, γ and $Y_{f,i}$ are constants that depend on the system's design parameters and the fluid properties. The (non-dimensionalized) variables are: the pressure (P), temperature (T), and fuel mass fraction (Y_f) in the combustion chamber and the fluid velocity (u) in the tailpipe. The system of equations can be expressed in vector form by

$$\frac{d\mathbf{y}}{dt} = \mathbf{f}(\mathbf{y}), \quad \text{where} \quad \mathbf{y} = (P, T, u, Y_f)^\top.$$

3 Parameter Study

3.1 Numerical Time Integration

Through numerical time integration it is possible to study the stability of the combustion process for given sets of parameters.

Figure 2 shows how the pressure evolves in time for three different values of the wall temperature (T_w) of the combustion chamber. The top figure shows steady combustion for $T_w = 1{,}200$ K. The middle figure shows pulse combustion for $T_w = 1{,}000$ K. The bottom figure shows that flame extinction occurs for $T_w = 750$ K.

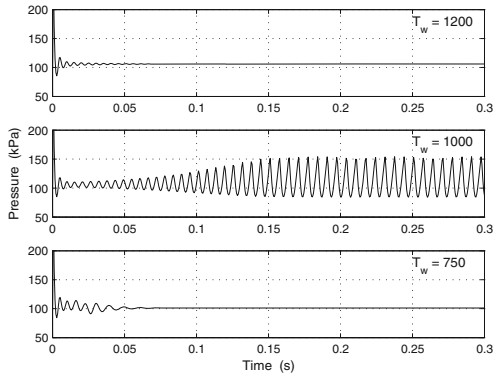


Fig. 2. Pressure signals in combustion chamber from numerical simulations for three wall temperatures

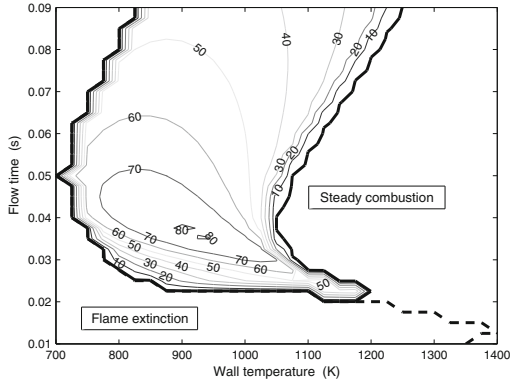


Fig. 3. Peak-to-peak pressure amplitudes (in kPa) obtained from numerical simulation for various combinations of wall temperature and flow time

Figure 3 shows how the peak-to-peak amplitude of the pressure oscillations depends on the wall temperature T_w and the flow time τ_f, which is inversely proportional to the fuel mass inflow rate. The figure indicates for which combinations of these two parameters stable pulse operation occurs.

3.2 Stability Analysis

This two parameter analysis is already very time consuming. In order to explore the parameter space in an insightful and computationally inexpensive way, we perform a stability analysis of the steady-state solutions. These are found by solving $\mathbf{f}(\mathbf{y}) = \mathbf{0}$.

Several steady-state solutions may exist for any given set of parameter values. Figure 4 shows the steady-state temperature(s) for given wall temperatures. Clearly either one, two, or three steady-state solutions exist, depending on the wall temperature.

The stability of the steady states can be determined by calculating the four eigenvalues of the corresponding Jacobian matrices. If all four eigenvalues have a negative real part, the steady state is stable and no pulse combustion can occur in its neighbourhood.

Figure 5 shows the largest of the real parts of the eigenvalues corresponding to the steady-state solution with the highest temperature. It indicates for which pairs of the parameters T_w and τ_f the steady-state solution is stable or unstable. The boundary for flame extinction results from bifurcation of the steady state, while the boundary for steady combustion results from a sign change of the real part of the eigenvalue. Note the qualitative correspondence with Fig. 3.

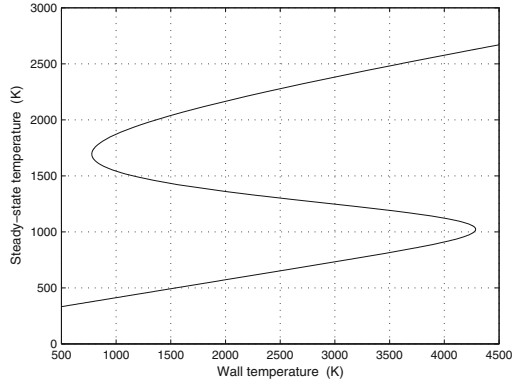


Fig. 4. Steady-state temperatures as function of the wall temperature

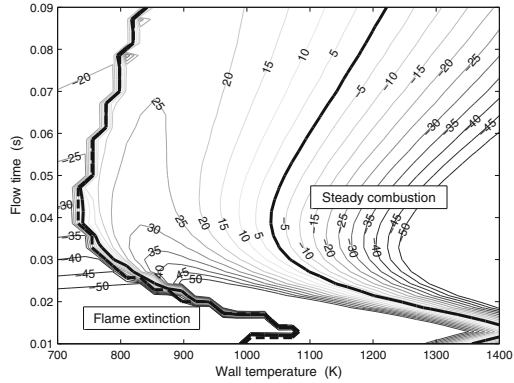


Fig. 5. Largest of the real parts of the four eigenvalues corresponding to the steady-state solution with the highest temperature

4 Concluding Remarks

4.1 Discussion

The stability analysis provides a useful tool for the parameter study. For state variables close enough to a stable steady state, the stability analysis gives us definitive information that pulse combustion is not possible. In general, the stability analysis gives a strong indication for which system parameters pulse combustion is not possible. For other system parameter values the stability analysis is not conclusive. Figure 3, for example, shows regions where flame extinction occurs, while the corresponding regions in Fig. 5 show unstable steady states. However, the stability analysis always gives good guidelines to determine which values of the design parameters should be investigated further for obtaining stable pulse operation.

4.2 Validity

The model in [3] describes *thermal* pulse combustion. The model can be adapted for valved pulse combustion, see for example [1] and [4], but then only one steady state exists: flame extinction. The model of Richards et al. also has serious limitations, see our analysis in [6]. Despite its limitations, experiments with different heat release rates in the model, and extending it with variable air/fuel ratio and stochastic noise, suggest that good agreement with experimental data for valved combustion can be obtained.

4.3 Future Research

The stability analysis has provided useful insight into the behaviour of a pulse combustor as modelled by Richards et al. As a next step in our research, we want to extend this stability analysis to more advanced models, and we also hope to gain more insight into its relation with Rayleigh's criterion. Furthermore, the model of Richards et al. can be improved by including additional physics (non-stoichiometric and pressurized combustion, combustion noise). Finally, we want to use the resulting model to study chaotic behaviour in pulse combustion operation, as for example in [5].

Acknowledgement

The authors thank H. Corstens for his stimulating comments on many aspects of this work. This joint research project of DLF Sustainable and the Delft University of Technology is sponsored by SenterNovem.

References

1. Narayanaswami, L., Richards, G.A.: J. Eng. Gas Turbines and Power, **118**(3), 461–468 (1996)
2. Rayleigh: Nature **18**(455), 319–321 (1878)
3. Richards, G.A., Morris, G.J., Shaw, D.W., Keeley, S.A., Welter, M.J.: Combust. Sci. Tech. **94**, 57–85 (1993)
4. Richards, G.A., Gemmen, R.S.: J. Eng. Gas Turbines and Power **118**(3), 469–473 (1996)
5. Daw, C.S., Thomas, J.F., Richards, G.A., Narayanaswami, L.L.: Chaos **5**(4), 662–670 (1995)
6. van Heerbeek, P.A.: Master's Thesis, Delft University of Technology (2008) (http://ta.twi.tudelft.nl/nw/users/vuik/numanal/heerbeek_eng.html)

Optimal Control of Buoyant Flows with Temperature-Dependent Viscosity

H. Herrero and F. Pla

Departamento de Matemáticas, Universidad de Castilla-La Mancha, 13071 Ciudad Real, Spain, Francisco.pla@uclm.es, Henar.Herrero@uclm.es

Summary. This paper shows the effects of a boundary control on pattern formation in a Rayleigh–Bénard problem with temperature-dependent viscosity. In particular, a rectangular domain infinite in one of the horizontal dimensions is considered. The conductive state bifurcates to a stationary pattern for the constant viscosity case. And the boundary control hinders instability up to the point where it is inhibited for the value of the control at which the gradient disappears. For the variable viscosity case, the conductive state bifurcates to a different stationary pattern, and the critical threshold is lower. The boundary control changes the critical wave number and favors instability up to the point where it is inhibited for the value of the control at which the gradient disappears.

1 Introduction

Thermoconvective flows often appear in nature. For instance, thermoconvective instabilities are responsible for the development of many geophysical phenomena like mantle convection, plate tectonics, etc. Classically, the problem is stated as a fluid layer heated uniformly from below [2,3]. A conductive state becomes unstable at temperature gradients beyond a certain threshold. Two different effects are responsible for the onset of motion: these are gravity (Rayleigh–Bénard problem) and capillary forces (Marangoni problem). Most studies consider a constant viscosity [5], although interest in convection problems with temperature dependent viscosity has increased [6,9] since this dependence is a fundamental feature of mantle convection. Optimal control techniques are useful for finding ways to avoid convection [1]. Navarro and Herrero [7] proves that an optimal boundary control consists of heating at the top with the same shape as at the bottom.

In this paper we propose the numerical study of a Rayleigh–Bénard problem in a rectangular domain that is infinite in one of the horizontal dimensions. Viscosity depends exponentially on temperature. And we apply a boundary control with constant heating at the top. We then compare the influence of

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*, Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_141,

the control on the instabilities to problems with constant and variable viscosities. The conductive state bifurcates to a stationary pattern for the constant viscosity case. And the boundary control hinders instability up to the point where it is inhibited for the value of the control at which the gradient disappears. For the variable viscosity case, the conductive state bifurcates to a different stationary pattern, and the critical threshold is lower. The boundary control changes the critical wave number and favors instability up to the point where it is inhibited for the value of the control at which the gradient disappears.

2 Formulation of the Problem

The physical setup consists of a horizontal fluid layer in a rectangular container l wide (x coordinate), d deep (z coordinate) and infinite in the y direction. The upper surface is free and the bottom plate is rigid. At $z = 0$ a constant temperature is imposed $T_{\max}$. The upper surface temperature is $T_{\min}$. We define $\triangle T = T_{\max} - T_{\min}$.

In the equations governing the system, u_x, u_y and u_z are the components of the velocity field $\mathbf{u}$, T is the temperature, p is the pressure, $\mathbf{x}$ is the space coordinate, and t is the time. The magnitudes are expressed in dimensionless form after rescaling in the following way: $\mathbf{x}' = \mathbf{x}/d$, $t' = \kappa t/d^2$, $\mathbf{u}' = d\mathbf{u}/\kappa$, $p' = d^2 p/\left(\rho_0 \kappa \nu_0\right)$, $\Theta = \left(T - T_{\min}\right)/\triangle T$. Here κ is the thermal diffusivity, ν_0 the kinematic viscosity of the liquid at temperature $T_{\min}$, and ρ_0 is the mean density at the temperature $T_{\min}$. After rescaling the domain $\Omega_1 = [0, l] \times [0, d] \times I\!R$ is transformed into $\Omega_2 = [0, \Gamma] \times [0, 1] \times I\!R$ where $\Gamma = l/d$ is the aspect ratio, which we set at $\Gamma = 2.891$ since it is a good representation of the general behavior shown in [8].

The system evolves in accordance with the momentum and the mass balance equations and the energy conservation principle, which in dimensionless form are (the primes in the corresponding fields have been dropped),

$$\nabla \cdot \mathbf{u} = 0, \tag{1}$$

$$\partial_t \Theta + \mathbf{u} \cdot \nabla \Theta = \nabla^2 \Theta, \tag{2}$$

$$\partial_t \mathbf{u} + (\mathbf{u} \cdot \nabla)\, \mathbf{u} = \mathrm{Pr}\left(-\nabla p + div\left(\nu(\Theta) \cdot \left(\nabla \mathbf{u} + (\nabla \mathbf{u})^T\right)\right) + \mathrm{R}\Theta \mathbf{e}_z\right), \tag{3}$$

where the Oberbeck–Boussinesq approximation has been used [8]. Here $\mathbf{e}_z$ is the unit vector in the z direction. The following dimensionless numbers have been introduced: the Prandtl number $\mathrm{Pr} = \nu/\kappa$, which accounts for the characteristics of the fluid and can be considered infinite as an approximation to the mantle property, and the Rayleigh number $R = g\alpha \triangle T d^3/\kappa\nu_0$, which represents the buoyancy effect. In these expressions α is the thermal expansion coefficient and g is the gravity constant. The viscosity, $\nu(\Theta)$, is assumed to be exponentially dependent on temperature,

$$\nu(\Theta) = \exp(-\gamma R \Theta), \tag{4}$$

where γ is the exponential rate. Only the values $\gamma = 0$ (constant viscosity case) and $\gamma = 0.0862$ (variable viscosity case) will be considered. This may be considered a large viscosity variation across the fluid layer [8].

We now turn our attention to the boundary conditions (bc). The boundaries are rigid at the bottom plate and free at the upper and lateral surfaces, so

$$u_x = u_y = u_z = 0 \ \text{ on } \ z = 0, \tag{5}$$

$$\partial_z u_x = \partial_z u_y = u_z = 0 \ \text{ on } \ z = 1, \tag{6}$$

$$u_x = u_y = \partial_x u_z = 0 \ \text{ on } \ x = 0 \text{ and } x = \Gamma. \tag{7}$$

For temperature we consider a constant value at the upper surface, C, that will be used as boundary control; the lateral walls are insulating and at the bottom a constant temperature is imposed,

$$\Theta = C \ \text{ on } \ z = 1; \ \ \partial_x \Theta = 0 \ \ \text{ on } \ \ x = 0 \ \text{ and } \ x = \Gamma; \ \ \Theta = 1 \ \ \text{ on } \ z = 0. \tag{8}$$

Additional boundary conditions due to pressure are included as explained in [4].

3 Basic States and Linear Stability Analysis

Of the hydrodynamic equations the conductive solution is the simplest. The temperature only depends on the vertical component and the fluid is at rest,

$$\Theta^b(z) = 1 - z, \ p^b(z) = p_0 + Rz - \frac{R}{2} z^2, \ \mathbf{u}^b = \mathbf{0}. \tag{9}$$

Here the superscript b indicates basic state. The stability of the basic states is studied by perturbing them with a vector field depending on the x, y and z coordinates, in a fully 3D analysis: $u_x(x, y, z) = u_x^b(x, z) + \bar{u}_x(x, z) \exp(iky + \sigma t)$, and similarly for the rest of the fields. The bar refers to the perturbation. We considered Fourier mode expansions in the direction y, because in this direction the boundary conditions are infinite. Expressions for the perturbed fields are replaced in the basic equations (1)–(3) and the resulting system is linearized. Boundary conditions for the perturbations $(\bar{u}_x, \bar{u}_y, \bar{u}_z, \overline{\Theta}, \overline{p})$ are found by substituting the perturbed fields in (5)–(8) and the additional boundary conditions due to pressure.

The resulting problem is an eigenvalue one in σ. If $Re(\sigma) < 0$ for any eigenvalue the basic state is stable, but if there exists a value of σ such that $Re(\sigma) > 0$ then the basic state becomes unstable. The condition $Re(\sigma) = 0$ may be satisfied for certain values of the external parameters, (R, γ, C), which define the critical threshold. At the critical threshold, a stationary bifurcation takes place if $Im(\sigma) = 0$, whereas it is a Hopf bifurcation if $Im(\sigma) \neq 0$.

The eigenvalue problem is discretized by expanding any unknown perturbation field $\mathbf{x}$ in a truncated series of orthonormal Chebyshev polynomials,

Table 1. (k_{c1} R_c) for different order expansions in chebyshew polynomials for $r = 0.0862$, $\Gamma = 2.891$ and $c = 0.996$

	$L = 14$	$L = 16$	$L = 18$	$L = 20$	$L = 22$
$N = 12$	(1,93.360)	(2,93.368)	(2,93.369)	(2,93.370)	(2,93.371)
$N = 14$	(1,93.355)	(2,93.367)	(2,93.369)	(2,93.369)	(2,93.370)
$N = 16$	(1,93.359)	(2,93.367)	(2,93.369)	(2,93.369)	(2,93.370)
$N = 18$	(2,93.368)	(2,93.368)	(2,93.369)	(2,93.370)	(2,93.370)

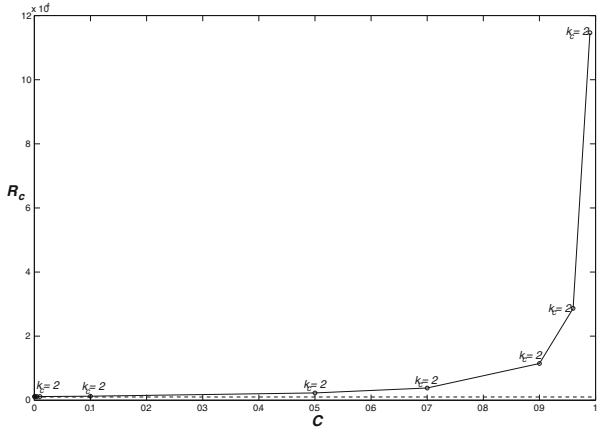


Fig. 1. Critical Rayleigh number depending on the control parameter C for $\gamma = 0$. The critical wave number is indicated above the *circles*

$$\mathbf{x} = \sum_{n=0}^{N-1} \sum_{l=0}^{L-1} a_{nl}^{\mathbf{x}} T_n(x) T_l(z). \tag{10}$$

For computational convenience the domain $\Omega_2 = [0, \Gamma] \times [0, 1] \times I\!R$ is transformed into $\Omega = [-1, 1] \times [-1, 1] \times I\!R$. This change of coordinates introduces scaling factors in equations and boundary conditions which are not explicitly given here. There are $P = 4 \times N \times L$ unknowns, which are determined by a collocation method. In particular, expansions (10) are replaced in the linearized, stationary, axisymmetric equations and boundary conditions (1)–(8), and these are posed at the Gauss–Lobatto collocation points according to the rules explained in [8]. The problem is then transformed into its discrete form and a generalized eigenvalue problem is obtained,

$$Aw = \sigma Bw, \tag{11}$$

where w is a vector that contains P unknowns and A and B are $P \times P$ matrices. The discrete generalized eigenvalue problem (11) has a finite number of eigenvalues σ_i**.** The stability condition explained above must now be imposed on $\sigma_{\max}$, where $\sigma_{\max} = \max Re(\sigma_i)$.

The convergence of the numerical method is tested by comparing the differences in the value of the critical Rayleigh number R_c and the critical

wave number k_c for different orders of expansions in Chebyshev polynomials. These critical wave number and Rayleigh number values for $\gamma = 0.0862$ and $C = 0.996$ are shown in Table 1 for several consecutive expansions on varying the number of polynomials taken in the x (N) and z (L) coordinates. Convergence is reached within a relative degree of precision for R_c in the order of 10^{-4}. Convergence is satisfactory from $N = 14$ and $L = 18$, and these are the orders used in the numerical computations throughout the paper.

4 Numerical Results

We analysed the linear stability of those states for four cases, constant and variable viscosity and without and with control.

4.1 Constant Viscosity Case

The conductive state bifurcates to a stationary pattern with wave number $k_c = 2$ at the critical Rayleigh number threshold $R_c = 1146.50$. By applying the boundary control, i.e. increasing the value of C, the critical wave number does not change, $k_c = 2$, and the critical R_c increases until it becomes infinity for $C = 1$ (Fig. 1). The value $C = 1$ is the same for the temperatures at the bottom and at the top, so there is no temperature gradient; the fluid has constant temperature and hence buoyancy has no effect and the instability disappears.

4.2 Variable Viscosity Case

For the variable viscosity case with $\gamma = 0.0862$ (Fig. 2) the conductive state bifurcates to a stationary pattern with wave number $k_c = 0$, and the critical

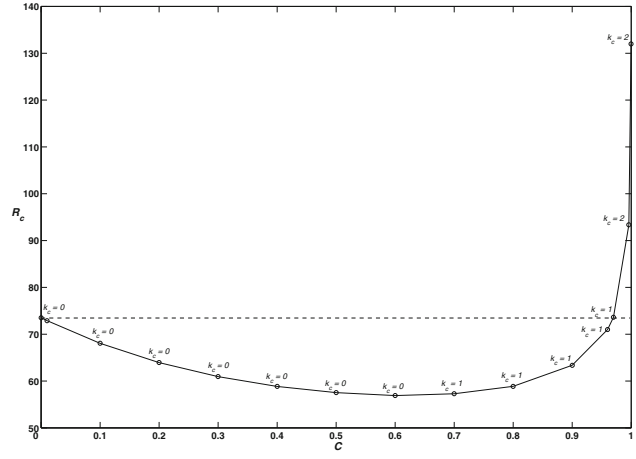


Fig. 2. Critical Rayleigh number depending on the control parameter C for $\gamma = 0.0862$. The critical wave number is indicated above the *circles*

threshold is lower: $R_c = 73.51$. The boundary control has more influence on the instabilities. There are three scenarios. For $0 < C \leq 0.6$ the critical wave number is the same, $k_c = 0$, and the critical Rayleigh number decreases until $R_c = 57$. For $0.6 < C \leq 0.97$ the critical wave number increases, $k_c = 1$, and R_c increases but is less than 73.51 (R_c without control). For $C > 0.97$, $k_c = 2$, and R_c tend to infinity at $C = 1$. This situation is similar to the constant viscosity case. So the control tends to equalise both cases with constant and variable viscosities.

5 Conclusions

In this paper we proposed a numerical study of a Rayleigh–Bénard problem in a rectangular domain infinite in one of the horizontal dimensions, with temperature dependent viscosity and a boundary control. Viscosity depends exponentially on temperature, and the boundary control is constant heating at the top. We compared the influence of the control on the instabilities of problems with constant and variable viscosities. The models with constant and variable viscosities behave differently with respect to the control. In the constant viscosity case the bifurcation does not change with the control: $k_c = 2$ for all C. But in the variable viscosity case it does: k_c changes with C. These results could be applied to planets, so that in a planet with the aspect ratio considered here and with variable viscosity, the mantle convection could be influenced by heating on the lithosphere, but the states could only be changed by strong heating.

Acknowledgments

This work was partially supported by Research Grants MCYT (Spanish Government) MTM2006-14843-C02-01 and CCYT (Junta de Comunidades de Castilla-La Mancha) PAI08-0269-1261, which include RDEF funds.

References

1. Abergel, F., Casas, E.: Rairo M2AN **27**, 223–247 (1993)
2. Bénard, H.: Rev. Gen. Sci. Pures Appl. Bull. Assoc. **11**, 1261–1271 (1900)
3. Bodenschatz, E., Pesch, W., Ahlers, G.: Ann. Rev. Fluid Mech. **32**, 709–778 (2000)
4. Herrero, H., Mancho, A.M.: Int. J. Numer. Meth. Fluids **39**, 391–402 (2002)
5. Koschmieder, E.L.: *Bénard Cells and Taylor Vortices*. Cambridge University Press, Cambridge (1993)
6. Moresi, L.-N., Solomatov, V.S.: Phys. Fluids **7**(9) 2154–2162 (1995)
7. Navarro, M.C., Herrero, H.: Phys. Rev. E 067203-1-4 (2007)
8. Pla, F., Mancho, A.M., Herrero, H.: Physica D **238**, 572–580 (2009)
9. Trompert, R., Hansen, U.: Nature **395**, 686–689 (1998)

Minimum Time Optimal Rendezvous on Circular and Elliptical Orbits

V. Istratie

National Institute for Aerospace Research, Bucharest, Romania,
istratie@incas.ro

Summary. This work studies the minimum rendezvous time of two space vehicles. The target vehicle is on a given circular or elliptical orbit and the surveyor vehicle is equipped with a low thrust motor. Optimal control variables are the magnitude and direction of the acceleration of the surveyor vehicle. This problem is solved by applying the Pontriagin maximum principle. By means of the Second-Order Optimality Conditions, it is demonstrated that the formulated optimization problem is indeed a maximum problem. The calculations are performed both for circular and elliptical orbits around the Earth.

1 Introduction

An interception problem [5] is solved in the planar case, with time constrained, single impulse return trajectories followed by low thrust, power limited return trajectories that minimize the total propellant consumed. This problem is an extension of problems [2,3] in which in these problem was solved the problem minimum fuel optimal rendezvous

2 Problem Formulation: Motion Equations of the State

Considering the $OXZY$ inertial planocentric system of axes and the $Axzy$ system (the z axis in the direction of the vectorial radius $\mathbf{r}_0$ – position vector of target, the x axis is counter rotating, and the y axis perpendicular on them) related to the A target which evolves on an already known circular or elliptical orbit (Fig. 1), starting from Newton's 2nd law (because the motion takes place out of the atmosphere the aerodynamic forces are neglected, and we divided by the mass),

$$\frac{d^2\left(\mathbf{r}_0+\mathbf{r}\right)}{dt^2}=\mathbf{g}+\mathbf{a},$$

where $\mathbf{g}$ is the gravity acceleration in the $Axzy$ axis system, $\mathbf{a}$ – the acceleration due to the thrust and $\mathbf{r}$ – position vector of the surveyor in the $Axzy$

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*,
Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_142,

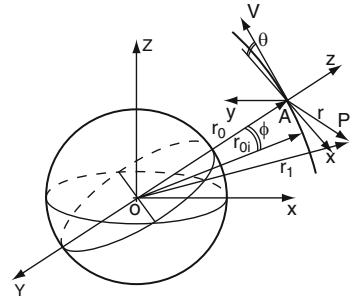


Fig. 1. Motion geometry

(also, in Fig. 1, $\mathbf{r}_1$ – position vector of the surveyor in the $OXZY$ and $\mathbf{r}_{0_i}$ – initial position of $\mathbf{r}_0$) the motion equations of the surveyor P are [2,3]:

$$\frac{dx}{dt} = V_x, \tag{1}$$

$$\frac{dV_x}{dt} = -\frac{2\mu\varepsilon\sin\varphi}{r_0^3}z + \frac{\mu p}{r_0^4}x + 2\frac{\sqrt{\mu p}}{r_0^2}V_z - \mu\frac{x}{\left[x^2+(r_0+z)^2+y^2\right]^{3/2}} + a_x, \tag{2}$$

$$\frac{dz}{dt} = V_z, \tag{3}$$

$$\frac{dV_z}{dt} = \frac{\mu}{r_0^2} + \frac{2\mu\varepsilon\sin\varphi}{r_0^3}x + \frac{\mu p}{r_0^4}z - 2\frac{\sqrt{\mu p}}{r_0^2}V_x - \mu\frac{r_0+z}{\left[x^2+(r_0+z)^2+y^2\right]^{3/2}} + a_z, \tag{4}$$

$$\frac{dy}{dt} = V_y, \tag{5}$$

$$\frac{dV_y}{dt} = -\mu\frac{y}{\left[x^2+(r_0+z)^2+y^2\right]^{3/2}} + a_y, \tag{6}$$

$$\frac{d\varphi}{dt} = \sqrt{\frac{\mu}{p^3}}\left(1+\varepsilon\cos\varphi\right)^2, \tag{7}$$

where μ – gravitational parameter, p – focal parameter, ε – eccentricity orbit ($\varepsilon = 0$ for circular orbit), a_x, a_z, a_y – control variables (the accelerations due to the thrust) and (the state variables) are: x, z, y – coordinates, V_x, V_z, V_y – velocities, φ – elliptical anomaly, with the initial conditions at the time $t = 0$:

$$\begin{gathered} x(0) = x_0, \quad z(0) = z_0, \quad y(0) = y_0, \\ V_{x'}(0) = V_{x_0},\ V_{z'}(0) = V_{z_0},\ V_{y'}(0) = V_{y_0},\ \varphi(0) = 105 \text{ degrees} \end{gathered} \tag{8}$$

and final conditions at $t = t_f$:

$$x(t_f) = z(t_f) = y(t_f) = 0. \tag{9}$$

The values $V_x(t_f)$, $V_z(t_f)$, $V_y(t_f)$, $\varphi(t_f)$ are free. The control variables are α and β – angles acceleration direction, from:

$$a_x = a\cos\beta\cos\alpha, \quad a_z = a\cos\beta\sin\alpha, \quad a_y = a\sin\beta, \tag{10}$$

where a is the module of acceleration and for orbits in around of Earth, $a = 10^{-3}g$, g – the gravity acceleration.

The problem which may be stated is: Let us find a control function

$$\mathbf{u} = (\alpha, \beta) : [0, t_f] \to \mathbb{R}^3$$

and a state function

$$\mathbf{x} = (x, z, y, V_x, V_z, V_y) : [0, t_f] \to \mathbb{R}^6$$

for the circular orbits and

$$\mathbf{x} = (x, z, y, V_x, V_z, V_y, \varphi) : [0, t_f] \to \mathbb{R}^7$$

for elliptical orbits, which minimize functional $J(\mathbf{u}) = \int_0^{t_f} dt$ subject to the differential equations of motion (1)–(7) with the initial conditions (8) and the final conditions (9).

3 Optimizing Problem: Boundary Value Problem

Optimizing the problem by the minimum principle. The above defined problem of optimal control is transformed in a well-known [1, 4, 6] into a two point boundary problem. For this, the Hamiltonian is:

$$H = -1 + p_x f_x + p_{V_x} f_{V_x} + p_z f_z + p_{V_z} f_{V_z} + p_y f_y + p_{V_y} f_{V_y} + p_\varphi f_\varphi,$$

where p_x, p_{V_x}, p_z, p_{V_z}, p_y, p_{V_y}, p_φ are the adjoint variables corresponding of the state variables x, V_x, z, V_z, y, V_y, φ and, also, f_x, f_{V_x}, f_z, f_{V_z}, f_y, f_{V_y}, f_φ are the functions defining the motion equations system of the state variables x, V_x, z, V_z, y, V_y, φ, respectively. By means of the Hamiltonian, the canonic equations that is the differential equations of the state variables (the above mentioned equations), the differential equations of the adjoint variables. The equations of the controls (the optimality conditions) are deduced for (α, β) the interior point. For H from $\mathbf{H}_u = 0$,

$$\alpha = a\tan\frac{\lambda_{V_x}}{\lambda_{V_z}}, \ 0 \le \alpha \le \frac{\pi}{2}, \quad \beta = a\tan\frac{\lambda_{V_y}}{\sqrt{\lambda_{V_x}^2 + \lambda_{V_z}^2}}, \ -\frac{\pi}{2} \le \beta \le \frac{\pi}{2}. \tag{11}$$

Therefore,

$$\begin{aligned} a_x &= a\lambda_{V_x} \Big/ \sqrt{\lambda_{V_x}^2 + \lambda_{V_z}^2 + \lambda_{V_y}^2}, \\ a_z &= a\lambda_{V_z} \Big/ \sqrt{\lambda_{V_x}^2 + \lambda_{V_z}^2 + \lambda_{V_y}^2}, \\ a_y &= a\lambda_{V_y} \Big/ \sqrt{\lambda_{V_x}^2 + \lambda_{V_z}^2 + \lambda_{V_y}^2} \end{aligned} \tag{12}$$

and must fulfill the Legendre–Clebsch condition (the Second-Order Optimality Conditions), that is $\mathbf{H}_{uu} > 0$.

4 Solving of the Problem: Numerical Application

Practically, the analytical solution could not be determined due to the nonlinear structure of the equations which form this system, so because of this we shall also solve this problem using the *shooting* type numerical method [3]. Calculations were performed for circular and elliptical ($\varepsilon = 0.005$) orbits around the Earth, the final time being deduced.

4.1 Circular Orbits (Figs. 2–4)

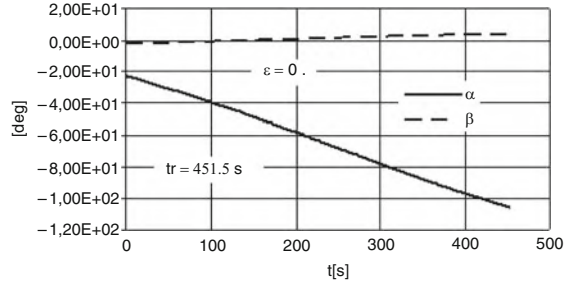


Fig. 2. Controls (Accelerations)

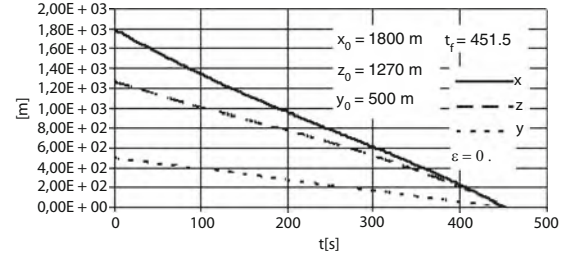


Fig. 3. Coordinates

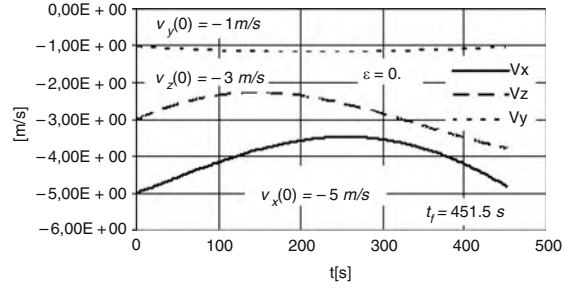


Fig. 4. Velocities

4.2 Elliptical Orbits (Figs. 5–7)

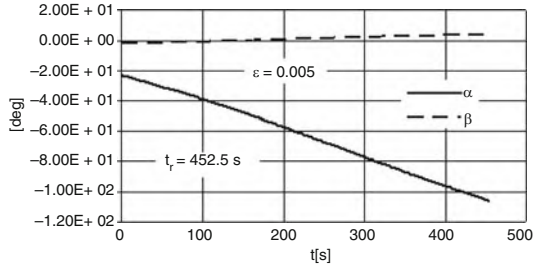


Fig. 5. Controls (Accelerations)

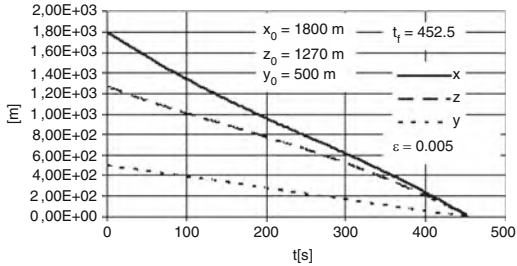


Fig. 6. Coordinates

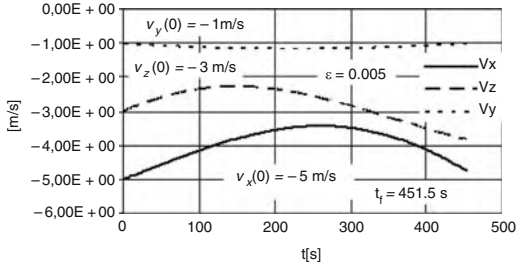


Fig. 7. Velocities

5 Conclusions

The calculations performed using based on the non-linear theory presented in this work show that the differences between the results obtained for circular orbits and the elliptical ones are small if the eccentricity of them are also small.

References

1. Bryson, A.E., Ho, Y.C.: Applied Optimal Control. Revised Printing, Hemisphere Publishing Corp., Washington D.C. (1975)
2. Istratie, V.: (AIAA-96-3646), Vol. AIAA/AAS, Astrodynamics Conference, a Collection of Technical Papers, pp. 667–675, San Diego, CA, USA (1996)
3. Istratie, V.: Ph.D. Thesis, Bucharest University (2000)
4. Leitman, G. (ed.): An Introduction to Optimal Control. Nauka, Moscow (1968)
5. Lembek, C.G., Prussing, J.E.: Guid. Control Dyn. J. **16**, 426–433 (1993)
6. Potriaguine, L., Boltianski, V., Gamkrelidze, E., Michtchenko, E. (ed.): The Mathematical Theory of Optimal Process. Mir, Moscow (1974)

Distributed Particle Swarm Intelligence for Optimization in the Water Industry

J. Izquierdo[1], I. Montalvo[1], R. Pérez[1], M.M. Tung[2] and M. Tavera[3]

[1] Centro Multidisciplinar de Modelación de Fluidos, jizquier@gmmf.upv.es, imontalvo@gmmf.upv.es, rperez@gmmf.upv.es
[2] Instituto de Matemática Multidisciplinar Universidad Politécnica de Valencia, Camino de Vera, s/n, 46022 Valencia, Spain, mtung@imm.upv.es
[3] Wasser S.A.E., Vía de las dos Castillas 33, edificio Atica 3, 28224 Pozuelo de Alarcón (Madrid), Spain, mtavera@wasser.es

Summary. Calibration and leak identification in Water Distribution Networks are of paramount importance in Water Industry. In this paper, Particle Swarm Optimization (*PSO*) is applied to tackle these problems. Standard *PSO*'s main drawback is that it is difficult to keep good levels of population diversity and to balance local and global searches. The formulation proposed, however, is able to find optimum or near-optimum solutions efficiently with considerably low computational effort because of the richer population diversity it introduces. Requiring only a low number of generations is a major advantage in real systems, where costs and time constraints prohibit too many iterations and hydraulic evaluations.

1 Introduction

Many problems in the Water Industry can be cast in the form of optimization problems. Before, we have considered the design of Water Distribution Networks (*WDN*) and Wastewater Systems [1–3]. But, taking into account the uncertainty of data (especially in existing configurations), it is frequently necessary to solve difficult inverse problems where optimization techniques are also of paramount importance. The calibration, identification and detection of leaks in a *WDN* can be reformulated as optimization problems. In fact, these problems are of essential interest in the Water Industry due to the great concern for finding mechanisms of sustainable water supply at a reasonable cost.

Classical methods of optimization involve the use of gradients or higher-order derivatives of the fitness function. But they are not well suited for many real world problems since they are not able to process inaccurate, noisy, discrete and complex data. Thus, more robust methods of optimization are often required to generate suitable results.

For the last decade, many researchers in the water field have embarked on the implementation of Evolutionary Algorithms: Genetic Algorithms [4–6];

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*,
Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_143,

Ant Colony Optimization [7]; Simulated Annealing [8]; Shuffled Complex Evolution [9]; and Harmony Search [10], among others. One of the evolutionary algorithms that has demonstrated a great potential for the solution of various optimization problems is *PSO*. The *PSO* algorithm was developed by Kennedy and Eberhart [11] and is a multi-agent optimization system inspired by the social behavior of a group of migrating birds trying to reach an unknown destination. This algorithm, with several modifications, is used in the present work to find solutions for calibration and leak detection problems in water systems. We provide the results of an application to a selected case-study.

2 PSO and Diversity-Increasing Variant

All evolutionary algorithms share two prominent features. First, they are all population-based. In *PSO*, each bird of the flock (swarm or population) represents a potential solution and is referred to as a *particle*. Second, there exists communication and therefore information exchange among the individuals. In this framework, the birds, besides having individual intelligence, also develop some social behavior and coordinate their movement towards a destination [11]. Initially, the process starts from a swarm of particles, in which each of them contains a candidate solution to the problem that is generated randomly, and then one searches the optimal solution by iteration. The performance of each particle is measured using a predefined fitness function, according to the problem at hand. The i-th particle is associated with (a) its current position, $X_i = (x_{i1} \dots x_{iD})$, where D is the number of variables involved in the problem; (b) its best position, $Y_i = (y_{i1} \dots y_{iD})$, reached in previous cycles; and (c) its flight velocity $V_i = (v_{i1} \dots v_{iD})$, which makes it evolve. In each cycle, the position of the best bird in the swarm, $Y^\star$, is updated. Then, the swarm is manipulated according to the equations

$$V_i' = \omega\, V_i + c_1\, \text{rand}(\,)\,(Y_i - X_i) + c_2\, \text{rand}(\,)\,(Y^* - X_i)\,, \tag{1}$$

$$X_i' = X_i + V_i, \tag{2}$$

where the prime denotes the new values. Here, c_1 and $c_2 > 0$ are two positive constants called learning factors or rates; rand( ) creates two independent random numbers between 0 and 1; ω is a factor of inertia suggested by Shi and Eberhart [12] which controls the impact of the velocity history into the new velocity. The ω factor permits to balance out global and local searches. It was suggested to have it decrease linearly with time, usually in a way to first emphasize global search and then prioritize local search. Equation (1) is used to calculate the i-th particle's new velocity, a description which considers three main ingredients: the particle's previous velocity, the distance of the particle's current position from its own best position, and the distance of the particle's current position from the swarm's *best experience* (position of the best particle). Thus, each particle or potential solution moves to a new position

according to (2). For each dimension, particle velocities are constrained by minimum and maximum velocities

$$V_{\min} \leq V_j \leq V_{\max}, \tag{3}$$

which are user defined parameters to control excessive roaming of particles outside of the search space. These important parameters determine the resolution with which regions between the present position and the target (best so far) positions are searched. If V_j is too big, particles might fly through good solutions. If V_j is too small, particles may not explore sufficiently beyond locally good regions and could easily be trapped in local optima.

PSO's main drawback is that it is difficult to maintain acceptable levels of population diversity and to balance local and global searches, and hence suboptimal solutions are prematurely obtained. In general, the random character, a typical feature of evolutionary algorithms, adds a degree of diversity to the manipulated populations. Nevertheless, in standard *PSO* these random components are unable to add a sufficient amount of diversity. As shown in [3], frequent collisions of birds occur in the search space, especially onto the leader. This, in fact, caused the effective population size to be lower and consequently produced a loss in the algorithm's effectiveness. The study in [13] presents a *PSO* variant in which a few of the best birds are selected to check collisions. Additionally, birds are re-generated completely at random when collision occurs. This random re-generation of the many birds which tend to collide with the best birds avoids premature convergence as it prevents clone populations from dominating the search. The inclusion of this procedure into *PSO* greatly increases diversity and improves convergence and quality of the final solutions.

The parameters have been selected after preliminary tuning following some suggestions [12,14]:

- $c_1 = 3, c_2 = 2;\ \omega = 0.5 + \dfrac{1}{2(\ln k + 1)}$, where k is the iteration number;
- $V_{\max} = 7\%$ and $V_{\min} = -V_{\max}$.

The termination condition stopped the process, if after 200 iterations no improvement in the solution had been obtained. A population of 300 particles was used.

3 Calibration of a WDN and Leak Identification

In the past, various optimization techniques have been considered to deal with the calibration and the identification of leaks in a *WDN* (see e.g. [15]). In this section, we show how the *PSO* algorithm can be used to tackle this problem.

Computational modeling of *WDN* is of great importance for *WDN* authorities. The complete set of equations for one of the models may be written using block matrix notation [16,17]:

$$\begin{pmatrix} A_{11}(q) & A_{12} \\ A_{12}^t & 0 \end{pmatrix} \begin{pmatrix} q \\ H \end{pmatrix} = \begin{pmatrix} -A_{10}H_f \\ Q \end{pmatrix}, \tag{4}$$

where A_{12} is the so-called connectivity matrix describing the way demand nodes are connected through the lines. Its size is $L \times N_p$, with L being the number of lines and N_p the number of demand nodes. Furthermore, q is the vector of the flowrates through the lines, H the vector of unknown heads at demand nodes; A_{10} is an $L \times N_f$ matrix, N_f being the number of fixed head nodes with known head H_f, and Q is the N_p-dimensional vector of demands. Finally, $A_{11}(q)$ is an $L \times L$ diagonal matrix. System (4) is a non-linear problem, whose solution is the state vector $x = (q, H)^t$ of the system.

In real-life applications, however, it is a very difficult task to create precise conditions for *WDN* hydraulic models. To be really useful, any such model should first be calibrated. Calibration is the process in which a certain number of model parameters are adjusted until this model closely mimics the behavior of the real system. Typical *WDN* parameters with significant uncertainty are pipe roughness and leaks.

For new pipes, roughness can be assessed directly through lab tests. But for an already existing *WDN*, these old manual methods are not accurate at all. Also, in order to obtain good values for the system losses, the flowrates through the pipes must be known with sufficient accuracy [18]. But these flowrates cannot be accurately determined, if there are leaks in the system (leaks can be considered as unknown demands). Frequently, leak identification can only be carried out globally through lumped audits. Much better results can be achieved, if the calibration of the analyzed *WDN* model is formulated via optimization problems using an objective function. The objective function attempts to reconcile the differences between the actually measured pressures and the predicted pressures of the hydraulic simulation which the computer model yields for a set of predefined system parameters. Roughness coefficients and leak magnitudes will be the variables of the fitness function. The discrepancy among measured and theoretical (given by the model) pressure heads is then minimized by using the following fitness function:

$$F = \sum_{j=1}^{N} p_j \cdot \left| H_j^m - H_j^c \right|. \tag{5}$$

Here, N denotes the number of demand junctions where pressure measures were taken (a limited number of all the junctions), and p_j the penalty for the discrepancy between the measured piezometric head H_j^m at node j and the calculated piezometric head H_j^c at node j. Furthermore, the penalty factor is taken as 10^5, if $\left| H_j^m - H_j^c \right|$ is bigger than the tolerance threshold allowed for node j, and otherwise 0. In the process of minimizing (5), the problem variables approach the values of their corresponding real parameters.

The proposed procedure has been applied to the Hanoi network [19]. The network topology uses the design data given in [19] for the length of the

pipes and the demand at the junctions. As stated in [19], pipe diameters were unknown, since it was a design problem. Here, we have assigned design values to the diameters of the pipes obtained in [3]. Also, a roughness coefficient of 130 (C of Hazen–Williams) has been assigned to all the pipes. Additionally, five new demand nodes have been considered in order to mimic the system leaks. By using *EPANET2* [20], the network was analyzed and the computed head values at the junctions were stored. These pressure heads, together with the assigned Hazen–Williams coefficient and the localization and magnitude of the leaks, represent synthetically the real (measured) values of the network. To measure the performance of the algorithm for the original network with the identification of leaking pipes, roughness coefficients were allowed to vary between 80 and 140, and leak exponents between 0 and 1.5. Running the algorithm 100 times, the difference between measured and calculated pressure heads was always found to be smaller than 0.15 mca.

4 Conclusions

Optimization problems in the field of urban hydraulics are complex in nature and difficult to solve by conventional optimization techniques. In particular for large *WDN*s, the optimization process of construction and maintenance needs the allocation of many resources every year. Also, a growing concern has arisen nowadays over water loss in existing *WDN*s, quite common with many aging elements, so that the calibration of friction and leakage is of paramount importance in drinking water systems.

In this work, *PSO* has been applied to the problem of calibration and leak identification in *WDN*s. Good solutions have been found for the case-study considered. A new feature which effectively increases the diversity of the population of birds has been included. This feature makes the algorithm converge with lesser iterations, and thus saves time, something of primary concern in real *WDN* problems.

The same approach should be extended to other networks, since this is a problem that has only received minor attention in the literature. In addition, parametric studies should be carried out in order to fine-tune the behavior of *PSO*. On the other hand, the results we have shown here are very promising and show that *PSO* is a reliable algorithm that must be considered as an excellent alternative to face optimization problems in water systems.

Acknowledgments

This work has been performed with support from the Grant MAEC-AECI 0000202066, awarded to one of the authors by the Ministerio de Asuntos Exteriores y Cooperación of Spain.

References

1. Montalvo, I., Izquierdo, J., Pérez, R., Tung, M.M.: Comput. Math. Appl. **56**(3), 769–776 (2007)
2. Izquierdo, J., Montalvo, I., Pérez, R., Fuertes, V.S.: Comput. Math. Appl. **56**(3), 777–784 (2007)
3. Montalvo, I., Izquierdo, J., Pérez, R., Iglesias, P.L.: Eng. Optim. **40**(7), 655–668 (2007)
4. Savic, D.A., Walters, G.A.: J. Water Resour. Plann. Manag. **123**(2), 67–77 (1997)
5. Matías, A.S.: Diseño de redes de distribución de agua contemplando la fiabilidad, mediante algoritmos genéticos, PhD thesis, Universidad Politécnica de Valencia, Valencia, Spain (2003)
6. Wu, Z.Y., Walski, T.: J. Water Resour. Plann. Manag. **131**(3), 181–192 (2005)
7. Zecchin, A.C., Simpson, A.R., Maier, H.R., Nixon, J.B.: IEEE Trans. Evol. Comput. **9**(2), 175–191 (2005)
8. Cunha, M.C., Sousa, J.: J. Water Resour. Plann. Manag. **125**(4), 214–221 (1999)
9. Liong, S.Y., Atiquzzaman, M.: J. Inst. Eng. Singapore **44**(1), 93–107 (2004)
10. Geem, Z.W.: Eng. Optim. **38**(3), 259–280 (2006)
11. Kennedy, J., Eberhart, R.C.: Particle Swarm Optimization in Proceedings of the *IEEE* International Conference on Neural Networks, pp. 1942–1948. Piscataway, NJ (1995)
12. Shi, Y., Eberhart, R.C.: A Modified Particle Swarm Optimizer in Proceedigns of the *IEEE* Congress on Evolutionary Computation, pp. 69–73. Piscataway, NJ (1998)
13. Herrera, M., Izquierdo, J., Montalvo, I., Pérez, R.: A Derivative of Particle Swarm Optimization with Enriched Diversity, Mathematical and Computer Modelling (2010) (submitted for publication)
14. Shi, X.H., Liang, Y.C., Lee, H.P., Lu, C., Wang, Q.X.: Inf. Process. Lett. **103**(5), 169–176 (2007)
15. Mariles, Ó.A.F., Nava, A.P.: Calibración del factor de fricción y localización de fugas en una red de tuberías de agua potable in: VII SEREA – Seminario Iberoamericano sobre Planificación, Proyecto y Operación de Sistemas de Abastecimiento de Agua, 2007, Morelia, México
16. Izquierdo, J., Pérez, R., Iglesias, P.L.: Math. Comp. Model. **39**(11–12), 1353–1374 (2004)
17. Izquierdo, J., Tung, M.M., Pérez, R., Martínez, F.J.: Estimation of fuzzy anomalies in Water Distribution Systems, in: Progress in Industrial Mathematics at *ECMI* 2006 (Series: Mathematics in Industry. Subseries: The European Consortium for Mathematics in Industry), vol. 12, pp. 801–805. Springer, Berlin (2008)
18. Izquierdo, J., López, P.A., Martínez, F.J., Pérez, R.: Math. Comput. Model. **46**(3–4), 341–350 (2007)
19. Wu, Z.Y., Simpson, A.R.: J. Comput. Civ. Eng. **15**(2), 89–101 (2001)
20. Rossman, L.A.: *EPANET* User's Manual, U.S. Environmental Protection Agency, Cincinnati (2000)

Application of the Method of Auxiliary Sources in Optical Diffraction Microscopy

M. Karamehmedović[1], M.-P. Sørensen[1], P.-E. Hansen[2], and A. Lavrinenko[3]

[1] DTU Mathematics, Technical University of Denmark, Matematiktorvet 303S, DK-2800 Kgs. Lyngby, Denmark, M.Karamehmedovic@mat.dtu.dk, m.p.soerensen@mat.dtu.dk
[2] Danish Fundamental Metrology, Technical University of Denmark, Matematiktorvet 307, DK-2800 Kgs. Lyngby, Denmark, peh@dfm.dtu.dk
[3] DTU Fotonik, Technical University of Denmark, Matematiktorvet 303S, DK-2800 Kgs. Lyngby, Denmark, alav@fotonik.dtu.dk

Summary. The Method of Auxiliary Sources is used for characterisation of grating defects. Grating profiles are characterised by best fit matching of a library of diffraction efficiencies with numerical simulated diffraction efficiencies with defects. It is shown that the presented method can solve the inverse problem with an accuracy usually thought to require rigorous electromagnetic theories.

1 Characterisation of Micro and Nano Structures Embedded in Materials

Functional materials with embedded micro and nano structures find application in such diverse areas of technology as optical telecommunication components, self-cleaning windows, medical equipment, and the technology of mass production of electronics and digital watermarks. The main useful properties of such materials are not intrinsic, but rather stem from the introduced modifications on or just beneath the surface of the material. The modifications are, e.g., insertion of particles or air holes of micro or nano scale under the material surface, and alterations of the topology of the surface, such as the introduction of surface gratings or deposition of particles, on micro and nano scale. The design process and industrial use of functional materials require rapid and non-destructive techniques of characterisation of the embedded micro and nano structures. Among several physically distinct methods, we focus on the combined spectroscopic and angular resolved scatterometry technique called Optical Diffraction Microscopy (ODM) [1, 2, 5–7]. Here, specific features of the sample under investigation are reconstructed from the measured optical power in the scattered far field. The method thus requires the solution of an inverse scattering problem, and ultimately of a nonlinear optimisation problem; however, in an industrial context such as quality control,

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*, Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_144,

the principal features of the scatterer may be well-known, and one needs rapid interpretation of measurement results to identify only relatively small perturbations, e.g., manufacturing errors, in these features. The structures of interest are typically small in terms of the wavelength of the illuminating light, and it is therefore relevant to address the inverse scattering problem using the full classical electromagnetic model, rather than asymptotic formulations. The Method of Auxiliary Sources (MAS) is an efficient numerical, non-asymptotic technique of solution of boundary problems; see [4,8] and references therein. In the following, the method is used to approximate the solution of example inverse problems which arise in Optical Diffraction Microscopy.

2 The Method of Auxiliary Sources

In the context of two-dimensional, time-harmonic forward electromagnetic scattering, the Method of Auxiliary Sources (MAS) is a variational method characterised by the choice of fundamental solutions of the Helmholtz equation in $\mathbb{R}^2$ for the expansion vectors of the scattered field, and the Dirac delta functions for the test vectors. Recall that, for every positive k, an outgoing[1] fundamental solution of the Helmholtz equation $(\Delta+k^2)u = 0$ in $\mathbb{R}^2$, with singularity at $x' \in \mathbb{R}^2$, is the Hankel function $H_0^{(2)}(k|x-x'|)$ of order zero and of second kind. Figures 1 and 2 show a model time-harmonic Dirichlet scattering problem in $\mathbb{R}^2$ and a corresponding MAS formulation. The constant k is the wave number $2\pi/\lambda$, where λ is the operating wavelength. In MAS, all employed fundamental solutions have singularities in the interior of the scatterer. The current sources of the approximation of the exact scattered field – the so-called *auxiliary sources* – are hence Delta functions in $\mathbb{R}^2$ with singularities in the interior of the scatterer, and, in the transverse electric (TE) case, the scattered field E^{s} is approximated in the exterior Ω by a finite linear combination of the form $E^{\mathrm{MAS}}(x) = \sum_{j=1}^{N} C_j H_0^{(2)}(k|x-x'_j|)$, $x \in \Omega$. The weights (complex numbers C_j) occurring in the linear combination are determined by enforcing the boundary condition at selected points x_l, $l = 1, \ldots, N$, on the scatterer boundary Γ. The classical inverse scattering problem which arises in the ODM consists in finding a surface Γ and a surface current distribution J on Γ such that the corresponding radiated far field has the same power pattern as the measured field. Evaluation of the objective function of this nonlinear optimisation problem necessarily involves the evaluation of the intermediate scattered far fields. In this context, the MAS representation of scattered fields holds two major advantages over the traditional surface integrals which originate from boundary layer potential formulations of scattering problems. First, with the MAS formulation, there is no need for numerical integration of surface currents, whereas the electric field radiated by a z-directed,

[1] That is, satisfying the outgoing radiation condition.

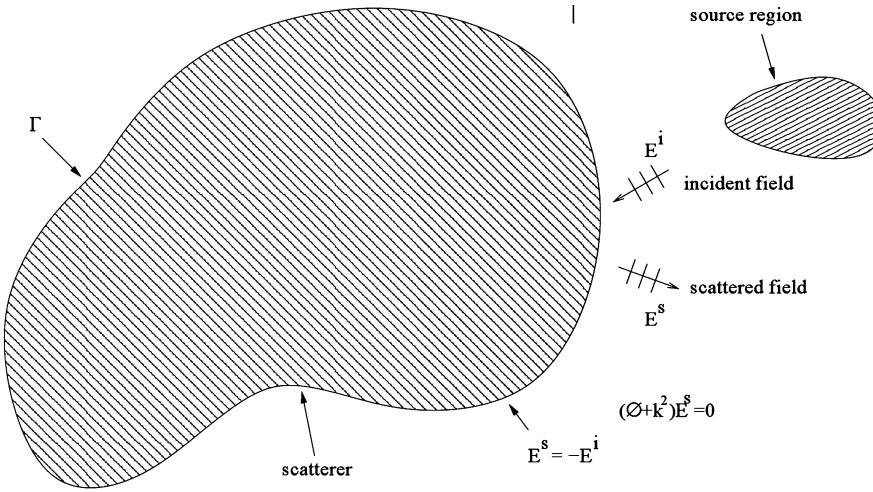


Fig. 1. The geometry of a scattering problem in a subset Ω of $\mathbb{R}^2$

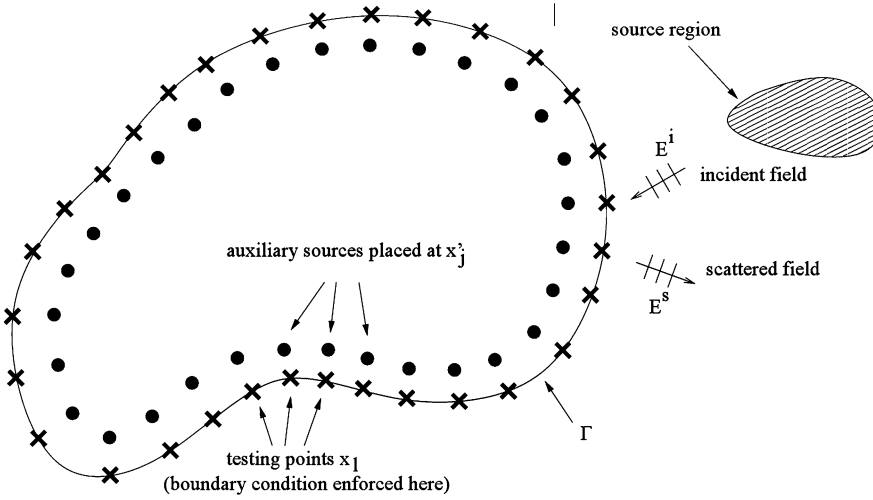


Fig. 2. A MAS setup used to approximate the solution of the considered boundary problem

time-harmonic electric current distribution J on a boundary Γ in $\mathbb{R}^2$ is proportional to the integral $\int_\Gamma H_0^{(2)}(k|x-x'|)J(x')d\Gamma(x')$, for x in the exterior of Γ. The second major advantage of the MAS is that the scatterer topology is identified only with the auxiliary sources, rather than with the sources *and* with a supporting boundary Γ. In the above-mentioned integral, the domain of integration Γ is, in general, a parameter of optimisation, and hence needs to be changed with each iteration. In conclusion, when MAS is used, the

optimisation problem involves an objective function which is simply a finite sum independent of the actual geometry of the scatterer surface, as opposed to an integral taken over a generally variable surface. In our implementation, described in Sect. 3, a number of scattered far field power patterns are stored in a library, together with the corresponding sets of auxiliary sources. (The latter are represented by the locations x'_j, $j = 1, \ldots, n$, and the complex amplitudes C_j, $j = 1, \ldots, n$; these sources radiate suitable approximations of the stored far field patterns.) Each far field power pattern corresponds to a well-defined perturbation of the basic topology of the scatterer. With elements x and x' of $\mathbb{R}^2$ represented by $(|x|, \phi)$ and $(|x'|, \phi')$, respectively, in the usual cylindrical coordinates, the function $\frac{1+i}{\sqrt{\pi k|x-x'|}} e^{-ik|x-x'|} e^{ik|x'|\cos(\phi-\phi')}$ is the asymptotic form of the Hankel function $H_0^{(2)}(k|x-x'|)$ of order zero and of second kind, valid for $|x-x'| \gg \lambda$. We use the phase function $e^{ik|x'|\cos(\phi-\phi')}$ of this asymptotic form for the auxiliary sources in our implementation. The procedure first compares the measured far-field power pattern with the direct samples in the library, using a distance function of the form $\sum \left| |E^{\text{library}}(\phi_l)| - |E^{\text{m}}(\phi_l)| \right|^2$, where $|E^{\text{m}}(\phi_l)|$ is the measured magnitude of the far field at angle ϕ_l. After the best match is found, simple interpolation is used to refine this solution of the inverse scattering problem. The auxiliary sources corresponding to the best match, as well as those corresponding to the two entries in the library which are closest to the best match, are fetched; these sources are represented by complex amplitude vectors C_0, C_{-1} and C_1 in $\mathbb{C}^N$, respectively. The objective function, which is a finite sum of the form $\sum \left| |E_t^{\text{MAS}}(\phi_l)| - |E^{\text{m}}(\phi_l)| \right|^2$, is then minimised with respect to the parameter $t \in [-1, 1]$; the field E_t^{MAS} is radiated by the auxiliary sources represented by the complex amplitude vector $C(t) = -tC_{-1} + (1+t)C_0$ when $t \in [-1, 0]$, and by $C(t) = (1-t)C_0 + tC_1$ when $t \in [0, 1]$. It is here assumed that the library entries are sufficiently close such that the error is, to a good approximation, a linear function of the perturbation of the scatterer geometry. The optimum value of the parameter t is therefore directly interpreted as a normalised deviation of the measured geometry from the library entries.

3 Results

Figure 3 shows the two-dimensional scattering problem considered here, and the type of the measured deviations in the scatterer topology. The scatterer, a piece of corrugated silicon, is immersed in air and illuminated by a time-harmonic, uniform plane wave of transverse electric (TE) polarisation and unit amplitude. The incident field propagates in the negative x direction. The operating wavelength is denoted λ. We want to measure the elongation of a specific protrusion on the scatterer. Our numerical experiment does not use actual field measurements; rather, the amplitude of the scattered electric far field is calculated using the COMSOL software [3,9]. The library entries are

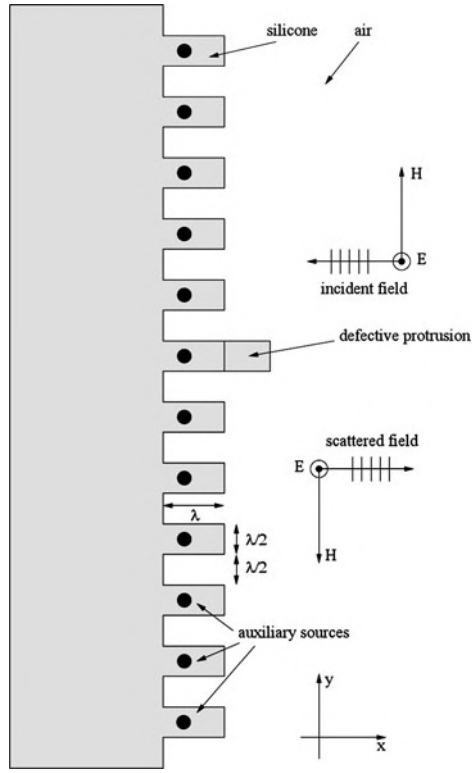


Fig. 3. The type of grating defects to be characterised

samples of the magnitude of the scattered electric far field, taken over the angle of 30° symmetrically with respect to the x-axis. A total of only 12 auxiliary sources are used for the interpolation of the far fields. Table 1 shows the results of the numerical experiment. The actual and the estimated elongations in the table are shown normalised with respect to the operating wavelength. Negative (positive) elongations correspond to the specific protrusion being shorter (longer) than the nominal one wavelength λ. Relative error 1 and relative error 2 show the error in the estimate relative to the actual elongation and relative to the nominal protrusion length, respectively. For the results in Table 1, the average absolute value of relative error 1 is 14.7%, and the average absolute value of relative error 2 is 5.4%. Of course, the elongations already represented in the library are measured with zero error, which improves the overall accuracy estimate for the method. However, it also turns out that the elongations of 0.2λ, 0.4λ and 0.6λ match well the library entries of 0.875λ and 1λ, which suggests that, in general, an appropriate a priori estimate is needed of the possible range of the elongation under measurement. After forcing the correct initial (library) values for the three above-mentioned elongations, the

Table 1. Accuracy of the estimates of the protrusion elongation

Actual elongation	Estimated elongation	Relative error 1 (%)	Relative error 2 (%)
−0.9	−0.9206	−2.3	−2.1
−0.3	−0.1800	40.0	12.0
−0.2	−0.1800	10.0	2.0
−0.1875	−0.1800	4.0	0.75
0.1	0.0681	−31.9	−3.2
0.3125	0.3250	4.0	1.3
0.8125	0.6681	−17.8	−14.4
0.9	0.9713	7.9	7.1

interpolation produces estimates with relative error 2 at 12.5%, −18.8% and −6.0%, respectively.

4 Conclusions and Further Work

It was demonstrated that the Method of Auxiliary Sources can be used for efficient numerical approximation of solution of certain inverse scattering problems occurring in two-dimensional monochromatic Optical Diffraction Microscopy. The method was tested on a number of relevant two-dimensional inverse problems involving the elongation of a specific protrusion in a grating. Future work includes the generalisation of the presented method to three-dimensional measurement, and to polychromatic measurement (in time domain).

Acknowledgement

We acknowledge the financial support from the innovation consortium FINST under the Danish Agency for Science, Technology and Innovation.

References

1. Agersnap, N., Hansen, P.-E., Petersen, J.C., Garnaes, J., Destouches, N., Parriaux, O.: Proc. SPIE Int. Soc. Opt. Eng. **5965**, 68–78 (2005)
2. Borsetto, F., Carneiro, K., Davi, I., Garnaes, J., Petersen, J.C., Agersnap, N., Hansen, P.-E., Holm, J., Christensen, L.H.: Proceedings of the 7th International Conference and 8th General Meeting of the European Society for Precision Engineering and Nanotechnology, Baaden, May 2006
3. COMSOL Multiphysics demonstration CD-ROM can be requested at http://www.comsol.com
4. Fairweather, G., Karageorghis, A.: Adv. Comput. Math. **9**, 69–95 (1998)
5. Garnaes, J., Hansen, P.-E., Agersnap, N., Davi, I., Petersen, J.C., Kuhle, A., Holm, J., Christensen, L.H.: Proc. SPIE Int. Soc. Opt. Eng. **5878**, 1–9 (2005)

6. Garnaes, J., Hansen, P.-E., Agersnap, N., Holm, J., Borsetto, F., Kuhle, A.: Appl. Opt. **45**, 3201–3212 (2006)
7. Hansen, P.-E., Nielsen, L.: Mater. Sci. Eng. B **65**, 165–168 (2009)
8. Kaklamani, D.I., Anastassiu, H.T.: IEEE Antennas Propag. **44**, 48–64 (2002)
9. Zimmerman, W.B.J.: Multiphysics Modelling with Finite Element Methods. World Scientific, New Jersey (2006)

Radial Basis Function (RBF) Solution of the Motz Problem

Manuel Kindelan and Francisco Bernal

G. Millán Institute of Modeling, Simulation and Industrial Mathematics
Universidad Carlos III de Madrid, 28911 Leganés, kinde@ing.uc3m.es,
bernal@maths.ox.ac.uk

Summary. The Motz problem can be considered as a benchmark problem for testing the performance of numerical methods in the solution of elliptic problems with boundary singularities. In this work, we address the solution of the Motz problem using the Radial Basis Function (RBF) method. We show that the accuracy of the solution can be significantly increased by using special functions which capture the behavior of the singularity.

1 Introduction

Standard numerical methods (finite element, boundary element, finite difference, spectral methods) are very efficient in solving elliptic partial differential equations, except for problems containing singularities, when their high-order convergence rates deteriorate. Unfortunately, these singularities are often present in problems of engineering interest, either due to an abrupt change in the boundary condition, or due to the presence of re-entrant corners.

Motz's problem can be considered as a prototype to check the efficiency of numerical schemes in solving this type of problems. It was first proposed by Motz [9] in 1947 and later modified by Wait and Mitchell [12]. It consists in finding a solution to the Laplace equation in a rectangular domain with the following boundary conditions,

$$u|_{x<0,y=0} = 0\,, \qquad u|_{x=1} = 500$$
$$u_y|_{y=1} = u_y|_{y=0,x>0} = u_x|_{x=-1} = 0$$

The solution has a singularity at the origin due to the change from Dirichlet to Neumann boundary conditions. It is representative of many problems of engineering interest containing this type of boundary conditions which lead to singularities in the first derivatives at the origin.

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*,
Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_145,

In fact, the solution can be expressed as,

$$u = \sum_{i=1}^{\infty} A_i \, r^{(i-1/2)} \, \cos\left[\left(i - \frac{1}{2}\right)\theta\right] \tag{1}$$

where r and θ are polar coordinates centered at the singular point. Since the convergence radius of (1) is 2 [11], the above expansion is valid in the entire solution domain. The coefficients A_i have been accurately computed by several authors [7, 11].

In this paper, we use the Radial Basis Function (RBF) [5, 6] method to compute the solution of Motz's problem. We show that the accuracy of the solution is significantly improved by using singular functions which capture the behavior of the solution near the discontinuity. To this end, we enlarge the functional space spanned by the RBF basis functions by adding singular functions which capture the behavior of the local singular solution. Similar ideas have been used in the past by Platte and Driscoll [10] in the solution of an eigenvalue problem, and by Hu et al. [4] using inverse multiquadrics (IMQRB) and Gaussian (GRB) radial basis functions.

2 Global RBF Solution

To solve Motz's problem with the RBF method, we look for a solution in the space spanned by the RBF multiquadric functions,

$$u(\mathbf{x}) = \sum_{i=1}^{N} \alpha_i \, \phi_i(\mathbf{x}), \quad \phi_i(\mathbf{x}) = \sqrt{\| \mathbf{x} - \mathbf{x}_i \|^2 + c^2}, \qquad i = 1 \ldots N \tag{2}$$

where c is the shape parameter. The RBF nodes $\mathbf{x}_i$ are located in an equi-spaced grid. The coefficients α_i are computed by collocation of the PDE in the interior nodes and collocation of the boundary conditions at the boundary nodes. To characterize the accuracy of the solution we use a fine grid of 5,000 evaluation nodes, and compute the mean square error, ϵ, on those nodes.

The left side of Fig. 1 shows the RBF solution obtained with a grid of 21×41 RBF nodes ($N = 861$) and an optimum value of the shape parameter $c = 0.3$. Notice the oscillations occurring in the vicinity of the singular point. Inaccuracy of the RBF solution at the boundaries has been often reported in the past [3]. This degradation is especially severe in the presence of discontinuous boundary conditions. In fact, the exact solution exhibits a sharp feature that does not belong to the interpolation space of the smooth RBFs.

The right side of Fig. 1 shows the dependence of the mean square error on the shape parameter, c, and on resolution. Notice that the error decreases with increasing shape parameter until for a certain value of c the system becomes ill-conditioned leading to a significant increase in error.

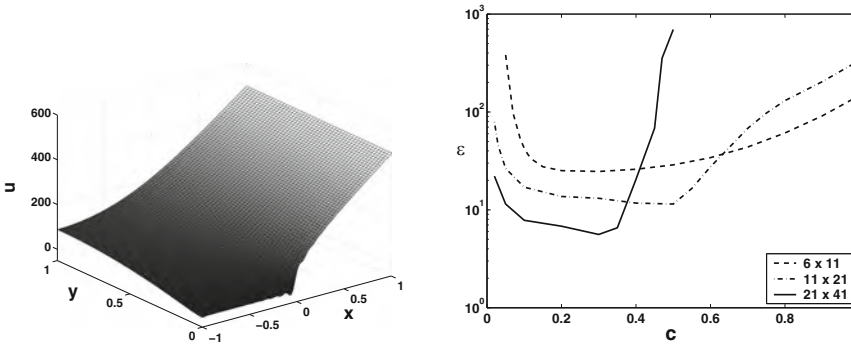


Fig. 1. *Left*: RBF solution. *Right*: Mean square error as a function of c

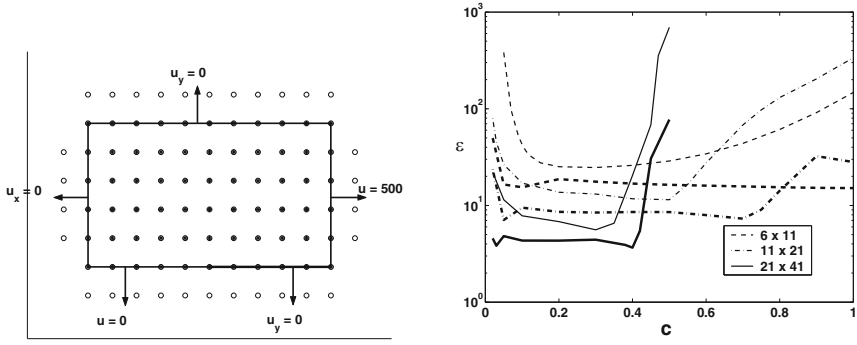


Fig. 2. *Left*: RBF centers (*open circle*) and Collocation nodes (*asterisk*). *Right*: Mean square error as a function of c. RBF solution (*thin*), PDE collocation at boundary (*thick*)

The accuracy of the solution, can be improved by enforcing collocation of the PDE also in boundary nodes [2]. However, since the number of equations increases, it is necessary to introduce additional RBF centers to match the number of unknown coefficients α_i. The left side of Fig. 2 shows the set of RBF centers (open circle) and the set of collocation nodes (asterisk).

The right side of Fig. 2 compares the dependence of mean square error as a function of c, using collocation of the Laplace equation at boundary nodes (thick lines), with that obtained with the standard RBF method (thin lines). A significant increase in the accuracy of the solution is observed.

3 Use of Singularity Capturing Functions

The errors in the RBF solution are concentrated in the vicinity of the singularity, and are due to the inability of the RBF expansion to capture them. To improve the accuracy of the numerical solution it is convenient to enlarge

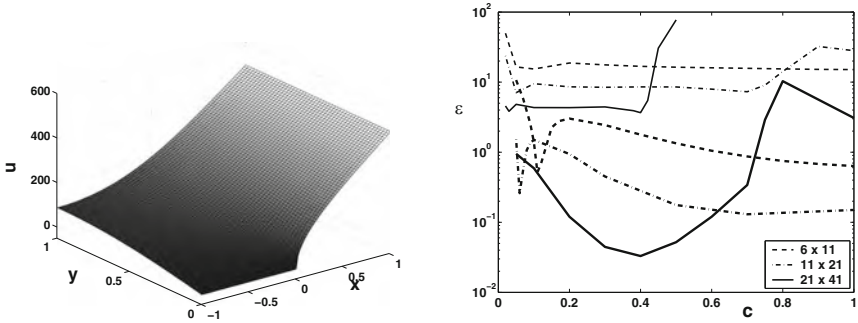


Fig. 3. *Left*: solution with 1 special function (21×41, $c = 0.3$). *Right*: Mean square error as a function of shape parameter c

the functional space spanned by the RBFs, by including a new function which captures the discontinuity in the boundary condition. Thus, we include as additional function the first term in the asymptotic expansion (1), namely

$$\phi_{N+1}(\mathbf{x}) = \sqrt{r}\cos\left(\frac{\theta}{2}\right), \qquad \theta = \arctan\left(\frac{y}{x}\right)$$

where ϕ_{N+1} is the first function in (1). Notice that ϕ_{N+1} satisfies the Laplace equation and the boundary conditions at $y = 0$.

Since there are $N+1$ unknowns, α_i, and N collocation nodes, an additional equation is required. We follow [10] in requiring the compatibility condition,

$$\sum_{i=1}^{N} \alpha_i \, \phi_{N+1}(\mathbf{x}_i) = 0$$

The left side of Fig. 3 shows the solution obtained with the RBF method enhanced with a special function to capture the singularity. Notice that the oscillations occurring in the vicinity of the singular point have completely disappeared. The right side of Fig. 3 compares the dependence of mean square error on c, using collocation of the Laplace equation at boundary nodes plus one special singularity capturing function (thick lines), with that obtained without special functions (thin lines). Significant improvements in accuracy are observed for all three grid resolutions. Including additional special functions leads to further improvements in accuracy but these are significantly lower than those obtained when adding the first special function.

However, the main result that appears in Fig. 3, is that the exponential convergence of the RBF method is recovered. Madych [8] gave an error estimate for multiquadrics as $O(e^{ac}\,\lambda^{c/h})$, where h is the mesh size. This result has a great impact in the numerical solution of PDEs, because it implies that solution accuracy can be increased either by refining the mesh (and increasing

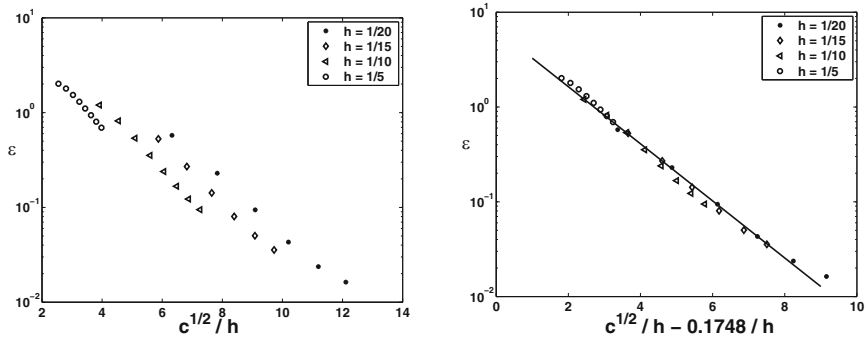


Fig. 4. *Left*: mean square error as a function of $\sqrt{c}/h$. *Right*: mean square error as a function of $\sqrt{c}\,/\,h\,-\,0.1478\,/\,h$, and best linear fit to the data

the computational cost), or by simply increasing the shape parameter c. However, as c increases the multiquadric functions become flat and the resulting system becomes ill-conditioned. This is shown in Fig. 3; accuracy increases with c until it reaches the breakdown point caused by machine roundoff error.

Although Madych [8] results were derived for the interpolation problem, several authors have have also found exponential convergence in the solution of PDEs. However, there is no agreement in the exact form of the dependence of the error on c and h. Hu et al. [4] follow Madych in considering that the error is $O(\lambda^{c/h})$, while Cheng et al. [1] results converge as $O(\lambda^{\sqrt{c}/h})$.

To check the dependence of the error on c and h for our data, we carried out several numerical experiments. The left side of Fig. 4 shows the mean square error of these experiments as a function of $\sqrt{c}/h$. The exponential convergence of the error is readily apparent. However, the experiments for different resolutions fail to collapse into a single curve. Therefore, the error convergence is not simply $O(\lambda^{\sqrt{c}/h})$, as was the case in Cheng et al. [1] experiments, but an additional dependence on h is apparent. In fact, if we plot our results as a function of $\sqrt{c}\,/\,h\,-\,0.1478\,/\,h$ the error of all the different numerical experiments coalesce into a single curve (see right side of figure 4),

$$\epsilon\,=\,6.54\,\cdot\,0.5^{\left(\sqrt{c}/h\,-\,0.1478\,/\,h\right)}$$

4 Conclusions

We analyze the performance of the RBF method in the solution of Motz's problem. However, the results obtained are applicable to a wide range of problems with boundary conditions that lead to singularities in the first derivatives of the solution. We show, that the exponential convergence which is typical of the RBF method, is lost in this type of problems containing singularities. The accuracy of the solution can be increased by collocation of the PDE at

boundary nodes. However, in order to restore the exponential convergence of the RBF method, it is necessary to use special functions which capture the behavior of the solution near the discontinuity.

Acknowledgements

MECD grants MAT2005-05730, FIS2004-03767 and FIS2007-62673 and by Madrid Autonomous Region grant S-0505.

References

1. Cheng, A.H-D., Goldberg, M.A., Kansa, E.J., Zammito, G.: Exponential convergence and H-c multiquadric collocation method for partial differential equations. Numer. Meth. Part. Differ. Equat. **19**, 571–594 (2003)
2. Fedoseyev, A.I., Friedman, M.J., Kansa, E.J.: Improved multiquadric method for elliptic partial differential equations via PDE collocation on the boundary. Comput. Math. Appl. **43**, 439–455 (2002)
3. Fornberg, B., Driscoll, T.A., Wright, G., Charles, R.: Observation of the behavior of radial basis function approximations near boundaries. Comput. Math. Appl. **43**, 473–490 (2002)
4. Hu, H.Y., Li, Z.C., Cheng, A.H.D.: Radial basis collocation methods for elliptic boundary value problems. Comput. Math. Appl. **50**, 289–320 (2005)
5. Kansa, E.J.: Multiquadrics, a scattered data approximation scheme with applications to computational fluid dynamics. I. Surface approximations and partial derivatives estimates, Comput. Math. Appl. **19**, 127–145 (1990)
6. Kansa, E.J.: Multiquadrics, a scattered data approximation scheme with applications to computational fluid dynamics. II. Solutions to parabolic, hyperbolic and elliptic partial differential equations. Comput. Math. Appl. **19**, 147–161 (1990)
7. Lu, T.T., Hu, H.Y., Li, Z.C.: Highly accurate solutions of Motz's and the cracked beam problems. Eng. Anal. Bound. Elem. **28**, 1387–1403 (2004)
8. Madych, W.D.: Miscellaneous error bounds for multiquadric and related interpolators. Comput. Math. Appl. **24**, 121–138 (1992)
9. Motz, H.: The treatment of singularities of partial differential equations by relaxation methods. Quart. Appl. Math. **4**, 373–377 (1947)
10. Platte, R., Driscoll, T.A.: Computing eigenmodes of elliptic operators using radial basis functions. Comput. Math. Appl. **48**, 561–576 (2004)
11. Rosser, J.B., Papamichael, N.: A power series solution of a harmonic mixed boundary value problem, MRC Technical Summary Report 1405, University of Wisconsin (1975)
12. Wait, R., Mitchell, A.R.: Corner singularities in elliptic problems by finite element methods. J. Comput. Phys. **8**, 45 (1971)

Bilevel Optimization of Container Cranes

M. Knauer and C. Büskens

Center of Industrial Mathematics, Universität Bremen, Germany
knauer@math.uni-bremen.de, bueskens@math.uni-bremen.de

Summary. Bilevel optimal control problems are presented as an extension of classical optimal control problems. Hereby, additional constraints are considered for the primary problem, which depend on the optimal solutions of secondary optimal control problems.

The numerical solution of the bilevel approach is illustrated by an application of a container crane. Time and energy optimal trajectories are calculated under the terminal condition that the crane system comes to be at rest at a predefined location. Additional bilevel constraints ensure that the crane system can be brought optimally to a rest position at a free location from any state of the trajectory. [1]

1 Container Cranes in Warehouses

Container cranes are an efficient alternative to commonly used forklift trucks if looking for special automated solutions in high rack warehousing.

A trolley at the top of a rack performs the horizontal movement. A payload is attached to it with wire ropes for vertical movement. After fixing the payload to the rack, the goods can be loaded or unloaded with a fork-like construction. As the trolleys can be attached to different heights in the rack, the usage of more trolleys in one rack can increase the overall performance.

When moving the trolley, however, the payload will start to swing. The task of trajectory planning is to move the whole system from one position in the rack to another and to ensure that there will be no oscillation when reaching the terminal point.

To make this system work in industrial facilities, further safety requirements have to be considered.

If using several cranes in one rack, the trajectories have to be generated avoiding collision between the cranes. If there is only one trolley, it still has to consider walls or corridors.

[1] This paper is based on a cooperation project with Westfalia, Borgholzhausen.

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*,
Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_146,

If in case of a power failure, the system switches off immediately, the swinging payloads must not collide neither.

Finally, the user might want to invoke a controlled braking. Here, the crane system has to stop fast without any oscillation remaining, no matter where the final position is. This means, that alternative trajectories have to be already available online, such that the control unit can switch to these.

The calculation of the original trajectory and its so-called safety stop trajectories can be interpreted as a bilevel problem:

The existence of an alternative trajectory is required at each point of the main trajectory. On the other hand the initial points of the safety stop trajectories depend on the main trajectory.

2 Bilevel Optimal Programming

In optimal programming, the task is to minimize a function under some constraints. When considering bilevel optimal programming, additional optimization problems also occur in the constraints:

$$\begin{aligned} &\min_{x \in X} F(x, y) \\ &\text{subject to } C(x, y) \geq 0, \\ &\qquad \text{for each fixed } x,\ y = y(x) \text{ is a solution to} \\ &\min_{y \in Y} f(x, y) \\ &\text{subject to } c(x, y) \geq 0 \end{aligned} \tag{1}$$

In game theory, this is interpreted as two decision makers or two players. The player on the upper level is called the leader, the player on the lower level is the follower. Both follow different objectives F and f and both have their own set of parameters x and y, influencing both objective functions. The constraints for the players are C and c, respectively.

The solution of a bilevel programming problem (1) is generally NP-hard. There exist methods for special cases or some heuristics to get approximate solutions.

If differentiability properties hold, the first-order necessary conditions for the lower-level problem of (1) can be used to replace it, such that a single-level problem remains:

$$\begin{aligned} &\min_{x \in X} F(x, y) \\ &\text{subject to } C(x, y) \geq 0 \\ &\nabla_y \ell(x, y, \lambda) = 0, \\ &c(x, y) \geq 0, \\ &\lambda^T \geq 0, \\ &\lambda^T c(x, y) = 0, \end{aligned} \tag{2}$$

with the Lagrange multiplier vector λ, and $\ell(x, y, \lambda) = f(x, y) - \lambda^T c(x, y)$.

3 Bilevel Optimal Control Problem

To generate a trajectory for a crane system, an optimal control problem has to be solved. This can be extended to a bilevel optimal control problem, whose solution consists of a trajectory calculated at the upper level and an alternative trajectory calculated at the lower level:

$$
\begin{aligned}
&\min_{x,u} G(x(0), x(t_f), y, v) + \int_0^{t_f} F_0(x(t), u(t), y, v)\, dt \\
&\text{subject to } \dot{x}(t) = F(x(t), u(t), y, v), \quad \text{for all } t \in [0, t_f], \\
&\qquad x_i(0) = A_i, \qquad\qquad \text{for } \; i \in I \subset \{1, \ldots, n\}, \\
&\qquad x_j(\tau_f) = \Omega_j, \qquad\qquad \text{for } \; j \in J \subset \{1, \ldots, n\}, \\
&\qquad C(x(t), u(t), y, v) \le 0, \qquad \text{for all } t \in [0, t_f], \\
&\qquad \text{for each fixed } x \text{ and } u, \\
&\qquad y(\tau) = y(\tau; x, u) \text{ and } v(\tau) = v(\tau; x, u) \text{ are a solution to} \\
&\min_{y,v} g(x, u, y(0), y(\tau_f)) + \int_0^{\tau_f} f_0(x, u, y(\tau), v(\tau))\, d\tau \\
&\text{subject to } \dot{y}(\tau) = f(x, u, y(\tau), v(\tau)), \quad \text{for all } \tau \in [0, \tau_f], \\
&\qquad y_i(0) = \alpha_i, \qquad\qquad \text{for } \; i \in \iota \subset \{1, \ldots, n\}, \\
&\qquad y_j(\tau_f) = \omega_j, \qquad\qquad \text{for } \; j \in \phi \subset \{1, \ldots, n\}, \\
&\qquad c(x, u, y(\tau), v(\tau)) \le 0, \qquad \text{for all } \tau \in [0, \tau_f].
\end{aligned}
\tag{3}
$$

The upper-level problem is to optimize the control vector u depending on time t, such that an objective function is minimized. The state vector x holds the system of differential equations F. Initial conditions A and terminal conditions Ω as well as constraints C can be included.

On the lower level, a similar problem is given. Both problems are coupled via the information in state vectors x and y and the control vectors u and v.

Similarly to (2), the first-order necessary conditions for the lower-level problem of (3) can be formulated as

$$
\begin{aligned}
\dot{\lambda}(\tau) &= -H_y, \\
H_v &= 0, \\
\lambda_i(0) &= 0, \qquad \text{if } y_i(0) \text{ free}, \\
\lambda_j(\tau_f) &= 0, \qquad \text{if } y_j(\tau_f) \text{ free}, \\
c(x, u, y(\tau), v(\tau)) &\le 0, \qquad \text{for all } \tau \in [0, \tau_f],
\end{aligned}
$$

with the Hamiltonian $H = l_0 f_0 + \lambda^t(\tau) f$.

In this formulation, only one follower is considered, and for that reason only one alternative trajectory is calculated. This means, that only at one point in time a safety stop can be requested.

Ideally, a bilevel problem with an infinite number of followers for every point in time of the upper-level problem has to be formulated. To solve this

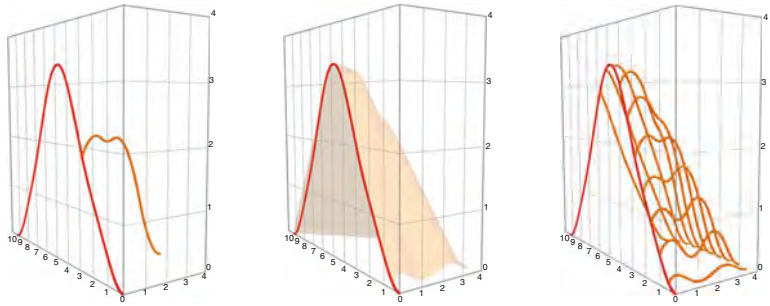


Fig. 1. Velocity of the optimal trajectory. From *left to right*: bilevel problem with one follower, bilevel problem with infinite number of followers, bilevel problem with reduced number of followers

problem numerically, memory requirements and calculation times have to be reduced, and hence the problem is restricted to a finite number of followers. A transformation of all problems to one fixed time scale finally allows the simultaneous calculation of the upper- and all lower-level problems, see Fig. 1.

The k lower-level problems, invoked from the upper-level at the discrete points in time $T_1, \ldots, T_k$, can be put together to one large lower-level problem by combining the control vectors v_{T_i} and the state vectors y_{T_i} to

$$Y(t) = \begin{pmatrix} y_{T_1} \\ \vdots \\ y_{T_k} \end{pmatrix}, \quad V(t) = \begin{pmatrix} v_{T_1} \\ \vdots \\ v_{T_k} \end{pmatrix}.$$

By using the first-order necessary conditions, a bilevel problem with several followers can be reduced to the single level problem

$$\begin{aligned}
&\min_{x,u,Y,\Lambda} G(x(0), x(1), Y, V) + \int_0^1 F_0(x(t), u(t), Y, V)\, dt \quad (OCP_+) \\
&\text{subject to } \dot{x}(t) = F(x(t), u(t), Y, V), \text{ for all } t \in [0,1], \\
&\qquad \Psi(x(0), x(1), Y, V) = 0, \\
&\qquad C(x(t), u(t), Y, V) \leq 0, \qquad \text{for all } t \in [0,1], \\
&\qquad \dot{\Lambda}(\tau) = \mathcal{H}_Y, \\
&\qquad \mathcal{H}_V = 0, \\
&\qquad \Lambda_i(0) = 0, \qquad \text{if } Y_i(0) \text{ free}, \\
&\qquad \Lambda_j(1) = 0, \qquad \text{if } Y_j(1) \text{ free}, \\
&\qquad \mathcal{C}(x, u, Y(t), V(t)) \leq 0, \qquad \text{for all } t \in [0,1].
\end{aligned}$$

This optimal control problem – now in standard formulation – can be transformed by direct transcription methods (e.g. NUDOCCCS [1]) into a nonlinear programming problem, solvable with SQP methods.

4 Numerical Results

The model of a container crane, which was used for these numerical results is represented by the following system of differential equations:

$$\begin{aligned} \ddot{s} &= u_1 \\ \ddot{l} &= u_2 \\ \ddot{d} &= u_1 - \tfrac{g-u_2}{l} \cdot d \end{aligned} \tag{4}$$

The absolute position of the trolley in the rack is denoted by s, and the length of the wire rope by l. The relative position of the payload with respect to the trolley is d. The control variables u_1 and u_2 represent the motor accelerations for the trolley and the wire rope. g is Earth's gravity.

The differential equations (4) are used in the optimal control problem for container cranes (5), where the crane has to move from one position at rest (s_0, l_0) to another position at rest (s_f, l_f) minimizing process time and energy consumption:

$$\begin{aligned} \min_{u,t_f} I[u] &= t_f + \int_0^{t_f} \|u(t)\|_2^2 \, dt \\ \text{subject to } \ddot{s} = u_1, \quad \ddot{l} &= u_2, \quad \ddot{d} = u_1 - \tfrac{g-u_2}{l} \cdot d \\ C(\text{states}(t), u(t)) &\le 0, \quad t \in [0, t_f] \\ s(0) = s_0, \quad \dot{s}(0) &= 0, \\ l(0) = l_0, \quad \dot{l}(0) &= 0, \\ d(0) = 0, \quad \dot{d}(0) &= 0, \\ s(t_f) = s_f, \, \dot{s}(t_f) &= 0, \\ l(t_f) = l_f, \, \dot{l}(t_f) &= 0, \\ d(t_f) = 0, \, \dot{d}(t_f) &= 0. \end{aligned} \tag{5}$$

The problem at the lower level is to find an alternative trajectory from the current state at time t reaching an unknown position in fixed time τ_f without any oscillation remaining:

$$\begin{aligned} \min_{v_t} I_t[v_t] &= \int_0^{\tau_f} \|v_t(\tau)\|_2^2 \, d\tau \\ \text{subject to } \ddot{s}_t = v_{t,1}, \quad \ddot{l}_t &= v_{t,2}, \quad \ddot{d}_t = v_{t,1} - \tfrac{g-v_{t,2}}{l_t} \cdot d_t \\ c_t(\text{states}(\tau), v_t(\tau)) &\le 0, \quad \tau \in [0, \tau_{f_t}] \\ s_t(0) = s(t), \quad \dot{s}_t(0) &= \dot{s}(t), \\ l_t(0) = l(t), \quad \dot{l}_t(0) &= \dot{l}(t), \\ d_t(0) = d(t), \quad \dot{d}_t(0) &= \dot{d}(t), \\ s_t(\tau_{f_t}) = \text{free}, \, \dot{s}_t(\tau_{f_t}) &= 0, \\ l_t(\tau_{f_t}) = \text{free}, \, \dot{l}_t(\tau_{f_t}) &= 0, \\ d_t(\tau_{f_t}) = 0, \quad \dot{d}_t(\tau_{f_t}) &= 0. \end{aligned}$$

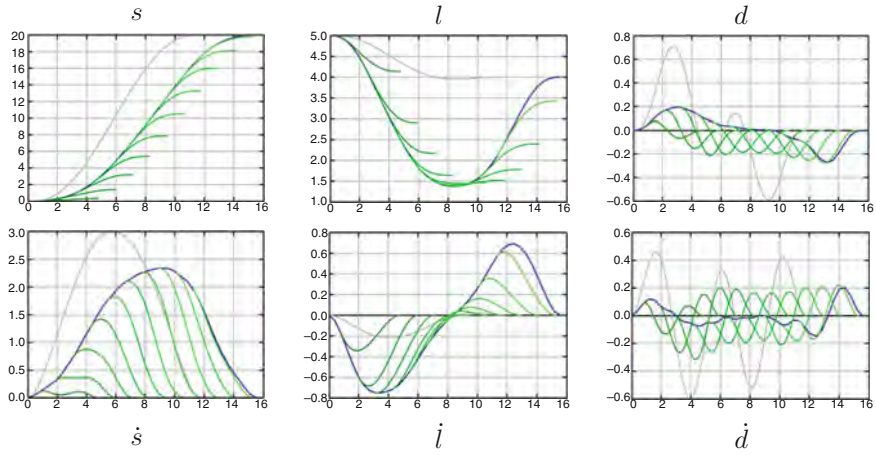


Fig. 2. Optimal solution of the main trajectory and ten alternative trajectories. The grey isolated line is the solution of the single-level problem

Figure 2 shows the optimal trajectory from the initial position $s_0 = 0$, $l_0 = 5$ to the terminal position $s_f = 20, l_f = 4$, considering ten safety stop trajectories, such that the system comes to be at rest within $4\,s$.

5 Conclusions

Bilevel optimal control problems combine optimal control theory and direct methods.

In our example, considering the lower-level problems, a reduction of the average amplitude of the main trajectory could be observed. For a fast calculation of stoppable trajectories, a simple criterion reducing the average amplitude can be used instead of the lower-level problem formulation.

References

1. Büskens, C.: Optimierungsmethoden und Sensitivitätsanalyse für optimale Steuerungsprozesse mit Steuer- und Zustandsbeschränkungen, PhD Thesis, Institut für Numerische Mathematik, Universität Münster (1998)

Optimization of Satellite Constellations

M. Knauer and C. Büskens

Center of Industrial Mathematics, Universität Bremen, Germany
knauer@math.uni-bremen.de, bueskens@math.uni-bremen.de

Summary. Constellations of satellites are used for a steady and efficient monitoring of selected regions of the Earth.

The orbit parameters for each satellite in a constellation define its flight path and the area covered by its swath. An optimal choice of the parameters could maximize the covered area, or find uniform coverages over time.

In order to use sequential quadratic programming methods to find the best constellation, a differentiable formulation of the coverage depending on the orbit parameters of each satellite has been developed. The areas monitored by the satellites' sensors are modelled as unions of convex polygons on a sphere.[1]

1 Satellite Mission Analysis

Satellites are used as a reliable source of data for global monitoring for environment and security. Depending on the types of sensors installed on a satellite, different kinds of data can be recorded.

In order to ensure the availability of recent data over a short period, constellations of more satellites are used which, once established, keep on their prearranged orbits for their whole lifetime. The control unit on the satellites is only available to compensate errors in the model or to avoid collisions with space debris.

A common task of satellite mission analysis is to find a constellation of satellites, such that a given target on Earth can be monitored at least once within a fixed repeating time frame with as few satellites as possible.

2 Optimization Problem

To find such a constellation, trajectories of the satellites have to be calculated. From orbital mechanics it is well known, that a satellite moves on an elliptical

[1] This paper is based on a cooperation project with OHB System AG, Bremen.

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*,
Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_147,

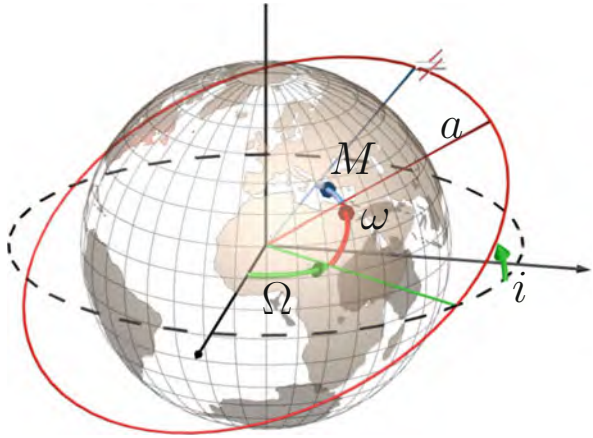


Fig. 1. Orbital elements: longitude of ascending node Ω, inclination i, argument of perigee ω, semimajor axis a and mean anomaly M. Eccentricity e is not shown

orbit where the Earth is at one focus. Its current position is generally noted by the six orbital elements (see Fig. 1).

Using Kepler's Second Law

$$\dot{M} = \text{const},$$

the flight path of a satellite around a point mass can be calculated. However, in order to consider the oblateness of Earth and its inhomogeneous geopotential a more complex system of differential equations has to be solved with a Runge–Kutta method. For the detailed system of the Klinkrad trajectory generator see [1].

In our model each satellite is equipped with left and/or right looking sensors with fixed opening angles. As the satellites follow their paths, the monitored region of the Earth's surface, the so-called swath, is continuously growing.

For satellite mission analysis, a constellation has to be determined such that the union of the swaths of all satellites after a given time is maximized.

Formulated as an sequential quadratic programming problem (SQP), this can be stated as

$$\begin{aligned} &\min_x \; -\text{Area}(x) \\ &\text{s.t. } g(x) \leq 0, \end{aligned}$$

where x is a vector of free orbit parameters of m satellites, and $g(x)$ denotes all constraints. These include e.g. box constraints for useful settings of allowed parameter intervals or constraints for coupling of satellites. The objective function $\text{Area}(x)$ for the SQP method, which evaluates the area covered by the constellation's satellites, has to be continuously differentiable.

A straightforward attempt for such an objective function would be to implement a discrete grid over the surface of the Earth and to count the

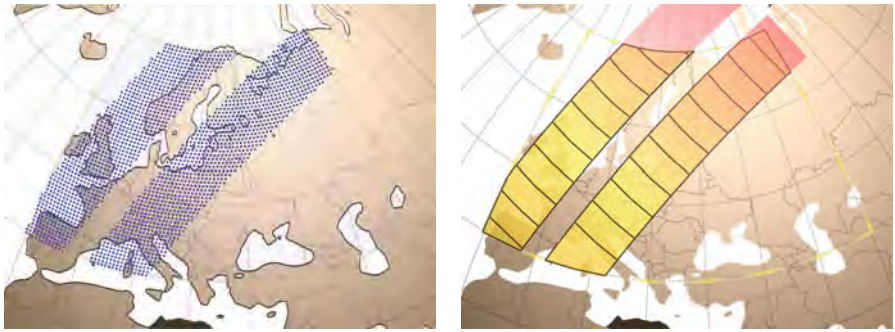


Fig. 2. Two methods for measuring the area. On the *left*, the approximative result is based on a discrete grid, on the *right* an exact result is found with the polygon method

grid points covered by the swaths. By increasing the number of grid points, a higher precision can be reached. The necessary differentiability properties with respect to the orbit parameters will not be fulfilled, unless a lot of memory is used.

Additionally it is not guaranteed that the region between adjacent covered grid points has really been covered (Fig. 2).

3 Polygon Coverage

In order to get an area function with the required differentiability properties, the area covered by the satellites' sensors is stored as a set of convex disjoint polygons.

As the swath increases, new polygons are added to the set, which reuse existing corners, and automatically remove overlapping parts.

In order to keep memory usage small, the set of polygons is simplified permanently, as unused corners are removed, and adjacent polygons are merged.

Every time the set of convex polygons is changed, a lot of corner points have to be tested for convexity. The following lemma provides an efficient method.

Lemma 1. *The corner* $C_i(x_i, y_i)$ *of a polygon is convex if and only if*

$$\begin{vmatrix} x_{i-1} & x_i & x_{i+1} \\ y_{i-1} & y_i & y_{i+1} \\ 1 & 1 & 1 \end{vmatrix} \geq 0,$$

where $C_{i-1}(x_{i-1}, y_{i-1})$ *and* $C_{i+1}(x_{i+1}, y_{i+1})$ *denote its previous and next corner in counterclockwise order.*

3.1 Polygon Coverage on Sphere

To apply the polygon coverage on a sphere S, or more generally to a manifold, an atlas $\mathcal{A} = \{(U_j, \varphi_j)\}$ has to be chosen, such that the domains of its charts (U_j, φ_j) cover S. A chart is defined as a homeomorphism

$$\varphi_j : U_j \in S \to V_j \in \mathbb{R}^n.$$

For a polygon on a plane, a border connecting two adjacent corners $C_i(x_i, y_i)$ and $C_{i-1}(x_{i-1}, y_{i-1})$ is the line segment $\overline{C_{i-1}C_i}$:

$$\overline{C_{i-1}C_i} = \left\{ \lambda \begin{pmatrix} x_i \\ y_i \end{pmatrix} + (1-\lambda) \begin{pmatrix} x_{i-1} \\ y_{i-1} \end{pmatrix} \mid \lambda \in [0,1] \right\}$$

For a polygon on a two-dimensional manifold, the boundary curve $B(C_{i-1}, C_i)$ connecting two corners $C_i, \tilde{C}_{i-1} \in S$ can be calculated as

$$B(C_{i-1}, C_i) = \left\{ \varphi_j^{-1}(\xi) \mid \xi \in \overline{\varphi_j(C_{i-1})\, \varphi_j(C_i)} \right\}$$

using the linear interpolation in the plane. The chart (U_j, φ_j) has to be selected, so that the line segment $\overline{\varphi_j(C_{i-1})\, \varphi_j(C_i)}$ is a subset of the codomain V_j.

For our application, an atlas of three charts is enough to construct reasonable polygons on a unit sphere. One chart is selected for each region around the two poles of the sphere, a third chart is provided for the remaining part of the surface.

The standard chart for the last case consists of the domain

$$U_3 = S \cap \{x \in \mathbb{R}^3 \mid |x_3| \leq \alpha\}$$

with a fixed parameter $\alpha \in (0,1)$, and the function $\varphi_3 : U_3 \to V_3$. As φ_3 is a homeomorphism, it can also be explained by its inverse

$$\varphi_3^{-1} : (\zeta, \nu) \mapsto \begin{pmatrix} \sin\nu\cos\zeta \\ \sin\nu\sin\nu \\ \cos\zeta \end{pmatrix}.$$

In the region of the sphere representing the area around the north pole, a chart can be constructed using polar coordinates as follows:

$$U_1 = S \cap \{x \in \mathbb{R}^3 \mid x_3 \geq \alpha\}$$

$$\varphi_1^{-1} : (\zeta, \nu) \mapsto \begin{pmatrix} \zeta \\ \nu \\ \sqrt{1-\zeta^2-\nu^2} \end{pmatrix}$$

A similar chart can be declared for the region around the south pole.

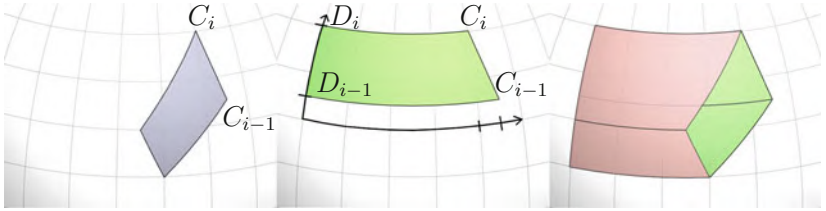


Fig. 3. For each border of the polygon (*left*) a trapezoid can be constructed (*middle*). The sum of all areas of the trapezoids is the area of the polygon (*right*)

3.2 Area of Polygon on Sphere

The set of polygons represents the region covered by the satellites' sensors. During the optimization process the area of large sets of polygons have to be calculated several times. Hence, a fast and efficient calculation of the area of a polygon is needed.

The surface integral $\mathrm{Area}(P)$ of a polygon P on the sphere with corners $C_i, i = 1, \ldots, n$ which is in the domain of the chart (U_3, φ_3) can be calculated as the sum of the surface $\mathrm{Area}(T_i)$ of trapezoids T_i with corner points

$$C_{i-1}, \quad C_i, \quad D_i = \varphi_3^{-1}((\varphi_3(C_i))_1, \tau), \quad D_{i-1} = \varphi_3^{-1}((\varphi_3(C_{i-1}))_1, \tau)$$

with any constant parameter τ, e.g. $\tau = 0$ (Fig. 3).

The surface integral on the sphere for each trapezoid T_i results in

$$\begin{aligned}\mathrm{Area}(T_i) &= \iint_{T_i} d\sigma = \iint_{\varphi_3(T_i)} \sin u \, du dv \\ &= \int_{\nu_{i-1}}^{\nu_i} \int_0^{\zeta_{i-1} + \frac{\zeta_i - \zeta_{i-1}}{\nu_i - \nu_{i-1}}(u - \nu_{i-1})} \sin u \, dv du,\end{aligned}$$

where $(\zeta_i, \nu_i) = \varphi_3(C_i)$ and $(\zeta_{i-1}, \nu_{i-1}) = \varphi_3(C_{i-1})$. Note that these tuples are simply the latitude and the longitude of the corner points on the sphere.

Here the value of $\mathrm{Area}(T_i)$ is positive if the direction of the polygon border which induces the trapezoid shows upwards, or negative otherwise.

The sum of the signed areas of these trapezoids is the area of the polygon on the sphere, which can be simplified to

$$\begin{aligned}\mathrm{Area}(P) &= \sum_{i=1}^{n} \mathrm{Area}(T_i) \\ &= \sum_{i=0}^{n} (\sin \nu_{i+1} - \sin \nu_i) \tfrac{\zeta_{i+1} - \zeta_i}{\nu_{i+1} - \nu_i}.\end{aligned}$$

4 Results

As a simplified numerical test example, the coverage of two satellites during one day should be optimized. Only the inclination angles of both satellites

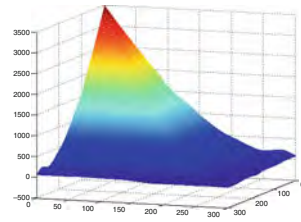


Fig. 4. Objective function for problem with two parameters

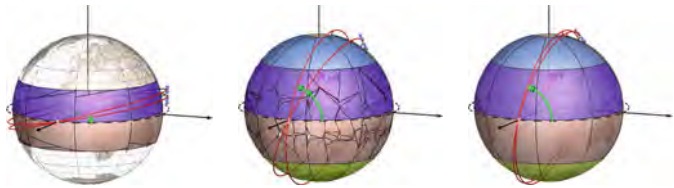


Fig. 5. Initial, intermediate and final state of optimization process

are used as free parameters. The objective function for this two-dimensional problem is shown in Fig. 4.

The initial values for the two optimization parameters are chosen so that the orbits of the satellites are close to the equator (Fig. 5). During the optimization process, the orbit of the satellites is rotated fast towards the poles so that – in this case – a complete coverage can be found.

5 Conclusions

For a proper formulation of the satellite optimization problem as a nonlinear optimization problem, we start with a parameter vector x defining a satellite constellation. Integration over time and projection of trajectory data results in the swath data. The polygon coverage module transforms this to a set of polygons. The objective function sums up the areas of all polygons and can be used with efficient SQP methods, as it is now differentiable with respect to the parameter vector x. The result is a satellite constellation with an optimal coverage.

References

1. Flury, W.: Raumfahrtmechanik, Lecture, Technische Universität Darmstadt (2002)

Moving Penalty Functions for Optimal Control with PDEs on Networks

O. Kolb[1], P. Domschke[1], and J. Lang[1,2,3]

[1] Department of Mathematics, Technische Universität Darmstadt, Dolivostr, 15, D-64293 Darmstadt, Germany, kolb@mathematik.tu-darmstadt.de
[2] Center of Smart Interfaces, Technische Universität Darmstadt, Petersenstr. 30, D-64287 Darmstadt, Germany, domschke@mathematikk.tu-darmstadt.de
[3] Graduate School Computational Engineering, Technische Universität Darmstadt, Dolivostr. 15, D-64293 Darmstadt, Germany, lang@mathematik.tu-darmstadt.de

Summary. An adaptive penalty technique to find feasible solutions of mixed integer nonlinear optimal control problems on networks is introduced. This new approach is applied to problems arising in the operation of gas and water supply networks.

1 Introduction

The operation of gas and water supply networks causes high costs. Likewise, the desire for cost-efficient control of those networks whereas all consumers' demands are satisfied is present. But the task of transient technical optimization results in complex mixed integer nonlinear problems. While the gas and water dynamics in the pipes of the network introduce partial differential equations (PDEs) as constraints, there are also combinatorial aspects concerning discrete processes like switching compressor stations and pumps on or off. Even finding a single feasible solution is a difficult task.

There are different approaches to solve such a kind of problems.[4] Our approach is based on a sequential quadratic programming (SQP) algorithm to solve nonlinear continuous optimization problems. Therefore, the discrete processes/variables have to be fixed, like in a branching algorithm, or relaxed. To enforce the relaxed variables to the feasible set, we introduce an adaptive penalty technique.

[4] A comparison of different approaches to the solution of optimal control problems for gas networks can be found in [1].

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*, Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_148,

2 Modelling

We model gas and water supply networks as directed graphs, where the edges of the graph correspond to the various components of the network, like pipes, compressor stations, pumps and valves. The vertices are inner/coupling or boundary nodes.

For both types of networks, the governing equations in the pipes are hyperbolic PDEs. In the gas networks, we apply the isothermal Euler equations supplemented by a suitable equation of state:

$$\partial_t \rho + \partial_x(\rho v) = 0 \tag{1}$$

$$\partial_t(\rho v) + \partial_x(\rho v^2) + \partial_x p = -g\rho\partial_x h - \frac{\lambda}{2D}\rho|v|v \tag{2}$$

$$\rho = \frac{p}{z(p,\bar{T})R_0\bar{T}}\,. \tag{3}$$

For the water dynamics, we apply the so-called water hammer equations:

$$\frac{gA}{a^2}\frac{\partial h}{\partial t} + \frac{\partial q}{\partial x} = 0 \tag{4}$$

$$\frac{\partial q}{\partial t} + gA\frac{\partial h}{\partial x} = -f\frac{q|q|}{2DA}\,. \tag{5}$$

Gas and water supply networks are operated in the subsonic flow region, that is, the velocity of the gas/water is smaller than the speed of sound in the considered medium. In this case, implicit box schemes are known to work very effectively. They are conservative and usually stable under mild conditions. Therefore, we apply a spatially symmetric implicit box scheme for the discretization of the isothermal Euler equations and the water hammer equations. For a general balance law of the form $u_t + f(u)_x = g(u)$, our discretization scheme reads as follows:[5]

$$\frac{u_{j-1}^{n+1} + u_j^{n+1}}{2} = \frac{u_{j-1}^{n} + u_j^{n}}{2} - \frac{\Delta t}{\Delta x}\left(f_j^{n+1} - f_{j-1}^{n+1}\right) + \Delta t\frac{g_{j-1}^{n+1} + g_j^{n+1}}{2}\,. \tag{6}$$

In addition to the gas/water dynamics inside the pipes, our main focus lies on the controllable elements, particularly compressor stations in the gas networks and pumps in the water supply networks. Both have in common that the feasible domain of the associated control variables consists of two disjoint sets. A compressor station can be switched off (the control variable equals zero), and when it is switched on/active, it has to run with a minimum power and below the technical maximum. Thus, the feasible domain is of the form

$$\{0\} \cup [minCtrl, maxCtrl] \tag{7}$$

with $minCtrl > 0$. Water pumps operate in a similar way. For pumps with fixed speed, the control variable is even binary since *minCtrl* equals *maxCtrl*.

[5] A stability and convergence analysis of our scheme can be found in [3].

Thus, in the gas network setting as well as the water supply networks, we are dealing with mixed integer optimal control problems. Here, the objective function typically consists of the costs caused by the controllable elements. These are the accumulated fuel gas consumption of the compressor stations and the power consumption of the pumps.

Besides the pipes and the controllable elements, there is a multitude of other components in gas and water supply networks, which are modelled by algebraic or ordinary differential equations.[6] For the numerical solution, these are discretized at the same time steps as the PDEs and the control variables. Altogether, this results in a coupled system of nonlinear equations, which we solve with an adapted version of Newton's method. Especially, we take advantage of the sparse structure of the Jacobian matrix of the underlying equations by using a particular solver [2].

3 Moving Penalty Functions

As we have seen in the previous section, the optimal control task for gas and water supply networks consists of discrete and continuous control decisions. The combination of discrete decision variables with a highly nonlinear PDE-constrained continuous task makes the entire problem very difficult to solve. The presented approach begins with the relaxation of the mixed integer problem: The feasible domain (7) of each control variable is expanded to $[0, maxCtrl]$. Here, we have to take care of a consistent extension of the equations describing the controllable elements. Afterwards, we solve the relaxed continuous optimal control task with a state-of-the-art SQP method [5].

Usually, the solution of the relaxed problem is not feasible for the original problem. Here, we apply our approach of "moving penalty functions" (MPF). The basic idea of MPF is to add a variable penalty term to the cost function of each relaxed switchable element, that is, each controllable element where switching decisions have to be made and the corresponding control variables have been relaxed for the optimization process. A prototype of our penalty functions, which are introduced for each binary variable, is plotted in Fig. 1. In the course of the optimization process, the position and the height of the peak are varied, giving this technique its name.

The basic algorithm from the view of a single penalty term is the following: When the penalty function is initialized, the position x_m of the peak of the penalty function is set to x_0 and the maximum value y_m to y_0. After each run of the optimization tool, we check whether the relaxed control variable x is within a so-called fixing region, that is, $x \leq x_{off}$ or $x \geq x_{on}$. In this case, the switching decision is fixed. Otherwise, we increase the penalty term and move the position of the peak in the direction of the current control (plus

[6] A detailed description of the modelling of the main components of a gas supply network can be found in [1] and [4].

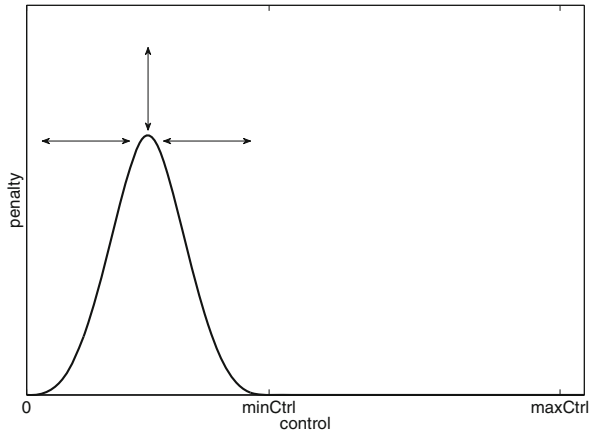


Fig. 1. Plot of a penalty function

β_r or minus β_l).[7] Then, the optimization tool is run again until all switching decisions are fixed or a maximum number of iterations is reached.

Although our basic algorithm already yields useful results, there are lots of challenges depending on the given task. One very important aspect is the choice of the parameters. Just as an example, consider the parameters x_{off} and x_{on}. On the one hand, we would like to have large regions where the binary decisions get fixed as fast as possible. But on the other hand, fixing the wrong variables too early might lead to an infeasible remaining task. As an improvement of our basic algorithm, we tackle the latter problem by introducing some kind of active set strategy, that is, if we cannot find a feasible solution of the remaining (relaxed) problem, we release some previously fixed binary variables which have influence on the violated constraints.

Another improvement of our basic algorithm deals with the parameters β_l and β_r for the moving of the position of the penalty peak. Small values for β_l or β_r can result in lots of iterations of the algorithm. On the other hand, large parameters for the moving can result in jumping around the current solution without the penalty functions being able to affect the control in neither direction. Such situations typically occur if two or more control variables have to be influenced in opposite directions to get a feasible solution of the original (not relaxed) problem. Our strategy is to resize the moving parameters β_l and β_r after a specified number of steps into the same direction (x_m is always increased or decreased) or when x_m jumps from one side of x to the other and back for a certain number of times. In the first case, either β_l or β_r is increased and in the second case, β_l and β_r are decreased.

[7] It is possible and also intended that x_m overtakes x.

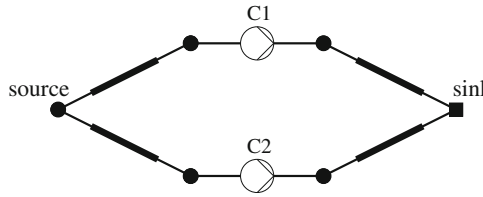


Fig. 2. Example gas network

Table 1. Optimization results for the gas network example

	t_0	t_1	t_2	t_3	t_4
Power of compressor C1	0	0	0	500	891
Power of compressor C2	0	0	532	733	1154

4 Results

4.1 Gas Network

Figure 2 shows one of our example gas networks. It consists of four pipes and two compressor stations. While the pressure is constant at the source node, there is an increasing flow demand at the sink. This causes higher friction losses, which the two compressor stations have to compensate for due to a minimum pressure constraint at the sink. The optimization horizon is four hours and it is equally discretized into four parts.

For both compressor stations, we have $minCtrl = 500$ and $maxCtrl = 1500$, but compressor C2 has a higher efficiency. Therefore, we expect the second compressor to do most of the necessary work. Table 1 shows the results of the optimization process.

4.2 Water Network

Figure 3 shows a water supply network with 20 pipes, three pumps, ten surge vessels and two water tanks. The four consumers on the right hand side have varying demands and obtain water from the two tanks, which are located on a hill. The (identical) fixed speed pumps on the other side of the hill are responsible for keeping certain stage constraints at the water tanks.

Here, the optimization horizon is 12 h and it is equally discretized into 24 parts. Table 2 shows the results of the optimization process. The bullets indicate time steps in which the pump is switched on, the dashes indicate switched off pumps, respectively.

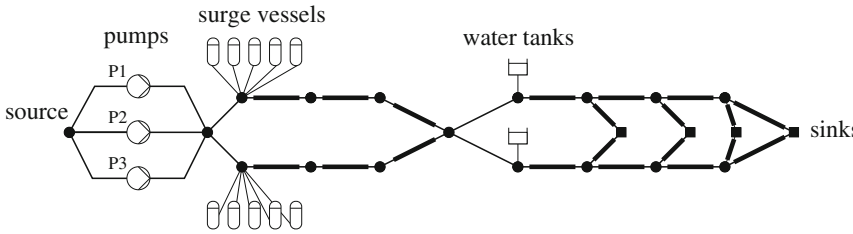


Fig. 3. Example water network

Table 2. Optimization results for the water network example

	t_0 until t_{24}
Pump P1	– – – ● ● – ● ● ● – – ● ● – ● ● ● – ● ● – ● ● – ●
Pump P2	– – – – ● ● ● – – ● – ● – ● ● – ● ● – ● ● – ● – ●
Pump P3	– – – – – – – – – – – ● ● – – ● – ● – – ● – ● – ● -

5 Conclusions and Further Work

We have introduced a new approach to find feasible solutions of mixed integer nonlinear optimal control problems on networks. First, we have relaxed the mixed integer problem and solved the resulting continuous optimization task with an SQP method. The feasibility of the obtained solution is enforced by adding variable penalty terms in further runs of the optimization tool. Here, we fix control variables which are within a so-called fixing region around the feasible discrete values. Additionally, previously fixed variables can be released again if necessary and the parameters of the applied MPF are adjusted in the course of the optimization process. Finally, we tested our new algorithm on problems arising in the operation of gas and water supply networks.

The next steps of our work will be the analysis of further strategies for the solution of mixed integer nonlinear optimal control problems on networks and the integration of the presented algorithm as heuristic in a branch and bound framework.

Acknowledgement

This work was supported by the DFG (grant LA1372/5-1) and the BMBF (grant 03MAPAK1). Part of the data was provided by the Siemens AG.

References

1. Bales, P., Geißler, B., Kolb, O., Lang, J., Martin, A., Morsi, A.: Dep. Math., Technische Universität Darmstadt, preprint 2552 (2008)
2. Davis, T.A.: Direct Methods for Sparse Linear Systems. SIAM, Philadelphia (2006)

3. Kolb, O., Lang, J., Bales, P.: An implicit box scheme for subsonic compressible flow with dissipative source term, Numer. Algorith. **53**(2), 293–307 (2010). doi:10.1007/s11075-009-9287-y
4. Moritz, S.: A Mixed Integer Approach for the Transient Case of Gas Network Optimization, PhD thesis, Technische Universität Darmstadt (2006)
5. Spellucci, P.: donlp2 users guide, Technische Universität Darmstadt

Numerical Analysis of Geometrical Characteristics of Machine Elements Obtained with CMM Scanning

P. Krawiec

Chair of Basic of Machine Design, Poznan University of Technology, Poznan, Poland, piotr.krawiec@put.poznan.pl

Summary. In industrial practice it is often necessary to obtain information regarding the geometrical characteristics of machine elements whose documentation does not exist or is unavailable. One way to obtain such information is by applying reverse engineering methods. This paper presents the process of digitization on a numerically controlled coordinate measuring machine CMM. An important problem is performing a correct numerical analysis of points obtained as a result of scanning of profiles having small radii of curvature, to be used for drawing parametric curves. Interpolation methods are necessary to obtain correct information about geometrical characteristics of machine parts.

1 Introduction

Product formation always is preceded by its design. For obtaining of correct feedback between designer and process engineer there is necessary the unequivocal and readable record of a design that is information about material, geometrical and strength features. For that reason there is necessary using of rules of design classical recording and creating of three-dimensional model. However sometimes product design form is difficult or impossible to unequivocal characterization. There exists some products of which geometrical features are not commonly and fully described. Such products are elements created with usage of free surfaces, hybrid models and others. Sometimes lack of full information about geometrical features may preclude the widening of products usage spectrum. The example of that may be semicircular and involute profiles of belt pulleys. In order to use these profiles for making non-typical (out-of-round) wheels there is necessary determining of full geometrical features characteristics. The possible solution of that task is making models with usage of Reverse Engineering methods. It involves usage of coordinate measuring machines CMM or so called digitizers (with contact and contact-less). Obtained data, as a result of measurements, most often in form of points or

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*, Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_149,

triangles network are always burdened with errors. Values of obtained differences determine the way of designing, manufacturing and final inspection of the product.

2 Measurements on Coordinate Measuring Machine

One of the measurement methods of machine elements is scanning on Coordinate Measuring Machines (Fig. 1). During this process there are recorded the sequential positions of centre of contact tip which is moving along the profile being measured. Recorded coordinates describe the geometrical features of elements which are not on actual profile but are on equidistant curve. The distance between measured curve and actual profile is equal to radius of spherical gauging point. Professional CMM are equipped with software for automatic exchange of measured curve into curve of actual profile. The significant problem is correctness of execution of this transformation. It is recommended to elaborate the smooth procedures of object profile. One of the methods can be the application of spline curve or approximation to close tangent arcs (circles) (Fig. 2).

Fig. 1. Measurements with CMM

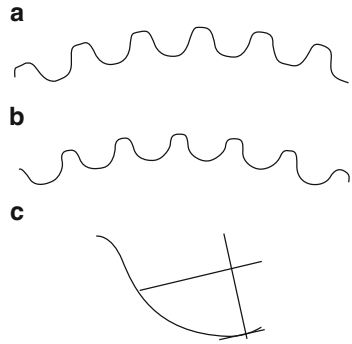


Fig. 2. Tooth profile obtained as a result of digitization: (**a**) involute profile, (**b**) semicircular profile, (**c**) enlarges profile of semicircular tooth

In presented example there were digitalized the semicircle and involute profiles of belt pulleys. After determination of equidistant curve the designed tooth profiles will be spaced in a proper way on wheels with non-circular profile of wheel rim. For curve generation on the basis of measurement point there was used Newton's interpolation formula of n-th order. In Cartesian system it takes the form:

$$W_n(x) = f(x_0) + f(x_0, x_1)\omega_0(x) + \\ + f(x_0, x_1, x_2)\omega_1(x) + \ldots + f(x_0, x_1, \ldots, x_n)\omega_{n-1}(x), \quad (1)$$

where: the divided difference

$$f(x_i, x_{i+1}, \ldots, x_{i+n}) = \frac{f(x_{i+1}, x_{i+2}, \ldots, x_{i+n}) - f(x_i, x_{i+1}, \ldots, x_{i+n-1})}{x_{i+n} - x_i} \quad (2)$$

for $n = 1, 2, \ldots$ and $i = 0, 1, 2, \ldots$

$$\omega_k(x) = (x - x_0)(x - x_1)\ldots(x - x_{k-1}), \quad (3)$$

where $k = 0, 1, \ldots, n-1$. Reproducibility error can be determined on the basis of

$$\varepsilon_i = \frac{\Delta^3(y_{i-1})}{3!(x_i - x_{i-1})^3}(x - x_{i-1})(x - x_i)(x - x_{i+1}) \quad (4)$$

where $(x_i - x_{i-1})$ – increment of independent variable, $\Delta^3(y_{i-1})$ – progressive difference of third order in point (x_{i-1}, y_{i-1}).

From workshop practice of manufacturing of belt pulleys teeth there results that their profiles should be described not with points' coordinates, but with curves segments (arc, involute, etc.). Therefore there is necessary determining of proper curves keeping tangency conditions. For that there is possible to use osculating curve (circle) definition and plane curve curvature.

For plane curve determined in parametric form with equations $x = \varphi(t)$, $y = \psi(t)$. Coordinates x_C, y_C of their centre of curvation C and curvature k and radius R of curve K curvature in point $M_1(x_1, y_1)$ may be determined with the following formulas [4]

$$x_C = x_1 - \frac{({x_1}'^2 + {y_1}'^2) \cdot {y_1}'}{|{x_1}'{y_1}'' - {x_1}''{y_1}'|}, \quad y_C = y_1 + \frac{({x_1}'^2 + {y_1}'^2) \cdot {x_1}'}{|{x_1}'{y_1}'' - {x_1}''{y_1}'|}, \quad (5)$$

$$k = \frac{|{x_1}'{y_1}'' - {x_1}''{y_1}'|}{({x_1}'^2 + {y_1}'^2)^{3/2}}, \quad R = \frac{1}{k} = \frac{({x_1}'^2 + {y_1}'^2)^{3/2}}{|{x_1}'{y_1}'' - {x_1}''{y_1}'|}, \quad (6)$$

where: $x_1 = \varphi(t_1)$, $y_1 = \psi(t_1)$, ${x_1}' = \varphi'(t_1)$, ${y_1}' = \psi'(t_1)$, ${x_1}'' = \varphi''(t_1)$, ${y_1}'' = \psi''(t_1)$

After determination of nominal profile there is necessary determination of real profile from teeth. It can be carried out using offset curves – parallel to curves. Offset curves profile for ellipse and noncircular wheel are shown in Fig. 3.

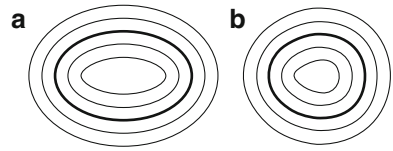


Fig. 3. Offset curves (**a**) for ellipse, (**b**) for noncircular wheel

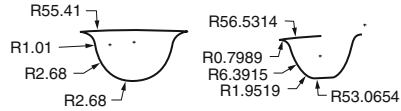


Fig. 4. Determined geometrical shape of tooth space

An offset curve is the set of all points that lie in a perpendicular distance d from a given curve in C^2. The scalar ρ is called the offset radius. If the parametric equation of the given curve is

$$P(t) = \big(x(t), y(t)\big). \tag{7}$$

Then the offset curve with offset radius ρ is given by formula [2,3]

$$\Omega(\rho, P(t)) = P(t) + \rho \frac{(y'(t)) - x'(t)}{\sqrt{x'^2(t) + y'^2(t)}}. \tag{8}$$

If P is an offset of Q , the reverse is generally true, as long as the offset radius is everywhere less than the radius of curvature of both curve segments:

$$\Omega(-\rho, \Omega(\rho, P(t))) = P(t). \tag{9}$$

In general, offset curves cannot be represented in Bezier form because contains a square root of a polynomial. The obvious exceptions are circles and straight lines. A non-obvious exception is that the offset of any parabola can be represented as a degree eight rational Bezier curve [1].

Demonstration semicircular and involute profile shapes (simplified to arc segments) presented in Fig. 4.

3 Manufacturing of Wheels with Rapid Prototyping Methods

Since introducing stereolithography in 1987 as the first method of Rapid Prototyping there was elaborated many others methods but the most often used is the principle of incremental laminar creating of model. Referring to their universality, in technique, there are commonly used the following methods: SLA (stereolithography), SGC (Solid Ground Curing), SLS (Selective Laser

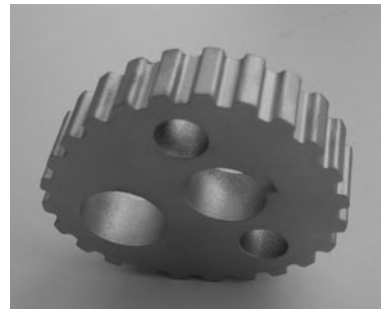


Fig. 5. Demonstration tooth profiles and non-circular wheel manufactured by Selective Laser Sintering SLS method

Sintering), LOM (Laminated Object Manufacturing), 3D Painting and others. For manufacturing of wheel presented in Fig. 5 there was used method SLS referring to that there is no necessity to use supporting and fixing elements and referring to that manufactured with this method products are full valuable elements of structure referring to strength of materials. Next very important advantage of this method is high accuracy of reproduction of given model CAD.

4 Verification of Elaborated Profiles on Measuring Stand

After wheels group manufacturing there was necessary the experimental verification of correctness of taken tooth profile shape, used interpolation method of coordinates describing wheel profile and taken manufacturing method. For this task there was used measuring machine of company Zeiss with diamond gauging point of diameter 1.8 mm (Fig. 6a). In measuring process there was put model CAD in format .sat to measuring machine's software (Fig. 6b), and then there was carried out measurement of profile envelope in three parallel planes (Fig. 6c). Results of measurements shows that maximal differences between measured values in reference to model CAD geometrical features are 0.01 mm. From this results that wheels manufactured with this method may be successfully used in uneven-running transmissions.

5 Summary

In this elaboration presented process of machine elements geometrical characteristics numerical analysis obtained with CMM scanning. There shown basic mathematical relationships used for interpolation of data obtained from measurements carried out on coordinate machine. And there was given the procedure of getting offset curves for obtaining wheels real profile. There were

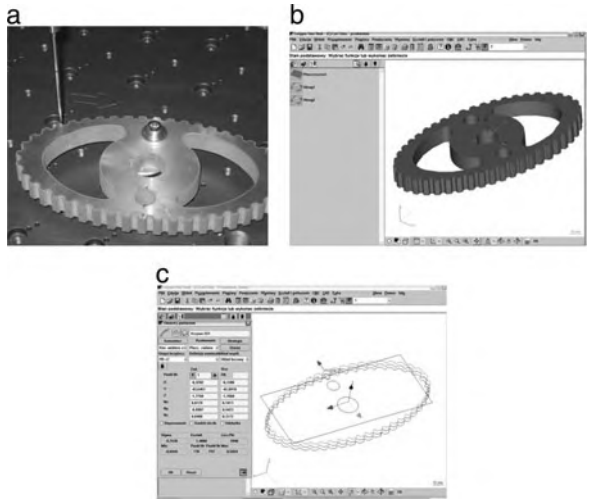


Fig. 6. Numerical analysis of manufactured non-circular wheels (**a**) measurement of wheel profile, (**b**) model CAD in measuring software, (**c**) measured profiles of wheel

included methods of noncircular wheels manufacturing with usage of systems CAD and CAD/CAM. In the end of this work there illustrated process of wheels geometrical features verification with usage of professional equipment and software in accordance with measuring method elaborated by author.

References

1. Bezier, P.: Numerical Control: Mathematics and Applications. Wiley, London (1972)
2. Klass, R.: Comput. Aided Des. **20**, 471–474 (1988)
3. Pham, B.: Comput. Aided Des. **15**, 297–299 (1983)
4. Niczyporowicz, E.: Krzywe plaskie. Wybrane zagadnienia z geometrii analitycznej i rozniczkowej. PWN, Warszawa (1991)

Plastic Yield of Particulate Materials Under the Effect of Temperature

I. Malujda

Poznan University of Technology, Chair of Machine Design Fundamentals, Poland, 60-965 Poznan, Piotrowo 3, ireneusz.malujda@put.poznan.pl

Summary. The parameter of primary importance for modelling the compression of particulate materials, in particular sawdust, is the critical stress which initiates plastic flow of the material. This value depends on the thermo-mechanical parameters of the material and the fundamental parameters of the process, namely compression pressure in the chamber, forming time and temperature. Compression of sawdust involves increase of temperature, which significantly reduces the compressive strength of the material. The focus of this paper is on the distribution of temperature in the superficial layer of briquette. It will enable to allow for the effect of temperature on the yield point, which is the strength criterion and the main element of a mathematical model describing the process of briquetting.

1 Strength Criterion in Compression of Particulate Materials

In the process described in this paper briquettes are formed in an open chamber in binderless process (Fig. 1). Cohesion and sufficient consolidation depends on the plastic flow state, which in the thin outer layer (crust) depends on heat. The work of friction related to pushing of sawdust through the main chamber and through the forming sleeve (Fig. 1) results in an increase of temperature in the outer layer of briquette. The highest temperature is noted at the forming sleeve surface and it drops quickly towards the centre of briquette. At depths at which the temperature has risen to ca. 130°C plasticisation of wood lignin occurs. On the briquette surface a crust is formed, which after cooling to ambient temperature provides a uniform and smooth consolidating structure.

Plastic flow in a thin layer of wood was analysed on the basis of the hypothesis of limit energy of shape deformation formulated by Huber-Mises-Hencky [2, 4, 5, 9].

A solution of this problem has been presented in the referenced publications [1, 2, 8, 9]. The theory of plasticity has been applied, specifically its

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*,
Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_150,

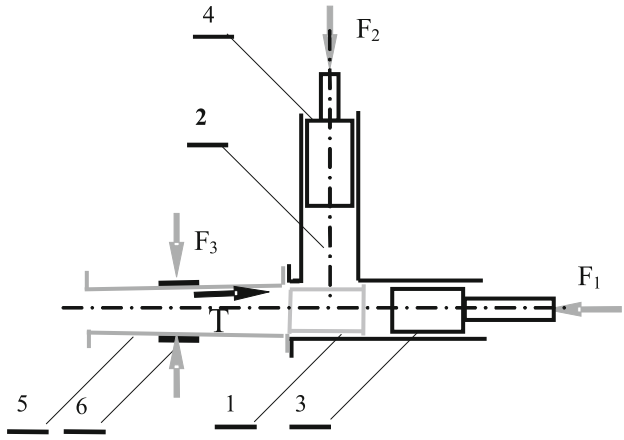


Fig. 1. Kinematic diagram of the pressing unit: 1 – main chamber, 2 – initial densification chamber, 3 – main piston, 4 – initial pressing piston, 5 – forming sleeve, 6 – pressing resistance adjustment control, F_1 – main pressing force, F_2 – initial pressing force, F_3 – pressing resistance adjustment force, T – friction force

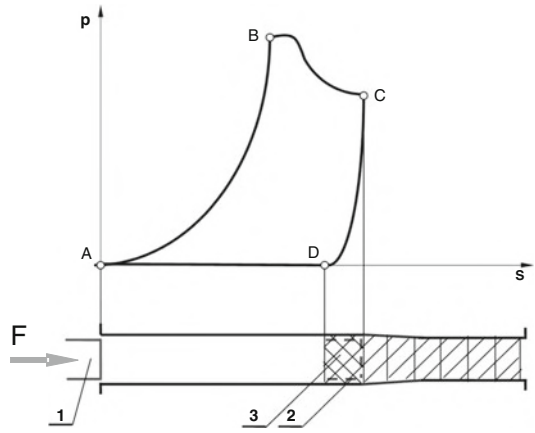


Fig. 2. Change of briquetting pressure p as a function of piston travel s during uniaxial compression of briquette in an extruder barrel with adjustable taper, where: 1 – start of piston travel, 2 – end of piston travel, 3 – pressure relief phase (withdrawal of piston) resulting in undesirable decompression of briquette. Curve BC illustrates pushing of briquette with characteristic drop of pressure due to friction-generated temperature

limit theorems. The yield point defining the critical stress has been determined empirically from the compressive strength test as the strength criterion for the mathematical model describing the process under analysis [1, 7, 8]. The intent here was to determine the force defining the stress causing plastic flow of the compressed material. It was assumed that area ABCD defining the actual pressing work (Fig. 2) can be replaced with an equivalent rectangular work area.

One of its sides shall have the length equal to $\sqrt{3}k_T$ corresponding to the equivalent yield point of an ideally plastic body and the second largest deformation ε_k was determined from the actual material curve. k_T stands for the average stress of plastic flow – depending on the influence of heat – equivalent to the yield point on shearing of the material in briquettes.

Subsequently the briquette pressing work determined from the actual diagram was compared with the work defined by the area of the rectangle [1, 6]:

$$\sqrt{3}k_T\varepsilon_k = \int_0^{\varepsilon_k} \sigma\varepsilon d\varepsilon, \tag{1}$$

where: ε_k – final deformation determined from the actual material curve, σ – temperature dependent yield point, determined from the actual material curve during compression. This relation was used to determine the value of k_T the average plastic flow stress, which was equivalent to the yield point of the compressed material.

Finally, we obtain the following equation for calculating the value of critical pressing force P in the conical main chamber (Fig. 3) and in the forming sleeve:

$$P = 3\pi k_T R_1^2 \left[\sqrt{3}\ln\frac{R_1}{R_2} + \frac{2}{3}\left(f_1\frac{\sqrt{1+(\operatorname{tg}\alpha)^2}}{\cos\alpha}\ln\frac{R_1}{R_2} + f_2\frac{l}{R_2}\right)\right]. \tag{2}$$

The value of critical stress k_T drops with the increase of temperature. As a direct measurement of temperature, especially during operation of the

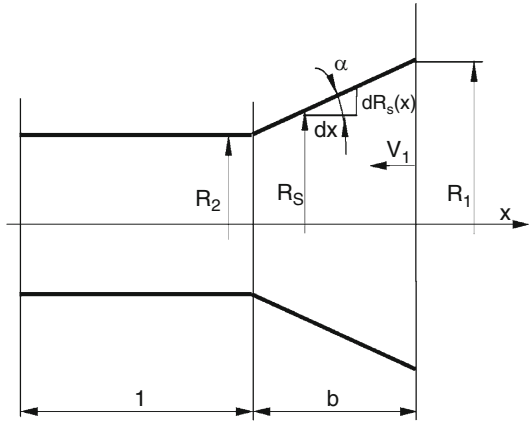


Fig. 3. Geometrical characteristics of the pressing channel, where: R_1 – radius of the opening on the entry into the main chamber, R_2 – radius of the opening on the exit from the main chamber, R_3 – radius as a function of the main chamber's length, l – length of the forming sleeve, α – half of main chamber's taper angle, b – length of the main chamber's

machine, is very difficult, an attempt was made to determine the distribution of temperature as a function of time in a layer of briquette by calculations.

2 Mathematical Model of Heat Transfer

The temperature has been the key parameter in formulation of the constitutive equations describing the analysed process. There are various ways in which heat penetrates inside a body. Here, we will briefly describe two of them, which are relevant to the process under analysis due to the actual transfer of heat between the hot wall of the forming sleeve and closely abutting side of briquette. Unsteady heat conduction (which is the case here) is described by the second Fourier's law [3]:

$$\frac{\partial T}{\partial t} = \alpha \nabla^2 T, \qquad a = \frac{\lambda}{c_p \gamma}, \tag{3}$$

where: a – thermal diffusivity, T – temperature, t – time,γ – specific gravity, c_p – specific heat, λ – thermal conductivity.

It is highly relevant to the process under analysis to consider convective transfer of heat, described by the following equation:

$$Q = \alpha(T_p - T_0), \qquad B_i = \frac{\alpha d}{\lambda} < 1, \tag{4}$$

where: T_p – surface temperature, T_0 – temperature of the boundary layer, α – heat transfer coefficient, B_i – Biot number, d – layer thickness.

Convection is more appropriate than conduction in describing the heat transfer when the Biot number is less than one (4), and for $B_i > 1$ heat is transferred by conduction. Transfer of heat by unsteady conduction is the closest approximation of the actual transfer of heat between the hot wall of the forming sleeve and the briquette. The hot wall of the forming sleeve is in contact with the surface of the formed briquette, and thus we can assume that the heat transfer coefficient α tends to infinity, and consequently the Biot number (4) is greater than 1. Thus, the criterion for conductive heat transfer is met. Therefore, the analysis of the constitutive relations describing the heat transfer between the wall of forming sleeve and the briquette will be related to conduction only.

Subsequent transformations yield the following differential equation

$$\frac{\partial T}{\partial t} = \alpha \left(\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} + \frac{\partial^2 T}{\partial z^2} \right). \tag{5}$$

The above term is known as Fourier's law describing the change of temperature in time and as a function of temperature gradient variation in 3D space. With correct boundary and initial conditions it allows for determination of time dependent temperature at any point throughout the analysed layer of

material. The transformed (5) draws our attention to the importance of the temperature conduction coefficient α (and, specifically, to the three parameters describing it, namely: c_p,γ, λ) for accuracy of the result. For solving the problem of heat conduction in the layer of briquette the approximate solutions method was used. Following the spatial digitisation of (5) selection of appropriate initial and boundary conditions and approximation of the temperature field with the finite element shape function a system of simple differential equations was obtained as a function of node temperatures and their time derivatives. The system of equations may be expressed with the following differential equation:

$$[C]\frac{d}{dt}\{T\} + [K]\{T\} = \{F\}, \tag{6}$$

where: $[C]$ – heat capacity matrix, $[K]$ – conductivity matrix, $\{F\}$ – thermal force vector, $\{T\}$ – nodal temperature vector.

3 Numerical Analysis of Heat Conduction in a Layer of Briquette

I-DEAS software program has been used to determine the distribution of temperature in the layer of briquette. The input data were the geometrical characteristics, process parameters and the sawdust properties, as used on the briquetting machine actually operating in a furniture factory. Dirichlet boundary condition has been adopted, assuming known briquette surface temperature and Cauchy initial conditions, i.e., at the time $t = 0$ the sample temperature is equal to the ambient temperature. For more accurate solution of the temperature distribution in the briquette crust the grid density has been increased at the point of contact between the hot forming sleeve and the surface of plasticized material. As the laminar representation of the spatial distribution of temperature does not allow for qualitative evaluation of the calculation results, they have been presented as curves (Fig. 4b), related to the flat cross-section of a briquette, representing the increase of temperature as a function of time in the subsequent layers (1,2,3,4) in the direction inwards the briquette.

The calculated temperature distribution inside the briquette shows that for the adopted input data the temperature increases in the respective layers to ca. 130°C (0–0.5 mm depth) and ca. 110°C (2 mm depth). We can assume that to ca. 1 mm depth the layer of sawdust becomes plasticized and at the same time overheated to the plastic flow point of lignin. Briquetting of sawdust requires its densification thorough and plasticization in the thin superficial layer under the effect of temperature.

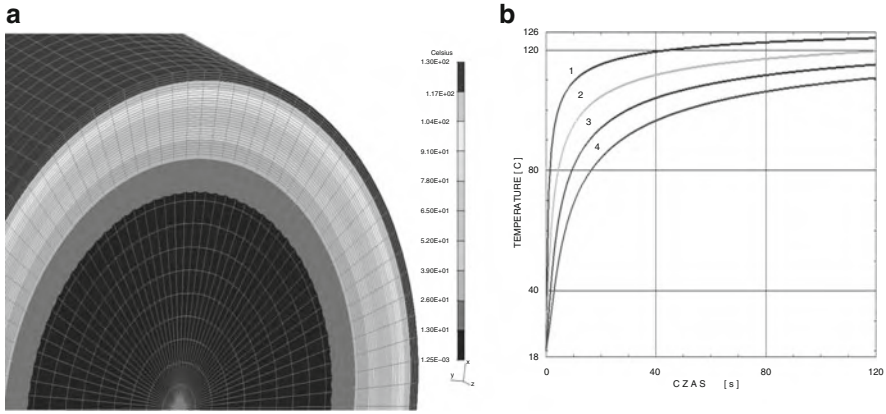


Fig. 4. Calculation result: (**a**) A fragment of grid showing temperature layers, (**b**) Curves representing the change of temperature as a function of time and depth of the respective layers, where the temperatures at the respective depths are: 1 – 0.5 mm, 2 – 1 mm, 3 – 1.5 mm, 4 – 2 mm

References

1. Dudziak, M., Malujda, I., Meler, I.: Zeszyty Naukowe Politechniki Poznanskiej, No.37, pp. 180–192. (1992)
2. Hill, R.: Clarendon, Oxford (1956)
3. Hobler, T.: Ruch ciepla i wymienniki. WNT, W-wa (1968)
4. Malczewski, J.: Oficyna Wydawnicza Politechniki Warszawskiej, Warszawa (1994)
5. Malujda, I.: Mielniczuk, J.: Mater. Eng. **28**(4), 35–40 (2004)
6. Malujda, I.: Mach. Dyn. Probl. **30**(4), 48–59 (2006)
7. Piwnik, J.: Rozprawy Inzynierskie **32**, 2 (1987)
8. Piwnik, J., Obrobka Plastyczna, T. XXV, z.2 (1986)
9. Walczak, J., PWN, Warszawa–Krakow (1978)

A Model for Spray Droplet Adhesion, Bounce or Shatter at a Crop Leaf Surface

Geoffry N. Mercer[1], Winston L. Sweatman[2], and W. Alison Forster[3]

[1] University of New South Wales at ADFA, Canberra, Australia and Australian National University, Canberra, Australia Geoff.Mercer@anu.edu.au
[2] Institute of Information and Mathematical Sciences, Massey University, Auckland, New Zealand w.sweatman@massey.ac.nz
[3] Plant Protection Chemistry NZ Ltd., Rotorua, New Zealand Alison.Forster@ppcnz.co.nz

Summary. Improvements in crop spray methods can result in great environmental and cost benefits. By developing a greater understanding of the physical processes involved, it should be possible to tailor spray formulations to maximise retention by plant foliage. This would enable the reduction of the chemical active required to achieve agrochemical efficacy. In the present paper one important aspect of the retention process is considered: a droplet-leaf impaction model is presented allowing for bounce, shatter or adhesion of the droplets by the leaf surface.

1 Introduction

During the process of spraying crops, there is usually some off-target loss of the spray. This can occur due to processes such as drift on the wind, droplet evaporation or the spray passing through the leaf canopy to reach the ground in the sprayed area. The spray droplet can also bounce from the plant surface after impaction. The off-target component due to these processes can be greater than the actual crop retention. This can be costly, both environmentally and financially. It is necessary to ensure that the correct quantity of the spray formulation reaches the plants to achieve the intended purpose.

This paper does not consider the probability that spray droplets hit or miss the plant, but rather considers the impaction process when droplets do hit the plant. Adhesion is the "stickability" of droplets on initial impact. Retention is the overall capture by plant surfaces of spray droplets either on initial or subsequent impact, and after loss due to run-off. Although adhesion may be low, retention may be much higher due to re-capture of bouncing or shattered droplets. Empirical models of individual spray performance can be constructed from field trial data [2]. However, in order to obtain more robust and transferable models, there is a need to incorporate the physical processes involved.

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*,
Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_151,

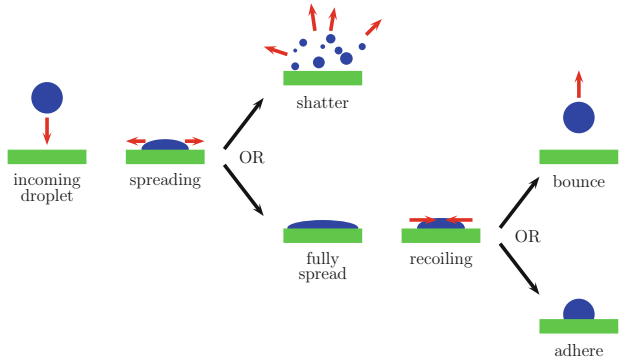


Fig. 1. The processes involved in the impact between a spray droplet and a leaf surface showing the three possible outcomes

We consider a single aspect of crop spraying: the immediate result of a collision between a spray droplet and the surface of a leaf. The current discussion will be restricted to where the leaf surface is horizontal and the spray droplet, falling vertically, impacts it at 90°. It will also be assumed that the interaction remains within the edges of the leaf. (If the interaction extends beyond this boundary or if the leaf is tilted there can be further complications such as the run-off of the spray droplet from the leaf [3].) Three outcomes of the initial collision are possible (Fig. 1):

- Adhesion, where the droplet spreads out on the leaf surface and remains there.
- Bounce, where the droplet after spreading on the leaf rebounds to leave the surface.
- Shatter, where the droplet is broken into a number of smaller droplets which then may leave the surface.

In practice it may be challenging to distinguish among interactions that are at the borderlines of these different initial outcomes. For instance they may lead to the same final result: a droplet that initially bounces can return under gravity and be retained on the leaf surface through secondary impaction. Even for cases where initial adhesion is low, re-capture of bouncing or shattered droplets may produce high retention.

2 Physical Parameters, Contact Angles and Their Measurement

Previous work has led to a number of models for the dynamics of a droplet impacting with a solid surface [1]. From these a number of key physical parameters have been identified. These can be measured in the laboratory

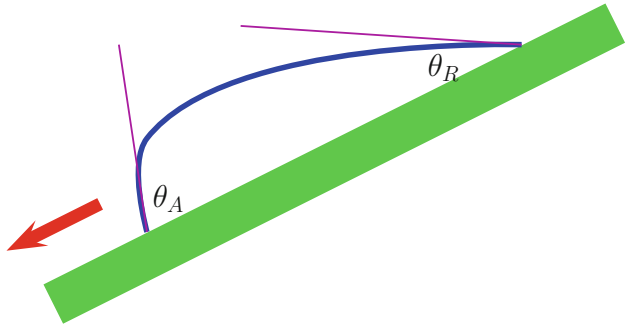


Fig. 2. A droplet flowing down a slope indicating advancing (θ_A) and receding (θ_R) contact angles

and incorporated into the model. Properties of the droplet fluid include its dynamic viscosity μ, surface tension σ and density ρ. The droplet itself can be taken to be initially spherical of diameter D_0, with a downward vertical velocity V_0. Characteristic dimensionless parameters can be constructed from these physical parameters: the Weber, Ohnesorge and Reynolds numbers ($\mathrm{We} = \rho D_0 V_0^2/\sigma$, $\mathrm{Oh} = \mu/\sqrt{\rho\sigma D_0}$, $\mathrm{Re} = \sqrt{\mathrm{We}}/\mathrm{Oh}$). In addition to these parameters, the advancing contact angle (θ_A) and the receding contact angle (θ_R) characterise the interaction of the droplet with the surface.

The contact angles for various formulations, on the leaves of each of the plant species studied, were measured using a KSV CAM 200 optical contact angle instrument with an automated tilting stage and Basler digital video camera. On a horizontal surface a resting droplet has the equilibrium contact angle all around its edge. However, as the surface is tilted the downslope and upslope contact angles begin to differ. For a wide range of tilts it was found that the difference between front and rear contact angles remains relatively constant. Finally, the tilt may be reached for which the droplet flows down the surface and the downslope and upslope contact angles may be taken as the advancing and receding contact angles, respectively (cf. Fig. 2). In practice, for some formulations, it is not obvious that the droplet has begun to run down the slope, and an alternative strategy is used to estimate the advancing and receding contact angles. The model described in Sect. 4 is comparatively insensitive to the exact value of the advancing contact angle and it is adequate to take the advancing contact angle to be equal to the equilibrium contact angle. The receding contact angle is computed by subtracting the relatively constant difference between upslope and downslope angles on a tilted surface from the equilibrium contact angle. For both of these angles a number of measurements were made and there was found to be a natural statistical variation between individual droplets of only a few percent. The effect of this variability is illustrated when the theory is compared with the results (Sect. 6).

3 Experimental Determination of Adhesion or Bounce

An impulse-jet droplet generator was used to produce monosized droplets. These were individually fired straight down onto a horizontal leaf surface. It was noted whether the *initial* outcome of the collision was that the droplet remained on the surface (adhesion) or alternatively left the surface (either as a bounce or through shatter). The experiment was repeated (ten droplets onto each of five replicate leaves, i.e. 50 droplet impactions) for each set of conditions for statistical robustness. The study utilised different plant species, covering a range of leaf surface roughness characteristics (50% acetone droplet contact angles varied from 0° to 108°); formulations to provide a range of surface tensions (33–72 $\mathrm{mNm^{-1}}$); droplet sizes (ca. 300–900 µm) and droplet impact velocities (ca. 1–3.5 $\mathrm{ms^{-1}}$).

4 The Bounce Model

Assuming for the present that a droplet does not shatter (this process is considered in Sect. 5), the process of collision between droplet and surface is modelled using conservation of energy. The falling droplet contains kinetic energy, due to its vertical velocity, and potential energy, which is mainly due to its surface tension. As the droplet impacts a leaf, spreads out and flattens on the leaf surface, some of the initial kinetic energy is converted into potential energy through increases in droplet surface area. The droplet reaches a point of maximum spread from which it then recedes back towards a more spherical shape returning under the force due to the surface tension. During both the expansion and recession phases the droplet loses energy by friction. If the loss is small enough then sufficient energy remains for the droplet to leave the leaf surface (bounce) as it retreats from its spreading phase. If friction energy losses are larger the droplet is retained (adhesion).

We use the Attané–Girard–Morin (AGM) model [1] to describe the droplet's behaviour on the leaf surface. The spreading (or receding) droplet is modelled by a rimmed cylinder that satisfies the conservation of energy equations:

$$\frac{1}{12}\frac{d}{dt}\left[\left(\frac{2}{3}+\frac{1}{45}\frac{1}{r^6}\right)\left(\frac{dr}{dt}\right)^2\right]+\frac{d}{dt}\left[r^2(1-\cos(\theta))+\frac{1}{3r}\right] +4\,\mathrm{Oh}\left(3r^4+\frac{2}{3}\frac{1}{r^2}+sr\right)\left(\frac{dr}{dt}\right)^2=0. \qquad (1)$$

There is only one free parameter, $s = 1.41\mathrm{Oh}^{-2/3}$, that has been fitted empirically [1]. The cylinder radius (r) and time (t) have been nondimensionalised, and θ is the advancing contact angle when the droplet is spreading ($dr/dt > 0$) and the receding contact angle when the droplet is receding ($dr/dt < 0$). The

initial conditions arise by replacing the falling spherical droplet with a cylinder on the leaf surface which has the same potential and kinetic energies:

$$\left[r^2(1-\cos(\theta))+\frac{1}{3r}\right]_{t=0}=1 \quad \text{and} \quad \left[\frac{dr}{dt}\right]_{t=0}=\sqrt{\text{We}}\left[\frac{2}{3}+\frac{1}{45}\frac{1}{r^6}\right]_{t=0}^{-1/2}. \quad (2)$$

However, for $\theta > 109°$, the pragmatic approach of assigning $r(0) = 0.39$ is taken as in this range $r(0)$ ceases to have a real positive root [1].

For the spreading phase it is appropriate to take θ to be the advancing contact angle, however, the equilibrium contact angle is an adequate estimate of this. In the subsequent phase of recession, the same model is valid, providing we modify the angle θ so that it is now the receding contact angle. (The estimation of these angles in practice is described in Sect. 2.)

We extend the AGM model to include a new criterion for bounce: if the receding droplet returns to the original cylinder radius $r(0)$ then bounce occurs. This approximation mimicks the initialisation (as justified for the AGM model [1]). Here, again, the droplet on the leaf has the same surface energy as a spherical droplet above the leaf surface and the remaining kinetic energy can be converted into vertical motion. If, instead, the droplet comes to rest at $r > r(0)$ then adhesion will occur. The final oscillatory motion of the droplet can be studied by allowing the droplet to continue to expand and contract further, changing the angle θ from the receding contact angle to the advancing contact angle and vice versa, as appropriate.

5 The Shatter Model

The AGM model is invalid if the droplet shatters. Then, during the spreading phase, the droplet breaks into smaller droplets which may leave the leaf. Modelling of this process is at a more preliminary stage [3]. Again by considering the energy balance, Mundo et al. [4], obtained a criterion for shatter

$$\text{Oh}\,(\text{Re})^{1.25} > K \quad (3)$$

where K is a constant. Their analysis found a value of $K = 57.7$ was valid over a wide range of surface roughnesses, however, Yoon et al. [5] observed that there should be some further variability and obtained $K = 152$ for water on a paraffin wax surface. At present, for individual cases, K can be determined by experiment in the laboratory. However, in the future it is planned to incorporate surface roughness and the contact angles into the formula. For comparison with the experimental results (Sect. 6), the present criterion has been used with an approximate value $K = 100$.

6 Results

Figure 3 illustrates the accuracy of the model predictions for the plant species *Pisum Sativum* (pea). Experimental results are shown with symbols: the

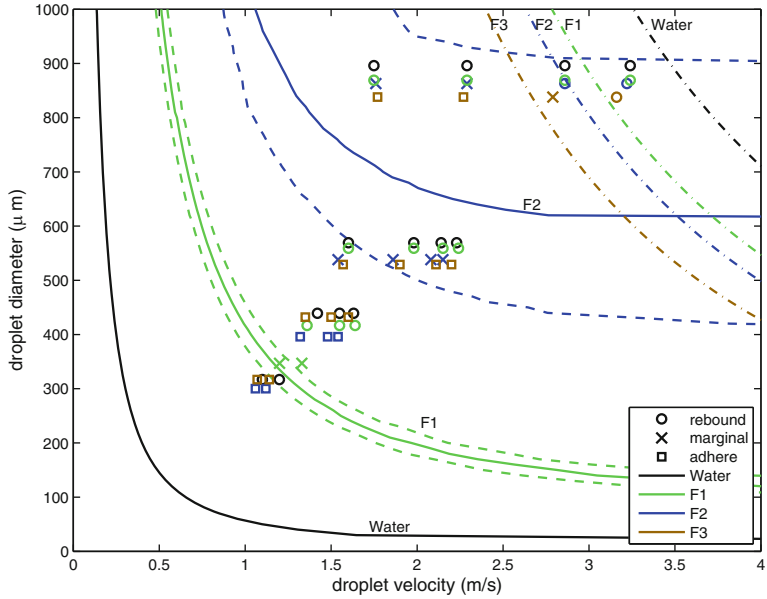


Fig. 3. Comparison of experimental results (*symbols*) with the model (*lines*). Plant species: *Pisum Sativum* (*pea*). Water ($\sigma = 72$) and formulations: F1 ($\sigma = 49$), F2 ($\sigma = 42$), and F3 ($\sigma = 33$)

'rebound' symbol indicates that the droplet did not adhere (i.e. it bounced or shattered) and the 'marginal' symbol indicates that in some trials the droplet adhered and in others it did not. Solid lines show the theoretical value (obtained from the AGM model) above which bounce can occur. Associated dashed lines enclose a borderline region due to natural variability and are computed using the variation in the contact angle determination. The dot-dashed lines represent the model value above which shatter occurs.

The water ($\sigma = 72$) results are above the bounce threshold and no droplets adhere. Variation of the receding contact angle for water is very small and the borderline region is not plotted. The formulation F1 droplet ($\sigma = 49$) outcomes are marginal near the borderline region and rebound above. The formulation F2 droplets ($\sigma = 42$) adhere below the bounce threshold and have marginal outcomes in the borderline region until non-adhesion occurs above the shatter threshold. Finally, the formulation F3 droplets ($\sigma = 33$) are below the bounce threshold (not shown as it is off the graph scale) and adhere until the shatter threshold is crossed where there is first a marginal outcome and then rebound.

7 Discussion and Conclusions

The current model shows excellent agreement with experimentally obtained values for adhesion when applied to one particular plant species (*Pisum Sativum*). Although not shown here similar good correlations have been found with other plant species. Currently the results for shatter are fitted with a free parameter K. The method will be further extended to include a more physical reason for the choice of this parameter to enable the model to be used as a predictive tool for both droplet adhesion and shatter.

References

1. Attané, P., Girard, F., Morin, V.: Phys. Fluids **19**, 012101, 1–17 (2007)
2. Forster, W.A., Kimberley, M.O., Steele, K.D., Haslett, M.R., Zabkiewicz, J.A.: J. ASTM Int. **3**(6). Paper ID JAI13528 (2006)
3. Mercer, G., Sweatman, W.L., Elvin, A., Caunce, J., Fulford, G., Harper, S., Pennifold, R.: In: Wake, G.C. (ed.) Proceedings of the 2006 Mathematics-in-Industry Study Group, pp. 57–85. Massey University, Albany, New Zealand (2007)
4. Mundo, C.H.R., Sommerfeld, M., Tropea, C.: Int. J. Multiphase Flow **21**, 151–173 (1995)
5. Yoon, S.S., DesJardin, P.E., Presser, C., Hewson, J.C., Avedisian, C.T.: Int. J. Multiphase Flow **32**, 132–157 (2006)

Optimisation through Control in Static and Dynamic Traffic Networks

Richard Mounce

University of Sheffield, UK richardmounce@hotmail.com

Summary. There is clearly a need for optimising traffic systems in order to reduce congestion and improve network reliability. A system optimal assignment is a traffic flow pattern that minimises total network costs. In reality, travellers are not under any centralised control, but instead choose routes in order to minimise their own individual travel costs, which in general does not lead to a system optimal assignment. Travellers may be induced to choose routes that are closer to yielding a system optimal assignment through the use of tolls and signal control. The paper considers both approaches within a static traffic model (where flows and costs stay constant over time) and within a dynamic traffic model (where flows and costs vary over time).

1 Introduction

Traffic engineers wish to optimise flows in a network; generally they want to minimise travel costs through the use of signals and tolls. The system optimal problem is to route traffic through the network in such a way that the total cost summed over all travellers is minimised. Since travellers seek to minimise their own travel cost, the traffic network must be optimised subject to user equilibrium, where more costly routes are unused. In the static model detailed in Sect. 2, the system optimal flows can be achieved at user equilibrium by appropriate choice of tolls (assuming signals are fixed), but in the dynamic model the problem is much more complex. Optimising the network by signal control is preferable to using tolls because of the additional cost to local authorities to collect and enforce tolls. Also, to make tolls practicable they are usually kept uniform across areas or cordons rather than being link specific, meaning that they are more effective at reducing overall travel demand than affecting route choice. Allsop [1] first suggested that traffic engineers should take explicit account of the long run influence that their signal setting policies have on the pattern of traffic flow, and that this could be achieved by iterative optimisation assignment (i.e. alternately updating the signal settings for fixed flows and then solving the traffic equilibrium problem for fixed

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*,
Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_152,

signals) although Dickson [2] showed that this process does not necessarily reach an optimal point. Alternatively, both steps can be combined by specifying a dynamical system that incorporates both responsive signal control and travellers' rerouting as in Mounce [5]. In order for such a system to be at equilibrium, it must be at equilibrium with respect its own rules and with respect to travellers' rerouting. Section 2 details several signal setting policies for the static model. Section 3 details a dynamic queueing model and various signal policies, which are counterparts of the policies in Sect. 2.

2 The Static Traffic Model

The traffic network is considered to be a directed graph consisting of a set of nodes and a set of links. Suppose that a set K of origin-destination (OD) pairs is given and that there is a fixed (positive) demand for travel between each of these OD pairs; let ρ_k be the demand for travel between OD pair k. A route is defined to be any acyclic path connecting an OD pair. Denote the flow on route r by X_r and the route flow vector by $\mathbf{X} = (X_1, X_2, \dots X_N)$ where N is the number of routes in the network. Let R_k denote the set of all routes connecting origin-destination pair k. Then the set of feasible route flow vectors, denoted $\mathcal{D}$, is given by

$$\mathcal{D} = \left\{ X \in \mathbb{R}_+^N : \sum_{\Gamma \in R_k} X_\Gamma = \rho_k \ \forall_\parallel \in \mathcal{K} \right\}. \tag{1}$$

If the route-link incidence matrix is denoted $\mathbf{A}$ (where $A_{ij} = 1$ if route j traverses link i and 0 otherwise) then the link flow vector, denoted $\mathbf{x}$, can be specified in terms of the route flow vector by $\mathbf{x} = \mathbf{AX}$. Suppose that at each node there is a given compatibility matrix, which gives all the sets of approach links along which traffic may simultaneously flow, called stages. Suppose also that there is a stage green time vector λ specifying the green time proportions assigned to each stage. A green time vector is feasible if each component is bounded below by λ_{min} and stage green times sum to 1 at each node, i.e. if there are M stages in total, the set of feasible green time vectors is given by

$$\mathcal{G} = \left\{ \lambda \in \mathbb{R}^M : \lambda_m \geq \lambda_{min} > 0 \,\&\, \sum_{m \in S_n} \lambda_m = 1 \, \forall_n \right\}$$

where S_n is the set of stages at node n. For each link i, suppose that the link cost c_i is given as a function of link flow vector $\mathbf{x}$ and the green time vector λ. These link cost functions determine the link cost vector $\mathbf{c}(\mathbf{x}, \lambda) = (c_1(\mathbf{x}, \lambda), c_2(\mathbf{x}, \lambda), \dots c_n(\mathbf{x}, \lambda))$ where n is the number of links in the network. Then the route cost vector $\mathbf{C}(\mathbf{X}, \lambda) = (C_1(\mathbf{X}, \lambda), C_2(\mathbf{X}, \lambda), \dots C_N(\mathbf{X}, \lambda))$ is given by $\mathbf{C}(\mathbf{X}, \lambda) = \mathbf{A^T c}(\mathbf{AX}, \lambda)$. At user equilibrium, more costly routes are unused, i.e. $\forall k \in K$, $\forall r, s \in R_k$, $C_r(\mathbf{X}, \lambda) > C_s(\mathbf{X}, \lambda) \implies X_r = 0$, whereas a system optimal assignment minimises $\sum_r X_r C_r(X)$.

2.1 Optimisation Through Pricing

In this section it is assumed that λ is fixed, so that costs are a function of flow only. If the total cost $f(\mathbf{X}) = \sum_r X_r C_r(X)$ is minimised over the feasible set $\mathcal{D}$ subject to the constraints $g_k(\mathbf{X}) = \sum_{r \in R_k} X_r - \rho_k = 0$ for each OD pair $k \in K$, then by the method of Lagrange multipliers, $\nabla f = \sum_{k \in K} \lambda_k \nabla g_k$. Notice that $(g_k(\mathbf{X}))_r = 1$ if $r \in R_k$ and 0 otherwise. Also,

$$(\nabla f)_r = \frac{\partial}{\partial X_r}\left(\sum_s X_s C_s(\mathbf{X})\right) = C_r(\mathbf{X}) + \sum_s X_s \frac{\partial C_s(\mathbf{X})}{\partial X_r}.$$

This implies that if r and s connect the same OD pair (OD pair k say),

$$C_r(\mathbf{X}) + \sum_{u \in R_k} X_u \frac{\partial C_u(\mathbf{X})}{\partial X_r} = C_s(\mathbf{X}) + \sum_{u \in R_k} X_u \frac{\partial C_u(\mathbf{X})}{\partial X_s}. \tag{2}$$

Certainly in general (2) is not satisfied at a user equilibrium. However, if the toll T_r to traverse route r is chosen to be $T_r(\mathbf{X}) = \sum_s X_s \frac{\partial C_s(\mathbf{X})}{\partial X_r}$ (giving a new tolled cost $C_r^T(\mathbf{X}) = C_r(\mathbf{X}) + T_r(\mathbf{X})$) then for a user equilibrium of the tolled system, $C_r^T(\mathbf{X}) = C_s^T(\mathbf{X})$ for each pair of used routes r and s connecting the same OD pair, which reduces to (2). Therefore any user equilibrium of the tolled system is system optimal.

2.2 Optimisation Through Signal Control

By introducing the concept of the pressure of a stage, several signal policies can be incorporated into the same framework. Let $P_m^{X\lambda}$ denote the pressure of stage m when the route flow vector is $\mathbf{X}$ and the green time vector λ. Equilibrium of the signal control policy occurs when less pressurised stages receive minimum green time, i.e. if for all stages k and m at the same node,

$$P_k^{X\lambda} < P_m^{X\lambda} \implies \lambda_k = \lambda_{min}.$$

For the equisaturation policy, $P_m^{X\lambda} = \max_{i \in m} \frac{x_i}{\lambda_i s_i}$. For other policies, the stage pressures can be defined in terms of the link pressures $p_i^{X\lambda}$ by

$$P_m^{X\lambda} = \sum_{i \in m} p_i^{X\lambda}. \tag{3}$$

The delay-minimisation policy, which minimises the total delay at a junction, chooses $p_i^{X\lambda} = -x_i \frac{\partial d_i}{\partial \lambda_i}$ where d_i is the delay on link i. Although delay-minimisation performs best at a single junction, it may not perform best over a network. Policy P_0 in Smith [6] chooses $p_i^{X\lambda} = s_i d_i = \frac{q_i}{\lambda_i}$ where s_i is the saturation flow on link i. Properties and solution methods for these policies are discussed in Smith and Van Vuren [7].

3 A Dynamic Traffic Model with Signals and Queues

The underlying network is the same as in Sect. 2. However in the dynamic model flows and costs vary over within-day time, which is considered to be continuous and represented by the interval $[0,1]$. The inflow rate to route r, denoted X_r, is considered to be a real-valued, non-negative, essentially bounded and measurable function (which may or may not be continuous). The null sets are then quotiented out (i.e. $X_r = Y_r$ means that X_r and Y_r agree for almost all time $t \in [0,1]$) so that each route inflow is in $L^\infty[0,1]$. All of these route inflow functions are components in the route flow vector $\mathbf{X}$. Demand for travel between OD pair $k \in K$ is considered to be a fixed function $\rho_k \in L^\infty[0,1]$. The set of feasible route flow vectors is therefore

$$D = \left\{ X \in \oplus_{i=1}^{N} L^\infty[0,1] : X_r \geq 0 \,\forall\, r \& \sum_{r \in R_k} X_r = \rho_k \,\forall\, k \in K \right\}$$

with norm on $\oplus_{i=1}^{N} L^\infty[0,1]$ being the supremum norm on the space of cumulative inflows as in Mounce [4] (and the metric on D will be the metric induced by this norm). It is assumed that link saturation flows s_i are fixed (and positive), but that at signalised junctions the green times assigned to each stage can vary. The green time allocated to stage m at time t will be denoted $\lambda_m(t)$ and it will be supposed that λ_m is a Lipschitz continuous function of within-day time (with Lipschitz constant k_1). λ will be said to be feasible if each component is bounded below by some constant $\lambda_{min} > 0$ and if for each signalised node the stage green times sum to one at each within-day time. Therefore the set G of feasible green time vectors is given by

$$\left\{ \lambda \in \oplus_{i=1}^{M} C[0,1] : \forall m \lambda_m \in Lip(k_1) \& \lambda_m(t) \geq \lambda_{min}, \sum_{m \in S_n} \lambda_m(t) = 1 \forall n \forall t \in [0,1] \right\}.$$

The norm on G will be the supremum norm as in Mounce [5] and the distance on G will be the metric induced by this norm. Given a green time vector λ, the exit capacity for link i at a signalised node at time $t \in [0,1]$, denoted $\kappa_i^\lambda(t)$, is given by $\kappa_i^\lambda(t) = s_i \sum_{m:i \in m} \lambda_m(t)$. For links at unsignalised junctions, the exit capacity is assumed to be fixed but is similarly denoted by $\kappa_i^\lambda(t)$. Given any link inflow function x_i, suppose that the cost to traverse link i if entered at time t, denoted $c_i^{x\lambda}(t)$, is the sum of a constant (congestion-free) travel time c_i , a constant toll t_i (which will be converted into a cost in time units) and a bottleneck delay $d_i^{x\lambda}(t)$. The cost to traverse route r, denoted $C_r^{X\lambda}(t)$, can then be found by summing all of the link costs at the respective times that each link is reached, i.e. $C_r^{X\lambda}(t) = \sum_{i \in r} c_i^{x\lambda}(A_{ir}^{X\lambda}(t))$ where $i \in r$ means that link i is a link on route r and $A_{ir}^{X\lambda}(t)$ is the arrival time at link i when route r is entered at time t if the route flow vector is $\mathbf{X}$ and the green time vector is λ. Queueing occurs vertically at link exits when traffic flow exceeds capacity. If link i is congested at time t, then the queue

on link i, denoted $q_i^{x\lambda}(t)$ is given by $q_i^{x\lambda}(t) = \int_{b_i^{x\lambda}(t)}^{t}(x_i(u-c_i) - \kappa_i^{\lambda}(u))du$ where $b_i^{x\lambda}(t) = \sup\left\{u \in [0,t] : q_i^{x\lambda}(u) = 0\right\}$. The bottleneck delay $d_i^{x\lambda}$ at link i is connected to the bottleneck capacity κ_i^{λ} and the bottleneck inflow x_i by the equation

$$\int_{t_0-c_i}^{t-c_i} x_i(u)du = \int_{t_0}^{t+d_i^{x\lambda}(t)} \kappa_i^{\lambda}(u)du \tag{4}$$

for all t in some congested period $[t_0, t_1]$. Now let x_i be the inflow to link i, which depends on the route flow vector $\mathbf{X}$ and the green time vector λ. Let $\mathbf{x}$ be the vector consisting of all these link flow functions. If x_{ir} denotes the inflow at link i of traffic on route r, then clearly $\sum_{r:i\in r} x_{ir} = x_i$. If Ox_{ir}^{λ} represents the outflow from link i of traffic on route r when the route flow vector is $\mathbf{X}$ and the green time vector is λ, then

$$\int_0^t x_{ir}(u)du = \int_0^{t+c_i+d_i^{x\lambda}(t)} Ox_{ir}^{\lambda}(u)du \tag{5}$$

since traffic entering at time t exits at time $t + c_i + d_i^{x\lambda}(t)$. Given a particular route flow vector $\mathbf{X}$ and green time vector λ, the associated link flow vector $\mathbf{x}$ is defined to be the solution of the integral equations (4) and (5). $\mathbf{c}(\mathbf{x}, \lambda)$ is in $\oplus_{i=1}^{n} C[0,1]$ and $\mathbf{C}(\mathbf{X}, \lambda)$ is in $\oplus_{i=1}^{N} C[0,1]$ (Mounce [4]). The norm on $\oplus_{i=1}^{n} C[0,1]$ and on $\oplus_{i=1}^{N} C[0,1]$ will be the supremum norm. At dynamical user equilibrium, more costly routes are unused for all within-day time, i.e. for all routes r and s connecting the same OD-pair and for all $t \in [0,1]$,

$$C_r^{X\lambda}(t) > C_s^{X\lambda}(t) \implies X_r(t) = 0. \tag{6}$$

3.1 Optimisation Through Tolling

In the dynamic model, the system optimal problem is to minimise

$$\sum_r \int_0^1 X_r(u) C_r^X(u)du$$

over the feasible set D. Ghali and Smith [3] showed that in the dynamic model route marginal costs are not simply sums of link marginal costs. Hence the problem of determining optimal tolls for a general network is not so straightforward as in the static model.

3.2 Optimisation Through Signal Control

Stage pressures are defined as in the static model (but now are time-varying) and any responsive policy seeks to approach a time-varying stage green time vector λ such that less pressurised stages receive minimum green time for all within-day time, i.e.

$$\forall t \in [0,1] \& \forall n \forall k, m \in S_n, P_k^{X\lambda}(t) < P_m^{X\lambda}(t) \implies \lambda_k(t) = \lambda_{min}. \quad (7)$$

In order for a network incorporating responsive signal control to be at equilibrium, it must satisfy both (6) and (7). For the equisaturation policy $P_m^{X\lambda}(t) = \max_{i \in m} \frac{q_i^{x\lambda}(t)}{s_i \lambda_i(t)}$. For the delay-minimisation policy, since delays on link i accumulate over time at rate $q_i^{x\lambda}(t)$, total delays are minimised by defining the link pressure as

$$p_i^{x\lambda}(t) = \begin{cases} s_i & \text{if } q_i^{x\lambda}(t) > 0 \\ 0 & \text{if } q_i^{x\lambda}(t) = 0 \end{cases}$$

and let the link pressures sum to give stage pressures as in (3). For policy P_0, $p_i^{x\lambda}(t) = \frac{q_i^{x\lambda}(t)}{\lambda_i(t)}$ and then (3) is applied. Mounce [5] shows existence of equilibrium when the signal policy is equisaturation and P_0 (but not for delay-minimisation as is mistakenly claimed).

4 Conclusion

The paper considered the optimisation of traffic networks subject to equilibrium constraints through signal control and pricing, both in a static and in a dynamic traffic model. In the static model, optimal tolls can be calculated from the route marginal costs. In the dynamic model, difficulty in calculating route marginal costs makes it difficult to determine optimal tolls. A variety of signal control policies were outlined for both the static and dynamic models.

References

1. Allsop, R.E.: Delay-minimising settings for fixed-time traffic signals at a single road junction. J. Inst. Math. Appl. **8**, 164–185 (1971)
2. Dickson, T.J.: A note on traffic assignment and signal timings in a signal-controlled road network. Transportation Research Part B **15**, 267–271 (1981)
3. Ghali, M.O., Smith, M.J.: A model for the dynamic system optimum traffic assignment problem. Transp. Res. B **29**, 155–170 (1995)
4. Mounce, R.: Convergence in a continuous dynamic queueing model for traffic networks. Transp. Res. B **40**, 779–791 (2006)
5. Mounce, R.: Existence of equilibrium in a continuous dynamic queueing model for traffic networks with responsive signal control. In: Lam, William, H.K., Wong, S.C., Lo, H.K. (eds.) Transportation and Traffic Theory 2009: Golden Jubilee, pp. 327–344. Springer, New York (2009)
6. Smith, M.J.: Traffic control and traffic assignment in a signal-controlled network with queueing. Proceedings of the 10th International Symposium on Transportation and Traffic Theory, pp. 61–68 (1987)
7. Smith, M.J., Van Vuren, T.: Traffic equilibrium with responsive traffic control. Transp. Sci. **27**, 118–132 (1993)

The Science of Desire: A Systematic Approach to Mathematical Modeling

Kees van Overveld

Department of Technology Management, Eindhoven University of Technology, Eindhoven, PO Box 513, 5600 MB Eindhoven, the Netherlands,
k.van.overveld@wxs.nl

Summary. A systematic approach to developing mathematical models is presented. The approach applies to problems where (optimal) decisions should be found that lead to maximized benefit or yield, and/or minimized loss or disadvantage. The approach is characterized in that all occurring relations are regarded as *functional relationships*; the model is developed in the form of a non-cyclic, directed graph of variables where the edges represent functional dependency. Next, Genetic Programming can be used to obtain (approximate) optimal decisions.

1 Context: The Didactics of Applied Mathematics *vs.* Mathematical Modeling

Over the last decades a clear distinction developed between curricula for pure mathematics and curricula for applied or industrial mathematics. The purpose of *pure* mathematics is, to produce insights, to understand structures and to prove theorems within the realm of formal mathematical reasoning; in *applied* mathematics problems are solved that come from an external context. The starting point for a project in pure mathematics is a set of axioms and perhaps some conjectures; a successful result takes the form of a proven – and preferably deep – theorem. The success of an exercise in applied mathematics, to the contrary, is invariably measured with respect to its usefulness in a non-mathematical context: the solution must be translatable e.g. in terms of increased benefits or yields, or reduced risk or disadvantage.

This means that as a part of a project in applied mathematics, there is always a phase of *modeling*. 'Modeling' means: start with the original, non-mathematical version of a problem, take its industrial, commercial and/or social context into account, and propose a set of variables, relations, equations and/or constraints that in some sense adequately *represent* the problem situation at hand. Next, apply formal mathematical manipulations with these variables and equations to produce some mathematical results (say, numerical values), guided by needs inherent to the problem. Finally, these results need

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*,
Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_153,

to be interpreted in terms of the original problem setting: they are *mapped back* to the problem domain. The entire endeavor is successful if the process of (1: modeling, 2: manipulating, and 3: mapping) leads to some advantages in the original problem domain.

The first and third phases of this process are not governed by mathematical logic nor rigor. There is no provable correct formal way for building a model, nor for interpreting the mathematical result in terms of the original problem. Whereas the mathematical manipulations in phase 2 are governed and taught by the standard conventions of mathematics and mathematical didactics, the process of devising a set of meaningful variables, relations and equations is left entirely to the common sense, the intuition and the experience of the mathematical practitioner.

Junior and less experienced mathematicians and students in applied mathematics often feel uncomfortable with this lack of systematic techniques for phases 1 and 3. But even for senior applied mathematicians the ad-hoc nature of these phases may prove an obstacle to transparent communication with customers or colleagues regarding the followed route.

This paper aims to partially overcome this omission.

2 Design Situations: A Broad Class of Problems for Mathematical Modeling

In the literature on design methodology, numerous definitions of *design* can be found, ranging from quite confined (say, the design of the shape a mechanical component, involving mainly geometrical representations and spatial constraints) to the very broad (the design of a business model, perhaps involving marketing, socio-demographic arguments and logistics). For the scope of this expose, a design situation in general is defined as "the process of taking decisions such that the happiness of stakeholders increases". 'Taking a decision' here, means: assigning a value to some variable – assuming that the value is within the range of admissible values for that variable, and that the designer has sufficient mandate to decide that the variable shall attain the chosen value. In the terminology of this paper, such variables will be called *category-I variables*. Category-I variables can be ordinal or nominal; ordinal variables can be numeric; their types can be closed (e.g., a finite range of values, such as components from a catalogue) or open. Examples are: the choice of material for some component (= a nominal variable; closed); the geometrical dimension of some feature of the designed artifact (open or closed), or the selling price of some commodity (open). The term 'happiness' in the definition occurs to stress that design always starts with the observation that somebody *wants* something. There is a desire to obtain or achieve something – hence the title of this paper. The design process should be set up such that this desire will become close(r) to realization. Therefore, it is paramount that the degree of fulfillment of the desire is expressed in terms of *ordinal variables*- to be called

category-II variables. Category-II variables have to be ordinal: indeed, the success of the design process must be comparable – at least with the status quo, but the consequences of various sets of choices of category-I variables also must be comparable; there must at least exist a partial ordering, with respect to each category-II variable.

3 A Systematic Procedure for Constructing Mathematical Models for Design Situations

With the notion of category-II variables as a starting point, the process of constructing a mathematical model proceeds as follows. Suppose that a, for the start, a relatively small numbers (2 or 3, say) of category-II variables is given. One of them is taken, say $profit$, and an analysis is performed *which mechanism* is believed to be responsible for the value of $profit$. This mechanism will involve one or more variables, and first a *qualitative* statement about the kind of dependency. For instance, suppose $profit$ represents the expected profit of a company to be set up (obviously, the profit is an ordinal variable: profit should be preferably high – the more the better!). The mechanism that causes the value of $profit$ is (1) money comes in , and (2) money goes out. So it is plausible to introduce two new variables, inc (for 'income') and exp (for 'expenses'), both expressed in Euro's per year. Next we need a computable expression that relates $profit$ to inc and exp. Notice that, in many cases, there is a variety of computable expressions that can be chosen. For instance, small contributions can be ignored or not – does a linearized relationship suffice or should higher order terms be taken into account? As another example, an asymptotic behavior can be obtained either by means of a rational function, a tangent or a negative exponential function – beforehand it may not be obvious to pick out the 'right' one (or: '*a* right one'). It is advisable therefore to give a qualitative specification of the behavior first, and next choose the (mathematically) simplest formal realization of this behavior. In the present example, this could be as simple as

$$profit = inc - exp,$$

thereby first ignoring e.g. inflation, savings, and tax. Such 'second order effects' can easily be accounted for once a first version of the model has been completed. With the introduction of this first function, $profit = f_1(inc, exp)$, the variable $profit$ is formally and functionally defined; it is removed from the 'to-do' stack. Two new variables have been 'pushed on the stack', however: inc and exp next must be defined. None of the two is in category-I: it is unlikely that the designer can freely choose what the income will be; moreover, it is implausible that (s)he will make a free and independent decision how much (s)he is going to spend. Rather, the expenses *depend*, say on the number of

products that is going to be bought ($nrProd$) and the price of these products ($pricePP$). So:

$$exp = nrProd \times pricePP.$$

Again notice that this formula, however trivial it may seem, is merely one possibility: it assumes, for instance, that no discount is given for larger quantities, that there is only one type of product, and that payments are not delayed to a next booking period! During the process, every step of defining the precise, functional meaning of a variable is an excellent opportunity to ponder about these alternatives and to make the assumptions explicit – this improves the transparency of the entire modeling process. With stating $exp = nrProd \times pricePP$ another variable has been 'popped off' the stack (namely exp), and two new ones have been pushed onto the stack ($nrProd$ and $pricePP$). The categories of the new variables, $nrProd$ and $pricePP$ merit some attention: $nrProd$, indeed, is presumably a category-I variable. The designer (or the shop owner) is *free* to make any choice for $nrProd$. $pricePP$ is not in category-I, neither in category-II: its value is given from the context, and it does not depend on any choice the designer can do. This defines a third category, to be called category-III. Now the question can be answered to which category exp belongs: it is clearly not in category-II (indeed, reducing expenses is not going to make the shop owner happy: it is the *difference between income and expenses* that matters, not the expenses per se.) It is also not in category-III, since its value is not independent to category-I variables. It is a so-called *intermediate* or *auxiliary* variable. It serves to couple the various functions, while still allowing to study one function at the time. Variables of this kind reside in category IV.

Now the entire modeling process can be summarized:

- Initially, the to-do stack is empty.
- While there are still category-II variables that have not been pushed onto the to-do stack, or the to-do stack is not empty:
  - Push a new category-II variable onto the to-do stack, or
  - Take a variable, say y from the stack (this is category-II or category-IV)
  - In the latter case, identify the simplest possible mechanism that determines y's value
  - Find a minimal set of variables, say $x_1, x_2, ...$ that are needed to explain or describe this mechanism (these variables may have been defined earlier in the process)
  - Choose an appropriate computational expression that models the dependency $y = f(x_1, x_2, ...)$
  - Establish the categories of the x_i. If they are in categories I or III, they can be removed from the to-do stack. Otherwise they are in category-IV and they are pushed onto the to-do stack.
- Once the to-do stack is empty, every category-II variable is expressed, perhaps by means of a network of functions of category-IV variables, in terms of category-I variables (choices) and/or category-III variables (constants).

The to-do stack being empty is the stopping criterion for the process. The resulting network of functions is the mathematical model sought for; in Sect. 4 it is demonstrated how the model is used to find 'optimal' solutions.

We conclude this section with some remarks:

- During the construction of the model, in every step the focus is on only one variable at the time, and on the mechanism that determines the value of that single variable. This stimulates separation of concerns [1] – considered as good engineering habit in computer science;
- The resulting model is an a-cyclic graph of functional dependencies. This means that any issues regarding 'what depends on what' need to be explicitly resolved beforehand. This is possible in virtue of the nature of the problems: in design problems, there is the assumption that the resulting happiness of the stakeholder ultimately *depends* on the values of category-I variables. In ad-hoc modeling, circular reasoning is a notorious pitfall!
- Since every simplification and simplifying assumption is encountered (and hopefully annotated) together with the variable in which it occurs, expanding the model to account for more detail can also done in an incremental and systematic way, thereby leaving the directed a-cyclic nature of the graph in tact;
- During the construction of the model, discussions may occur about the nature of a variable (category-I or category-III?) These questions are highly relevant: all too often a design situation is of limited use since too many (unnecessary) implicit assumptions are used – in other words, the systematic approach as outlined above helps to exploit the full design space.

4 Solution Methods

In the a-cyclic, directed graph that results from the procedure in Sect. 3, arbitrary functions can occur. These functions need not be differentiable or even continuous: indeed, they will often contain conditional expressions, or the calculation may involve database table access. Moreover, there will be, in general, more than one category-II variable. That means that standard mathematical optimization techniques, geared towards finding stationary points in a single function f, such as solving for $\nabla f = 0$ won't often work. Similar, numerical techniques such as simplex methods or local descent methods or even combined continuous and discrete methods are usually too limited.

Fortunately, there is a practical alternative: the SPEA (=Strength Pareto Evolutionary Algorithm) technique as proposed by Ziztler et al. ([2]). It hinges around the notion of *Pareto Optimality*, ([3]) and *dominance*. A solution, i.e., a set of values for category-I variables, $\{x_1, x_2, x_3, ...\}$ with corresponding category-II variables, $\{y_1, y_2, y_3, ...\}$ is said to *dominate* an other solution $\{x'_1, x'_2, x'_3, ...\}$ with $\{y'_1, y'_2, y'_3, ...\}$ iff $\forall i : y_i \geq y'_i$ – assuming that all y_i need to

be maximized[1]. The collection of non-dominated solutions is called the *Pareto Front*. An exact calculation of the Pareto Front is in all but trivial cases infeasible. However, a practical approach to obtain an approximate Pareto Front has been proposed in [2]. There, *genetic optimization* is used: an initial population of random solution sets is allowed random mutations and cross-over interbreeding against a fitness criterion, where the fitness for a solution X is defined[2] as minus the number of solutions that dominate X. This fitness function is called 'Strength'- hence the name of the algorithm. A software system, aimed at supporting the process of mathematic modeling, which features the interactive, incremental construction of the directed a-cyclic graph cf. Sect. 3, together with a generic implementation of the SPEA algorithm, has first been described in [4]. In various courses on modeling by the author, a more recent version of a similar software system has been used in a variety of design projects.

5 Conclusions and Further Work

For a large class of practical problems, a systematic modeling procedure has been proposed that consists of incrementally building a directed, a-cyclic graph of variables and functional relationships. A mathematical model is obtained that represents the problem at hand in the form of a network of executable functions – perhaps represented in terms of a spreadsheet or (other) computer program. The mathematical manipulations for the actual resolution of the original problem either amount to algebraic or (more often) numerical techniques such as SPEA. Over the last few years, ample experience has been gained with this approach in classroom settings; currently, the method is being used in more realistic small-to-medium sized real life business cases.

References

1. Dijkstra, E.W.: On the role of scientific thought. In: Dijkstra, E.W. (ed.) Selected Writings on Computing: A Personal Perspective. Springer, New York (1982)
2. Ziztler, E., Deb, K., Thiele, L.: Evol. Comput. **8**(2), 173–195 (2000)
3. Fudenberg, D., Tirole, J.: Game Theory. MIT, Cambridge, Chapter 1, Section 2.4 (1983)
4. Ivashkov, M.: ACCEL: a Tool for Supporting Concept Generation in the Early Design Phase. PhD thesis, Eindhoven University of Technology (2004)

[1] Obviously, for a category-II variable y_j that should be minimized, the condition reads $y'_j \leq y_j$.

[2] This definition is a simplification; practical fitness functions are usually somewhat more subtle.

Modeling, Analysis and Simulations of Case Hardening of Steel

L. Panizzi[1], A. Fasano[2] and D. Hömberg[3]

[1] Scuola Normale Superiore, Italy
[2] Department of Mathematics, Florence University, Italy
[3] Weierstrass Institute for Applied Analysis and Stochastics, Berlin, Germany, hoemberg@wias-berlin.de

Summary. A mathematical model for the case hardening of steel is presented. Carbon is dissolved in the surface layer of a low-carbon steel part at a temperature sufficient to render the steel austenitic, followed by quenching to form a martensitic microstructure. The model consists of a nonlinear evolution equation for the temperature, coupled with a nonlinear evolution equation for the carbon concentration, both coupled with two ordinary differential equations to describe the evolution of phase fractions. Existence and uniqueness of solutions are investigated and some numerical simulations are presented.

1 Case Hardening in the Metallurgical Industry

Steel is still the basic material for a modern industrial society. A distinct feature of steel is that one can change its physical properties by thermal interference. The reason for this behavior lies in the occurring solid-solid phase transitions. It is indeed utilized in the heat treatment of steel, which is a process of controlled heating and cooling to achieve a desired distribution of metallurgical phases corresponding to locally varying, heterogeneous physical properties.

Case hardening is a special heat treatment of steel. It is a widely used process in industry aimed to obtain a special sort of steel with a hard case and soft and ductile core.

There are many types of case hardening, the most widely used is called carburizing and we will concentrate on this one. The process can be shortly described as follows: steel is heated up to austenitizing temperature, where the solubility of carbon in iron is high, subsequently it is immersed in a carbon-rich atmosphere for hours in order to allow the diffusion of carbon into the workpiece and finally it is rapidly cooled down to obtain the desired hardening effect.

Many important processes in the metallurgical industry rely on this heat treatment. Hence, there is a special demand for its mathematical modeling and

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*,
Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_154,

simulation. Regarding carburizing, despite its worldwide application, the current process performance faces some challenges regarding the process control. The industrial approach to solving such problems often involves trial and error methods and empirical analysis, both of which are expensive and time consuming. With the present work we aim to derive and analyze a mathematical model, in order to understand the mechanism of case hardening and to predict the carbon concentration profile and case depth during the heat treatment process.

2 The Mathematical Model

The mathematical description of solid-solid phase transitions in steel started with the works of Avrami and Kolmogorov (see, for instance [1]) in the 1930s. Since then the subject has been widely studied and, stimulated by the development of ever-faster computer hardware, numerous papers were published on the numerical simulation of the diffusion controlled phase transitions in steel. The first analytical investigation of phase transitions in steel, concerned with the austenite-pearlite transformation, dates back to the 1980s [4]. The model we propose here is a phenomenological model which takes into account all the relevant parameters and the physical quantities, which are: temperature, carbon concentration and phase fractions which form during the heating and cooling process.

2.1 Kinetics of Phase Transitions in Steel

The kinetics of the phase change can be briefly described as follows. Depending on temperature, two different lattice structures can occur: a body-centered-cubic (b.c.c.) and a face-centered-cubic (f.c.c) lattice. Above a certain temperature A_s steel is in the austenitic phase, a solid solution of carbon in f.c.c. iron. Below A_s this lattice is no longer stable. But before the lattice can change its configuration to form a b.c.c structure, carbon atoms have to diffuse, due to the higher solubility of carbon in the f.c.c lattice. The result is pearlite, a lamellar aggregate of ferrite and cementite, soft and ductile. Upon high cooling rate carbon has no time to diffuse and is trapped, forming a tetragonally distorted b.c.c. lattice, called martensite.

The transformation diagrams of interest for the modelling of the phase fractions evolution (see (1a,b) below), during the cooling process, are called indeed *continuous cooling transformation* (CCT) diagrams and describe the transformation of austenite as a function of time for a continuously decreasing temperature. In other words a sample is austenitized and then cooled at a predetermined rate and the degree of transformation is measured. The start of transformation is defined as the temperature at which 1% of the new microstructure has formed. The transformation is completed when only 1% of the original austenite is left.

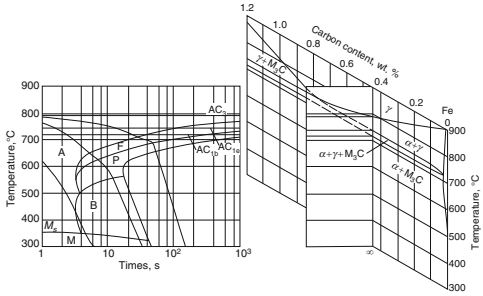


Fig. 1. Equilibrium diagram of the system iron-carbon (*right*) as limit of the CCT-diagram with infinite low cooling rate

In carburized steels the process is strongly influenced by the carbon content, which varies from the carbon-enriched superficial layer to the core. Thus, it cannot be described by only one continuous-cooling-transformation diagram. Figure 1 shows a continuous cooling diagram describing, for a given austenitizing condition, the transformation at all carbon levels in a carburized specimen. The cross sections for fixed carbon percentages give CCT diagrams of the type of the one plotted in Fig. 1 on the left. To avoid unnecessary technicalities for the modelling, we assume that the cooling takes place from the high temperature phase austenite with phase fraction a to two different product phases, pearlite with fraction p and martensite with fraction m. A more elaborate model accounting for all the phases occurring during the heat treatment of steel can be found in [3].

The evolution of the phases p and m can be described by the following system:

$$\dot{p} = (1 - p - m)g_1(\theta, c) \tag{1a}$$

$$\dot{m} = [\min\{\overline{m}(\theta, c); 1 - p\} - m]_+ g_2(\theta, c) \tag{1b}$$

$$p(0) = 0 \tag{1c}$$

$$m(0) = 0 \tag{1d}$$

where c is the concentration of carbon. Here the bracket $[\ ]_+$ denotes the positive part function $[x]_+ = \max\{x, 0\}$ and the dot means the derivative with respect to t. While the growth rate of pearlite $\dot{p}$ is assumed to be proportional to the remaining austenite fraction, the rate of martensite growth $\dot{m}$ is zero if m exceeds either the non-perlitic fraction $1-p$, or the threshold $\overline{m}$ depending on both temperature and carbon concentration. Indeed martensite is produced at temperatures less than a value M_s but complete transformation to martensite can be obtained only below some other temperature threshold M_f. Both these temperatures depend on the local value of carbon concentration. The quantity $\overline{m}(\theta, c)$ represents the maximum attainable value of martensite fraction and can be defined as:

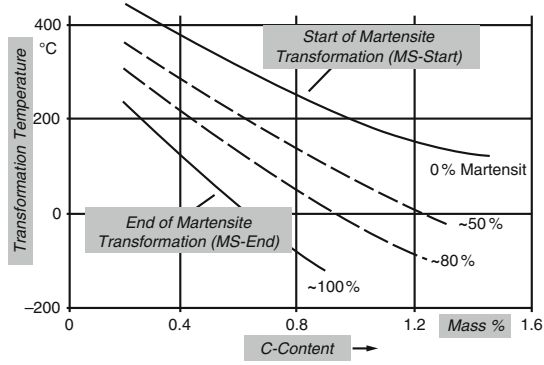


Fig. 2. Level curves of function $\overline{m}(\theta, c)$

$$\overline{m}(\theta, c) = \begin{cases} 0 & \theta > M_s(c) \\ 1 & \theta < M_f(c) \end{cases}$$

and by interpolation for intermediate temperatures. Since there is no phase transition from pearlite to martensite, the term $\min\{\overline{m}(\theta, c); 1-p\}$ represents the maximal fraction of martensite that can be reached at time t. The functions g_1 and g_2 are positive given functions that can be identified from the time-temperature-transformation diagrams described before. The process of carbon diffusion is governed by the following nonlinear parabolic equation:

$$\frac{\partial c}{\partial t} - div((1-p-m)D(\theta, c)\nabla c) = 0.$$

The factor $(1-p-m)$ in front of the diffusion coefficient $D(\theta, c)$ reflects the fact that enrichment with carbon only takes place in the austenite phase. The difference in carbon potential between the surface and the workpiece provides the driving force for carbon diffusion into the piece. The carbon potential of the furnace atmosphere must be greater than the carbon potential of the surface of the workpiece for carburizing to occur. Hence we have the following boundary condition:

$$-(1-p-m)D(\theta, c)\frac{\partial c}{\partial \nu} = \beta(c - c_p)$$

where β, the mass transfer coefficient, controls the rate at which carbon is absorbed by the steel during carburizing and c_p is the carbon concentration in the furnace, usually named carbon potential of the gas. $\frac{\partial c}{\partial \nu}$ denotes the outward normal derivative. The evolution of temperature during the entire process is described by the following nonlinear problem

$$\begin{aligned} \rho\alpha(\theta)\frac{\partial \theta}{\partial t} - div\big(k\nabla\theta\big) &= \rho L_p(\theta)\dot{p} + \rho L_m(\theta)\dot{m} \\ -k\frac{\partial \theta}{\partial \nu} &= h(\theta - \theta_\Gamma) \\ \theta(x, 0) &= \theta_0. \end{aligned}$$

Here ρ is the mass density, α the specific heat, k the heat conductivity of the material. L_p and L_m denote latent heats of the austenite-pearlite and the austenite-martensite phase changes, respectively. θ_Γ is the temperature of the coolant and $\theta_0(x)$ is the temperature at the beginning of the process. In the technical process, we have three different time stages:

- Stage 1: carburization in a furnace, hence $\beta \neq 0$ and $h = 0$.
- Stage 2: diffusion period, with $\beta = 0$ and $h \neq 0$, serving as a linearized radiation law.
- Stage 3: quenching with $\beta = 0$ and $h \neq 0$.

From the mathematical point of view, without loss of generality, we will assume that β and h are time independent functions. Then, the mathematical result to be formulated in the following section can be applied subsequently to the three process stages, covering the complete case hardening process.

2.2 System of Governing Equations

Let $\Omega \subset \mathbb{R}^3$ be an open bounded set with C^2-boundary $\partial\Omega$ and $Q_T := \Omega \times (0,T)$ the corresponding time cylinder. After the considerations made in the previous paragraph, we come to consider the following system of equations governing our process:

$$\rho\alpha(\theta)\frac{\partial\theta}{\partial t} - div\big(k\nabla\theta\big) = \rho L_p(\theta)p_t + \rho L_m(\theta)m_t \quad \text{in } Q_T \tag{2a}$$

$$\frac{\partial c}{\partial t} - div((1-p-m)D(\theta,c)\nabla c) = 0 \quad \text{in } Q_T \tag{2b}$$

$$p_t = (1-p-m)g_1(\theta,c) \quad \text{in } Q_T \tag{2c}$$

$$m_t = [\min\{\overline{m}(\theta,c); 1-p\} - m]_+ g_2(\theta,c) \quad \text{in } Q_T \tag{2d}$$

$$-k\frac{\partial\theta}{\partial\nu} = h(\theta-\theta_\Gamma) \text{on } \partial\Omega \times (0,T) \tag{2e}$$

$$-(1-p-m)D(\theta,c)\frac{\partial c}{\partial\nu} = \beta(c-c_p) \quad \text{on } \partial\Omega \times (0,T) \tag{2f}$$

$$\theta(x,0) = \theta_0, \quad c(x,0) = c_0, \quad p(0) = 0, \quad m(0) = 0 \quad \text{in } \Omega. \tag{2g}$$

3 Results

Under standard assumptions on the data, the following theorems hold:

Theorem 1 (Existence of a weak solution). *There exists a weak solution* (θ, c, p, m) *to problem (2a-g) such that* $\theta \in H^{2,1}(Q_T)$, $c \in W(0,T)$, $p, m \in W^{1,\infty}(0,T;L^\infty(\Omega))$.

Theorem 2 (Uniqueness). *Assume moreover that* α *is constant,* $D = D(\theta)$, $h, \beta \in W_5^1(\partial\Omega), \theta_0, c_0 \in W_5^2(\Omega)$. *Then the solution to (2a-g) is unique.*

For details about assumptions, the definitions of the respective Sobolev spaces and proofs, we refer to the paper [2].

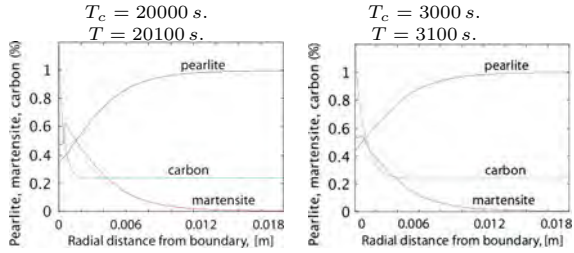


Fig. 3. Phase fractions of martensite, pearlite and carbon percentage curve, plotted against the radius of the *circle*, for different carburizing times T_c and end times T, after a quenching time of 100 s

4 Some Numerical Results

As a sample configuration, we consider the cross section of a cylinder of radius 50 mm. Material parameters are taken from the data tables for the low-carbon steel AISI 4130. An example of the simulation work can be seen in Fig. 3, where we can observe the distribution of phase fractions at the end of a cycle of carburizing and quenching. In the same figure we can see how the formation of martensite depends on the carbon concentration, in accordance with the graphic of Fig. 2 of the first section, obtained from experimental data. The simulations were performed with the finite element software *Comsol Multiphysics*.

5 Conclusions and Further Work

In the present work we have discussed a mathematical model of case hardening. From a mathematical point of view, we have proved the existence of a unique solution. First numerical results confirm qualitative agreement with experiments. A more detailed comparison requires more precise data. To this end a cooperation with some engineering institutes has been started. The development of an optimal control strategy is also under study.

References

1. Avrami, M.: J. Chem. Phys. **7–8–9** (1939, 1940, 1941)
2. Fasano, A., Hömberg, D., Panizzi, L.: Math. Models Methods Appl. Sci. (2009)
3. Hömberg, D., Wolff, W.: IEEE Trans. Control Syst. Technol. **14**, 896–904 (2006)
4. Visintin, A.: IMA J. Appl. Math. **30**, 143–157 (1987)

Surface Recording of His-Purkinje Activity by One-Beat Wavelet Analysis in Atrial Fibrillation and Flutter

V. Pezza[1], B. Pezza[2], E. Pezza[1], L. Pezza[3], M. Curione[4], and V. Sanguigni[5]

[1] USL/FR4 Frosinone, Italy
[2] Ospedale Civile SS. Trinità di Sora (FR), Italy
[3] Dipartimento Me. Mo. Mat., Università di Roma "La Sapienza", via Scarpa, n. 16. 00161 Rome, Italy pezza@dmmm.uniroma1.it
[4] Università di Roma "La Sapienza", Italy
[5] Università di Tor Vergata, Roma, Italy

Dedicated to the memory of Prof. Vincenzo Pezza with infinite love.

Summary. After the first report concerning the invasive recording of the His bundle activity, several efforts have been conducted in order to identify His-Purkinje potential from body surface. The problem of achieving an adequate signal-to-noise ratio, however, is not yet resolved. Only recently the Wavelet Transform System (WTS) has been suggested to bridge the gap. The purpose of the present study is to employ such a method for recording His potentials in the atrial fibrillation and flutter in order to deeply evaluate these arrhythmias.

1 Background

High-frequency components of ECG, including His bundle activity, can be analyzed by signal averaging (SA) or by beat-to-beat recording of High-Resolution ECG. However, the SA method has three major limitations: (1) it is not able to detect dynamic (beat-to-beat) change in the signal; (2) the SA ECG cannot be recorded during complex cardiac arrhythmias; and (3) SA tends to blunt deflections, even with the most precise trigger mechanism [1,2]. On the other hand, even if traditional beat-to-beat recording (TBR) overcomes these limitations, it implies high noise level that, at the moment, is eliminated by Fourier based digital filters. Nevertheless, these filters, may produce a phase shift of atrial signals and, at times, may cause ringing at the end of the atrial waveform [11]. Recently the WTS of one-beat signal has been proposed and validated as a method to overcome these limitations [9]. The aim of this study is to employ such a method for recording the His signal in the atrial fibrillation

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*,
Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_155,

and flutter in order to deeply evaluate these arrhythmias (e.g., to distinguish the premature ventricular contraction from the aberrant beat).

2 Methods

A number of 12 patients (mean age of approximately 48 years) were studied. Seven subjects had atrial fibrillation, and five had atrial flutter. Every patient was taking digitalis.

2.1 Surface Recording and Analog-Digital Signal Converting

All the subjects, after careful skin preparation and application of silver/silver chloride electrodes, were studied by orthogonal X, Y, Z leads recording and by three unipolar precordial leads with electrodes positioned as follows: lead I at the third intercostal space and right sternal border, lead II at the third intercostal space and left sternal border and lead III at the fifth intercostal space and left sternal border. The ECG was amplified, sampled at 1,000 Hz and digitised with resolution of 1 mV by a 12-bit analog to digital converter.

2.2 Digital Signal Processing

Several mathematical methods are used to record micropotentials. Among these, SA and TBR of High-Resolution ECG, both subsequently processed by Fourier based digital bandpass filters, are the most commonly chosen [2]. However, as mentioned above, the SA method has three major limitations: (1) it is not able to detect dynamic (beat-to-beat) change in the signal; (2) the SA ECG cannot be recorded during complex cardiac arrhythmias; and (3) SA tends to blunt deflections, even with the most precise trigger mechanism. On the other hand, traditional beat-to-beat recording (TBR), even if it overcomes these limitations, it means high noise level. At the moment, some researchers [5, 7] try to eliminate this noise, without any appreciable result, by high-pass filters of 80 and 100 Hz [5] or by the so-called "Spatial averaging" [7].

It must be taken into consideration that the averaging process reinforces the identical potentials and attenuates the different ones. Therefore, the averaging technique consists in averaging simultaneous recorded ECG signals from two electrodes placed at a distance small enough that the ECG potentials are similar and, hence, reinforced and ample enough that the noises are different and then attenuated. Obviously such a compromise is not easily attainable. Furthermore, with the two above mentioned methods, SA and TBR, there is the common problem of distinguishing late atrial depolarization and atrial repolarization, from His bundle signals. As atrial waveforms contain a greater representation of the lower frequencies, a high-pass filter of 30 Hz has been used to overcome such a problem. Some portion of His-Purkinje signal, however, is eliminated with this high-pass filter, which, however, may produce a

phase shift of atrial signals and, at times, may cause ringing at the end of atrial waveform. Obviously same limitations are present in all of the Fourier-based filters, even in the band-pass ones which, as above mentioned, are used with both methods in the electrocardiogram for the identification of the His bundle potentials. Therefore, we propose Wavelet Transform Systems (WTS) of one-beat signal as a method to overcome all these limitations. We have used such mathematical model for two purposes: (1) to de-noise the signal achieving high signal/noise ratio, and (2) to extract the characteristic frequencies, or specific oscillations, of the de-noised signal.

2.3 De-Noising the Signal

For such aims wavelet transform offers two complementary interesting features [12]. First, the wavelet transform allows for a temporary localized sliding analysis of the signal, thus giving access at any time to its analysis. Second, the shape of the basis elements used in the wavelet transform differ from the fixed sinusoidal shape of the Fourier transform and can be designed to better fit the shape of the analyzed signal. This allows for a better quantitative measurement. By means of the first feature, wavelet analysis allows to follow the temporal evolution of the spectrum of the frequencies contained in the signal. Such feature is demonstrated by a comparative experiment (see [10, pp. 301–311]) where a signal is analyzed performing both the Fourier and Wavelet transform on two pairs of electrocardiogram of different morphology but of same duration (250 msec). For both methods the signal is decomposed in basic components (harmonics and levels respectively) whose sum reconstructs the original signal. In the case of Fourier analysis the difference of morphology of the two electrocardiograms has no influence in the frequency of the harmonics. Indeed, such frequency, F, is given only by signal duration in accordance with the formula $F = 1/D$ Hz, where D is the signal duration in seconds. For both the electrocardiographic signals, the first harmonic has a frequency of 4 Hz, the second one has a frequency of 8 Hz, the third has a frequency of 12 Hz and so on. The unique difference concerns the magnitude and the phase of the harmonics. Therefore there is no reference to the temporal evolution of the signal. Instead, in the case of Wavelet transform analysis, the same difference determines a distinct appearance of the levels of the decomposition. This yields the evidence of temporal evolution of the signal, while the definition on the frequency domain is limited, even if not fully abrogated, as it will be specified on the discussion. Such concepts are summarized and elucidated as follows. Assume we have two signals composed by two sine waves of different frequency, say, of 50 and 100 Hz (or, 10 and 20 periods, respectively). Assume signal S_1 contains the low-frequency sinusoidal signal first and then the high-frequency signal; whereas, signal S_2 contains the same waves but with inverted order. The FFT of both S_1 and S_2 have an identical spectrum which is flat except for two peaks representing the two above mentioned frequencies. Conversely to Fourier, wavelet analysis displays a different result

for the two signals, clearly showing the exact location in time of where the two frequencies change (by showing a peak), and where they are too. Therefore, by allowing the temporal evaluation of the spectrum, wavelet analysis is capable of revealing frequency breakdown, breakdown points, discontinuities in higher derivatives, proximal discontinuity and so on. Hence, WT analysis is particularly effective for the extraction of very low magnitude signals such as His potential. Moreover, wavelet analysis can often de-noise a signal without appreciable distortion. The de-noising of the signal has been achieved by fixed and minimax form using a wavelet called "Symlet 3" [4] and then the best de-noised signal was chosen. Signal processing performed by the WTS de-noising method differs from one used by the Fourier based digital filters. With the first method the signal is studied to achieve some statistical parameters (correlation, spectrum, and distribution) and then is de-noised by using frequency bands deducted by the above-mentioned parameters [5]. Conversely, with the second method the frequency bands are chosen a priori regardless from the statistical parameters, with the risk of eliminating some portion of His signal.

2.4 Extracting of the Characteristic Frequencies

To extract the characteristic frequencies, or specific oscillations, the signal was analyzed performing a three level, rarely four level, decomposition by using a wavelet function called "Daubechies 7" [4]. Such a wavelet has been chosen because, in accordance with the second above mentioned WTS feature, it seemed to be the most similar to the His bundle signal. The above mentioned high-pass filter of 30 Hz has not been used because distinguishing late atrial depolarization and repolarization, from His bundle signals is not a problem in this case. A deflection in the PR segment was considered as a His bundle potential provided it satisfied two criteria. First, it had to be at least 2 V in amplitude. Second, a relatively isoelectric segment of at least 10 msec was required between the terminal atrial activity and the deflection [6].

3 Results

By wavelet de-noising we have achieved a good removal of the noise, which has been reduced from 6 to 1 V, and a best preservation of the shapes of very sharp peaks. Furthermore, in all of the cases His bundle potential could be recorded. Often it was best identified at 2 and 3 levels and rarely at 3 and 4 levels of the wavelet decomposition, it was 15–25 msec in duration and 2–12 mV in amplitude. The H-Q interval ranged from 30 to 65 m sec. Often, the simple inspection of the de-noised ECG allows to detect the His bundle potential which looks as a monophasic (positive or negative) or biphasic deflection between P and QRS waves of 15–25 m sec. Often, the His bundle potential detection is followed by a notch which, in our opinion, is identifiable with the right bundle-branch potential. Indeed it is absent in right bundle

branch block. However the wavelet decomposition leads to an easier, safer, more constant and reliable identification of His bundle potential.

His bundle recordings were obtained in all seven cases of atrial fibrillation, in which a single His bundle deflection preceded each QRS complex (except for ventricular premature beats). The H-Q intervals were constant from beat to beat during atrial fibrillation. In the case of premature ventricular contraction, the second complex is not preceded by His bundle deflection. Therefore it represents a ventricular extrasystole. Block below the His bundle was not observed in any of the cases. The findings in all five cases of atrial flutter were similar. Each QRS complex was preceded by a single His deflection, and the absence of His deflections in the nonconducted beats indicates that the level of block was proximal to the common bundle. In a case of ventricular extrasystole, the first QRS complex is wide and is not preceded by a His deflection. This it is due to a premature ventricular contraction. Conversely, in a case of aberrantly conducted beat the second complex is of wide duration and is preceded by a His deflection. In this case, the signal was previously submitted to a high-pass filter of 25 Hz to eliminate the f wave which obscured the His deflection. Our data are in accordance with invasive study [8].

4 Discussion

The wavelet transform enables noise reduction by allowing elective use of frequency bands with high signal-to-noise ratio for time feature extraction; therefore automatic estimation of time parameters is robust. The noise reduction from 6 to 1 mV, performed without averaging, is very effective because performing the next step wavelet decomposition also means further filtering of the signal. This filtering occurs because the different frequencies predominate at different levels. That is, with sampling at 1,000 Hz of analogic signal, the frequencies of 500–1,000 Hz predominate at the first level, frequencies of 250–500 Hz predominate at second level, 125–250 Hz at third and 62.5–125 Hz at fourth. Performing both wavelet de-noising and decomposition, in our opinion, is crucial for the identification, without averaging, of very low amplitude signals. Up to now for such a purpose only the de-noising or only the decomposition have been used [6]. With our method a His bundle potential could be readily identified in all of the 12 patients. Hence, the method can be used as a valid tool for recognition of aberrant conduction in atrial fibrillation and flutter. Aberrant conduction may be mistaken as premature ventricular contraction and no rule has been proposed for differentiating the two pathologic conditions [3]. Note that the diagnosis is difficult in the absence of recognizable P waves, as in atrial fibrillation. Therefore our method appears to be the easiest one for the diagnosis of aberrant conduction and for the resolution of the consequent therapeutic problems. Moreover the procedure is very simple to be performed and can therefore find a large clinical application. Indeed, in our opinion, the clinical application of our noninvasive method can be very

large including acute alterations of AV conduction, arrhythmias secondary to acute myocardial infarction, sudden changes in AV conduction after previously mild chronic abnormality, long term follow-up, study of the natural history of cardiac conduction system diseases and so on. On synthesis, all indications for internal His bundle recording may apply to our method which, on the other hand, can be extended to all diseases in which an invasive investigation is contraindicated.

Acknowledgments

We thank Antonio De Falco and Luca G. Tallini for their valuable suggestions and help.

References

1. Berbari, E.J., Scherlag, B.J., Lazzara, R.A.: Computerized technique to record new components of the electrocardiogram. IEEE Trans. Biomed. Eng. **65**, 799–804 (1977)
2. Chen, W.C., Zeng, Z.R., Chow, C., Xine, Q.Z., Kou, L.C.: Application of a new spatial signal-averaging device for the beat-to-beat detection of cardiac late potentials. Clin. Cardiol. **9**, 263–267 (1985)
3. Cohen, S.I., Lau, S.H., Haft, J., Damato, A.N.: Experimental production of aberrant ventricular conduction in man. Circulation **36**, 673–685 (1967)
4. Daubechies, I.: Ten Lectures on Wavelets. CBMF Conference Series in Applied Mathematics, vol. 61. SIAM, Philadelphia (1992)
5. Donoho, D.L.: De-Noising by soft tresholding. IEEE Trans. Inf. Theory **41**, 613–627 (1995)
6. El Sherif, N., Mehra, R., Restivo, M.: Beat to beat surface recording high resolution electrocardiogram: technical and clinical aspects. In: El Sherif, N. (ed.) High Resolution Electrocardiography. Mount Kisco, Futura Pub. Co. Inc., New York (1992)
7. Gomez, J.A., Winters, S.L., Stewart, D., Targonski, A.: Optimal bandpass filters for time-domain analysis of the signal-average electrocardiogram. Am. J. Cardiol. **60**, 1290–1298 (1987)
8. Lau, S.H., Damato, A.N., Berkowitz, W.D., Pattorn, R.D.: Study of atrioventricular conduction in atrial fibrillation and flutter in man using His bundle recording. Circulation **40**, 71–78 (1969)
9. Pezza, V., Pezza, E., Pezza, B., Curione, M., Pezza, L.: Surface recognition of His-Purkinje activity by one-beat analysis wavelet transform system. Circulation **110**, 457 (2004)
10. Santoro, M.T., Pezza, L. (ed.): Vincenzo Pezza – Raccolta Pubblicazioni Scientifiche. CESI, Roma (2009)
11. Simson, M.B.: Use of signals in the terminal QRS complex to identify patients with ventricular tachycardia after myocardial infarction. Circulation **64**, 235–242 (1981)
12. Wiklund, U., Akay, M., Niklasson, U.: Short-term analysis of heart rate variability by adapted wavelet transform. IEEE Eng. Med. Biol. Mag. **16**, 113–118 (1997)

Application of FEM in Analysis of Spigot Joint Contact Problems

T. Podolski[1] and J. Kroczak[2]

[1] Chair or Basics of Machine Design, Poznan University of Technology, ul. Piotrowo 3, 60-965 Poznan, POLAND tomasz.podolski@put.poznan.pl
[2] Chair or Basics of Machine Design, Poznan University of Technology, ul. Piotrowo 3, 60-965 Poznan, POLAND jacek.kroczak@put.poznan.pl

Summary. The conventional designing process of axial-symmetrical connections does not take into consideration the manufacturing errors of connection elements. Here we take measurements of elements that are parts of pin and spigot joints, showing the incomplete contact of surfaces of interlocking parts. Computational models are used to analyse the cooperation of elements of axial-symmetrical connection. We make use of the engineering system I-DEAS with FEM and contact element modules. The resulting parametric computational models enable us to correct the experimental values and to evaluate the contact stresses more accurately.

1 Introduction

Axial-symmetrical joints are ones of basic methods of machines elements connections. Quality of the whole machine depends on their manufacturing accuracy. Issues of their designing are integrally connected to geometrical characteristics recording in technical documentation, to manufacturing of elements and metrology measuring possibilities.

At present standards are solely recommendations concerning manufacturing of particular joint elements, and their usage by a designer is not necessary connected to technical premises. If a designer does not make use of their recommendations then he should be conscious of significant strength, functional and exploitation changes resulted of inadequate choosing of tolerances and fit [1]. In this article there is presented the attempt of evaluation of contact stresses in spigot joint for wide range of fittings.

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*,
Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_156,

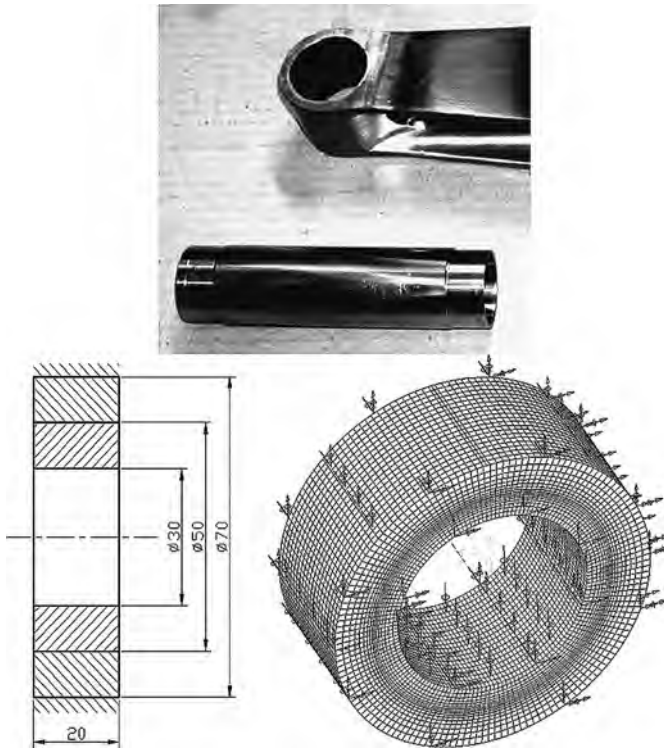


Fig. 1. Pin joint, dimensions of connection and computational model FEM grid with boundary conditions

2 Research Object and Computational Model

There was chosen the spigot joint free fitted characterized with wide tolerance range.

- Connecting rod $\phi 38,075^{+0,030}$.
- Gudgeon pin $\phi 38,025_{-0,010}$.
- Clearance $0,050 \div 0,090$ mm.

There were elaborated the set of geometrical models corresponding to limiting values of the hole and shaft tolerances. Geometric models were the basis of elaboration of FEM models, which were formed by putting finite elements mesh connections on simplified geometry. As the finite elements there were taken the linear element type 'brick' (Fig. 1). Referring to accessibility and simplicity of finite elements mesh generating as well as rich base of finite elements there were chosen system I-DEAS [3]. On the created topology of finite elements there were defined boundary conditions (Table 1). System generated contact elements on cooperating surfaces. Outer surface of a hole (outer ring) has taken away all degrees of freedom, whereas a shaft (inner ring) can

Table 1. Parameters of computational model FEM

Node Label Range	1 - 49518, 49140 Total
Element Label Range	1 - 41600, 41600 Total
Element Types	solid linear brick : 41600
Physical Properties	1 - SOLID1 : 41600
Materials	1 - GENERIC_ISOTROPIC_STEEL : 41600
Solution Set	1 - SOLUTION SET1
	1 - BOUNDARY CONDITION SET 1
	Restraint:
	1 - RESTRAINT SET 1
	Contact:
Boundary Conditions	1 - CONT
	Global contact search distance lower bound : -1
	Global contact search distance upper bound : 1
	Global friction : OFF
	Load:
	1 - LOAD SET 1
Model/Analysis Type	Structural / Static
	1 - B.C. 1,DISPLACEMENT_1,LOAD SET 1
	2 - B.C. 1,REACTION FORCE_2,LOAD SET 1
	3 - B.C. 1,STRESS_3,LOAD SET 1
Results	4 - B.C. 1,STRAIN ENERGY_4,LOAD SET 1
	5 - B.C. 1,CONTACT STRESS_5,LOAD SET 1
	6 - B.C. 1,CONTACT FRICTION STRESS_6,LOAD S
	7 - B.C. 1,CONTACT PRESSURE_7,LOAD SET 1

move in vertical direction only. In vertical direction on the inner surface of a shaft there was imposed a load. The purpose of computational models elaboration was estimating of spigot joint stress state for assumed dimensional tolerances. There were evaluated the influence of a hole and a shaft diameters on contact stresses distribution for boundary tolerations dimensions at assumed constant load conditions. Smaller diameter of inner ring was loaded with pressure 80 MPa.

3 Results of Computer Simulation

As the result of computer simulation there obtained coloured charts of stresses distribution acc. to Huber-Misses-Hencky Hypothesis and distribution of distortions and values of restrain forces, pressures and contact stresses. For further analysis there were used results and contact stresses charts (Fig. 2). For results evaluation facilitation all contact stresses charts were presented in equal scale (460 MPa). Results analyses were done and there were established maximal and average values of contact stresses for particular values of diameters. Also there was proposed parameter Ne determining quotient of general nodes quantity to quantity of nodes having non-zero values of stresses.

Table 2. Stress results for parameter Ne

		Connection[mm]			
Chosen results		Hole ϕ38,105, Shaft ϕ38,015	Hole ϕ38,105, Shaft ϕ38,025	Hole ϕ38,075, Shaft ϕ38,015	Hole ϕ38,075, Shaft ϕ38,025
Max. values of contact stresses [MPa]		459,1	451,3	430,9	416,4
Average value of contact stresses [MPa]		52,41	52,56	51,56	51,85
Value of parameter Ne		0,1864	0,1921	0,1957	0,2027
Number of nodes	above 400 MPa	344	271	110	64
with values of	above 300 MPa	608	698	836	671
contact stresses	above 200 MPa	511	553	618	898
properly grouped	above 100 MPa	428	415	420	409
	Up to 100 MPa	144	161	154	172

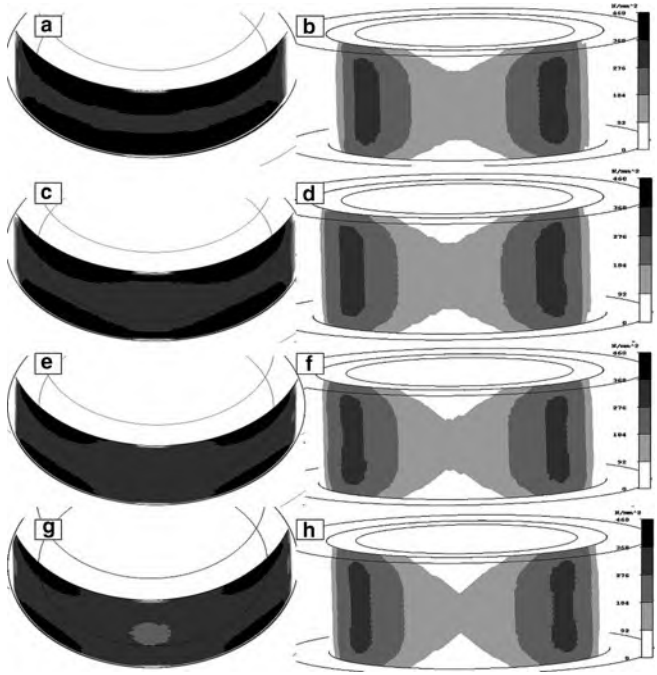


Fig. 2. Top and bottom views of results of contact stresses distributions of pinjoint: (**a**,**b**) for deviations, hole ϕ 38,105, shaft ϕ 38,015; (**c**,**d**) for deviations, hole ϕ 38,105, shaft ϕ 38,025; (**e**,**f**) for deviations, hole ϕ 38,075, shaft ϕ 38,015; (**g**,**h**) for deviations, hole ϕ 38,075, shaft ϕ 38,025

This parameter correlates to contact surface (Table 2). For assumed changes of diameters values there occurred 10% increase of maximal value of contact stresses, 2% increase of mean value, but also the change of contact stresses distribution character that could be seen in particular charts. For assumed method of load there is changing a contact cooperation from outer edges in models of bigger values of clearances to more complete cooperation in middle zone for smaller values of clearance.

4 Conclusions

Carried out analysis has shown the purposefulness of usage of finite elements method and special contact elements as a tool enabling for evaluation of contact stresses values. Made set of computer simulations enabled for evaluation of cooperation of important machine connections. Considerations were focused on the influence of dimensional tolerance of particular connection elements on contact parameters. There evaluated the changing of maximal and mean values of contact stresses.

Nodes of non-zero contact stresses values were grouped. From this grouping arises conclusion that there are changing quantities of nodes in particular groups with tendency to decreasing of maximal values and increasing of nodes quantities in ranges 200–400 MPa for decreasing value of connection clearance. With the decreasing of clearance value there increased the quantity of non-zero nodes increases of contact surface. There also drops maximal value of contact stresses.

In the future there should be enlarged the range of analysis on simulation of connection elements cylindricity because these factors additionally decrease real contact surface.

References

1. Dudziak, M., Podolski, T., Kołodziej, A.: Wybrane problemy wykorzystania elementów kontaktowych w modelowaniu połaczeń osiowosymetrycznych. XV Konferencja nt. Metody i Środki Projektowania Wspomaganego Komputerowo, Kazimierz Dolny, 2005
2. Podolski, T.: Walidacja stanu naprężeń kontaktowych w połaczeniach osiowosymetrycznych. I Miedzyuczelniane Seminarium Studenckich Kół Naukowych i Studiów Doktoranckich, 2006
3. IDEAS HELP, instruction manuals (e-book) edited by SDRC

Fractional Cauchy Problem with Applications to Anomalous Diffusion

E. Popescu

Technical University of Civil Engineering, Bd. Lacul Tei 124, 38RO-020396 Bucharest, Romania epopescu@utcb.ro

Summary. Starting from the Cauchy problem associated with a Feller semigroup, some expressions of solutions of the fractional Cauchy problem are presented. The fractional Cauchy problem is applied in physics for modeling anomalous diffusion, in which particles spread slower than is predicted by the classical diffusion model.

1 Introduction

An anomalous diffusion is the phenomenon, met in disordered or fractal media, according to which the displacement variance is no longer linear in time but proportional to a power α of time with $0 < \alpha < 2$. The particles spread in a different manner than the prediction of the classical diffusion equation. A known model for an anomalous diffusion is the fractional diffusion equation, where the usual second derivative in space is replaced by a fractional derivative of order α, $0 < \alpha < 2$,

$$\frac{\partial u}{\partial t}(x,t) = D\frac{\partial^\alpha u}{\partial x^\alpha}(x,t), \quad u(x,0) = f(x).$$

We can extend this equation to the fractional Cauchy problem

$$\frac{\partial^\beta u}{\partial t^\beta}(x,t) = \left(Au(\cdot,t)\right)(x), \quad u(x,0) = f(x),$$

where A is a pseudodifferential operator. In [1] was shown that the solution of this problem can be expressed as an integral transform of the solution to the usual Cauchy problem. In this paper, starting from this integral transform we give some formulae for the solution of the fractional Cauchy problem.

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*, Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_157,

2 Integral Representation of the Operators which form a Feller Semigroup

Let $(X, \|\cdot\|)$ be a Banach space. $\{T(t)\}_{t\geq 0}$ is a strongly continuous semigroup on X. $u(t) = T(t)f$ solves the abstract Cauchy problem

$$\frac{d}{dt}u(t) = Au(t),\ u(0) = f,$$

for $f \in D(A)$.

A continuous negative definite function a is described by Lévy-Khinchin formula

$$a(\xi) = c + ib\cdot\xi + q(\xi) + \int\limits_{\mathbf{R}^n\setminus\{0\}} \left[1 - e^{-i\xi\cdot y} - i\frac{\xi\cdot y}{1+|y|^2}\right]\frac{1+|y|^2}{|y|^2}d\mu(y)$$

with $c \geq 0$, $b \in \mathbf{R}^n$, q a continuous non-negative definite quadratic form on $\mathbf{R}^n$ and μ a non-negative finite measure on $\mathbf{R}^n \setminus \{0\}$.

In the following, we denote by $C_\infty(\mathbf{R}^n)$ the Banach space of all continuous functions on $\mathbf{R}^n$ vanishing at infinity with the supremum norm $\|\cdot\|_\infty$ and by $C_0^\infty(\mathbf{R}^n)$ the set of all C^∞-functions on $\mathbf{R}^n$ with compact support. $S(\mathbf{R}^n)$ will be the Schwartz space, i.e. the set of all functions $\varphi \in C^\infty(\mathbf{R}^n)$ such that $\sup_{x\in\mathbf{R}^n}|x^\beta\partial^\alpha\varphi(x)| < \infty$ for all multi-indices α and β. $S(\mathbf{R}^n)$ is dense in $C_\infty(\mathbf{R}^n)$.

The general form of a *pseudo-differential operator* is

$$p(x,D)\varphi(x) = (2\pi)^{-(n/2)}\int\limits_{\mathbf{R}^n} e^{ix\cdot\xi}p(x,\xi)\widehat{\varphi}(\xi)d\xi,$$

for $\varphi \in C_0^\infty(\mathbf{R}^n)$, where $\overset{\wedge}{\varphi}(\xi) = (2\pi)^{-(n/2)}\int\limits_{\mathbf{R}^n} e^{-ix\cdot\xi}\varphi(x)dx$ is the Fourier transform. $p(x,\xi)$ is called the *symbol* of the operator $p(x,D)$ (see [3]).

Let $A : D(A) \to C_\infty(\mathbf{R}^n)$ be a linear operator, where $D(A)$ is a linear dense subspace of $C_\infty(\mathbf{R}^n)$. A satisfies *the positive maximum principle* on $D(A)$ if for all $u \in D(A)$ and $x_0 \in \mathbf{R}^n$ such that $\sup_{x\in\mathbf{R}^n} u(x) = u(x_0) \geq 0$ it follows that $Au(x_0) \leq 0$.

Theorem 2.1.[2] Let $A : C_0^\infty(\mathbf{R}^n) \to C_b(\mathbf{R}^n)$ be a linear operator satisfying the positive maximum principle. Then

$$Au(x) = -(2\pi)^{-(n/2)}\int_{\mathbf{R}^n} e^{ix\cdot\xi}a(x,\xi)\widehat{u}(\xi)d\xi,$$

where $a : \mathbf{R}^n \times \mathbf{R}^n \to \mathbf{C}$ is a locally bounded function such that for any fixed $x \in \mathbf{R}^n$, $\xi \to a(x,\xi)$ is a continuous negative definite function.

The *convolution semigroup* on $C_\infty(\mathbf{R}^n)$ generated by a is defined by the formula

$$T(t)u(x) = (2\pi)^{-(n/2)} \int_{\mathbf{R}^n} e^{ix\cdot\xi} p_t(\xi)\widehat{u}(\xi)d\xi,$$

for each $t > 0$ and $u \in S(\mathbf{R}^n)$, where $p_t(\xi) = e^{-ta(\xi)}$. In this case, we observe that for any $t > 0$ the symbol is p_t (note that there is no x-dependence). The function $\xi \to p_t(\xi)$ is a positive definite function and the infinitesimal generator of $\{T(t)\}_{t\geq 0}$ is

$$Au(x) = -(2\pi)^{-(n/2)} \int_{\mathbf{R}^n} e^{ix\cdot\xi} a(\xi)\widehat{u}\,(\xi)\,d\xi,$$

for all $u \in C_0^\infty(\mathbf{R}^n)$, $x \in \mathbf{R}^n$.

Let $\{T(t)\}_{t\geq 0}$ be a strongly continuous semigroup on $C_\infty(\mathbf{R}^n)$.

If $\|T(t)u\| \leq \|u\|$ for all $u \in C_\infty(\mathbf{R}^n)$ and $t \geq 0$, then $\{T(t)\}_{t\geq 0}$ is a contraction semigroup. A strongly continuous positive contraction semigroup on $C_\infty(\mathbf{R}^n)$ is called a *Feller semigroup* on $\mathbf{R}^n$. We have an integral representation of the operators which form a Feller semigroup (see [4]).

Theorem 2.2. Let $\{T(t)\}_{t\geq 0}$ be a Feller semigroup on $\mathbf{R}^n$. For any $t \geq 0$ there exists a unique function $p_t : \mathbf{R}^n \times \mathbf{R}^n \to \mathbf{C}$ measurable, locally bounded and such that for any fixed $x \in \mathbf{R}^n$, $\xi \to p_t(x,\xi)$ is a continuous positive definite function with the property that for any $u \in S(\mathbf{R}^n)$,

$$T(t)u(x) = (2\pi)^{-(n/2)} \int_{\mathbf{R}^n} e^{ix\cdot\xi} p_t(x,\xi)\widehat{u}(\xi)d\xi.$$

For $u \in S(\mathbf{R}^n)$, the infinitesimal generator A of $\{T(t)\}_{t\geq 0}$ is

$$Au(x) = (2\pi)^{-(n/2)} \int_{\mathbf{R}^n} e^{ix\cdot\xi} a(x,\xi)\widehat{u}(\xi)d\xi,$$

where

$$a : \mathbf{R}^n \times \mathbf{R}^n \to \mathbf{C}, \quad a(x,\xi) = \frac{d}{dt}\, p_t(x,\xi)\mid_{t=0}.$$

Moreover, we deduce the following result.

Theorem 2.3. Let $\{T(t)\}_{t\geq 0}$ be a Feller semigroup on $\mathbf{R}^n$. For any $t \geq 0$ and $u \in C_b^2(\mathbf{R}^n)$, $T(t)u(x) = C(t)u(x) + D(t)u(x)$, with

$$C(t)u(x) = \sum_{i,j=1}^{n} a_{ij}^{(t)}(x)\frac{\partial^2 u}{\partial x_i \partial x_j}(x) + \sum_{i=1}^{n} b_i^{(t)}(x)\,\frac{\partial u}{\partial x_i}(x) + \gamma^{(t)}(x)u(x)$$

and

$$D(t)u(x) = \int_{\mathbf{R}^n} N^{(t)}(x,dy)\left\{u(y) - \sigma_x^{(t)}(y)\left[u(x) + \sum_{i=1}^{n}\frac{\partial u}{\partial x_i}(x)\cdot(y_i - x_i)\right]\right\},$$

where $\gamma^{(t)}(x) = c^{(t)}(x)+d^{(t)}(x)+1$, $a_{ij}^{(t)}$, $b_i^{(t)}$, $c^{(t)}$, $d^{(t)}$ are continuous functions, $a_{ij}^{(t)} = a_{ji}^{(t)}$, $\sum_{i,j=1}^{n} a_{ij}^{(t)}(x)\xi_i\xi_j \geq 0$, $c^{(t)} \leq 0$, $\sigma_x^{(t)}$ is a certain cut-off function and $N^{(t)}(x, dy)$ is a certain Lévy kernel such that $d^{(t)}(x) + \int_{\mathbf{R}^n} N^{(t)}(x, dy)\{1 - \sigma_x^{(t)}(y)\} \leq 0$, for all $x \in \mathbf{R}^n$.

Proof. Indeed, $T(t) - I$ satisfies the positive maximum principle on $C_\infty(\mathbf{R}^n)$ for every $t \geq 0$. The above formula follows from the last assertion of Lemma 3.3 ([2], pp. 2–34) and Corollary 3 ([2], pp. 2–10). □

3 Fractional Cauchy Problem

We consider the fractional Cauchy problem

$$\frac{\partial^\beta}{\partial t^\beta}u(x,t) = Au(x,t),\ u(x,0) = f(x),$$

where $0 < \beta < 1$, $t \geq 0$ and A is the generator of bounded continuous semigroup $\{T(t)\}_{t\geq 0}$ on the Banach space X. For a function g with $\widetilde{g}(s) := \int_0^\infty e^{-st}g(t)dt$ the Laplace transform, $\frac{\partial^\beta}{\partial t^\beta}g(t)$ is the Caputo fractional derivative in time, which can be defined as inverse Laplace transform of $s^\beta\widetilde{g}(s) - s^{\beta-1}g(0)$. We observe that $p(t,x) = T(t)f(x)$ is the unique solution to the abstract Cauchy problem $\frac{\partial}{\partial t}p(x,t) = Ap(x,t),\ p(x,0) = f(x)$, for any f in the domain of A.

We note that the fractional Cauchy problem can be written in several equivalent forms (see [1]).

Proposition 3.1. Assume $0 < \beta < 1$. Let A be the generator of a strongly continuous semigroup $\{T(t)\}_{t\geq 0}$ on the Banach space X and $g \in C([0,\infty]\times X)$ be Laplace transformable. Then for all $h \in X$ the following are equivalent:

(1) For all $t > 0$, the Riemann-Liouville derivative of g exists, $g(t) \in D(A)$, the Laplace transform of $D_t^\beta g(t)$ exists, and

$$D_t^\beta g(t) = Ag(t) + \frac{t^{-\beta}}{\Gamma(1-\beta)}h.$$

(2) For all $t > 0$, the Caputo derivative of g exists, $g(t) \in D(A)$, the Laplace transform of $\frac{\partial^\beta}{\partial t^\beta}g(t)$ exists, and

$$\frac{\partial^\beta}{\partial t^\beta}g(t) = Ag(t),\ g(0) = h.$$

(3) For all $t > 0$, the function g is differentiable, $g(t) \in D(A)$, the Laplace transform of $\frac{\partial}{\partial t}g(t)$ exists, and

$$\frac{\partial}{\partial t}g(t) = D_t^{1-\beta}Ag(t),\ g(0) = h.$$

(4) The function $g(t)$ is analytic on $0 < t < \infty$, satisfies $\|g(t)\| \leq Me^{\omega t}$ on $0 < t < \infty$ for some $M, \omega \geq 0$ and

$$g(t) = \int_0^\infty \frac{t}{\beta s^{1+1/\beta}} g_\beta\left(\frac{t}{s^{1/\beta}}\right) T(s)\, h ds,$$

where g_β is such that $\int_0^\infty e^{-\lambda t} g_\beta(t)dt = e^{-\lambda^\beta}$.

In the above proposition $D_t^\beta g(t) = \frac{d^m}{dt^m}\int_0^t \frac{(t-u)^{m-\beta-1}}{\Gamma(m-\beta)} g(u)du$, $m = [\beta]$, is the Riemann-Liouville fractional derivative of order β. On the other hand,

$$\frac{\partial^\beta}{\partial t^\beta}g(t) = \int_0^t \frac{(t-u)^{m-\beta-1}}{\Gamma(m-\beta)} g^{(m)}(u)\, du,\ m = [\beta].$$

If β is a positive integer then $D_t^\beta = \frac{\partial^\beta}{\partial t^\beta}$ is the usual derivative operator.

In the framework of Proposition 3.1, we define the family of bounded, strongly continuous linear operators on X,

$$S(t)h(x) := \int_0^\infty \frac{t}{\beta s^{1+1/\beta}} g_\beta\left(\frac{t}{s^{1/\beta}}\right) T(s)\, h(x) ds, \qquad t \geq 0.$$

In view of (4) the function $g(t) = S(t)h$ defines a solution to the fractional Cauchy problem for any initial condition $h \in X$ and this solution depends continuously on the initial condition h.

In the sequel we consider $X = C_\infty(\mathbf{R}^n)$.

Proposition 3.2. Let A be the generator of a Feller semigroup $\{T(t)\}_{t\geq 0}$ on $\mathbf{R}^n$. Then for any $t \geq 0$ and $h \in S(\mathbf{R}^n)$,

$$S(t)h(x) = (2\pi)^{-n/2} \int_0^\infty \int_{\mathbf{R}^n} \int_{\mathbf{R}^n} \frac{t}{\beta s^{1+1/\beta}} g_\beta\left(\frac{t}{s^{1/\beta}}\right) e^{i(x-y)\cdot\xi} p_s(x,\xi) h(y) dy d\xi ds,$$

where $p_s : \mathbf{R}^n \times \mathbf{R}^n \to \mathbf{C}$ is measurable, locally bounded and such that for any fixed $x \in \mathbf{R}^n$, $\xi \to p_s(x,\xi)$ is a continuous positive definite function.

If $k_s(x,y) := (2\pi)^{-n/2} \int_{\mathbf{R}^n} e^{i(x-y)\cdot\xi} p_s(x,\xi) d\xi$ is finite, then

$$S(t)h(x) = \int_0^\infty \int_{\mathbf{R}^n} \frac{t}{\beta s^{1+1/\beta}} g_\beta\left(\frac{t}{s^{1/\beta}}\right) k_s(x,y)\, h(y) dy ds, \qquad t \geq 0.$$

Proof. In the formula of the definition of $S(t)h(x)$ we apply Theorem 2.2 for the semigroup $\{T(t)\}_{t\geq 0}$ on $\mathbf{R}^n$. □

Remark 3.3. Using the relation from Theorem 2.3, we obtain the "structure" of each operator $S(t)$, $t \geq 0$. Since $T(t)h(x) = C(t)h(x) + D(t)h(x)$, we have

$$S(t)h(x) = \int_0^\infty \frac{t}{\beta s^{1+1/\beta}} g_\beta\left(\frac{t}{s^{1/\beta}}\right) C(s)\, h(x) ds + \\ + \int_0^\infty \frac{t}{\beta s^{1+1/\beta}} g_\beta\left(\frac{t}{s^{1/\beta}}\right) D(s) h(x) ds.$$

Thus we can interpret the solution $g(t) = S(t)h$ of the fractional Cauchy problem for the initial condition h as the sum of "diffusion part" and "Lévy part".

Example 3.4. We consider an anomalous diffusion given by the equation

$$\frac{\partial^\beta u}{\partial t^\beta}(x,t) = D\Delta u(x,t), \qquad u(x,0) = h(x), \qquad x \in \mathbf{R}^n,\ t \geq 0.$$

where $\Delta = \sum_{i=1}^n \dfrac{\partial^2}{\partial x_i^2}$ and D is a constant.

We denote with $|\cdot|$ the norm in $\mathbf{R}^n$. For $p_s(x,\xi) := e^{-D|\xi|^2 s}$, we have $k_s(x,y) = \frac{1}{(2\sqrt{\pi D s})^n} e^{-\frac{|y-x|^2}{4sD}}$ and we deduce

$$S(t)h(x) = \frac{t}{\beta\left(2\sqrt{\pi D}\right)^n} \int_0^\infty \int_{\mathbf{R}^n} \frac{1}{s^{1+1/\beta+n/2}} g_\beta\left(\frac{t}{s^{1/\beta}}\right) e^{-\frac{|y-x|^2}{4sD}} h(y) dy ds.$$

Example 3.5. For the equation

$$\frac{\partial^\beta u}{\partial t^\beta}(x,t) = \frac{1}{2}(\Delta - x \cdot grad) u(x,t), \qquad u(x,0) = h(x), \qquad x \in \mathbf{R}^n,\ t \geq 0,$$

we have $p_s(x,\xi) = e^{i(e^{-s/2}-1)x\cdot\xi - (1-e^{-s})\cdot|\xi|^2/2}$.

Indeed, the Ornstein-Uhlenbeck semigroup is defined, for each $t > 0$, by

$$U_t f(x) = \int_{\mathbf{R}^n} f(xe^{-t/2} + y\sqrt{1-e^{-t}})\mu(dy),$$

where μ is the Gaussian measure on $\mathbf{R}^n$, whose Fourier transform is $\widehat{\mu}(u) = e^{-|u|^2/2}$. On $S(\mathbf{R}^n)$,

$$U_t f(x) = (2\pi)^{-n/2} \int_{\mathbf{R}^n} e^{ix\cdot\xi} p_t(x,\xi) \widehat{f}(\xi) d\xi,$$

with $p_s(x,\xi)$ as above. Then $S(t)h(x)$ is equal with

$$\frac{1}{(2\pi)^n} \int_0^\infty \int_{\mathbf{R}^n} \int_{\mathbf{R}^n} \frac{t}{\beta s^{1+1/\beta}} g_\beta\left(\frac{t}{s^{1/\beta}}\right) e^{-iy\cdot\xi + ie^{-s/2}x\cdot\xi - (1-e^{-s})\cdot|\xi|^2/2} h(y) dy d\xi ds.$$

References

1. Baeumer, B., Kurita, S., Meerschaert, M.M.: Fractional Calculus Appl. Anal. **8**, 371–386 (2005)
2. Courrège, P.: Sém. Théorie du Potentiel, 38p. (1965/1966)
3. Jacob, N.: Pseudo-Differential Operators and Markov Processes. Mathematical Research, vol. 94. Akademie Verlag, Berlin (1996)
4. Popescu, E.: Potential Analysis, vol. 14, pp. 207–209. (2001)

Multi-scale Modeling of the Interplanetary Magnetic Field

N.A. Popescu[1] and E. Popescu[2]

[1] Astronomical Institute of Romanian Academy, Cutitul de Argint 5, Bucharest, Romania nedelia@aira.astro.ro
[2] Technical University of Civil Engineering, Bd. Lacul Tei 124, 38RO-020396 Bucharest, Romania epopescu@utcb.ro

Summary. Models for heavy-tailed data with applications to the study of multiscale behaviour of the interplanetary magnetic field are presented. Numerical aspects are given in the case of the data obtained by Ulysses mission (magnetometer VHM/FGM). This approach yields probabilistic predictions of the dynamics and multiscale behaviour of the interplanetary magnetic field.

1 Models for Heavy-Tailed Data

The study of statistical properties of the interplanetary magnetic field fluctuations represents an important topic in space research. These fluctuations are related to acceleration processes and energy transport in the solar wind, and can provide an important insight into the solar wind turbulent cascade.

The fractional diffusion equations, which are used to represent the complexity, provide a suitable mathematical framework for the multiscale behavior. A space-time fractional diffusion equation is obtained from the standard diffusion equation by replacing the second order space-derivative by a fractional Riesz derivative of order $\alpha > 0$ and skewness θ, and the first order time-derivative by a fractional derivative of order $\beta > 0$ in Caputo or Riemann–Liouville sense. In the cases $0 < \alpha < 2$, $\beta = 1$ or $\alpha = 2$, $0 < \beta < 2$ or $0 < \alpha = \beta \leq 2$, the fundamental solution (or Green's function) $F_{\alpha,\beta}(x,t)$ of the equation can be interpreted as a spatial probability density function (PDF) evolving in time. It is well known that for the standard diffusion ($\alpha = 2$, $\beta = 1$) the Green's function is the Gaussian PDF. The scaling property of the Green's function allows to express it in terms of a function of a single variable, the reduced Green's function $R_{\alpha,\beta}(x)$ (see [2]). We observe that $R_{\alpha,1}(x)$ are the stable distributions. Some computational forms of $R_{\alpha,\beta}(x)$ are as follows:

(a) if $\alpha = \beta$ then

$$R_{\alpha,\alpha}(x) = \frac{1}{\pi x} \sum_{n=0}^{\infty} \left(-x^{\alpha}\right)^{n} \sin\left[\frac{n\pi}{2}(\theta - \alpha)\right], \; 0 < x < 1 \tag{1}$$

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*, Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_158,

and

$$R_{\alpha,\alpha}(x) = \frac{1}{\pi x}\sum_{n=0}^{\infty}(-x^{-\alpha})^n \sin\left[\frac{n\pi}{2}(\theta-\alpha)\right], x > 1; \quad (2)$$

(b) for $\alpha < \beta$ we have

$$R_{\alpha,\beta}(x) = \frac{1}{\pi x}\sum_{n=1}^{\infty}(-x^{-\alpha})^n \frac{\Gamma(1+n\alpha)}{\Gamma(1+n\beta)} \sin\left[\frac{n\pi}{2}(\theta-\alpha)\right], x > 0; \quad (3)$$

A purpose of this study is to show that the PDF's are well adapted to model the random characteristic of the interplanetary magnetic field in different cases of configurations.

2 Statistical Scaling Properties of the Magnetic Field Intensity Fluctuations

Using the interplanetary magnetic field data provided by the Ulysses mission (magnetometer VHM/FGM) for 3 years 2002–2004, we analyze the changes in the magnetic field intensity, $B(t)$, at different scales. In this interval of time Ulysses mission, being at the beginning of its third solar orbit, was situated at heliocentric distances between 2.5859 and 5.4044 AU (the maximum heliocentric distance reached by the mission). At this distance, Ulysses obtained information on the Jupiter's magnetosphere. Also, we consider the Ulysses Solar Wind Plasma Investigation data (SWOOPS) for analyzing the distribution of proton number density and solar wind velocity for the years 2002–2004. In Figs. 1 and 2 are presented the magnetic field intensity profile (top panel) and solar wind velocity profile bottom panel) during 2002–2003, respectively 2003–2004 (a total of 1,096 days of data recorded by Ulysses).

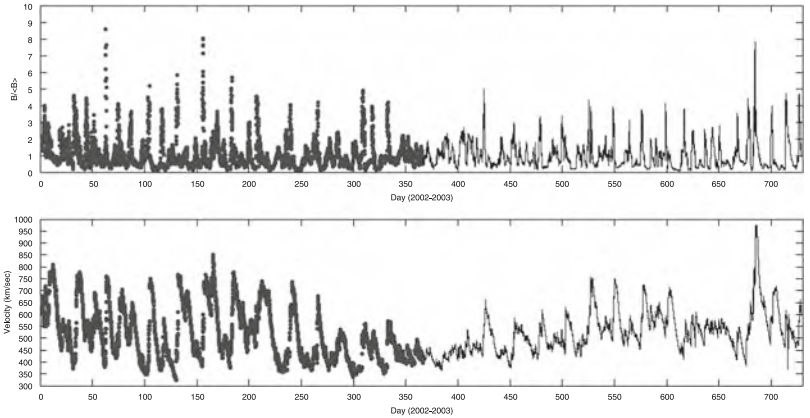


Fig. 1. Time series of magnetic field intensity B (*top panel*); solar wind velocity v (*bottom panel*) during 2002–2003 years of data recorded by Ulysses

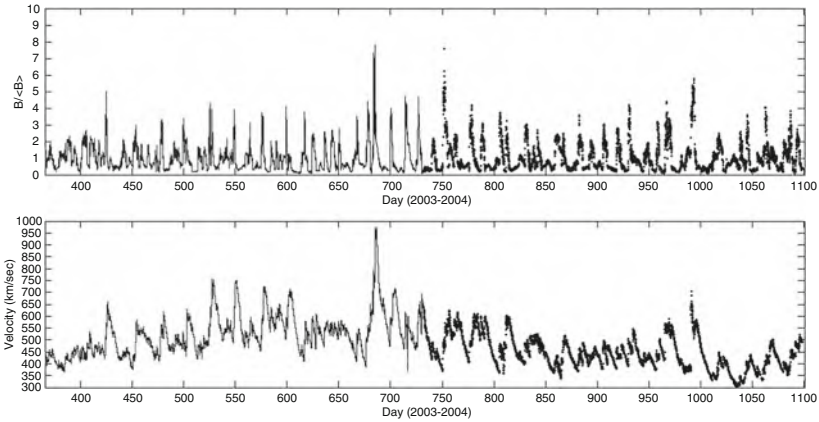


Fig. 2. Time series of magnetic field intensity B (*top panel*); solar wind velocity v (*bottom panel*) during 2003–2004 years of data recorded by Ulysses

2.1 Finite Size Scaling Technique

We apply the finite size scaling technique on Ulysses data in order to study the scaling and intermittency of the magnetic field intensity. Although intermittency refers to the statistical behavior of the fluctuations in the spatial domain, time differences are equivalent to space differences when the Taylor hypothesis is valid (see [3]) – i.e. a turbulent structure transits the space craft at a time which is small in comparison with its own evolution.

In our study the considered technique is based on differencing of the original time series over a range of temporal scales τ. The fluctuations on temporale scale τ can be captured by a set of differences $dS(t,\tau) = S(t+\tau) - S(t)$, where $S(t)$ represents a given time series (see [1]). The basic quantity considered in this section is the change in the magnetic field intensity B, at different scales (time lags $\tau_n = 2^n$ days, $n = 0, 1, 2, \ldots$).

First step in our calculation is represented by the determination of magnetic field intensity increments at a given scale τ_n through:

$$dB_n = dB_n(t_i, \tau_n) = [B(t_i + \tau_n) - B(t_i)]\,, \tag{4}$$

where t_i is the time (days of the year); $B(t_i)$ is the daily average of B.

The second step consists in the normalization of these quantities (which represent characteristic fluctuations across eddies at the scales τ_n) to their variance, σ^2, obtaining data sets of normalized fluctuations of the magnetic field intensity.

2.2 Results: Magnetic Field Fluctuation Analysis

Figure 3 presents the magnetic field strength signal observed by Ulysses during 2002–2004 at heliocentric distances between 2.5859 and 5.4044 AU. We

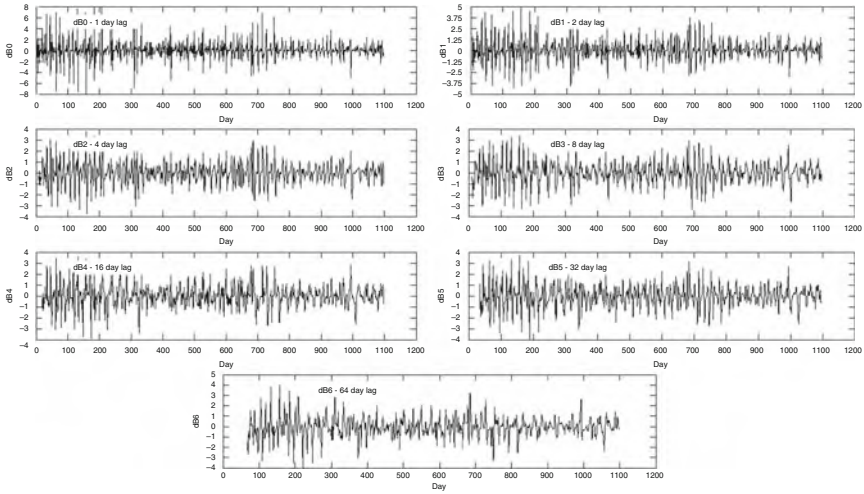


Fig. 3. The determined dB_n for the scales $n = 0, \ldots, 6$

considered scales ranging from $\tau_0 = 1$ day to $\tau_6 = 2^6 = 64$ days, and obtained seven normalized data sets $dB_n(t_i, \tau_n)$.

At scales of 1, 2, and 4 days, the fluctuations $dB_0(t)$, $dB_1(t)$ and $dB_2(t)$ are intermittent with spikes (pulse of extremely short duration), superposed on a signal with bursts (abrupt increase in the amplitude of the signal) of fluctuations, especially for the year 2002 (days 1–365). At larger scales the fluctuations are less intermittent (fewer bursts and spikes).

We have to mention that the intermittency (which is connected with sudden occurrence of large amplitude variations of magnetic field intensity) is usually pointed out as a departure of the PDFs from a Gaussian distribution. The intermittency is the triggering process for increased probabilities of large amplitude fluctuations at smaller scales.

Magnetic field intensity fluctuations at different time scales are quantitatively described by the PDFs. These are represented by the normalized histograms of $dB_n(t_i, \tau_n)$. In Fig. 4, from top to bottom, are presented the PDFs (denoted fdB_n) of the normalized fluctuations of the magnetic field intensity, dB_n, measured by Ulysses, on time scales of 1, 2, 4, 8, 16, 32 and 64 days. In order to reveal the shape of fdB_n, these are separated by 0.1 differences in Fig. 4. The presence of heavy tails is obvious for all time scales.

Figure 5 presents the observational (points) and theoretical PDFs (continuous lines) for 1, 8, 32 and 64 days time scales. The theoretical PDF fdB_0 is obtained for $\alpha = 0.35$, $\beta = 0.55$, $\theta = -0.35$ with formula (3). The theoretical PDFs fdB_3 and fdB_5 are obtained for $\alpha = \beta \leq 0.2$ using the formulae (1) and (2). In order to point out the heavy tails, the logarithmic scale is considered. At large scales the PDFs are almost Gaussian, and the tails of the distributions grow up as the scale becomes smaller. The presence of intermittency

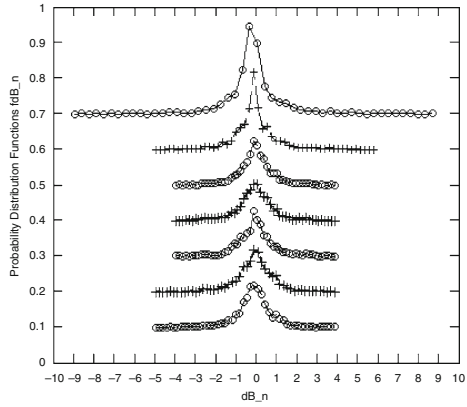


Fig. 4. The observational PDFs – fdB_n – for the scales $n = 0, 1, 2, \ldots, 6$

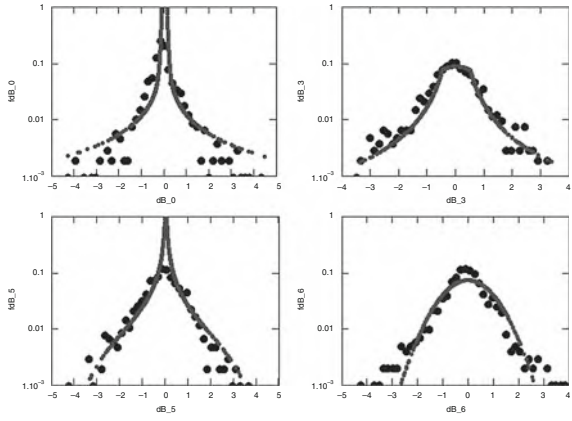


Fig. 5. The observational (points) and theoretical PDFs – fdB_n – for the scales $n = 0, 3, 5, 6$

is obvious for fdB_0 and fdB_3), where the PDFs of fluctuations increasingly depart from a Gaussian distribution with decreasing temporal scale.

2.3 Proton Number Density Distribution

The proton number density distribution can be fitted with the formula (3) from the previous subsection, for $\alpha = 0.4$, $\beta = 1$. In Fig. 6 are presented the observed proton number density distribution (histogram), and the representation of the stable distribution for data recorded by Ulysses.

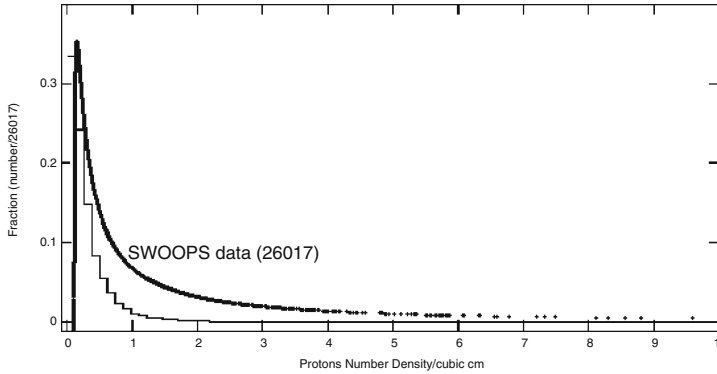


Fig. 6. Proton number density distribution for 26017 data in the interval 2002–2004

3 Conclusions

Analysis of the probability distribution functions of the magnetic field intensity fluctuations has underlined their non-Gaussian properties on small time scales, and uncorrelated features at large scale. Numerical solutions of space-time fractional diffusion equations have been used to analyze the presence or absence of heavy tails, typically associated with multiscale behaviour, in the case of the interplanetary magnetic field data obtained by Ulysses mission (for the years 2002–2004). At larger scales the fluctuations are less intermittent than at small scales, where the fluctuations present spikes and bursts indicating intermittency. The changes of B at different scales have been studied by means of PDFs, good fits of the observational PDFs being obtained. In this mode, probabilistic predictions can be done for the dynamics and multiscale behaviour of the interplanetary magnetic field.

Acknowledgements

A. Balogh from the National Space Science Data Center and D.J. McComas from SWRI, for providing the VHM/FGM data, respectively SWOOPS data.

References

1. Frisch, U.: Turbulence. The legacy of A.N. Kolmogorov Cambridge University Press, Cambridge (1995)
2. Mainardi, F., Luchko, Y., Pagnini, G.: Fractional Calculus Appl. Anal. **4**, 153–192 (2001)
3. Taylor, G.I.: Proc. R. Soc.Lon.Ser-A **164**, 476 (1938)

Analytical and Numerical Modelling of Thermoviscous Shocks and Their Interactions in Nonlinear Fluids Including Dissipation

A.R. Rasmussen[1], M.P. Sørensen[1], Yu.B. Gaididei[2], and P.L. Christiansen[3]

[1] Department of Mathematics, Technical University of Denmark, DK-2800 Kongens Lyngby, Denmark, A.R.Rasmussen@mat.dtu.dk, m.p.soerensen@mat.dtu.dk
[2] Bogolyubov Institute for Theoretical Physics, 03680 Kiev, Ukraine, ybg@bitp.kiev.ua
[3] Department of Informatics and Mathematical Modelling and Department of Physics, Technical University of Denmark, DK-2800 Kongens Lyngby, Denmark, plc@imm.dtu.dk

Summary. A wave equation, that governs finite amplitude acoustic disturbances in a thermoviscous Newtonian fluid, and includes nonlinear terms up to second order, is proposed. The equation preserves the Hamiltonian structure of the fundamental fluid dynamical equations in the non-dissipative limit. An exact thermoviscous shock solution is derived. This solution is, in an overall sense, equivalent to the Taylor shock solution of the Burgers equation. However, in contrast to the Burgers equation, the model equation considered here is capable to describe waves propagating in opposite directions. Studies of head-on colliding thermoviscous shocks demonstrate that the propagation speed changes upon collision.

1 Introduction

The "classical" equation of nonlinear acoustics, the so-called Kuznetsov equation [7], governs finite amplitude acoustic disturbances in a Newtonian, homogeneous, viscous, and heat conducting fluid. This equation arises in the modelling of biomedical ultrasound [5] and modelling of jet engines [2], to mention a few examples. The derivations of the Kuznetsov equation [7] and related model equations [8] are based on the complete system of the equations of fluid dynamics. It has been demonstrated that this system of equations is of Hamiltonian structure in the absence of dissipation [9]. However, in the non-dissipative limit, the Kuznetsov equation does not retain the Hamiltonian structure.

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*, Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_159,

In this paper we propose a nonlinear wave equation, which, in the non-dissipative limit, preserves the Hamiltonian structure of the fundamental equations. We present the derivation and analysis of an exact thermoviscous shock solution. The derivation of the exact solution is based on a *generalized* travelling wave assumption, which leads to a wider class of exact solutions compared to the one reported by Jordan [6] for the Kuznetsov equation. Furthermore, the introduction of the generalized assumption is necessary in order to interpret the results of numerical simulations of head-on colliding thermoviscous shocks presented in this paper.

2 Nonlinear Wave Equations

Equations governing finite amplitude acoustic disturbances in a Newtonian, homogeneous, viscous and heat conducting fluid may be derived from four equations of fluid dynamics. Namely, the equation of motion, the equation of continuity, the heat transfer equation and an equation of state.

To obtain a nonlinear wave equation all dependent variables except one are eliminated from this system of equations, resulting in a nonlinear wave equation for that single variable. Retaining nonlinear terms up to the second order, we obtained a nonlinear wave equation, which we write here for the case of one-dimensional plane fields

$$\psi_{tt} - c_0^2\psi_{xx} = \psi_t\psi_{xx} + \frac{\partial}{\partial t}\left(b\psi_{xx} + (\psi_x)^2 + \frac{B/A-1}{2c_0^2}(\psi_t)^2\right). \tag{1}$$

From the velocity potential $\psi = \psi(x,t)$ one can obtain the fluid particle velocity as $u = -\psi_x$ and the acoustic pressure as $p \approx \psi_t$. The parameter b is the diffusivity of sound (or thermoviscous dissipation parameter) [4], c_0 is the small-signal sound speed, and B/A is the fluid nonlinearity parameter [1]. In the first order approximation (1) reduces to $\psi_{tt} = c_0^2\psi_{xx}$. Introducing this in the first term on the right hand side of (1), the Kuznetsov equation [7]

$$\psi_{tt} - c_0^2\psi_{xx} = \frac{\partial}{\partial t}\left(b\psi_{xx} + (\psi_x)^2 + \frac{B/A}{2c_0^2}(\psi_t)^2\right), \tag{2}$$

is obtained.

The Euler equations of fluid dynamics possess Hamiltonian structure [9]. This property is, however, *not* retained in (2) with $b = 0$, i.e. the non-dissipative limit of the Kuznetsov equation is not Hamiltonian. In contrast, (1) *does* retain the Hamiltonian structure in the non-dissipative limit. Accordingly, the equation may be derived from the Lagrangian density

$$\mathcal{L} = \frac{(\psi_t)^2}{2} - c_0^2\frac{(\psi_x)^2}{2} - \frac{B/A-1}{6c_0^2}(\psi_t)^3 - \frac{\psi_t(\psi_x)^2}{2}, \tag{3}$$

using the Euler–Lagrange equation. Using the Legendre transformation the corresponding Hamiltonian density can be obtained

$$\mathcal{H} = \frac{\partial \mathcal{L}}{\partial \psi_t}\,\psi_t - \mathcal{L}. \tag{4}$$

3 Exact Thermoviscous Shock Solution

Recently, a standard travelling wave approach was applied to the one-dimensional approximation of the Kuznetsov equation (2) to reveal an exact travelling wave solution [6]. In this section we extend the standard approach by introducing the following generalized travelling wave assumption

$$\begin{aligned} \psi(x,t) &= \Psi(x - vt) - \lambda x + \sigma t \\ &\equiv \Psi(\xi) - \lambda x + \sigma t, \end{aligned} \tag{5}$$

where λ and σ are arbitrary constants, v denotes the wave propagation velocity, and $\xi \equiv x - vt$ is a wave variable. The inclusion of $-\lambda x + \sigma t$ in (5) leads to a wider class of exact solutions, compared to the one obtained from the assumption $\psi = \Psi(x - vt)$, which is the standard one. Furthermore, the introduction of the generalized assumption is necessary in order to interpret the results of numerical simulations of head-on colliding thermoviscous shocks presented in Sect. 4. Inserting (5) into the nonlinear wave equation (1), integrating once and introducing $\Phi \equiv -\Psi'$ we obtain the ordinary differential equation

$$\begin{aligned} C = {} & vb\Phi' - \left(\frac{3}{2} + \frac{B/A - 1}{2c_0^2}v^2\right) v\Phi^2 \\ & + \left\{\left(1 - \frac{B/A - 1}{c_0^2}\sigma\right) v^2 - 2\lambda v - c_0^2 - \sigma\right\} \Phi, \end{aligned} \tag{6}$$

where prime denotes differentiation with respect to ξ and C is a constant of integration. Requiring that the solution satisfy $\Phi' \to 0$ as $\xi \to \pm\infty$, and either

$$\Phi \to \begin{cases} \theta, \xi \to +\infty \\ 0, \xi \to -\infty \end{cases} \quad \text{or} \quad \Phi \to \begin{cases} 0, \xi \to +\infty \\ \theta, \xi \to -\infty \end{cases}, \tag{7}$$

where θ is an arbitrary constant, lead us to $C = 0$ and

$$\frac{B/A - 1}{2c_0^2}\theta v^3 - \left(1 - \frac{B/A - 1}{c_0^2}\sigma\right) v^2 + \left(\frac{3}{2}\theta + 2\lambda\right) v + c_0^2 + \sigma = 0. \tag{8}$$

In order to obtain our travelling wave solution we solve (6) subject to $C = 0$ by separation of variables, and by invoking (8) we find the solution to be

$$\Phi = \frac{\theta}{2}\left\{1 - \tanh\left(\frac{2\,(\xi - x_0)}{l}\right)\right\}, \tag{9}$$

$$l \equiv \frac{4b}{\left(\dfrac{B/A - 1}{2c_0^2}v^2 + \dfrac{3}{2}\right)\theta}, \tag{10}$$

where x_0 is an integration constant, $|l|$ is the shock thickness, and $0 < \Phi < \theta$. Finally, using $\Phi = -\Psi'$ and inserting (9) into (5) we obtain (apart from an arbitrary constant of integration)

$$\psi(x,t) = -\frac{\theta}{2}\left\{\xi - \frac{l}{2}\ln\left(\cosh\frac{2(\xi - x_0)}{l}\right)\right\} - \lambda x + \sigma t, \tag{11}$$

which is the exact solution for the velocity potential.

Travelling tanh solutions, such as the solution (9), are often called Taylor shocks or thermoviscous shocks. The existence of an exact solution of this type to the classical Burgers equation is a well known result [3]. However, the Burgers equation is restricted to wave propagation *either* to the left *or* to the right. The model equation considered in this paper does not suffer from this limitation, as shall be illustrated in Sect. 4.

Taking the partial derivatives of (11), we find that the fluid particle velocity, $u = -\psi_x$, and the acoustic pressure, $p \approx -\psi_t$, are given by

$$-\psi_x = \Phi + \lambda \quad \text{and} \quad \psi_t = v\Phi + \sigma. \tag{12}$$

Note that, according to (7) and (12), the asymptotic boundary conditions for $-\psi_x$ and ψ_t are determined by v, θ, λ, and σ, which must satisfy (8).

4 Head-on Colliding Thermoviscous Shocks

The numerical simulation presented in Fig. 1 shows the result of a head-on collision between two thermoviscous shocks.[1] From the simulation we observe that two new waves emerge upon the collision. The contour plot reveals that these travel at a higher speed, compared to the speed of the waves before the collision. For other choices of initial condition, we found the outcome of similar head-on collisions to be two thermoviscous shocks travelling at lower speed, compared to that before the collision.

In order to analyze solutions of (1) that comprise two thermoviscous shocks, we assume that each of these belong to the class of exact solutions derived in Sect. 3. Investigations of the thermoviscous shocks that emerge upon a head-on collision have made it clear that this assumption is true, only when the generalized travelling wave assumption is used, in contrast to the standard travelling wave assumption. For each of the two thermoviscous shocks in the solution we introduce four new parameters, v, θ, λ, and σ, which must satisfy (8) as

$$\begin{aligned}\frac{B/A-1}{2}\theta_i v_i^3 - (1-(B/A-1)\sigma_i)\,v_i^2 \\ + \left(\frac{3}{2}\theta_i + 2\lambda_i\right)v_i + \sigma_i + 1 = 0, \quad i = 1,2,\end{aligned} \tag{13}$$

[1] Non-dimensional variables $\tilde{x} = c_0 x/b$, $\tilde{t} = c_0^2 t/b$, and $\tilde{\psi}(\tilde{x},\tilde{t}) = \psi(x,t)/b$ were introduced prior to the numerical computation.

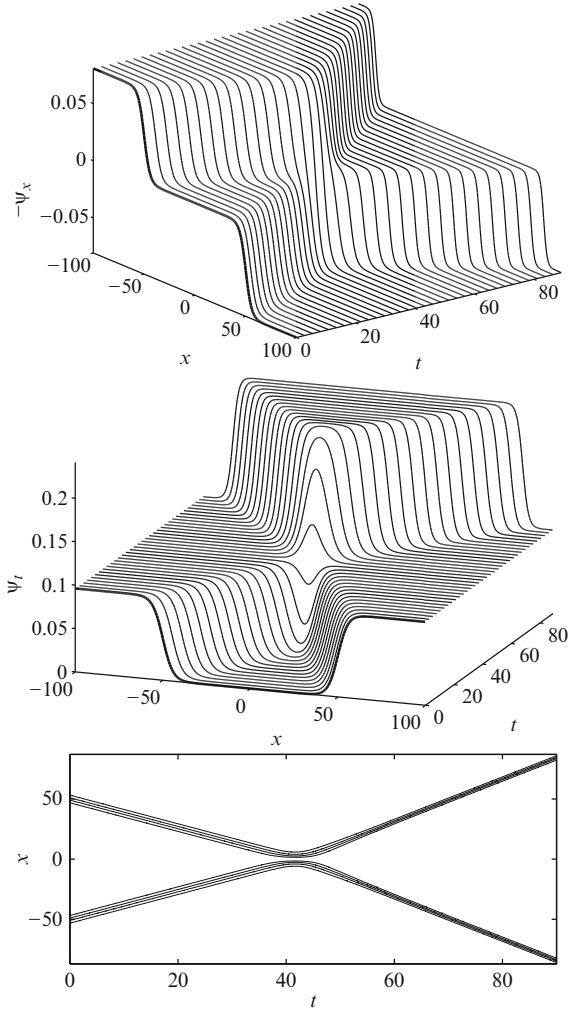


Fig. 1. The initial condition (*bold lines* in the two *topmost* plots) corresponds to two thermoviscous shocks that travel at $v = \pm 1.19$ and make a head-on collision at $t = 42$. The nonlinearity parameter was set to $B/A = 5$. *Lowermost*: contour lines given by $-\psi_x = Z$, where Z takes four equidistantly spaced values across each wave

where subscript 1 and 2 denote parameters associated with waves positioned to the left and right, respectively. Furthermore, we require that solutions comprising two waves are (I) continuous and (II) satisfy the following set of arbitrary boundary conditions

$$-\psi_x \to \begin{cases} P, \ x \to -\infty \\ Q, \ x \to +\infty \end{cases}, \quad \psi_t \to \begin{cases} R, \ x \to -\infty \\ S, \ x \to +\infty \end{cases}. \tag{14}$$

Assuming that $l_1 > 0$ and $l_2 < 0$, we find (using the boundary conditions for each of the two waves) that the two requirements lead to the following conditions

$$\lambda_1 = \lambda_2, \qquad \sigma_1 = \sigma_2, \tag{15a}$$

$$P = \theta_1 + \lambda_1, \qquad Q = \theta_2 + \lambda_2, \tag{15b}$$

$$R = v_1\theta_1 + \sigma_1, \qquad S = v_2\theta_2 + \sigma_2, \tag{15c}$$

Finally, we substitute the boundary values of $-\psi_x$ and ψ_t at $x = \pm 100$ in Fig. 1 for P, Q, R, and S in (15), substitute the value of B/A into (13), and solve the resulting system of equations for v_1, θ_1, λ_1, σ_1, v_2, θ_2, λ_2, and σ_2. Following these steps we find that the waves after the collision travel at the velocities $-v_1 = v_2 = 1.76$ compared to $v_1 = -v_2 = 1.19$ before the collision. This finding is fine in agreement with the velocities determined from the slope of the contour lines in Fig. 1.

5 Conclusions

An exact thermoviscous shock solutions has been obtained using a generalized travelling wave assumption. This generalized assumption leads to a wider class of exact solutions compared to the one obtained from a standard travelling wave assumption and in turn this enable us to predict the outcome of two head-on colliding shocks. Analytical results for the wave speeds after the collision was in fine agreement with numerical observations. In future studies, it would be rewarding to further investigate interacting thermoviscous shocks, e.g. collisions between shocks travelling in the same directions.

References

1. Beyer, R.T.: In: Hamilton, M.F., Blackstock, D.T. (eds.) Nonlinear Acoustics. Academic, San Diego (1998)
2. Fernando, R., Marchiano, R., Coulouvrat, F., Druon, Y.: Nonlinear Acoustics–Fundamentals and Applications, Proceedings of the 18th ISNA, pp. 99–102. (2008)
3. Hamilton, M.F., Blackstock, D.T., Pierce, A.D.: In: Hamilton, M.F., Blackstock, D.T. (eds.)Nonlinear Acoustics. Academic, San Diego (1998)
4. Hamilton, M.F. and Morfey, C.L.: In: Hamilton, M.F., Blackstock, D.T. (eds.) Nonlinear Acoustics. Academic, San Diego (1998)
5. Hoffelner, J., Landes, H., Kaltenbacher, M., Lerch, R.: IEEE Trans. Ultrason. Ferroelect. Freq. Contr. **48**, 779–786 (2001)
6. Jordan, P.M.: Phys. Lett. A **326** 77–84 (2004)
7. Kuznetsov, V.P.: Sov. Phys. Acoust. **16**, 467–470 (1971)
8. Naugolnykh, K., Ostrovsky, L.: Nonlinear Wave Processes in Acoustics. Cambridge University Press, Cambridge (1998)
9. Zakharov, V.E., Kuznetsov, E.A.: Physics-Uspekhi **40**, 1087–1116 (1997)

Study on Development of the Seated Human Body System Exposed to Vehicular Ride Vibration Environment

S. Rodean and M. Arghir

Department of Mechanics and Computer Programming, Technical University of Cluj-Napoca, Romania srodean@staff.utcluj.ro, Mariana.Arghir@mep.utcluj.ro

Summary. This paper tries to find an appropriate structure of human model, which can better represent the characteristics of the real human body, using the apparent mass (APMS) and head transmissibility (STHT) in vertical vibrations. The model parameters were identified through minimizing an error function comprising the measured and model response in terms of magnitude and phase characteristics of APMS and STHT.

1 Introduction

In sitting posture the vibration, exciting the hip and thigh, is transmitted to the head through the entire body. In this way, the vibration transmissibility to the head and driving point mechanical impedance or apparent mass of the human body are important characteristics to express the vibration characteristics of a body. The apparent mass can show the driving point characteristics, while the head transmissibility can show the end point characteristics of the body. The apparent mass at the head and the head transmissibility are related to the comfort feeling.

2 Development of Human Driver Model

The development of complex models of the human body response requires an understanding of the modes of the body oscillation . Since the proposed human body model is to be used in the study of seating dynamics, the structure should be simple and the number of degrees-of-freedom should be low for its convenient use. Various biodynamic models have been developed to depict human motion from single DOF to multi-DOF models. These models can be divided into lumped parameter and distributed models. The lumped parameter models consider the human body as several rigid bodies and springs-dampers.

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*, Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_160,

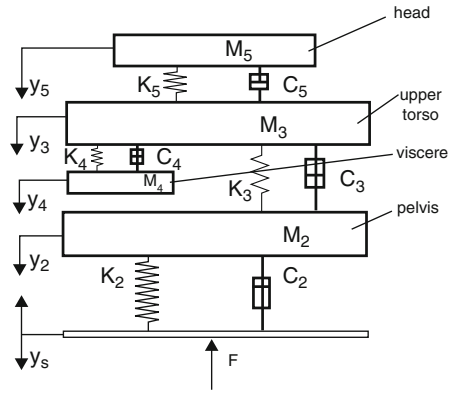


Fig. 1. 4-DOF linear biodynamic model

Considering the human body as a mechanical system, at low frequencies (less than 100 Hz) and low vibration levels, it may be roughly approximated by linear lumped parameter systems. For this study we used a mechanical model with 4 DOFs [2], corresponding to the following segments of the human body: pelvis, upper torso, viscera and head, respectively (Fig. 1).

The response of this system is given by:

$$\begin{aligned} M_2\ddot{y}_2 + K_2(y_2 - y_s) + C_2(\dot{y}_2 - \dot{y}_s) - K_3(y_3 - y_2) - C_3(\dot{y}_3 - \dot{y}_2) = 0 \\ M_3\ddot{y}_3 + K_3(y_3 - y_2) + C_3(\dot{y}_3 - \dot{y}_2) - K_4(y_4 - y_3) \\ - C_4(\dot{y}_4 - \dot{y}_3) - K_5(y_5 - y_3) - C_5(\dot{y}_5 - \dot{y}_3) = 0 \\ M_4\ddot{y}_4 + K_4(y_4 - y_3) + C_4(\dot{y}_4 - \dot{y}_3) = 0 \\ M_5\ddot{y}_5 + K_5(y_5 - y_3) + C_5(\dot{y}_5 - \dot{y}_3) = 0 \end{aligned} \quad (1)$$

The apparent mass response is derived from the resultant force at mass m_o and the driving-point acceleration . The resultant force F at the lower mass can be computed from the equation of motion for mass m_0.

$$F(t) = m_0\ddot{y}_s + K_2(y_s - y_2) + C_2(\dot{y}_s - \dot{y}_2) \quad (2)$$

The solution of (1) and (2) yields

$$F(t) = m_0\ddot{x}_s + m_2\ddot{x}_2 + m_3\ddot{x}_3 + m_4\ddot{x}_4 + m_5\ddot{x}_5 \quad (3)$$

Using the Laplace transforms, the (1) become written in the frequency domain. The APMS response of the model can then be derived as follows:

$$M(s) = \frac{F(s)}{s^2(X_s(s))} = m_0 + m_2\frac{X_2(s)}{X_s(s)} + m_3\frac{X_3(s)}{X_s(s)} + m_4\frac{X_4(s)}{X_s(s)} + m_5\frac{X_5(s)}{X_s(s)} \quad (4)$$

The STHT response of the model is computed from:

$$T(s) = \frac{X_5(s)}{(X_s(s))} \quad (5)$$

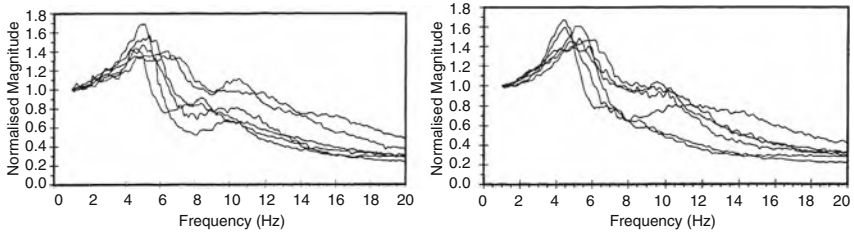


Fig. 2. The normalized apparent mass of subjects measured under ENS posture at $1\,\mathrm{m/s^2}$ rms (**a**) and $2\,\mathrm{m/s^2}$ rms (**b**) acceleration excitation

3 Experimental Set-Up

A rigid seat structure was designed to mimic typical automotive seat geometry and the sitting posture. The seat assembly was installed on a vertical vibration simulator (VVS) through a force platform, with a maximum displacement of 20 cm. A resonance search test was performed at excitation frequencies below 20 Hz. An accelerometer was attached to the seat pan to measure the acceleration transmitted to the human body. The VVS generated two acceleration levels: $1\,\mathrm{m/s^2}$ rms and $2\,\mathrm{m/s^2}$ rms. The vertical vibration of the head was measured using a bite-bar located at the corner of the mouth. The posture is defined as erect seated when only the lower back is in contact with the backrest – ENS posture. The foot is on the foot-plate and the hands held in driving position. The duration of vibration exposure did not exceed 90 s. The analysis is limited to the $0.4 - 20$ Hz frequency range.

3.1 Characteristics of the Test Subject Population

Subject	Age (years)	Height (m)	Mass(kg) standing	Mass(kg) sitting	% of the weight supported by the seat
A	31	1.70	77.8	57.0	73.26
G	39	1.76	98.7	77.3	78.31
D	40	1.80	75.9	54.1	71.27
P	41	1.83	105.33	78.5	74.52
F	38	1.90	81.17	72.1	88.82
S	32	1.68	54.0	39.6	73.33

3.2 Analysis of the Measured APMS Data

An increase in the mass parameters tends to reduce the primary resonant frequency, while an increase in the stiffness parameters tends to increase the primary resonant frequency for the APMS. An increase in the damping coefficients tends to decrease the primary resonant frequency derived from the APMS. The APMS can be conveniently normalized to the static mass supported by the seat to reduce the extent of variations attributed to the body mass (Fig. 2).

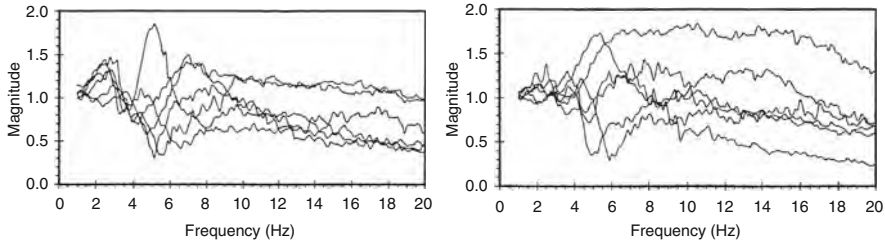


Fig. 3. The measured seat-to-head transmissibility of subjects under $1\,\mathrm{m/s^2}$ rms (**a**) and $2\,\mathrm{m/s^2}$ rms (**b**) acceleration excitation

3.3 Analysis of the Measured STHT Data

Some studies have concluded that increased mass or increased body size can be associated with lower STHT magnitude over a wide frequency range. From the data, it can be observed that the ENS posture yields larger variations among different subjects. Under an ENS posture, however, the STHT magnitudes for several subjects tend to fall below unity at all frequencies above 5 Hz, while others show values constantly in excess of unity up to 20 Hz with the exception of a dip in the $4 - 4.5$ Hz frequency range (Fig. 3).

4 Model Parameter Identification Methodology

A parametric optimization technique [1] was used to determine the model parameters. An objective function was defined to minimize the error between the computed and the measured values of the two biodynamic response functions over a specific frequency range. The objective function is defined as the weighted sum of the squared magnitude and phase errors associated with the APMS or STHT functions, respectively, and expressed as:

$$U(\chi) = \mathit{minimize}\,[\alpha U_M(\chi) + \beta U_T(\chi)] \tag{6}$$

where $U_M(\chi)$ and $U_T(\chi)$ are sums of squared errors resulting from APMS and STHT, respectively, given by:

$$U_M(\chi) = \lambda_1 \sum_{i=1}^{N} [|M(\omega_i)| - |M_t(\omega_i)|]^2 + \lambda_2 \sum_{i=1}^{N} [|\Phi_M(\omega_i)| - |\Phi_{M_t}(\omega_i)|]^2 \tag{7}$$

$$U_T(\chi) = \psi_1 \sum_{i=1}^{N} [|T(\omega_i)| - |T_t(\omega_i)|]^2 + \psi_2 \sum_{i=1}^{N} [|\Phi_T(\omega_i)| - |\Phi_{T_t}(\omega_i)|]^2 \tag{8}$$

where: – $M(\omega_i)$ and $\phi_M(\omega_i)$ are the magnitude and phase of the APMS response of the model corresponding to excitation frequency ω_i; – $M_t(\omega_i)$

and $\phi_{M_t}(\omega_i)$ are the corresponding measured values; – $T(\omega_i)$ and $\phi_T(\omega_i)$ are the magnitude and phase of the STHT response of the model; – $T_t(\omega_i)$ and $\phi_{T_t}(\omega_i)$ are the corresponding measured values; – N is the number of discrete frequencies selected in the 0.4–20 Hz frequency range; – χ is a vector of model parameters to be identified, expressed as:

$$\chi = \{m_0, m_2, m_3, m_4, m_5, c_2, c_3, c_4, c_5, k_2, k_3, k_4, k_5\}^T;$$

$-\lambda_1$,λ_2 and ψ_1, ψ_2 are weighting factors used in the APMS and STHT error functions, respectively, to ensure somewhat comparable contributions of magnitude and phase errors in the objective function; -α and β weighting factors – are chosen to emphasize the contributions due to either apparent mass or seat-to-head transmissibility functions to the total error function.

The minimization problem expressed in the (6) is solved subject to constraints applied on the total model mass. Since the mean measured data is related to mean body mass of 63.1 kg, supported by the seat, the limit constraints are expressed as 10% variations about the mean segment masses, while the total body mass is expressed as an equality constraint: $26.1\,\text{kg} \leq m_2 \leq 31.9\,\text{kg}$; $19.62\,\text{kg} \leq m_3 \leq 23.98\,\text{kg}$; $6.12\,\text{kg} \leq m_4 \leq 7.48\,\text{kg}$; $4.95\,\text{kg} \leq m_5 \leq 6.05\,\text{kg}$ and $\sum_{i=2}^{5} m_i = 6.31\,\text{kg}$. The optimization function is further subject to have the positive parameters.

The development of a human body model involves complexities associated with identification of its restoring and dissipative properties.

5 Numerical Model: The Solution of the Constrained Optimization Problem

The optimization problem defined in (7) and (8) is solved using the optimization software MATLAB based on sequential search. The differential equations of motion (1) are solved for unit displacement excitation to derive the apparent mass magnitude and phase and seat-to-head transmissibility magnitude and phase respectively. Different optimization runs corresponding to different starting values converged to similar values of model parameter and the error function. The model parameters, thus identified, are summarized below: $m_0 = 2\,kg$; $m_2 = 29\,kg$; $c_2 = 108.42\,Ns/m$; $k_2 = 16.21\,e^4\,N/m$; $m_3 = 21.8\,\text{kg}$; $c_3 = 199.72\,\text{Ns/m}$; $k_3 = 3.78\,e^4\,\text{N/m}$; $m_4 = 6.8\,\text{kg}$; $c_4 = 138.74\,\text{Ns/m}$; $k_4 = 0.28\,e^4\,\text{N/m}$; $m_5 = 5.5\,\text{kg}$; $c_5 = 210.95\,\text{Ns/m}$; $k_5 = 20.22\,e^4\,\text{N/m}$.

The analytical model of the seated body is evaluated to derive the response characteristics in terms of STHT and APMS, using equations of $M(s)$ and $T(s)$ respectively. The computed response, as shown in following figures and characteristics are compared with the measured to examine the effectiveness of the proposed model.

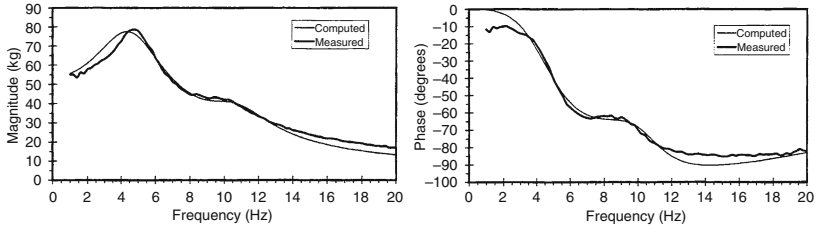


Fig. 4. Comparison of the computed APMS with the measured data

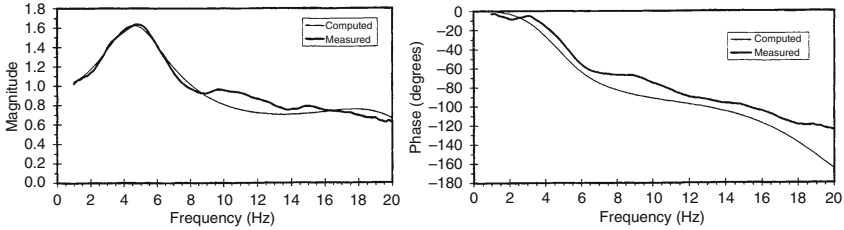


Fig. 5. Comparison of the computed APMS with the measured data

6 Results and Discussion

The computed APMS function reveals two resonant peaks in the vicinity of 5 Hz and 10 Hz, which are quite consistent from the measured data. The resonance in the vicinity 5 Hz is primarily associated with the deflection modes of the coupled subsystem comprising m_5 and m_4, while the resonance near 10 Hz is associated with the deflection mode for the mass m_2. The apparent mass is more closely related to the seat to head transmissibility than the mechanical impedance in yielding the primary resonant frequency of the body (Fig. 4).

The STHT is solely attributed to the dynamic response due to subsystem comprising m_4 and m_5. Similarly to the measured data, the computed STHT function reveals a resonance peak near 5 Hz and near 10 Hz for the second (Fig. 5).

7 Conclusions

These results suggest that if a model is to be based on both driving-point force-motion relation transfer function and vibration transmission function, APMS and STHT functions should be selected. The results also suggest that it is possible to develop a seated body model with relatively lower degrees-of-freedom, on the basis of analytical functions and measured data. The optimum form of the 4 DOFs model was determined by curve fitting to the experimental data obtained. After minimizing the objective function $U(\chi)$, comprising magnitude and phase components of both response functions APMS and STHT,

using the measured data in range of frequency $0.4 - 20$ Hz the parameters in the equations of the model were redefined.

References

1. Boileau, P.E., Rakheja, S.: Int. J. Ind. Ergon. (1997)
2. Payne, R.R., Band, E.G.U.: Aerospace Medical Research Lab. Report AMRL-TR-70-35, Wright-Patterson Air Force Base, OH (1971)

Surrogate Modeling for Geometry Optimization

M. Rojas[1], Y.B. Abraham[2], N.A.W. Holzwarth[3], and R.J. Plemmons[4]

[1] Delft Institute of Applied Mathematics, Delft University of Technology, P.O. Box 5031, 2600 GA Delft, The Netherlands narielba.rojas@tudelft.nl
[2] Computer Science and Physics, Wake Forest University, USA abrahayb@wfu.edu. The author is currently with Quantlab Financial in Houston, Texas, USA
[3] Physics, Wake Forest University, USA natalie@wfu.edu
[4] Mathematics and Computer Science, Wake Forest University, USA plemmons@wfu.edu

Summary. A new approach for optimizing the nuclear geometry of an atomic system is described. Instead of the original expensive objective function (energy functional), a small number of simpler surrogates is used.

1 The Problem

We consider the problem of finding a configuration or geometry that minimizes the total energy of an atomic system. We use the Born–Oppenheimer approximation [4] which considers the motion of the (heavier) nuclei and of the (lighter) electrons separately. Therefore, in this work, atomic configuration or geometry refers to configuration or geometry of the atomic nuclei.

Configurations with minimum energy determine the electronic structure of an atomic system. In turn, the electronic structure determines all properties of materials including elastic, magnetic and optical properties. Therefore, the computational study and design of materials often begins by finding an equilibrium geometry: one that yields a minimum of the energy (hyper)surface.

For a system of N atoms, our problem can be formulated as the following unconstrained minimization problem:

$$\min_{X} E(X) \tag{1}$$

where E is the energy, and $X = (X_1, X_2, \ldots, X_N)$ denotes the geometry: the spatial coordinates of the N nuclei.

The X_i's are either Cartesian coordinates or so-called internal coordinates (angles and distances) which are often preferred in practice. Once symmetries and other properties are taken into account, the number of variables in (1) is

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*, Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_161,

significantly smaller than $3 \times N$ (when working with Cartesian coordinates). The typical dimension of (1) is approximately 20 for a system of 40 atoms.

Unfortunately, no expression exists for the energy in terms of the nuclear coordinates only. In most models, the energy depends on both the nuclear geometry and the electronic wave functions or orbitals. For a system with N atoms and M electrons, the problem that we need to solve is:

$$\min_{X,\Psi} \; E(X;\Psi) \tag{2}$$

where X is as before and $\Psi = (\Psi_1, \Psi_2, \ldots, \Psi_M)$ denotes the electronic wave functions.

One of the most popular and successful models for the energy is the total-energy functional from Density Functional Theory (DFT) [11,12] which offers a good compromise between complexity (cost) and accuracy. In this model, the expression for the energy is known as the DFT Total-Energy Functional which depends on the nuclear geometry and the charge density of the system. Since the charge density depends on the electronic wave functions, the total-energy functional can be written, albeit in simplified form, as:

$$E(X;\Psi) = E_{kinetic}(\Psi) + E_{Coulomb}(X;\Psi) + E_{xc}(\Psi) \tag{3}$$

where $E_{kinetic}$ denotes the kinetic energy; $E_{Coulomb}$ denotes the Coulomb energy involved in the interaction between nuclei and electrons (attraction) and among electrons (repulsion); and E_{xc} is the Exchange-Correlation energy due to quantum-mechanical effects. Note that E_{xc} is unknown and must be approximated. The Local Density Approximation (or LDA) is a common choice.

In this work, we use (3) as objective function and transform problem (2) into the following two-step minimization problem:

$$\min_{X} \{\min_{\Psi} E(X;\Psi)\} \tag{4}$$

The solution of (4) requires at least the evaluation of (3) and possibly of its derivatives. The evaluation of the DFT total-energy functional and its derivatives at a given geometry is usually done by means of the Self-Consistent-Field (SCF) method [7,9,10,21]. SCF is a lengthy procedure that may require weeks of computations to evaluate the total-energy functional at a *single* geometry. Therefore, classical optimization techniques such as Quasi-Newton and Nonlinear Conjugate Gradients Methods [16] applied to (4) are usually very expensive. Moreover, these techniques cannot take advantage of previously-computed energy values. An alternative is the molecular-dynamics approach or Car-Parrinello (CP) Method [5] in which both the geometry and the electronic structure are computed simultaneously. The CP method is very efficient for some systems. However, the method is not robust and can be very expensive for large systems. Moreover, it cannot make use of previously-computed energy values.

In the next section, we describe an approach for finding an approximate solution to (4) that makes use of previously-computed function values and requires a low number of evaluations of the total-energy functional.

2 Proposed Approach

We propose to use surrogate modeling [3] (see also [2]) for solving (4). In surrogate modeling, we construct a simple and inexpensive model (surrogate) of an expensive, sometimes unknown, objective function. Classical optimization techniques can then be applied to the surrogate and the resulting information can be used to construct a more accurate, but still inexpensive surrogate. Surrogate modeling for geometry optimization was first used in [1].

Constructing a surrogate requires an initial set of true function values. Some approaches also require derivative values at those points. Spline interpolation (see, for example [6]) and statistical interpolation such as kriging [15] are popular techniques for constructing surrogates. In [1] and in this work, we used both splines and kriging interpolation.

Our approach consists of first, building an energy surrogate from an initial set of points (design sites), then minimizing the surrogate, and finally, adding the minimizer and the true energy value at this point to the design sites to build a new surrogate. These steps are repeated until a satisfactory geometry is found. The procedure is presented in Fig. 1. The approach has the advantage of making use of energy values previously (and costly) calculated.

The procedure in Fig. 1 usually yields reasonable approximations after few evaluations of the total-energy functional (typically, 1 or 2 iterations are needed).

Input: design sites $< X_\ell, E_\ell >,\ \ell = 1, \ldots, m$
Output: X_* (geometry)
1. $k = 0$; convergence = **false**;
2. while not convergence
2.1 Construct surrogate $e(X)$
2.2 Compute minimizer X_k^* of e
2.3 Compute E_k^*: energy value at X_k^*
2.4 Add $< X_k^*, E_k^* >$ to design sites
2.5 $k = k + 1$; update convergence;
end
3. $X^* = X_k^*$

Fig. 1. Procedure for geometry optimization using surrogates

3 Results

All our experiments were conducted in MATLAB [19] on a SUN Blade station running Solaris. The spline surrogates were computed with MATLAB's `interpn` function for n-dimensional spline interpolation. The kriging surrogates were computed with the DACE package [15]. The minimization of the surrogates was accomplished by means of the direct-search Nelder–Mead algorithm [14] as implemented in MATLAB's `fminsearch` function.

The proposed strategy produced very good results on simple test problems including the Lennard–Jones (1D) and Gaussian (2D) potentials. Splines surrogates for these cases are shown in Figs. 2 and 3, respectively.

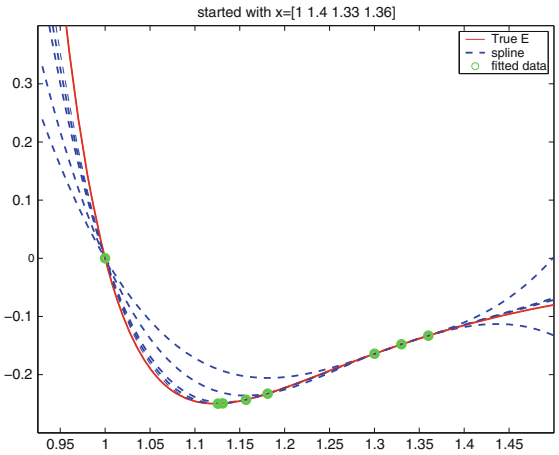


Fig. 2. Spline surrogates for the Lennard–Jones potential: $E(r) = \frac{1}{r^{12}} - \frac{1}{r^6}$

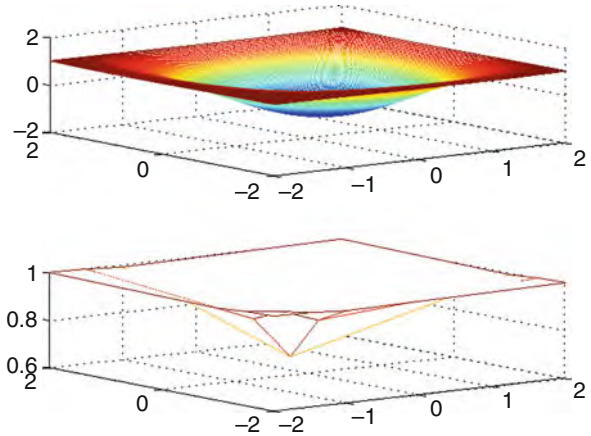


Fig. 3. A spline surrogate (bottom) for the Gaussian potential $E(r_1, r_2) = (1 - e^{-(r_1^2 + r_2^2)})^2$ (top)

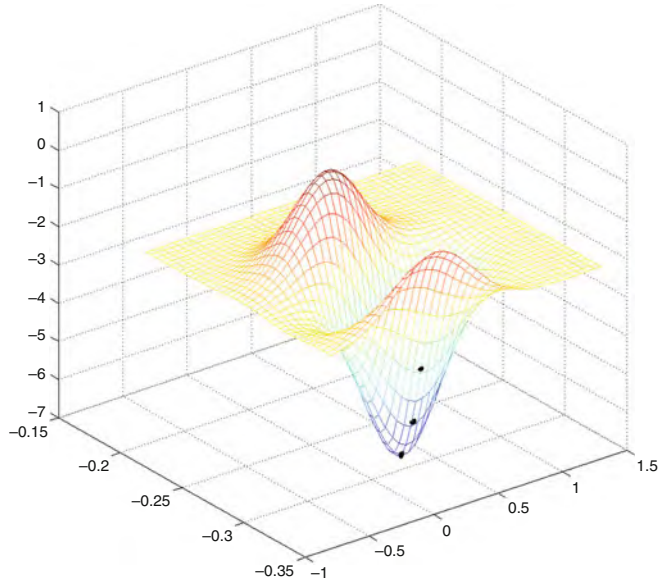


Fig. 4. A kriging surrogate for a 2D diamond sheet

We have also tested our approach on a real material. The goal was to optimize the geometry of a 2D diamond sheet. The plot in Fig. 4 shows the kriging surrogate constructed with DACE after nine evaluations of the total-energy functional, when we fixed all but two variables.

In all our tests, and in particular in the case of the real material, the kriging surrogates computed with DACE outperformed the spline surrogates.

4 Concluding Remarks

We have presented a surrogate-modeling approach for optimizing the geometry of atomic systems. Our approach takes advantage of available energy values which are expensive to compute and produces satisfactory results for practitioners at a lower cost than conventional techniques.

Future work includes the use of improvements on SCF such as [8,13,17,18,20,22] for solving the inner minimization problem in (4), and the development of an automatic stopping criterion.

References

1. Abraham, Y.: Optimization with Surrogates for Electronic-Structure Calculations, Master's Thesis, Department of Computer Science, http://hdl.handle.net/10339/190, Wake Forest University, Winston-Salem, North Carolina, USA, May 2004

2. Bandler, J., Madsen, K.: Optim. Eng. **2**, 367–368 (2001)
3. Booker, A., Dennis, J.E., Frank, P., Serafini, D., Torczon, V., Trosset, M.: Struct. Multidiscip. Optim. **17**(1), 1–13 (1999)
4. Born, M., Oppenheimer, R.: Ann. Phys. **84**, 457–484 (1927)
5. Car, R., Parrinello, M.: Phys. Rev. Lett. **55**, 2471–2474 (1985)
6. Cheney, W., Kincaid, D.: Numerical Mathematics and Computing, 5th edn. Thomson Learning, Belmont (2004)
7. Fock, V.: Z. Physik **61**, 126 (1930)
8. Francisco, J.B., Martínez, J.M., Martínez, L.: J. Math. Chem. **40**, 349–377 (2006)
9. Hall, G.G.: Proc. Roy. Soc. **A205**, 541 (1951)
10. Hartree, D.R.: Proc. Cambridge Phil. Soc. **24**, 89–111 (1928)
11. Hohenberg, P., Kohn, W.: Phys. Rev. **136**, B864–B871 (1964)
12. Kohn, W., Sham, L.: Phys. Rev. **140**, A1133–A1138 (1965)
13. Kudin, K.N., Scuseria, G.E., Cances, E.: J. Chem. Phys. **116**, 8255 (2002)
14. Lagarias, J.C., Reeds, J.A., Wright, M.H., Wright, P.E.: SIAM J. Optim. **9**(1), 112–147 (1998)
15. Lophaven, S., Nielsen, H.B., Søndergaard, J.: DACE – A MATLAB Kriging Toolbox, Version 2.0, Technical Report IMM-TR-2002-12, Informatics and Mathematical Modelling, Technical University of Denmark, Lyngby, Denmark, August 2002
16. Nocedal, J., Wright, S.J.: Numerical Optimization, 2nd edn. Springer, New York (2006)
17. Pulay, P.: Chem. Phys. Lett. **73**, 393 (1980)
18. Pulay, P.: J. Comput. Chem. **3**, 556 (1982)
19. The MathWorks: MATLAB: The Language of Technical Computing, Version 7, The MathWorks, Inc., (2006)
20. Thøgersen, L., Olsen, J., Yeager, D.P., Jørgensen, P. Salek, P., Helgaker, T.: J. Chem. Phys. **1**, 16–27 (2002)
21. Roothaan, C.C.J.: Rev. Mod. Phys. **23**, 69 (1951)
22. Yang, C., Meza, J.C., Wang, L.: SIAM J. Sci. Comput. **29**, 1854–1875 (2007)

Variational Optimization of Power Yield in Industrial Systems

Stanislaw Sieniutycz

Faculty of Chemical and Process Engineering, Warsaw TU, 1 Waryńskiego, Poland
sieniutycz@ichip.pw.edu.pl

Summary. Variational optimization is applied to simulation and modeling of dynamical energy converters, in particular thermal and solar engines. Basic thermodynamic principles lead to expressions for converter's efficiency and limiting work. Work generated is a cumulative effect obtained in a system of a resource fluid, engines, and an infinite bath. The limiting work function depends on thermal coordinates and a dissipation index, h, in fact the Hamiltonian of the optimization problem of minimum entropy production. Bounds on work delivery implied by the limiting function are stronger than those predicted by the reversible work potential.

1 Introduction

Power limits in dynamical energy systems are determined by flows and properties of propelling fluids which play the role of resources. A power limit is an upper (lower) bound on power produced (consumed) in the system. A resource is a valuable substance or energy used in a process; its value can be quantified by specifying its availability function, a maximum work that can be obtained when the resource relaxes to the equilibrium. Reversible relaxation of the resource is associated with classical availability; when dissipative phenomena are allowed generalized availabilities emerge which quantify deviations of the system's efficiency from the Carnot efficiency. An availability is obtained as the value function to the variational problem of extremum work. Other components of the variational solution are optimal trajectory and optimal control. In thermal systems the trajectory is characterized by the temperature of the resource fluid, $T(t)$, whereas the control is Carnot temperature $T'(t)$ defined in our previous work [4]. For the reader's convenience basic properties of T' are outlined in the Appendix. Whenever $T'(t)$ differs from $T(t)$ the resource is downgraded or upgraded with a finite rate, and with an efficiency different from the Carnot efficiency. Only when $T'(t) = T(t)$ the efficiency is Carnot, but this corresponds with an infinitely slow relaxation rate of the resource to the equilibrium.

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*,
Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_162,

In this paper work functionals are formulated and extremized to find the optimal resource temperature, power output, and controls. Recent expressions for efficiency of imperfect converters [6, 7] are used to derive and solve Hamilton–Jacobi equations describing upgrading and downgrading of the resource fluid. Optimal trajectories and controls which maximize work yield are evaluated by various optimization methods.

2 Finite Resources and Dynamical Problems of Power Optimization

From the optimization viewpoint the dynamical process is every one in which one can distinguish sequential changes of state, either in the chronological time or in (spatial) holdup time. Power yield during the resource's relaxation to the environment is such a sequential process which is accompanied by the decrease of the resource's temperature in time.

The great deal of research on power limits published to date deals with stationary systems, in which case both reservoirs are infinite. To this case refer steady-state analyses of the Chambadal-Novikov-Curzon-Ahlborn engine (CNCA engine; [2]), in which energy exchange is described by Newtonian law of cooling, or the Stefan–Boltzmann engine, a system with the radiation fluids and the energy exchange governed by the Stefan–Boltzmann law [3]. Because of their stationarity (caused by the infiniteness of both reservoirs), controls maximizing power are represented by fixed points in the control space. Yet, the prediction of a dynamical energy yield requires the evaluation of an extremal curve rather than an extremum point. This leads to variational methods (to handle functional extrema) in place of static optimization methods (to handle extrema of functions). For example, the use of a pseudo-Newtonian model to quantify the dynamical energy yield from radiation, gives rise to an extremal curve describing the radiation relaxation to the equilibrium. This curve is non-exponential, the consequence of the nonlinear properties of the relaxation dynamics. Non-exponential are also other curves describing the radiation relaxation, e.g. those following from exact models applying the Stefan–Boltzmann equation (symmetric and hybrid; [3]).

Analytical difficulties associated with the dynamical optimization of nonlinear systems may be severe; this is why diverse models of power yield and diverse numerical approaches are applied. Various control variables may be used in modeling since the process analysis using a particular control can be substantially easier than the analysis in terms of another one.

Optimal (i.e. power-maximizing) relaxation curve $T(t)$ is associated with the optimal control curve $T'(t)$; they both are components of the (dynamic) optimization solution to a continuous problem. In the corresponding discrete problem, formulated for numerical purposes via a suitable discretization, one searches for optimal temperature sequences $\{T^n\}$ and $\{T'^n\}$; optimization methods lead to optimal sequences $\{T^n\}$ and $\{T'^n\}$.

Minimum work supplied to the system is described in a suitable way by function sequences $R^n(T^n, t^n)$, whereas maximum work produced – by functions $V^n(T^n, t^n)$. Profit-type performance function V and cost-type performance function R describe, respectively, optimal work production and consumption. They both differ by sign, i.e. $V^n(T^n, t^n) = -R^n(T^n, t^n)$. A beginner may find the change from symbol V to symbol R and back unnecessary and confusing. Yet, each function is positive in its own, natural regime of control variables (V is positive in the engine range and R is positive in the heat pump range).

To obtain the classical availability from a work function it is sufficient to assume that the thermal efficiency of the system is identical with the Carnot efficiency. On the other hand, non-Carnot efficiencies lead to generalized availabilities.

3 Two Modes of Control and Finite Rate Availabilities

For appropriate boundary conditions, the principal function of the variational problem of extremum work coincides with an availability function, a quantity that characterizes the quality of finite resources.

Two different works, the first associated with the resource downgrading during its relaxation to the equilibrium and the second – with the reverse process of resource upgrading, are essential (Fig. 1). The resource is downgraded during the approach to the equilibrium; then $T'(t) < T(t)$ and the *engine mode* of the system takes place in which work is released. The resource is upgraded during the departure from the equilibrium; then $T'(t) > T(t)$ and the *heat-pump mode* occurs in which work is supplied. Work W delivered in the engine mode is positive by assumption ("engine convention"). A sequence of irreversible engines (CNCA or Stefan–Boltzmann) serves to determine a rate-dependent availability that extends the classical availability for irreversible, finite rate processes. Before the work maximization, process efficiency η has to be expressed as a function of state T and a control variable, i.e. energy flux q or rate $\mathrm{d}T/\mathrm{d}\tau$, to assure the functional property (path dependence) of the work integral. The optimal work is sought in the form of a potential function that depends on the end states and duration. Each small step is a work-producing (consuming) CNCA stage with the energy exchange between two fluids and the thermal machine through finite "conductances" (products of transfer coefficients and related areas). For radiation engines, it follows from the Stefan–Boltzmann law that the effective transfer coefficient α_l of the "driving" (radiation) fluid is necessarily temperature dependent, $\alpha_l = T_l^3$. The optimizer's task is to find an optimal temperature of the resource fluid along the path that extremizes the work consumed or delivered. For traditional fluids (constant c_v) an optimal path is known to be exponential [4]. Yet, no exponential decay of temperature occurs for nonlinear fluids.

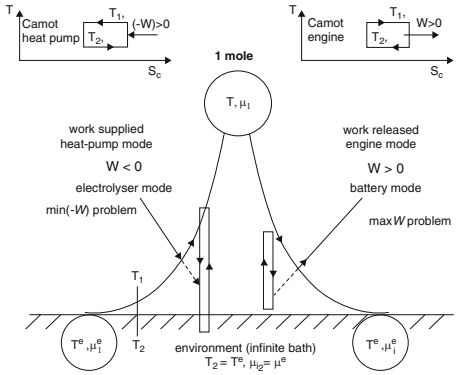


Fig. 1. Limiting works produced and consumed are different in an irreversible process

Total power obtained from an infinite number of infinitesimal engines is determined as the Lagrange functional of the following structure

$$\dot{W}[\mathbf{T}^i, \mathbf{T}^f] = \int_{t^i}^{t^f} f_0(T, T')\,\mathrm{d}t = -\int_{t^i}^{t^f} \dot{G}c(T)\eta(T, T')\dot{T}\,\mathrm{d}t \tag{1}$$

where f_0 is power generation intensity, $\dot{G}$ is resource flux, $c(T)$ is specific heat, $\eta(T, T')$ is efficiency in terms of state T and control T', further $\mathbf{T}$ is enlarged state vector comprising temperature and time, t is the time variable (residence time or holdup time) for the resource contacting with heat transfer surface. Often one uses a non-dimensional time τ, identical with the so-called number of the heat transfer units. For a constant mass flow $\dot{G}$ of a resource, one can extremize power per unit mass flux (the quantity of work dimension). In this case (1) describes a problem of extremum work at flow. Integrand f_0 is common for both modes, yet the numerical results it generates differ by sign (positive for engine mode; "engine convention"). Power generation function f_0 can be replaced by power consumption function $l_0 = -f_0$. Formally, l_0 plays the role of a process Lagrangian. f_0 in (1) contains thermal efficiency, η, described by a practical counterpart of the Carnot formula. Whenever $T > T^{\mathrm{e}}$, efficiency η decreases in the engine mode above Carnot η_c and increases in the heat-pump mode below η_c. At the limit of vanishing rates, $\mathrm{d}T/\mathrm{d}t = 0$ and $T' \to T$ and we obtain the integral of the classical availability. Thus, (1) leads to a generalized availability for finite rate processes. In problems with a constant specific heat, fluid's specific work at flow, w, is described by an equation

$$w[\mathbf{T}^i, \mathbf{T}^f] = \frac{\dot{W}}{\dot{G}} = -\int_{T^i}^{T^f} c(T)\left(1 - \frac{T^{\mathrm{e}}}{T}\right)\mathrm{d}T - T^{\mathrm{e}}\int_{t^i}^{t^f} c(T)\frac{(T'-T)^2}{T'T}\mathrm{d}\tau, \tag{2}$$

where

$$\tau \equiv \frac{x}{H_{TU}} = \frac{\alpha' a_v F}{\dot{G}c} x = \frac{\alpha' a_v F v}{\dot{G}c} t = \frac{1}{\chi} \tag{3}$$

is *non-dimensional* time of the process. The functional (2) has two additive parts: the classical (potential) part and a non-potential, dissipative part which depends on the history of the process. Equation (3) assumes that a resource fluid flows with velocity v through cross-section F and contacts with the heat transfer exchange surface per unit volume a_v. Quantity τ is identical with the so-called *number of the heat transfer units*.

Solutions to work extremum problems are obtained by:

(a) variational methods, i.e. via Euler–Lagrange equation of variational calculus

$$\frac{\partial L}{\partial T} - \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial L}{\partial \dot{T}}\right) = 0 \tag{4}$$

In the example considered above, i.e. for a thermal system with linear kinetics

$$T\frac{\mathrm{d}^2 T}{\mathrm{d}\tau^2} - \left(\frac{\mathrm{d}T}{\mathrm{d}\tau}\right)^2 = 0 \tag{5}$$

which corresponds with the optimal trajectory

$$T(\tau, \tau^f, T^i, T^f) = T^i (T^f/T^i)^{\tau/\tau^f}. \tag{6}$$

($\tau^i = 0$ is assumed in (6).) However, the solution of the Euler-Lagrange equation does not contain any information about the optimal work function. This property is assured by solving the Hamilton-Jacobi-Bellman equation (HJB equation).

(b) dynamic programming via HJB equation for the 'principal function' (V or R), also called extremum work function. For the example described by (2)

$$\frac{\partial V}{\tau} + \min_{T'}\left\{\left(\frac{\partial V}{\partial T} + c\left(1 - \frac{T^{\mathrm{e}}}{T'}\right)(T' - T)\right)\right\} = 0. \tag{7}$$

The extremal work function V is a function of the final state T and total duration. After evaluation of optimal control and its substitution into one obtains a nonlinear equation

$$\frac{\partial V}{\partial \tau} - c\left\{\sqrt{T^{\mathrm{e}}} - \sqrt{T\left(1 + c^{-1}\frac{\partial V}{\partial T}\right)}\right\}^2 = 0 \tag{8}$$

which is the Hamilton-Jacobi equation of the problem. Its solution can be found by the integration of work intensity along an optimal path, between limits T^i and T^f. A reversible (path independent) component of V is the classical availability $A(T, T^{\mathrm{e}}, 0)$.

Details of models of multistage power production in sequences of infinitesimal engines are known from our previous publications [4,6,7]. These models provide both power generation functions f_0 (or thermal Lagrangians $l_0 = -f_0$) and dynamical constraints. Numerical methods apply discrete models for rates f_0 (or l_0) and $\mathbf{f}$. With discrete models the theory can be restated and live with its own life in the realm of difference equations, sums, recurrence relations, etc., often achieving a form dissimilar while still equivalent to the original HJB theory [5].

4 Analytical Solutions in Systems with Linear Kinetics

In the HJB formalism Hamiltonians are defined in the enlarged state space (T, τ) or (T, t) rather than in the phase space (T, z, τ) or (T, z, t). Yet, the Pontryagin's Hamiltonian of the linear system (for the Newtonian energy flow in time τ rather than t) is

$$H = \left[z - c\left(1 - \frac{T^{\mathrm{e}}}{T'}\right)\right](T' - T) = \left[z - c\left(1 - \frac{T^{\mathrm{e}}}{T + u}\right)\right]u. \tag{9}$$

In fact, $z = \partial R/\partial T$, i.e. the temperature adjoint is the gradient of R (or negative gradient of V). Optimal driving temperature T' is obtained as a quantity maximizing Hamiltonian (9) with respect to T' at each point of the path. The maximization of H leads to an equation

$$\frac{\partial R}{\partial T} - \frac{\partial l_0(T, T')}{\partial T'} = \frac{\partial R}{\partial T} - c\left(1 - \frac{T^{\mathrm{e}}T}{T'^2}\right) = 0 \tag{10}$$

that expresses the optimal control T' in terms of T and z or $\partial R/\partial T$ and holds along with the original HJB equation (7) without extremizing operation. In terms of R rather than V

$$\frac{\partial R}{\partial \tau} + \frac{\partial R}{\partial T}(T' - T) - c\left(1 - \frac{T_2}{T'}\right)(T' - T) = 0 \tag{11}$$

To obtain optimal control function $T'(z, T)$ one should solve the second equality in (10) in terms of T'. The result is optimal Carnot control T' in terms of T and $z = \partial R/\partial T$,

$$T' = \left[T^{\mathrm{e}}T\left(1 - c^{-1}\frac{\partial R}{\partial T}\right)^{-1}\right]^{1/2}. \tag{12}$$

This is next substituted into (11); the result is the nonlinear Hamilton–Jacobi equation

$$\frac{\partial R}{\partial \tau} + cT\left(\sqrt{1 - c^{-1}\frac{\partial R}{\partial T}} - \sqrt{\frac{T^{\mathrm{e}}}{T}}\right)^2 = 0 \tag{13}$$

which is related to H of (9) for $z = \partial R/\partial T$. Assuming a numerical value of $H = h$,

$$cT\left(\sqrt{1-c^{-1}z}-\sqrt{T^{\mathrm{e}}/T}\right)^2 = h \tag{14}$$

one can exploit the constancy of autonomous H to eliminate adjoint z. Next, combining (14) with optimal control (12), yields optimal rate $u = \dot{T}$ in terms of T and constant h

$$\dot{T} = \xi(h, T^{\mathrm{e}})T, \tag{15}$$

where

$$\xi(h, T^{\mathrm{e}}) \equiv \pm\sqrt{h/cT^{\mathrm{e}}}(1 - \pm\sqrt{h/cT^{\mathrm{e}}})^{-1} \tag{16}$$

is a process intensity index. Positive ξ refer to heating of the resource in heat-pump mode, and the negative to cooling in engine mode. Equation (15) describes the optimal trajectory in terms of state variable T and constant h. The corresponding Carnot control is

$$T' = [\xi(h, T^{\mathrm{e}}) + 1]\, T. \tag{17}$$

Now one can find the (solution to the problem of) Hamiltonian representation of extremal work. Substituting temperature control (17) into work functional (2) and integrating along an optimal path yields the extremal work function

$$\begin{aligned} V(T^i, T^f, h) &= c(T^i - T^f) - cT^{\mathrm{e}} \ln \frac{T^i}{T^f} + cT^{\mathrm{e}} \frac{\xi(h)}{1+\xi(h)} \ln \frac{T^i}{T^f} \\ &= c(T^i - T^f) - cT^{\mathrm{e}} \ln \frac{T^i}{T^f} - cT^{\mathrm{e}} \sqrt{\frac{h}{cT^{\mathrm{e}}}} \ln \frac{T^i}{T^f} \end{aligned} \tag{18}$$

This expression is valid for every process mode. Integration of (15) subject to boundary conditions $T(\tau^i) = T^i$ and $T(\tau^f) = T^f$ allows us to express (18) in terms of the process duration

$$V(T^i, T^f, \tau^i, \tau^f) = c(T^i - T^f) - cT^{\mathrm{e}} \ln \frac{T^i}{T^f} - \frac{cT^{\mathrm{e}}[\ln(T^i/T^f)]^2}{\tau^f - \tau^i - \ln(T^i/T^f)}. \tag{19}$$

5 Final Remarks

Applications of HJB theory lead to solutions which describe finite-rate generalizations of the standard availability. Generalized availabilities are irreversible extensions of the reversible work potential including minimally irreversible processes. Limits for energy yield or consumption provided by generalized availabilities are stronger than those defined by the classical availability. An essential decrease of the maximum work received from an engine system and an increase of the minimum work added to a heat pump system has been shown in the high-rate regimes and for short durations. Finite rates increase a

minimum work that must be supplied to the system and decrease a maximum work that can be produced by the system. These results help an engineer in better evaluation of energy limits in practical processes, especially in those undergoing in thermal engines and solar driven heat pumps.

6 Appendix: Carnot Temperature Control

For the reader's convenience we recall here the definition of Carnot temperature whose derivation and applications are presented in our previous work [4]. For brevity we restrict ourselves to a steady endoreversible cycle characterized by reservoir temperatures T_1 and T_2 and temperatures of circulating fluid T_1' and T_2'. In engine mode $T_1 > T_1' > T_2' > T_2$. Evaluating entropy production σ_s as the difference of outlet and inlet entropy fluxes yields

$$\sigma_s = \frac{q_2}{T_2} - \frac{q_1}{T_1} = \frac{(1-\eta)q_1}{T_2} - \frac{q_1}{T_1} = \frac{q_1}{T_2}\left(1 - \eta - \frac{T_2}{T_1}\right) \tag{20}$$

where η is the first-law efficiency. Since $\eta = 1 - T_2/T_1$, we obtain in terms of the temperatures of fluids circulating in the engine

$$\sigma_s = \frac{q_1}{T_2}\left(\frac{T_2{}'}{T_1{}'} - \frac{T_2}{T_1}\right). \tag{21}$$

Therefore after introducing an effective temperature called Carnot temperature

$$T' \equiv T_2 \frac{T_1{}'}{T_2{}'} \tag{22}$$

endoreversible entropy production (21) takes the following simple form

$$\sigma_s = q_1\left(\frac{1}{T'} - \frac{1}{T_1}\right) \tag{23}$$

This form is identical with the familiar expression obtained for the process of purely dissipative heat exchange between two bodies with temperatures T_1 and T'.

The endoreversible efficiency $\eta = 1 - T_2{}'/T_1{}'$ takes in terms of T' the classical Carnot form

$$\eta = 1 - \frac{T_2}{T'} \tag{24}$$

which substantiates the name "Carnot temperature" for T'. Moreover, power produced in the endoreversible system also takes the classical form

$$p = \eta q_1 = \left(1 - \frac{T_2}{T'}\right) q_1 \tag{25}$$

It is essential that the derivation of (20)–(25) does not require any specific assumptions regarding the nature of heat transfer kinetics. Kinetic aspects of

Carnot temperature are discussed elsewhere [4, 5]. Abandoning the endoreversibility assumption requires the knowledge of the experimental data of internal entropy production [1].

References

1. Chua, H.T., Ng, K.C., Gordon, J.M.: Experimental Study of the Fundamental Properties of Reciprocating Chillers and its Relation to Thermodynamic Modelling and Chiller Design. Intern. J. Heat Mass Transf. **39**, 2195–2204 (1996)
2. Curzon, F.L., Ahlborn, B.: Efficiency of Carnot Engine at Maximum Power Output. American J. Phys. **43**, 22–24 (1975)
3. Kuran, P.: Nonlinear Models of Production of Mechanical Energy in Non-Ideal Generators Driven by Thermal or Solar Energy. PhD thesis, Warsaw University (2006)
4. Sieniutycz, S.: Carnot Controls to Unify Traditional and Work-assisted Operations with Heat and Mass Transfer. Int. J. Appl. Thermodyn. **6**, 59–67 (2003)
5. Sieniutycz, S.: Dynamic Programming and Lagrange Multipliers for Active Relaxation of Resources in Nonlinear Non-Equilibrium Systems. Appl. Math Model. **33**, 1457–1478 (2009)
6. Sieniutycz, S., Kuran, P.: Nonlinear Models for Mechanical Energy Production in Imperfect Generators Driven by Thermal or Solar Energy. Int. J. Heat Mass Transf. **48**, 719–730 (2005)
7. Sieniutycz, S., Kuran, P.: Modeling Thermal Behavior and Work Flux in Finite-Rate Systems with Radiation. Int. J. Heat Mass Transf. **49**, 3264–3283 (2006)

An Age-Dependent Metapopulation Model

Jacques A.L. Silva[1] and Edgar Pereira[2]

[1] Departamento de Matematica, Universidade Federal do Rio Grande do Sul, Porto Alegre-RS, Brazil jaqx@mat.ufrgs.br
[2] DI - Instituto de Telecomunicações Universidade da Beira Interior, Covilhã, Portugal edgar@di.ubi.pt

Summary. A dynamic model of a network of coupled populations is developed. In each of the n sites in the network a local demographic model of Leslie type is used. Connection between locations is modeled through density-independent migration which is a function of age and also on the connecting sites. We perform a simulation illustrating a policy of reduction of excessive migration.

1 The Model

The model is divided in two parts. The first considering only the local dynamics while the second part deals with migratory aspects.

1.1 The Local Dynamics

We model the population of a region, state or province as a dynamic network of local populations. The local populations can be thought as being cities or villages. Within a local population, individuals are subject to survival and reproduction processes encompassing what we call local dynamics. Typical human demographic data based in life tables contain information about these two processes since the survivorship function $l(x)$ (the probability of survival from birth to age x) and the maternity function $m(x)$ (mean number of offspring per individual aged x per unit time) are basic entries in life tables. After the local dynamics the individuals are allowed to move to another location in the network and the cycle continues. We assume there are n local populations labeled as $1, 2, \ldots, n$. Within each city or village the population is divided into age classes of the same duration. Let $X_j^t = [x_{1j}^t x_{2j}^t \ldots x_{kj}^t]^T$ be the population vector of city j at time t. The entries $x_{ij}^t, i = 1, 2, \ldots, k$ represent the number of female individuals of age class i living in the location j at time t. The whole dynamical system, often called metapopulation model, consists of nk equations describing the time evolution of each cohort at each

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*,
Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_163,

location of the network. We will use discrete time equations, thus the time step ($t = 0, 1, 2, \ldots$) has the same duration of an age class.

For a while assume there is no migration between local populations. The local dynamics will be assumed to be described by a projection matrix model of Leslie type [5]. Individuals in any age class except the first at time step $t+1$ must be the survivors of the previous age class at time t. Thus the ageing of individuals in site j are accounted by

$$x_{ij}^{t+1} = p_{i-1j} x_{i-1j}^{t}, \quad i = 2, 3, \ldots, k, \tag{1}$$

where p_{i-1j} is the transition probability from age class $i-1$ to age class i in location j. In terms of the survivorship function, the transition probabilities are given by

$$p_{ij} = \frac{l_{i+1j}}{l_{ij}} = \frac{\int_{i}^{i+1} l_j(x)dx}{\int_{i-1}^{i} l_j(x)dx}, \quad i = 1, 2, \ldots, k-1, \tag{2}$$

where $l_j(x)$ is the probability of survival from birth to age x in the local population j. Newborns enter the population at the age class 1, thus for a fixed j

$$x_{ij}^{t+1} = \sum_{i=1}^{k} f_{ij} x_{ij}^{t}, \tag{3}$$

where the fertility coefficients f_{ij} , give the number of daughters per female in age class i at the location j that survive through the time step in which they were born. Since births occur continuously over the time step, according to [4] and [1], the fertility coefficients assume the form

$$f_{ij} = l_j(0.5)\left(\frac{m_{ij} + p_{ij} m_{i+1j}}{2}\right), \tag{4}$$

where $m_{ij} = \int_{i-1}^{i} m_j(x)dx$, and $m_j(x)$ is the maternity function the site j.

The discrete dynamical system given by (1) and (3), describe the time evolution of the population at site j in the absence of migration. These k scalar equations can be written in vector form

$$X_j^{t+1} = L_j X_j^{t}, \quad j = 1, 2, \ldots, n, \tag{5}$$

with the projection matrix of site j written as

$$L_j = \begin{bmatrix} f_{1j} & f_{2j} & f_{3j} & \cdots & f_{kj} \\ p_{1j} & 0 & 0 & \cdots & 0 \\ 0 & p_{2j} & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & \cdots & 0 & p_{k-1j} & 0 \end{bmatrix}, \tag{6}$$

Equations (5) and (6) describe the local dynamics of an isolated local population j. Interesting properties of (5) are extensively discussed in [1] while nonlinear version of (5) are studied in [6,7] and [8].

1.2 The Metapopulation Dynamics

We now incorporate movement into the model. During each time step, following the local dynamics, individuals are allowed to move to other locations in the network. The migration process has two basic components. The first is based only on the decision to leave the current city or village and it is described by the $k \times k$ diagonal matrix M_j given by

$$M_j = \begin{bmatrix} m_{1j} & 0 & \cdots & 0 \\ 0 & m_{2j} & \cdots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & m_{kj} \end{bmatrix}, \tag{7}$$

where m_{ij}, $i = 1, 2, \ldots, k$, represent the probability of an individual of age class i to leave location j. Thus, at time step t, the components of the vector $M_j L_J X_j^t \in \mathbb{R}^k$ list the number of individuals of each age class leaving location j at time t. The second basic component of the migration process is based on the decision of choosing where to go. Individuals leaving site location j have to distribute among the other $n-1$ locations and their preferences is expressed in the numbers c_{ij}, $i = 1, 2, \ldots, n$, representing the percentage of individuals that left location j and decided to establish themselves in the location i. Of course, $c_{ij} > 0$, and $c_{ii} = 0$ for all $i, j = 1, 2, \ldots, n$. This approach was considered in [2]. The $n \times n$ matrix C with entries c_{ij}, $i = 1, 2, \ldots, n$ is the interaction matrix of the network of n local populations. This implies that choosing where to settle does not depend on age. It may be reasonable for certain special cases of animal dispersal, but is certainly unrealistic in human demography. For example, some cities could be industrial towns therefore more attractive to adults, while college towns are more attractive to young people, while other cities could be more attractive to retired people and so forth. In fact, the entries c_{ij} depend on the age class, thus we write c_{ij} as a $k \times k$ diagonal matrix

$$C_{ij} = \begin{bmatrix} c_{ij}(1) & 0 & \cdots & 0 \\ 0 & c_{ij}(2) & \cdots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & c_{ij}(k) \end{bmatrix}, \quad i \neq j, \tag{8}$$

where the $(1), (2) \ldots, (k)$ represent the age classes, furthermore C_{ii} is the zero $k \times k$ matrix. Assuming that are no loss of persons during the migration process, we require that $\sum_{j=1}^{n} C_{ij} = I_k, j = 1, 2, \ldots, n$, where I_k is the $k \times k$ identity matrix. The components of the vector $C_{ij} M_j L_j X_j^t \in \mathbb{R}^k$ list the number of individuals in each age class that left location j to settle at location i at time t. Clearly,$(I_k - M_i) L_i X_i^t$ list the number of persons in each age class that did not leave location i at time t. Thus the evolution equations are

$$X_i^{t+1} = (I_k - M_i)L_i X_i^t + \sum_{j=1}^{n} C_{ij} M_j L_j X_j^t, \quad j = 1, 2, \ldots, n. \tag{9}$$

Equation (9) can be written in more concise way if we consider the metapopulation vector $X^t = \left[X_1^t \; X_2^t \cdots X_n^t \right]^T \in \mathbb{R}^{nk}$, the diagonal block matrix

$$L = \begin{bmatrix} L_1 & 0 & \cdots & 0 \\ 0 & L_2 & \cdots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & L_n \end{bmatrix}, \tag{10}$$

where each $k \times k$ block L_j is given by (6) , and the block matrix A, with the $k \times k$ block entries given by

$$A_{ij} = \begin{cases} I_k - M_i, & i = j \\ C_{ij} M_j, & i \neq j \end{cases} \tag{11}$$

The matrix L contains all the information on the survival and reproduction in each local population while the matrix A has the information about the movement of the individuals. Clearly we have

$$X^{t+1} = ALX^t \tag{12}$$

2 Simulations

Next we perform numerical simulations to obtain a projection of the network of populations in a realistic context. As an example, we use a network composed of six sites and divide the each population in fourteen age classes. First we assume the migration matrices according to the characteristics of each location. In a second stage we apply a reduction factor to reflect a policy of diminishing undesirable migration.

2.1 Computational Considerations

Despite the fact that (12) represents the dynamics in a rather simplified way, it is not very efficient for simulation purposes. We observe that in the jth block column of the product AL, the product $M_j L_j$ appears n times which means that in the ith block row $L_i X_i^t$ also occurs n times. In order to avoid unnecessary computations we developed an algorithm composed of 3 stages.

Algorithm

For $t = 0, 1, \ldots, p$ and initial data X_i^0, $i = 1, 2, \ldots, n$, do:

1. $Y_i^t = L_i X_i^t$, $i = 1, 2, \ldots, n$
2. $N_i = M_i Y_i^t$, $i = 1, 2, \ldots, n$
3. $X_i^{t+1} = Y_i^t - N_i + \sum_{j=1}^{n} C_{ij} N_j$, $i = 1, 2, \ldots, n$

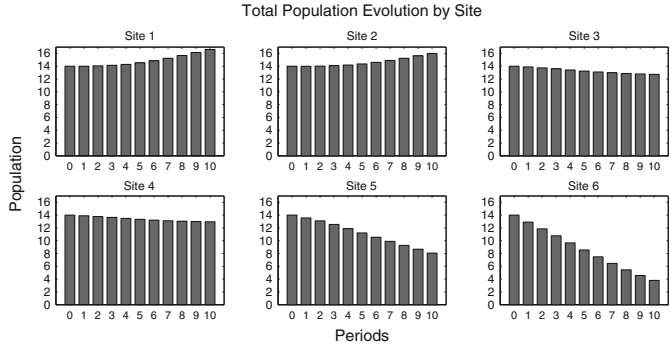


Fig. 1. Time evolution of the total population in each location considering all locations with same initial population. In the first two locations the population increases while in the other sites it decreases, specially in the last one where it is about to go extinct

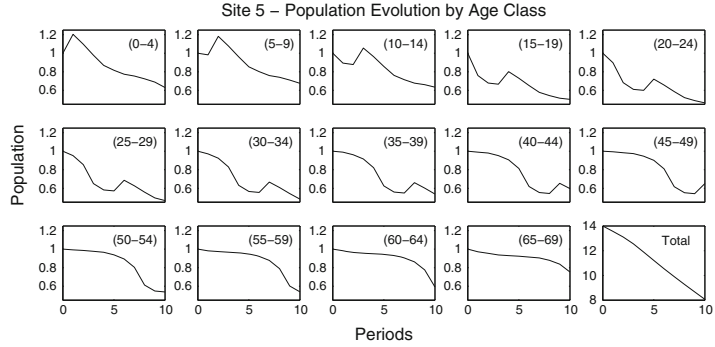


Fig. 2. Time evolution of all age classes. We observe travelling wave along the age classes due to the fact that the net reproductive rate is larger than one

2.2 Results

In our example $n = 6$, that is, we consider six locations classified as follows: 1-Large Metropolitan Area, 2- Industry Municipality, 3- University Municipality, 4- Medium Size Municipality, 5- Small Municipality I, and 6- Small Municipality II. The difference between locations 5 and 6 is that later consists of a population in danger of extinction due to excessive migration. For simplicity we assume that the fertilities and transition probabilities are location independent, which means we have only one Leslie matrix for the whole network, that is, $L_j = \overline{L}$, $j = 1, 2, \ldots, 6$. We choose parameter values such that the net reproductive rate is $R_0 = \sum_{i=1}^{14} f_i l_i = 1.1309$. Each of the 14 age classes consists of 5 year groups $(0 : 4, 5 : 9, \ldots, 65 : 69)$. We assume $p = 10$, therefore the projection is made for 50 years. In Fig. 1 the evolution of the total population in each of the 6 locations is shown while in Fig. 2 we depict the time evolution of each age class of location 5 for the 50 years.

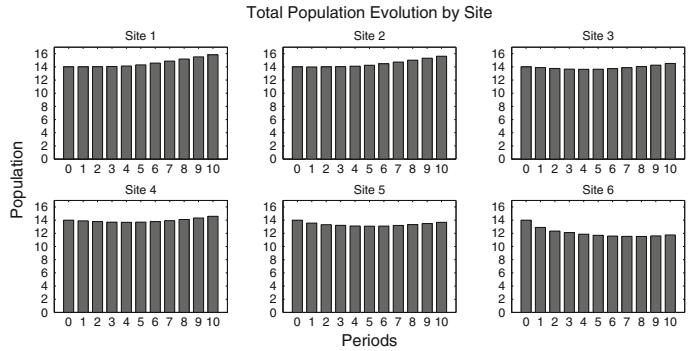


Fig. 3. Time evolution of the total population in each location after the reduction migration policy is applied

The policy of reducing migration is incorporated in the model by considering a factor $q_i^0 = 1$ and in each iteration we let $q_i^{t+1} = q_i^t - 0.1$, and $M_i^{t+1} = q_i^t M_i^t$. The results are illustrated in Fig. 3. Comparing with Fig. 1 we notice that the reducing migration policy can be very effective in preventing dangerous population reduction in certain locations.

3 Final Remarks

We presented the main aspects of the coupled population network dynamical cohort model framework. This can serve as basis for more complex models that include more realistic features as density-dependent migration and sex ratio different than 1:1. Metapopulation dynamical models involving age classes are of great importance in studies related to demographic projections, e.g. [3].

References

1. Caswell, H.: Matrix Population Models: Construction, Analysis, and Interpretation. Sinauer, Sunderland (1989)
2. DeCastro, M.L., Silva, J.A.L., Justo, D.A.R.: J. Math. Biol. **52**, 183–203 (2006)
3. Department of City Planning, New York City Population Projections by Age/Sex & Borough 2000–2030. The City of New York (2006)
4. Keyfitz, N.: Introduction to the Mathematics of Populations. Addison-Wesley, Reading (1968)
5. Leslie, P.H.: Biometrika **33**, 183–212 (1945)
6. Silva, J.A.L., Hallam, T.G.: Math. Biosci. **110**, 67–101 (1992)
7. Silva, J.A.L., Hallam, T.G.: J. Math. Biol. **31**, 367–395 (1993)
8. Wilkan, A., Mjflhas, E.: J. Theor. Biol. **173**, 109–119 (1995)

Two-Layer Shallow Water Equations with Complete Coriolis Force and Topography

A.L. Stewart and P.J. Dellar

Oxford Centre for Industrial and Applied Mathematics, Mathematical Institute, 24–29 St Giles', Oxford OX1 3LB, UK stewart@maths.ox.ac.uk, dellar@maths.ox.ac.uk

Summary. Equations are presented that describe two superposed shallow layers of inviscid fluid flowing over topography in a rotating frame, with a complete treatment of the Coriolis force. Motivated by applications to the Earth's equatorial oceans, these equations offer a physically reasonable alternative to the empirical friction currently used to regularise existing shallow water models at the equator.

1 Introduction

Layered shallow water equations describe the behaviour of several superposed layers of inviscid fluid of different, constant densities flowing over bottom topography, as illustrated in Fig. 1. This structure captures something of the density stratification of the oceans, which makes it a useful idealised setting for studying the interactions between stratification and rotation that govern the large-scale dynamics of the oceans [2, 4, 10].

Shallow water equations may be derived from the three-dimensional Euler equations by averaging the horizontal fluid velocity across each layer. The standard approach neglects the Coriolis terms due to the horizontal components of the Earth's rotation vector, and also the vertical acceleration. These are referred to as the 'traditional' and 'hydrostatic' approximations respectively [10, 12]. Both may be justified in the limit of a vanishingly small ocean depth. However, recent work [4] includes the 'nontraditional' components of the rotation vector in a single-layer shallow water model, and suggests that there may be significant effects associated with these components. This is consistent with the findings of the UK Meteorological Office, who in 1992 abandoned the traditional approximation in their unified model [3, 12].

This paper presents a set of two-layer shallow water equations that incorporate the nontraditional Coriolis terms, but omit the vertical acceleration. They thus correspond to a layered analogue of the quasihydrostatic approximation for a continuously stratified fluid [11, 12]. These equations are

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*, Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_164,

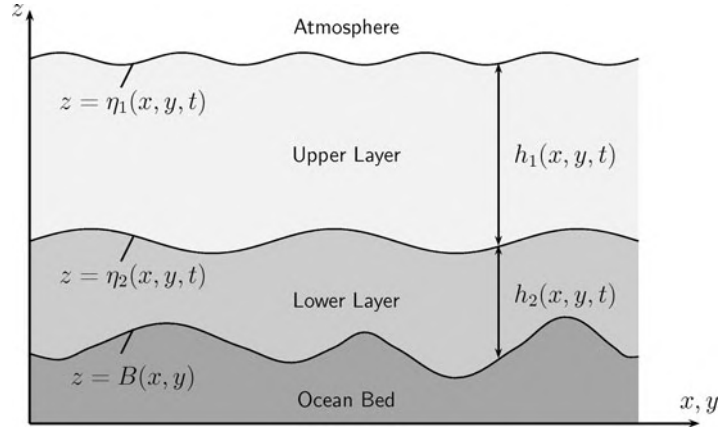


Fig. 1. Structure of the two-layer ocean model

particularly relevant for the study of deep ocean currents, such as the Antarctic Bottom Water, whose behaviour in equatorial regions is not well described by traditional models [2].

2 Formulation and Derivation

We begin our derivation by writing down the Euler equations for two fluid layers, each of constant density ϱ_i, in a rotating frame,

$$\frac{\partial \tilde{\mathbf{u}}_i}{\partial \tilde{t}} + \left(\tilde{\mathbf{u}}_i \cdot \tilde{\nabla}\right) \tilde{\mathbf{u}}_i + \tilde{w}_i \frac{\partial \tilde{\mathbf{u}}_i}{\partial \tilde{z}} + 2\,\tilde{\Omega}_z \tilde{\mathbf{z}} \times \tilde{\mathbf{u}}_i + 2\,\tilde{\mathbf{\Omega}} \times \hat{\mathbf{z}}\,\tilde{w}_i + \frac{1}{\varrho_i} \tilde{\nabla} \tilde{p}_i = 0, \quad \text{(1a)}$$

$$\frac{\partial \tilde{w}_i}{\partial \tilde{t}} + \tilde{\mathbf{u}}_i \cdot \tilde{\nabla} \tilde{w}_i + \tilde{w}_i \frac{\partial \tilde{w}_i}{\partial \tilde{z}} + 2\big(\tilde{v}_i \tilde{\Omega}_x - \tilde{u}_i \tilde{\Omega}_y\big) + \frac{1}{\varrho_i} \frac{\partial \tilde{p}_i}{\partial \tilde{z}} + g = 0, \quad \text{(1b)}$$

$$\tilde{\nabla} \cdot \tilde{\mathbf{u}}_i + \frac{\partial \tilde{w}_i}{\partial \tilde{z}} = 0. \quad \text{(1c)}$$

Here $i = 1{,}2$ denotes the upper and lower layers respectively, and the superscript tildes ($\tilde{}$) indicate dimensional variables. The horizontal velocity within each layer is $\tilde{\mathbf{u}}_i = (\tilde{u}_i, \tilde{v}_i)^T$, the vertical velocity is $\tilde{w}_i$, and the pressure is $\tilde{p}_i$. These quantities all depend on $\tilde{x}$, $\tilde{y}$, $\tilde{z}$ and $\tilde{t}$, but $\tilde{\nabla} = (\partial_{\tilde{x}}, \partial_{\tilde{y}})$ is a horizontal derivative. The gravitational acceleration is g, and $\tilde{\mathbf{\Omega}} = (\tilde{\Omega}_x, \tilde{\Omega}_y)^T$ and $\tilde{\Omega}_z$ are the horizontal and vertical components of the rotation vector.

In applying this model we approximate the curved surface of the Earth using a flat plane. The Cartesian coordinates are constructed such that the combination of centrifugal acceleration and gravity acts vertically [10], as represented by the g term in (1b). However, we allow for the spatial variation of the rotation vector with latitude, the so-called β-plane approximation [7, 10]. We thus consider $\tilde{\mathbf{\Omega}} = \tilde{\mathbf{\Omega}}(\tilde{x}, \tilde{y})$ and $\tilde{\Omega}_z = \tilde{\Omega}_z(\tilde{x}, \tilde{y}, \tilde{z})$. In general,

$\tilde{\Omega}_z$ must depend on $\tilde{z}$ to make the three-dimensional rotation vector non-divergent, $\tilde{\nabla} \cdot \tilde{\mathbf{\Omega}} + \partial_z \tilde{\Omega}_z = 0$. This ensures conservation of potential vorticity [7]. Integrating with respect to $\tilde{z}$ yields the following expression for $\tilde{\Omega}_z$, where $\tilde{\Omega}_{z0} = \tilde{\Omega}_z|_{\tilde{z}=0}$,

$$\tilde{\Omega}_z(\tilde{x}, \tilde{y}, \tilde{z}) = \tilde{\Omega}_{z0}(\tilde{x}, \tilde{y}) - (\tilde{\nabla} \cdot \tilde{\mathbf{\Omega}})\tilde{z}. \tag{2}$$

We assume that the upper surface is stress-free ($\tilde{p}_1 = 0$ on $\tilde{z} = \tilde{\eta}_1$) and that the pressure is continuous at the internal surface ($\tilde{p}_1 = \tilde{p}_2$ on $\tilde{z} = \tilde{\eta}_2$). However, we allow for a discontinuous horizontal fluid velocity between the layers, so the kinematic boundary conditions become,

$$\begin{gathered} \tilde{w}_2 = \tilde{\mathbf{u}}_2 \cdot \tilde{\nabla} \tilde{B} \quad \text{on} \quad \tilde{z} = \tilde{B}, \qquad \tilde{w}_1 = \frac{\partial \tilde{\eta}_1}{\partial \tilde{t}} + \tilde{\mathbf{u}}_1 \cdot \tilde{\nabla} \tilde{\eta}_1 \quad \text{on} \quad \tilde{z} = \tilde{\eta}_1, \\ \tilde{w}_2 - \tilde{\mathbf{u}}_2 \cdot \tilde{\nabla} \tilde{\eta}_2 = \frac{\partial \tilde{\eta}_2}{\partial \tilde{t}} = \tilde{w}_1 - \tilde{\mathbf{u}}_1 \cdot \tilde{\nabla} \tilde{\eta}_2 \quad \text{on} \quad \tilde{z} = \tilde{\eta}_2. \end{gathered} \tag{3}$$

We now derive the two-layer shallow water equations by averaging the Euler equations over each layer. We follow a procedure similar to that described in [1, 4], nondimensionalising the governing equations and introducing $\delta = H/L$, the ratio of the vertical to horizontal length scales. We shall take $\delta \ll 1$ below. The resulting set of dimensionless equations is

$$Ro \left(\frac{\partial \mathbf{u}_i}{\partial t} + (\mathbf{u}_i \cdot \nabla) \mathbf{u}_i + w_i \frac{\partial \mathbf{u}_i}{\partial z} \right) + \Omega_z \hat{\mathbf{z}} \times \mathbf{u}_i + \delta \mathbf{\Omega} \times \hat{\mathbf{z}} w_i + \nabla p_i = 0, \tag{4a}$$

$$\delta^2 Ro \left(\frac{\partial w_i}{\partial t} + \mathbf{u}_i \cdot \nabla w_i + w_i \frac{\partial w_i}{\partial z} \right) + \delta(v_i \Omega_x - u_i \Omega_y) + \frac{\partial p_i}{\partial z} + Bu = 0, \tag{4b}$$

$$\nabla \cdot \mathbf{u}_i + \frac{\partial w_i}{\partial z} = 0. \tag{4c}$$

Here $Ro = U/(2\Omega L)$ and $Bu = gH/(2\Omega UL)$ are the Rossby and Burger numbers respectively, which we assume are both O(1). Exploiting $\delta \ll 1$ for shallow layers, we pose asymptotic expansions of $\mathbf{u}_i$, w_i and p_i in the form $\mathbf{u}_i = \mathbf{u}_i^{(0)} + \delta \mathbf{u}_i^{(1)} + \cdots$. Equation (4b) implies that the leading-order pressure $p^{(0)}$ is hydrostatic, so (4a) is satisfied at leading order by a z-independent $\mathbf{u}_i^{(0)}$. We may thus obtain expressions for $w_i^{(0)}$ and $p_i^{(0)} + \delta p_i^{(1)}$ from (4c) and (4b) respectively. The lower layer acquires a contribution to its pressure from the upper layer, which we find from the boundary condition at the interface, $p_2^{(0)} + \delta p_2^{(1)} = (\varrho_1/\varrho_2)(p_1^{(0)} + \delta p_1^{(1)})$ $z = \eta_2$. Similarly, the vertical velocity in the upper layer acquires a contribution from the vertical velocity in the lower layer, $w_1^{(0)} = w_2^{(0)} - \mathbf{u}_2^{(0)} \cdot \nabla \eta_2 + \mathbf{u}_1^{(0)} \cdot \nabla \eta_1$ on $z = \eta_2$.

Applying the layer averaging formula from [13] to the continuity equation (4c), we derive evolution equations for the layer depths,

$$\frac{\partial h_i}{\partial t} + \nabla \cdot (h_i \overline{\mathbf{u}}_i) = 0, \tag{5}$$

where an overbar ($\bar{}$) denotes a layer average,

$$\overline{\mathbf{u}}_1 = \frac{1}{h_1}\int_{\eta_2}^{\eta_1} \mathbf{u}_1 \mathrm{d}z, \quad \overline{\mathbf{u}}_2 = \frac{1}{h_2}\int_{B}^{\eta_2} \mathbf{u}_2 \mathrm{d}z. \tag{6}$$

The depths of the two layers are $h_1 = \eta_1 - \eta_2$ and $h_2 = \eta_2 - B$.

Averaging (4a) over each layer, as described in [1,4], we obtain

$$\begin{aligned} Ro\left(\frac{\partial}{\partial t}\left(h_i\overline{\mathbf{u}}_i\right) + \nabla\cdot\left(h_i\overline{\mathbf{u}_i\mathbf{u}_i}\right)\right) + h_i\hat{\mathbf{z}}\times\overline{\Omega_z\mathbf{u}_i} \\ + \delta\boldsymbol{\Omega}\times\hat{\mathbf{z}}h_i\overline{w_i^{(0)}} + h_i\overline{\nabla\left(p_i^{(0)} + \delta p_i^{(1)}\right)} = \mathrm{O}(\delta^2). \end{aligned} \tag{7}$$

The average pressure gradient may be computed from $p_i^{(0)} + \delta p_i^{(1)}$, as found above. To complete the derivation, we note that we may factorise the averages of products of quantities that are z-independent to leading order [1,9]. Since $\mathbf{u}_i = \mathbf{u}_i^{(0)} + \mathrm{O}(\delta)$, $\overline{\mathbf{u}_i\mathbf{u}_i} = \overline{\mathbf{u}}_i\overline{\mathbf{u}}_i + \mathrm{O}(\delta^2)$. Similarly, $\overline{\Omega_z\mathbf{u}_i} = \overline{\Omega_z}\overline{\mathbf{u}}_i + \mathrm{O}(\delta^2)$, we may evaluate $\overline{\Omega_z}$ using (2), and $\mathbf{u}_i^{(0)} = \overline{\mathbf{u}}_i + \mathrm{O}(\delta)$. Neglecting terms of $\mathrm{O}(\delta^2)$ and above, we obtain the following averaged momentum equations,

$$\begin{aligned} Ro\left(\frac{\partial\mathbf{u}_1}{\partial t} + (\mathbf{u}_1\cdot\nabla)\mathbf{u}_1\right) + \left[\Omega_{z0} - \delta\nabla\cdot\left(\left(B + h_2 + \tfrac{1}{2}h_1\right)\boldsymbol{\Omega}\right)\right]\hat{\mathbf{z}}\times\mathbf{u}_1 \\ + \nabla\left[Bu\left(B + h_2 + h_1\right) + \tfrac{1}{2}\delta h_1(v_1\Omega_x - u_1\Omega_y)\right] \\ - \delta\,\boldsymbol{\Omega}\times\hat{\mathbf{z}}\,\nabla\cdot\left(h_2\mathbf{u}_2 + \tfrac{1}{2}h_1\mathbf{u}_1\right) = 0, \end{aligned} \tag{8a}$$

$$\begin{aligned} Ro\left(\frac{\partial\mathbf{u}_2}{\partial t} + (\mathbf{u}_2\cdot\nabla)\mathbf{u}_2\right) + \left[\Omega_{z0} - \delta\nabla\cdot\left(\left(B + \tfrac{1}{2}h_2\right)\boldsymbol{\Omega}\right)\right]\hat{\mathbf{z}}\times\mathbf{u}_2 \\ + \nabla\big[Bu\left(B + h_2 + \varrho_r h_1\right) + \tfrac{1}{2}\delta h_2(v_2\Omega_x - u_2\Omega_y) \\ + \delta\varrho_r h_1\left(v_1\Omega_x - u_1\Omega_y\right)\big] - \delta\,\boldsymbol{\Omega}\times\hat{\mathbf{z}}\,\nabla\cdot\left(\tfrac{1}{2}h_2\mathbf{u_2}\right) = 0. \end{aligned} \tag{8b}$$

Here we have dropped the overbars on averaged velocities, and introduced the density ratio $\varrho_r = \varrho_1/\varrho_2$. We thus obtain the traditional two layer shallow water equations [10] plus several additional terms proportional to Ω_x and Ω_y.

3 Conservation Properties

The 'nontraditional' two-layer shallow water equations inherit the conservation laws of the full three-dimensional equations. In particular, there are two materially conserved potential vorticities, $\partial_t q_i + \mathbf{u}_i\cdot\nabla q_i = 0$ for $i = 1{,}2$, with

$$q_i = \frac{1}{h_i}\left\{\left[\Omega_{z0} - \delta\nabla\cdot\left(\left(\eta_i - \frac{h_i}{2}\right)\boldsymbol{\Omega}\right)\right] + Ro\left(\frac{\partial v_i}{\partial x} - \frac{\partial u_i}{\partial y}\right)\right\}. \tag{9}$$

These q_i differ by terms proportional to Ω_x and Ω_y from the traditional potential vorticities given in [10]. This modification may provide useful insight into

the dynamics of cross-equatorial ocean currents. The contributions from Ω_{z0} change sign at the equator, which severely constrains the ability of fluid parcels to cross the equator [8]. This constraint may be at least partly alleviated by the additional contribution to the q_i from the interaction of topography and nontraditional rotation.

Conservation laws for the energy and momentum of the fluid may also be obtained. The energy density is unchanged by rotation, whilst the energy flux and momentum density acquire additional terms containing nontraditional components of the rotation vector.

4 Linear Plane Waves

Some important properties of the extended shallow water equations may be highlighted by considering linear plane wave solutions. Taking the usual GFD axes (y pointing north, x pointing east) and a non-traditional f-plane approximation [10], such that $\Omega_x = 0$ and Ω_y and Ω_z are constants, we linearise (8a), (8b) and (5) by assuming that the dependent variables are small perturbations to a state of rest: $\mathbf{u}_1 = \mathbf{u}_1'$, $\mathbf{u}_2 = \mathbf{u}_2'$, $h_1 = H_1 + h_1'$ and $h_2 = H_2 + h_2'$. By neglecting products of these variables and seeking solutions of the form $\exp(\mathrm{i}(kx + ly - \omega t))$, we obtain a dispersion relation for the waves.

This dispersion relation is plotted in Fig. 2. We have taken the layers to be of equal mean depth ($H_1 = H_2$), and the aspect ratio to be $\delta = 0.2$, a little larger than typical for long internal waves ($0.02 \lesssim \delta \lesssim 0.14$). Our density ratio is $\varrho_r = 0.9$. The realistic value $\varrho_r = 0.98$ makes it impossible to show the two wave branches on the same plot. We have also set $Ro = Bu = 1$ for the purpose of illustration. The waves with higher frequency propagate on the internal and upper surfaces simultaneously. The lower frequency waves propagate primarily on the interface between the layers, with the upper surface remaining approximately flat. The nontraditional Coriolis effects cause a distinct shift in the frequencies, creating a left/right asymmetry. More importantly, nontraditional effects create a range of waves with frequencies below the inertial frequency, the smallest allowable frequency under the traditional approximation. These so-called subinertial waves have been observed in previous studies of nontraditional Coriolis effects in continuously stratified fluids [5,6], and provide an additional source of energy for mixing in the deep ocean.

5 Conclusions

We have derived two-layer shallow water equations that include additional terms arising from the nontraditional components of the Coriolis force. They may be shown to retain the expected conservation laws for energy, momentum, and potential vorticity. We have illustrated some deviations from traditional behaviour, such as the existence of subinertial waves, caused by the additional

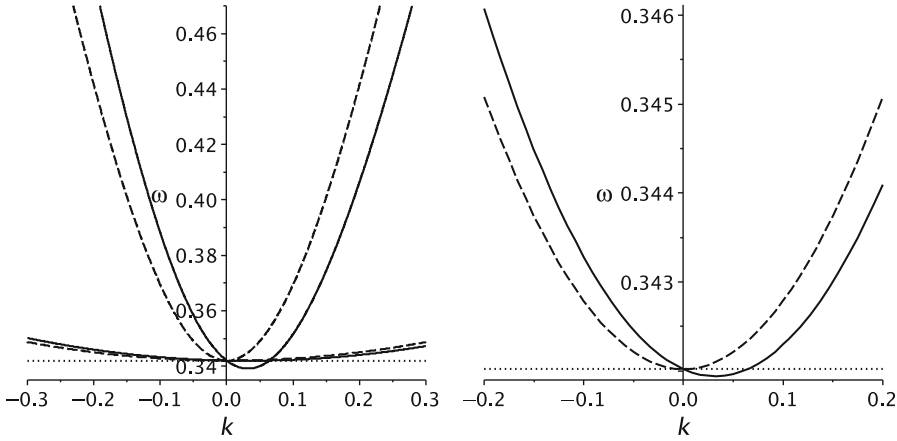


Fig. 2. Dispersion relation for waves propagating zonally at a latitude of 20° North in the traditional (*dashed line*) and nontraditional (*solid line*) two-layer shallow water equations. *Left*: all wave modes. *Right*: internal wave modes. Notice the band of internal waves with frequencies below the inertial frequency (*dotted line*)

part of the Coriolis force. These equations will serve as a useful prototype for investigating the dynamics of cross-equatorial ocean currents over topography.

Acknowledgement

ALS is supported by an EPSRC Doctoral Training Account award. PJD's research is supported by an EPSRC ARF, grant number EP/E054625/1.

References

1. Camassa, R., Holm, D.D., Levermore, C.D.: J. Fluid Mech. **349**, 173–189 (1997)
2. Choboter, P.F., Swaters, G.E.: Canad. Appl. Math. Quart. **8**, 367–385 (2000)
3. Cullen, M.J.P.: Met. Mag. **122**, 81–94 (1993)
4. Dellar, P.J., Salmon, R.: Phys. Fluids **17**, 106601–19 (2005)
5. Gerkema, T., Shrira, V.I.: J. Fluid Mech. **529**, 192–219 (2005)
6. Gerkema, T., Zimmerman, J.T.F., Maas, L.R.M., van Haren, H.: Rev. Geophys. **46**, 2004–33 (2008)
7. Grimshaw, R.: Tellus **27**, 351–357 (1975)
8. Stommel, H., Arons, A.B.: Deep-Sea Res. **6**, 217–233 (1960)
9. Su, C.H., Gardner, C.S.: J. Math. Phys. **10**, 536–539 (1969)
10. Vallis, G.K.: Atmospheric and Oceanic Fluid Dynamics: Fundamentals and Large-Scale Circulation. Cambridge University Press, Cambridge (1996)
11. White, A.A., Bromley, R.A.: Quart. J. Roy. Met. Soc. **121**, 399–418 (1995)
12. White, A.A., Hoskins, B.J., Roulstone, I., Staniforth, A.: Quart. J. Roy. Met. Soc. **131**, 2081–2107 (2005)
13. Wu, T.Y.: J. Eng. Mech. Div. ASCE **107**, 501–522 (1981)

Optimising for Wind Power Contributions in an Electricity Grid

Winston L. Sweatman[1], Geoff Pritchard[2], Bill Whiten[3], Mike Camden[4], and Kim Nan[2]

[1] Institute of Information and Mathematical Sciences, Massey University, Auckland, New Zealand w.sweatman@massey.ac.nz
[2] Department of Statistics, University of Auckland, New Zealand
[3] University of Queensland, Australia
[4] Statistics New Zealand, New Zealand

Summary. This paper is based on a Mathematics-in-Industry Study Group project from MISG 2007 in Wollongong. This project considered management issues for a national electric power grid that may arise as a result of using larger amounts of wind power generation. The variability of this power source has implications for both the maintenance of power supply and its transmission upon finite capacity power lines. A number of approaches and simple models were used to study these aspects of production and transmission.

1 Introduction

Wind power generation in New Zealand is anticipated to rise over the next decade in response to the demands for renewable energy and low carbon emissions. This will introduce a number of challenges due to the variability and degree of uncertainty in this power source. A Mathematics-in-Industry Study Group at MISG 2007 in Wollongong, sponsored by Transpower and the Energy Efficiency and Conservation Authority, New Zealand, considered several aspects of the problem [2]. Here, we summarise some approaches using simple models to explore the effect of variability on first production and secondly transmission. The problems are treated as being essentially economic in nature, we are seeking an optimum solution.

2 Optimising Generation Across Three Different Sources

To explore the effect of variability on power production, we consider three power generators in close proximity to a constant load so that there are no significant transmission losses. The generators consist of a wind farm whose

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*, Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_165,

power cost is taken to be zero with fluctuations that take immediate effect; a low-cost (thermal) generator with cost c_ℓ per unit time that can only adjust at ramp rate r to new power levels, and a fast-ramping but high-cost (hydro) generator with cost c_h per unit time and essentially instantaneous ramping. The power outputs of the generators are adjusted to meet the load while at the same time attempting to minimise the cost of the long-term operation. In general, it is assumed that changes in the wind are not anticipated although this could be added to the models.

If wind-power was constant then the sourcing of power would be straightforward, however, in practice we anticipate change and alternative strategies may be more cost effective. The operation of the power generators can be divided into three main states:

- Adjustment to an increase in wind-power.
- Adjustment to a decrease in wind-power.
- Steady state.

Increases and decreases in potential wind power are fundamentally different: it is possible to spill additional wind without altering the combination of power sources, however, if wind-power is fully-utilised and then decreases, the difference must be met from somewhere else. Using additional low-cost power can provide a buffer for sharp decreases at the expense of not using the full potential wind power. Using high-cost power to replace some low-cost power enables more rapid utilisation of increases in wind power.

2.1 The Initial Model

As a simple model to illustrate some of the effects of wind-power variability, the potential wind-power generation P_w is taken to fluctuate between the values where it can meet the entire load L and zero. We consider a single upwards change in wind power followed by a downwards change during a time period T:

$$P_w(t) = \begin{cases} L & \text{for } 0 < t < \beta \\ 0 & \text{for } \beta < t < T \end{cases} \tag{1}$$

Although focused upon this single time interval T, for further simplicity we treat the pattern as if it were periodic ($P_w(t) = P_w(t + T)$), i.e. a square wave. Supplies of low-cost and high-cost power are taken to be sufficient to meet the load. This model captures the effect that the wind-power generation varies rapidly and will only be available for a proportion of the time.

Depending upon the prices of the different sources of power it may not be cost effective to use all potential wind power. We take h to denote the maximum amount of wind-power that is used and seek to optimise this.

Denoting wind power used P_u, low-cost power P_ℓ and high-cost power P_h, for relatively expensive high-cost power and a relatively rapid slow-ramping rate, the solution is

$$[P_u, P_\ell, P_h](t) = \begin{cases} [rt, L - rt, 0], & \text{if } 0 < t < h/r \\ [h, L - h, 0], & \text{if } h/r < t < \beta \\ [0, L - h + r(t - \beta), h - r(t - \beta)], & \text{if } \beta < t < \beta + h/r \\ [0, L, 0], & \text{if } \beta + h/r < t < T \end{cases}$$

$$[P_u, P_\ell, P_h](t) = [P_u, P_\ell, P_h](t + T). \tag{2}$$

Initially, the entire load is met by the low-cost generator. This is ramped down and replaced with wind power on the onset of the wind. On the cessation of wind the high-cost power must meet the immediate demand until the low-cost generator can ramp up. The total cost over the time period T is

$$c_\ell(LT - h\beta) + c_h h^2/(2r), \tag{3}$$

with a minimum at $h = r\beta c_\ell/c_h$, corresponding to a total cost:

$$c_\ell LT - c_\ell^2 \beta^2 r/(2c_h). \tag{4}$$

If the difference between low- and high-cost generators is too small (specifically $c_\ell T < c_h(T - \beta)$) then it is cheaper to switch directly between wind and high-cost power with no use of low-cost power. If the slow ramping is too slow (i.e. either $h/r > \beta$ or $h/r > T - \beta$, or both), then the total cost becomes a linear function of h and reaches its extreme values at the limits of the domain. For no ramping $r = 0$, the optimum value of h will be either 0 or L corresponding to per period costs of $c_\ell LT$ or $c_h L(T - \beta)$, respectively.

The simple square-wave model can also be used to illustrate savings if changes in wind generation can be predicted. Excluding the cases considered above where the slow-ramping is too slow or the high-cost power too cheap, suppose that the rises in wind power are anticipated (but not the falls). Then high-cost power can be used to replace h_2 of the low-cost power prior to the onset of the wind so that h_2 of wind power can be used immediately. The modified cost function is

$$c_\ell(LT - h\beta - hh_2/r) + c_h(h^2 + h_2^2)/(2r) \tag{5}$$

with a minimum at a higher value $h = r\beta c_\ell/(c_h - c_\ell^2/c_h)$ with $h_2 = c_\ell h/c_h$ and total cost $c_\ell LT - c_\ell^2\beta^2 r/(2(c_h - c_\ell^2/c_h))$. This is a saving of $c_\ell^4\beta^2 r/(2c_h(c_h^2 - c_\ell^2))$ per period T over not anticipating the wind.

2.2 Optimum Settings for More General Wind-Power Forms

Now we consider changes in wind power which occur as a random process. Initially we assume a single step change in wind power ΔP_w in a period T.

If high-cost power is sufficiently expensive then using additional low-cost power (ΔP_ℓ) to replace some wind power is advantageous as it provides a buffer for sudden drops in wind power($\Delta P_w < 0$). In this case, the total cost is

$$c_\ell \int_0^T P_\ell(t)dt + c_\ell \Delta P_\ell T + \frac{c_h}{2r}(-\Delta P_w - \Delta P_\ell)^2. \tag{6}$$

This has a minimum value when

$$\Delta P_\ell = -\Delta P_w - rTc_\ell/c_h. \tag{7}$$

The optimum ΔP_ℓ must always lie between zero and $|\Delta P_w|$ and also depends on the time interval T. For longer time intervals between drops in wind power, less excess low-cost power is justified.

Generalising this case for different sizes of step change over the time interval T, we assume that for step changes such that $\Delta P_w < -\Delta P_\ell$ the size is determined by the probability: $\mathrm{P}(\Delta P_w(u) : \Delta P_w(u) < -\Delta P_\ell; T)$. Instead of (6), the total cost is now:

$$c_\ell \textstyle\int_0^T P_\ell(t)dt + c_\ell \Delta P_\ell T + \\ \frac{c_h}{2r} \int_0^1 \mathrm{P}(\Delta P_w(u) : \Delta P_w(u) < -\Delta P_\ell, T)(-\Delta P_w(u) - \Delta P_\ell)^2 du. \tag{8}$$

To obtain an approximate minimum cost, the probability term is assumed essentially constant, whence by differentiation with respect to ΔP_ℓ, and equating to zero, we obtain:

$$\Delta P_\ell = \frac{-\int_0^1 \mathrm{P}(\Delta P_w(u) : \Delta P_w(u) < -\Delta P_\ell, T)\Delta P_w(u)du - rTc_\ell/c_h}{\int_0^1 \mathrm{P}(\Delta P_w(u) : \Delta P_w(u) < -\Delta P_\ell, T)du}. \tag{9}$$

The first contribution on the right-hand side is the average size of the larger step changes, while the second adjusts to account for the likelihood of a downwards step change. For specific cases, this implicit equation for ΔP_ℓ is readily solved numerically.

To take advantage of the rapid increases in wind-power generation ($\Delta P_w > 0$), some low-cost power may be replaced with high-cost power ΔP_h if the price difference is not too large. For $\Delta P_h < \Delta P_w$ the total cost is:

$$c_\ell \int_0^T P_\ell(t)dt - c_\ell \Delta P_h T + c_h \Delta P_h T - \frac{c_h}{2r}\Delta P_w^2 + \frac{c_h}{2r}(\Delta P_w - \Delta P_h)^2. \tag{10}$$

Using differentiation, the value of ΔP_h at the minimum is:

$$\Delta P_h = \Delta P_w - r(c_h - c_\ell)T/c_h. \tag{11}$$

This always gives $\Delta P_h < \Delta P_w$ when $c_h > c_\ell$.

Now consider a cycle of a step up followed by a step down. Suppose in both cases that there is time for the low-cost power to adjust ($T > 2\Delta P_w/r$). From (11)

$$\Delta P_h < \Delta P_w(2c_\ell/c_h - 1). \tag{12}$$

This implies that for a positive ΔP_h we require $c_h < 2c_\ell$.

When $\Delta P_w(u)$ is given by a probability distribution we have:

$$\Delta P_h = \frac{\int_0^1 \mathrm{P}(\Delta P_w(u) : \Delta P_w(u) < \Delta P_h, T)\Delta P_w(u)du - r(c_h - c_\ell)T/c_h}{\int_0^1 \mathrm{P}(\Delta P_w(u) : \Delta P_w(u) < \Delta P_h, T)du} \quad (13)$$

and so ΔP_h is always bounded above by a weighted sum of the ΔP_w values.

The quantity ΔP_h gives the target for the low-cost generator power reduction. The low-cost generator moves towards this at its limited slow ramp rate.

In [2], an illustration is given of how these formulae may be combined with numerical simulations of wind-power generation to estimate the optimum strategies for power production. This work used models from [3] and a standard wind-power relationship, however, more recent models for wind-power generation are presented in [1]. The probability distribution of wind power over the time interval T is estimated. From this, (9) and (13) are solved to obtain target values of ΔP_ℓ and ΔP_h. As the low-cost generator may need to be ramped towards the target, any deficit is made up first by wind power, if available, and then by high-cost power. The effect of T can be determined by using a number of repeats with the same random wind profile.

3 Optimal Dispatch on a Network of Finite Capacity

Apart from the problem of providing electric power generation, there is also the issue of ensuring its transmission across the power grid network which has a limited capacity. As before, it is assumed that there are three kinds of power generation: wind, fast-ramping and slow-ramping. The cost of wind power is again taken as zero, however, in this model the other power generators offer power in tranches with different prices. Changes in the output of wind power and fast-ramping generators are again taken to be instantaneous, however, the slow-ramping power stations are taken to have zero ramp rate $r = 0$ and their dispatch is maintained constant during the time period T considered.

For illustration we consider a two-node network connected with a lossless transmission line of capacity M. (Further examples are contained in [2].) At node N_L, time-varying wind power $P_w(t)$ is offered at zero price (marginal cost), and a slow-ramping station (Thermal 1) offers unlimited power at price c_a. On the other side of the transmission line, a load L is met at node N_R. Other generators also supply power directly at node N_R: a fast-ramping (hydro) station offers a relatively cheap tranche (H_1) of quantity P_H at price c_b, and a more expensive tranche (H_2) at price c_e. Another slow-ramping station (Thermal 2) offers unlimited quantities at price c_d. We assume $M < L$, $0 < c_a < c_d$, and $0 < c_b < c_d < c_e$.

If the wind were constant, the least-cost solution would dispatch the generators in the order of their offer prices, starting with the cheapest. Once all the wind power had been allocated, Thermal 1 could be used to the extent

allowed by the transmission capacity and the remaining balance is met from the cheaper hydro power (H_1) and then by Thermal 2, both of which are located next to the load.

However, we should consider possible rises in wind power. If the transmission capacity is fully used and the wind output subsequently rises then the excess cannot be used and the wind must be spilt. To allow for this possibility it may be more cost-effective to leave some unused capacity or headroom x in the transmission line, by reducing the dispatch from Thermal 1 (by x) and instead using power from Thermal 2 (which is located adjacent to the load). A subsequent rise in potential wind power may then be used to displace hydro generation H_1.

Decreases in wind-power are also possible. However, for the network illustration here, this problem does not involve the transmission line and is essentially of the same kind as considered in Sect. 2. (To avoid increased hydro generation from the expensive H_2 tranche, one may reserve spare capacity directly within the H_1 hydro tranche, using extra power from Thermal 2 if necessary.)

Suppose there is a headroom $x \geq 0$ in the transmission line. Then the ongoing additional cost (per unit time) is $(c_d - c_a)x$, relative to the constant wind solution ($x = 0$). At a time $t > 0$, the instantaneous additional cost is

$$f(t) = c_e(P_w(0) - P_w(t))_+ - c_b \min(x, (P_w(t) - P_w(0))_+), \tag{14}$$

where the notation z_+ denotes max $(z, 0)$. At time 0, the expected average cost (per unit time) over the time interval $0 \leq t \leq T$ is:

$$\begin{aligned} C(x) &= \mathrm{E}\left[\frac{1}{T}\int_0^T \left((c_d - c_a)x + f(t)\right)\, dt\right] \\ &= \mathrm{E}\left[(c_d - c_a)x + c_e(-\delta)_+ - c_b \min(x, \delta_+)\right], \end{aligned}$$

where $\delta = P_w(\tau) - P_w(0)$, with τ a random variable independent of $(P_w(t))$ and distributed uniformly on $[0, T]$.

For simplicity $P_w(t)$ is chosen to have a continuous probability distribution. Then, for $x > 0$,

$$\frac{d}{dx}C(x) = (c_d - c_a) - c_b\,\mathrm{E}\left[\frac{d}{dx}\min(x, \delta_+)\right] = (c_d - c_a) - c_b\,\mathrm{P}(\delta > x).$$

If $\mathrm{P}(\delta > 0) \leq (c_d - c_a)/c_b$ then $C(x)$ increases with x for all $x > 0$, and so the cheapest option is to have zero headroom ($x = 0$). Otherwise, the optimal x is given by the solution to

$$\mathrm{P}(\delta > x) = (c_d - c_a)/c_b. \tag{15}$$

In some situations it may be reasonable to assume that the wind-power has a similar likelihood of rising as of falling ($\mathrm{P}(\delta > 0) = \frac{1}{2}$). It is only worth reserving headroom then if $c_d - c_a < c_b/2$ as the first unit of headroom costs $c_d - c_a$ to create, whereas, for half of the time, the average cost of not having this headroom is c_b.

4 Conclusions

In the early part of this paper (Sect. 2) we consider a simple model of power generation to meet a load using different sources. These include a wind farm taken to have negligible relative cost and a power output with sudden changes, a high-cost generator that can change its output as rapidly, and a low-cost generator that can only ramp slowly to a new power output. For a sufficiently large gap between high-cost and low-cost prices, there can be a significant benefit in scheduling a margin of low-cost power to buffer for possible drops in wind-power generation. Preparation to better utilise a possible increase in wind power by the replacement of some low-cost power with high-cost power, is only of an advantage when the prices are close.

In Sect. 3, we instead consider the effect of limited capacity in the transmission network. An illustration is used to show how variation in wind-power output may be planned for by reserving spare capacity in the transmission line.

References

1. Allwright, D., Andrade, V., Avramidis, T., Dewynne, J., Howison, S., Lacey, A., Licea, A., Patidar, S., Pototsky, A., Smith, H., Tambue, A.: Simulating the distribution and cross-correlation of wind farm output. 64th European Study Group with Industry 2008, Final Report (2008)
2. Pritchard, G., Sweatman, W.L., Nan, K., Camden, M., Whiten, W.: Maximizing the contribution of wind power in an electric power grid. In: Merchant, T., Edwards, M., Mercer, G. (eds.) Proceedings of the 2007 Mathematics and Statistics in Industry Study Group, pp. 114–139. University of Wollongong, Wollongong University, Australia (2008)
3. Whiten, W., Tsoularis, T.: Prediction of power generation from a wind farm. In: Wake, G. (ed.) Proceedings of the 2004 Mathematics and Statistics in Industry Study Group, pp. 61–87. Massey University, Albany, New Zealand (2005)

A Novel Solution Method for Tokamak Plasma Force Balance

A. Thyagaraja and P.J. Knight

EURATOM/CCFE Fusion Association, Culham Science Centre, Abingdon, Oxfordshire, OX14 3DB, UK a.thyagaraja@ccfe.ac.uk, peter.knight@ccfe.ac.uk

Summary. Turbulence is thought to play a key role in the transport of particles and energy within thermonuclear plasmas, and a number of codes have been developed to study the phenomena involved. A novel algorithm for calculating the force balance within such a code is presented. This involves the solution of a non-linear elliptic boundary value problem by considering it as a steady-state limit of a parabolic (heat) equation.

1 Introduction

One of the grand scientific challenges of our time is the understanding and control of electromagnetic turbulence in thermonuclear plasmas typically encountered in tokamak experiments. Spherical tokamaks (e.g. the Mega Ampère Spherical Tokamak, MAST [7] in the UK) in particular have uncovered fascinating new insights and regimes, with the potential of providing new and efficient approaches to building practical fusion power plants and associated materials testing/development [4]. The fluid approach to plasma turbulence simulations seeks to evolve a set of conservation equations (similar in form but differing in detail from the classic compressible Navier–Stokes equations of neutral gas dynamics) together with the Maxwell equations in three spatial dimensions and time. The fluid models are similar in character and akin in their computational philosophy and techniques to studies of long term (on time-scales of decades or even centuries) climatic dynamics and changes over the whole globe. Experimental measurements suggest that electromagnetic turbulence must play a vital role, and advances in computing are enabling rapid strides in modelling the basic underlying mechanisms. Success in this venture will facilitate improvements to the prospects for economic fusion power.

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*, Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_166,

2 The CENTORI Plasma Turbulence Code

The CENTORI code is a fully toroidal, two-fluid (ions + electrons) electromagnetic turbulence simulation code, and has been developed by CCFE in collaboration with colleagues at the University of Edinburgh (EPCC) [3]. It is designed to simulate tokamak plasma turbulence in realistic geometries and conditions such as those found in the present day machines MAST and the European flagship Joint European Torus experiment (JET) [2], and in the forthcoming international fusion experiment ITER [1]. CENTORI self-consistently co-evolves the global plasma equilibrium and the electromagnetic turbulence driven by sources of particles, heat, momentum and currents via the gradients generated thereby in plasma quantities like the pressure and temperature. Powerful parallel processing techniques allow CENTORI to use sufficiently high spatial and temporal resolutions to enable the modelling of scales varying from the system size to the experimentally relevant ion-gyro radius scales. Here, we describe the algorithm used to solve the force balance within the code.

3 Equilibrium Force Balance

3.1 The Grad Shafranov Equation

CENTORI models a toroidal plasma confined by magnetic fields due to external coils and by driving a toroidal current in the plasma. The equilibrium (force balance) equation relating the magnetic field $\mathbf{B}$ (the sum of external coil fields and the plasma generated one) and the plasma pressure gradient, in Gaussian units, is:

$$\frac{\mathbf{J} \times \mathbf{B}}{c} = \nabla p \tag{1}$$

where p is the sum of the electron and ion pressures, c is the speed of light and $\mathbf{J}$ is the current density within the plasma, given by Ampère's Law:

$$\frac{4\pi}{c}\mathbf{J} = \nabla \times \mathbf{B} \tag{2}$$

We use a right-handed cylindrical coordinate system (R, Z, ζ), where R is the *major radius*, or distance from the machine's vertical axis of symmetry, Z defines the vertical direction, parallel to the machine's axis, and ζ is the azimuthal/toroidal angle. This is shown in the left-hand plot of Fig. 1.

The equilibrium magnetic field is azimuthally symmetric and is described by two functions, $\psi(R, Z), F(\psi)$:

$$\mathbf{B} = (\nabla\zeta \times \nabla\psi) + F\,\nabla\zeta \tag{3}$$

The flux function $\psi(R, Z)$ describes the poloidal magnetic field. The level curves of ψ in the (R, Z) plane form nested, closed contours in the region

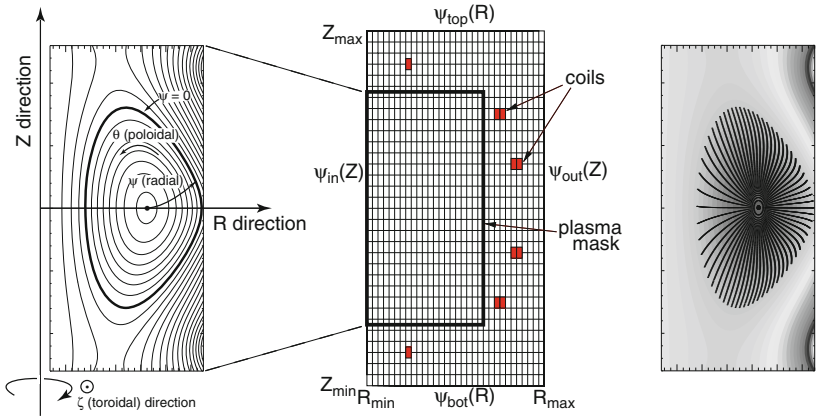


Fig. 1. *Left*: A typical plot of ψ contours over the poloidal (R, Z) plane as calculated by the `GRASS` equilibrium solver, showing diagrammatically the laboratory coordinates (R, Z, ζ) and the plasma coordinates (ψ, θ, ζ) used within `CENTORI`. *Middle*: Schematic diagram of the computational domain of `GRASS`, showing the main solution grid and the plasma mask. *Right*: A typical set of (ψ, θ) grid points, superimposed on the original $\psi(R, Z)$ grid

containing the plasma. These form toroidal flux surfaces. The minimum value of ψ within these closed surfaces lies at the centre of the plasma, and defines the location of the so-called *magnetic axis*, along the circle (R_0, Z_0, ζ).

The function F depends only upon ψ and defines the toroidal magnetic field in (3). Using the standard definition of the gradient operator, the gradients in the ψ and ζ directions are:

$$\nabla\psi = \frac{\partial\psi}{\partial R}\mathbf{e_R} + \frac{\partial\psi}{\partial Z}\mathbf{e_Z}, \quad \nabla\zeta = \frac{1}{R}\,\mathbf{e}_\zeta \tag{4}$$

where $(\mathbf{e_R}, \mathbf{e_Z}, \mathbf{e}_\zeta)$ are the unit vectors in the coordinate directions.

Combining (1)–(4) we obtain the simplest version of the Grad–Shafranov equation for equilibrium force balance:

$$\left[R\,\frac{\partial}{\partial R}\left(\frac{1}{R}\frac{\partial\psi}{\partial R}\right) + \frac{\partial^2\psi}{\partial Z^2}\right] = -4\pi R^2 p' - FF' = \frac{4\pi}{c}RJ_{tor} \tag{5}$$

where the left hand side operator is denoted by $\Delta^*\psi$, and $'$ is the usual notation for $\partial/\partial\psi$. We next outline a novel algorithm for solving this equation for $\psi(R, Z)$, given $p'(\psi)$ and $FF'(\psi)$. These are determined by solving appropriate transport equations (not described here).

3.2 The `GRASS` Equilibrium Solver

The `CENTORI` source code includes the free boundary Grad Shafranov equilibrium solver, named `GRASS` which solves (5) subject to Dirichlet conditions,

taking into account currents in the coils located in the vicinity of the plasma. Figure 1 (middle plot) shows the layout of the computational domain of GRASS. The solver uses two rectangular grids:

1. The main solution grid, within the domain (R_{min}, Z_{min}) to (R_{max}, Z_{max}). The plasma and the coils are assumed to lie wholly within this grid, and the ψ values on the grid boundaries are supplied by the user.
2. The plasma mask represents the rectangular sub-region $(R_{p\,min}, Z_{p\,min})$ to $(R_{p\,max}, Z_{p\,max})$. The (hot) plasma is assumed to lie wholly within the plasma mask, but no coils must be present inside it.

Each coil current density, J_c (which can, of course, vary between coils), is assigned to a number of grid cells, to approximate the coil location and cross-sectional area. The following equation is solved over the main solution grid:

$$\Delta^* \psi = \frac{4\pi}{c} R J_{tor}$$

where J_{tor} is a function of ψ, R in the plasma, and $J_{tor} = J_c$ at the coil locations. That is, we can rewrite the equation as:

$$\Delta^* \psi = \frac{4\pi}{c} \left(R J_t(\psi, R)\, H + R J_c \right)$$

$$\text{where} \qquad H = \begin{cases} 1 & \text{inside plasma mask} \\ 0 & \text{elsewhere} \end{cases}$$

and $J_t(\psi, R) = -c\, R^2 p' - \frac{c}{4\pi} F F'$ is the toroidal component of $\mathbf{J}$ within the plasma. The functional forms of $p'(\psi)$ and $FF'(\psi)$ are evolved externally by CENTORI to include the effects of turbulence, with suitable averaging over the flux surfaces. The physics ensures that these functions fall off fast enough with ψ so that there is only a negligible amount of residual plasma current outside of the chosen edge plasma contour.

We use a tridiagonal matrix algorithm and "imbed" the above elliptic equation in a parabolic equation, as described below. A preliminary transformation makes the operator symmetric and increases the diagonal dominance of the matrix equation, by setting, $\psi = R^{\frac{1}{2}} u \iff u = \frac{\psi}{R^{\frac{1}{2}}}$ and applying the boundary conditions in terms of u instead of ψ. Now, let the total ψ be written as the sum of two components: ψ_1, which is defined to have zero boundary conditions at $Z = Z_{min}$ and $Z = Z_{max}$; then ψ_2, which "absorbs" the rest of ψ. Thus, $\psi = \psi_1 + \psi_2$, where, $\psi_2 \equiv \frac{Z - Z_{min}}{h} \psi_{top} + \frac{Z_{max} - Z}{h} \psi_{bot}$ and $h = Z_{max} - Z_{min}$ which, by inspection, has the desired behaviour. The transformation of u follows. Note that, $\frac{\Delta^* \psi_1}{R^{\frac{1}{2}}} = \frac{\Delta^* \psi}{R^{\frac{1}{2}}} - \frac{\Delta^* \psi_2}{R^{\frac{1}{2}}}$ and we can define a new operator $\Delta_u^* u$ (note the subscript u):$\Delta_u^* u \equiv \frac{1}{R^{\frac{1}{2}}} \Delta^* \psi$. It follows that, $\Delta_u^* u = \frac{\partial^2 u}{\partial R^2} + \frac{\partial^2 u}{\partial Z^2} - \frac{3}{4R^2} u$, leading to,

$$\frac{\partial^2 u_1}{\partial R^2} + \frac{\partial^2 u_1}{\partial Z^2} - \frac{3}{4R^2} u_1 = \frac{4\pi}{c} \left(\frac{R J_t(\psi)\, H}{R^{\frac{1}{2}}} + R^{\frac{1}{2}} J_c \right) - \Delta_u^* u_2 \qquad (6)$$

Note that $\Delta_u^* u_2$ can be evaluated using finite differences straightforwardly, once at the beginning of the calculation. From the definition of u_2 there is no $\partial^2 u_2/\partial Z^2$ term.

The novel approach in `GRASS` is to use a heat-like equation to solve this equation. Consider:

$$\frac{\partial u_1}{\partial \tau} = \epsilon \left(\Delta_u^* u_1 - G\right) \tag{7}$$

where G represents the right hand side of (6), and τ is a pseudo-time variable. At steady-state, $\partial u_1/\partial \tau = 0$, so the quantities inside the parentheses become equal, as desired. We convert (7) into a finite difference equation in Fourier space using the sine transform in the Z direction to maintain the zero boundary conditions, denoting the k_Z^{th} sine transform coefficients by $\widehat{\ldots}$, with i labelling the i^{th} grid point in the R direction and N the iteration count. This leads to a 1-D tridiagonal matrix equation in the R direction, of the form:

$$A_i\, \hat{u}_{1\,i-1}^{N+1} + B_i\, \hat{u}_{1\,i}^{N+1} + C_i\, \hat{u}_{1\,i+1}^{N+1} = \left(\hat{u}_{1\,i}^{N} - \Delta\tau.\epsilon\, \widehat{G_i}\right)$$

where

$$A_i = C_i = -\frac{\Delta\tau.\epsilon}{(\Delta R)^2}$$

$$B_i = 1 + \Delta\tau.\epsilon \left\{ \frac{2}{(\Delta R)^2} + \frac{\pi^2 k_Z^2}{h^2} + \frac{3}{4R_i^2} \right\}$$

This is readily solved to obtain $\hat{u}_1^{N+1}$; the inverse sine transform is used to obtain u_1 and thereby ψ_1. The total ψ is recovered by adding back ψ_2, and the process is repeated until ψ over the grid does not change significantly between successive iterations. Typically, convergence is achieved after a few hundred iterations, depending on the initial FF' and p' profiles specified.

The final stage of `GRASS` is to redefine the ψ within the plasma mask for convenience, so that the edge of the plasma (determined using an algorithm that takes into account the presence or absence of saddle ("X-") points in the $\psi(R, Z)$ contours) is defined to be the $\psi = 0$ contour; modifying ψ by an additive constant everywhere does not affect its gradients, i.e. the magnetic field. `CENTORI` is passed only this modified $\psi(R, Z)$ within the masked region (thus excluding the coils), interpolated using Chebyshev fits in R and Z.

The plasma quantities that `CENTORI` evolves to model the turbulence are stored in arrays at a set of computational grid points chosen for their physical and numerical convenience. The method used to form the plasma coordinate system based on these grid points will be described in detail elsewhere. It involves the use of Chebyshev polynomial approximations to accurately represent the ψ contours. The coordinate system and a suitable choice of Jacobian is made to produce a mesh system which is optimal for the turbulence evolutionary dynamics (cf. Fig. 1, right-hand plot).

4 Conclusions and Further Work

In this brief description we have only touched upon the method of solving the non-linear elliptic equations of plasma equilibria in the **absence** of significant flows in the system. Modern tokamaks often have large sheared toroidal and poloidal flows which can modify the equilibrium. In earlier analytical work [5,6] we have investigated such effects in detail and found extensions of the Grad-Shafranov equations involving rigid or Keplerian toroidal rotation. The methods we have described in this paper are now being adapted to include these important effects. The ultimate aim of our research is to co-evolve the plasma equilibrium given the experimental sources of current, particles, energy and momentum and compute on the time-scales of interest. This can only be accomplished with the use of a robust and accurate scheme to obtain the equilibrium and construct the curvilinear plasma coordinates needed to formulate the turbulence evolution equations. The scheme described in this paper enables us to carry out this process in a reliable and effective manner.

Acknowledgements

This work was funded by the United Kingdom Engineering and Physical Sciences Research Council under grant EP/G003955 and the European Communities under the contract of Association between EURATOM and CCFE. The views and opinions expressed herein do not necessarily reflect those of the European Commission.

References

1. Aymar, R.: Fusion Eng. Des. **55**, 107–118 (2001)
2. Keilhacker, M., Gibson, A., Gormezano, C., Rebut, P.H.: Nucl. Fusion **41**, 1925–1966 (2001)
3. Knight, P.J., Roach, C.M., Thyagaraja, A., Applegate, D.J., Joiner, N.: Capability Comput. **6**, 4–6 (2005)
4. Morris, A.W., Akers, R.J., Counsell, G.F., Hender, T.C., Lloyd, B., Sykes, A., Voss, G.M., Wilson, H.R.: Fusion Eng. Des. **74**, 67–75 (2005)
5. Thyagaraja, A., McClements, K.G.: Mon. Not. R. Astron. Soc. **323**, 733–742 (2002)
6. Thyagaraja, A., McClements, K.G.: Phys. Plasmas **13**, 062502 (2006)
7. Valovič, M., et al.: Nucl. Fusion **48**, 075006 (2008)

A Differential-Geometric Approach to Model Isotropic Diffusion on Circular Conic Surfaces in Uniform Rotation

M.M. Tung and A. Hervás

Instituto de Matemática Multidisciplinar, Universidad Politécnica de Valencia, Camino de Vera s/n, 46022 Valencia, Spain mtung@mat.upv.es, ahervas@mat.upv.es

Summary. We outline a differential-geometric approach to analytically solve the diffusion equation on a static circular conic surface assuming isotropic and sourceless diffusion. We also extend the proposed technique to find general solutions for a cone in arbitrary axisymmetric and uniform rotation.

The new, analytical expressions for these solutions rely on the construction of the kernel function for the diffusion operator on the corresponding Riemannian manifold. Given particular boundary conditions, the resulting series expansions may for practical purposes be approximated numerically, providing a valuable tool for diffusion models.

1 Introduction and Review

Diffusion phenomena are an essential ingredient in mathematical models and simulation techniques for particle and fluid dynamics, or heat transport processes. Many industrial devices take particularly simple shapes, such as the form of circular cones.

In a classic work, Carslaw and Jaeger have studied the heat flow *in* a sphere and cone [2]. Further examples of analytical solutions for diffusion *on* a sphere are found in Zwillinger's standard reference [8].

In this work, we follow the differential-geometric approach outlined in [7] to analytically solve the diffusion equation on a static circular conic surface assuming isotropic and sourceless diffusion. We also extend the proposed technique to find general solutions for a cone in arbitrary axisymmetric and uniform rotation.

The diffusion-advection equation is a non-homogeneous parabolic partial differential equation which can be derived from a Lagrangian function via a variational principle [7]. This inherently covariant approach makes it possible to tackle diffusion processes on smooth manifolds, such as the surface of a circular cone.

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*, Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_167,

Contrary to classical mechanics, for the diffusion case we require the partial derivatives of a configuration with respect to all *spacetime* coordinates. Therefore, the corresponding Lagrangian function will be a mapping

$$\mathcal{L} : J^1 N \to \mathbb{R},$$

where $\mathcal{L} : J^1 N$ is the jet bundle with coordinates $(C, C^*, \dot{C}, \dot{C}^*, C_{;i}, C^*{}_{;i}) \simeq \mathbb{R}^{10}$ and N is the configuration space with parameters $(x^1, x^2, x^3, t, C, C^*)$. As usual, x^i for $i = 1, 2, 3$ denote the spatial components in a local coordinate frame of point $p \in M$, where $(M, \mathbf{g})$ is a smooth 3-dimensional Riemannian manifold with a given metric $\mathbf{g}$. Time is represented by $t \in \mathbb{R}_+$ and its corresponding derivatives by dotted symbols. All covariant derivatives with respect to x^i are denoted by the common semicolon notation.

Following the terminology of Marsden et al. [5], we further introduce base space $B = M \times \mathbb{R}_+$, which constitutes standard spacetime, and ambient space P, given by the two concentrations $C, C^* : B \to \mathbb{R}_+$.[1] Thus, $N = B \times P$.

Within this framework, any particular configuration of the system is described by a mapping $B \to N$. Provided with the explicit form of the diffusion Lagrangian [7], we will be in the position to study isotropic diffusion on the cone.

2 Lagrangian Formalism

For diffusion, the equation of motion, *i.e.* the diffusion equation itself, is obtained from the following action integral over a bounded and closed set $V \subset M$

$$L = \int_V \mathcal{L} \sqrt{g}\, d\tau, \tag{1}$$

where $\sqrt{g}\, d\tau = \sqrt{g}\, dx^1 dx^2 dx^3$ is the invariant volume element with $g = \det \mathbf{g}$, and [7]

$$\mathcal{L} = -D^{ij} C_{;i}\, C^*{}_{;j} - \tfrac{1}{2}\big(\dot{C} C^* - C \dot{C}^*\big) + S\big(C + C^*\big). \tag{2}$$

Here, the molecular diffusion tensor $\mathbf{D}$ is of type $\mathbf{D}_p : (T_p^* M)^2 \to \mathbb{R}$ for $p \in M$, and a general source/reaction term is a scalar defined by $S : M \to \mathbb{R}$.

Then, the stationary solution of the action (1) under any variation of the generalized coordinate C yields [7]

$$\frac{\delta L}{\delta C} = 0 \quad \Rightarrow \quad \dot{C} = \Big(D^{ij} C_{;i}\Big)_{;j} + S, \tag{3}$$

which is the all-inclusive (not necessarily isotropic and with arbitrary sources) diffusion equation in covariant form.

[1] Note that C^* is a mirror symmetry which has to be introduced to satisfy energy conservation [6].

3 Diffusion on a Static Cone

For isotropic diffusion the molecular diffusion tensor reduces to

$$\mathbf{D}_p = D\,\mathbf{g}_p \quad \forall p \in M \quad \text{with} \quad D \in \mathbb{R}_+. \tag{4}$$

Hence, the fundamental differential equation which governs sourceless isotropic diffusion ($S = 0, D = 1$) becomes

$$\dot{C} = \Delta_M C, \tag{5}$$

where Δ_M is the Laplace–Beltrami operator on the given manifold $(M, \mathbf{g})$.

For an open cone with its peak located at the origin and a constant radius-to-height parameter $a > 0$, the metric is

$$(g_{ij}) = \begin{pmatrix} 1+a^2 & 0 \\ 0 & a^2 z^2 \end{pmatrix} \quad \text{with} \quad \sqrt{g} = a\sqrt{1+a^2}\,z, \tag{6}$$

where $z \in]0, \infty[$ and $\varphi \in [0, 2\pi[$ are the local coordinates on the surface referring to $i, j = 1, 2$, respectively. Note that there are only two independent Christoffel symbols, $\Gamma^1_{22} = -a^2 z/(1+a^2)$ and $\Gamma^2_{12} = 1/z$, which produce after a short calculation

$$\begin{aligned} \Delta_M C &= g^{ij} C_{;ij} = g^{ij}\left(C_{,ij} - \Gamma^k_{ij} C_{,k}\right) \\ &= \frac{1}{1+a^2}\frac{1}{z}\frac{\partial}{\partial z}\left(z\frac{\partial C}{\partial z}\right) + \frac{1}{a^2 z^2}\frac{\partial^2 C}{\partial \varphi^2}. \end{aligned} \tag{7}$$

3.1 Formal Self-Adjointness

To investigate further properties of Δ_M, we introduce the following inner product

$$\langle f_1, f_2\rangle_{(M,g)} = \int_M f_1(x) f_2(x)\sqrt{g}\,d\tau(x), \quad \forall f_1, f_2 \in \{f : M \to \mathbb{R}; f \in \mathcal{C}^\infty\}. \tag{8}$$

Then, by applying integration by parts, one can show that

$$\langle \Delta_M f_1, f_2\rangle_{(M,g)} = \langle f_1, \Delta_M f_2\rangle_{(M,g)}, \tag{9}$$

i.e. the operator Δ_M is *formally self-adjoint*. As a consequence the time-independent eigenvalue problem

$$(\Delta_M + \lambda)\Phi(\lambda, x) = 0, \quad \forall x \in M, \tag{10}$$

has a continuous non-negative spectrum with a complete and orthogonal set of eigenfunctions $\Phi(\lambda, x)$. Continuity follows from the unbound integration domain of the differential equation [3].

Thus, we may expand the solutions of the diffusion equation in terms of normalized eigenfunctions $\hat{\Phi}(\lambda, x)$:

$$C(x,t) = \int_0^\infty a(\lambda)\, e^{-\lambda t}\hat{\Phi}(\lambda, x)\,d\lambda. \tag{11}$$

3.2 Formal Solutions for the Circular Conic Surface

Normalization for the eigenfunction system $\{\hat{\Phi}(\lambda,\cdot)\}$ solving (10) is chosen such that

$$\langle\hat{\Phi}(\lambda,x),\hat{\Phi}(\mu,x)\rangle_{(M,g)} = \delta(\lambda-\mu), \tag{12}$$

and given a continuous boundary function $C(x,0) : M \times [0,\infty[\to \mathbb{R}$, the expansion coefficients of the solution are

$$a(\lambda) = \langle C(x,0),\hat{\Phi}(\lambda,x)\rangle_{(M,g)}. \tag{13}$$

Inserting the coefficients $a(\lambda)$ into the solution expansion readily yields the full solution as an evolution equation:

$$C(x,t) = \langle K(x,y,t), C(y,0)\rangle_{(M,g)}. \tag{14}$$

Here, $K(x,y,t)$ is the kernel function of the diffusion operator $(\partial_t + \Delta_M)$ on $M \times M \times]\,0,\infty\,[$ and is given by the expression

$$K(x,y,t) = \int_0^\infty e^{-\lambda t}\hat{\Phi}(\lambda,x)\,\hat{\Phi}(\lambda,y)\,d\lambda. \tag{15}$$

3.3 Explicit Solutions for the Circular Conic Surface

Explicit solutions for the circular conic surface M may be obtained by employing the separation-of-variable method (see e.g. [4]) to determine the diffusion kernel (15). A lengthy but straightforward calculation produces the following two solutions for the eigenvalue problem

$$\begin{aligned}\Phi_m^{(1)}(\lambda,z,\varphi) &\sim \cos(m\varphi)\,J_{\frac{\sqrt{1+a^2}}{a}m}\left(\sqrt{1+a^2}\,z\sqrt{\lambda}\right),\\ \Phi_m^{(2)}(\lambda,z,\varphi) &\sim \sin(m\varphi)\,J_{\frac{\sqrt{1+a^2}}{a}m}\left(\sqrt{1+a^2}\,z\sqrt{\lambda}\right),\end{aligned} \tag{16}$$

where $m \in \mathbb{N}_0$. The functions $J_\beta(\alpha z)$ with $\alpha = \sqrt{1+a^2}\sqrt{\lambda}$ and $\beta = m\sqrt{1+a^2}/a > 0$ are the Bessel functions of the first kind. The solutions $J_{-\beta}(\alpha z)$ have been discarded, since they behave like $z^{-\beta}$ at $z = 0$ and are unbounded.

The *normalized* eigenfunctions will have to satisfy

$$\langle\hat{\Phi}_m^{(i)}(\lambda,x),\hat{\Phi}_m^{(j)}(\mu,x)\rangle_{(M,g)} = \delta_{ij}\,\delta(\lambda-\mu), \quad m \neq 0. \tag{17}$$

and the Bessel closure relation [1]. Furthermore, the standard transformation formulae for the δ-distributions help to determine the normalization constants.

In summary, given the boundary problem

$$\begin{aligned}\dot{C} &= \Delta_M C,\\ B(x) &= \lim_{t\to 0_+} C(x,t)\ \forall x \in M \text{ and } B \in \mathcal{C}(M),\end{aligned} \tag{18}$$

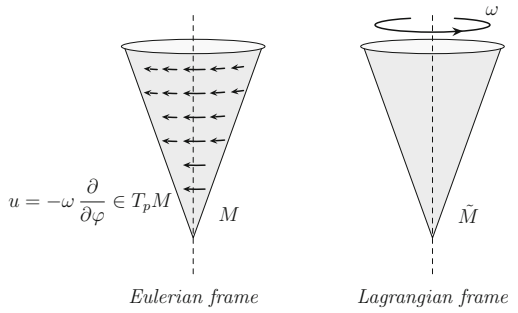


Fig. 1. Schematic view of diffusion on a cone in different frames

we obtain the corresponding solution for an infinitely extended cone using the kernel $K \in \mathcal{C}^{2,2,1}(M \times M \times]\,0, \infty\,[\,)$:

$$C(x,t) = \langle K(x,y,t), B(y) \rangle_{(M,g)}. \tag{19}$$

Here, the explicit form for the fundamental solution (15) is

$$K(x,y,t) = \int_0^\infty e^{-\lambda t} \sum_{i,j=1,2} \sum_{m=0}^{\infty} \hat{\Phi}_m^{(i)}(\lambda, x)\, \hat{\Phi}_m^{(j)}(\lambda, y)\; d\lambda, \tag{20}$$

where the normalized eigenfunctions are for $a > 0$:

$$\begin{cases} \hat{\Phi}_0^{(1)}(\lambda, x) = \dfrac{(1+a^2)^{1/4}}{2\sqrt{\pi a}} J_0\left(\sqrt{1+a^2}\, z\sqrt{\lambda}\right) \\ \hat{\Phi}_0^{(2)}(\lambda, x) = 0 \\ \hat{\Phi}_m^{(1)}(\lambda, x) = \dfrac{(1+a^2)^{1/4}}{\sqrt{2\pi a}} \cos(m\varphi)\, J_{\frac{\sqrt{1+a^2}}{a} m}\left(\sqrt{1+a^2}\, z\sqrt{\lambda}\right), \quad m \in \mathbb{N} \\ \hat{\Phi}_m^{(2)}(\lambda, x) = \dfrac{(1+a^2)^{1/4}}{\sqrt{2\pi a}} \sin(m\varphi)\, J_{\frac{\sqrt{1+a^2}}{a} m}\left(\sqrt{1+a^2}\, z\sqrt{\lambda}\right), \quad m \in \mathbb{N} \end{cases} \tag{21}$$

For approximate boundary conditions, this general solution can be evaluated analytically or approximated by suitable numerical schemes.

4 Diffusion on a Uniformly Rotating Cone

We now consider diffusion with axisymmetric and uniform rotation of the cone (see Fig. 1). In the *Eulerian frame* the isotropic diffusion equation will contain a transport term $u \neq 0$ such that

$$\dot{C} = \Delta_M C - u^i C_{;i} \tag{22}$$

with $u = -\omega\,\partial/\partial\varphi \in T_pM$ in local coordinates for all $p \in M$, and $\omega > 0$. We cannot proceed with a solution for (22) as outlined in Sect. 3, since the operator

$$\left(\Delta_M - \mathbf{u}\cdot\nabla_M\right) \tag{23}$$

is not self-adjoint.[2] However, moving to the *Lagrangian frame* with $\tilde{M}$ (see Fig. 1), the global transformation becomes

$$\varphi = \tilde{\varphi} + \omega t, \tag{24}$$

which effectively removes the influence of the transport field, and one only has to solve

$$\dot{\tilde{C}} = \Delta_{\tilde{M}}\tilde{C}, \tag{25}$$

where $\tilde{C} = C(\tilde{x}, t)$ and $\tilde{B}(\tilde{x}) = B(x)$, with $x \in M$ and $\tilde{x} \in \tilde{M}$. Hence, the rotating solution is given in terms of the already known, static result of Sect. 3:

$$C(z,\varphi,t) = \langle K(x,y,t), B(y)\rangle_{(\tilde{M},g)} = \langle K(z,\varphi+\omega t;\tilde{z},\tilde{\varphi};t), B(\tilde{z},\tilde{\varphi})\rangle. \tag{26}$$

5 Conclusion and Outlook

We have presented a differential-geometric approach to deal with diffusion processes on curved surfaces using a general Lagrangian on smooth manifolds. The fundamental solutions for diffusion on a static and a rotating cone are derived by constructing the kernel function for the corresponding Laplace–Beltrami operator and adopting an appropriate reference frame. These solutions may provide the foundation for future diffusion models with related geometry.

Acknowledgment

This work has been partially supported by projects PAID-06-07/3283 of the Universidad Politécnica de Valencia and MTM2009-08587 by the Spanish ministry.

References

1. Arfken, G.B.: Mathematical Methods for Physicists. Academic Press, San Diego (1985)
2. Carslaw, H.S., Jaeger, J.C.: Conduction of Heat in Solids. Oxford University Press, Oxford (1986)
3. Courant, R., Hilbert, D.: Methoden der mathematischen Physik. Springer, Berlin (1993)

[2] This becomes clear because in this case $\langle \mathbf{u}\cdot\nabla_M f_1, f_2\rangle_{(M,g)} \neq \langle f_1, \mathbf{u}\cdot\nabla_M f_2\rangle_{(M,g)}$.

4. Lanczos, C.: Linear Differential Operators. Society for Industrial Mathematics, Philadelphia (1987)
5. Lew, A., Marsden, J.E., Ortiz, M., West, M.: Arch. Ration. Mech. Anal. **167**, 85–146 (2003)
6. Morse, P.M., Feshbach, H.: Methods of Theoretical Physics McGraw-Hill Publishing Company, New york (1953)
7. Tung, M.M.: Basics of a differential-geometric approach to diffusion: Uniting Lagrangian and Eulerian models on a manifold. In: Bonilla, L.L., Moscoso, M.A., Platero, G., Vega, J.M. (eds.) Progress in Industrial Mathematics at ECMI 2006, vol. 12 of Mathematics in Industry, pp. 897–901. Springer, (2007)
8. Zwillinger, D.: Handbook of Differential Equations. Academic, Boston (1998)

A General Model of Lung Tumour Motion

P.L. Wilson[1] and J. Meyer[2]

[1] Department of Mathematics & Statistics, University of Canterbury, Private Bag 4800, Christchurch 8140, New Zealand p.wilson@math.canterbury.ac.nz
[2] Department of Physics & Astronomy, University of Canterbury, Private Bag 4800, Christchurch 8140, New Zealand juergen.meyer@canterbury.ac.nz

Summary. A limiting factor for the effective delivery of radiotherapy to lung tumours is the tumour motion as the patient breathes. If the tumour position is known at all times then treatment parameters may be adjusted accordingly. We formulate a general approach to model the spatial relationship between an external respiratory signal and the tumour position. The model treats the tumour as a point mass attached to a spring-dashpot system driven by abdominal motion. We present the model and show results of numerical computations based on clinical data.

1 Introduction

Cancer of the respiratory tract is a common disease with relatively poor prognosis. Respiratory-induced tumour motion is a major cause of unfavourable treatment responses to radiation therapy, since the motion necessitates relatively large treatment margins [4]. Consequently, a larger volume of healthy tissue is irradiated, thereby considerably increasing side effects from the treatment. This forces a treatment dose lower than that required for adequate tumour control.

An accurate knowledge of the tumour position at all times during irradiation would enable treatment parameters to be adjusted and therefore a more effective dose to be delivered [2, 5]. In this fashion, higher overall doses could be given to the tumour without significantly increasing the side effects. However, in practice it is difficult to directly and non-invasively determine the tumour position in real-time during treatment [1, 6, 8, 9, 11, 12, 14].

This technical difficulty has led most non-invasive approaches to resort to measuring only the respiratory signal of the patient. Accurate and reliable models are required to relate respiratory motion to that of the lung tumour, in order that the position of the tumour can be known at all times. Previous "grey box" type models of this relationship [7] have been successfully applied to clinical cases, but intrinsically do not have the ability to model certain behaviours, such as when the tumour trajectory exhibits a hysteresis [11].

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*,
Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_168,

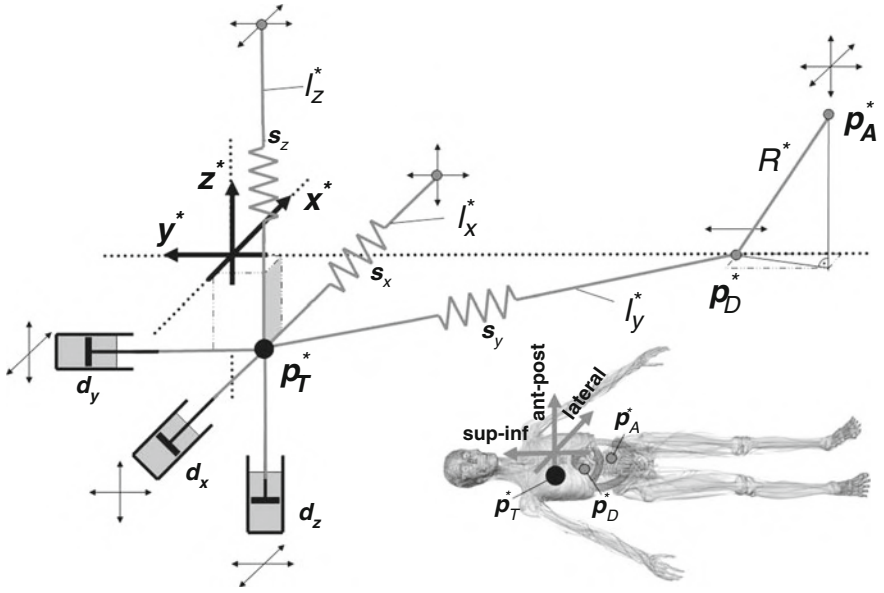


Fig. 1. The spring-dashpot system. The coordinate axes, spring and dashpot labels, and the position of the tumour (subscript T), diaphragm (D), and abdomen (A) are shown. Degrees of freedom are indicated by *arrows*

Here we present a novel 3D model aimed at physically modelling the spatial relationship between an external abdominal breathing signal and the motion of a lung tumour. The approach uses a system of springs and dashpots arranged to model tumours at a general lung location. While the mathematical model is general for any lung tumour, the model parameters are patient- and tumour location-specific. The model is analysed by means of computer simulations and validated against real clinical tracking data.

2 The Model

With reference to Fig. 1, the tumour is modelled as being attached to three springs and three dashpots, one spring-dashpot pair for each room coordinate, reflecting the mechanical and dynamical properties of human anatomy. At a general time t^*, where a superscript star indicates a dimensional quantity, the tumour lies at position $\mathbf{p}_T^*(t) = (x_T^*(t^*), y_T^*(t^*), z_T^*(t^*))$ where (x^*, y^*, z^*) forms a right-handed triad as shown, with corresponding unit vectors $\mathbf{i}$, $\mathbf{j}$, $\mathbf{k}$. The superior-inferior (head-foot, or sup-inf) spring is additionally attached to a point representing the diaphragm, which in turn is attached by a rigid rod to a point representing the abdomen. This abdomen point is free to move in a fully three-dimensional way. The springs are assumed to follow Hooke's law, while the dashpots provide friction proportional to velocity.

In more detail, spring s_X is of natural length l_X^* and has a pivot point free to move in the plane $x^* = l_X^*$. Dashpot d_X damps the x^*-motion, and is also free to move in a (y^*, z^*)-plane to track the (y^*, z^*)-components of the tumour position. A similar arrangement holds for spring s_Z. On the other hand, spring s_Y of natural length l_Y^* connects the tumour to the diaphragm point $\mathbf{p}_D^*(t^*)$, which we assume is constrained to the line $x^* = 0, z^* = 0$. This point drives the tumour motion by means of being connected by a rigid rod of length R^* to the point $\mathbf{p}_A^*(t^*)$, which models the abdominal position in three dimensions. The dashpot d_Y damps the y^*-motion, and is free to move in an (x^*, z^*)-plane to track the (x^*, z^*)-components of the tumour motion.

Next, the three coupled equations of motion are formulated and non-dimensionalised, yielding the 3D-3D model: three coupled equations describing the 3D motion of the system forced by a 3D abdominal breathing signal. The input to this model is a three-dimensional breathing signal; the parameters of the model need to be optimised for each patient; and the output is a predicted lung tumour motion.

3 Governing Equations

The angle which s_Y makes with the plane $x = 0$ is denoted θ_X, and that which s_Y makes with the plane $z = 0$ is denoted θ_Z. Then for a tumour of constant mass m^*, Newton's second law gives the governing dimensional equations

$$m^* \ddot{x}_T^* = k_X^* x_T^* + k_Y^* \xi_Y^* \sin\theta_X \cos\theta_Z - \beta_X^* \dot{x}_T^*, \tag{1}$$

$$m^* \ddot{y}_T^* = -k_Y^* \xi_Y^* \cos\theta_X \cos\theta_Z - \beta_Y^* \dot{y}_T^*, \tag{2}$$

$$m^* \ddot{z}_T^* = k_Z^* z_T^* + k_Y^* \xi_Y^* \sin\theta_Z - \beta_Z^* \dot{z}_T^*, \tag{3}$$

where

$$\xi_Y^* = \left\{ x_T^{*2} + \left[y_A^* - y_T^* + \left(R^{*2} - x_A^{*2} - z_A^{*2} \right)^{\frac{1}{2}} \right]^2 + z_T^{*2} \right\}^{\frac{1}{2}} - l_Y^*, \tag{4}$$

and where $\beta_X^*, \beta_Y^*, \beta_Z^*$ are the friction coefficients in the x^*-,y^*-, and z^*-directions, respectively, and a dot denotes differentiation with respect to time.

We now non-dimensionalise these equations (non-dimensional variables will be written without a superscript star). A representative length is taken to be l_Y^*, while τ^* denotes a general representative time scale. The non-dimensional equations are

$$\ddot{x}_T = \omega_X^2 x_T + \omega_Y^2 \xi_Y \sin\theta_X \cos\theta_Z - 2\lambda_X \dot{x}_T, \tag{5}$$

$$\ddot{y}_T = -\omega_Y^2 \xi_Y \cos\theta_X \cos\theta_Z - 2\lambda_Y \dot{y}_T, \tag{6}$$

$$\ddot{z}_T = \omega_Z^2 z_T + \omega_Y^2 \xi_Y \sin\theta_Z - 2\lambda_Z \dot{z}_T, \tag{7}$$

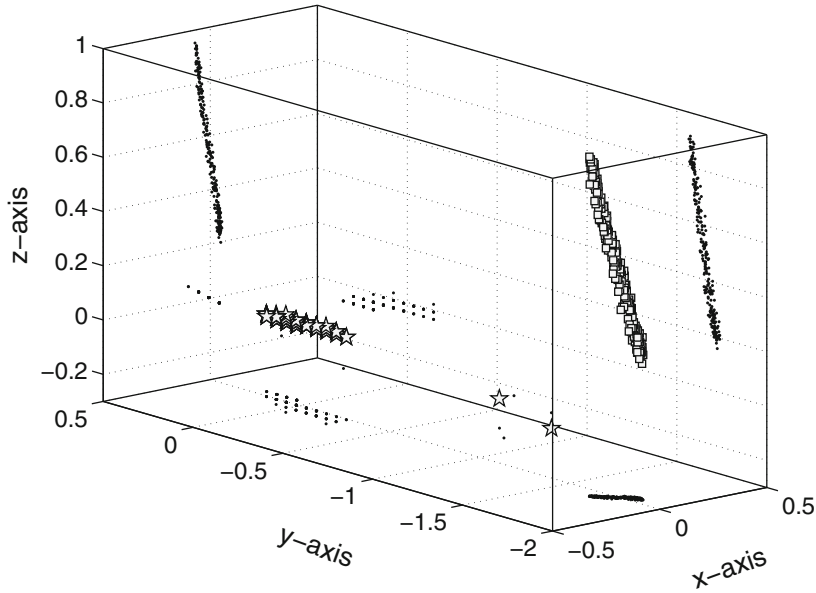


Fig. 2. Spatial distribution of clinical respiratory data (*squares*) and corresponding tumour data (*stars*). The data are also projected onto the (x, y)-, (x, z)-, and (y, z)-planes (*dots*)

where we have formed the six dimensionless groups

$$\omega^2_{X,Y,Z} = \frac{\tau^{*2} k^*_{X,Y,Z}}{m^*},\ 2\lambda_{X,Y,Z} = \frac{\tau^* \beta^*_{X,Y,Z}}{m^*}, \tag{8}$$

and where

$$\xi_Y = \left\{ x_T^2 + \left[y_A - y_T + \left(R^2 - x_A^2 - z_A^2 \right)^{\frac{1}{2}} \right]^2 + z_T^2 \right\}^{\frac{1}{2}} - 1. \tag{9}$$

4 Numerical Solutions

4.1 Numerical Method and Clinical Data

Typically, the main component of tumour motion is in the sup-inf direction, along the y-axis. In the limit of small lateral and transverse motion, asymptotic analysis reveals that the equations decouple, to leading order [13]. This leading-order equation for y_T was solved numerically using MATLAB, subject to the 3D breathing signal shown as squares in Fig. 2. We called this the 1D-3D system, since the breathing signal is still fully three-dimensional. Governing parameters were optimised iteratively based on a cost function to fit results to the clinical data. The numerical scheme was validated by comparison to limit-case analytical results [13].

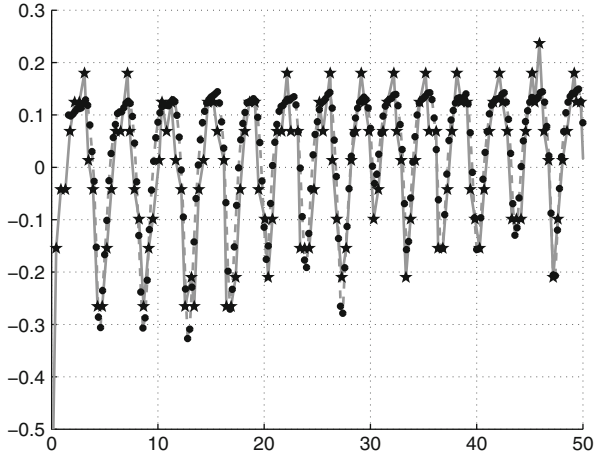


Fig. 3. Calculated and measured tumour signal y_T as a function of time. The *stars* and *solid line* correspond to the measured data and the *dots* and *dashed line* correspond to the calculated data

4.2 Results

The stars and solid line in Fig. 3 show the normalised and zeroed y-component of the tumour data in Fig. 2 versus time. Also plotted in Fig. 3 is the model output (dots and dashed line) for this data set, likewise normalised and zeroed. Excellent agreement can be seen at points between local extrema. These parts of the tumour motion are of greatest clinical relevance for real-time adjustment of treatment parameters. Small differences apparent close to some local extrema are likely due to experimental noise and sampling frequency (see the discussion in [13]), and even if genuine are likely to have minimal impact [3, 10].

5 Conclusions and Further Work

These results in the limiting case of small lateral and transverse motion have shown that it is indeed possible to use this spring-dashpot approach to model the spatial relationship between abdominal and lung tumour motion on a patient-specific basis. The 1D-3D model parameters can automatically be optimised using a simple cost function.

Our belief is that such models will be superior to more common grey box approaches, such as least-square models. Further work will include full 3D simulations of different breathing patterns using clinical data, and will also address the practical aspects of implementing such an approach in a clinical environment.

References

1. Berbeco, R.I., Mostafavi, H., Sharp, G.C., Jiang, S.B.: Phys. Med. Biol. **50**, 4481–4490 (2005)
2. D'Souza, W.D., Naqvi, S.A., Yu, C.X.: Phys. Med. Biol. **50**, 4021–4033 (2005)
3. Engelsman, M., Sharp, G.C., Bortfeld, T., Onimaru, R., Shirato, H.: Phys. Med. Biol. **50**, 477–490 (2005)
4. ICRU Report 50, Bethesda, MD, USA (1993)
5. Keall, P.J., Joshi, S., Vedam, S.S., Siebers, J.V., Kini, V.R., Mohan, R.: Med. Phys. **32**, 942–951 (2005)
6. Meyer, J., Richter, A., Baier, K., Wilbert, J., Guckenberger, M., Flentje, M.: Med. Phys. **33**, 1275–1280 (2006)
7. Meyer, J., Baier, K., Wilbert, J., Guckenberger, M., Richter, A., Flentje, M.: Acta Oncol. **45**, 923–934 (2006)
8. Murphy, M.J.: Semin. Radiat. Oncol. **14**, 91–100 (2004)
9. Parikh, P., Hubenschmidt, J., Vertatschitsch, E., Dimmer, S., Wright, J., Low, D.: Med. Phys. **32**, 2112–2113 (2005)
10. Rietzel, E., Liu, A.K., Doppke, K.P., Wolfgang, J.A., Chen, A.B., Chen, G.T., Choi, N.C.: Int. J. Radiat. Oncol. Biol. Phys. **66**, 287–95 (2006)
11. Seppenwoolde, Y., Shirato, H., Kitamura, K., Shimizu, S., van Herk, M., Lebesque, J.V., Miyasaka, K.: Int. J. Radiat. Oncol. Biol. Phys. **53**, 822–834 (2002)
12. Shimizu, S., Shirato, H., Ogura, S., Akita-Dosaka, H., Kitamura, K., Nishioka, T., Kagei, K., Nishimura, M., Miyasaka, K.: Int. J. Radiat. Oncol. Biol. Phys. **51**, 304–310 (2001)
13. Wilson, P.L., Meyer, J.: Comput. Math. Methods Med. **11**(1), 13–26 (2010)
14. Zhang, J., Wu, Y., Stepaniak, C., Liu, W., Gore, E., Li, X.: Med. Phys. **33**, 2046–2046 (2006)

The Lipid Bilayer at the Mesoscale: A Physical Continuum Model

P.L. Wilson[1], S. Takagi[2,3], and H. Huang[4]

[1] Department of Mathematics & Statistics, University of Canterbury, Private Bag 4800, Christchurch 8140, New Zealand p.wilson@math.canterbury.ac.nz
[2] Department of Mechanical Engineering, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan takagi@mech.t.u-tokyo.ac.jp
[3] RIKEN, 2-1 Hirosawa, Wako, Saitama 351-0198, Japan
[4] Mathematics & Statistics Department, York University, 4700 Keele Street, Toronto, Ontario, Canada M3J 1P3 hhuang@yorku.ca

Summary. A new model of inter-molecular interactions is introduced into a continuum paradigm for the lipid bilayer membrane. The model promotes the hydrogen bond network responsible for the hydrophobic effect. Physically-realistic numerical bilayers are obtained from the model.

1 Lipid Bilayers

Every cell in the human body is defined, internally divided, and has its contents maintained by membranes composed of a double layer of lipid molecules [7]. These lipid bilayer membranes, or simply lipid bilayers, have a dual nature. Seen from a continuum point of view as elastic solids, they are soft materials which yet have great strength despite being only two molecules in thickness. However, from a molecular viewpoint, the individual lipid molecules are free to drift past one another in their own layer – the membrane can also be seen as a quasi-two dimensional viscous fluid.

Lipid molecules ("lipids") in solution will spontaneously aggregate into forms dependent on a variety of factors, of which lipid geometry and concentration are the two most important [2]. The aggregates form not by strong molecular bonding but by a "soft" entropic force, the *hydrophobic force*. Since the bilayer molecules are only weakly bound to one another, it is the hydrophobic force which is responsible for giving the bilayer integrity [3].

The aim of the current work is to extend a continuum paradigm of the lipid bilayer by modelling the hydrophobic force in a more physical way. This is important not only for our first-principles understanding of lipid bilayers and their emergent properties, but also for multiscale simulations which must pass data between the different scales of the simulation. Indeed, we work at

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*,
Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_169,

what is loosely termed the mesoscale, being for the purposes of this paper a length and time scale at which both molecular and continuum descriptions of the bilayer are valid.

2 The Modelling

2.1 The Paradigm of Blom & Peletier

Reference [1] introduces a continuum paradigm of the lipid bilayer based on the mesoscopic dynamics framework of [4]. The aim is to minimise the intrinsic free energy of a system of water and lipid molecules with respect to the constraint that the molecular distribution gives rise to the continuous volume fractions, the variables of the paradigm. The underlying assumption is that the (unobservable) microstate relaxes to equilibrium over the relatively long time scale of the continuous description. We work in one dimension, which is more amenable to analysis and computation.

Figure 1 shows the general setup of the system in one dimension. Lipids are represented by a head and tail bead of zero dimension, connected by a rigid rod of length ϵ. Water molecules are represented by points. The single direction x is normal to the plane of the membrane, and the lipids fall into one of two classes: those whose tails, having density $u(x)$, point towards positive x, and those whose tails, having density $v(x)$, point towards negative x. The total tail density is $u+v$, while the total head density is $\tau_{-\epsilon}u+\tau_{\epsilon}v$, where $\tau_{-\epsilon}u(x) = u(x+\epsilon)$ and so on.

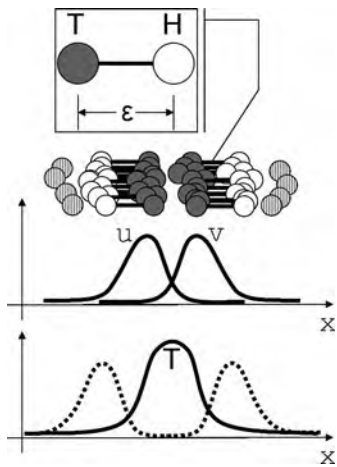


Fig. 1. Cartoon of the setup, showing the basic lipid structure, the parameter ϵ, the direction x normal to the bilayer plane, and the water molecules (*hatched circles*). The densities u, v of the two tail groups are sketched, along with total tail (*solid line*) and head (*dashed line*) densities

The total (Gibbs) free energy of the system of lipids and waters has three terms, modelling the entropy, compressibility, and intermolecular interactions in turn. In the next subsection, we outline the physical nature of the hydrophobic force, before formulating the free energy functional in Sect. 2.3.

2.2 The Hydrophobic Force

Liquid water is a dynamic hydrogen bond network in which each water molecule forms up to four hydrogen bonds with its neighbours. The non-zero dipole moment of lipid heads enables them to accept hydrogen bonds from waters (but unable to bond to one another): they are hydrophilic. By contrast, the hydrophobic lipid tails are unable to form hydrogen bonds. Any thermodynamic or electrostatic interactions between molecules are ignored here, since in liquid water at room temperature the hydrogen-bond energy is typically an order of magnitude stronger [5, 7].

Introducing a hydrophobic moiety creates a cavity with a structured "surface" in the hydrogen bond network, causing a decrease in the system entropy [6]. An entropic "force" acts to gather together hydrophobic moieties so as to minimize the disruption to the hydrogen bond network. However, the physical origin of this hydrophobic effect is that water molecules close to a sufficiently large hydrophobic moiety no longer participate in four hydrogen bonds; with no attractive force towards the hydrophobic moiety, these molecules' remaining bonds now draw them away from the moiety. The basis of the model is this fact that lipids aggregate because lipid heads can be, and tails cannot be, nodes in the hydrogen bond network.

Moreover, first principles arguments and molecular-level simulations feature only attractive forces (other than at very small distances), making preferable a model dealing with attractive forces between molecules. In [1], the original model of the hydrophobic interaction moved tails away from heads and waters by penalising proximity between them, mimicking the *effect* of the hydrophobic force but not the underlying *cause*, which ultimately rests on the attractive forces of the hydrogen bond network. By contrast, our approach is to promote water-water and water-head (but not head-head) proximity, modelling the hydrogen bond network, and effectively to ignore the hydrophobic tails. The relative strength of the water-water bonding preference to the water-head bonding preference is controlled by a parameter γ.

2.3 The Free Energy Functional

The ideal part of the free energy, roughly corresponding to the Helmholtz free energy, deals with connectivity interactions, and is given by the first two integrals in (1) which are the same as those in [1]. The new non-ideal term, here modelling the hydrophobic force, is the final term in (1).

$$E = T\int [\eta(u) + \eta(v) + \eta(w)]\ dx + \frac{p}{2}\int (1 - u - v - \tau_{-\epsilon}u - \tau_{\epsilon}v - w)^2\ dx$$
$$+\alpha \int w\hat{\kappa} * [w + \gamma(\tau_{-\epsilon}u + \tau_{\epsilon}v)]\ dx. \tag{1}$$

The first integral is an entropic term which encourages mixing, where $\eta(s) = \log s$ for $s > 0$ and $\eta(s) = \infty$ otherwise, and T is the system temperature. The second integral, in which p is the system pressure, is a potential energy due to compressibility. Herein we take $p = \infty$, corresponding to an assumption of incompressibility. This requires the integrand of the second term to be zero, which we later use to eliminate w.

In the third term, being our model of the hydrophobic force, $*$ represents convolution of the form

$$(f\hat{\kappa} * g)(x) = \int f(x)\hat{\kappa}(x - y)g(y)\ dy,$$

and the interaction kernel $\hat{\kappa}$ is given by

$$\hat{\kappa}(s) = \kappa_0 - \frac{1}{2\beta}e^{-\frac{|s|}{\beta}}, \tag{2}$$

for a constant κ_0. Since, according to our earlier discussion, the free energy is minimised subject to certain constraints, this third term promotes water-water proximity $w\hat{\kappa} * w$ and water-head proximity $w\hat{\kappa} * \gamma(\tau_{-\epsilon}u + \tau_{\epsilon}v)$.

The next step is to formally apply a variational calculus approach to minimise the free energy of the water-lipid mixture.

2.4 Calculus of Variations

To simplify the analysis, we first neglect the entropy of the water molecules, since its effect on the solvation of lipid molecules is small [6]. Next, using the assumption of incompressibility and a maximum normalised total density of unity, we rewrite the water density w as a linear combination of the lipid densities. Finally, we scale T into α, such that α represents (the inverse of) temperature effects, and set $\gamma = 1$, meaning that heads can be seen as "attached waters". This can be generalised in future work.

Choosing $\kappa_0 = (2 - e^{-L/\beta})/2L$ in (2), where the integration interval is $[-L, L]$, the energy functional becomes

$$E_I = \int [\eta(u) + \eta(v)]\ dx + \alpha\left(1 - 2c_0 - \frac{m}{2L}\right)\int (1 - u - v - \tau_{-\epsilon}u - \tau_{\epsilon}v)\ dx$$
$$-\alpha \int (1 - u - v - \tau_{-\epsilon}u - \tau_{\epsilon}v)\kappa * (1 - u - v)\ dx. \tag{3}$$

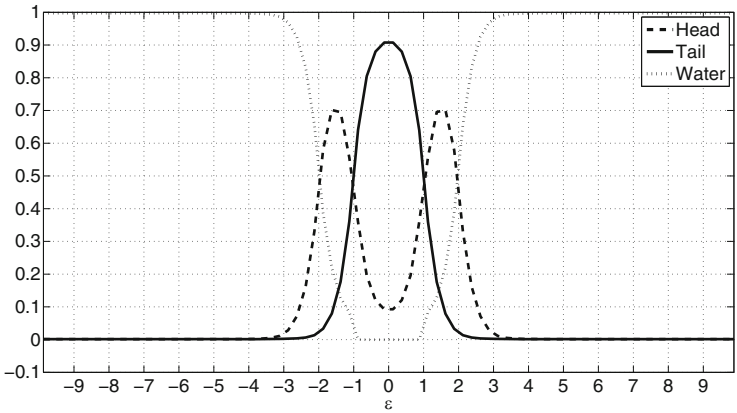


Fig. 2. A sample "bilayer" for the parameter set $\alpha = 3, \epsilon = 2, \beta = 1, c_0 = 0.024, m = 0.05 * 2L$

We use the method of Lagrange multipliers to rewrite this as

$$E_T = E_I + \frac{K}{2}\int \mu^2 \, dx + \lambda_+ \left(m - \int u + v - 2c_0 \, dx \right) + \lambda_- \left(\int u + v - 2c_0 \, dx - m \right), \tag{4}$$

where K and $\lambda_\pm$ are Lagrange multipliers and $\mu = (u + v + \tau_{-\epsilon}u + \tau_\epsilon v - 1)_+$, with $(\cdot)_+ = \max\{\cdot, 0\}$. The second term on the right hand side is the condition of non-negative water density, and the final two terms represent the mass conservation condition.

We apply formal methods of variational calculus to derive the Euler-Lagrange equations

$$0 = \log u - \alpha\kappa * (2u + 2v + 2\tau_{-\epsilon}u + \tau_{-\epsilon}v + \tau_\epsilon v) + K\mu + K\mu(x + \epsilon) + \lambda, \tag{5}$$

$$0 = \log v - \alpha\kappa * (2u + 2v + \tau_{-\epsilon}u + \tau_\epsilon u + 2\tau_\epsilon v) + K\mu + K\mu(x - \epsilon) + \lambda, \tag{6}$$

where $\lambda = \lambda_- - \lambda_+ + 1 + 3\alpha - 2\alpha(1 - 2c_0 - m/2L)$.

3 Numerical Results

We solve the Euler–Lagrange equations numerically by replacing them with evolution equations based on gradient flows, namely $u_t = -(\delta E/\delta u)$ and $v_t = -(\delta E/\delta v)$, respectively. The resulting equations are solved on a quasi-periodic domain of period $2L$.

A sample numerical result is shown in Fig. 2. The lipids have formed a well-defined bilayer structure, with a central hydrophobic tail zone from which water is excluded, separated from a water zone by two clear peaks in the head density.

4 Discussion

4.1 The Physical Nature of the Results

In a physical system, the ratio of the thickness of the head zone to that of the tail zone depends on the choice of lipid molecule (other factors such as temperature being equal) and so characterizes the bilayer properties for our purposes. In [8] we show that by varying the model parameters we can find a solution corresponding closely to a desired physical bilayer, and summarise the method of doing so. Indeed, our model is shown to behave physically in terms of the temperature effects, in contrast to the original model of [1].

4.2 Analytical Tool

A "short-range" interaction approximation takes $\beta \to 0$, so that the interaction kernel (2) approaches $\kappa_0 - \delta(s)$, where $\delta(s) = 1$ if $s = 0$ and 0 otherwise. The Euler–Lagrange equations (5,6) can then be solved analytically, and general existence conditions derived. Furthermore, β is a useful tool for smoothly adjusting numerical results to match with experimental data. More detail can be found in [8].

5 Conclusions and Further Work

We introduced a more general model of the hydrophobic effect into the continuum paradigm of [1]; the resulting one-dimensional numerical solutions resemble physical lipid bilayers. The new model is based on a consideration of the physical nature of the hydrogen bond network. A brief discussion of the physical nature of the model and analytical approximations was given. For example, the new parameter β introduces a short-range interaction approximation. The main aim of future work is to consider higher dimensions, and include compressibility effects by allowing p to vary. Other parameters can also be studied, and their underlying physical significance explored, leading if possible to a priori values.

References

1. Blom, J.G., Peletier, M.A.: Eur. J. Appl. Math. **15**, 487–508 (2004)
2. Boal, D.: Mechanics of the Cell. CUP, Cambridge (2002)
3. Chandler, D.: Nature **437**, 640–647 (2005)
4. Fraaije, J.G.E.M.: J. Chem. Phys. **11**, 9202–9212 (1993)
5. Immergut, E.H.: Encyclopedia of Applied Physics, vol. 18. VCH, Berlin (1991)
6. Kronberg, B., Costas, M., Silveston, R.: Pure Appl. Chem. **67**, 897–902 (1995)
7. Mouritsen, O.G.: Life – as a Matter of Fat. Springer, Berlin (2005)
8. Wilson, P.L., Huang, H., Takagi, S.: Commun. Comput. Phys. **6**(3), 655–672 (2009)

Wavelet Transform in Speech Segmentation

M. Ziółko,[1] J. Gałka,[1] and T. Drwiega[2]

[1] Department of Electronics, AGH University of Science and Technology, Kraków, Poland, ziolko@agh.edu.pl, jgalka@agh.edu.pl
[2] Faculty of Applied Mathematics, AGH University of Science and Technology, Kraków, Poland, drwiega@wms.mat.agh.edu.pl

Summary. A non-uniform speech segmentation method based on discrete wavelet transform is used for the localization of phoneme boundaries. A vector of real values representing the digital speech signal is decomposed into phone-like units by placing segment borders according to the result of the multiresolution analysis. The final decision on localization of boundaries is taken by analysis of the energy flow among the decomposition levels. Distribution-like event functions indicate events, regarded as the segment boundaries.

1 Introduction

Many speech segmentation algorithms (see [1, 2]) have been used in systems built for the speech technology, but only a few use the wavelet spectra [1, 5]. Wavelet methods are known to be very useful in the time-frequency analysis of signals. Wavelet transform combines the best properties of classic frequency and time analysis in a common tool.

Most of the segmentation methods utilise some kind of statistical modelling of the signals and use optimisation methods (Viterbi decoding or dynamic time warping (DTW))(see [4]). These methods can only be used if the proper models of the language are known. This assumption leads to the necessity of preparing such models what usually is rough and time-consuming task. The algorithm proposed in this paper is feature-driven and thus does not need any additional language models. Phonetically annotated database of spoken Polish – *Corpora'97* was used for tuning and testing the method.

2 Wavelet Decomposition

The discrete wavelet transformation (DWT) belongs to the group of frequency transformations and is used to obtain a time-frequency spectrum (see [3, 8]) of signal $\{s(n)\}$. This encourages us to use the DWT as an artificial method

A.D. Fitt et al. (eds.), *Progress in Industrial Mathematics at ECMI 2008*,
Mathematics in Industry 15, DOI 10.1007/978-3-642-12110-4_170,

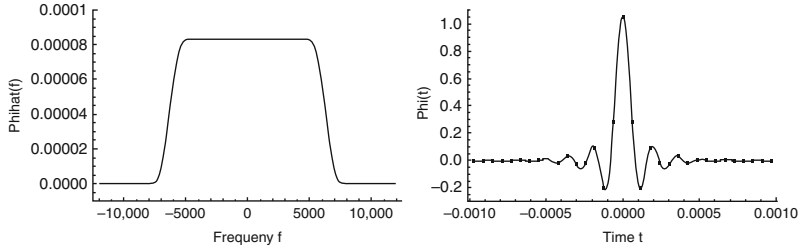


Fig. 1. Spectrum (*left figure*) and its Meyer scale function with $N = 33$ samples (*right figure*)

of speech analysis. Dyadic frequency division makes the DWT much more compatible with the principles of the operation of human hearing system, equipped with subsystem for frequency analysis (to reveal the information important for speech recognition ability), than other methods.

In order to obtain the DWT, the coefficients $c_{m+1,i}$ of series

$$s(n) = \sum_i c_{m+1,i} \phi_{m+1,i}(n) \tag{1}$$

are computed for $m = M, M-1, \ldots, 1$, where

$$\phi_{m,i}(n) = 2^{\frac{m}{2}} \phi\left(2^m n \Delta t - i\right) \tag{2}$$

is the ith wavelet function at the mth resolution level and Δt is the sampling density. An example of wavelet function $\phi(t)$ and its spectrum is presented in Fig. 1. Due to the orthogonality of wavelet functions $\{\phi_{m+1,i}\}_i$ we obtain

$$\begin{aligned} c_{m+1,i} &= 2^{\frac{m+1}{2}} \int_{-\infty}^{+\infty} s_a(t) \phi\left(2^{m+1} t - i\right) dt \\ &= 2^{\frac{m+1}{2}} \sum_{n=-\infty}^{+\infty} s_a(n) \int_{-\infty}^{+\infty} \phi\left(2^{m+1} t - i\right) \frac{\sin\left(\pi\left(t - n\Delta t\right)/\Delta t\right)}{\pi\left(t - n\Delta t\right)/\Delta t} dt, \end{aligned} \tag{3}$$

where $s_a(t)$ is an analog signal and its samples create the digital signal, *i.e.* $s_a(n\Delta t) = s(n)$.

Formula (3) has two disadvantages which are very important from the computational point of view. Firstly, it is difficult to compute integrals numerically when wavelet supports are unlimited. Secondly, the numerical computations of integrals are time-consuming, because the high quality standard needs 16,000 series (1) for each second of the recorded speech signal. Therefore instead of formula (3), we used approximation

$$c_{m+1,i} = \sum_{n \in D_i} s(n) \phi_{m+1,i}(n), \tag{4}$$

where D_i are compact supports of $\phi_{m+1,i}$.

The support of scale function $\phi(t)$ must be compact to provide the fast calculations in the real time. It is common feature of the scale functions that $\phi(t) \longrightarrow 0$ very fast as $|t| \longrightarrow +\infty$. In practice the support can be limited to the segment $[-T, T]$ where

$$T = \max\{t \in \mathbb{R} : |\phi(t)| \geq h\}. \tag{5}$$

The threshold h should depend on the extreme value of the scale function. We choose condition $h = \alpha \cdot \max_t |\phi(t)|$, where α can be taken arbitrary, e.g. $\alpha = 0.001$. In that way, the support of scale function was bounded to obtain the reasonable compromise: fast computations in real time and relatively small errors.

The number of samples should be the smallest integer value N which satisfies inequality $(N-1)\,\Delta t \geq 2T$, that is $N \geq 1 + 32{,}000T$ because the sampling frequency $f_s = 1/\Delta t = 16{,}000\,\text{Hz}$. The sampling density in the frequency domain $\Delta f = 0.5/T$ and $(N-1)\,\Delta f \geq 16{,}000\,\text{Hz}$ because the whole frequency band is spread from $-8{,}000$ to $8{,}000\,\text{Hz}$.

The coefficients of the lower level are calculated by applying the well known (see [3, 9]) formulae

$$c_{m,n} = \sum_i h_{i-2n} c_{m+1,i} \tag{6}$$

$$d_{m,n} = \sum_i g_{i-2n} c_{m+1,i} \tag{7}$$

where $\{h_i\}$ and $\{g_i\}$ are the coefficients which depend on the assumed pair: scale function ϕ and wavelet ψ. In other words, the speech spectrum is decomposed using digital filtering and downsampling procedures defined by (6) and (7). It means that given the wavelet coefficients $c_{m+1,i}$ of the $(m+1)$th resolution level, (6) and (7) are applied to compute the coefficients of the mth resolution level. The coefficients of next resolution levels are calculated recursively by applying formulae (6) and (7). The multiresolution analysis gives a hierarchical and fast scheme for the computation of the wavelet spectrum for a given signal s.

The undertaken experiments show that the speech signal decomposition into six levels is sufficient (see Table 1) to cover the frequency band of voice. he energy of the speech signal above 8 kHz and below 125 Hz is very low and can be neglected.

The above presented wavelet decomposition leads to series

$$s(n) = \sum_i c_{1,i}\phi_{1,i}(n) + \sum_{m=1}^{M}\sum_i d_{m,i}\psi_{m,i}(n) \tag{8}$$

where

$$\phi_{1,i}(n) = 2^{(1-M)/2} \begin{cases} \phi\left(\left(2^{1-M}n - i\right)\Delta t\right) & \text{if} \quad 0 \leq 2^{1-M}n - i \leq N-1 \\ 0 & \text{for other} \quad 2^{1-M}n - i \end{cases} \tag{9}$$

Table 1. Frequency division obtained for $M = 6$ levels of dyadic wavelet decomposition. Sampling frequency $f_s = 16\,\text{kHz}$

Decomposition level m	Frequency band [Hz]
6	4,000–8,000
5	2,000–4,000
4	1,000–2,000
3	500–1,000
2	250–500
1	125–250
Approximation	0–125

and

$$\psi_{m,i}(n) = 2^{(m-M)/2} \begin{cases} \psi\left(\left(2^{1-M}n - i\right)\Delta t\right) & \text{if} \quad 0 \le 2^{m-M}n - i \le N-1 \\ 0 & \text{for other} \quad 2^{m-M}n - i \end{cases} \tag{10}$$

The elements of the DWT for a mth level may be collected into a vector $\mathbf{d}_m = (d_{m,1}, d_{m,2}, \ldots)^T$. In this way the values of DWT for $M+1$ levels can be obtained. It means that discrete wavelet spectrum

$$\text{DWT}(s) = \{\mathbf{d}_M, \mathbf{d}_{M-1}, \ldots, \mathbf{d}_1, \mathbf{c}_1\} \tag{11}$$

is created from the coefficients of series (8).

3 Segmentation Scheme

The role of the segmentation algorithm is to detect significant transitions of the energy among the wavelet sub-bands. When significant enough transition is found, it is marked and scored as a spectral-phonetic event. It is assumed that events occur when the energy transition changes the order of the power-sorted bands.

The non-uniform segmentation algorithm consists of the following steps:

1. Decompose signal s into the six levels of DWT $= \{\mathbf{d}_{6,n}, \mathbf{d}_{5,n}, \ldots, \mathbf{d}_{1,n}\}$.
2. Calculate the sum of power samples in all frequency sub-bands according to rule

$$B_{m,k} = \sum_{n=(k-1)\cdot 2^{6-m}+1}^{k\cdot 2^{6-m}} d_{m,n}^2. \tag{12}$$

3. Calculate the power envelopes as a running mean values

$$B_{m,k}^{env} = \frac{1}{2\cdot\left\lfloor\frac{K}{2}\right\rfloor + 1} \sum_{n=k-\left\lfloor\frac{K}{2}\right\rfloor}^{k+\left\lfloor\frac{K}{2}\right\rfloor} B_{m,n}, \tag{13}$$

where $K = 2^{-M}\Delta t_\mu \cdot f_s$ for expected mean duration Δt_μ of the segment of speech. For the given $\Delta t_\mu = 100$ ms, $f_s = 16$ kHz and $M = 6$ we obtain $K = 25$ samples.

4. Generate importance matrix $\mathbf{M} = [M_{m,k}] \in \mathbb{R}^{6\times L}$ of frequency bands by sorting the envelopes in each time k position *i.e.*

$$\mathbf{M}_k = \{m_i\}_{i=1}^{6} : \; B^{env}_{m_1,k} \geq B^{env}_{m_2,k} \geq B^{env}_{m_3,k} \geq B^{env}_{m_4,k} \geq B^{env}_{m_5,k} \geq B^{env}_{m_6,k}$$

where L depends on the length of the speech signal.

5. Compute event-function

$$f(k) = \sum_{m=1}^{6} \frac{|M_{m,k+1} - M_{m,k}|}{m}. \tag{14}$$

6. Segment border's locations can now be extracted from $f(k)$ by choosing its local maxima, which fulfill two conditions:
   - Each of the chosen maximum has to be the highest value within the neighborhood of Δt_{min} milliseconds, which is related to minimal assumed segment duration,
   - Local maximum is greater than specified threshold f_{tr}.

Time-range condition rejects multiple changes related to the same border and segments shorter than Δt_{min}. Threshold adjusts sensitivity of the segmentation. By increasing its value we reduce the number of chosen events. It is reasonable to set its value on-line, according to

$$f_{tr}(k) = \frac{\beta \cdot \sum_{n=-P}^{P} f(k-n)}{2P}, \tag{15}$$

where P is adaptation range corresponding to 100 ms.

4 Conclusions

Presented algorithm was tested using Polish annotated speech database – *Corpora'97*. The speech of five different persons, with 1825 utterances were used for evaluation. These utterances include all of the 37 phonemes of Polish language and its natural concatenations. Reference phonetic annotation of speech was known, since it had been prepared earlier. Various values of the detection parameters Δt_{min} and β were used in order to find the combination producing the less number of errors.

The best results were obtained for parameter Δt_{min} set in the range 10–20 ms. In this range phone recognition, insertion and deletion rates are taking their best values. Threshold adaptation factor β does not affect mentioned rates when is set within 0–1. When β obtains the values greater than

1, results degrade considerably because of increase the rate of deletions, which are the most corrupting errors in speech segmentation (see [6]).

It must be mentioned, that segmentation procedure uses acoustic, not phonetic features of speech. It will result in increased level of insertion rate because some phonemes are not acoustically uniform . This feature, however, does not affect overall performance of speech recognition systems (see [6,7]).

The use of wavelet analysis turns out to be an effective tool in finding the boundaries between two phonemes. The use of non-uniform segmentation reduces total number of segments to be processed by higher-level parts of ASR systems (HMM modeling). The effect is a significant decrease of Viterbi decoding search-space and computational cost.

Acknowledgments

We would like to thank Stefan Grocholewski form Institute of Computer Science, Poznań University of Technology for providing a corpus of spoken Polish – *Corpora'97*. This work was supported by grant R00 035 02.

References

1. Alani, A., Deriche, M.: Proceedings of The Fifth International Symposium on Signal Processing and its Applications, pp. 127–130 (1999)
2. Cheng, S., Wang, H.: Proceedings of 8th European Conference on Speech Communication and Technology – EUROSPEECH, pp. 945–948 (2003)
3. Daubechies, I.: Ten Lectures on Wavelets. SIAM, Philadelphia (1992)
4. Demuynck, K., Laureys, T.: Proceedings of the 5th International Conference on Text, Speech and Dialogue, pp. 277–284 (2002)
5. Farooq, O., Datta, S.: IEE Proc. Vis. Image Signal Process. **151**(3), 187–193 (2004)
6. Gaka, J., Ziko, B.: NAUN Int. J. Circuits Syst. Signal Process. **2**(1), 167–172 (2007)
7. Grocholewski, S.: Proceedings of International Conference on Language Resources and Evaluation, pp. 1059–1062 (1998)
8. Meyer, Y.: Wavelets and Applications. Masson, Paris (1991)
9. Rioul, O., Vetterli, M.: IEEE Signal Process. Mag. **8**, 11–38 (1991)

Author Index