Lecture Notes in Computational Science and Engineering

66

Björn Engquist • Per Lötstedt • Olof Runborg
Editors

Multiscale Modeling and Simulation in Science

With 109 Figures and 4 Tables

Springer

Björn Engquist
Department of Mathematics
The University of Texas at Austin
1 University Station C1200
Austin, TX 78712-0257
USA
engquist@math.utexas.edu

Per Lötstedt
Department of Information Technology
Uppsala University
751 05 Uppsala
Sweden
perl@it.uu.se

Olof Runborg
Department of Numerical Analysis
and Computer Science
Royal Institute of Technology
100 44 Stockholm
Sweden
olofr@nada.kth.se

ISBN 978-3-540-88856-7 e-ISBN 978-3-540-88857-4

Lecture Notes in Computational Science and Engineering ISSN 1439-7358

Library of Congress Control Number: 2008939216

Mathematics Subject Classification (2000): 65-01, 65P99, 35B27, 65R20, 70-08, 42C40, 76S05, 76T20

Cover design: deblik, Berlin

Printed on acid-free paper

9 8 7 6 5 4 3 2 1

springer.com

Preface

Most problems in science involve many scales in time and space. An example is turbulent flow where the important large scale quantities of lift and drag of a wing depend on the behavior of the small vortices in the boundary layer. Another example is chemical reactions with concentrations of the species varying over seconds and hours while the time scale of the oscillations of the chemical bonds is of the order of femtoseconds. A third example from structural mechanics is the stress and strain in a solid beam which is well described by macroscopic equations but at the tip of a crack modeling details on a microscale are needed.

A common difficulty with the simulation of these problems and many others in physics, chemistry and biology is that an attempt to represent all scales will lead to an enormous computational problem with unacceptably long computation times and large memory requirements. On the other hand, if the discretization at a coarse level ignores the fine scale information then the solution will not be physically meaningful. The influence of the fine scales must be incorporated into the model.

This volume is the result of a Summer School on Multiscale Modeling and Simulation in Science held at Bosön, Lidingö outside Stockholm, Sweden, in June 2007. Sixty PhD students from applied mathematics, the sciences and engineering participated in the summer school.

The purpose of the summer school was to bring together leading scientists in computational physics, computational chemistry and computational biology and in scientific computing with PhD students in these fields to solve problems with multiple scales of research interest. By training the students to work in teams together with other students with a different background to solve real life problems they will be better prepared for their future work in academia, institutes, or industry. The importance of interdisciplinary science will certainly grow in the coming years.

There were lectures on computational multiscale techniques in the morning sessions of the first week. Most of these lectures are found in the first, tutorial part of this volume. The afternoons were devoted to the solution of mathematical and computational exercises in small groups. The exercises are interspersed in the articles in the first part. The speakers and the titles of their lectures were:

- Jørg Aarnes, Department of Applied Mathematics, SINTEF, Oslo: *Multiscale Methods for Subsurface Flow*
- Björn Engquist, Department of Numerical Analysis, KTH, Stockholm, and Department of Mathematics, University of Texas, Austin: *Introduction to Analytical and Numerical Multiscale Modeling*
- Heinz-Otto Kreiss, Department of Numerical Analysis, KTH, Stockholm: *Ordinary and Partial Differential Equations with Different Time Scales*
- Claude Le Bris, CERMICS, École Nationale des Ponts et Chaussées, Marne la Vallée: *Complex Fluids*
- Olof Runborg, Department of Numerical Analysis, KTH, Stockholm: *Introduction to Wavelets and Wavelet Based Homogenization*
- Richard Tsai, Department of Mathematics, University of Texas, Austin: *Heterogeneous Multiscale Method for ODEs*
- Lexing Ying, Department of Mathematics, University of Texas, Austin: *Fast Algorithms for Boundary Integral Equations*

In the second week, nine realistic problems from applications in astronomy, biology, chemistry, and physics were solved in collaborations between senior researchers and the PhD students. The problems were presented by experts in the applications in short lectures. Groups of students with different backgrounds worked together on the solutions with guidance from an expert. The week ended with oral presentations of the results and written papers. The student papers are found at the homepage of the summer school `www.ngssc.vr.se/S2M2S2`. The students received credit points at their home university for their participation as a part of the course work for the PhD degree. As a break from the problem solving sessions, there were three invited one hour talks on timely topics:

- Tom Abel, Department of Physics, Stanford University: *First Stars in the Universe*
- Lennart Bengtsson, Max Planck Institut für Meteorologie, Hamburg: *Climate Modeling*
- Yannis Kevrekidis, Department of Chemical Engineering, Princeton University: *Equation-free Computation for Complex and Multiscale Systems*

These are the nine different projects with the project leaders:

- **Climate Modeling**
  - Erland Källén, Heiner Körnich, Department of Meteorology, Stockholm University: *Climate Dynamics and Modelling* (two projects)
- **Solid State Physics**
  - Peter Zahn, Department of Physics, Martin-Luther-Universität, Halle-Wittenberg: *Complex Band Structures of Spintronics Materials*
  - Erik Koch, Eva Pavarini, Institut für Festkörperforschung, Forschungszentrum Jülich, Jülich: *Orbital Ordering in Transition Metal Oxides*
- **Astrophysics**
  - Garrelt Mellema, Stockholm Observatory, Stockholm University: *Photo-Ionization Dynamics Simulation*

  - Axel Brandenburg, Nordita, Stockholm: *Turbulent dynamo simulation*
- **Quantum Chemistry**
  - Yngve Öhrn, Erik Deumens, Department of Chemistry and Physics, University of Florida, Gainesville: *Molecular Reaction Dynamics with Explicit Electron Dynamics*
- **Molecular Biology**
  - Håkan Hugosson, Hans Ågren, Department of Theoretical Chemistry, KTH, Stockholm: *Quantum Mechanics - Molecular Mechanics Modeling of an Enzyme Catalytic Reaction*
- **Flow in Porous Media**
  - James Lambers, Department of Energy Resources, Stanford University: *Coarse-scale Modelling of Flow in Gas-Injection Processes for Enhanced Oil Recovery*

The projects were chosen to contain a research problem that could be at least partly solved in a week by a group of students with guidance from a senior researcher. The problems had multiple scales where the finest scale cannot be ignored. Part two of this volume contains a short description of the projects mentioned above.

The summer school was organized by the Department of Numerical Analysis and Computer Science (NADA), KTH, Stockholm, the Department of Information Technology and the Centre for Dynamical Processes and Structure Formation (CDP) at Uppsala University with an organizing committee consisting of Timo Eirola, Helsinki, Björn Engquist, Stockholm, Bengt Gustafsson, Uppsala, Sverker Holmgren, Uppsala, Henrik Kalisch, Bergen, Per Lötstedt, Uppsala, Anna Önehag, Uppsala, Brynjulf Owren, Trondheim, Olof Runborg, Stockholm, Anna-Karin Tornberg, Stockholm.

Financial support was received from the Swedish National Graduate School in Scientific Computing (NGSSC), Swedish Foundation for Strategic Research (SSF), Centre for Dynamical Processes and Structure Formation (CDP), Nordita, NordForsk, Research Council of Norway, and Comsol. Computing resources were provided by Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX).

Stockholm, Uppsala,
September 2008

Björn Engquist
Per Lötstedt
Olof Runborg

Contents

Part I Tutorials

Part II Projects

List of Contributors

Gil Ariel
Department of Mathematics
University of Texas at Austin
Austin, TX 78712, USA
ariel@math.utexas.edu

Jørg E. Aarnes
SINTEF ICT
Dept. of Applied Mathematics
P.O. Box 124, Blindern
N-0314 Oslo, Norway
Jorg.Aarnes@sintef.no

Hans Ågren
Dept. of Theoretical Chemistry
School of Biotechnology,
Royal Institute of Technology
SE-100 44 Stockholm, Sweden
agren@theochem.kth.se

Erik Deumens
Quantum Theory Project
Departments of Chemistry and Physics
University of Florida, Gainesville
FL 32611-8435, USA
deumens@qtp.ufl.edu

Björn Engquist
Department of Mathematics
University of Texas at Austin
Austin, TX 78712, USA
engquist@math.utexas.edu

Håkan W. Hugosson
Dept. of Theoretical Chemistry
School of Biotechnology
Royal Institute of Technology
SE-100 44 Stockholm, Sweden
hakan@theochem.kth.se

Tomas Johnson
Department of Mathematics
Uppsala University
P O Box 480
SE-751 06 Uppsala, Sweden
johnson@math.uu.se

Erland Källén
Department of Meteorology
Stockholm University,
SE-106 91 Stockholm, Sweden
erland@misu.su.se

Vegard Kippe
SINTEF ICT
Dept. of Applied Mathematics
P.O. Box 124, Blindern
N-0314 Oslo, Norway

Stein Krogstad
SINTEF ICT
Dept. of Applied Mathematics
P.O. Box 124, Blindern
N-0314 Oslo, Norway

Erik Koch
Institut für Festkörperforschung
Forschungszentrum Jülich
D-52425 Jülich, Germany
E.Koch@fz-juelich.de

Heiner Körnich
Department of Meteorology
Stockholm University,
SE-106 91 Stockholm, Sweden
heiner@misu.su.se

Heinz-Otto Kreiss
Träskö-Storö Institute of Mathematics
Sweden

James Lambers
Department of Energy Resources Engineering
Stanford University
Stanford, CA 94305-2220, USA
lambers@stanford.edu

Claude Le Bris
CERMICS – ENPC
École Nationale des Ponts et Chaussées
6 et 8 avenue Blaise Pascal
Cité Descartes – Champs sur Marne
F-77455 Marne la Vallée Cedex 2
France
lebris@cermics.enpc.fr

Tony Lelièvre
CERMICS – ENPC
École Nationale des Ponts et Chaussées
6 et 8 avenue Blaise Pascal
Cité Descartes – Champs sur Marne
F-77455 Marne la Vallée Cedex 2
France
lelievre@cermics.enpc.fr

Knut–Andreas Lie
SINTEF ICT
Dept. of Applied Mathematics
P.O. Box 124, Blindern
N-0314 Oslo, Norway
Knut-Andreas.Lie@sintef.no

Garrelt Mellema
Stockholm Observatory
AlbaNova University Centre
Stockholm University
SE-10691 Stockholm, Sweden
garrelt@astro.su.se

Yngve Öhrn
Quantum Theory Project
Departments of Chemistry and Physics
University of Florida
Gainesville
FL 32611-8435, USA
ohrn@qtp.ufl.edu

Eva Pavarini
Institut für Festkörperforschung
Forschungszentrum Jülich
D-52425 Jülich, Germany
E.Pavarini@fz-juelich.de

Olof Runborg
Department of Numerical Analysis
KTH
SE-100 44 Stockholm, Sweden
olofr@nada.kth.se

Patrik Thunström
Department of Physics
Theoretical Magnetism Group
Uppsala University
P O Box 530
SE-751 21 Uppsala, Sweden
patrik.thunstrom@fysik.uu.se

Richard Tsai
Department of Mathematics
University of Texas at Austin
Austin, TX 78712, USA
`ytsai@math.utexas.edu`

Lexing Ying
Department of Mathematics
University of Texas
Austin, TX 78712, USA
`lexing@math.utexas.edu`

Peter Zahn
Department of Physik
Martin-Luther-Universität Halle-Wittenberg
D-06099 Halle, Germany
`peter.zahn@physik.uni-halle.de`

Multiscale Methods for Subsurface Flow

Jørg E. Aarnes, Knut–Andreas Lie, Vegard Kippe, and Stein Krogstad

SINTEF ICT, Dept. of Applied Mathematics, Oslo Norway

Modelling of flow processes in the subsurface is important for many applications. In fact, subsurface flow phenomena cover some of the most important technological challenges of our time. To illustrate, we quote the UN's Human Development Report 2006:

> "There is a growing recognition that the world faces a water crisis that, left unchecked, will derail the progress towards the Millennium Development Goals and hold back human development. Some 1.4 billion people live in river basins in which water use exceeds recharge rates. The symptoms of overuse are disturbingly clear: rivers are drying up, groundwater tables are falling and water-based ecosystems are being rapidly degraded. Put bluntly, the world is running down one of its most precious natural resources and running up an unsustainable ecological debt that will be inherited by future generations."

The road toward sustainable use and management of the earth's groundwater reserves necessarily involves modelling of groundwater hydrological systems. In particular, modelling is used to acquire general knowledge of groundwater basins, quantify limits of sustainable use, and to monitor transport of pollutants in the subsurface.

A perhaps equally important problem is how to reduce emission of greenhouse gases, such as CO_2, into the atmosphere. Indeed, the recent report from the UN Intergovernmental Panel on Climate Change (see e.g., `www.ipcc.ch`) draws a frightening scenario of possible implications of human-induced emissions of greenhouse gases. Carbon sequestration in porous media has been suggested as a possible means. Schrag [46] claims that

> "Carbon sequestration (...) is an essential component of any serious plan to avoid catastrophic impacts of human-induced climate change. Scientific and economical challenges still exist, but none are serious enough to suggest that carbon capture and storage (in underground repositories) will not work at the scale required to offset trillions of tons of CO_2 emissions over the next century."

The primary concern related to storage of CO_2 in subsurface repositories is related to how fast the CO_2 will escape. Repositories do not need to store CO_2 forever, just long enough to allow the natural carbon cycle to reduce the atmospheric CO_2 to near pre-industrial level. Nevertheless, making a qualified estimate of the leakage rates from potential CO_2 storage facilities is a non-trivial task, and demands interdisciplinary research and software based on state-of-the art numerical methods for modelling subsurface flow.

These examples illustrate that the demand for software modelling subsurface flow will not diminish with the decline of the oil and gas era. In fact, the need for tools that help us understand flow processes in the subsurface is probably greater than ever, and increasing. Nevertheless, more than 50 years of prior research in this area has led to some degree of agreement in terms of how subsurface flow processes can be modelled adequately with numerical simulation technology. Because most of the prior research in this area targets reservoir simulation, i.e., modelling flow in oil and gas reservoirs, we will focus on this application in the remainder of this paper. However, the general modelling framework, and the numerical methods that are discussed, apply also to modelling flow in groundwater reservoirs and CO_2 storage facilities.

To describe the subsurface flow processes mathematically, two types of models are needed. First, one needs a mathematical model that describes how fluids flow in a porous medium. These models are typically given as a set of partial differential equations describing the mass-conservation of fluid phases. In addition, one needs a geological model that describes the given porous rock formation (the reservoir). The geological model is used as input to the flow model, and together they make up the reservoir simulation model.

Unfortunately, geological models are generally too large for flow simulation, meaning that the number of grid cells exceed the capabilities of current flow simulators (usually by orders of magnitude) due to limitations in memory and processing power. The traditional, and still default, way to build a reservoir simulation model therefore starts by converting the initial geomodel (a conceptual model of the reservoir rock with a plausible distribution of geological parameters) to a model with a resolution that is suitable for simulation. This process is called upscaling. Upscaling methods aim to preserve the small-scale effects in the large-scale computations (as well as possible), but because small-scale features often have a profound impact on flow occurring on much larger scales, devising robust upscaling techniques is a non-trivial task.

Multiscale methods are a new and promising alternative to traditional upscaling. Whereas upscaling techniques are used to derive coarse-scale equations with a reduced set of parameters, multiscale methods attempt to incorporate fine-scale information directly into the coarse-scale equations. Multiscale methods are rapidly growing in popularity, and have started to gain recognition as a viable alternative to upscaling, also by industry. The primary purpose of this paper is to provide an easily accessible introduction to multiscale methods for subsurface flow, and to clarify how these methods relate to some standard, but widely used, upscaling methods.

We start by giving a crash course in reservoir simulation. Next, we describe briefly some basic discretisation techniques for computing reservoir pressure and velocity fields. We then provide a brief introduction to upscaling, and present some of the most commonly used methods for upscaling the pressure equation. The final part of the paper is devoted to multiscale methods for computing pressure and velocity fields for subsurface flow applications.

1 Introduction to Reservoir Simulation

Reservoir simulation is the means by which we use a numerical model of the petrophysical characteristics of a hydrocarbon reservoir to analyse and predict fluid behaviour in the reservoir over time. For nearly half a century, reservoir simulation has been an integrated part of oil-reservoir management. Today, simulations are used to estimate production characteristics, calibrate reservoir parameters, visualise reservoir flow patterns, etc. The main purpose is to provide an information database that can help the oil companies to position and manage wells and well trajectories in order to maximize the oil and gas recovery. Unfortunately, obtaining an accurate prediction of reservoir flow scenarios is a difficult task. One of the reasons is that we can never get a complete and accurate characterisation of the rock parameters that influence the flow pattern. And even if we did, we would not be able to run simulations that exploit all available information, since this would require a tremendous amount of computer resources that exceed by far the capabilities of modern multi-processor computers. On the other hand, we do not need, nor do we seek a simultaneous description of the flow scenario on all scales down to the pore scale. For reservoir management it is usually sufficient to describe the general trends in the reservoir flow pattern.

In this section we attempt only to briefly summarise some aspects of the art of modelling porous media flow and motivate a more detailed study of some of the related topics. More details can be found in one of the general textbooks describing modelling of flow in porous media, e.g., [10, 21, 26, 30, 41, 43, 23].

1.1 The Reservoir Description

Natural petroleum reservoirs typically consist of a subsurface body of sedimentary rock having sufficient porosity and permeability to store and transmit fluids. Sedimentary rocks are formed through deposition of sediments and typically have a layered structure with different mixtures of rock types. In its simplest form, a sedimentary rock consists of a stack of sedimentary beds that extend in the lateral direction. Due to differences in deposition and compaction, the thickness and inclination of each bed will vary in the lateral directions. In fact, during the deposition, parts of the beds may have been weathered down or completely eroded away. In addition, the layered structure of the beds may have been disrupted due to geological activity, introducing fractures and faults. Fractures are cracks or breakage in the rock, across which there has been no movement. Faults are fractures across which the layers in the rock have been displaced.

Oil and gas in the subsurface stem from layers of compressed organic material that was deposited millions of years ago, and, with time, eventually turned into water and different hydrocarbon components. Normally the lightest hydrocarbons (methane, ethane, etc.) escaped quickly, whilst the heavier oils moved slowly towards the surface, but at certain sites geological activity had created and bent layers of low-permeable (or non-permeable) rock, so that the migrating hydrocarbons were trapped. It is these quantities of trapped hydrocarbons that form today's oil and gas reservoirs.

Rock formations found in natural petroleum reservoirs are typically heterogeneous at all length scales, from the micrometre scale of pore channels between sand grains to the kilometre scale of the full reservoir. To obtain a geological description of these reservoirs, one builds models that attempt to reproduce the true geological heterogeneity in the reservoir rock. However, it is generally not possible to account for all pertinent scales that impact the flow. Instead one has to create models for studying phenomena occurring at a reduced span of scales. In reservoir engineering, the reservoir is modelled in terms of a three-dimensional grid, in which the layered structure of sedimentary beds (a small unit of rock distinguishable from adjacent rock units) in the reservoir is reflected in the geometry of the grid cells. The physical properties of the rock (porosity and permeability) are represented as constant values inside each grid cell. The size of a grid block in a typical geological grid-model is in the range 10–50 m in the horizontal direction and 0.1–1 m in the vertical direction. Thus, a geological model is clearly too coarse to resolve small-scale features such as the micro-structure of the pores.

Rock Parameters

The rock *porosity*, usually denoted by ϕ, is the void volume fraction of the medium; i.e., $0 \leq \phi < 1$. The porosity usually depends on the pressure; the rock is *compressible*, and the rock compressibility is defined by:

$$c_r \quad \frac{1}{\phi}\frac{d\phi}{dp},$$

where p is the reservoir pressure. For simplified models it is customary to neglect the rock compressibility. If compressibility cannot be neglected, it is common to use a linearisation so that:

$$\phi \quad \phi_0\big(1 + c_r(p - p_0)\big),$$

where p_0 is a specified reference pressure and $\phi_0 \quad \phi(p_0)$.

The (absolute) *permeability*, denoted by K, is a measure of the rock's ability to transmit a single fluid at certain conditions. Since the orientation and interconnection of the pores are essential for flow, the permeability is not necessarily proportional to the porosity, but K is normally strongly correlated to ϕ. Rock formations like sandstones tend to have many large or well-connected pores and therefore transmit fluids readily. They are therefore described as permeable. Other formations, like shales, may have smaller, fewer or less interconnected pores, e.g., due to a high content of

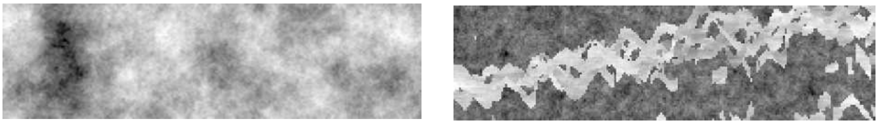


Fig. 1. Examples of two permeability fields: a shallow-marine Tarbert formation (left) and a fluvial Upper Ness formation (right).

clay. Such formations are described as impermeable. Although the SI-unit for permeability is m^2, it is commonly represented in Darcy (D), or milli-Darcy (mD). The precise definition of 1D ($\approx 0.987 \cdot 10^{-12}\,\mathrm{m}^2$) involves transmission of a 1cp fluid (see below) through a homogeneous rock at a speed of $1\,\mathrm{cm/s}$ due to a pressure gradient of $1\,\mathrm{atm/cm}$. Translated to reservoir conditions, 1D is a relatively high permeability.

In general, K is a tensor, which means that the permeability in the different directions depends on the permeability in the other directions. We say that the medium is isotropic (as opposed to anisotropic) if K can be represented as a scalar function, e.g., if the horizontal permeability is equal to the vertical permeability. Moreover, due to transitions between different rock types, the permeability may vary rapidly over several orders of magnitude, local variations in the range 1 mD to 10 D are not unusual in a typical field.

The heterogeneous structure of a porous rock formation is a result of the deposition history and will therefore vary strongly from one formation to another. In Fig. 1 we show two permeability realisations sampled from two different formations in the Brent sequence from the North Sea. Both formations are characterised by large permeability variations, 8–12 orders of magnitude, but are qualitatively different. The Tarbert formation is the result of a shallow-marine deposition and has relatively smooth permeability variations. The Upper Ness formation is fluvial and has been deposited by rivers or running water, leading to a spaghetti of well-sorted high-permeable channels of long correlation length imposed on low-permeable background.

Grids

As described above, the rock parameters ϕ and K are usually given on a grid that also gives the geometrical description of the underlying rock formations. The most widespread way to model the geometry of rock layers is by so-called *corner-point grids*. A corner-point grid consists of a set of hexahedral cells that are aligned in a logical Cartesian fashion. One horizontal layer in the grid is then assigned to each sedimentary bed to be modelled. In its simplest form, a corner-point grid is specified in terms of a set of vertical or inclined pillars defined over an areal Cartesian 2D mesh in the lateral direction. Each cell in the volumetric corner-point grid is restricted by four pillars and is defined by specifying the eight corner points of the cell, two on each pillar. Figure 2 shows a side-view of such a corner-point grid. Notice the occurrence of degenerate cells with less than eight non-identical corners where the beds are partially eroded away. Some cells also disappear completely and hence

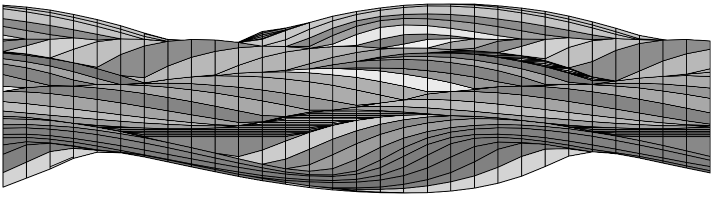


Fig. 2. Side view in the xz-plane of corner-point grid with vertical pillars modelling a stack of sedimentary beds (each layer indicated by a different colour).

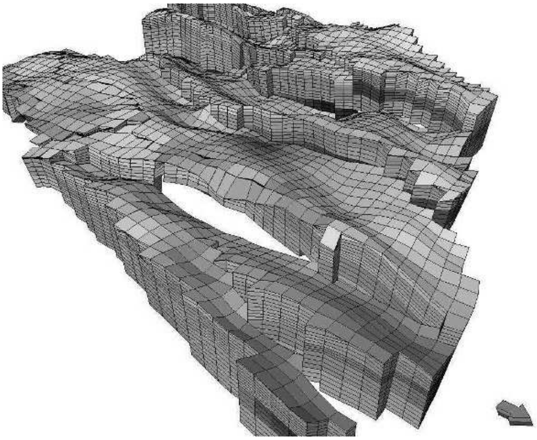


Fig. 3. Example of a geological grid model.

introduce new connections between cells that are not neighbours in the underlying logical Cartesian grid.

The corner-point format easily allows for degeneracies in the cells and discontinuities (fractures/faults) across faces. Hence, using the corner-point format it is possible to construct very complex geological models that match the geologist's perception of the underlying rock formations, e.g., as seen in Fig. 3. Due to their many appealing features, corner-point grids are now an industry standard and the format is supported in most commercial software for reservoir modelling and simulation.

1.2 Flow Parameters

The void in the porous medium is assumed to be filled with different phases. The volume fraction *s* occupied by each phase is the *saturation* of that phase. Thus,

$$\sum_{\text{all phases}} s_i \quad 1. \tag{1}$$

Here only three phases are considered; aqueous (a), liquid (l), and vapour (v). Each phase contains one or more *components*. A hydrocarbon component is a unique chemical species (methane, ethane, propane, etc). Since the number of hydrocarbon components can be quite large, it is common to group components into pseudo-components, e.g., water (w), oil (o), and gas (g).

Due to the varying conditions in a reservoir, the hydrocarbon composition of the different phases may change throughout a simulation. The mass fraction of component α in phase j is denoted by $m_{\alpha,j}$. In each of the phases, the mass fractions should add up to unity, so that for N different components, we have:

$$\sum_{\alpha\quad 1}^{N} m_{\alpha,j} \qquad 1.$$

The *density* ρ and *viscosity* μ of each phase are functions of *phase pressure* p_i ($i\quad a, l, v$) and the component composition. That is, for vapour

$$\rho_v \quad \rho_v(p_v, \{m_{\alpha,v}\}), \qquad \mu_v \quad \mu_v(p_v, \{m_{\alpha,v}\}),$$

and similarly for the other phases. These dependencies are most important for the vapour phase, and are usually ignored for the aqueous phase.

The compressibility of the phase is defined as for rock compressibility:

$$c_i \quad \frac{1}{\rho_i}\frac{d\rho_i}{dp_i}, \quad i \quad a, l, v.$$

Compressibility effects are more important for gas than for fluids. In simplified models, the compressibility of the aqueous phase is usually neglected.

Due to interfacial tensions, the phase pressures are different, defining the *capillary pressure*,

$$p_{ij}^{c} \quad p_i - p_j,$$

for $i, j \quad a, l, v$. Although other dependencies are reported, it is usually assumed that the capillary pressure is a function of the saturations only.

Even though phases do not really mix, we assume that all phases may be present at the same location. The ability of one phase to move will then depend on the environment at the actual location. That is, the permeability experienced by one phase depends on the saturation of the other phases at that specific location, as well as the phases' interaction with the pore walls. Thus, we introduce a property called *relative permeability*, denoted by $k_{ri}, i \quad a, l, v$, which describes how one phase flows in the presence of the two others. Thus, in general, and by the closure relation (1), we may assume that

$$k_{ri} \quad k_{ri}(s_a, s_v),$$

where subscript r stands for *relative* and i denotes one of the phases a, l, or v. Thus, the (effective) permeability experienced by phase i is $K_i \quad Kk_{ri}$. It is important to note that the relative permeabilities are nonlinear functions of the saturations, so that the sum of the relative permeabilities at a specific location (with a specific composition) is not necessarily equal to one. In general, relative permeabilities may depend on the pore-size distribution, the fluid viscosity, and the interfacial forces between the fluids. These features, which are carefully reviewed by Demond and Roberts [27], are usually ignored. Of greater importance to oil recovery is probably the temperature dependency [42], which may be significant, but very case-related.

Other parameters of importance are the bubble-point pressures for the various components. At given temperature, the bubble-point pressures signify the pressures where the respective phases start to boil. Below the bubble-point pressures, gas is released and we get transition of the components between the phases. For most realistic models, even if we do not distinguish between all the components, one allows gas to be dissolved in oil. For such models, an important pressure-dependent parameter is the solution gas-oil ratio r_l for the gas dissolved in oil at reservoir conditions. It is also common to introduce so-called formation volume factors that model the pressure dependent ratio of bulk volumes at reservoir and surface conditions. We will introduce these parameters later when presenting the three-phase black-oil model.

1.3 Production Processes

Initially, a hydrocarbon reservoir is at equilibrium, and contains gas, oil, and water, separated by gravity. This equilibrium has been established over millions of years with gravitational separation and geological and geothermal processes. When a well is drilled through the upper non-permeable layer and penetrates the upper hydrocarbon cap, this equilibrium is immediately disturbed. The reservoir is usually connected to the well and surface production facilities by a set of valves. If there were no production valves to stop the flow, we would have a “blow out” since the reservoir is usually under a high pressure. As the well is ready to produce, the valves are opened slightly, and hydrocarbons flow out of the reservoir due to over-pressure. This in turn, sets up a flow inside the reservoir and hydrocarbons flow towards the well, which in turn may induce gravitational instabilities. Capillary pressures will also act as a (minor) driving mechanism, resulting in local perturbations of the situation. During this stage, perhaps 20 percent of the hydrocarbons present are produced until a new equilibrium is achieved. We call this *primary production* by natural drives. One should note that a sudden drop in pressure also may have numerous other intrinsic effects. Particularly in complex, composite systems this may be the case, as pressure-dependent parameters experience such drops. This may give non-convective transport and phase transfers, as vapour and gaseous hydrocarbons may suddenly condensate.

As pressure drops, less oil and gas is flowing, and eventually the production is no longer economically sustainable. Then the operating company may start *secondary production*, by engineered drives. These are processes based on injecting water or gas into the reservoir. The reason for doing this is twofold; some of the pressure is rebuilt or even increased, and secondly one tries to push out more profitable hydrocarbons with the injected substance. One may perhaps produce another 20 percent of the oil by such processes, and engineered drives are standard procedure at most locations in the North Sea today.

In order to produce even more oil, *Enhanced Oil Recovery* (EOR, or tertiary recovery) techniques may be employed. Among these are heating the reservoir or injection of sophisticated substances like foam, polymers or solvents. Polymers are supposed to change the flow properties of water, and thereby to more efficiently push out oil. Similarly, solvents change the flow properties of the hydrocarbons, for instance by developing miscibility with an injected gas. In some sense, one tries to wash the pore walls for most of the remaining hydrocarbons. The other technique is based on injecting steam, which will heat the rock matrix, and thereby, hopefully, change the flow properties of the hydrocarbons. At present, such EOR techniques are considered too expensive for large-scale commercial use, but several studies have been conducted and the mathematical foundations are being carefully investigated, and at smaller scales EOR is being performed.

One should note that the terms primary, secondary, and tertiary are ambiguous. EOR techniques may be applied during primary production, and secondary production may be performed from the first day of production.

2 Mathematical Models

In this section we will present two mathematical models, first a simple single-phase model that incorporates much of the complexities that arise due to heterogeneities in the porous rock formations. Then we present the classical black-oil model, which incorporates more complex flow physics.

2.1 Incompressible Single-Phase Flow

The simplest possible way to describe the displacement of fluids in a reservoir is by a single-phase model. This model gives an equation for the pressure distribution in the reservoir and is used for many early-stage and simplified flow studies. Single-phase models are used to identify flow directions; identify connections between producers and injectors; in flow-based upscaling; in history matching; and in preliminary model studies.

Assume that we want to model the filtration of a fluid through a porous medium of some kind. The basic equation describing this process is the continuity equation which states that mass is conserved

$$\frac{\partial(\phi\rho)}{\partial t} + \nabla\cdot(\rho v) \quad q. \tag{2}$$

Here the source term q models sources and sinks, that is, outflow and inflow per volume at designated well locations.

For low velocities v, filtration through porous media is modelled with an empirical relation called Darcy's law after the French engineer Henri Darcy. Darcy discovered in 1856, through a series of experiments, that the filtration velocity is proportional to a combination of the gradient of the fluid pressure and pull-down effects due to gravity. More precisely, the volumetric flow density v (which we henceforth will refer to as flow velocity) is related to pressure p and gravity forces through the following gradient law:

$$v \quad -\frac{K}{\mu}(\nabla p+\rho g\nabla z). \tag{3}$$

Here g is the magnitude of the gravitational acceleration and z is the spatial coordinate in the upward vertical direction. For brevity we write $G \quad -g\nabla z$ for the gravitational pull-down force. We note that Darcy's law is analogous to Fourier's law of heat conduction (in which K is replaced with the heat conductivity tensor) and Ohm's law of electrical conduction (in which K is the inverse of the electrical resistance). However, whereas there is only one driving force in thermal and electrical conduction, there are two driving forces in porous media flow: gravity and the pressure gradient.

As an illustrative example, we will now present an equation that models flow of an incompressible fluid, say, water, through a rigid and incompressible porous medium characterised by a permeability field K and a corresponding porosity distribution ϕ. For an incompressible medium, the temporal derivative term in (2) vanishes and we obtain the following elliptic equation for the water pressure:

$$\nabla\cdot v \quad \nabla\cdot\left[-\frac{K}{\mu}(\nabla p-\rho G)\right] \quad \frac{q}{\rho}. \tag{4}$$

To close the model, we must specify boundary conditions. Unless stated otherwise we shall follow common practice and use no-flow boundary conditions. Hence, on the reservoir boundary $\partial\Omega$ we impose $v\cdot n \quad 0$, where n is the normal vector pointing out of the boundary $\partial\Omega$. This gives an isolated flow system where no water can enter or exit the reservoir.

2.2 Three-Phase Black-Oil Model

The most commonly used model in reservoir simulation is the so-called black oil model. Here we present the three-phase black-oil model, in which there are three components; water (w), oil (o), and gas (g), and three phases; aqueous (a), liquid (l), and vapour (v). The aqueous phase contains only water, but oil and gas may exist in both the liquid phase and the vapour phase. The three-phase black-oil model is governed by mass-balance equations for each component

$$\sum_{j\quad a,l,v}\left\{\frac{\partial}{dt}\left(\phi m_{\alpha,j}\rho_j s_j\right)+\nabla\cdot\left(m_{\alpha,j}\rho_j v_j\right)\right\} \quad q_\alpha, \qquad \alpha \quad w,o,g, \tag{5}$$

where the Darcy velocities v_j are given by

$$v_j \quad -\frac{Kk_{rj}}{\mu_j}\left(\nabla p_j-\rho_j G\right), \qquad j \quad a,l,v. \tag{6}$$

Here q_α is a source term and p_j denotes the phase pressure.

We now introduce the volume formation factors $b_\alpha \quad V_{\alpha s}/V_\alpha$, where $V_{\alpha s}$ and V_α are volumes occupied by a bulk of component α at surface and reservoir conditions, respectively; the phase densities at surface conditions ρ_{js}; $r_l \quad V_{gs}/V_{os}$, the ratio of the volumes of gas and oil in the liquid phase at surface conditions; and $r_v \quad V_{os}/V_{gs}$, the ratio of the volumes of oil and gas in the vapour phase at surface conditions. Recalling that water does not mix into the liquid and vapour phases, we derive

$$\begin{array}{lll}
m_{w,a}\rho_a \quad b_w\rho_{ws}, & m_{o,a} \quad 0, & m_{g,a} \quad 0,\\
m_{w,l} \quad 0, & m_{o,l}\rho_l \quad b_o\rho_{os}, & m_{g,l}\rho_l \quad r_l b_o\rho_{gs},\\
m_{w,v} \quad 0, & m_{o,v}\rho_v \quad r_v b_g\rho_{os}, & m_{g,v}\rho_v \quad b_g\rho_{gs}.
\end{array}$$

Inserting these expressions into (5) gives

$$\frac{\partial}{dt}(\phi A\ s_{j_\ulcorner}) + \nabla\cdot(A\ v_{j_\ulcorner}) \qquad q_{\alpha_\ulcorner}, \tag{7}$$

where $\xi_{j_\ulcorner} \quad (\xi_a,\xi_l,\xi_v)^t$, $\xi_{\alpha_\ulcorner} \quad (\xi_w,\xi_o,\xi_g)^t$, and

$$A \quad \begin{bmatrix} b_w\rho_{ws} & 0 & 0\\ 0 & b_o\rho_{os} & r_v b_g\rho_{os}\\ 0 & r_l b_o\rho_{gs} & b_g\rho_{gs}\end{bmatrix} \quad \begin{bmatrix}\rho_{ws} & 0 & 0\\ 0 & \rho_{os} & 0\\ 0 & 0 & \rho_{gs}\end{bmatrix}\begin{bmatrix}1 & 0 & 0\\ 0 & 1 & r_v\\ 0 & r_l & 1\end{bmatrix}\begin{bmatrix}b_w & 0 & 0\\ 0 & b_o & 0\\ 0 & 0 & b_g\end{bmatrix}.$$

Premultiplying (7) with $\mathbf{1}^t A^{-1}$, expanding $\partial/\partial\xi \quad (\partial/\partial p_l)(\partial p_l/\partial\xi)$, and assuming $\mathbf{1}^t\ s_{j_\ulcorner} \quad 1$, i.e., that the three phases occupy the void space completely, gives an equation of the following form:

$$\frac{\partial\phi}{\partial p_l} + \phi\sum_j c_j s_j \frac{\partial p_l}{\partial t} + \nabla\cdot\left(\sum_j v_j\right) + \sum_j c_j v_j\cdot\nabla p_l \quad q. \tag{8}$$

Exercise 1. Derive (8) from (7) and show that q and the phase compressibilities c_j are defined by

$$q \quad \mathbf{1}^t A^{-1}\ q_{\alpha_\ulcorner} \quad \frac{q_w}{b_w\rho_{ws}} + \frac{1}{1-r_v r_l}\left(\left(\frac{1}{b_o}-\frac{r_l}{b_g}\right)\frac{q_o}{\rho_{os}} + \left(\frac{1}{b_g}-\frac{r_v}{b_o}\right)\frac{q_g}{\rho_{gs}}\right).$$

and

$$c_a \quad \frac{\partial\ln b_w}{\partial p_l}, \quad c_l \quad \frac{\partial\ln b_o}{\partial p_l} + \frac{1}{b_g}\frac{b_o - r_v b_g}{1-r_v r_l}\frac{\partial r_l}{\partial p_l},$$
$$c_v \quad \frac{\partial\ln b_g}{\partial p_l} + \frac{1}{b_o}\frac{b_g - r_l b_o}{1-r_v r_l}\frac{\partial r_v}{\partial p_l}.$$

3 Discretisation of Elliptic Pressure Equations

In this section we present four different numerical methods for solving elliptic pressure equations on the form (4). We only consider mass-conservative methods, meaning that each method provides velocity fields that satisfy the following mass-balance equation:

$$\int_{\Omega_i}\nabla\cdot v\,dx \quad \int_{\partial\Omega_i} v\cdot n\,ds \quad \int_{\Omega_i}\frac{q}{\rho}\,dx \tag{9}$$

for each grid cell Ω_i in Ω (the reservoir). Here n denotes the outward-pointing unit normal on $\partial\Omega_i$ and ds is the surface area measure. We first present the two-point flux-approximation (TPFA) scheme, a very simple discretisation technique that is widely used in the oil-industry.

3.1 The Two-Point Flux-Approximation (TPFA) Scheme

In classical finite-difference methods, partial differential equations (PDEs) are approximated by replacing the partial derivatives with appropriate divided differences between point-values on a discrete set of points in the domain. Finite-volume methods, on the other hand, have a more physical motivation and are derived from conservation of (physical) quantities over cell volumes. Thus, in a finite-volume method the unknown functions are represented in terms of average values over a set of finite volumes, over which the integrated PDE model is required to hold in an averaged sense.

Although finite-difference and finite-volume methods have fundamentally different interpretation and derivation, the two labels are used interchangeably in the scientific literature. We therefore choose to not make a clear distinction between the two discretisation techniques here. Instead we ask the reader to think of a finite-volume method as a conservative finite-difference scheme that treats the grid cells as control volumes. In fact, there exist several finite-volume and finite-difference schemes of low order, for which the cell-centred values obtained with a finite-difference scheme coincide with cell averages obtained with the corresponding finite-volume scheme.

To derive a set of finite-volume mass-balance equations for (4), consider Equation (9). Finite-volume methods are obtained by approximating the pressure p with a cell-wise constant function $\{p_{w,i}\}$ and estimating the normal velocity $v \cdot n$ across cell interfaces $\gamma_{ij} \quad \partial\Omega_i \cap \partial\Omega_j$ from a set of neighbouring cell pressures. To formulate the TPFA scheme it is convenient to reformulate equation (4) slightly, so that we get an equation of the following form:

$$-\nabla \cdot \lambda \nabla u \quad f, \tag{10}$$

where $\lambda \quad K/\mu$. To this end, we have two options: we can either introduce a flow potential $u \quad p + \rho g z$ and express our model as an equation for u

$$-\nabla \cdot \lambda \nabla u \quad \frac{q}{\rho},$$

or we can move the gravity term $\nabla \cdot (\lambda \rho G)$ to the right-hand side. Hence, we might as well assume that we want to solve (10) for u.

As the name suggests, the TPFA scheme uses two points, the cell-averages u_i and u_j, to approximate the flux $F_{ij} \quad -\int_{\gamma_{ij}} (\lambda \nabla u) \cdot n \, ds$. To be more specific, let us consider a regular hexahedral grid with gridlines aligned with the principal coordinate axes. Moreover, assume that γ_{ij} is an interface between adjacent cells in the x–coordinate direction so that the interface normal n_{ij} equals $(1,0,0)^T$. The gradient ∇u on γ_{ij} in the TPFA method is now replaced with

$$(\nabla u \cdot n)|_{\gamma_{ij}} \approx \frac{2(u_j - u_i)}{\Delta x_i + \Delta x_j}, \tag{11}$$

where Δx_i and Δx_j denote the respective cell dimensions in the x-coordinate direction. Thus, we obtain the following expression for F_{ij}:

$$F_{ij} = -\frac{2(u_j - u_i)}{\Delta x_i + \Delta x_j} \int_{\gamma_{ij}} \lambda \, ds.$$

However, in most reservoir simulation models, the permeability K is cell-wise constant, and hence not well-defined at the interfaces. This means that we also have to approximate λ on γ_{ij}. In the TPFA method this is done by taking a distance-weighted harmonic average of the respective directional cell permeabilities, $\lambda_{i,ij} = n_{ij} \cdot \lambda_i n_{ij}$ and $\lambda_{j,ij} = n_{ij} \cdot \lambda_j n_{ij}$. To be precise, the n_{ij}–directional permeability λ_{ij} on γ_{ij} is computed as follows:

$$\lambda_{ij} = (\Delta x_i + \Delta x_j) \left(\frac{\Delta x_i}{\lambda_{i,ij}} + \frac{\Delta x_j}{\lambda_{j,ij}} \right)^{-1}.$$

Hence, for orthogonal grids with gridlines aligned with the coordinate axes, one approximates the flux F_{ij} in the TPFA method in the following way:

$$F_{ij} = -|\gamma_{ij}| \lambda_{ij} (\nabla u \cdot n)|_{\gamma_{ij}} = 2|\gamma_{ij}| \left(\frac{\Delta x_i}{\lambda_{i,ij}} + \frac{\Delta x_j}{\lambda_{j,ij}} \right)^{-1} (u_i - u_j). \qquad (12)$$

Finally, summing over all interfaces, we get an approximation to $\int_{\partial \Omega_i} v \cdot n \, ds$, and the associated TPFA method is obtained by requiring the mass-balance equation (9) to be fulfilled for each grid cell $\Omega_i \in \Omega$.

In the literature on finite-volume methods it is common to express the flux F_{ij} in a more compact form than we have done in (12). Terms that do not involve the cell potentials u_i are usually gathered into an interface transmissibility t_{ij}. For the current TPFA method the transmissibilities are defined by:

$$t_{ij} = 2|\gamma_{ij}| \left(\frac{\Delta x_i}{\lambda_{i,ij}} + \frac{\Delta x_j}{\lambda_{j,ij}} \right)^{-1}.$$

Thus by inserting the expression for t_{ij} into (12), we see that the TPFA scheme for equation (10), in compact form, seeks a cell-wise constant function $\mathbf{u} = \{u_i\}$ that satisfies the following system of equations:

$$\sum_j t_{ij} (u_i - u_j) = \int_{\Omega_i} f \, dx, \qquad \forall \Omega_i \subset \Omega. \qquad (13)$$

We have now derived a system of linear equations $\mathbf{Au} = \mathbf{f}$, where the matrix $\mathbf{A} = [a_{ik}]$ is given by

$$a_{ik} = \begin{cases} \sum_j t_{ij} & \text{if } k = i, \\ -t_{ik} & \text{if } k \neq i. \end{cases}$$

This system is symmetric, and a solution is, as for the continuous problem, defined up to an arbitrary constant. The system is made positive definite, and symmetry is preserved, by forcing $u_1 = 0$, for instance. That is, by adding a positive constant to the first diagonal of the matrix. In [2] we present a simple, but yet efficient, MATLAB implementation of the TPFA scheme, which we have used in the following example:

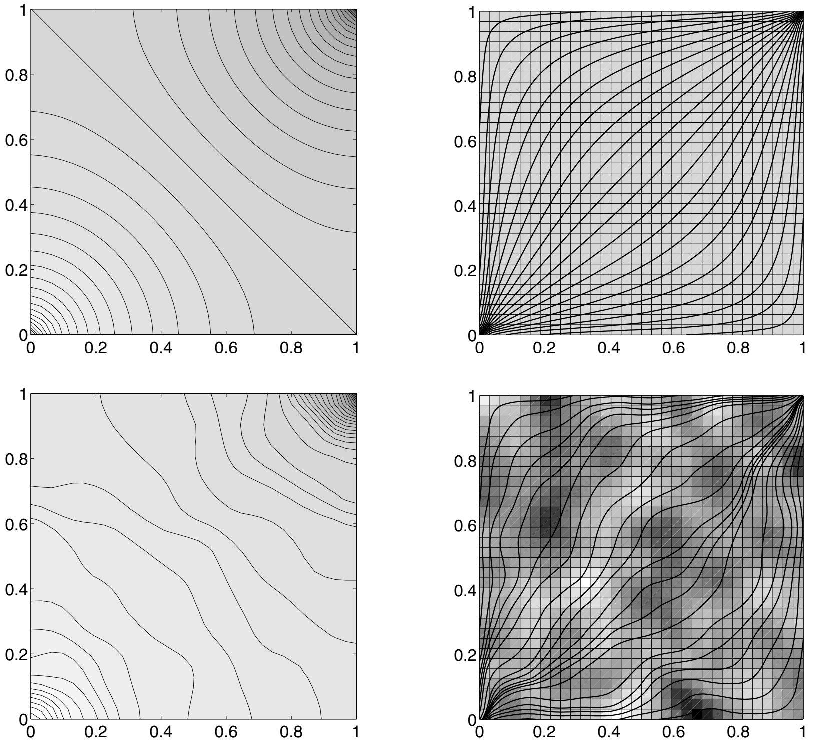


Fig. 4. Pressure contours and streamlines for the classical quarter five-spot test case with a homogeneous and a log-normal permeability field (top and bottom row, respectively).

Example 1. Our first example is the so-called *quarter five-spot* test case, which is the most widespread test case within reservoir simulation. The reservoir is the unit square with an injector at $(0,0)$, a producer at $(1,1)$, and no-flow boundary conditions. Figure 4 shows pressure contours and streamlines for two different isotropic 32×32 permeability fields. The first field is homogeneous, whereas the other is sampled from a log-normal distribution. The pressure and velocity field are symmetric about both diagonals for the homogeneous field. For the heterogeneous field, the flow field is no longer symmetric since the fluids will seek to flow in the most high-permeable regions.

3.2 Multipoint Flux-Approximation (MPFA) Schemes

The TPFA finite-volume scheme presented above is convergent only if each grid cell is a parallelepiped and

$$n_{ij} \cdot K n_{ik} \quad 0, \qquad \forall \Omega_i \subset \Omega, \qquad n_{ij} \;/\; \pm n_{ik}, \tag{14}$$

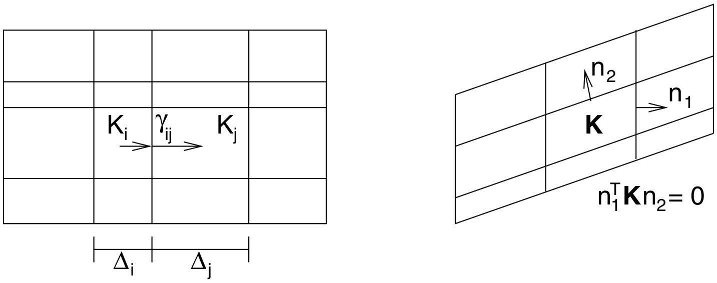


Fig. 5. The grid in the left plot is orthogonal with gridlines aligned with the principal coordinate axes. The grid in the right plot is a K-orthogonal grid.

where n_{ij} and n_{ik} denote normal vectors into two neighbouring grid cells. A grid consisting of parallelepipeds satisfying (14) is said to be K-orthogonal. Orthogonal grids are, for example, K-orthogonal with respect to diagonal permeability tensors, but not with respect to full tensor permeabilities. Figure 5 shows a schematic of an orthogonal grid and a K-orthogonal grid.

If the TPFA method is used to discretise (10) on grids that are not K-orthogonal, the scheme will produce different results depending on the orientation of the grid (so-called grid-orientation effects) and will generally converge to a wrong solution. Despite this shortcoming of the TPFA method, it is still the dominant (and default) method for practical reservoir simulation, owing to its simplicity and computational speed. We now present a class of so-called *multi-point flux-approximation (MPFA) schemes* that aim to amend the shortcomings of the TPFA scheme.

Consider an orthogonal grid and assume that $K \quad K^{\xi,\zeta}{}_{\xi,\zeta \quad x,y,z}$, is a constant tensor with nonzero off-diagonal terms and let γ_{ij} be an interface between two adjacent grid cells in the x–coordinate direction. Then for a given function u, the corresponding flux across γ_{ij} is given by:

$$\int_{\gamma_{ij}} v \cdot n_{ij}\, ds \quad -\int_{\gamma_{ij}} \frac{1}{\mu}\left(K^{x,x}\partial_x u + K^{x,y}\partial_y u + K^{x,z}\partial_z u\right) ds.$$

This expression involves derivatives in three orthogonal coordinate directions. Evidently, two point values can only be used to estimate a derivative in one direction. In particular, the two cell averages u_i and u_j can not be used to estimate the derivative of u in the y and z-directions. Hence, the TPFA scheme neglects the flux contribution from $K^{x,y}\partial_y u$ and $K^{x,z}\partial_z u$.

To obtain consistent interfacial fluxes for grids that are not **K**-orthogonal, one must also estimate partial derivatives in coordinate directions parallel to the interfaces. For this purpose, more than two point values, or cell averages, are needed. This leads to schemes that approximate F_{ij} using multiple cell averages, that is, with a linear expression on the form:

$$F_{ij} \quad \sum_k t_{ij}^k g_{ij}^k(\mathbf{u}).$$

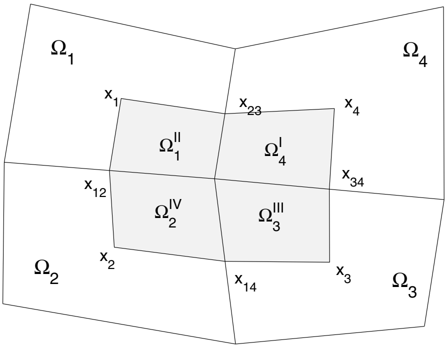


Fig. 6. The shaded region represents the interaction region for the O-method on a two-dimensional quadrilateral grid associated with cells Ω_1, Ω_2, Ω_3, and Ω_4.

Here $\{t_{ij}^k\}_k$ are the transmissibilities associated with γ_{ij} and $\{g_{ij}^k(\mathbf{u})\}_k$ are the corresponding multi-point pressure or flow potential dependencies. Thus, we see that MPFA schemes for (10) can be written on the form:

$$\sum_{j,k} t_{ij}^k g_{ij}^k(\mathbf{u}) \quad \int_{\Omega_i} f\,dx, \qquad \forall \Omega_i \subset \Omega. \tag{15}$$

MPFA schemes can, for instance, be designed by simply estimating each of the partial derivatives $\partial_\xi u$ from neighbouring cell averages. However, most MPFA schemes have a more physical motivation and are derived by imposing certain continuity requirements. We will now outline very briefly one such method, called the O-method [6, 7], for irregular, quadrilateral, matching grids in two spatial dimensions.

The O-method is constructed by defining an interaction region around each corner point in the grid. For a two-dimensional quadrilateral grid, this interaction region is the area bounded by the lines that connect the cell-centres with the midpoints on the cell interfaces, see Fig. 6. Thus, the interaction region consists of four sub-quadrilaterals ($\Omega_1^{\mathrm{II}}, \Omega_2^{\mathrm{IV}}, \Omega_3^{\mathrm{III}}$, and Ω_4^{I}) from four neighbouring cells (Ω_1, Ω_2, Ω_3, and Ω_4) that share a common corner point. For each interaction region, define

$$U_{\mathrm{IR}} \quad \mathrm{span}\{U_i^{\mathrm{J}} : i \quad 1,\ldots,4, \quad \mathrm{J{=}I,\ldots,IV}\},$$

where $\{U_i^{\mathrm{J}}\}$ are linear functions on the respective four sub-quadrilaterals. With this definition, U_{IR} has twelve degrees of freedom. Indeed, note that each U_i^{J} can be expressed in the following non-dimensional form

$$U_i^{\mathrm{J}}(x) \quad u_i + \nabla U_i^{\mathrm{J}} \cdot (x - x_i),$$

where x_i is the cell centre in Ω_i. The cell-centre values u_i thus account for four degrees of freedom and the (constant) gradients ∇U_i^{J} for additional eight.

Next we require that functions in U_{IR} are: (i) continuous at the midpoints of the cell interfaces, and (ii) flux-continuous across the interface segments that lie inside the interaction region. To obtain a globally coupled system, we first use (i) and (ii) to express the gradients ∇U_i^{J}, and hence also the corresponding fluxes across the interface segments of the interaction region, in terms of the unknown cell-centre potentials u_i. This requires solution of a local system of equations. Finally, the cell-centre potentials are determined (up to an arbitrary constant for no-flow boundary conditions) by summing the fluxes across all interface segments of the interaction region and requiring that the mass-balance equations (9) hold. In this process, transmissibilities are assembled to obtain a globally coupled system for the unknown pressures over the whole domain.

We note that this construction leads to an MPFA scheme where the flux across an interface γ_{ij} depends on the potentials u_j in a total of six neighbouring cells (eighteen in three dimensions). Notice also that the transmissibilities $\{t_{ij}^k\}$ that we obtain when eliminating the gradients of the interaction region now account for grid-cell geometries in addition to full-tensor permeabilities.

3.3 A Mixed Finite-Element Method (FEM)

Whereas finite-volume methods treat velocities as functions of the unknown discrete pressures, mixed FEMs [18] obtain the velocity directly. The underlying idea is to consider both the pressure and the velocity as unknowns and express them in terms of basis functions. To this end, we return to the original formulation and describe how to discretise the following system of differential equations with mixed FEMs:

$$v \quad -\lambda(\nabla p - \rho G), \qquad \nabla\cdot v \quad q. \tag{16}$$

As before we impose no-flow boundary conditions on $\partial\Omega$. To derive the mixed formulation, we first define the following Sobolev space

$$H_0^{\mathrm{div}}(\Omega) \quad \{v \in (L^2(\Omega))^d : \nabla\cdot v \in L^2(\Omega) \text{ and } v\cdot n \quad 0 \text{ on } \partial\Omega\}.$$

The mixed formulation of (16) with no-flow boundary conditions now reads: find $(p,v) \in L^2(\Omega)\times H_0^{\mathrm{div}}(\Omega)$ such that

$$\int_\Omega v\cdot\lambda^{-1}u\,dx - \int_\Omega p\,\nabla\cdot u\,dx \quad \int_\Omega \rho G\cdot u\,dx, \tag{17}$$

$$\int_\Omega l\,\nabla\cdot v\,dx \quad \int_\Omega ql\,dx, \tag{18}$$

for all $u \in H_0^{\mathrm{div}}(\Omega)$ and $l \in L^2(\Omega)$. We observe again that, since no-flow boundary conditions are imposed, an extra constraint must be added to make (17)–(18) well-posed. A common choice is to use $\int_\Omega p\,dx \quad 0$.

In mixed FEMs, (17)–(18) are discretised by replacing $L^2(\Omega)$ and $H_0^{\mathrm{div}}(\Omega)$ with finite-dimensional subspaces U and V, respectively. For instance, in the Raviart–Thomas mixed FEM [44] of lowest order (for triangular, tetrahedral, or regular parallelepiped grids), $L^2(\Omega)$ is replaced by

$$U \quad \{p \in L^2(\Omega) : p|_{\Omega_i} \text{ is constant } \forall \Omega_i \in \Omega\}$$

and $H_0^{\text{div}}(\Omega)$ is replaced by

$$\begin{aligned} V \quad & \{v \in H_0^{\text{div}}(\Omega) : v|_{\Omega_i} \text{ has linear components } \forall \Omega_i \in \Omega, \\ & (v \cdot n_{ij})|_{\gamma_{ij}} \text{ is constant } \forall \gamma_{ij} \in \Omega, \text{ and } v \cdot n_{ij} \text{ is continuous across } \gamma_{ij}\}. \end{aligned}$$

Here n_{ij} is the unit normal to γ_{ij} pointing from Ω_i to Ω_j. The corresponding Raviart–Thomas mixed FEM thus seeks

$$(p, v) \in U \times V \text{ such that (17)–(18) hold for all } u \in V \text{ and } q \in U. \tag{19}$$

To express (19) as a linear system, observe first that functions in V are, for admissible grids, spanned by base functions $\{\psi_{ij}\}$ that are defined by

$$\psi_{ij} \in \mathscr{P}_1(\Omega_i)^d \cup \mathscr{P}_1(\Omega_j)^d \quad \text{and} \quad (\psi_{ij} \cdot n_{kl})|_{\gamma_{kl}} \quad \begin{cases} 1, & \text{if} \quad \gamma_{kl} \quad \gamma_{ij}, \\ 0, & \text{else}, \end{cases}$$

where $\mathscr{P}_1(B)$ is the set of linear functions on B. Similarly,

$$U \quad \text{span}\{\chi_m\} \quad \text{where} \quad \chi_m \quad \begin{cases} 1, & \text{if} \quad x \in \Omega_m, \\ 0, & \text{else}. \end{cases}$$

Thus, writing $p \quad \sum_{\Omega_m} p_m \chi_m$ and $v \quad \sum_{\gamma_{ij}} v_{ij} \psi_{ij}$, allows us to write (19) as a linear system in $\mathbf{p} \quad \{p_m\}$ and $\mathbf{v} \quad \{v_{ij}\}$. This system takes the form

$$\begin{bmatrix} \mathbf{B} & -\mathbf{C}^T \\ \mathbf{C} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{v} \\ \mathbf{p} \end{bmatrix} \quad \begin{bmatrix} \mathbf{g} \\ \mathbf{f} \end{bmatrix}. \tag{20}$$

Here $\mathbf{f} \quad f_m$, $\mathbf{g} \quad g_{kl}$, $\mathbf{B} \quad b_{ij,kl}$ and $\mathbf{C} \quad c_{m,kl}$, where:

$$\begin{aligned} g_{kl} \quad & \left[\int_{\Omega} \rho G \cdot \psi_{kl}\, dx\right], & f_m \quad & \left[\int_{\Omega_m} f\, dx\right], \\ b_{ij,kl} \quad & \left[\int_{\Omega} \psi_{ij} \cdot \lambda^{-1} \psi_{kl}\, dx\right], & c_{m,kl} \quad & \left[\int_{\Omega_m} \nabla \cdot \psi_{kl}\, dx\right]. \end{aligned}$$

A drawback with the mixed FEM is that it produces an indefinite linear system. These systems are in general harder to solve than the positive definite systems that arise, e.g., from the TPFA and MPFA schemes described in Sects. 3.1 and 3.2. However, for second-order elliptic equations of the form (4) it is common to use a so-called hybrid formulation. This method leads to a positive definite system where the unknowns correspond to pressures at grid-cell interfaces. The solution to the linear system arising from the mixed FEM can now easily be obtained from the solution to the hybrid system by performing only local algebraic calculations.

3.4 A Mimetic Finite Difference Method (FDM)

The current mimetic FDM [19, 20] is based on the same principles as the above mixed FEM, but the approximation space $V \subset H^{\text{div}}(\Omega)$ is replaced with a space $M \subset$

$L^2(\cup_i \partial\Omega_i)$, and the L^2 inner product on $H^{\text{div}}(\Omega)$ is replaced with an approximative form $m(\cdot,\cdot)$ that acts on $L^2(\cup_i \partial\Omega_i)$. Moreover, whereas functions in V represent velocities, functions in M represent fluxes across grid cell boundaries. Thus, for the current mimetic FDM

$$M \quad \text{span}\{\psi_{ij}\}, \qquad \psi_{ij} \quad \begin{cases} 1, & \text{on } \gamma_{ij}, \\ 0, & \text{on } \gamma_{kl},\ kl\ /\ ij, \end{cases}$$

where one interprets ψ_{ij} to be a basis function that represents a quantity of flow with unit velocity across γ_{ij} in the direction of the unit normal n_{ij}, and zero flow across all other interfaces. Hence, conceptually, the only difference between these basis functions and the Raviart–Thomas basis functions is that we here do not associate a corresponding velocity field in Ω_i and Ω_j.

Next, we present an inner-product $m(u,v)$ on M that *mimics* or "approximates" the L^2 inner-product $(u,\lambda^{-1}v)$ on $H^{\text{div}}(\Omega)$. That is, if $u,v \in H^{\text{div}}(\Omega)$, then we want to derive an inner-product $m(\cdot,\cdot)$ so that

$$(u,\lambda^{-1}v) \approx m(u,v) \quad \sum_k \sum_{i,j} u_{ki} v_{kj} m(\psi_{ki},\psi_{kj}) \quad \sum_k \mathbf{u}_k^t \mathbf{M}_k \mathbf{v}_k, \tag{21}$$

where u_{ki} and v_{ki} are the average velocities across γ_{ki} corresponding to u and v, respectively, and $\mathbf{u}_k \quad u_{ki_\ulcorner i}$, $\mathbf{v}_k \quad v_{ki_\ulcorner i}$. Furthermore, $\mathbf{M}_k$ is defined by

$$\mathbf{M}_k \quad \frac{1}{|\Omega_k|}\mathbf{C}_k \lambda^{-1} \mathbf{C}_k^t + \frac{|\Omega_k|}{2\text{trace}(\lambda)}(\mathbf{I} - \mathbf{Q}_k \mathbf{Q}_k^t), \tag{22}$$

where the matrices $\mathbf{C}_k$, and $\mathbf{Q}_k$ are defined as follows:

$\mathbf{N}_k$: row i is defined by

$$\mathbf{n}_{k,i} \quad \frac{1}{|\gamma_{ki}|}\int_{\gamma_{ki}} (n_{ki})^t \, ds,$$

$\mathbf{C}_k$: row i is defined by

$$\mathbf{c}_{k,i} \quad \int_{\gamma_{ki}} (x - x_k)^t \, ds,$$

where x_k is the mass centre of Ω_k,

$\mathbf{Q}_k$: columns form an orthonormal basis for the column space of $\mathbf{N}_k$.

The discrete system that arises from this mimetic FDM is of the same form as (20). The only difference at the discrete level is that the entries in $\mathbf{B}$ and $\mathbf{g}$ are computed using the $m(\cdot,\cdot)$ inner-product instead of the L^2 inner-product $(u,\lambda^{-1}v)$ on $H^{\text{div}}(\Omega)$. Thus, for the mimetic FDM we have

$$g_{kl} \quad \Big[m(\rho\Xi,\psi_{kl}) \Big], \qquad b_{ij,kl} \quad \Big[m(\psi_{ij},\psi_{kl}) \Big],$$

where $\Xi \quad \sum_{ij} \xi_{ij}\psi_{ij}$ and $\xi_{ij} \quad \frac{1}{|\gamma_{ij}|}\int_{\gamma_{ij}} G \cdot n_{ij} \, ds$.

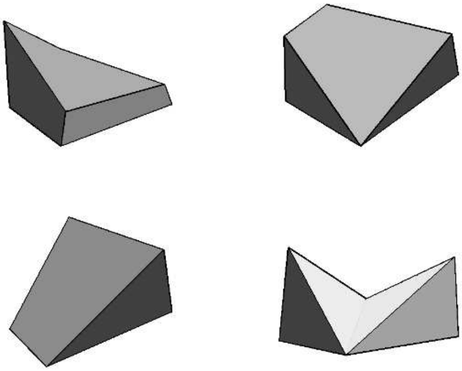


Fig. 7. Examples of deformed and degenerate hexahedral cells arising in corner-point grid models.

3.5 General Remarks

Using geological models as input to flow simulation introduces several numerical difficulties. First of all, typical reservoirs extend several hundred or thousand metres in the lateral direction, but the zones carrying hydrocarbon may be just a few tens of metres in the vertical direction and consist of several layers with different rock properties. Geological models therefore have grid-cells with very high aspect ratios and often the majority of the flow in and out of a cell occurs across the faces with the smallest area. Similarly, the possible presence of strong heterogeneities and anisotropies in the permeability fields typically introduces large conditions numbers in the discretised flow equations. These difficulties are observed even for grid models consisting of regular hexahedral cells.

The flexibility in cell geometry of the industry-standard corner-point format introduces additional difficulties. First of all, since each face of a grid cell is specified by four (arbitrary) points, the cell interfaces in the grid will generally be bilinear surfaces and possibly be strongly curved. Secondly, corner-point cells may have zero volume, which introduces coupling between non-neighbouring cells and gives rise to discretisation matrices with complex sparsity patterns. Moreover, the presence of degenerate cells, in which the corner-points collapse in pairs, means that the cells will generally be polyhedral and possibly contain both triangular and quadrilateral faces (see Fig. 7). Finally, non-conforming grids arise, using the corner-point format, in fault zones where a displacement along a hyperplane has occurred, see Fig. 8. Altogether, this calls for a very flexible discretisation that is not sensitive to the geometry of each cell or the number of faces and corner points.

Having said this, it is appropriate with some brief remarks on the applicability of the methods presented above.

TPFA: Most commercial reservoir simulators use traditional finite-difference methods like the TPFA scheme. These methods were not designed to cope with the

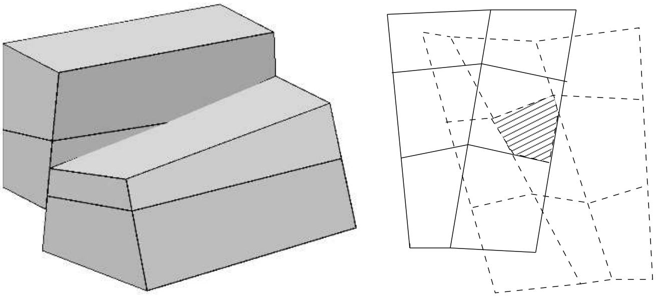


Fig. 8. Two examples of fault surface in a three-dimensional model with non-matching interfaces across the faults. (Left) Three-dimensional view. (Right) Two-dimensional view, where the shaded patch illustrates a "sub-interface" on the fault surface.

type of grid models that are built today using modern geomodelling tools. Hence, if one is interested in *accurate* solutions, two-point schemes should be avoided.

MPFA methods amend shortcomings of two-point scheme, but are unfortunately hard to implement for general grids, especially if the grid is non-conforming with non-matching faces.

Mixed FEMs are more accurate than two-point schemes and generally quite robust. However, the different cells in geological models are generally not diffeomorphic. One therefore needs to introduce a reference element and a corresponding Piola transform for each topological case. This complicates the implementation of a mixed FEM considerably. Moreover, mixed FEMs gives rise to larger linear systems than TPFA and MPFA.

Mimetic FDMs have similar accuracy to MPFA methods and low-order mixed FEMs. But unlike MPFA methods and mixed FEMs, mimetic FDMs are quite easy to formulate and implement for grids with general polyhedral cells. In particular, it is relatively straightforward to handle grids with irregular cell geometries and non-matching faces.

4 Upscaling for Reservoir Simulation

The basic motivation behind upscaling is to create simulation models that produce flow scenarios that are in close correspondence with the flow scenarios that one would obtain by running simulations directly on the geomodels. The literature on upscaling techniques is extensive, ranging from simple averaging techniques, e.g., [37], via local simulation techniques [14, 28], to multiscale methods [1, 8, 9, 22, 33, 34] and homogenisation techniques for periodic structures [15, 32, 36]. It is not within our scope to give a complete overview over the many upscaling techniques that have been applied in reservoir simulation. Instead, we refer the reader to the many review papers that have been devoted to this topic, e.g., [13, 24, 45, 48]. Here we give only a brief introduction to upscaling rock permeability for the pressure equation.

The process of upscaling permeability for the pressure equation (4) or (8) is often termed single-phase upscaling. Most single-phase upscaling techniques seek homogeneous block permeabilities that reproduce the same total flow through each coarse grid-block as one would get if the pressure equation was solved on the underlying fine grid with the correct fine-scale heterogeneous structures. However, designing upscaling techniques that preserve averaged fine-scale flow-rates is in general nontrivial because the heterogeneity at all scales have a significant effect on the large-scale flow pattern. A proper coarse-scale reservoir model must therefore capture the impact of heterogeneous structures at all scales that are not resolved by the coarse grid.

To illustrate the concept behind single-phase upscaling, let p be the solution that we obtain by solving

$$-\nabla\cdot K\nabla p \quad q, \qquad \text{in } \Omega \tag{23}$$

on a fine grid with a suitable numerical method, e.g., a TPFA scheme of the form (13). To reproduce the same total flow through a grid-block V we have to find a homogenised tensor K_V^* such that

$$\int_V K\nabla p\,dx \quad K_V^* \int_V \nabla p\,dx. \tag{24}$$

This equation states that the net flow-rate $\bar{v}$ through V is related to the average pressure gradient $\overline{\nabla p}$ in V through the upscaled Darcy law $\bar{v} \quad -K^*\overline{\nabla p}$.

Note that for a given pressure field p, the upscaled permeability tensor K_V^* is not uniquely defined by (24). Conversely, there does not exist a K_V^* such that (24) holds for any pressure field. This reflects that K_V^* depends on the flow through V. Of course, one does not know a priori what flow scenario V will be subject to. However, the aim is not to replicate a particular flow regime, but to compute coarse-scale permeability tensors that give reasonably accurate results for a wide range of flow scenarios. We now review some of the most commonly used single-phase upscaling methods.

Averaging Methods

The simplest method to upscale permeability is to compute the average of the permeabilities inside the coarse block. To this end, power averaging is a popular technique

$$K_V^{*,p} \quad \left(\frac{1}{|V|}\int_V K(x)^p\,dx\right)^{1/p}, \quad -1 \le p \le 1.$$

Special cases include the arithmetic average ($p \quad 1$), the harmonic average ($p \quad -1$), and the geometric average ($p \to 0$).

The use of power averaging can be motivated by the so-called Wiener-bounds [49], which state that for a statistically homogeneous medium, the correct upscaled permeability will be bounded above and below by the arithmetic and harmonic mean, respectively. This result has a more intuitive explanation. To see this, consider the one-dimensional pressure equation:

$$-\partial_x(K(x)p'(x)) \quad 0 \quad \text{in } (0,1), \quad p(0) \quad p_0,\ p(1) \quad p_1.$$

Integrating once, we see that the corresponding Darcy velocity is constant. This implies that $p'(x)$ must scale proportional to the inverse of $K(x)$. Hence, we derive

$$p'(x) \quad \frac{p_1-p_0}{K(x)}\left[\int_0^1 \frac{dx}{K(x)}\right]^{-1} \quad \frac{p_1-p_0}{K(x)}K_V^{*,-1}.$$

If we insert this expression into (24) we find that the correct upscaled permeability K_V^* is identical to the harmonic mean $K_V^{*,-1}$.

The same argument applies to the special case of a perfectly stratified isotropic medium; for instance, with layers perpendicular to the x–axis so that $K(x,\cdot,\cdot)$ is constant for each x. Now, consider a uniform flow in the x–direction:

$$\begin{aligned} -\nabla\cdot K\nabla p \quad & 0 \quad \text{in } V \quad (0,1)^3, \\ p(0,y,z) \quad & p_0, \quad p(1,y,z) \quad p_1, \\ (-K\nabla p)\cdot n \quad & 0 \quad \text{for } y,z\in\{0,1\}, \end{aligned} \tag{25}$$

where n is the outward unit normal on ∂V. This means that for each pair $(y,z)\in(0,1)^2$ the one-dimensional function $p_{y,z} \quad p(\cdot,y,z)$ satisfies

$$-\partial_x\big(Kp'_{y,z}(x)\big) \quad 0 \text{ in } (0,1), \quad p_{y,z}(0) \quad p_0,\ p_{y,z}(1) \quad p_1,$$

from which it follows that

$$-K(x)\nabla p \quad -(K(x)p'_{y,z}(x),0,0)^T \quad -K_V^{*,-1}(p_1-p_0,0,0)^T.$$

Hence, the correct upscaled permeability is equal to the harmonic mean.

Exercise 2. Show that if K instead models a stratified isotropic medium with layers perpendicular to the y or z–axis, then the correct upscaled permeability for uniform flow in the x–direction would be equal to the arithmetic mean.

The discussion above shows that averaging techniques can be appropriate in special cases. However, if we consider the model problem (25) with a less idealised heterogeneous structures, or with the same heterogeneous structures but with other boundary conditions, then both the arithmetic and harmonic average will generally give wrong net flow-rates. Indeed, these averages give correct upscaled permeability only for cases with essentially one-dimensional flow. To try to model flow in more than one direction, one could generate a diagonal permeability tensor with the following diagonal components:

$$K^{x,x} \quad \mu_a^z(\mu_a^y(\mu_h^x)), \quad K^{y,y} \quad \mu_a^z(\mu_a^x(\mu_h^y)), \quad K^{z,z} \quad \mu_a^x(\mu_a^y(\mu_h^z)).$$

Here μ_a^ξ and μ_h^ξ represent the arithmetic and harmonic means, respectively, in the ξ-coordinate direction. Thus, in this method one starts by taking a harmonic average along grid cells that are aligned in one coordinate-direction. One then computes

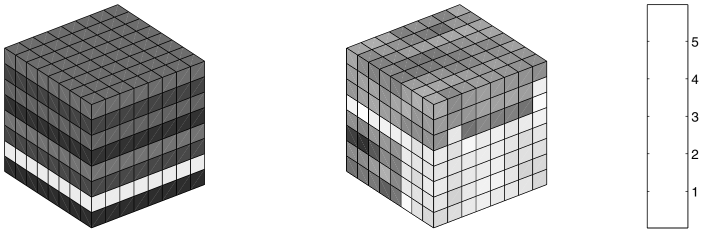


Fig. 9. Logarithm of permeability: the left cube is a layered medium, whereas the right cube is extracted from the lower part of the fluvial Upper Ness formation from Model 2 of the 10th SPE Comparative Solution Project [25].

the corresponding diagonal by taking the arithmetic mean of all "one dimensional" harmonic means. This average is sometimes called the harmonic-arithmetic average and may give good results if, for instance, the reservoir is layered and the primary direction of flow is along the layers.

Despite the fact that averaging techniques can give correct upscaling in special cases, they tend to perform poorly in practice since the averages do not reflect the structure or orientation of the heterogeneous structures. It is also difficult to decide which averaging technique to use since the best average depends both on the heterogeneity of the media and on the flow process we want to model (flow direction, boundary conditions, etc). To illustrate the dependence on the flow process we consider an example.

Example 2 (from [2]). Consider a reservoir in the unit cube $0,1^{3}$ with two different geomodels that each consist of a $8 \times 8 \times 8$ uniform grid blocks and permeability distribution as depicted in Fig. 9. We consider three different upscaling methods: harmonic average (H), arithmetic average (A), and harmonic-arithmetic average (HA). The geomodels are upscaled to a single grid-block, which is then subjected to three different boundary conditions:

BC1: $p\quad 1$ at $(x,y,0)$, $p\quad 0$ at $(x,y,1)$, no-flow elsewhere.
BC2: $p\quad 1$ at $(0,0,z)$, $p\quad 0$ at $(1,1,z)$, no-flow elsewhere.
BC3: $p\quad 1$ at $(0,0,0)$, $p\quad 0$ at $(1,1,1)$, no-flow elsewhere.

Table 1 compares the observed coarse-block rates with the flow-rate obtained by direct simulation on the $8 \times 8 \times 8$ grid. For the layered model, harmonic and harmonic-arithmetic averaging correctly reproduce the vertical flow normal to the layers for BC1. Arithmetic and harmonic-arithmetic averaging correctly reproduce the flow along the layers for BC2. Harmonic-arithmetic averaging also performs well for corner-to-corner flow (BC3). For model two, however, all methods produce significant errors, and none of the methods are able to produce an accurate flow-rate for boundary conditions BC1 and BC3.

Table 1. Flow-rates relative to the reference rate Q_R on the fine grid.

	Model 1			Model 2		
	BC1	BC2	BC3	BC1	BC2	BC3
Q_H/Q_R	1	2.31e−04	5.52e−02	1.10e−02	3.82e−06	9.94e−04
Q_A/Q_R	4.33e+03	1	2.39e+02	2.33e+04	8.22	2.13e+03
Q_{HA}/Q_R	1	1	1.14	8.14e−02	1.00	1.55e−01

Flow-Based Upscaling

A popular class of methods are so-called flow-based upscaling methods as first suggested by Begg et al. [14]. In this approach one solves a set of homogeneous pressure equations on the form

$$-\nabla \cdot K\nabla p \quad 0 \quad \text{in } V,$$

for each grid block V with prescribed boundary conditions that induce a desired flow pattern. Each member of this class of methods differ in the way boundary conditions are prescribed.

A simple and popular choice is to impose a pressure drop in one of the coordinate directions and no-flow conditions along the other faces, as in (25) for flow in the x–direction. This gives us a set of three flow-rates for each grid block that can be used to compute an effective diagonal permeability tensor with components

$$K^{x,x} \quad -Q_x L_x/\Delta P_x, \qquad K^{y,y} \quad -Q_y L_y/\Delta P_y, \qquad K^{z,z} \quad -Q_z L_z/\Delta P_z.$$

Here Q_ξ, L_ξ and ΔP_ξ are the net flow, the length between opposite sides, and the pressure drop in the ξ-direction inside V, respectively.

Another popular option is to choose periodic boundary conditions. That is, one assumes that each grid block is a unit cell in a periodic medium and imposes full correspondence between the pressures and velocities at opposite sides of the block; that is, to compute $K^{x,x}$, $K^{x,y}$, and $K^{x,z}$ we impose the following boundary conditions:

$$\begin{aligned} p(1,y,z) \quad p(0,y,z) - \Delta p, \quad p(x,1,z) \quad p(x,0,z), \qquad p(x,y,1) \quad p(x,y,0),\\ v(1,y,z) \quad v(0,y,z), \quad v(x,1,z) \quad v(x,0,z), \qquad v(x,y,1) \quad v(x,y,0), \end{aligned}$$

and define $K^{x,\xi} \quad -Q_\xi L_\xi/\Delta p$. This approach yields a symmetric and positive definite tensor [28], and is usually more robust than the directional flow boundary conditions.

Example 3 (from [2]). We revisit the test-cases considered in Example 2, but now we compare harmonic-arithmetic averaging (HA) with the flow-based techniques using directional (D) and periodic (P) boundary conditions. The latter method gives rise to full permeability tensors, but for the cases considered here the off-diagonal terms in the upscaled permeability tensors are small, and are therefore neglected for simplicity.

Table 2. Flow-rates relative to the reference rate Q_R on the fine grid.

	Model 1			Model 2		
	BC1	BC2	BC3	BC1	BC2	BC3
Q_{HA}/Q_R	1	1	1.143	0.081	1.003	0.155
Q_D/Q_R	1	1	1.143	1	1.375	1.893
Q_P/Q_R	1	1	1.143	0.986	1.321	1.867

Table 2 compares the observed coarse-block rates with the flow-rate obtained by direct simulation on the $8 \times 8 \times 8$ grid. For the layered model, all methods give the same diagonal permeability tensor, and hence give exactly the same results. For Model 2 we see that the numerical pressure computation methods give significantly better results than the harmonic-arithmetic average. Indeed, the worst results for the pressure computation method, which were obtained for corner-to-corner flow, is within a factor two, whereas the harmonic-arithmetic average underestimates the flow rates for BC1 and BC3 by almost an order of magnitude.

It should be noted that in the discrete case, the appropriate upscaling technique depends on the underlying numerical method. For instance, if the pressure equation is discretised by a TPFA scheme of the form (13), then grid-block permeabilities are used only to compute interface transmissibilities at the coarse scale. Upscaling methods for this method may therefore instead be targeted at computing coarse-scale transmissibilities (that reproduce a fine-scale flow field in an averaged sense) directly. Procedures for computing coarse-scale transmissibilities similar to the averaging and numerical pressure computation techniques have been proposed in [38] and, e.g., [31], respectively.

5 Multiscale Methods the Pressure Equation

Subsurface flow problems represent an important application that calls for a more mathematically rigorous treatment of the way the large span of permeability values and correlation lengths impact the solution. Conventional methods are inadequate for this problem because the heterogeneity in natural porous media does not have clearly separated scales of variation, and because permeability variations occurring at small length scales (e.g., smaller scale than the grid resolution) may have strong impact on the flow at much larger scales. This makes subsurface flow problems a natural target for a new class of methods called multiscale methods – methods that attempt to model physical phenomena on coarse grids while honouring small-scale features that impact the coarse grid solution in an appropriate way, e.g., by incorporating subgrid information into numerical schemes for partial differential equations in a way that is consistent with the local property of the differential operator.

A large number of multiscale methods have appeared in the literature on computational science and engineering. Among these, there are a variety of methods (e.g., [1, 8, 9, 22, 33, 34]) that target solving elliptic equations of the same form as the pressure equation for incompressible subsurface flow. Upscaling methods that derive coarse-grid properties from numerical subgrid calculations may also in a certain sense be viewed as multiscale methods, but the way the upscaled properties are incorporated into the coarse-scale systems is not necessarily consistent with the properties of the differential operator.

In this section we present three selected multiscale methods. The main idea is to show how multiscale methods are built, and how subgrid information is embedded into the coarse-scale system. For presentational brevity and enhanced readability we consider only elliptic (incompressible flow) equations, and disregard capillary forces so that $\nabla p_j \quad \nabla p$ for all phases j.

Let Ω denote our reservoir. Furthermore, let $\mathscr{B} \quad \{B_i\}$ be a partitioning of Ω into polyhedral grid-blocks and let $\{\Gamma_{ij} \quad \partial B_i \cap \partial B_j\}$ be the corresponding set of non-degenerate interfaces. Throughout we implicitly assume that all grid-blocks B_i are divided into smaller grid cells that form a sub-partitioning of Ω. Without compressibility and capillary forces, the pressure equation for the three-phase black-oil model now reads:

$$v \quad -K(\lambda \nabla p - \lambda_G G), \qquad \nabla \cdot v \quad q \qquad \text{in } \Omega. \tag{26}$$

where we have inserted $v \quad \sum_j v_j$, $\lambda \quad \sum_j \frac{k_{rj}}{\mu_j}$, and $\lambda_G \quad \sum_j \rho_j \frac{k_{rj}}{\mu_j}$ for brevity. We assume that no-flow boundary conditions $v \cdot n \quad 0$ are imposed on $\partial\Omega$, and that p is uniquely determined by adding the constraint $\int_\Omega p\,dx \quad 0$.

5.1 The Multiscale Finite-Element Method (MsFEM) in 1D

Before we introduce multiscale methods for solving (26) in three-dimensional domains, we start with an instrumental example in one spatial dimension. To this end, we consider the following elliptic problem:

$$\partial_x \big(K(x)p'(x)\big) \quad f, \text{ in } \Omega \quad (0,1), \qquad p(0) \quad p(1) \quad 0, \tag{27}$$

where $f, K \in L^2(\Omega)$ and K is bounded above and below by positive constants.

The MsFEM was first introduced by Hou and Wu [33], but the basic idea goes back to earlier work by Babuška and Osborn [12] for 1D problems and Babuška, Caloz, and Osborn [11] for special 2D problems. The method is, like standard FEMs, based on a variational formulation. In the variational formulation of (27) we seek $p \in H_0^1(\Omega)$ such that

$$a(p,v) \quad (f,v) \quad \text{for all } v \in H_0^1(\Omega), \tag{28}$$

where $(\cdot,\cdot)$ is the L^2 inner-product and

$$a(p,v) \quad \int_\Omega K(x)u'(x)v'(x)\,dx.$$

Now, let $\mathscr{N}_{\mathscr{B}} = \{0 = x_0 < x_1 < \ldots < x_{n-1} < x_n = 1\}$ be a set of nodal points and define $B_i = (x_{i-1}, x_i)$. For each x_i, $i = 1, \ldots, n-1$ we associate a corresponding basis function $\phi^i \in H_0^1(\Omega)$ defined by

$$a(\phi^i, v) = 0 \quad \text{for all } v \in H_0^1(B_i \cup B_{i+1}), \qquad \phi_i(x_j) = \delta_{ij}, \tag{29}$$

where δ_{ij} is the Kronecker delta. The multiscale finite-element method seeks the unique function p_0 in

$$V^{\mathrm{ms}} = \mathrm{span}\{\phi_i\} = \{u \in H_0^1(\Omega) : a(u,v) = 0 \text{ for all } v \in H_0^1(\cup_i B_i)\} \tag{30}$$

satisfying

$$a(p_0, v) = (f, v) \quad \text{for all } v \in V^{\mathrm{ms}}. \tag{31}$$

We now show that the solution p of (28) can be written as a sum of p_0 and a family of solutions to independent local subgrid problems. To this end, we first show that $p_0 = p_I$, where p_I is the unique function in V^{ms} with $p_I(x) = p(x)$, $x \in \mathscr{N}_{\mathscr{B}}$. Indeed, since $p - p_I$ vanishes on $\mathscr{N}_{\mathscr{B}}$, we have $p - p_I \in H_0^1(\cup_i B_i)$. Hence, it follows from (28) and the mutual orthogonality of V^{ms} and $H_0^1(\cup_i B_i)$ with respect to $a(\cdot,\cdot)$ that

$$a(p_I, v) = a(p, v) = (f, v) \quad \text{for all } v \in V^{\mathrm{ms}}.$$

Thus, in particular, by (31) and choosing $v = p_I - p_0$ we obtain

$$a(p_I - p_0, p_I - p_0) = 0,$$

which implies $p_0 = p_I$. Thus, $p = p_0 + \sum_{i>0} p_i$ where $p_i \in H_0^1(B_i)$ is defined by

$$a(p_i, v) = (f, v) \quad \text{for all } v \in H_0^1(B_i).$$

Hence, as promised, the solution of (28) is a sum of p_0 and solutions to independent local subgrid problems. This result can also be seen directly by noting that p_0 is, by definition, the orthogonal projection onto V^{ms} with respect to the inner-product $a(\cdot,\cdot)$ and noting that $H_0^1(\Omega) = V^{\mathrm{ms}} \oplus H_0^1(\cup_i B_i)$.

Exercise 3. Show that

$$a(\phi_i, \phi_j) = \begin{cases} K_i^{*,-1}/(x_i - x_{i-1}) + K_{i+1}^{*,-1}/(x_{i+1} - x_i), & \text{if } i = j, \\ -K_{\max(i,j)}^{*,-1}/|x_i - x_j|, & \text{if } |i-j| = 1, \\ 0, & \text{if } |i-j| > 1, \end{cases} \tag{32}$$

where $K_i^{*,-1}$ is the harmonic mean of K over the interval $[x_{i-1}, x_i]$, i.e.,

$$K_i^{*,-1} = \frac{x_i - x_{i-1}}{\int_{x_{i-1}}^{x_i} K(x)^{-1}\, dx}.$$

Consider next the standard nodal basis functions used in the linear FEM. Here the basis functions ϕ_i are linear on each interval and satisfy $\phi_i(x_j) = \delta_{ij}$. Show that the corresponding coefficients for this method is obtained by replacing the harmonic means in (32) with the associated arithmetic means.

The multiscale finite-element method can also be extended to higher dimensions, but does not give locally mass-conservative velocity fields. Next we present a multiscale finite-volume method that is essentially a control-volume finite-element version of the MsFEM. Control-volume finite-element methods seek solutions in designated finite-element approximation spaces (on a dual-grid), but rather than formulating the global problem in a variational framework, they employ a finite-volume formulation (on a primal grid) that gives mass-conservative velocity fields.

5.2 The Multiscale Finite-Volume Method (MsFVM)

The multiscale finite-volume method [34] employs numerical subgrid calculations (analogous to those in [33]) to derive a multi-point stencil for solving (26) on a coarse grid. The method then proceeds and reconstructs a mass-conservative velocity field on a fine grid as a superposition of local subgrid solutions, where the weights are obtained from the coarse-grid solution.

The derivation of the coarse-scale equations in the MsFVM is essentially an upscaling procedure for generating coarse-scale transmissibilities. The first step is to solve a set of homogeneous boundary-value problems of the form

$$-\nabla\cdot K\lambda\nabla\phi_i^k \quad 0, \quad \text{in } R, \qquad \phi_i^k \quad v_i^k, \quad \text{on } \partial R, \tag{33}$$

where R are so-called interaction regions as illustrated in Fig. 10 and v_i^k are boundary conditions to be specified below. Subscript i in ϕ_i^k denotes a corner-point in the coarse grid (x_i in the figure) and the superscript k runs over all corner points of the interaction region (x^k in the figure). Thus, for each interaction region associated with e.g., a hexahedral grid in three dimensions we have to solve a total of eight local boundary-value problems of the form (33). The idea behind the MsFVM is to express the global pressure as a superposition of these local pressure solutions ϕ_i^k. Thus, inside each interaction region R one assumes that the pressure is a superposition of the local subgrid solutions $\{\phi_i^k\}$, where k ranges over all corner-points in the interaction region (i.e., over the cell-centres of the coarse-grid blocks).

First, we define the boundary conditions v_i^k in (33). These are defined by solving a reduced-dimensional flow problem on each face F of the interaction region

$$-\nabla\cdot K\lambda\nabla v_i^k \quad 0 \quad \text{in } F, \tag{34}$$

with boundary conditions given by $v_i^k(x^l) \quad \delta_{kl}$ at the corner points of the interaction region. (In 3D, the corner-point values are first extended to the edges of F by linear interpolation). Once v_i^k are computed, the local pressure solutions ϕ_i^k can be computed from (33).

The next step is to identify basis functions for the multiscale method. To this end, we observe that the cell centers x^k constitute a corner point for four interaction regions in 2D and for eight interaction regions in 3D (for a regular hexahedral grid). Moreover, for all corner-points x_i of the coarse grid, the corresponding boundary

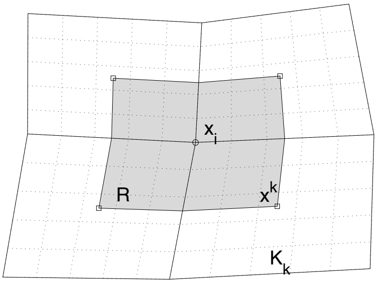


Fig. 10. The shaded region represents the interaction region R for the MsFVM, where x_i denotes corner-points and x^k the midpoints of the coarse grid-blocks. The midpoints x^k are the corner-points of the interaction region.

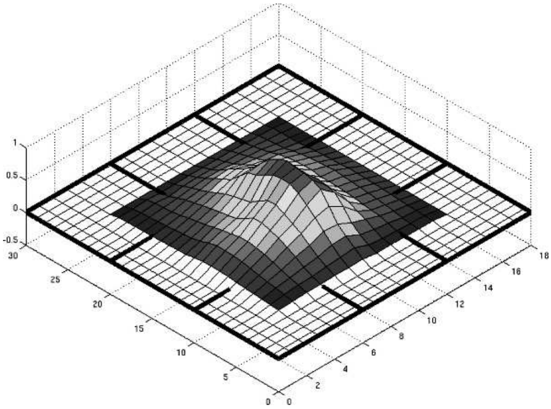


Fig. 11. Pressure basis function ϕ^k for the MsFVM in two-dimensional space.

conditions v_i^k for the different pressure equations coincide on the respective faces of the interaction regions that share the corner point x^k. This implies that the basis function

$$\phi^k \quad \sum_i \phi_i^k \tag{35}$$

is continuous (in a discrete sense), see Fig. 11. In the following construction, the base functions defined in (35) will serve as building blocks that are used to construct a global "continuous" pressure solution.

Thus, define now the approximation space $U^{\mathrm{ms}} \quad \mathrm{span}\{\phi^k\}$ and observe that all basis functions vanish at all but one of the grid block centres x^k. This implies that, given a set of pressure values $\{p^k\}$, there exists a unique extension $\{p^k\} \longrightarrow p \in U^{\mathrm{ms}}$ with $p(x^k) \quad p^k$. This extension is defined by

$$p \quad \sum_k p^k \phi^k \quad \sum_{i,k} p^k \phi_i^k . \tag{36}$$

A multi-point stencil can now be defined by assembling the flux contribution across the grid-block boundaries from each basis function. Thus, let

$$f_{k,l} \quad -\int_{\partial B_l} n \cdot K\lambda \nabla \phi^k \, ds$$

be the local flux out of grid-block B_l induced by ϕ^k. The MsFVM for solving (26) then seeks constant grid-block pressures $\{p^k\}$ satisfying

$$\sum_k p^k f_{k,l} \quad \int_{B_l} \big(q - \nabla \cdot K\lambda_G G\big)\, dx \qquad \forall l.$$

To reconstruct a mass-conservative velocity field on a fine scale, notice first that the expansion (36) produces a mass-conservative velocity field on the coarse grid. Unfortunately, this velocity field will not preserve mass across the boundaries of the interaction regions. Thus, to obtain a velocity field that is also mass conservative on the fine grid we will use the subgrid fluxes obtained from p as boundary conditions for solving a local flow problem inside each coarse block B_l to *reconstruct* a fine-scale velocity v_l. That is, solve

$$v_l \quad -K(\lambda \nabla p_l - \lambda_G G), \quad \nabla \cdot v_l \quad \frac{1}{|B_l|} \int_{B_l} q \, dx \quad \text{in } B_l, \tag{37}$$

with boundary conditions obtained from (36), i.e.,

$$v_l \quad -K\lambda \nabla p \quad \text{on } \partial B_l, \tag{38}$$

where p is the expanded pressure defined by (36). If these subgrid problems are solved with a conservative scheme, then the global velocity field $v \quad \sum_{B_l} v_l$ will be mass conservative. Note, however, that since the subgrid problems (37)–(38) are solved independently we loose continuity of the global pressure solution, which is now defined by $p \quad \sum_{B_l} p_l$.

Remark 1. The present form of the MsFVM, which was developed by Jenny et al. [34], does not model sources at the subgrid scale. Indeed, the source term in (37) is equally distributed within the grid-block. Thus, to use the induced velocity field to simulate the phase transport one has to treat the wells as a uniform source within the entire well block. However, a more detailed representation of flow around wells can be obtained by replacing (37) by

$$v_l \quad -K(\lambda \nabla p_l - \lambda_G G), \qquad \nabla \cdot v_l \quad q \quad \text{in } B_l \tag{39}$$

in grid blocks containing a well, i.e., for all B_l in which q is nonzero.

5.3 A Multiscale Mixed Finite-Element Method (MsMFEM)

Recall that mixed finite-element discretisations of elliptic equations on the form (26) seek a solution (p, v) to the mixed equations

$$\int_{\Omega} u\cdot(K\lambda)^{-1}v\,dx-\int_{\Omega} p\,\nabla\cdot u\,dx \quad \int_{\Omega} \lambda_G G\cdot u\,dx, \tag{40}$$

$$\int_{\Omega} l\,\nabla\cdot v\,dx \quad \int_{\Omega} ql\,dx, \tag{41}$$

in a finite-dimensional product space $U\times V\subset L^2(\Omega)\times H_0^{1,\mathrm{div}}(\Omega)$. If the subspaces $U\subset L^2(\Omega)$ and $V\subset H_0^{1,\mathrm{div}}(\Omega)$ are properly balanced (see, e.g., [16, 17, 18]), then p and v are defined (up to an additive constant for p) by requiring that (40)–(41) holds for all $(l,u)\in U\times V$.

In MsMFEMs one constructs a special approximation space for the velocity v that reflects the important subgrid information. For instance, instead of seeking velocities in a simple approximation space spanned by basis functions with polynomial components, one computes special multiscale basis functions Ψ in a manner analogous to the MsFVM, and defines a corresponding multiscale approximation space by $V^{\mathrm{ms}}\quad \mathrm{span}\{\Psi\}$. The pressure approximation space consists simply of piecewise constant functions on the coarse grid, i.e.,

$$U\quad \{p\in L^2(\Omega): p|_B \text{ is constant for all } B\in\mathscr{B}\}.$$

Hence, in the MsMFEM we seek

$$p\in U,\ v\in V^{\mathrm{ms}} \quad \text{such that (40)–(41) holds for} \quad \forall l\in U,\ \forall u\in V^{\mathrm{ms}}. \tag{42}$$

The MsMFEM thus resolves subgrid-scales locally through the construction of special multiscale basis functions, whereas the large scales are resolved by solving the discretised equations on a coarse-grid level.

An approximation space for the pressure p that reflects subgrid structures can be defined in a similar manner. However, whereas velocity fields for flow in porous media may fluctuate rapidly, the pressure is usually relatively smooth. It is therefore often sufficient to model pressure with low resolution as long as it does not significantly degrade the accuracy of the velocity solution. Thus, because the MsMFEM treats the pressure and velocities as separate decoupled variables, it is natural to use a high-resolution space for velocity and a low-resolution space for pressure. In other words, the computational effort can be spent where it is most needed. Moreover, the approximation spaces can not be chosen arbitrarily. Indeed, the convergence theory for mixed finite element methods, the so-called *Ladyshenskaja–Babuška–Brezzi* theory (see [16, 17, 18]) states that the approximation spaces must satisfy a relation called the *inf-sup* condition, or the LBB (*Ladyshenskaja–Babuška–Brezzi*) condition. Using a multiscale approximation space, also for the pressure variable, can cause the LBB condition to be violated.

Exercise 4. Show that if the velocity solution v of (17)–(18) is contained in V^{ms}, then the velocity solution of (42) coincides with v.

Approximation Space for the Darcy Velocity

Consider a coarse grid that overlays a fine (sub)grid, for instance as illustrated in Fig. 12. For the velocity we associate one vector of basis functions with each non-

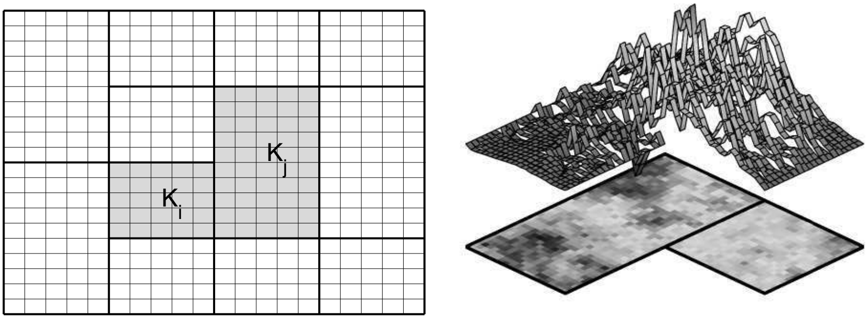


Fig. 12. Left: Schematic of the coarse and fine grid for the MsMFEM. The shaded region denotes the support of the velocity basis function associated with the edge between the two grid-blocks B_i and B_j. Right: x-component of a MsMFEM basis function associated with an interface between two rectangular (two-dimensional) grid-blocks.

degenerate interface Γ_{ij} between two neighbouring grid-blocks B_i and B_j. To be precise, for each interface Γ_{ij} we define a basis function Ψ_{ij} by

$$\Psi_{ij} = -K\nabla\phi_{ij}, \qquad \text{in } B_i \cup B_j, \tag{43}$$

where ϕ_{ij} is determined by

$$(\nabla\cdot\Psi_{ij})|_{B_i} = \ell(x)/\int_{B_i}\ell(x)\,dx, \tag{44}$$

$$(\nabla\cdot\Psi_{ij})|_{B_j} = -\ell(x)/\int_{B_j}\ell(x)\,dx. \tag{45}$$

with no-flow boundary conditions along the edges $\partial B_i \cup \partial B_j \backslash \Gamma_{ij}$.

The function ℓ in (44)–(45) is a positive function that can be defined in various ways. Chen and Hou [22] simply used $\ell(x) = 1$, which produces mass-conservative velocity fields at the coarse-scale level and on the fine scale for all blocks where the source term q is zero. For blocks with nonzero source term q, the fine-scale velocity is not conservative unless q is treated as a constant within each grid block (analogous to the way sources are modelled in the original MsFVM [34]). In reservoir simulation, however, this way of treating sources is inadequate. Indeed, here the source term q represents wells that are point- or line-sources, and modelling flow correctly in the near-well region is considered to be very important. However, since this issue is linked specifically to the reservoir simulation application, we will discuss how ℓ can be defined to handle wells along with other implementational issues in Sect. 6.

To obtain a mass-conservative velocity field on a subgrid scale we need to solve the subgrid problems (43)–(45) with a mass conservative scheme. Fig. 12 displays the x-component of a velocity basis function for the case with $\ell(x) = 1$ computed using the lowest order Raviart–Thomas mixed FEM. We clearly see strong fluctuations in the velocity that reflect the fine-scale heterogeneity. Note also that the basis functions Ψ_{ij} are defined to be time-independent. This implies that the computation of the multiscale basis functions may be made part of a preprocessing step, also for flows with large variations in the total mobility λ. In other words, a single set of basis functions may be used throughout the entire simulation. The reason why it is not necessary to include the total mobility in (43) is that mobility variations within a single block are usually small relative to the jumps in the permeability. Therefore, by including only K we account for the dominant part of the fine-grid variability in the coefficients $K\lambda$. The coarse grid variability of the total mobility is taken into account by reassembling the coarse grid system at each time step.

Remark 2. For the MsFVM one can also use a single set of basis functions throughout entire simulations. However, to account for coarse-grid variability of the total mobility one needs to update the upscaled MsFVM transmissibilities, e.g., by multiplying the initial transmissibilities with a factor that reflects the change in total mobility. This implies that one can not escape from solving the local subproblems (37) or (39) in order to obtain a mass conservative velocity field on the fine grid. This feature generally makes the MsFVM more computationally expensive for multi-phase flows than the MsMFEM.

5.4 Numerical Examples

Both MsMFEM and MsFVM solve a coarse-scale equation globally while trying to resolve fine-scale variations by using special multiscale basis functions. Next, we demonstrate that the accuracy of the generated velocity solutions is not very sensitive to the dimension of the coarse grid.

Example 4 (from [3]). Consider a horizontal, two-dimensional reservoir with 60 × 220 grid cells with permeability from the bottom layer of Model 2 in the 10th SPE Comparative Solution Project [25]. We inject water in the centre of the domain and produce oil and water at each of the four corners. The pressure equation is solved using the MsFVM and the MsMFEM with various coarse-grid dimensions. For comparison, we also compute two reference solutions using the TPFA scheme, one on the original 60 × 220 grid, and one on a grid that is refined four times in each direction. Employing the corresponding velocity fields, we solve an equation modelling transport of an incompressible fluid using an upstream finite-volume method on the underlying fine grid.

Fig. 13 shows the resulting saturation fields when the total volume of the water that has been injected is equal to 30% of the total accessible pore volume. We observe that all saturation plots are quite similar to the saturation plots obtained using the reference velocity fields. We therefore also quantify the errors in the respective saturation fields by

$$\delta(S) \quad \frac{\varepsilon(S)}{\varepsilon(S_{\text{ref}})}, \qquad \varepsilon(S) \quad \frac{\|S - \mathscr{I}(S_{\text{ref}}^{4\times})\|_{L^1}}{\|\mathscr{I}(S_{\text{ref}}^{4\times})\|_{L^1}},$$

where $\mathscr{I}$ is an operator that maps the saturation solution on the refined 240 × 880 grid onto the original 60 × 220 grid. The results displayed in Table 3 show that there

Table 3. Relative saturation error $\delta(S)$ for a five-spot simulation in Layer 85 of Model 2 of the 10th SPE Comparative Solution Project for various coarse grids.

	30×110	15×55	10×44	5×11
MsMFEM	1.0916	1.2957	1.6415	1.9177
MsFVM	1.0287	1.6176	2.4224	3.0583

Table 4. Runtimes for Model 2 of the 10th SPE Comparative Solution Project using a streamline simulator with TPFA or MsMFEM pressure solver measured on a workstation PC with a 2.4 GHz Intel Core 2 Duo processor with 4 Mb cache and 3 Gb memory.

	Pressure	Streamline	Total
TPFA	465 sec	51 sec	516 sec
MsMFEM	91 sec	51 sec	142 sec

is some degradation of solution quality when the grid is coarsened, but the errors are not very sensitive to coarse-grid size.

When the pressure equation (26) needs to be solved once, the multiscale methods described above can only offer limited speed-up relative to the time spent on solving the full problem on the fine grid using state-of-the-art linear solvers, e.g., algebraic multigrid methods [47]. However, for two-phase flow simulations, where the pressure equation needs to be solved repeatedly, it has been demonstrated that the basis functions need to be computed only once, or updated infrequently [1, 35, 39]. This means that the main computational task is related to solving the global coarse-grid system, which is significantly less expensive than solving the full fine-grid system. This is illustrated by the following example.

Example 5 (from [40]). Consider now the full SPE 10 model, which consists of $60 \times 220 \times 85$ uniform cells. The top 35 layers are from a smooth Tarbert formation, whereas the bottom 50 layers are from a fluvial Upper Ness formation, see Fig. 1. The reservoir is produced using a five-spot pattern of vertical wells with an injector in the middle; see [25] for more details.

To simulate the production process we use a streamline simulator with two different pressure solvers: (i) TPFA with an algebraic multigrid linear solver [47], and (ii) MsMFEM on a $5 \times 11 \times 17$ coarse grid. Streamline solvers are known to be very efficient compared to conventional (finite-difference) reservoir simulators, for which computing the full 3D SPE10 model is out of bounds using a single processor and takes several hours on a parallel processor. The key to the high efficiency of streamline solvers is underlying operator splitting used to separate the solution of pressure/velocity from the solution of the fluid transport, which here is solved along 1D streamlines (i.e., in Lagrangian coordinates) and mapped back to the Eulerian grid used to compute pressure and velocities.

Table 4 reports runtimes for two simulations of 2 000 days of production for the whole model. In both runs the simulator used 5 000 streamlines and 25 times

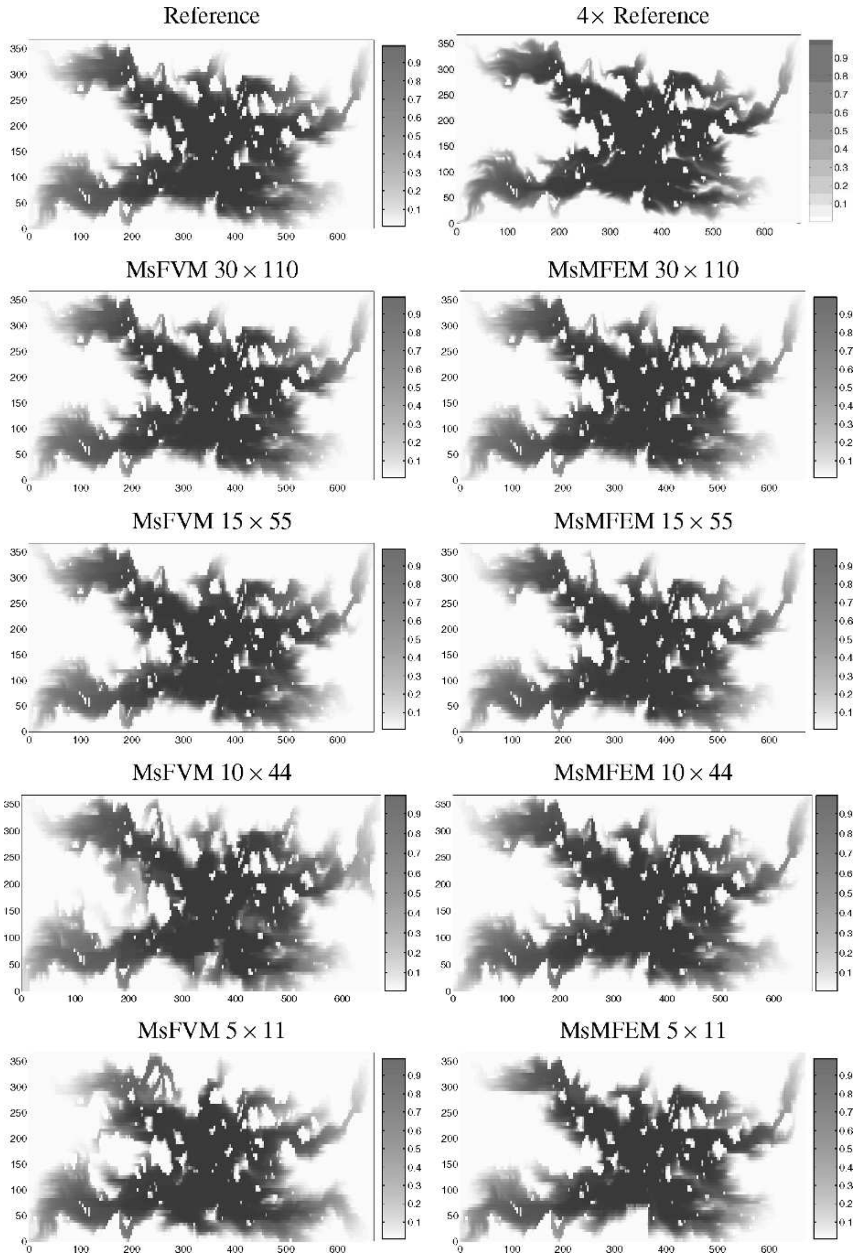


Fig. 13. Saturation solutions computed using velocity fields obtained with MsMFEM and MsFVM on various coarse grids (c–j), TPFA on the original fine grid (a), and TPFA on the grid that is refined four times in each direction (b).

steps. The time spent on the transport step includes tracing of streamlines, solving 1D transport equations, and mapping solutions back and forth between the pressure and the streamline grid. The time spent in the multiscale pressure solver includes initial computation of basis functions and assembly and solution of coarse-grid system for each time step. Using the MsMFEM pressure solver gives a speedup of 5.1 for the pressure solution and 3.6 for the overall computation. Moreover, with a total runtime of 2 minutes and 22 seconds, simulating a million-cell reservoir model has become an (almost) interactive task using the the multiscale–streamline solver.

Remark 3. Note that the basis function can be computed independently, which means that the computation of basis functions is a so-called embarrassingly parallel task. Even further speedup should therefore be expected for parallel implementations, using e.g., the multi-core processors that are becoming available in modern PCs.

6 Implementational Issues for MsMFEM

In this section we discuss some of the implementational issues that need to be addressed when implementing the MsMFEM. We start by discussing what considerations one should take into account when generating the coarse grid. Next we explain how the coarse-grid system can be assembled efficiently, and the implications that this has on the choice of numerical method used for computing the multiscale velocity basis functions. We then discuss the role of the function ℓ in the definition of the basis functions, and how it impacts the MsMFEM solution. Finally, we describe briefly how to build global information into the basis functions to more accurately resolve flow near large-scale heterogeneous structures that have a strong impact on the flow regime.

6.1 Generation of Coarse Grids

It has been demonstrated in [4, 5] that MsMFEM is *very* flexible with respect to the geometry and topology of the coarse grid. A bit simplified, the grid flexibility can be stated as follows: given an appropriate solver for the local flow problems on a particular type of fine grids, the MsMFEM can be formulated on any coarse grid where each grid block consists of an arbitrary collection of connected fine-grid cells. To illustrate, consider a small model where Ω is defined as the union of the three blocks depicted in Fig. 14. Although these blocks are stacked on top of each other, each pair of blocks has a common interface. Thus, in the multiscale formulation we construct three basis functions for this set of blocks, one for each pair depicted in Fig. 14.

Extensive tests, some of which are reported in [4, 5], show that the accuracy of the MsMFEM is generally not very sensitive to the shape of the blocks. In fact, accurate results are obtained for grids containing blocks with rather ’exotic’ shapes, see e.g., [4, 5]. In the next three examples we will show some examples of coarse grids to substantiate this claim. The reader is referred to [4, 5] for a more thorough discussion of the numerical accuracy obtained using this kind of coarse grids.

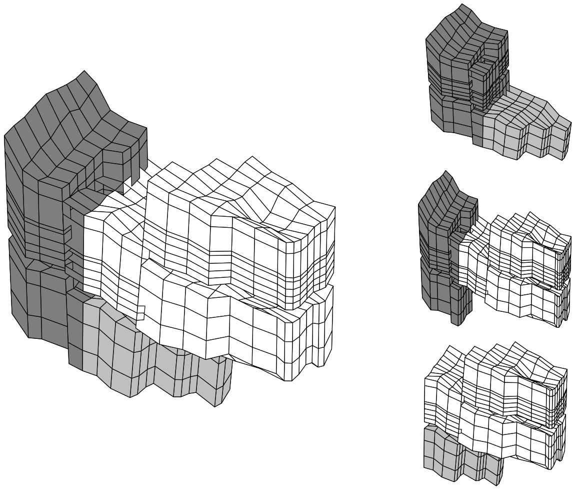


Fig. 14. A three-block domain and the corresponding subdomains constituting the support of the resulting MsMFEM basis functions.

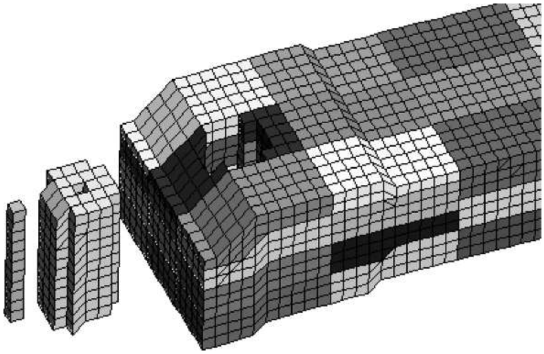


Fig. 15. A coarse grid defined on top of a structured corner-point fine grid. The cells in the coarse grid are given by different colours.

Example 6 (Near-well grid). Figure 15 shows a vertical well penetrating a structured corner-point grid with eroded layers. On the coarse grid, the well is confined to a single cell consisting of all cells in the fine grid penetrated by the well. Moreover, notice the single neighbouring block shaped like a 'cylinder' with a hole.

Example 7 (Barriers). Figure 16 shows a subsection of the SPE10 model, in which we have inserted a few flow barriers with very low permeability. In [4] it was shown that MsMFEM becomes inaccurate if coarse grid-cells are cut into two (or more) non-communicating parts by a flow barrier. Fortunately, this can be automatically detected when generating basis functions, and the resolution can be improved by using some form of grid refinement. The figure shows two different approaches: (i)

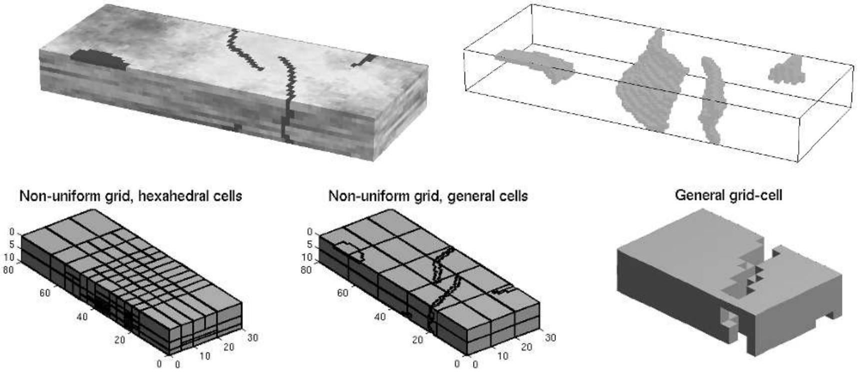

Non-uniform grid, hexahedral cells
Non-uniform grid, general cells
General grid-cell
0
5
10
80
60
40
20
0
0
10
20
30

Fig. 16. The upper row shows the permeability field (right), and the interior barriers (left). The lower row shows a hierarchically refined grid (left), the barrier grid (middle), and a coarse grid-block in the barrier grid (right).

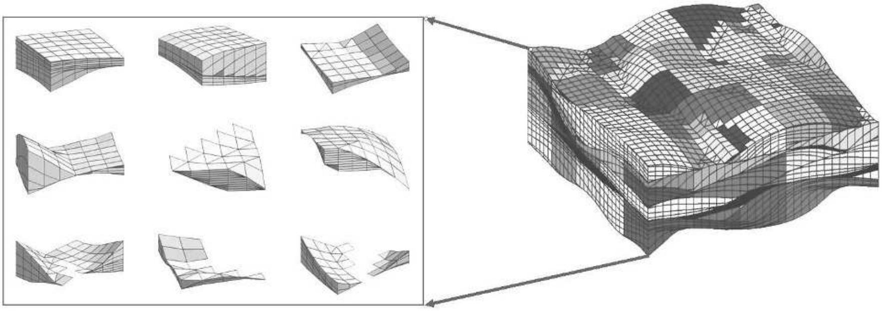


Fig. 17. Uniform partitioning in index space of a corner-point model containing a large number of eroded layers.

structured, hierarchical refinement, and (ii) direct incorporation of the flow barriers as extra coarse grid-blocks intersecting a uniform 3 × 5 × 2 grid. This results in rather exotic coarse cells, e.g., as shown in the figure, where the original rectangular cell consisting of 10 × 16 × 5 fine cells is almost split in two by the barrier, and the resulting coarse cell is only connected through a single cell in the fine grid. Although the number of grid cells in the barrier grid is five times less than for the hierarchically refined grid, the errors in the production curves are comparable, indicating that MsMFEM is robust with respect to the shape of the coarse cells.

Example 8 (Eroded layers). Figure 17 shows a uniform partitioning in index space of a corner-point grid modelling a wavy depositional bed on a meter-scale. The corner-point grid is described by vertical pillars that form a uniform 30 × 30 in the horizontal plane and 100 very thin layers, out of which many collapse to a hyper-plane in some regions. The figure also shows the shape in physical space of some of the coarse

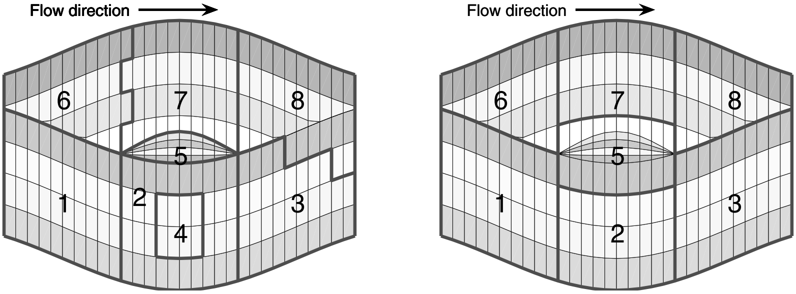


Fig. 18. Illustration of some of the guidelines for choosing a good coarse grid. In the left plot, all blocks except for Block 1 violate at least one of the guidelines each. In the right plot, the blocks have been improved at the expense of more couplings in the coarse-grid system.

blocks resulting from the uniform partitioning in index space. All blocks are used directly in the simulation, except for the block in the lower-right corner, which has two disconnected parts and thus can be split in two automatically.

The complex coarse blocks arising from the straightforward partitioning in index space will in fact give more accurate results than what is obtained from more sophisticated up-gridding schemes trying e.g., to make each cell be as close to a regular hexahedral box as possible. The reason is that the flow will follow the layered structure of the medium and therefore is resolved most accurately by coarse grids that reflect the layering.

The fact that MsMFEM is rather insensitive to the number and the shape of the blocks in the coarse grid means that the process of generating a coarse simulation grid from a complex geological model can be greatly simplified, especially when the fine grid is fully unstructured or has geometrical complications due to faults, throws, and eroded cells; e.g., as seen in Figs. 3 and 8. However, MsMFEM does have some limitations, as identified in [4]. Here it was observed that barriers (low-permeable obstacles) may cause inaccurate results unless the coarse grid adapts to the barrier structures. In addition it was demonstrated that MsMFEM in its present form has limited ability to model bidirectional flow across coarse-grid interfaces; fine-grid fluxes at coarse-grid interfaces in the reconstructed flow field will usually go in the same direction.

As a remedy for the limitations identified in [4], it is possible to exploit global information (e.g., from an initial fine-scale pressure solve) when constructing the basis functions [1], see also Sect. 6.4. However, our experience indicates that accurate results are also obtained if the coarse grid obeys certain guidelines; see the left plot in Fig. 18 for illustrations:

1. The coarse grid should preferably minimise the occurrence of bidirectional flow across coarse-grid interfaces. Examples of grid structures that increase the likelihood for bidirectional flow are:
   - Coarse-grid faces with (highly) irregular shapes, like the ’saw-tooth’ faces between Blocks 6 and 7 and Blocks 3 and 8.

  - Blocks that do not contain source terms and have only one neighbour, like Block 4. (A simple remedy for this is to split the interface into at least two sub-faces, and define a basis function for each sub-face.)
  - Blocks having interfaces only along and not transverse to the major flow directions, like Block 5. (To represent flow in a certain direction, there must be at least one non-tangential face that defines a basis function in the given flow direction.)
2. Blocks and faces in the coarse grid should follow geological layers whenever possible. This is not fulfilled for Blocks 3 and 8.
3. Blocks in the coarse-grid should adapt to flow obstacles (shale barriers, etc.) whenever possible; see [4].
4. For parabolic (compressible flow) problems, e.g., three-phase black-oil models, one should model point-sources (and line-sources) at the subgrid level. For instance, for reservoir simulation one should assign a separate grid block to each cell in the original grid with an open well perforation[1].

In addition, to enhance the efficiency of the method, one should try to keep the number of connections between coarse-grid blocks as low as possible to minimise the bandwidth of the coarse-scale system, and avoid having too many small blocks as this increases the dimension of the coarse-scale system, but does not necessarily improve accuracy significantly.

In the right plot of Fig. 18, we have used the guidelines above to improve the coarse grid from the left plot. In particular, we joined Blocks 2 and 4 and have have increased the size of Block 5 to homogenise the block volumes and introduce basis functions in the major flow direction for this block. In doing so, we increase the number of couplings from nine to twelve (by removing the coupling between Blocks 2 and 4 and introducing extra coupling among Blocks 1, 3, 5, 6, and 8). In general it may be difficult to obtain an 'optimal' coarse grid, since guidelines may be in conflict with each other. On the other hand, this is seldom necessary, since the MsMFEM is relatively robust with respect to the choice of coarse grid.

6.2 Computing Basis Functions and Assembling the Linear System

In principle, any conservative numerical method may be used to construct the basis functions, e.g., any of the four methods discussed in Sect. 2.1. However, computing the entries in the coarse-grid linear system requires evaluating the following inner-products between the multiscale basis functions:

[1] For reservoir simulation there is also another reason, apart from compressibility, to why it is preferable to assign separate blocks to each cell with an open well perforation. Indeed, the source q in reservoir simulation models is generally not known a priori, but determined by so-called well-models that relate the well-rates to the pressure in the associated well-block. To compute the rates "correctly" one needs to get the pressure in the well-block correct. The MsMFEM provides a pressure value for each coarse grid-block. Thus, by assigning a block to each cell with an open well perforation, we extract values that represent the actual pressure in these cells. In other words, the pressure at the wells is modelled with subgrid resolution.

$$\int_\Omega \Psi_{ij} \cdot (K\lambda)^{-1} \Psi_{kl}\, dx. \tag{46}$$

Alternatively, one can use an approximate inner product like the one used in the mimetic formulation discussed in Sect. 3.4.

If a finite-volume method is used, a computational routine for computing these inner-products, either exactly or approximately, is generally not available. Thus, to implement the MsMFEM one needs to add an extra feature in the numerical implementation. When a mixed FEM or mimetic FDM is used, on the other hand, a routine for calculating the inner-product (46) is part of the implementation of the subgrid solver. In fact, in this case the integral (46) can be expressed as a vector-matrix-vector product.

Let $\mathbf{R}$ be the matrix formed with columns $\mathbf{r}_{ij}$ holding the coefficients r_{kl}^{ij} in the following expansion:

$$\Psi_{ij} \quad \sum_{\gamma_{kl}} r_{kl}^{ij} \psi_{kl}.$$

Furthermore, let $\mathbf{B}$ be the $\mathbf{B}$-matrix in a system of the form (20) that stems from a Raviart–Thomas mixed FEM or a mimetic FDM on a fine grid. Then

$$\int_\Omega \Psi_{ij} \cdot (K\lambda)^{-1} \Psi_{kl}\, dx \quad \mathbf{r}_{ij}^t \mathbf{B} \mathbf{r}_{ij}. \tag{47}$$

Thus, the coarse-grid system for the MsMFEM may be expressed as follows:

$$\mathbf{B}^{\mathrm{ms}} \quad \mathbf{R}^t\mathbf{B}\mathbf{R}, \qquad \mathbf{g}^{\mathrm{ms}} \quad \mathbf{R}^t\mathbf{g}.$$

The right hand side $\mathbf{q}^{\mathrm{ms}}$ in the multiscale system is formed by integrating q over each grid block, and the matrix $\mathbf{C}^{\mathrm{ms}} \quad c_{m,kl}$ is given by

$$c_{m,kl} \quad \int_{B_m} \nabla \cdot \Psi_{kl}\, dx \quad \begin{cases} 1, & \text{if } k \quad m, \\ -1, & \text{if } l \quad m, \\ 0, & \text{otherwise.} \end{cases}$$

6.3 Role of the Weighting Function

The weighting function ℓ in (44)–(45) has been defined in different ways

- $\ell \quad 1$ in [22];
- $\ell \quad q$ if $\int_{B_m} q \ / \ 0$ and $\ell \quad 1$ elsewhere in [1]; and
- $\ell \quad q$ if $\int_{B_m} q \ / \ 0$ and $\ell \quad \mathrm{trace}(K)$ elsewhere in [4, 5].

To understand how these definitions have come into play, recall first that the MsMFEM velocity solution is a linear superposition of the velocity basis functions. Hence,

$$\begin{aligned} (\nabla \cdot v)|_{B_i} \quad & \sum_j v_{ij} \nabla \cdot \Psi_{ij} \quad \frac{\ell}{\int_{B_i} \ell\, dx} \sum_j v_{ij} \\ & \frac{\ell}{\int_{B_i} \ell\, dx} \int_{\partial B_i} v \cdot n\, ds \quad \frac{\ell}{\int_{B_i} \ell\, dx} \int_{B_i} \nabla \cdot v\, dx. \end{aligned}$$

One can therefore say that the primary role of ℓ is to distribute the divergence of the velocity field onto the fine grid in an appropriate way.

For incompressible flow problems $\mathrm{div}(v)$ is non-zero only in blocks with a source. For blocks where $\int_{B_i} q \;/\; 0$, the choice $\ell \quad q$ stems from the fact that it gives mass conservative velocity fields on the subgrid. For blocks without a source (where the velocity is divergence free) ℓ can be chosen nearly arbitrarily. The idea of letting the weight function scale with the trace of the mobility was introduced in [4] as a way of avoiding unnaturally large amount of flow through low-permeable zones and in particular through flow barriers. In general, however, using $\ell \quad 1$ gives (almost) equally accurate results.

For compressible flow (e.g., (8)) we may no longer choose ℓ arbitrarily. For instance, defining base functions using $\ell \quad q$ would concentrate all compressibility effects where q is nonzero. To avoid this, one has to separate the contribution to the divergence field stemming from sources and from compressibility. This can be achieved, as we have proposed in Sect. 6.1, by assigning one "coarse" grid block to each cell in the fine grid with a source or sink. By doing so, we may, in principle, choose $\ell \quad 1$ everywhere. But, for the three-phase black-oil model (cf. Sect. 2.2), we have

$$\nabla \cdot v \quad q - c_t \frac{\partial p}{dt} - \sum_j c_j v_j \cdot \nabla p_l. \tag{48}$$

Hence, ℓ should ideally be proportional to the right hand side of (48). Although the right hand side of (48) can be estimated from local computations, we do not propose using this strategy to define ℓ. Indeed, the multiscale concept is not to try to replicate fine-scale solutions by trying to account for all subgrid information. The important thing is to account for the subgrid effects that strongly influence flow on the coarse-grid level, and subgrid variability in the velocity divergence field is generally not among these effects.

Our own numerical experience so far indicates that good accuracy is obtained by taking ℓ to be the porosity ϕ. To motivate this choice, we note that c_t is proportional to ϕ when the saturations are smooth. Moreover, using $\ell \quad \phi$ is in accordance with the idea behind using $\ell \quad \mathrm{trace}(\lambda)$. Indeed, regions with very low permeability also tend to have low porosity, so by choosing $\ell \quad \phi$ one should (to some extent) avoid forcing too much flow through low-permeable barriers, [4]. Using $\ell \quad \mathrm{trace}(K)$, on the other hand, will generally give velocity solutions for which $\mathrm{div}(v)$ oscillates too much, i.e., is underestimated in low-permeable regions and overestimated in high-permeable regions.

6.4 Incorporating Global Information

All multiscale methods essentially attempt to decouple the global problem into a coarse-grid system and a set of independent local problems. In Sect. 5.1 it was shown that in the one-dimensional case there is an exact splitting. That is, the global solution (of the variational formulation) can be expressed as the sum of the MsFEM solution and solutions of independent local problems. In higher dimensions, however, decoupling the system into a low-dimensional coarse-grid system and independent local subproblems is not possible in general. But it is possible to invoke global information, e.g., from a single-phase flow solution computed at initial time, to specify better boundary conditions for the local flow problems and thereby improve the multiscale solutions, as was shown in [1] for MsMFEM and in [29] for MsFVM.

For many problems, invoking global information may have little effect, and will, for multi-phase flow problems, only give an incremental improvement in accuracy. But for certain problems, such as for models with large scale near-impermeable shale barriers that force the flow to take a detour around the barrier, invoking global information can improve accuracy quite significantly, and should be viewed as an alternative to grid refinement.

Since MsMFEM allows running entire simulations with a single set of basis functions, solving the pressure equation once on a fine grid in order to improve the accuracy of the multiscale solution is easily justified. To this end, one needs to split each of the subgrid problems (43)–(45) into two independent problems in B_i and B_j, respectively, with a common Neumann boundary condition on the interface Γ_{ij}. In particular, if v is the initial fine-scale velocity solution, the following boundary condition should be imposed on Γ_{ij}:

$$\Psi_{ij} \cdot n_{ij} \qquad \frac{v \cdot n_{ij}}{\int_{\Gamma_{ij}} v \cdot n_{ij}\, ds}. \tag{49}$$

The method that stems from defining the multiscale basis functions with this formulation is usually referred to as the global, as opposed to local, MsMFEM.

Exercise 5. Assign one grid block to each cell with a source and let $\ell \quad 1$. Alternatively let $\ell \quad q$ if $\int_{B_i} q \ / \ 0$ and $\ell \quad 1$ elsewhere. Show that if the multiscale basis functions are defined by (43)–(45) and (49), then $v \in \operatorname{span}\{\Psi_{ij}\}$.

References

1. J. E. Aarnes. On the use of a mixed multiscale finite element method for greater flexibility and increased speed or improved accuracy in reservoir simulation. *Multiscale Model. Simul.*, 2(3):421–439, 2004.
2. J. E. Aarnes, T. Gimse, and K.-A. Lie. An introduction to the numerics of flow in porous media using MATLAB. In G. Hasle, K.-A. Lie, and E. Quak, editors, *Geometrical Modeling, Numerical Simulation, and Optimization: Industrial Mathematics at SINTEF*, pages 265–306. Springer Verlag, 2007.
3. J. E. Aarnes, V. Kippe, K.-A. Lie, and A. Rustad. Modelling of multiscale structures in flow simulations for petroleum reservoirs. In G. Hasle, K.-A. Lie, and E. Quak, editors, *Geometrical Modeling, Numerical Simulation, and Optimization: Industrial Mathematics at SINTEF*, pages 307–360. Springer Verlag, 2007.
4. J. E. Aarnes, S. Krogstad, and K.-A. Lie. A hierarchical multiscale method for two-phase flow based upon mixed finite elements and nonuniform coarse grids. *Multiscale Model. Simul.*, 5(2):337–363, 2006.
5. J. E. Aarnes, S. Krogstad, and K.-A. Lie. Multiscale mixed/mimetic methods on corner-point grids. *Comput. Geosci*, 12(3):297–315, 2008.

6. I. Aavatsmark, T. Barkve, Ø. Bøe, and T. Mannseth. Discretization on unstructured grids for inhomogeneous, anisotropic media. Part I: Derivation of the methods. *SIAM J. Sci. Comp.*, 19(5):1700–1716, 1998.
7. I. Aavatsmark, T. Barkve, Ø. Bøe, and T. Mannseth. Discretization on unstructured grids for inhomogeneous, anisotropic media. Part II: Discussion and numerical results. *SIAM J. Sci. Comp.*, 19(5):1717–1736, 1998.
8. T. Arbogast. Numerical subgrid upscaling of two-phase flow in porous media. In Z. Chen, R. E. Ewing, and Z.-C. Shi, editors, *Numerical Treatment of Multiphase Flows in Porous Media (Beijing, 1999)*, Lecture Notes in Phys., pages 35–49. Springer-Verlag, Berlin, 2000.
9. T. Arbogast. Analysis of a two-scale, locally conservative subgrid upscaling for elliptic problems. *SIAM J. Numer. Anal.*, 42(2):576–598, 2004.
10. K. Aziz and A. Settari. *Petroleum reservoir simulation*. Elsevier, London and New York, 1979.
11. I. Babuška, G. Caloz, and E. Osborn. Special finite element methods for a class of second order elliptic problems with rough coefficients. *SIAM J. Numer. Anal.*, 31:945–981, 1994.
12. I. Babuška and E. Osborn. Generalized finite element methods: Their performance and their relation to mixed methods. *SIAM J. Numer. Anal.*, 20:510–536, 1983.
13. J. W. Barker and S. Thibeau. A critical review of the use of pseudorelative permeabilities for upscaling. *SPE Reservoir Eng.*, 12:138–143, 1997.
14. S. H. Begg, R. R. Carter, and P. Dranfield. Assigning effective values to simulator grid-block parameters for heterogeneous reservoirs. *SPE Reservoir Eng.*, pages 455–463, 1989.
15. A. Benesoussan, J.-L. Lions, and G. Papanicolaou. *Asymptotic Analysis for Periodic Structures*. Elsevier Science Publishers, Amsterdam, 1978.
16. D. Braess. *Finite elements: Theory fast solvers and applications in solid mechanics*. Cambridge University Press, Cambridge, 1997.
17. S. C. Brenner and L. R. Scott. *The mathematical theory of finite element methods*. Springer–Verlag, New York, 1994.
18. F. Brezzi and M. Fortin. *Mixed and hybrid finite element methods*. Computational Mathematics. Springer–Verlag, New York, 1991.
19. F. Brezzi, K. Lipnikov, M. Shashkov, and V. Simoncini. A new discretization methodology for diffusion problems on generalized polyhedral meshes. *Comput. Methods Appl. Mech. Engrg.*, 196(37-40):3682–3692, 2007.
20. F. Brezzi, K. Lipnikov, and V. Simoncini. A family of mimetic finite difference methods on polygonial and polyhedral meshes. *Math. Models Methods Appl. Sci.*, 15:1533–1553, 2005.
21. G. Chavent and J. Jaffre. *Mathematical models and finite elements for reservoir simulation*. North Holland, 1982.
22. Z. Chen and T. Y. Hou. A mixed multiscale finite element method for elliptic problems with oscillating coefficients. *Math. Comp.*, 72:541–576, 2003.
23. Z. Chen, G. Huan, and Y. Ma. *Computational methods for multiphase flows in porous media*. Computational Science & Engineering. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2006.
24. M. A. Christie. Upscaling for reservoir simulation. *J. Pet. Tech.*, 48:1004–1010, 1996.
25. M. A. Christie and M. J. Blunt. Tenth SPE comparative solution project: A comparison of upscaling techniques. *SPE Reserv. Eval. Eng*, 4(4):308–317, 2001. url: `http://www.spe.org/csp`.

26. L. P. Dake. *Fundamentals of reservoir engineering*. Elsevier, Amsterdam, 1978.
27. A. H. Demond and P. V. Roberts. An examination of relative permeability relations for two-phase flow in porous media. *Water Res. Bull.*, 23:617–628, 1987.
28. L. J. Durlofsky. Numerical calculations of equivalent gridblock permeability tensors for heterogeneous porous media. *Water Resour. Res.*, 27(5):699–708, 1991.
29. Y. Efendiev, V. Ginting, T. Hou, and R. Ewing. Accurate multiscale finite element methods for two-phase flow simulations. *J. Comput. Phys.*, 220(1):155–174, 2006.
30. R. E. Ewing. *The mathematics of reservoir simulation*. SIAM, 1983.
31. L. Holden and B. Nielsen. Global upscaling of permeability in heterogeneous reservoirs; the output least squares (ols) method. *Transp. Porous Media*, 40(2):115–143, 2000.
32. U. Hornung. *Homogenization and porous media*. Springer Verlag, New York, 1997.
33. T. Y. Hou and X.-H. Wu. A multiscale finite element method for elliptic problems in composite materials and porous media. *J. Comput. Phys.*, 134:169–189, 1997.
34. P. Jenny, S. H. Lee, and H. A. Tchelepi. Multi-scale finite-volume method for elliptic problems in subsurface flow simulation. *J. Comput. Phys.*, 187:47–67, 2003.
35. P. Jenny, S. H. Lee, and H. A. Tchelepi. Adaptive multiscale finite-volume method for multiphase flow and transport in porous media. *Multiscale Model. Simul.*, 3(1):50–64, 2004/05.
36. V. V. Jikov, S. M. Kozlov, and O. A. Oleinik. *Homogenization of differential operators and integral functionals*. Springer–Verlag, New York, 1994.
37. A. G. Journel, C. V. Deutsch, and A. J. Desbarats. Power averaging for block effective permeability. In *SPE California Regional Meeting*, Oakland, California, 2-4 April 1986. SPE 15128.
38. M. J. King, D. G. MacDonald, S. P. Todd, and H. Leung. Application of novel upscaling approaches to the Magnus and Andrew reservoirs. In *SPE European Petroleum Conference*, The Hague, Netherlands, 20-22 October 1998. SPE 50463.
39. V. Kippe, J. E. Aarnes, and K.-A. Lie. A comparison of multiscale methods for elliptic problems in porous media flow. *Comput. Geosci*, 12(3):377–398, 2008.
40. V. Kippe, H. Hægland, and K.-A. Lie. A method to improve the mass-balance in streamline methods. In *SPE Reservoir Simulation Symposium*, Houston, Texas U.S.A., February 26-28 2007. SPE 106250.
41. L. Lake. *Enhanced oil recovery*. Prentice Hall, Inglewood Cliffs, NJ, 1989.
42. B. B. Maini and T. Okazawa. Effects of temperature on heavy oil-water relative permeability. *J. Can. Petr. Tech*, 26:33–41, 1987.
43. D. W. Peaceman. *Fundamentals of numerical reservoir simulation*. Elsevier, Amsterdam, 1977.
44. P. A. Raviart and J. M. Thomas. A mixed finite element method for second order elliptic equations. In I. Galligani and E. Magenes, editors, *Mathematical Aspects of Finite Element Methods*, pages 292–315. Springer–Verlag, Berlin – Heidelberg – New York, 1977.
45. P. Renard and G. de Marsily. Calculating equivalent permeability. *Adv. Water Resour.*, 20:253–278, 1997.
46. D. P. Schrag. Preparing to capture carbon. *Science*, 315:812–813, 2007.
47. K. Stüben. *Multigrid*, chapter Algebraic Multigrid (AMG): An Introduction with Applications. Academic Press, 2000.
48. X.-H. Wen and J. J. Gómez-Hernández. Upscaling hydraulic conductivities in heterogeneous media: An overview. *J. Hydrol.*, 183:ix–xxxii, 1996.
49. O. Wiener. *Abhandlungen der Matematisch.* PhD thesis, Physischen Klasse der Königlichen Sächsischen Gesellscaft der Wissenschaften, 1912.

Multiscale Modelling of Complex Fluids: A Mathematical Initiation

Claude Le Bris[1,2] and Tony Lelièvre[1,2]

[1] CERMICS, École Nationale des Ponts et Chaussées, 6 & 8, avenue Blaise Pascal, F-77455 Marne-La-Vallée Cedex 2, France,
lelievre@cermics.enpc.fr, lebris@cermics.enpc.fr

[2] INRIA Rocquencourt, MICMAC project-team, Domaine de Voluceau, B.P. 105, F-78153 Le Chesnay Cedex, France,

Summary. We present a general introduction to the multiscale modelling and simulation of complex fluids. The perspective is mathematical. The level is elementary. For illustration purposes, we choose the context of incompressible flows of infinitely dilute solutions of flexible polymers, only briefly mentioning some other types of complex fluids. We describe the modelling steps, compare the multiscale approach and the purely macroscopic, more traditional, approach. We also introduce the reader with the appropriate mathematical and numerical tools. A complete set of codes for the numerical simulation is provided, in the simple situation of a Couette flow. This serves as a test-bed for the numerical strategies described in a more general context throughout the text. A dedicated section of our article addresses the mathematical challenges on the front of research.

Keywords: non-Newtonian flows, complex fluids, polymer flow, multiscale modelling, Couette flow, Hookean and FENE dumbbell models, Oldroyd-B model, Fokker-Planck equation, stochastic differential equation.

1 Introduction

This article presents a general introduction to the multiscale modelling and simulation of complex fluids. The perspective is mathematical. The level is elementary. For illustration purposes, we choose the context of incompressible flows of infinitely dilute solutions of flexible polymers. This category of fluids is that for which the mathematical understanding is the most comprehensive one to date. It is therefore an adequate prototypical context for explaining the recently developed multiscale approach for the modelling of complex fluids, and more precisely for that of fluids with microstructures. Other types of complex fluids, also with microstructures, such as liquid crystals, suspensions, blood, may also be modeled by such types of models. However the modelling is either less understood mathematically, or more intricate and technical to describe (or both). The former case is therefore more appropriate for an *initiation*.

We describe the modelling steps, compare the multiscale approach and the purely macroscopic, more traditional, approach. We also introduce the reader to the appropriate mathematical and numerical tools.

The readership we wish to reach with our text consists of two categories, and our purpose is thus twofold.

Our primary purpose is to describe to mathematics (or applied mathematics) students, typically at undergraduate level, or in their early years of graduate studies, the various steps involved in a modern modelling endeavor. The multiscale simulation of complex fluids is an excellent example for this. Thinking to this audience, we concentrate ourselves on key issues in the modelling, assuming only the knowledge of some basic notions of continuum mechanics (briefly recalled in Sect. 2) and elaborating on those in Sect. 3 to construct the simplest multiscale models for complex fluids. We also assume that these students are familiar with some standard notions about partial differential equations and the discretization techniques commonly used for their simulation. On the other hand, because we know from our teaching experience that such students often have only a limited knowledge in probability theory and stochastic analysis, we choose to give (in Sect. 4) a *crash course* on the elements of stochastic analysis needed to manipulate the stochastic versions of the models for complex fluids. The latter are introduced in the second part of Sect. 4. To illustrate the notions introduced on a very simple case, and to allow our readers to get into the heart in the matter, we devote the entire Sect. 5 to several possible variants of numerical approaches for the simulation of start-up Couette flows. This simple illustrative case serves as a test-bed for the numerical strategies described in a more general context throughout the text. A complete set of codes for the numerical simulation is provided, which we encourage the readers to work with like in a *hands-on* session.

The second category of readers we would like to get interested in the present article consists of practitioners of the field, namely experts in complex fluids rheology and mechanics, or chemical engineers. The present text could serve, we believe, as an introduction to a mathematical viewpoint on their activity. Clearly, the issues we, as mathematicians and computational scientists, emphasize, are somewhat different from those they consider on a regular basis. The perspective also is different. We are looking forward to their feedback on the text.

For both communities above, we are aware that an introductory text, although useful, is not fully satisfactory. This is the reason why we devote a section of our article, Sect. 6, to a short, however comprehensive, description of the mathematical and numerical challenges of the field. This section is clearly more technical, and more mathematical in nature, than the preceding ones. It is, hopefully, interesting for advanced graduate students and researchers, professionals in mathematics, applied mathematics or scientific computing. The other readers are of course welcome to discover there what the exciting unsolved questions of the field are.

Finally, because we do not want our readers to believe that the modelling of infinitely dilute solutions of flexible polymers is the only context within complex fluids science where mathematics and multiscale simulation can bring a lot, we close the loop, describing in our last Sect. 7 *some* other types of complex fluids where the same multiscale approach can be employed.

2 Incompressible fluid mechanics: Newtonian and non-Newtonian fluids

2.1 Basics

To begin with, we recall here some basic elements on the modelling of incompressible fluids.

Consider a viscous fluid with volumic mass (or density) ρ, flowing at the velocity $\boldsymbol{u}$. It experiences external forces $\boldsymbol{f}$ per unit mass. Denote by $\boldsymbol{T}$ the stress tensor.

The equation of conservation of mass for this fluid reads

$$\frac{\partial\rho}{\partial t} + \operatorname{div}(\rho\,\boldsymbol{u}) \quad 0. \tag{1}$$

On the other hand, the equation expressing the conservation of momentum is

$$\frac{\partial(\rho\,\boldsymbol{u})}{\partial t} + \operatorname{div}(\rho\,\boldsymbol{u}\otimes\boldsymbol{u}) - \operatorname{div}\boldsymbol{T} \quad \rho\boldsymbol{f}, \tag{2}$$

where $\otimes$ denotes the tensor product: for two vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ in $\mathbb{R}^d$, $\boldsymbol{u}\otimes\boldsymbol{v}$ is a $d\times d$ matrix with (i,j)-component $\boldsymbol{u}_i\boldsymbol{v}_j$. For such a viscous fluid, the stress tensor reads

$$\boldsymbol{T} \quad -pId + \boldsymbol{\tau}, \tag{3}$$

where p is the (hydrodynamic) pressure, and $\boldsymbol{\tau}$ is the tensor of viscous stresses. In order to close the above set of equations, a *constitutive law* (or *constitutive relation*) is needed, which relates the viscous stress $\boldsymbol{\tau}$ and the velocity field $\boldsymbol{u}$, namely

$$\boldsymbol{\tau} \quad \boldsymbol{\tau}(\boldsymbol{u},\rho,...). \tag{4}$$

Note that (4) is symbolic. A more precise formulation could involve derivatives in time, or in space, of the various fields $\boldsymbol{\tau}$, $\boldsymbol{u}$, ρ, ...

Assuming that $\boldsymbol{\tau}$ linearly depends on $\boldsymbol{u}$, that $\boldsymbol{\tau}$ is invariant under the change of Galilean referential, and that the fluid has isotropic physical properties, it may be shown that the relation between $\boldsymbol{\tau}$ and $\boldsymbol{u}$ necessarily takes the following form

$$\boldsymbol{\tau} \quad \lambda\,(\operatorname{div}\boldsymbol{u})Id + 2\eta\,\boldsymbol{d} \tag{5}$$

where λ and η are two scalar coefficients (called the *Lamé coefficients*). The latter depends, in full generality, on the density ρ and the temperature. In (5), $\boldsymbol{d}$ denotes the (linearized) rate of deformation tensor (or rate of strain tensor)

$$\boldsymbol{d} \quad \frac{1}{2}(\nabla\boldsymbol{u} + \nabla\boldsymbol{u}^T). \tag{6}$$

When a fluid obeys the above assumptions, it is called a *Newtonian fluid*. The kinetic theory of gases allows to show that

$$\lambda \quad -\frac{2}{3}\eta \tag{7}$$

and the common practice is to consider both coefficients λ and η constant.

The system of equations (1)-(2)-(3)-(5)-(6)-(7) allows then to describe the motion of the fluid. When accounting for temperature effects, or for compressible effects, the system is complemented by two additional equations, the energy equation and an equation of state (relating p, ρ and T). We will neglect such effects in the following and assume the temperature is constant and the fluid is incompressible:

$$\operatorname{div} \boldsymbol{u} \quad 0. \tag{8}$$

Then, equations (1)-(2)-(3)-(5)-(6)-(7)-(8) provide the complete description of the evolution of the Newtonian fluid.

Let us additionally assume the fluid has constant density

$$\rho \quad \rho_0.$$

Such a fluid is often called *homogeneous*. The equation of conservation of momentum then rewrites

$$\rho\left(\frac{\partial \boldsymbol{u}}{\partial t} + (\boldsymbol{u}\cdot\nabla)\,\boldsymbol{u}\right) - \eta\Delta\boldsymbol{u} + \nabla p \quad \rho\boldsymbol{f}. \tag{9}$$

It is supplied with the divergence-free condition

$$\operatorname{div} \boldsymbol{u} \quad 0. \tag{10}$$

The couple of equations (9)-(10) form is the celebrated *Navier-Stokes equation* for the motion of incompressible homogeneous viscous Newtonian fluids.

2.2 Non-Newtonian fluids

Some experimental observations

Non-Newtonian fluids, and, in particular, viscoelastic fluids are ubiquitous in industry (oil industry, food industry, rubber industry, for example), as well as in nature (blood is a viscoelastic fluid). As mentioned above, Newtonian fluids are characterized by the fact that the stress is proportional to the rate of deformation $\frac{1}{2}\left(\nabla\boldsymbol{u} + \nabla\boldsymbol{u}^T\right)$: this is *viscosity*. For elastic solids, the stress is proportional to the deformation (see the tensors (35) $\boldsymbol{C}_t$ or (36) $\boldsymbol{F}_t$ below for some measure of deformation): this is *elasticity*. The characteristic feature of viscoelastic fluids is that their behavior is both viscous and elastic. Polymeric fluids are one instance of *viscoelastic* fluids.

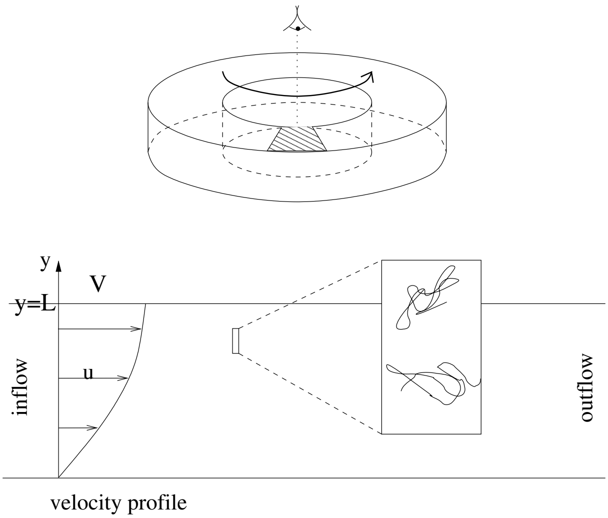


Fig. 1. Schematic representation of a rheometer. On an infinitesimal angular portion, seen from the top, the flow is a simple shear flow (Couette flow) confined between two planes with velocity profile $(u(t,y),0,0)$.

To explore the rheological behavior of viscoelastic fluids (*rheology* is the science studying why and how fluids flow), physicists study their response to so-called *simple flows* (typically flows in pipes or between two cylinders) to obtain so-called *material functions* (such as shear viscosity, differences of normal stress, see below). Typically, for such flows, the velocity field is known and is not influenced by the non-Newtonian features of the fluid. This owes to the fact that the velocity field is homogeneous, which means that $\nabla \boldsymbol{u}$ does not depend on the space variable. Such flows are called *homogeneous flows*. Two types of simple flows are very often used in practice: *simple shear flows* and *elongational flows* (see R.B. Bird, R.C. Armstrong and O. Hassager [11, Chap. 3]). We focus here on simple shear flows. In practical situations (in an industrial context for example), flows are generally more complicated than the simple flows used to characterize the rheological properties of the fluids: such flows are called *complex flows*. Complex flows are typically not homogeneous: $\nabla \boldsymbol{u}$ depends on the space variable $\boldsymbol{x}$.

In a simple shear flow, the velocity $\boldsymbol{u}$ has the following form:

$$\boldsymbol{u}(t,\boldsymbol{x}) \quad (\dot{\gamma}(t)y,0,0),$$

where $\boldsymbol{x} \quad (x,y,z)$ and $\dot{\gamma}$ is the shear rate. The *shear viscosity* η:

$$\eta(t) \quad \frac{\boldsymbol{\tau}_{x,y}(t)}{\dot{\gamma}(t)}, \tag{11}$$

and the *first and second differences of normal stress*:

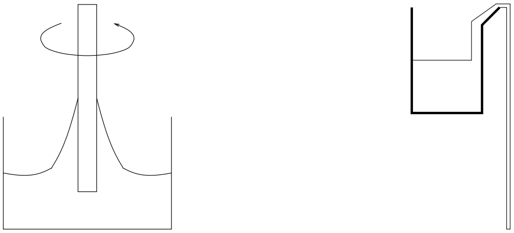


Fig. 2. Schematic representation of two unexpected, counterintuitive behaviors for some polymeric fluids: the *rod-climbing effect* (left) and the *open syphon effect* (right).

$$\begin{aligned} N_1(t) &\quad \boldsymbol{\tau}_{x,x}(t) - \boldsymbol{\tau}_{y,y}(t), \\ N_2(t) &\quad \boldsymbol{\tau}_{y,y}(t) - \boldsymbol{\tau}_{z,z}(t), \end{aligned} \tag{12}$$

may be measured experimentally. For Newtonian fluids, the shear viscosity is constant, and both N_1 and N_2 vanish. This is not the case in general for viscoelastic fluids. In particular, for many non-Newtonian fluids, η is a decreasing function of $\dot{\gamma}$ (this property is called *shear-thinning*), goes to a constant η_∞ when $\dot{\gamma}$ goes to infinity, and goes to some value η_0 (the *zero-shear rate viscosity*) when $\dot{\gamma}$ goes to zero.

In practice, such flows are studied in rheometers, the fluid being confined between two cylinders. The outer cylinder is fixed, the inner one is rotating (see Fig. 1). On an infinitesimal portion, the flow can be approximated by a simple shear flow. We will return to this in Sect. 5.

The simple shear flow may also be useful to study the dynamics of the fluid using an oscillating excitation: $\dot{\gamma}(t) \quad \gamma_0 \cos(\omega t)$. The in-phase response with the deformation is related to the elasticity of the fluid. The out-of-phase response is related to the viscosity of the fluid. This can be easily understood for example in the simple Maxwell model presented below, and an analogy with electric circuits (see Fig. 4).

Before addressing the modelling in details, let us mention some peculiar behaviors of some non-Newtonian fluids.

We first describe the *rod-climbing effect* (see Fig. 2 or R. G. Owens and T. N. Phillips [104, Fig. 1.9]). A rod is introduced in the fluid and is rotated: for a Newtonian fluid, inertia causes the fluid to dip near the rod and rise at the walls. For some non-Newtonian fluids, the fluid may actually climb the rod (this is called the *Weissenberg effect*). This phenomenon is related to non zero normal stress differences (see A.S. Lodge [91]).

Another experiment is the *open syphon effect* (see Fig. 2 or R.G. Owens and T.N. Phillips [104, Fig. 1.11]). A beaker is tilted so that a small thread of the fluid starts to flow over the edge, and then is put straight again. For some viscoelastic fluids, the liquid keeps on flowing out.

Another, simpler experiment, which we will be able to reproduce with a micro-macro model and a simple numerical computation (see Sect. 5) is the *start-up of shear flow*. A fluid initially at rest and confined between two plates is sheared (one

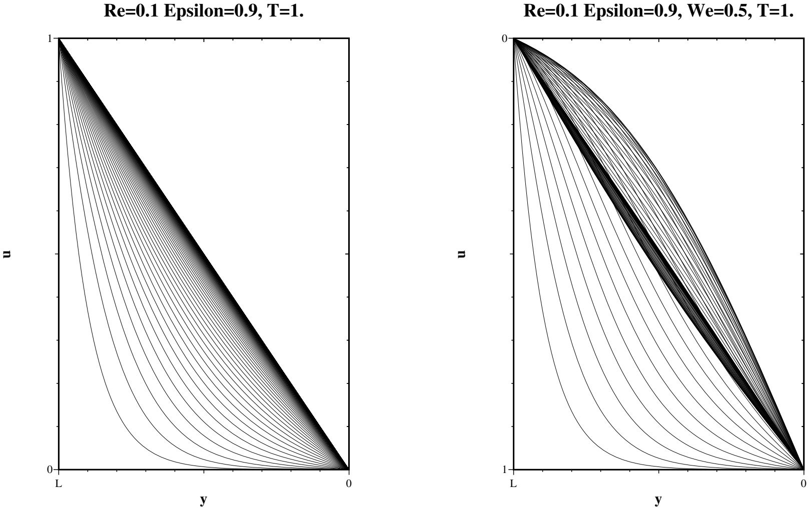


Fig. 3. Velocity profile as a function of time for a *start-up of shear flow*. The velocity profile (u as a function of y, see Fig. 1) is represented at various times in the time interval $0,1$. For polymeric fluids (on the right, case of the Hookean dumbbell micro-macro model) an overshoot of the velocity is observed, while this is not the case for Newtonian fluid (on the left).

plate is moving, and the other one is fixed) (see Figs. 1 and 3). For Newtonian fluids, the velocity profile progressively reaches monotonically the stationary state. For some polymeric fluids, the velocity goes beyond its stationary value: this is the *overshoot* phenomenon.

Modelling of non-Newtonian fluids

When the fluid, although viscous, incompressible and homogeneous, does not obey the simplifying assumptions leading to (5), the following system of equations is to be used, in lieu of (9)-(10):

$$\begin{cases} \rho\left(\dfrac{\partial \boldsymbol{u}}{\partial t}+(\boldsymbol{u}\cdot\nabla)\boldsymbol{u}\right)-\eta\Delta\boldsymbol{u}+\nabla p-\operatorname{div}\boldsymbol{\tau}_p \quad \rho\boldsymbol{f} \\ \operatorname{div}\boldsymbol{u} \quad 0 \end{cases} \tag{13}$$

where the stress $\boldsymbol{\tau}$ has been decomposed along

$$\boldsymbol{\tau} \quad \boldsymbol{\tau}_n+\boldsymbol{\tau}_p \tag{14}$$

giving birth to the terms $-\eta\Delta\boldsymbol{u}$, and $\operatorname{div}\boldsymbol{\tau}_p$, respectively. In (14), $\boldsymbol{\tau}_n$ denotes the Newtonian contribution (expressed as in (5)) and $\boldsymbol{\tau}_p$ denotes the part of the stress (called *non-Newtonian* or extra stress) which cannot be modelled as in (4). Our notation $\boldsymbol{\tau}_p$ refers to the fact we will mainly consider in the sequel fluids for which the non-Newtonian character owes to the presence of polymeric chains flowing in a solvent.

For non-Newtonian fluids, many purely macroscopic models exist. All are based upon considerations of continuum mechanics. The bottom line is to write an equation, in the vein of (4), ruling the evolution of the non-Newtonian contribution $\boldsymbol{\tau}_p$ to the stress tensor, and/or a relation between the latter and other quantities characterizing the fluid dynamics, such as the deformation tensor $\boldsymbol{d}$ or $\nabla\boldsymbol{u}$ itself. Such an equation may read

$$\frac{D\boldsymbol{\tau}_p}{Dt} \quad F(\boldsymbol{\tau}_p, \nabla\boldsymbol{u}), \tag{15}$$

where $\frac{D\bullet}{Dt}$ denotes an appropriate extension (for tensorial quantities, see next section) of the usual convected derivative for vectors

$$\frac{\partial\bullet}{\partial t} + (\boldsymbol{u}\cdot\nabla)\bullet.$$

A model such as (15) is called a *differential model* for the non-Newtonian fluid. One famous example is the *Oldroyd B model*. It will be made precise in the next section.

An alternative option is to resort to an *integral model*:

$$\boldsymbol{\tau}_p(t,\boldsymbol{x}) \quad \int_{-\infty}^{t} m(t-t')\boldsymbol{S}_{t'}\,dt', \tag{16}$$

where m is a so-called memory kernel (typically $m(s) \quad \exp(-s)$), $\boldsymbol{S}_{t'}$ denotes a quantity depending on $\nabla\boldsymbol{u}$, and where the integral is considered along a fluid trajectory (or pathline) ending at point $\boldsymbol{x}$. We shall also return to such models in the next section.

The major observation on both forms (15) and (16) is that, in contrast to the Newtonian case (5), $\boldsymbol{\tau}_p(t,\boldsymbol{x})$ does not only depend on the deformation at point $\boldsymbol{x}$ and at time t (as it would be the case in (5)), but also depends on the *history* of the deformation for all times $t' \leq t$, along the fluid trajectory leading to $\boldsymbol{x}$. It is particularly explicit on the form (16), but may also be seen on (15).

The complete system of equations modelling the fluid reads

$$\begin{cases} \rho\left(\frac{\partial\boldsymbol{u}}{\partial t} + (\boldsymbol{u}\cdot\nabla)\boldsymbol{u}\right) - \eta\Delta\boldsymbol{u} + \nabla p - \operatorname{div}\boldsymbol{\tau}_p \quad \rho\boldsymbol{f}, \\ \qquad\qquad\qquad\qquad\qquad\qquad \operatorname{div}\boldsymbol{u} \quad 0, \\ \qquad\qquad\qquad\qquad\qquad\qquad \frac{D\boldsymbol{\tau}_p}{Dt} \quad F(\boldsymbol{\tau}_p, \nabla\boldsymbol{u}). \end{cases} \tag{17}$$

This system is called a *three-field system*. It involves the velocity $\boldsymbol{u}$, the pressure p, and the stress $\boldsymbol{\tau}_p$.

Solving this three-field problem is much more difficult and computationally demanding than the 'simple' Newtonian problem (9)–(10), that is (13) where $\boldsymbol{\tau}_p \quad 0$ and only two fields, the velocity and the pressure, are to be determined. However, the major *scientific* difficulty is neither a mathematical one nor a computational one. The major difficulty is to *derive* an equation of the type (15) or (16). It requires a deep, qualitative and quantitative, understanding of the physical properties of the fluid under consideration. For many non-Newtonian fluids, complex in nature, reaching such an understanding is a challenge. Moreover, even if such an equation is approximately known, evaluating the impact of its possible flaws on the final result of the simulation is not an easy matter. It can only be done *a posteriori*, comparing the results to actual experimental observations, when the latter exist, and they do not always exist. The difficulty is all the more prominent that the non-Newtonian fluids are very diverse in nature. New materials appear on a daily basis. For traditional fluids considered under unusual circumstances, or for recently (or even not yet) synthesized fluids, reliable evolution equations are not necessarily available.

All this, in its own, motivates the need for alternative strategies, based on an explicit microscopic modelling of the fluid. This will give rise to the so-called *micro-macro models*, which are the main topic of this article. The lack of information at the macroscopic level is then circumvented by a multiscale strategy consisting in searching for the information at a finer level (where reliable models do exist, based on general conservation equations, posed *e.g.* on the microstructures of the fluids). The latter information is then inserted in the equations of conservation at the macroscopic level. At the end of the day, because the modelling assumptions are avoided as much as possible, a complete description is attained, based on a more reliable, however much more computationally demanding, model. Otherwise stated, a crucial step of the modelling is replaced by a numerical simulation. But before we turn to this, from Sect. 3 on, let us give some more details on the purely macroscopic models (17) for non-Newtonian fluids. They are today the most commonly used models (in particular because they are less demanding computationally). For our explanatory survey, we choose the context of *viscoelastic fluids*.

2.3 Some macroscopic models for viscoelastic fluids

Throughout this section, the stress tensor $\boldsymbol{\tau}$ is decomposed into a Newtonian part and a non-Newtonian part, as in (14). The former, $\boldsymbol{\tau}_n$, reads $\boldsymbol{\tau}_n \quad \eta\dot{\boldsymbol{\gamma}}$ where η is the viscosity, and $\dot{\boldsymbol{\gamma}}$ is given by

$$\dot{\boldsymbol{\gamma}} \quad \nabla\boldsymbol{u}+\nabla\boldsymbol{u}^T. \tag{18}$$

The latter is denoted by $\boldsymbol{\tau}_p$. The stress is the combination of the two, namely:

$$\boldsymbol{\tau} \quad \eta\dot{\boldsymbol{\gamma}}+\boldsymbol{\tau}_p. \tag{19}$$

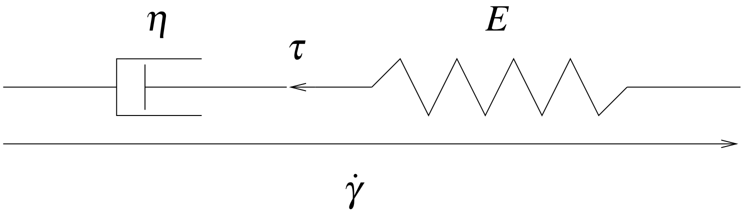


Fig. 4. One-dimensional Maxwell model. The analogy with an electric circuit is obvious, τ and $\dot{\gamma}$ playing the role of the current intensity and the voltage respectively, η and E that of the capacity of a capacitor and the conductivity of a resistor respectively.

More on differential models

The basic model for viscoelastic fluids is the *Maxwell model*. It combines a linear elasticity model and a linear viscosity model. In the former, the stress depends linearly on the deformation. It is the *Hooke law*. The proportionality constant is the *Young modulus E*. This part of the stress is to be thought of as a linear spring. On the other hand, the linear viscosity model assumes the stress depends linearly on the rate (or speed) of deformation, the proportionality constant being the viscosity η. Heuristically, this amounts to considering a piston. The one-dimensional Maxwell model combines the Hookean spring and the piston (see Fig. 4). Then, denoting the stress by τ and the deformation rate by $\dot{\gamma}$, the following ordinary differential equation is obtained:

$$\dot{\gamma} \quad \frac{1}{E}\frac{d\tau}{dt}+\frac{\tau}{\eta}, \tag{20}$$

that is,

$$\lambda\frac{d\tau}{dt}+\tau \quad \eta\dot{\gamma}, \tag{21}$$

where $\lambda \quad \frac{\eta}{E}$ denotes a characteristic relaxation time of the system.

Remark 1. The mathematically inclined reader should not be surprised by the elementary nature of the above arguments. *Modelling is simplifying*. Some excellent models of fluid mechanics (and other fields of the engineering and life sciences) are often obtained using such simple derivations. On the other hand, it is intuitively clear that the determination of the parameters of such models is often an issue, which limits their applicability, and that there is room for improvement using more advanced descriptions of matter. This will be the purpose of the multiscale models introduced in the present article.

Passing from the one-dimensional setting to higher dimensions requires to replace the time derivative in (21) by a convective derivative of a tensor. Based on invariance arguments, the following model is derived:

$$\lambda\left(\frac{\partial\boldsymbol{\tau}_p}{\partial t}+\boldsymbol{u}\cdot\nabla\boldsymbol{\tau}_p-\nabla\boldsymbol{u}\boldsymbol{\tau}_p-\boldsymbol{\tau}_p(\nabla\boldsymbol{u})^T\right)+\boldsymbol{\tau}_p \quad \eta_p\dot{\boldsymbol{\gamma}}, \tag{22}$$

where λ is a relaxation time, and η_p is an extra viscosity (due to the polymers in our context). Then the stress tensor $\boldsymbol{\tau}$ is given by (19). When $\eta \quad 0$, the model is called the *Upper Convected Maxwell model* (UCM). When $\eta \; / \; 0$, it is the *Oldroyd-B model*, also called the *Jeffreys model.*

For future reference, let us rewrite the complete system of equations for the Oldroyd-B model, in a non-dimensional form:

$$\begin{cases} \mathrm{Re}\left(\dfrac{\partial \boldsymbol{u}}{\partial t} + \boldsymbol{u}\cdot\nabla\boldsymbol{u}\right) \quad (1-\varepsilon)\Delta\boldsymbol{u} - \nabla p + \operatorname{div}\boldsymbol{\tau}_p, \\ \operatorname{div}\boldsymbol{u} \quad 0, \\ \dfrac{\partial \boldsymbol{\tau}_p}{\partial t} + \boldsymbol{u}\cdot\nabla\boldsymbol{\tau}_p - (\nabla\boldsymbol{u})\boldsymbol{\tau}_p - \boldsymbol{\tau}_p(\nabla\boldsymbol{u})^T \quad -\dfrac{1}{\mathrm{We}}\boldsymbol{\tau}_p + \dfrac{\varepsilon}{\mathrm{We}}\left(\nabla\boldsymbol{u} + (\nabla\boldsymbol{u})^T\right). \end{cases} \tag{23}$$

The *Reynolds number* $\mathrm{Re} > 0$, the *Weissenberg number* $\mathrm{We} > 0$ and $\varepsilon \in (0,1)$ are the non-dimensional numbers of the system (see Equations (97) below for precise definitions). The Weissenberg number (which is the ratio of the characteristic relaxation time of the microstructures in the fluid to the characteristic time of the fluid) plays a crucial role in the stability of numerical simulations (see Sect. 4.4).

Remark 2. In (22) and throughout this article, we denote by $(\nabla\boldsymbol{u})_{i,j} \quad \dfrac{\partial u_i}{\partial x_j}$. Other, and in fact many authors in the literature of non-Newtonian fluid mechanics (see *e.g.* R.B. Bird, C.F. Curtiss, R.C. Armstrong and O. Hassager [11, 12], R.G. Owens and T.N. Phillips [104], or H.C. Öttinger [102])), adopt the alternative convention: $(\nabla\boldsymbol{u})_{i,j} \quad \dfrac{\partial u_j}{\partial x_i}$. Our equations have to be modified correspondingly.

Remark 3. The convective derivative in (22) is called the *upper-convected derivative*. Other derivatives may be considered, such as the *lower-convected derivative*, or the *co-rotational derivative* (the latter being particularly interesting for mathematical purposes, see Sect. 6). All these derivatives obey the appropriate invariance laws of mechanics, but we have chosen the upper-convected derivative because it spontaneously arises when using the kinetic models (see Sect. 3). It is also the most commonly used derivative in macroscopic models, such as the Phan-Thien Tanner model, the Giesekus model or the FENE-P model. We shall return to such models later on. A discussion of the physical relevance of convective derivatives appears in D. Bernardin [10, Chap. 3]. See also R.B. Bird, R.C. Armstrong and O. Hassager [11, Chap. 9].

The Oldroyd B model has several flaws, as regards its ability to reproduce experimentally observed behaviors.

Refined macroscopic models for viscoelastic fluids have thus been derived, allowing for a better agreement between simulation and experience. In full generality, such models read:

$$\lambda\left(\frac{\partial \boldsymbol{\tau}_p}{\partial t} + \boldsymbol{u}\cdot\nabla\boldsymbol{\tau}_p - \nabla\boldsymbol{u}\boldsymbol{\tau}_p - \boldsymbol{\tau}_p(\nabla\boldsymbol{u})^T\right) + \boldsymbol{T}(\boldsymbol{\tau}_p, \dot{\boldsymbol{\gamma}}) \quad \eta_p\dot{\boldsymbol{\gamma}}, \tag{24}$$

where $\boldsymbol{T}(\boldsymbol{\tau}_p, \dot{\boldsymbol{\gamma}})$ typically depends nonlinearly on $\boldsymbol{\tau}_p$. The most commonly used models are the following three. The *Giesekus model*(see H. Giesekus [51]) involves a quadratic term:

$$\lambda\left(\frac{\partial \boldsymbol{\tau}_p}{\partial t} + \boldsymbol{u}\cdot\nabla\boldsymbol{\tau}_p - \nabla\boldsymbol{u}\boldsymbol{\tau}_p - \boldsymbol{\tau}_p(\nabla\boldsymbol{u})^T\right) + \boldsymbol{\tau}_p + \alpha\frac{\lambda}{\eta_p}\boldsymbol{\tau}_p\boldsymbol{\tau}_p \quad \eta_p\dot{\boldsymbol{\gamma}}. \tag{25}$$

where α is a fixed scalar.

The *Phan-Thien Tanner model* (PTT) is derived from a lattice model (see N. Phan-Thien and R.I. Tanner [106]). It writes:

$$\lambda\left(\frac{\partial \boldsymbol{\tau}_p}{\partial t} + \boldsymbol{u}\cdot\nabla\boldsymbol{\tau}_p - \nabla\boldsymbol{u}\boldsymbol{\tau}_p - \boldsymbol{\tau}_p(\nabla\boldsymbol{u})^T\right) + Z(\mathrm{tr}(\boldsymbol{\tau}_p))\boldsymbol{\tau}_p + \frac{\xi}{2}\lambda\left(\dot{\boldsymbol{\gamma}}\boldsymbol{\tau}_p + \boldsymbol{\tau}_p\dot{\boldsymbol{\gamma}}\right) \quad \eta_p\dot{\boldsymbol{\gamma}}, \tag{26}$$

with two choices for the function Z :

$$Z(\mathrm{tr}(\boldsymbol{\tau}_p)) \quad \begin{cases} 1 + \phi\lambda\dfrac{\mathrm{tr}(\boldsymbol{\tau}_p)}{\eta_p} \\ \exp\left(\phi\lambda\dfrac{\mathrm{tr}(\boldsymbol{\tau}_p)}{\eta_p}\right) \end{cases}, \tag{27}$$

where ξ and ϕ are fixed scalars.

The *FENE-P model*, which we will return to in Sect. 4.2, is derived from a kinetic model (see A. Peterlin [105] and R.B. Bird, P.J. Dotson and N.L. Johnson [13] and Sect. 4). It reads:

$$\begin{cases} \lambda\left(\dfrac{\partial \boldsymbol{\tau}_p}{\partial t} + \boldsymbol{u}\cdot\nabla\boldsymbol{\tau}_p - \nabla\boldsymbol{u}\boldsymbol{\tau}_p - \boldsymbol{\tau}_p(\nabla\boldsymbol{u})^T\right) + Z(\mathrm{tr}(\boldsymbol{\tau}_p))\boldsymbol{\tau}_p \\ \quad -\lambda\left(\boldsymbol{\tau}_p + \dfrac{\eta_p}{\lambda}\mathrm{Id}\right)\left(\left(\dfrac{\partial}{\partial t} + \boldsymbol{u}\cdot\nabla\right)\ln\left(Z(\mathrm{tr}(\boldsymbol{\tau}_p))\right)\right) \quad \eta_p\dot{\boldsymbol{\gamma}}, \end{cases} \tag{28}$$

with

$$Z(\mathrm{tr}(\boldsymbol{\tau}_p)) \quad 1 + \frac{d}{b}\left(1 + \lambda\frac{\mathrm{tr}(\boldsymbol{\tau}_p)}{d\,\eta_p}\right), \tag{29}$$

where d is the dimension of the ambient space and b is a scalar that is thought of as related to the maximal extensibility of the polymer chains embedded in the fluids (see the FENE force below, Equation (91)).

All these nonlinear models yield much better results than the Oldroyd B model, and satisfactorily agree with experiments on simple flows. They can be further generalized, considering several relaxation times λ_i and several viscosities $\eta_{p,i}$, but we will not proceed further in this direction in this introductory survey.

More on integral models

Let us return to the one-dimensional Maxwell model (21). Its solution may be explicitly written in terms of the integral:

$$\tau(t) \quad \tau(t_0)\exp\left(-\frac{t-t_0}{\lambda}\right)+\int_{t_0}^{t}\frac{\eta}{\lambda}\exp\left(-\frac{t-s}{\lambda}\right)\dot{\gamma}(s)\,ds. \tag{30}$$

Letting t_0 go to $-\infty$, and assuming τ bounded when $\dot{\gamma}$ is bounded, we obtain:

$$\tau(t) \quad \int_{-\infty}^{t}\frac{\eta}{\lambda}\exp\left(-\frac{t-s}{\lambda}\right)\dot{\gamma}(s)\,ds. \tag{31}$$

Denoting by:

$$\begin{cases} \dfrac{d}{dt}\gamma(t_0,t) \quad \dot{\gamma}(t) \\ \\ \gamma(t_0,t_0) \quad 0 \end{cases}, \tag{32}$$

and integrating by parts, we obtain a form equivalent to (31) :

$$\tau(t) \quad \int_{-\infty}^{t}\frac{\eta}{\lambda^2}\exp\left(-\frac{t-s}{\lambda}\right)\gamma(t,s)\,ds. \tag{33}$$

This form explicitly shows that, as announced earlier, the constraint at time t depends on the history of the deformation. The function $\frac{\eta}{\lambda^2}\exp\left(-\frac{t-s}{\lambda}\right)$ is often called a *memory function*.

The one-dimensional computation performed above can be generalized to higher dimensions and yields:

$$\boldsymbol{\tau}_p(t,\boldsymbol{x}) \quad -\int_{-\infty}^{t} M(t-s)f\left(\boldsymbol{C}_t^{-1}(s,\boldsymbol{x})\right)\left(\mathrm{Id}-\boldsymbol{C}_t^{-1}(s,\boldsymbol{x})\right)ds. \tag{34}$$

where M is a memory function, f is a given real valued function, and $\boldsymbol{C}_t^{-1}(s,\boldsymbol{x})$ denotes the so-called *Finger tensor* (at time t). The latter is the inverse tensor of the *Cauchy deformation tensor*:

$$\boldsymbol{C}_t(s,\boldsymbol{x}) \quad \boldsymbol{F}_t(s,\boldsymbol{x})^T\boldsymbol{F}_t(s,\boldsymbol{x}) \tag{35}$$

where

$$\boldsymbol{F}_t(s,\boldsymbol{x}) \quad \nabla(\boldsymbol{\chi}_t(s))(\boldsymbol{x}) \tag{36}$$

is the deformation tensor and $\boldsymbol{\chi}_t(s)$ is the flow chart (mapping positions at time t to positions at time s).

It is easily seen that the upper-convected derivative of the Finger tensor vanishes. When $M(t-s) \quad \dfrac{\eta_p}{\lambda^2}\exp\left(-\dfrac{t-s}{\lambda}\right)$ and $f \quad 1$, this shows that $\boldsymbol{\tau}_p$ defined by (34) satisfies (22). The parameter λ models the time needed by the system to "forget" the history of the deformation.

Remark 4. As in the case of differential models, there exist many generalizations and variants for the integral models introduced above. Alternative convective derivatives may be considered, several characteristic times λ_i and viscosities $\eta_{p,i}$ can be employed. See R.B. Bird, R.C. Armstrong and O. Hassager [11] or D. Bernardin [10] for such extensions.

3 Multiscale modelling of polymeric fluids

There exists an incredibly large variety of non-Newtonian fluids. We have briefly overviewed in the previous section the modelling of viscoelastic fluids. This is one category of non-Newtonian fluids. One important class of non-Newtonian fluids is the family of *fluids with microstructures*. For such fluids, the non-Newtonian character owes to the presence of microstructures, often more at a mesoscopic scale than at a truly microscopic one. Snow, blood, liquid concrete, etc, are examples of fluids with microstructures. Polymeric fluids form the category we will focus on in the sequel. Analogous scientific endeavors can be followed for other fluids with microstructures. The bottom line for the modelling remains: write an equation at the microscopic level that describes the evolution of the microstructures, then deduce the non-Newtonian contribution $\boldsymbol{\tau}_p$ to the stress. Thus a better quantitative understanding. Section 7 will give some insight on other types of fluids with microstructures.

The present section is only a brief introduction to the subject. To read more on the multiscale modelling of complex fluids, we refer to the monographs: R. Bird, Ch. Curtiss, C. Armstrong and O. Hassager [11, 12], H.C. Öttinger [102], R. Owens and T. Phillips [104]. Other relevant references from the physics perspective are F. Devreux [34]. M. Doi [35], M. Doi and S.F. Edwards [36], M.P. Allen and D.J. Tildesley [1], D. Frenkel and B. Smit [47].

Before we get to the heart of the matter, let us briefly introduce the reader to the specificities of polymeric fluids.

A *polymer* is, by definition, a molecular system formed by the repetition of a large number of molecular subsystems, the *monomers*, all bound together by intramolecular forces. If the subsystems are not all of the same chemical type, one speaks of *copolymers*. Polymeric materials are ubiquitous: they may be of natural origin, such as natural rubber, wood, leather, or artificially synthesized, such as vulcanized rubber or plastic. They can be classified according to their polymerization degree, that is the number N of monomers present in the chain: N 1 to 4 for gases, N 5 for oils, N 25 for brittle materials such as a candle, $N > 2000$ for ductile materials such as plastic films. As N grows, the fusion temperature grows and the polymeric properties become prominent: they are already significant for N 100, and obviously dominant for N 1000. The specific mechanical properties of the material stem from the long chains present inside. The length of the chain for instance prevents the material from organizing itself regularly when solidification occurs, thus the flexibility of the material (such as a tire). Likewise, the long chains give additional viscosity to liquid polymers, such as oils. Solvents may enjoy good, or bad, solvating properties for the polymers, depending whether the chains expand or retract in the solvent. For example, paints are solvated differently in water and oils.

As regards the concentration of polymers within the solvent, three cases may schematically arise. When the concentration is low, one speaks of *infinitely dilute polymeric fluids*. There, the chains basically ignore each other, interacting with one another only through the solvent. This is the case we will mainly consider in the sequel. The other extreme case is the case of *dense polymeric fluids*, also called *polymer* *melts*. In-between, one finds polymeric fluids with intermediate concentrations. Of the above three categories, polymer melts form indeed the most interesting one, technically and industrially. Their modelling has made great progress in the 1960s with the contributions by De Gennes, and his theory of *reptation*. Basically, it is considered that, owing to the density of polymeric chains present, each single chain moves in the presence of others like a snake in a dense bush, or a spaghetti in a plate of spaghettis.

Reptation models amount, mathematically, to equations for the evolution of microstructures similar in spirit to those that will be manipulated below. There are however significant differences. Macroscopic versions also exist. In any case, models for polymer melts are much less understood mathematically than models for infinitely dilute polymers. For this reason, we will not proceed further in this direction in the present mathematically biased text.

Remark 5. Let us give some details about the reptation model for polymer melts (see for example [102, Sect. 6]). In such a model, the microscopic variables are $(\boldsymbol{U}_t, S_t)$ (say at a fixed position in space $\boldsymbol{x}$), where $\boldsymbol{U}_t$ is a three dimensional unit vector representing the direction of the reptating polymer chain at the curvilinear abscissa S_t (S_t is a stochastic process with value in $(0,1)$). The Fokker-Planck equation ruling the evolution of $(\boldsymbol{U}_t, S_t)$ is:

$$\begin{aligned}&\frac{\partial \psi(t,\boldsymbol{x},\boldsymbol{U},S)}{\partial t} + \boldsymbol{u}(t,\boldsymbol{x}) \cdot \nabla_{\boldsymbol{x}} \psi(t,\boldsymbol{x},\boldsymbol{U},S)\\ &\quad -\mathrm{div}_{\boldsymbol{U}} \left((\nabla_{\boldsymbol{x}} \boldsymbol{u}(t,\boldsymbol{x}) \boldsymbol{U} - \nabla_{\boldsymbol{x}} \boldsymbol{u}(t,\boldsymbol{x}) : (\boldsymbol{U} \otimes \boldsymbol{U}) \boldsymbol{U}) \, \psi(t,\boldsymbol{x},\boldsymbol{U},S) \right)\\ &\quad + \frac{1}{\lambda} \frac{\partial^2 \psi(t,\boldsymbol{x},\boldsymbol{U},S)}{\partial S^2},\end{aligned}$$

where : denotes the Frobenius inner product: for two matrices A and B, $A : B \quad \mathrm{tr}(A^T B)$. The boundary conditions for $S \quad 0$ and $S \quad 1$ supplementing the Fokker-Planck equation are

$$\psi(t,\boldsymbol{x},\boldsymbol{U},0) \quad \psi(t,\boldsymbol{x},\boldsymbol{U},1) \quad \frac{1}{4\pi} \delta_{|\boldsymbol{U}| \quad 1},$$

where $\delta_{|\boldsymbol{U}| \quad 1}$ is the Lebesgue (surface) measure on the sphere. In terms of the stochastic process $(\boldsymbol{U}_t, S_t)$, this equation is formally equivalent to a deterministic evolution of the process $\boldsymbol{U}_t$ (the unit vector is rotated following the flow field) and a stochastic evolution of the index S_t as $dS_t + \boldsymbol{u} \cdot \nabla_{\boldsymbol{x}} S_t \, dt \quad \sqrt{2/\lambda} \, dW_t$. The only coupling between $\boldsymbol{U}_t$ and S_t arises when S_t reaches 0 or 1, in which case $\boldsymbol{U}_t$ is reinitialized randomly uniformly on the sphere. The contribution of the polymers to the stress tensor can then be computed using a Kramers formula (similar to (48)), and this closes the micro-macro model. An interesting open mathematical question is to define rigorously the dynamics of the process $(\boldsymbol{U}_t, S_t)$.

Remark 6. Also for concentrated polymers, a regime different from reptation can also be considered. When sufficiently numerous bridges are (chemically) created between

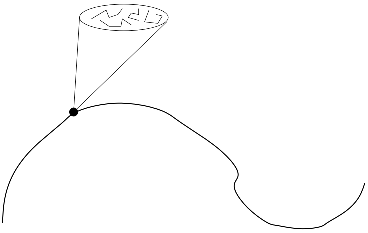


Fig. 5. A collection of polymeric chains lies, microscopically, at each macroscopic point of the trajectory of a fluid particle.

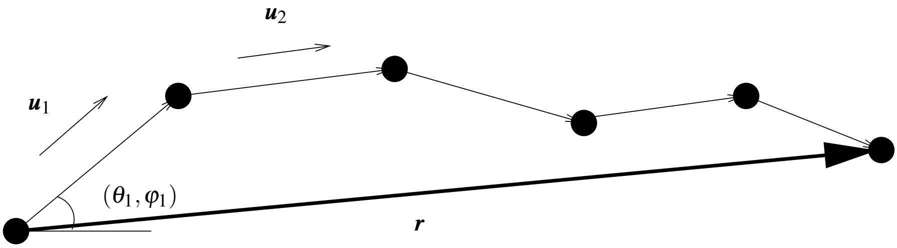


Fig. 6. A polymeric chain: $\boldsymbol{u}_j$ denote the unit vectors between the "atoms", each of them corresponds to a pair of angles (θ_i, φ_i) and has length a. The end-to-end vector is $\boldsymbol{r}$.

the entangled chains (this is exactly the purpose of the vulcanization process involved in the production of tires), the polymeric material turns into a lattice, often called a *reticulated polymer*. Its properties are intermediate between those of a fluid and those of a solid material (owing to the -slight- rigidity provided by the lattice). A multiscale modeling can be envisioned for such materials, but again this is not the purpose of this article. We refer for example to H. Gao and P. Klein [48], or S. Reese [108, 109].

In the sequel of this article (with the notable exception of Sect. 7), we consider infinitely dilute polymeric fluids.

In order to understand the contribution to the stress provided by this assembly of long polymeric chains, we zoom out on such a chain. We now want to write an evolution equation on this object. First we have to model the chain, then see the forces it experiences, and finally write an appropriate evolution equation.

As regards the modelling of a polymeric chain, the point to understand is that it is out of the question to explicitly model all the atoms of the chain. There are thousands of them. The interactions between atoms are incredibly expensive to model. Without even thinking to a model from quantum chemistry, the 'simple' consideration of a classical force-field for the molecular dynamics of an entire polymeric chain is too a computationally demanding task. It can be performed for some sufficiently short chains, considered alone, and not interacting with their environment. But the simulation *in situ*, over time frame relevant for the fluid mechanics simulation, of millions of long chains, is out of reach. Even if it was possible, there is no reason to believe that the actual motion of each single atom, and the precise description of the dynamics of each chain significantly impacts the overall rheology of the fluid.

So the two keywords here are *statistical mechanics* and *coarse-graining* (somehow, these terms are synonymous). The bottom line is to consider one single, hopefully representative chain, sitting at a macropoint of the fluid, then derive some sufficiently simple description of this chain, which carries enough physics to adequately model the impact of the chains onto the fluid, and conversely. This within the limit of our simulation capabilities.

Let us first obtain a coarse-grained model for the chain.

3.1 Statistical mechanics of the free chain

Generalities

As said above, it is not reasonable, and it is not the point, to simulate the actual dynamics of all atoms composing all the chains.

We first choose a representative chain. For simplicity, we assume the chain is a linear arrangement of *N beads* (as opposed to the case of *branched* polymers, where the chain has several branches). Each of these beads models a group of atoms, say 10 to 20 monomer units. They are milestones along the chain. They are assumed to be connected by massless bars with length a. This is the so-called *Kramers chain model* (see Fig. 6). The configuration of the chain, at time t and each macroscopic point $\boldsymbol{x}$, is described by a *probability density* ψ (momentarily implicitly indexed by t and $\boldsymbol{x}$), defined over the space

$$(\theta_1, \varphi_1, ..., \theta_{N-1}, \varphi_{N-1})$$

of Euler angles of the unit vectors $\boldsymbol{u}_i$ along this representative chain. Thus

$$\psi(\theta_1, \varphi_1, ..., \theta_{N-1}, \varphi_{N-1}) d\theta_1 d\varphi_1 \dots d\theta_{N-1} d\varphi_{N-1} \tag{37}$$

is the probability that the chain has angles between (θ_1, φ_1) and $(\theta_1 + d\theta_1, \varphi_1 + d\varphi_1)$ between the first two beads labeled 1 and 2, etc...

Some coarse graining has already been performed by considering these beads instead of the actual atoms, but we will now proceed one step further. We are going to only keep a very limited number of these beads, say N_b, and eliminate (using a limiting procedure) all the $N - N_b$ beads in-between. The typical number of beads kept is well below 100. The simplest possible case, that of $N_b \;\; 2$ beads, is the *dumbbell* case and we will in fact mainly concentrate on it in the sequel.

In order to reduce the description of the chain to the simple knowledge of $N_b \;\; 2$ beads, we are going to consider the vector $\boldsymbol{r}$ linking the first bead to the last one. This vector is called the *end-to-end vector* (see Fig. 6) and writes as the sum

$$\boldsymbol{r} \;\; \sum_{i \;\; 1}^{N-1} a\boldsymbol{u}_i, \tag{38}$$

where $\boldsymbol{u}_i$ is the unit vector describing the i-th link. Between the extreme two beads lies indeed a supposedly large number $N-2$ of beads. The chain is free to rotate around each of these beads: think typically to the arm of a mechanical robot. We first describe all the possible configurations of all the N beads. In a second step, we will pass to the limit $N \longrightarrow +\infty$ in order to obtained a reduced model for the two extreme beads, thereby obtaining a statistics on the end-to-end vector. The motivation for this limit process is of course that, given the two extreme beads, the $N-2$ other beads are in extremely large number.

At equilibrium (namely for zero velocity field for the surrounding solvent and at a fixed temperature), the probability density for the Euler angles (θ_i,φ_i) of the i-th link writes

$$\psi_i(\theta_i,\varphi_i) \quad \frac{1}{4\pi}\sin\theta_i,$$

simply by equiprobability of the orientation of this link. As the chain freely rotates around each bead, the orientations of links are independent from one link to another one, and thus the overall probability density for the sequence of angles $(\theta_1,\varphi_1,...,\theta_{N-1},\varphi_{N-1})$ is simply the product

$$\psi(\theta_1,\varphi_1,...,\theta_{N-1},\varphi_{N-1}) \quad \left(\frac{1}{4\pi}\right)^{N-1}\prod_{i\;\;1}^{N-1}\sin\theta_i. \tag{39}$$

Any statistical quantity (observable) B depending on the state of the chain thus has average value

$$\langle B\rangle \quad \int B(\theta^{N-1},\varphi^{N-1})\,\psi(\theta^{N-1},\varphi^{N-1})\,d\theta^{N-1}\,d\varphi^{N-1} \tag{40}$$

where $\theta^{N-1} \quad (\theta_1,...,\theta_{N-1})$ and $\varphi^{N-1} \quad (\varphi_1,...,\varphi_{N-1})$.

For instance, it is a simple calculation that

$$\langle|\boldsymbol{r}|^2\rangle \quad (N-1)a^2 \tag{41}$$

where a denotes the length between two beads.

It follows that the probability density for the end-to-end vector $\boldsymbol{r}$ reads:

$$P(\boldsymbol{r}) \quad \int\delta\left(\boldsymbol{r}-\sum_{i\;\;1}^{N-1}a\boldsymbol{u}_i\right)\psi(\theta^{N-1},\varphi^{N-1})\,d\theta^{N-1}\,d\varphi^{N-1}, \tag{42}$$

where δ is formally a Dirac mass and $\boldsymbol{u}_i$ the unit vector of Euler angles (θ_i,φ_i). Using (39), a simple but somewhat tedious calculation shows that an adequate approximation formula for P, in the limit of a large number $N-2$ of beads eliminated, is

$$P(\boldsymbol{r}) \overset{N\text{large}}{\approx} \left(\frac{3}{2\pi(N-1)a^2}\right)^{3/2}\exp\left(-\frac{3|\boldsymbol{r}|^2}{2(N-1)a^2}\right). \tag{43}$$

The right-hand side of (43) is now chosen to be the probability law of $\boldsymbol{r}$, which is consequently a Gaussian variable. From now on, only the end-to-end vector, and its probability, are kept as the statistical description of the entire chain.

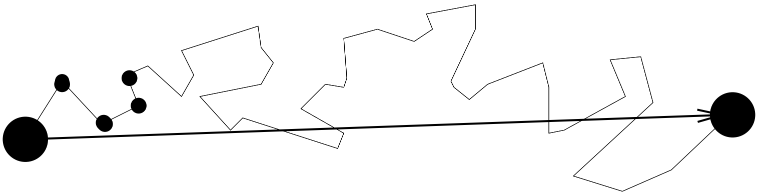


Fig. 7. A polymeric chain consisting of, say, thirty beads and its phenomenological representation as a dumbbell.

Remark 7. For some dedicated applications, chains with N_b 10 or N_b 20 beads are simulated. This is typically the case when one wants to model multiple relaxation time scales in the polymer chain or understand boundary effects. Consider a pipeline where the polymeric fluid flows. Some macroscopic model may provide good results for the inner flow, but they need to be supplied with adequate boundary conditions on the walls of the pipeline. Dumbbell models could be envisioned for this purpose, but since the complexity of the chain is a key issue for rheological properties near the boundaries, more sophisticated models with larger N_b have sometimes to be employed. Apart from such specific situations, it is considered that the dumbbell model already gives excellent answers. But this also depends upon the force fields that will be used. The purpose of the next section is exactly to introduce such a force. Others will be mentioned in Sect. 4.

The Hookean model

We now have our configuration space, namely that of a single end-to-end vector $\boldsymbol{r}$ equipped with a Gaussian probability at equilibrium. Let us next define the forces this end-to-end vector experiences.

We need to equip the vector $\boldsymbol{r}$ with some rigidity. Such a rigidity does not express a *mechanical* rigidity due to forces, of interatomic nature, holding between beads. It will rather model an *entropic* rigidity, related to the variations of the configurations of the actual entire chain when the end-to-end vector itself varies.

To understand this, let us only mention two extreme situations. If the end-to-end vector has length exactly $|\boldsymbol{r}| \quad (N-1)a$, there is one and only one configuration of the entire chain that corresponds to such an end-to-end vector, namely the chain fully extended as a straight line. In contrast, when the end-to-end vector has, say, length $|\boldsymbol{r}| \quad (N-1)a/2$, there is an enormous number of configurations, corresponding to various shapes of a chain of total length $(N-1)a$ that give rise to such an end-to-end vector. Entropy will thus favor short end-to-end vectors, rather than long ones. It remains now to quantitatively understand this.

We know from Statistical Mechanics arguments that for a system with probability law $P(\boldsymbol{r})$ (obtained from (43)), the free energy is given by

$$A(\boldsymbol{r}) \quad A^0 - kT \ln P(\boldsymbol{r})$$

where T denotes temperature, A^0 is a constant and k the Boltzmann constant. When the end-to-end vector is modified by $d\boldsymbol{r}$, the resulting free energy modification reads

$$\begin{aligned} dA &\quad -kT\,d\ln P(\boldsymbol{r}),\\ &\quad \frac{3kT}{(N-1)a^2}\,\boldsymbol{r}\cdot d\boldsymbol{r}. \end{aligned} \tag{44}$$

On the other hand, when temperature is kept constant, the free energy change is related to the tension $\boldsymbol{F}$ of the chain by

$$dA \quad \boldsymbol{F}(\boldsymbol{r})\cdot d\boldsymbol{r}. \tag{45}$$

Comparing (44) and (45), we obtain the tension

$$\boldsymbol{F}(\boldsymbol{r}) \quad \frac{3kT}{(N-1)a^2}\,\boldsymbol{r}. \tag{46}$$

In other words, the entropic force $\boldsymbol{F}$ expressed in terms of the end-to-end vector $\boldsymbol{r}$ is *defined* as the gradient of $\ln P$ with respect to $\boldsymbol{r}$ where P is the probability density of the end-to-end vector at equilibrium (zero velocity field for the surrounding solvent, and fixed temperature). This definition of the entropic force is consistent with the fact that P is indeed a stationary solution for the dynamics that will be defined on the probability density ψ of the end-to-end vector (see Equation (47) below) when the velocity field in the solvent is zero (equilibrium situation).

The end-to-end vector therefore acts as a linear spring, with stiffness

$$H \quad \frac{3kT}{(N-1)a^2}.$$

The model obtained is called the *Hookean dumbbell model.*

The above derivation is the simplest possible one, based on oversimplifying assumptions. Several improvements of the Hookean force (46) are indeed possible. We prefer to postpone the presentation of such improvements until Sect. 4. Let us momentarily assume we have a force $\boldsymbol{F}(\boldsymbol{r})$ at hand, the prototypical example being the Hookean force (46), and proceed further. On purpose, we do not make precise the expression of $\boldsymbol{F}(\boldsymbol{r})$ in the sequel.

3.2 The multiscale model

Let us now denote $\psi(t,\boldsymbol{x},\boldsymbol{r})$ the probability density for the end-to-end vectors of the polymer chains at macropoint $\boldsymbol{x}$ and time t.

The variation of ψ in time, calculated along a fluid trajectory, that is $\dfrac{\partial\psi}{\partial t}+\boldsymbol{u}\cdot\nabla_{\boldsymbol{x}}\psi$, follows from three different phenomena:

1. a hydrodynamic force: the dumbbell is elongated or shortened because of the interaction with the fluid ; Its two ends do not necessarily share the same macroscopic velocity, the slight difference in velocities (basically $\nabla\boldsymbol{u}(t,\boldsymbol{x})\,\boldsymbol{r}$) results in a force elongating the dumbbell $\zeta\nabla\boldsymbol{u}(t,\boldsymbol{x})\,\boldsymbol{r}$ where ζ denotes a friction coefficient;

2. the entropic force $\boldsymbol{F}$ issued from the coarse-graining procedure and which is reminiscent of the actual, much more complex, geometry of the entire polymeric chain;
3. a Brownian force, modelling the permanent collisions of the polymeric chain by solvent molecules, which (randomly) modifies its evolution.

We have gone into many details about the second phenomenon (the entropic force) in the previous section. We will momentarily admit the modelling proposed above for the first and third phenomena (the hydrodynamic force and the Brownian force, respectively) and proceed further. In Sect. 4, we will return to this in more details, explaining the intimate nature of these forces and motivating their actual mathematical form by rigorous arguments.

The overall conservation of momentum equation reads as the following evolution equation on ψ:

$$\begin{aligned}\frac{\partial \psi(t,\boldsymbol{x},\boldsymbol{r})}{\partial t} &+ \boldsymbol{u}(t,\boldsymbol{x}) \cdot \nabla_{\boldsymbol{x}} \psi(t,\boldsymbol{x},\boldsymbol{r}) \\ & -\mathrm{div}_{\boldsymbol{r}} \left(\left(\nabla_{\boldsymbol{x}} \boldsymbol{u}(t,\boldsymbol{x})\, \boldsymbol{r} - \frac{2}{\zeta} \boldsymbol{F}(\boldsymbol{r}) \right) \psi(t,\boldsymbol{x},\boldsymbol{r}) \right) + \frac{2kT}{\zeta} \Delta_{\boldsymbol{r}} \psi(t,\boldsymbol{x},\boldsymbol{r}).\end{aligned} \tag{47}$$

Equation (47) is called a *Fokker-Planck equation* (or also a *forward Kolmogorov equation*). The three terms of the right-hand side of (47) respectively correspond to the three phenomena listed above, in this order. A crucial point to make is that, in this right-hand side, all differential operators acting on ψ are related to the variable $\boldsymbol{r}$ of the configuration space, *not* of the ambient physical space. In contrast, the gradient of the left-hand side is the usual transport term in the physical space $\boldsymbol{u} \cdot \nabla_{\boldsymbol{x}}$. In the absence of such a transport term (this will indeed be the case for extremely simple geometries, such as that of a Couette flow), (47) is simply a family of Fokker-Planck equation posed in variables $(t,\boldsymbol{r})$ and *parameterized* in $\boldsymbol{x}$. These equations only speak to one another through the macroscopic field $\boldsymbol{u}$. When the transport term is present, (47) is a genuine partial differential equation in all variables $(t,\boldsymbol{x},\boldsymbol{r})$. It is intuitively clear that the latter case is much more difficult, computationally and mathematically.

Once ψ is obtained, we need to formalize its contribution to the total stress, and, further, its impact on the macroscopic flow.

Let us return to some basics of continuum mechanics. When defining the stress tensor, the commonly used mental image is the following: consider the material, cut it by a planar section into two pieces, try and separate the pieces. The reaction force experienced when separating the two pieces is $\boldsymbol{\tau n}$, where $\boldsymbol{\tau}$ is the stress tensor and $\boldsymbol{n}$ the unit vector normal to the cut plane. Varying the orientation of the cut planes, and thus $\boldsymbol{n}$, provides all the entries of $\boldsymbol{\tau}$. Applying the same 'methodology' to the polymeric fluid under consideration gives rise to two contributions (see Fig. 8): that, usually considered, of the solvent, which contributes as the usual Newtonian stress tensor, and that coming from all the polymeric chains reacting. The latter needs to be evaluated quantitatively. This is the purpose of the so-called *Kramers formula*.

$$\boldsymbol{\tau}_p(t,\boldsymbol{x}) \quad -n_p kT \mathrm{Id} + n_p \int (\boldsymbol{r} \otimes \boldsymbol{F}(\boldsymbol{r}))\, \psi(t,\boldsymbol{x},\boldsymbol{r})\, d\boldsymbol{r}, \tag{48}$$

where $\otimes$ denotes the tensor product ($\boldsymbol{r}\otimes\boldsymbol{F}(\boldsymbol{r})$ is a matrix with (i,j)-component $\boldsymbol{r}_i\boldsymbol{F}_j(\boldsymbol{r})$) and n_p denotes the total number of polymeric chain per unit volume. Note that the first term only changes the pressure by an additive constant.

The complete system of equation combines the equation of conservation of momentum at the macroscopic level, the incompressibility constraint, the Kramers formula, and the Fokker-Planck equation for the distribution of the end-to-end vector:

$$\begin{cases} \rho\left(\dfrac{\partial\boldsymbol{u}}{\partial t}+(\boldsymbol{u}\cdot\nabla)\boldsymbol{u}\right)-\eta\Delta\boldsymbol{u}+\nabla p-\operatorname{div}\boldsymbol{\tau}_p \quad \rho\boldsymbol{f}, \\ \operatorname{div}\boldsymbol{u} \quad 0, \\ \boldsymbol{\tau}_p(t,\boldsymbol{x}) \quad n_p\displaystyle\int(\boldsymbol{r}\otimes\boldsymbol{F}(\boldsymbol{r}))\,\psi(t,\boldsymbol{x},\boldsymbol{r})\,d\boldsymbol{r}-n_pkT\mathrm{Id}, \\ \dfrac{\partial\psi(t,\boldsymbol{x},\boldsymbol{r})}{\partial t}+\boldsymbol{u}(t,\boldsymbol{x})\cdot\nabla_{\boldsymbol{x}}\psi(t,\boldsymbol{x},\boldsymbol{r}) \\ \qquad -\operatorname{div}_{\boldsymbol{r}}\left(\left(\nabla_{\boldsymbol{x}}\boldsymbol{u}(t,\boldsymbol{x})\boldsymbol{r}-\dfrac{2}{\zeta}\boldsymbol{F}(\boldsymbol{r})\right)\psi(t,\boldsymbol{x},\boldsymbol{r})\right)+\dfrac{2kT}{\zeta}\Delta_{\boldsymbol{r}}\psi(t,\boldsymbol{x},\boldsymbol{r}). \end{cases} \tag{49}$$

For future reference, let us rewrite this system of equations in a non-dimensional form (see Sect. 4.3 and (97) for the derivation of the non-dimensional equations and the definition of the non-dimensional numbers Re, ε and We):

$$\boxed{\begin{cases} \mathrm{Re}\left(\dfrac{\partial\boldsymbol{u}}{\partial t}+(\boldsymbol{u}\cdot\nabla)\boldsymbol{u}\right)-(1-\varepsilon)\Delta\boldsymbol{u}+\nabla p-\operatorname{div}\boldsymbol{\tau}_p \quad \boldsymbol{f}, \\ \operatorname{div}\boldsymbol{u} \quad 0, \\ \boldsymbol{\tau}_p(t,\boldsymbol{x}) \quad \dfrac{\varepsilon}{\mathrm{We}}\left(\displaystyle\int(\boldsymbol{r}\otimes\boldsymbol{F}(\boldsymbol{r}))\,\psi(t,\boldsymbol{x},\boldsymbol{r})\,d\boldsymbol{r}-\mathrm{Id}\right), \\ \dfrac{\partial\psi(t,\boldsymbol{x},\boldsymbol{r})}{\partial t}+\boldsymbol{u}(t,\boldsymbol{x})\cdot\nabla_{\boldsymbol{x}}\psi(t,\boldsymbol{x},\boldsymbol{r}) \\ \qquad -\operatorname{div}_{\boldsymbol{r}}\left(\left(\nabla_{\boldsymbol{x}}\boldsymbol{u}(t,\boldsymbol{x})\boldsymbol{r}-\dfrac{1}{2\mathrm{We}}\boldsymbol{F}(\boldsymbol{r})\right)\psi(t,\boldsymbol{x},\boldsymbol{r})\right)+\dfrac{1}{2\mathrm{We}}\Delta_{\boldsymbol{r}}\psi(t,\boldsymbol{x},\boldsymbol{r}). \end{cases}} \tag{50}$$

The multiscale nature of this system is obvious. In the specific context of complex fluids, such a system is called a *micro-macro model*. It is equally obvious on (50) that the computational task will be demanding. Formally, system (50) couples a Navier-Stokes type equation (that is, an equation the simulation of which is one of the major challenges of scientific computing, and has been the topic of thousands of years of researchers efforts), and, at each point (that is, slightly anticipating the discretization, at each node of the mesh used for the space discretization of the macroequation), one parabolic partial differential equation set on the space of $\boldsymbol{r}$. It is thus intuitively clear that, in nature, such a micromacro strategy will be limited to as simple as possible test cases. We will return to this later.

With a view to generalizing the approach followed above to various other contexts, it is interesting to write system (50) as a particular form of a more abstract system. A purely macroscopic description of non-Newtonian fluids, issued from equations of the type (13)–(15) typically reads:

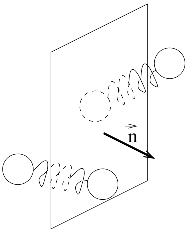


Fig. 8. Kramer Formula : the contribution of all polymeric chains to the stress is obtained summing over all chains cut by the plane considered.

$$\left\{\begin{array}{ll}\dfrac{D\boldsymbol{u}}{Dt} & \mathscr{F}(\boldsymbol{\tau}_p,\boldsymbol{u}),\\[2ex] \dfrac{D\boldsymbol{\tau}_p}{Dt} & \mathscr{G}(\boldsymbol{\tau}_p,\boldsymbol{u}).\end{array}\right. \tag{51}$$

In contrast, a multiscale approach introduces an additional intermediate step, where the stress tensor is calculated as an average value of a field Σ describing the microstructure. An evolution equation is written on the latter:

$$\left\{\begin{array}{ll}\dfrac{D\boldsymbol{u}}{Dt} & \mathscr{F}(\boldsymbol{\tau}_p,\boldsymbol{u}),\\[2ex] \boldsymbol{\tau}_p & \text{average over } \Sigma,\\[2ex] \dfrac{D\Sigma}{Dt} & \mathscr{G}_\mu(\Sigma,\boldsymbol{u}).\end{array}\right. \tag{52}$$

The structure of system (52) is a common denominator to all multiscale models for complex fluids. Beyond this, it also illustrates the nature of all multiscale approaches, in very different contexts (see C. Le Bris [77]). A global macroscopic equation is coupled with a local (microscopic) equation, via an averaging formula. For instance, the reader familiar with homogenization theory for materials recognizes in (52) the homogenized equation, the value of the homogenized tensor, and the corrector equation, respectively. On the numerical front, it is also a structure shared with multiscale algorithmic approaches: a global coarse solver coupled to a local fine one using an averaging process (think of the Godunov scheme for solving the Riemann problem in computational fluid dynamics).

4 The stochastic approach

We now need to complement the derivation of the previous section in three directions:

- We need to introduce a definite stochastic description of the polymeric chain that will justify the expressions employed for the elongation force and the Brownian force in (47).

- We need to provide other entropic forces, alternative to the simple Hookean force (46).
- We need to prepare for an efficient computational strategy that allows for the practical simulation of systems of the type (50) even when the configuration space for the chain is high-dimensional.

For all three aspects, stochastic analysis comes into play. This is why we devote the next section to a brief introduction of the major ingredients from stochastic analysis needed in the sequel. Such ingredients are traditionally not necessarily well known by readers more familiar with the classical analysis of partial differential equations and their discretization techniques in the engineering sciences (finite element methods, etc. . .). Of course, the reader already familiar with the basics of stochastic analysis may easily skip the next section and directly proceed to Sect. 4.2. In any event, Subsection 4.1 is no more than a surrogate for a more comprehensive course of Stochastic Analysis, as contained in the classical textbooks F. Comets and T. Meyre [26], I. Karatzas and S.E. Shreve [69], P.E. Kloeden and E. Platen [73], B. Øksendal [101], L.C.G. Rogers and D. Williams [114, 115], D. Revuz and M. Yor [113], D. Stroock and S.R.S. Varadhan [117]. We also refer to D.J. Higham [58] for an attractive practical initiation.

4.1 Initiation to Stochastic Differential Equations

We assume that the reader is familiar with the following elementary notions of Probability Theory: the notion of *probability space* $(\Omega,\mathscr{A},\mathbb{P})$, where Ω is the space, $\mathscr{A}$ is a σ-algebra, and $\mathbb{P}$ is the probability measure that equips the space; the notion of vector-valued or scalar-valued random variables defined on this probability space; the notion of expectation value and the notion of law.

A rather abstract notion we must define before getting to the heart of the matter is the notion of *filtration*: a *filtration* $(\mathscr{F}_t, t \geq 0)$ is an increasing sequence, indexed by time $t \in \mathbb{R}_+$, of subsets of the σ-algebra $\mathscr{A}$. The filtration $\mathscr{F}_t$ is to be thought of as the set of information available at time t.

The Monte Carlo method

The Monte Carlo method is a stochastic method to compute the expectation value of a random variable. Let X be a random variable with finite variance:

$$\mathrm{Var}(X) \quad \mathbb{E}\big((X-\mathbb{E}(X))^2\big) \quad \mathbb{E}(X^2)-(\mathbb{E}(X))^2 < \infty.$$

The principle of the Monte Carlo method is to approximate the expectation value $\mathbb{E}(X)$ by the *empirical mean*

$$I_K \quad \frac{1}{K}\sum_{k \ 1}^{K} X^k,$$

where $(X^k)_{k\geq 0}$ are independent identically distributed (i.i.d.) random variables, the law of X^k being the the law of X.

The foundation of the Monte Carlo method is based on two mathematical results. The *law of large numbers* states that (if $\mathbb{E}|X| < \infty$)

$$\text{almost surely, } \lim_{K\longrightarrow\infty} I_K \quad \mathbb{E}(X).$$

The *central limit theorem* gives the rate of convergence (if $\mathbb{E}((X)^2) < \infty$): $\forall a > 0$,

$$\lim_{K\longrightarrow\infty} \mathbb{P}\left(|I_K - \mathbb{E}(X)| \leq a\sqrt{\frac{\mathrm{Var}(X)}{K}} \right) \quad \frac{1}{\sqrt{2\pi}} \int_{-a}^{a} \exp(-x^2/2)\,dx.$$

This estimate enables to build *a posteriori* error estimates (so-called confidence intervals) by choosing typically $a \quad 1.96$ so that $\frac{1}{\sqrt{2\pi}} \int_{-a}^{a} \exp(-x^2/2)\,dx \simeq 95\%$, and estimating the variance $\mathrm{Var}(X)$ by the *empirical variance*

$$V_K \quad \frac{1}{K} \sum_{k\ 1}^{K} (X^k)^2 - (I_K)^2.$$

This estimate shows that the rate of convergence of a Monte Carlo method is of order $\sqrt{\frac{\mathrm{Var}(X)}{K}}$: to reduce the error, one needs to add more replicas (increase K), or reduce the variance of the random variable (which is the basis of *variance reduction* methods, see Sect. 5.4 below).

Stochastic processes, Brownian motion and simple stochastic differential equations

Let us now introduce the notion of a (continuous-in-time) *stochastic process*, as a family of random variables $(X_t)_{t\geq 0}$ indexed by time $t \in \mathbb{R}_+$. Given a stochastic process X_t, we may consider the *natural filtration* generated by X_t, that is the filtration $\mathscr{F}_t$ formed, for each t, by the smallest σ-algebra for which the maps $\omega \longrightarrow X_s(\omega)$, $0 \leq s \leq t$, are measurable functions.

Conversely, being given a filtration $\mathscr{F}_t$, a stochastic process such that, for all t, X_t is a measurable function with respect to $\mathscr{F}_t$, is called a $\mathscr{F}_t$*-adapted* stochastic process.

A remarkable random process is the Brownian motion, which we now briefly introduce.

The formal motivation for the introduction of the Brownian motion is the need for modelling random trajectories. For such trajectories, the random perturbations at time t should be independent of those at time $t' < t$, and essentially the same. By this we mean that the two should share the same law. The mathematical manner to formalize the above somewhat vague object is the notion of Brownian motion. There are several ways to define a Brownian motion. One way is to take the limit of random walks on lattices, with an adequate scaling law on the size of the lattice and time. The definition we choose to give here is the axiomatic definition. We define a *Brownian motion* as a real-valued random process enjoying the following three properties. First, its trajectories, that is, the maps $s \longrightarrow X_s(\omega)$ are continuous, for almost all $\omega \in \Omega$. Second, it has *independent increments*, that is, when $s \leq t$, the random variable $X_t - X_s$ is independent of the σ-algebra $\mathscr{F}_s$: otherwise stated, for all $A \in \mathscr{F}_s$, and all bounded measurable function f, $\mathbb{E}(1_A f(X_t - X_s)) \quad \mathbb{E}(f(X_t - X_s))\mathbb{P}(A)$. Third, it has *stationary increments*, that is when $s \leq t$, $X_t - X_s$ and $X_{t-s} - X_0$ share the same law. In fact, the conjunction of these three properties implies that, necessarily, $X_t - X_0$ is a Gaussian variable, with mean rt (for some r) and variance $\sigma^2 t$ (for some σ). When $r \quad 0$ and $\sigma \quad 1$, the Brownian motion is called a *standard* Brownian motion.

We now wish to define differential equations, typically modelling the motion of particles, which are subject to random perturbations. The adequate mathematical notion for this purpose is that of *stochastic differential equations*. Let us fix a probability space $(\Omega, \mathscr{A}, \mathbb{P})$, where it is sometimes useful to think of Ω as the product $\Omega \quad \Omega_1 \times \Omega_2$ where Ω_1 models the randomness due to the initial condition supplied for the differential equation, and Ω_2 models the randomness associated with the perturbations occurring at all positive times.

Let us also consider a filtration $\mathscr{F}_t$ and a $\mathscr{F}_t$-adapted Brownian motion B_t. Let $\sigma > 0$ denote a fixed parameter, called *diffusion*, and $b(t,x)$ a fixed regular function, called *drift*. As regards regularity issues, the most appropriate setting is to consider functions b measurable with respect to time t, Lipschitz with respect to the space variable x, and with a growth at most linear at infinity, that is $|f(t,x)| \leq C(1+|x|)$ for all t, x. For simplicity, the Lipschitz constant and the growth constant are assumed uniform on $t \in \ 0, T$. We then define the *stochastic differential equation*:

$$dX_t \quad b(t, X_t)\,dt + \sigma\, dB_t, \tag{53}$$

with initial condition $X_0(\omega_1)$. Equation (53) is formal. It is to be understood in the following sense: X_t is said a solution to (53) when

$$X_t(\omega_1, \omega_2) \quad X_0(\omega_1) + \int_0^t b(s, X_s(\omega_1, \omega_2))\,ds + \sigma B_t(\omega_2), \tag{54}$$

almost surely. Our setting in (53)–(54) is one dimensional, but the notion is readily extended to the higher dimensional context (see (64) below).

Note that we do not question here the existence and the uniqueness of the solutions to the above stochastic differential equations. This is beyond the scope of this simplified presentation. Let us only say that we assume for the rest of this expository survey that typically the Lipschitz regularity mentioned above is sufficient to define in a unique manner the solution to (53). For less regular drifts and related questions, we refer the interested reader to Sect. 6. The modelling of complex fluids may indeed naturally involve non-regular drifts.

Stochastic integration

The above form (53) is actually a simple form of a stochastic differential equations. This form is sufficient to deal with the context of *flexible polymers*, which is the main topic of this presentation. However, for rigid polymers, briefly addressed in Sect. 7, and some other types of complex fluids, it is useful to define the general form of stochastic differential equations. This is the purpose of this short section.

In addition, the consideration of this general form of stochastic differential equation will allow us to introduce a technical lemma which will be crucially useful, even in our simple setting.

Using a standard Brownian motion B_t, the *Itô integral* may be constructed. The construction of this notion of integral is similar to that of the Riemann integral, proceeding first for piecewise constant functions, and then generalizing the notion to more general functions by approximation. Consider a decomposition $\{s_0 \quad 0, ..., s_j, ..., s_n \quad t\}$ of the range $0,t$ and a piecewise constant process

$$Y_s(\omega) \quad \sum_{j \;\; 1}^{n} \tilde{Y}_{j-1}(\omega)\, 1_{s_{j-1},s_j}(s)$$

constructed from random variables $\tilde{Y}_j$ (such that $\mathbb{E}(|\tilde{Y}_j|) < +\infty$ and $\tilde{Y}_j$ is $\mathscr{F}_{s_j}$-measurable). Then we define

$$\int_0^t Y_s \, dB_s \quad \sum_{j \;\; 1}^{n} \tilde{Y}_{j-1} \left(B_{s_j} - B_{s_{j-1}}\right). \tag{55}$$

Next, for any arbitrary $\mathscr{F}_t$-adapted stochastic process $Y_t(\omega)$ such that, almost surely, $\int_0^T Y_t(\omega)^2 \, dt < +\infty$, this allows, by approximation, for the definition of the stochastic process

$$\int_0^t Y_s \, dB_s.$$

In the simple case when $Y_t \equiv 1$, this coincides with the already known notion $\int_0^t dB_s \quad B_t$. Notice that by taking the expectation of (55), we have, for all $t \in \; 0,T$

$$\mathbb{E}\left(\int_0^t Y_s \, dB_s\right) \quad 0, \tag{56}$$

which actually holds (by an approximation argument) for any arbitrary stochastic process Y_t such that $\mathbb{E}\left(\int_0^T Y_t(\omega)^2 \, dt\right) < +\infty$.

Having defined the Itô integral, we are in position, for any regular drift b and diffusion σ, to define the *stochastic differential equation*:

$$dX_t \quad b(t,X_t)\, dt + \sigma(t,X_t)\, dB_t, \tag{57}$$

supplied with the initial condition X_0. Mathematically:

$$X_t(\omega_1,\omega_2) \quad X_0(\omega_1) + \int_0^t b(s,X_s(\omega_1,\omega_2))\, ds + \left(\int_0^t \sigma(s,X_s)\, dB_s\right)(\omega_1,\omega_2), \tag{58}$$

almost surely. In the right-hand side, the first integral is the Lebesgue integral, the second one is a Itô integral.

Itô calculus and Fokker-Planck equation

We now wish to relate the above stochastic differential equation (57) with a partial differential equation. The latter is indeed the Fokker-Planck equation

$$\frac{\partial p}{\partial t}(t,x) + \frac{\partial}{\partial x}(b(t,x)\,p(t,x)) - \frac{\partial^2}{\partial x^2}\left(\frac{\sigma^2(t,x)}{2}\,p(t,x)\right) \quad 0. \tag{59}$$

In the context of deterministic equations, the reader is perhaps familiar with the intimate link between *ordinary* differential equations and *linear transport equations*. This is the famous method of characteristics, which we briefly recall here. Consider the linear transport equation

$$\frac{\partial u}{\partial t}(t,x) - b(x)\frac{\partial u}{\partial x}(t,x) \quad 0 \tag{60}$$

supplied with the initial condition u_0 at initial time. Its solution reads

$$u(t,x) \quad u_0(X(t;0,x)) \tag{61}$$

where $X(t;0,x)$ is the solution at time t of the ordinary differential equation

$$\frac{dX(t)}{dt} \quad b(X(t)) \tag{62}$$

starting from the initial condition $X(0) \quad x$. The proof of this fact is elementary. For $s \in \; 0,t$, we have (where $X(s) \quad X(s;0,x)$)

$$\begin{aligned}\frac{\partial}{\partial s}(u(t-s,X(s))) \quad & -\frac{\partial u}{\partial t}(t-s,X(s)) + \frac{dX}{dt}(s)\frac{\partial u}{\partial x}(t-s,X(s)),\\ & -\frac{\partial u}{\partial t}(t-s,X(s)) + b(X(s))\frac{\partial u}{\partial x}(t-s,X(s)) \quad 0.\end{aligned}$$

By integrating this relation from $s \quad 0$ to $s \quad t$, we thus obtain (61).

A similar type of argument, based on the so-called *Feynman-Kac Formula* would show the relation holding between the stochastic differential equation (57) and a partial differential equation, called the backward Kolmogorov equation. A dual viewpoint to the above one illustrates the relation between the stochastic differential equation (57) and the Fokker-Planck equation (59). We now present it.

First, we need to establish a chain rule formula in the context of stochastic processes. This is the purpose of the celebrated *Itô formula* (stated here in a simple one-dimensional setting).

Lemma 1. Itô Formula *Let X_t solve*

$$dX_t \quad b(t,X_t)\,dt + \sigma(t,X_t)\,dB_t,$$

in the sense of (58). Then, for all C^2 regular function φ,

$$d\varphi(X_t) \quad \left(\varphi'(X_t)b(t,X_t) + \frac{1}{2}\,\varphi''(X_t)\,\sigma(t,X_t)^2\right) dt + \varphi'(X_t)\,\sigma(t,X_t)\,dB_t$$

in the sense

$$\begin{aligned}\varphi(X_t) \quad & \varphi(X_0) + \int_0^t \left(\varphi'(X_s)b(s,X_s) + \frac{1}{2}\,\varphi''(X_s)\,\sigma(s,X_s)^2\right) ds \\ & + \int_0^t \varphi'(X_s)\,\sigma(s,X_s)\,dB_s. \end{aligned} \tag{63}$$

The point is of course to compare with the deterministic setting, corresponding to $\sigma \quad 0$, and for which no second derivatives of φ appears since

$$\frac{d}{dt}\varphi(X_t) \quad \varphi'(X_t)\frac{dX_t}{dt}.$$

We are now in position to relate (57) and (59). Assume that all conditions of regularity are satisfied, which gives sense to the formal manipulations we now perform. Let us assume that X_0, the initial condition for (57) has law p_0, where p_0 is the initial condition given to (59). Let us denote by $p(t,x)$ the probability density (with respect to the Lebesgue measure) of the random variable X_t.

For any arbitrary C^2 function φ, we write

$$\int \varphi(x)\frac{\partial p}{\partial t}(t,x)\,dx \quad \frac{d}{dt}\int \varphi(x)p(t,x)\,dx \quad \frac{d}{dt}\mathbb{E}\left(\varphi(X_t)\right).$$

Now, taking the expectation of (63), we obtain

$$\begin{aligned}\mathbb{E}\left(\varphi(X_t)\right) \quad & \mathbb{E}\left(\varphi(X_0)\right) + \mathbb{E}\left(\int_0^t \left(\varphi'(X_s)b(s,X_s) + \frac{1}{2}\,\varphi''(X_s)\,\sigma(s,X_s)^2\right) ds\right) \\ & + \mathbb{E}\left(\int_0^t \varphi'(X_s)\,\sigma(s,X_s)\,dB_s\right). \end{aligned}$$

Under suitable regularity assumptions, the last term vanishes for all times (see (56)). We thus have

$$\begin{aligned}\int \varphi(x)\frac{\partial p}{\partial t}(t,x)\,dx \quad & \mathbb{E}\left(\varphi'(X_t)b(t,X_t) + \frac{1}{2}\varphi''(X_t)\sigma(t,X_t)^2\right), \\ & \int \left(\varphi'(x)b(t,x) + \frac{1}{2}\varphi''(x)\sigma^2(t,x)\right) p(t,x)\,dx, \\ & \int \varphi(x)\left(-\frac{\partial}{\partial x}(pb)(t,x) + \frac{1}{2}\frac{\partial^2}{\partial x^2}(\sigma^2 p)\right) dx. \end{aligned}$$

This precisely shows that p is the solution to (59), which starts from p_0 at initial time.

A similar argument, based on the multi-dimensional Itô Formula (a straightforward extension of Lemma 1), allows to establish the same correspondence between, on the one-hand, the *vectorial* stochastic differential equation

$$d\boldsymbol{X}_t \quad \boldsymbol{b}(t,\boldsymbol{X}_t)\,dt + \boldsymbol{\sigma}(t,\boldsymbol{X}_t)\,d\boldsymbol{B}_t, \tag{64}$$

where $\boldsymbol{X}_t$ is a random process with values in $\mathbb{R}^N$, $\boldsymbol{b}(t,\cdot)$ is a vector field on $\mathbb{R}^N$ for all times, σ is $N \times K$ matrix valued function, and $\boldsymbol{B}_t$ is a K-dimensional Brownian motion, and, on the other hand, the Fokker-Planck equation.

$$\frac{\partial p}{\partial t}(t,\boldsymbol{x}) + \mathrm{div}\,((\boldsymbol{b}(t,\boldsymbol{x})\,p(t,\boldsymbol{x})) - \frac{1}{2}\nabla^2 : \left(\boldsymbol{\sigma}\boldsymbol{\sigma}^T p\right)(t,\boldsymbol{x}) \quad 0, \tag{65}$$

where $\nabla^2 : \left(\boldsymbol{\sigma}\boldsymbol{\sigma}^T p\right) \quad \sum_{i,j\ 1}^{N} \frac{\partial^2}{\partial x_i \partial x_j}\left(\sum_{k\ 1}^{K} \sigma_{i,k}\sigma_{j,k}\,p\right).$

Under appropriate conditions of regularity (which we have omitted to make precise above), we may therefore claim that the law of any process solving the stochastic differential equation solves the Fokker-Planck equation. The converse assertion is false. Let us give the following simple illustration. Consider the stochastic differential equation

$$dX_t \quad -\frac{1}{2}X_t\,dt + dB_t, \tag{66}$$

with initial condition X_0 normally distributed with zero mean and variance one (and independent of B_t), and the associated Fokker-Planck equation

$$\frac{\partial p(t,x)}{\partial t} - \frac{1}{2}\frac{\partial}{\partial x}(x\,p(t,x)) - \frac{1}{2}\frac{\partial^2}{\partial x^2}p(t,x) \quad 0. \tag{67}$$

Clearly, the solution to (66) reads

$$X_t \quad e^{-t/2}X_0 + \int_0^t e^{(s-t)/2}dB_s.$$

Therefore, for all $t \geq 0$, X_t is a Gaussian random variable with zero mean and variance one and of course, as the previous argument shows, $p(t,x) \quad \frac{1}{\sqrt{2\pi}}\exp(-x^2/2)$ indeed solves the Fokker-Planck equation (67). However, any random process Y_t such that its marginals in time (namely the law of Y_t, for fixed t) are normally distributed with zero mean and variance one, such as the *constant* process $Y_t \quad X_0$, does *not* solve (66). The process encodes more information than the law of the time marginals, and it is thus intuitively clear that the knowledge of the law of the time marginals is not sufficient to know *the trajectory* of the process. Otherwise stated, knowing the law of the time marginals allows to compute all expectation values of the type $\mathbb{E}(\varphi(X_t))$, but, *e.g.*, not quantities such as $\mathbb{E}(\psi(X_t,X_s))$.

Nevertheless, for most situations of interest, and in particular for many physically relevant situations, only the knowledge of expectation values such as $\mathbb{E}(\varphi(X_t))$ is sufficient. In such situations, solving the Fokker-Planck equation, *when it is practically feasible*, provides all the information needed. In our context of the modelling of complex fluids, we can therefore equivalently use the stochastic differential viewpoint, or the Fokker-Planck viewpoint. Efficiency considerations indicate which is the best strategy, depending on the dimension of the problem at hand, and other parameters. We will return to this below.

Discretization of SDEs

We now briefly give here some basic elements of numerical analysis for stochastic differential equations. For this purpose, we assume that the reader is familiar with the discretization techniques for *ordinary* differential equations and the associated analysis (see E. Hairer, S.P. Nørsett and G. Wanner [55, 56]).

For simplicity, we argue on the one-dimensional simple case (53), that is

$$dX_t \quad b(t,X_t)\,dt + \sigma\, dB_t,$$

with in addition, a constant σ. We leave aside questions related to the general case $dX_t \quad b(t,X_t)\,dt + \sigma(t,X_t)\,dB_t$, which, owing to the dependence of σ upon the solution X_t, might be significantly more technical than the simple case considered here (see Remark 8 below). Likewise, we assume that b is regular and that all questions of existence and uniqueness have been settled.

The crucial point to bear in mind is that, in contrast to the deterministic setting, there are *two* notions of convergence for a scheme discretizing a stochastic differential equation.

The notion of convergence analogous to the deterministic notion is:

Definition 1. *The numerical scheme is said* strongly convergent *and is said to have* strong order of convergence $\alpha > 0$ *when there exists a constant C, possibly depending on the interval of integration* $0,T$, *such that, for all timesteps* Δt *and for all integer* $n \in \ 0, T/\Delta t$,

$$\mathbb{E}\left(\left|\overline{X}_n - X_{t_n}\right|\right) \le C\,(\Delta t)^{\alpha}, \tag{68}$$

where X_{t_n} *denotes the exact solution at time* $t_n \quad n\Delta t$, *and* $\overline{X}_n$ *denotes its numerical approximation.*

A weaker notion, which is a better metric to assess convergence in practical situations, is:

Definition 2. *Under the same conditions as the above definition, the scheme is said* weakly convergent *and is said to have* weak order of convergence $\beta > 0$ *when for all integer* $n \in \ 0, T/\Delta t$,

$$\left|\mathbb{E}\left(\varphi(\overline{X}_n)\right) - \mathbb{E}\left(\varphi(X_{t_n})\right)\right| \le C\,(\Delta t)^{\beta}, \tag{69}$$

for all C^∞ *function* φ, *with polynomial growth at infinity, and such that all its derivatives also have polynomial growth at infinity.*

The latter definition, specific to the stochastic setting, is motivated by the fact that in many applications, as already mentioned above, the stochastic differential equation is simulated only to evaluate some expectation values $\mathbb{E}\left(\varphi(X_t)\right)$. This will be the case for complex fluid flows simulation (see the expression (82) of the stress tensor below). The notion of weak convergence is tailored for this purpose. In contrast to the strong convergence, it does not measure the accuracy of the approximation of each realization (each "trajectory") (note indeed that $\mathbb{P}(|\overline{X}_n - X_{t_n}| \geq a) \leq \dfrac{1}{a}\,\mathbb{E}(|\overline{X}_n - X_{t_n}|)$), but only the accuracy of the mean. Of course, strong convergence clearly implies weak convergence.

Let us now mention the simplest possible scheme for the numerical integration of (57). It is the *forward* (or *explicit*) *Euler scheme*:

$$\overline{X}_{n+1} \quad \overline{X}_n + b(t_n, \overline{X}_n)\,\Delta t + \sigma\,(B_{t_{n+1}} - B_{t_n}). \tag{70}$$

Since the increment $B_{t_{n+1}} - B_{t_n}$ is a centered Gaussian random variable with variance $t_{n+1} - t_n \quad \Delta t$, the scheme also writes

$$\overline{X}_{n+1} \quad \overline{X}_n + b(t_n, \overline{X}_n)\,\Delta t + \sigma\sqrt{\Delta t}\,G_n, \tag{71}$$

where $(G_n)_{n\geq 0}$ denote i.i.d. standard normal random variables.

It is easy to see that the scheme (70) arises from the approximation

$$\begin{aligned} X_{t_{n+1}} - X_{t_n} &\quad \int_{t_n}^{t_{n+1}} b(t, X_t)\,dt + \sigma \int_{t_n}^{t_{n+1}} dB_t, \\ &\approx b(t_n, X_{t_n})\,\Delta t + \sigma\,(B_{t_{n+1}} - B_{t_n}). \end{aligned}$$

The second integration in the right-hand side being exact, the precision order is exactly that of the approximation of the Lebesgue integral, and is therefore $\alpha \quad 1$. This is the strong order of convergence, and we leave to the reader the task to check that this is also the weak order of convergence.

Remark 8. Actually, the above argument is slightly misleading. It is specific to the case of a constant diffusion σ as in (53) or, more appropriately stated, to a deterministic diffusion σ that may depend on time, but that does not depend on the solution X_t. When the latter depends on the solution, that is

$$dX_t \quad b(t, X_t)\,dt \,+\, \sigma(X_t)\,dB_t,$$

then the Euler scheme

$$\overline{X}_{n+1} \quad \overline{X}_n + b(t_n, \overline{X}_n)\,\Delta t + \sigma(\overline{X}_n)\,(B_{t_{n+1}} - B_{t_n}) \tag{72}$$

(actually also called the *Euler-Maruyama scheme*) is only of strong order $\alpha \quad 1/2$, but it remains of weak order $\beta \quad 1$. The reason lies in the difference between the Itô calculus and the usual deterministic calculus. In fact, to obtain strong convergence with order 1, the adequate scheme to employ (at least for one-dimensional processes) is the *Euler-Milstein scheme*:

$$\begin{aligned} \overline{X}_{n+1} - \overline{X}_n \quad & b(t_n, \overline{X}_n)\,\Delta t + \sigma(\overline{X}_n)\,(B_{t_{n+1}} - B_{t_n}) \\ & + \frac{1}{2}\,\sigma(\overline{X}_n)\sigma'(\overline{X}_n)\,\big((B_{t_{n+1}} - B_{t_n})^2 - \Delta t\big). \end{aligned} \tag{73}$$

It is of strong order of convergence $\alpha \quad 1$, and of course agrees with the Euler-Maruyama scheme when σ is independent of X_t.

4.2 Back to the modelling

Given the notions of the previous section, we are now in position to return to some key issues in the modelling step, briefly addressed earlier in Sect. 3. Our purpose there was only to concentrate on the multiscale problem, and reach as soon as possible a prototypical form of such a system. This has been performed with (50) at the price of some simplifications and shortcuts. Let us now take a more pedestrian approach to the problem, and dwell into some issues, based on our present mathematical knowledge of the stochastic formalism.

The microscopic equation of motion

Let us first concentrate on the two forces exerted by the solvent onto the chain, namely the friction force elongating the chain and the Brownian force modeling collisions.

For this purpose, we isolate one single bead, denote by m its mass, $\boldsymbol{V}_t$ its velocity, and write the following equation of motion, called the *Langevin equation*:

$$m d\boldsymbol{V}_t \quad -\zeta \boldsymbol{V}_t\, dt + D\, d\boldsymbol{B}_t, \tag{74}$$

where $\boldsymbol{B}_t$ denotes a standard, d-dimensional, Brownian motion and D a scalar parameter to be determined. The solution of (74) is a so-called *Ornstein-Uhlenbeck process*:

$$\boldsymbol{V}_t \quad \boldsymbol{V}_0 \exp\left(-\frac{\zeta}{m}t\right) + \frac{D}{m}\int_0^t \exp\left(-\frac{\zeta}{m}(t-s)\right) d\boldsymbol{B}_s,$$

where $\boldsymbol{V}_0$ is the initial velocity, assumed independent of $\boldsymbol{B}_t$. Consequently, $\boldsymbol{V}_t$ is a Gaussian process with mean

$$\mathbb{E}(\boldsymbol{V}_t) \quad \mathbb{E}(\boldsymbol{V}_0)\exp\left(-\frac{\zeta}{m}t\right),$$

and covariance matrix

$$\begin{aligned} &\mathbb{E}\left((\boldsymbol{V}_t - \mathbb{E}(\boldsymbol{V}_t)) \otimes (\boldsymbol{V}_t - \mathbb{E}(\boldsymbol{V}_t))\right) \\ &\quad \mathbb{E}\left((\boldsymbol{V}_0 - \mathbb{E}(\boldsymbol{V}_0)) \otimes (\boldsymbol{V}_0 - \mathbb{E}(\boldsymbol{V}_0))\right)\exp\left(-\frac{2\zeta}{m}t\right) \\ &\quad + \frac{D^2}{2\zeta m}\left(1 - \exp\left(-\frac{2\zeta}{m}t\right)\right)\mathrm{Id}. \end{aligned} \tag{75}$$

For the above derivation, we have assumed that the fluid is at rest. The process $\boldsymbol{V}_t$ is thus expected to be stationary, which imposes:

$$\begin{cases} \mathbb{E}(\boldsymbol{V}_t) \quad \mathbb{E}(\boldsymbol{V}_0) \quad 0, \\ \mathbb{E}\left(\boldsymbol{V}_t \otimes \boldsymbol{V}_t\right) \quad \mathbb{E}\left(\boldsymbol{V}_0 \otimes \boldsymbol{V}_0\right) \quad \dfrac{D^2}{2\zeta m}\mathrm{Id}. \end{cases} \tag{76}$$

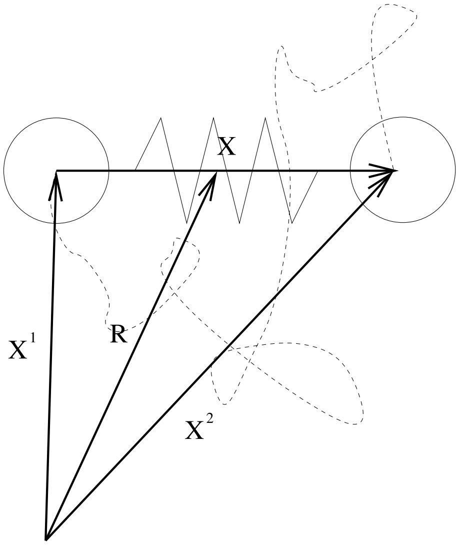


Fig. 9. The dumbbell model: the end-to-end vector $\boldsymbol{X}$ is the vector connecting the two beads, while $\boldsymbol{R}$ gives the position of the center of mass.

Using the principle of equipartition of energy, the mean kinetic energy $\frac{1}{2}m\mathbb{E}(\|\boldsymbol{V}_t\|^2)$ should be equal to $\frac{d}{2}kT$ (where d is the dimension of the ambient space) thus the *Nernst-Einstein relation*:

$$D \quad \sqrt{2kT\zeta}. \tag{77}$$

Let us next consider two beads, forming a dumbbell. We denote by $\boldsymbol{X}_t^i$ the (random) position of bead i, $i \quad 1,2$, and $\boldsymbol{X}_t \quad \boldsymbol{X}_t^2 - \boldsymbol{X}_t^1$ the end-to-end vector (see Fig. 9). We also denote $\boldsymbol{R} \quad \frac{1}{2}\left(\boldsymbol{X}^1 + \boldsymbol{X}^2\right)$ the position of the center of mass. In addition to the above two forces experienced by each of the beads, a force $\boldsymbol{F}(\boldsymbol{X}_t)$ of entropic nature is to be accounted for. We now know this well (see Sect. 3.1).

The Langevin equations for this simple two particle system reads:

$$\left\{\begin{array}{ll} m\,d\left(\dfrac{d\boldsymbol{X}_t^1}{dt}\right) & -\zeta\left(\dfrac{d\boldsymbol{X}_t^1}{dt} - \boldsymbol{u}(t,\boldsymbol{X}_t^1)\right)dt + \boldsymbol{F}(\boldsymbol{X}_t)\,dt + \sqrt{2kT\zeta}\,d\boldsymbol{B}_t^1, \\ m\,d\left(\dfrac{d\boldsymbol{X}_t^2}{dt}\right) & -\zeta\left(\dfrac{d\boldsymbol{X}_t^2}{dt} - \boldsymbol{u}(t,\boldsymbol{X}_t^2)\right)dt - \boldsymbol{F}(\boldsymbol{X}_t)\,dt + \sqrt{2kT\zeta}\,d\boldsymbol{B}_t^2, \end{array}\right. \tag{78}$$

where $\boldsymbol{B}_t^1$ and $\boldsymbol{B}_t^2$ are two independent, d-dimensional Brownian motions. In the limit of a vanishing $\frac{m}{\zeta}$, (that is when the characteristic timescale of relaxation to equilibrium for the end-to-end vector is far larger than this value), we obtain by linear combination of the above two Langevin equations:

$$\begin{cases} d\boldsymbol{X}_t \quad \left(\boldsymbol{u}(t,\boldsymbol{X}_t^2)-\boldsymbol{u}(t,\boldsymbol{X}_t^1)\right)dt-\dfrac{2}{\zeta}\boldsymbol{F}(\boldsymbol{X}_t)\,dt+2\sqrt{\dfrac{kT}{\zeta}}d\boldsymbol{W}_t^1, \\ d\boldsymbol{R}_t \quad \dfrac{1}{2}\left(\boldsymbol{u}(t,\boldsymbol{X}_t^1)+\boldsymbol{u}(t,\boldsymbol{X}_t^2)\right)dt+\sqrt{\dfrac{kT}{\zeta}}d\boldsymbol{W}_t^2, \end{cases} \tag{79}$$

where $\boldsymbol{W}_t^1 \quad \frac{1}{\sqrt{2}}\left(\boldsymbol{B}_t^2-\boldsymbol{B}_t^1\right)$ and $\boldsymbol{W}_t^2 \quad \frac{1}{\sqrt{2}}\left(\boldsymbol{B}_t^1+\boldsymbol{B}_t^2\right)$ are also two independent, d-dimensional Brownian motions. We assume they do not depend on space.

At this stage, the following assumptions are in order:

- as the length of the polymer is in any case far smaller than the spatial variations of the velocity of the solvent, we may perform the Taylor expansion

$$\boldsymbol{u}(t,\boldsymbol{X}_t^i)\simeq\boldsymbol{u}(t,\boldsymbol{R}_t)+\nabla\boldsymbol{u}(t,\boldsymbol{R}_t)(\boldsymbol{X}_t^i-\boldsymbol{R}_t)$$

  for $i \quad 1,2$,
- as $\frac{1}{2}\left(\boldsymbol{u}(t,\boldsymbol{X}_t^1)+\boldsymbol{u}(t,\boldsymbol{X}_t^2)\right)dt$ is of macroscopic size, in comparison to the microscopic variation $\sqrt{\frac{kT}{\zeta}}d\boldsymbol{W}_t^2$, the noise $\boldsymbol{W}_t^2 \quad 0$ is neglected.

Denoting by $\boldsymbol{W}_t \quad \boldsymbol{W}^1$, we obtain:

$$\begin{cases} d\boldsymbol{X}_t \quad \nabla\boldsymbol{u}(t,\boldsymbol{R}_t)\boldsymbol{X}_t\,dt-\dfrac{2}{\zeta}\boldsymbol{F}(\boldsymbol{X}_t)\,dt+2\sqrt{\dfrac{kT}{\zeta}}d\boldsymbol{W}_t, \\ d\boldsymbol{R}_t \quad \boldsymbol{u}(t,\boldsymbol{R}_t)\,dt. \end{cases} \tag{80}$$

The above system is supplied with initial conditions $\boldsymbol{X}_0$ and $\boldsymbol{R}_0$. The processes $\boldsymbol{X}_t$ and $\boldsymbol{W}_t$ are naturally indexed by the trajectories of fluid particles. The Eulerian description corresponding to the above Lagrangian description reads, for $\boldsymbol{X}_t(\boldsymbol{x})$ denoting the conformation at $\boldsymbol{x}$ at time t:

$$d\boldsymbol{X}_t(\boldsymbol{x})+\boldsymbol{u}(t,\boldsymbol{x}).\nabla\boldsymbol{X}_t(\boldsymbol{x})\,dt \quad \nabla\boldsymbol{u}(t,\boldsymbol{x})\boldsymbol{X}_t(\boldsymbol{x})\,dt-\frac{2}{\zeta}\boldsymbol{F}(\boldsymbol{X}_t(\boldsymbol{x}))\,dt+2\sqrt{\frac{kT}{\zeta}}d\boldsymbol{W}_t. \tag{81}$$

Equation (81) is simply the stochastic version of the model already introduced in Sect. 3 under the form of equation (47). Indeed, the latter is the Fokker-Planck associated to the stochastic differential (81). The function ψ solution to (47) is the probability density of $\boldsymbol{X}_t(\boldsymbol{x})$ solution to (81). We refer the reader to the previous section for more details on the ingredient of stochastic analysis needed for the proof of this fact (see Sect. 4.1).

The stress tensor

Using the definition of the stress tensor recalled in Sect. 3, the Kramers formula can be shown. In the stochastic language we adopt here, it reads

$$\boldsymbol{\tau}_p(t) \quad n_p\Big(\mathbb{E}(\boldsymbol{X}_t \otimes \boldsymbol{F}(\boldsymbol{X}_t)) - kT\mathrm{Id}\Big), \tag{82}$$

where $\otimes$ denotes the tensorial product and n_p is the concentration of polymers. See H.C. Öttinger [102, pp158–159], M. Doi and S.F. Edwards [36, section 3.7.4], R.B. Bird et al [12, section 13.3]. Of course, this expression is similar, in terms of $\boldsymbol{X}_t$, to the expression previously found in terms of the probability density function $\psi(t,\cdot)$ of $\boldsymbol{X}_t$, namely (48) in Sect. 3.

Using Itô calculus, an interesting alternative expression can be found for the stress tensor. Indeed, introducing the so-called *structure tensor* $\boldsymbol{X}_t(\boldsymbol{x}) \otimes \boldsymbol{X}_t(\boldsymbol{x})$, we have:

$$\begin{aligned} d(\boldsymbol{X}_t(\boldsymbol{x}) \otimes \boldsymbol{X}_t(\boldsymbol{x})) &\quad (d\boldsymbol{X}_t(\boldsymbol{x})) \otimes \boldsymbol{X}_t(\boldsymbol{x}) + \boldsymbol{X}_t(\boldsymbol{x}) \otimes (d\boldsymbol{X}_t(\boldsymbol{x})) + \frac{4kT}{\zeta}\mathrm{Id}\,dt \\ &\Big(- \boldsymbol{u}(t,\boldsymbol{x}).\nabla(\boldsymbol{X}_t(\boldsymbol{x}) \otimes \boldsymbol{X}_t(\boldsymbol{x})) \\ &+ \nabla\boldsymbol{u}(t,\boldsymbol{x})(\boldsymbol{X}_t(\boldsymbol{x}) \otimes \boldsymbol{X}_t(\boldsymbol{x})) + (\boldsymbol{X}_t(\boldsymbol{x}) \otimes \boldsymbol{X}_t(\boldsymbol{x}))(\nabla\boldsymbol{u}(t,\boldsymbol{x}))^T \\ &- \frac{2}{\zeta}\boldsymbol{F}(\boldsymbol{X}_t) \otimes \boldsymbol{X}_t - \frac{2}{\zeta}\boldsymbol{X}_t \otimes \boldsymbol{F}(\boldsymbol{X}_t) + \frac{4kT}{\zeta}\mathrm{Id}\Big)\,dt \\ &+ 2\sqrt{\frac{kT}{\zeta}}\left((\boldsymbol{X}_t(\boldsymbol{x}) \otimes d\boldsymbol{W}_t) + (d\boldsymbol{W}_t \otimes \boldsymbol{X}_t(\boldsymbol{x}))\right). \end{aligned} \tag{83}$$

The mean of the structure tensor

$$\boldsymbol{A}(t,\boldsymbol{x}) \quad \mathbb{E}(\boldsymbol{X}_t(\boldsymbol{x}) \otimes \boldsymbol{X}_t(\boldsymbol{x})) \tag{84}$$

therefore solves, under some mathematical assumptions on $\boldsymbol{X}_t$,

$$\begin{aligned} \frac{\partial \boldsymbol{A}}{\partial t}(t,\boldsymbol{x}) + \boldsymbol{u}(t,\boldsymbol{x}).\nabla\boldsymbol{A}(t,\boldsymbol{x}) - \nabla\boldsymbol{u}(t,\boldsymbol{x})\boldsymbol{A}(t,\boldsymbol{x}) - \boldsymbol{A}(t,\boldsymbol{x})(\nabla\boldsymbol{u}(t,\boldsymbol{x}))^T& \\ -\frac{4}{\zeta}\mathbb{E}(\boldsymbol{X}_t \otimes \boldsymbol{F}(\boldsymbol{X}_t)) + \frac{4kT}{\zeta}\mathrm{Id}.& \end{aligned} \tag{85}$$

Using (82), the following expression of the stress tensor, called the *Giesekus formula*, is obtained, which only explicitly depends on second moments of $\boldsymbol{X}_t$:

$$\begin{aligned} &\boldsymbol{\tau}_p(t,\boldsymbol{x}) \\ &\quad -\frac{\zeta}{4}n_p\left(\frac{\partial \boldsymbol{A}}{\partial t}(t,\boldsymbol{x}) + \boldsymbol{u}(t,\boldsymbol{x}).\nabla\boldsymbol{A}(t,\boldsymbol{x}) - \nabla\boldsymbol{u}(t,\boldsymbol{x})\boldsymbol{A}(t,\boldsymbol{x}) - \boldsymbol{A}(t,\boldsymbol{x})(\nabla\boldsymbol{u}(t,\boldsymbol{x}))^T\right). \end{aligned}$$

The stress $\boldsymbol{\tau}_p$ is thus proportional to the upper-convected derivative of $\boldsymbol{A}$.

The force

We now have to make the force $\boldsymbol{F}$ specific. In full generality, it is assumed that $\boldsymbol{F}$ is the gradient of a *convex, radially symmetric,* potential $\Pi(\boldsymbol{X}) \quad \pi(\|\boldsymbol{X}\|)$. Thus,

$$\boldsymbol{F}(\boldsymbol{X}) \quad \pi'(\|\boldsymbol{X}\|)\frac{\boldsymbol{X}}{\|\boldsymbol{X}\|}. \tag{86}$$

The convexity of $\Pi(\boldsymbol{X})$ with respect to $\boldsymbol{X}$ of course amounts to that of $\pi(l)$ with respect to l, together with $\pi'(0) \geq 0$.

The simplest example of potential π is the quadratic potential $\pi_{\mathrm{Hook}}(l) \quad H\frac{l^2}{2}$, which of course corresponds to the Hookean force introduced in (46). There are two major pitfalls with the Hookean dumbbell model: first it is *not* a multiscale model in nature, and second (and perhaps more importantly), it has a highly non physical feature.

Let us begin by verifying that the Hookean model is actually equivalent to the purely macroscopic Oldroyd B model introduced in (22).

More on the Hookean model

For Hookean dumbbell, we have: $\mathbb{E}(\boldsymbol{X} \otimes \boldsymbol{F}(\boldsymbol{X})) \quad H\mathbb{E}(\boldsymbol{X} \otimes \boldsymbol{X})$, thus the following equation is obtained on the structure tensor $\boldsymbol{A} \quad \mathbb{E}(\boldsymbol{X} \otimes \boldsymbol{X})$:

$$\begin{aligned}\frac{\partial \boldsymbol{A}}{\partial t}(t,\boldsymbol{x}) + \boldsymbol{u}(t,\boldsymbol{x}).\nabla \boldsymbol{A}(t,\boldsymbol{x}) - \nabla \boldsymbol{u}(t,\boldsymbol{x})\boldsymbol{A}(t,\boldsymbol{x}) - \boldsymbol{A}(t,\boldsymbol{x})(\nabla \boldsymbol{u}(t,\boldsymbol{x}))^T \\ -\frac{4H}{\zeta}\boldsymbol{A}(t,\boldsymbol{x}) + \frac{4kT}{\zeta}\mathrm{Id},\end{aligned} \tag{87}$$

that is, in terms of $\boldsymbol{\tau}_p$:

$$\begin{aligned}\frac{\zeta}{4H}\left(\frac{\partial \boldsymbol{\tau}_p}{\partial t}(t,\boldsymbol{x}) + \boldsymbol{u}(t,\boldsymbol{x}).\nabla \boldsymbol{\tau}_p(t,\boldsymbol{x}) - \nabla \boldsymbol{u}(t,\boldsymbol{x})\boldsymbol{\tau}_p(t,\boldsymbol{x}) - \boldsymbol{\tau}_p(t,\boldsymbol{x})(\nabla \boldsymbol{u}(t,\boldsymbol{x}))^T\right) \\ -\boldsymbol{\tau}_p(t,\boldsymbol{x}) + n_p kT\frac{\zeta}{4H}\left(\nabla \boldsymbol{u}(t,\boldsymbol{x}) + (\nabla \boldsymbol{u}(t,\boldsymbol{x}))^T\right).\end{aligned} \tag{88}$$

Introducing the relaxation time

$$\lambda \quad \frac{\zeta}{4H}, \tag{89}$$

and the viscosity

$$\eta_p \quad n_p kT\lambda, \tag{90}$$

we recognize the macroscopic Maxwell (or Oldroyd B) model (22), that is,

$$\lambda\left(\frac{\partial \boldsymbol{\tau}_p}{\partial t} + \boldsymbol{u}\cdot\nabla \boldsymbol{\tau}_p - \nabla \boldsymbol{u}\boldsymbol{\tau}_p - \boldsymbol{\tau}_p(\nabla \boldsymbol{u})^T\right) + \boldsymbol{\tau}_p \quad \eta_p\dot{\boldsymbol{\gamma}}.$$

A few other multiscale models have a macroscopic equivalent. This is for example the case of the FENE-P model (see Equation (92) below), which is deliberately built to have a macroscopic equivalent. But for most other multiscale models of real interest (in particular those involving FENE forces, see Equation (91) below), no macroscopic equivalent formulation is known. And it is believed that no such formulation exists. In this latter sense, multiscale models *are more powerful* than purely macroscopic models.

In addition to the above, a major theoretical flaw of the dumbbell model, exemplified in (46), is that nothing prevents the end-to-end vector, in the Hookean model, to reach arbitrarily large lengths $|\boldsymbol{r}|$. This is of course not consistent with the actual finite length of the chain. This indeed comes from the method of derivation where we have taken the limit of large N, each of the link being of length a. In the limit, the total length of the chain therefore explodes, thus the formula (46). For all the above reasons, the Hookean dumbbell model, although a perfect test case for preliminary mathematical arguments, is not a fully appropriate benchmark, physically, mathematically and numerically representative, for multiscale models.

Accounting for the finite extensibility of the chain is an important issue, for which adequate models exist. We now turn to two of them.

Other forces

The *FENE model*, where FENE is the acronym for *Finite Extensible Nonlinear Elastic*, is perhaps the most famous force field employed in the simulation of polymeric fluids. It corresponds to the potential (see Fig. 10):

$$\pi_{\mathrm{FENE}}(l) \quad -\frac{bkT}{2}\ln\left(1-\frac{l^2}{bkT/H}\right). \tag{91}$$

The success of this potential is well recognized. In this mathematical text, it is not our purpose to argue on the physical validity and relevance of the models. However, an interesting point to make is the following. The dumbbell model is a very coarse model of the polymer chain. Taking two beads to model a thousand-atom chain seems oversimplifying. When equipped with an appropriate entropic force, like the FENE force, this model nevertheless yields tremendously good results. From a general viewpoint, this shows that

- a multiscale model is much more powerful than a purely macroscopic model,
- the description of the microstructure does not need to be sophisticated to give excellent results,
- it only has to capture the right physics (see the FENE force in contrast to the Hookean force).

Note also that, as a counterpart to the above, the FENE model raises a huge number of challenging mathematical and numerical questions. We will address some of them in Sect. 6.

The FENE model cannot be rephrased under the form of a purely macroscopic model. There is no proof of this claim, but it is strongly believed to be the case. For some specific purposes, the idea has arisen to find a modification of the FENE model (a so-called *closure approximation*) which would have a macroscopic equivalent. This gives birth to the *FENE-P model*, where *P* stands for *Peterlin*. Following A. Peterlin [105] and R.B. Bird, P.J. Dotson and N.L. Johnson [13], it has indeed been proposed to replace the denominator of the FENE force (91) by a mean value of the elongation:

$$\boldsymbol{F}_{\text{FENE-P}}(\boldsymbol{X}_t) \quad \frac{H\boldsymbol{X}_t}{1-\frac{\mathbb{E}(\|\boldsymbol{X}_t\|^2)}{bkT/H}}. \tag{92}$$

Accordingly, the microscopic description of the fluids now reads:

$$\begin{cases} \boldsymbol{\tau}_p \quad n_p\left(\dfrac{H\mathbb{E}\left(\boldsymbol{X}_t\otimes\boldsymbol{X}_t\right)}{1-\mathbb{E}\left(\|\boldsymbol{X}_t\|^2\right)/(bkT/H)}-kT\mathrm{Id}\right), \\ d\boldsymbol{X}_t+\boldsymbol{u}\cdot\nabla\boldsymbol{X}_t\,dt \quad \left(\nabla\boldsymbol{u}\boldsymbol{X}_t-\dfrac{2H}{\zeta}\dfrac{\boldsymbol{X}_t}{1-\mathbb{E}\left(\|\boldsymbol{X}_t\|^2\right)/(bkT/H)}\right)dt \\ \qquad +2\sqrt{\dfrac{kT}{\zeta}}d\boldsymbol{W}_t. \end{cases} \tag{93}$$

Using the expression of $\boldsymbol{\tau}_p$, (82) and (87), we obtain:

$$\begin{aligned} &\frac{\partial\boldsymbol{A}}{\partial t}(t,\boldsymbol{x})+\boldsymbol{u}(t,\boldsymbol{x}).\nabla\boldsymbol{A}(t,\boldsymbol{x})-\nabla\boldsymbol{u}(t,\boldsymbol{x})\boldsymbol{A}(t,\boldsymbol{x})-\boldsymbol{A}(t,\boldsymbol{x})(\nabla\boldsymbol{u}(t,\boldsymbol{x}))^T \\ &\qquad -\frac{4H}{\zeta}\frac{\boldsymbol{A}(t)}{1-\mathrm{tr}(\boldsymbol{A}(t))/(bkt/H)}+\frac{4kT}{\zeta}\mathrm{Id}. \end{aligned} \tag{94}$$

Inserting this into:

$$\boldsymbol{A} \quad \frac{1}{HZ(\mathrm{tr}(\boldsymbol{\tau}_p))}\left(\frac{\boldsymbol{\tau}_p}{n_p}+kT\mathrm{Id}\right),$$

where Z is defined by (29), the following equation is obtained for $\boldsymbol{\tau}_p$:

$$\begin{aligned} &\frac{\zeta}{4H}\left(\frac{\partial\boldsymbol{\tau}_p}{\partial t}(t,\boldsymbol{x})+\boldsymbol{u}(t,\boldsymbol{x}).\nabla\boldsymbol{\tau}_p(t,\boldsymbol{x})-\nabla\boldsymbol{u}(t,\boldsymbol{x})\boldsymbol{\tau}_p(t,\boldsymbol{x})-\boldsymbol{\tau}_p(t,\boldsymbol{x})(\nabla\boldsymbol{u}(t,\boldsymbol{x}))^T\right) \\ &\quad +Z(\mathrm{tr}(\boldsymbol{\tau}_p))\boldsymbol{\tau}_p-\frac{\zeta}{4H}\left(\boldsymbol{\tau}_p+n_pkT\mathrm{Id}\right)\left(\frac{\partial}{\partial t}+\boldsymbol{u}.\nabla\right)\ln(Z(\mathrm{tr}(\boldsymbol{\tau}_p))) \\ &\qquad n_pkT\frac{\zeta}{4H}\left(\nabla\boldsymbol{u}(t,\boldsymbol{x})+(\nabla\boldsymbol{u}(t,\boldsymbol{x}))^T\right), \end{aligned} \tag{95}$$

which is exactly the FENE-P model mentioned in (28) (when λ and η_p are respectively given by (89) and (90)). The FENE-P model can thus be seen as a modification of the FENE model, in order to obtain a multiscale model that has an equivalent purely macroscopic formulation. Other variants of the FENE model exist in the literature.

4.3 The multiscale model

We now have all the bricks for the stochastic variant of our multiscale system (49). Collecting the material of the previous section, we obtain:

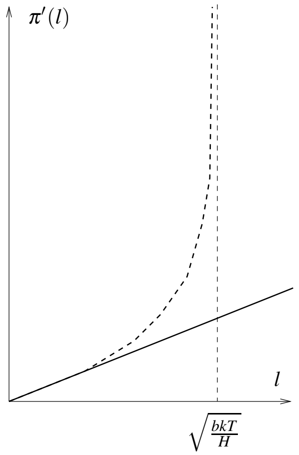


Fig. 10. Comparison of the Hookean force (continuous line) and the FENE force (dashed line).

$$\left\{\begin{aligned}
&\rho\left(\frac{\partial \boldsymbol{u}}{\partial t}(t,\boldsymbol{x})+\boldsymbol{u}(t,\boldsymbol{x})\cdot\nabla\boldsymbol{u}(t,\boldsymbol{x})\right)-\eta\Delta\boldsymbol{u}(t,\boldsymbol{x})+\nabla p(t,\boldsymbol{x})\\
&\qquad\quad \operatorname{div}\left(\boldsymbol{\tau}_p(t,\boldsymbol{x})\right)+\rho\boldsymbol{f}(t,\boldsymbol{x}),\\
&\operatorname{div}\left(\boldsymbol{u}(t,\boldsymbol{x})\right)\quad 0,\\
&\boldsymbol{\tau}_p(t,\boldsymbol{x})\quad n_p\Big(\mathbb{E}(\boldsymbol{X}_t(\boldsymbol{x})\otimes\boldsymbol{F}(\boldsymbol{X}_t(\boldsymbol{x})))-kT\mathrm{Id}\Big),\\
&d\boldsymbol{X}_t(\boldsymbol{x})+\boldsymbol{u}(t,\boldsymbol{x}).\nabla\boldsymbol{X}_t(\boldsymbol{x})\,dt\\
&\qquad\quad \nabla\boldsymbol{u}(t,\boldsymbol{x})\boldsymbol{X}_t(\boldsymbol{x})\,dt-\frac{2}{\zeta}\boldsymbol{F}(\boldsymbol{X}_t(\boldsymbol{x}))\,dt+2\sqrt{\frac{kT}{\zeta}}d\boldsymbol{W}_t.
\end{aligned}\right. \tag{96}$$

As was the case for the Fokker-Planck equation, the stochastic differential equations are to be solved at each point of the macroscopic flow. The process $\boldsymbol{X}_t$ therefore implicitly depends on $\boldsymbol{x}$.

It is well-known that the form of equations actually used in the numerical practice is a non-dimensional form. Because this involves the introduction of several non-dimensional numbers that have a physical meaning and are present in the literature, let us briefly establish now this non-dimensional form for (96) (and thus for (49), by analogy, see (50)).

We introduce the following characteristic quantities: U the characteristic velocity, L the characteristic length, $\lambda \quad \dfrac{\zeta}{4H}$, as in (89), the characteristic relaxation time, $\eta_p \quad n_p kT\lambda$, as in (90), the viscosity of polymers. Then, we consider the following non-dimensional numbers:

$$\begin{cases} \mathrm{Re} \quad \dfrac{\rho UL}{\eta}, \; \varepsilon \quad \dfrac{\eta_p}{\eta}, \\ \mathrm{We} \quad \dfrac{\lambda U}{L}, \; \mu \quad \dfrac{L^2 H}{kT}, \end{cases} \tag{97}$$

respectively the *Reynolds number* Re measuring the ratio of inertia over viscosity (usually for the complex fluids under consideration, $\mathrm{Re} \leq 10$), ε measuring the ratio of viscosity of the polymers over the total viscosity (usually $\varepsilon \approx 0.1$), We the *Weissenberg number* (also called *Deborah number*) which is the ratio of the relaxation time of the polymers versus the characteristic time of the flow (usually $0.1 \leq \mathrm{We} \leq 10$), and μ measuring a ratio of lengths.

Non-dimensionalizing also the force by $\overline{\boldsymbol{F}}(\overline{\boldsymbol{X}}) \quad \dfrac{\boldsymbol{F}(L\overline{\boldsymbol{X}})}{HL}$, and taking (which is the commonly used value) $\mu \quad 1$, we obtain:

$$\boxed{\begin{cases} \mathrm{Re}\left(\dfrac{\partial \boldsymbol{u}}{\partial_t} + \boldsymbol{u}\cdot\nabla\boldsymbol{u}\right) - (1-\varepsilon)\Delta\boldsymbol{u} + \nabla p \quad \operatorname{div}\boldsymbol{\tau}_p + \boldsymbol{f}, \\ \operatorname{div}\boldsymbol{u} \quad 0, \\ \boldsymbol{\tau}_p \quad \dfrac{\varepsilon}{\mathrm{We}}\Big(\mathbb{E}(\boldsymbol{X}_t \otimes \boldsymbol{F}(\boldsymbol{X}_t)) - \mathrm{Id}\Big), \\ d\boldsymbol{X}_t + \boldsymbol{u}.\nabla\boldsymbol{X}_t\, dt \quad \nabla\boldsymbol{u}\boldsymbol{X}_t\, dt - \dfrac{1}{2\mathrm{We}}\boldsymbol{F}(\boldsymbol{X}_t)\, dt + \dfrac{1}{\sqrt{\mathrm{We}}} d\boldsymbol{W}_t. \end{cases}} \tag{98}$$

An important practical remark stems from the actual range of parameters mentioned above. In contrast to the usual setting of computational fluid mechanics where the challenge is to deal with flows with high Reynolds numbers, the challenge here is *not* the Reynolds number (kept relatively small), but the Weissenberg number. Tremendous practical (and also, actually, theoretical) difficulties are associated with the so-called *High Weissenberg number problem* ("high" meaning exceeding, say, 10).

4.4 Schematic overview of the simulation

Our focus so far has been the modelling difficulties for viscoelastic fluids. Another question is the discretization of the models, and their numerical simulations. This has to be performed very carefully since a model is typically validated by some comparisons between experiments and numerical simulations on simple or complex flows.

The present section summarizes the issues and techniques, in a language accessible to readers familiar with scientific computing and numerical analysis. A much more elementary presentation will be given in Sect. 5.

Numerical methods

Most of the numerical methods are based upon a finite element discretization in space and Euler schemes in time, using a semi-explicit scheme: at each timestep, the velocity is first solved for a given stress, and then the stress is updated, for a fixed velocity.

In the case of micro-macro models such as (50) and (98), another discretization step is necessary to approximate the expectation or the integral in the definition of the stress tensor $\boldsymbol{\tau}_p$. There are basically two methods of discretization, depending on the formulation used: stochastic methods for (98), and deterministic methods for (50).

To discretize the expectation in (98), a Monte Carlo method is employed: at each macroscopic point $\boldsymbol{x}$ (*i.e.* at each node of the mesh once the problem is discretized) many replicas (or realizations) $(\boldsymbol{X}_t^{k,K})_{1\leq k\leq K}$ of the stochastic process $\boldsymbol{X}_t$ are simulated, driven by independent Brownian motions $(\boldsymbol{W}_t^k)_{k\geq 1}$, and the stress tensor is obtained as an empirical mean over these processes:

$$\boldsymbol{\tau}_p^K \quad \frac{\varepsilon}{\mathrm{We}}\left(\frac{1}{K}\sum_{k\ \ 1}^{K}\boldsymbol{X}_t^{k,K}\otimes\boldsymbol{F}(\boldsymbol{X}_t^{k,K})-\mathrm{Id}\right).$$

In this context, this discretization method coupling a finite element method and a Monte Carlo technique is called CONNFFESSIT for *Calculation Of Non-Newtonian Flow: Finite Elements and Stochastic SImulation Technique* (see M. Laso and H.C. Öttinger [75]). In Sect. 5, we will implement this method in a simple geometry. Let us already mention that one important feature of the discretization is that, at the discrete level, all the unknowns $(\boldsymbol{u},p,\boldsymbol{\tau}_p)$ become *random variables*. The consequence is that the variance of the results is typically the bottleneck for the accuracy of the method. In particular, variance reduction methods are very important.

To discretize the Fokker-Planck equation in (50), spectral methods are typically used (see A. Lozinski [92] or J.K.C. Suen, Y.L. Joo and R.C. Armstrong [118]). It is not easy to find a suitable variational formulation of the Fokker-Planck equation, and an appropriate discretization that satisfies the natural constraints on the probability density ψ (namely non negativity, and normalization). We refer to C. Chauvière and A. Lozinski [25, 93] for appropriate discretization in the FENE case. One major difficulty in the discretization of Fokker-Planck equations is when the configurational space is high-dimensional. In the context of polymeric fluid flow simulation, when the polymer chain is modelled by a chain of N beads linked by springs, the Fokker-Planck equation is a parabolic equation posed on a $3N$-dimensional domain. Some numerical methods have been developed to discretize such high dimensional problems. The idea is to use an appropriate Galerkin basis, whose dimension does not explode when dimension grows. We refer to P. Delaunay, A. Lozinski and R.G. Owens [33], T. von Petersdorff and C. Schwab [120], H.-J. Bungartz and M. Griebel [20] for the sparse-tensor product approach, to L. Machiels, Y. Maday, and A.T. Patera [94] for the reduced basis approach and to A. Ammar, B. Mokdad, F. Chinesta and R. Keunings [2, 3] for a method coupling a sparse-tensor product discretization with a reduced approximation basis approach.

Main difficulties

It actually turns out that the discretization of micro-macro models such as (50) and (98) or that of macro-macro models such as (23) is not trivial. Let us mention three kinds of difficulties:

1. Some *inf-sup condition* must be satisfied by the spaces respectively used for the discrete velocity, pressure and stress (if one wants the discretization to be stable for ε close to 1).
2. The *advection terms* need to be discretized correctly, in the conservation of momentum equations, in the equation on $\boldsymbol{\tau}_p$ in (23), in the equation on ψ in (50), on in the SDE in (98).
3. The *nonlinear terms* require, as always, a special care. On the one hand, some nonlinear terms stem from the coupling: $\nabla \boldsymbol{u} \boldsymbol{\tau}_p + \boldsymbol{\tau}_p (\nabla \boldsymbol{u})^T$ in (23), $\nabla \boldsymbol{u} \boldsymbol{X}_t$ in (98) or $\operatorname{div}_{\boldsymbol{X}} (\nabla \boldsymbol{u} \boldsymbol{X} \psi(t, \boldsymbol{x}, \boldsymbol{X}))$ in (50). On the other hand, for rheological models more complicated than Oldroyd-B or Hookean dumbbell, some nonlinear terms come from the model itself (see the entropic force $\boldsymbol{F}(\boldsymbol{X}_t)$ in (98) for FENE model for example).

Besides, for both micro-macro models and purely macroscopic models, one central difficulty of the simulation of viscoelastic fluids is the so-called *High Weissenberg Number Problem* (HWNP). It is indeed observed that numerical simulations do not converge when We is too large. The maximum value which can be actually correctly simulated depends on the geometry of the problem (4:1 contraction, flow past a cylinder,...), on the model (Oldroyd-B model, FENE model, ...) and also on the discretization method. Typically, it is observed that this maximum value decreases with mesh refinement.

We will return to these questions in Sect. 6.

4.5 Upsides and downsides of multiscale modelling for complex fluids

Micro-macro *vs* macro-macro modelling

We are now in position to compare the micro-macro approach and the macro-macro approach to simulate polymeric fluids (and more generally complex fluids). Figure 11 summarizes the main features of these approaches. Let us discuss this from two viewpoints: modelling and numerics.

From the modelling viewpoint, the interest of the micro-macro approach stems from the fact it is based on a clear understanding of the physics at play. The kinetic equations used to model the evolution of the polymers are well established and the limit of the validity of these equations is known. The constants involved in micro-macro models have a clear physical signification, and can be estimated from some microscopic properties of the polymer chains. From this point of view, the micro-macro approach seems more predictive, and enables an exploration of the link between the microscopic properties of the polymer chains (or more generally the microstructures in the fluid) and the macroscopic behavior of the complex fluid.

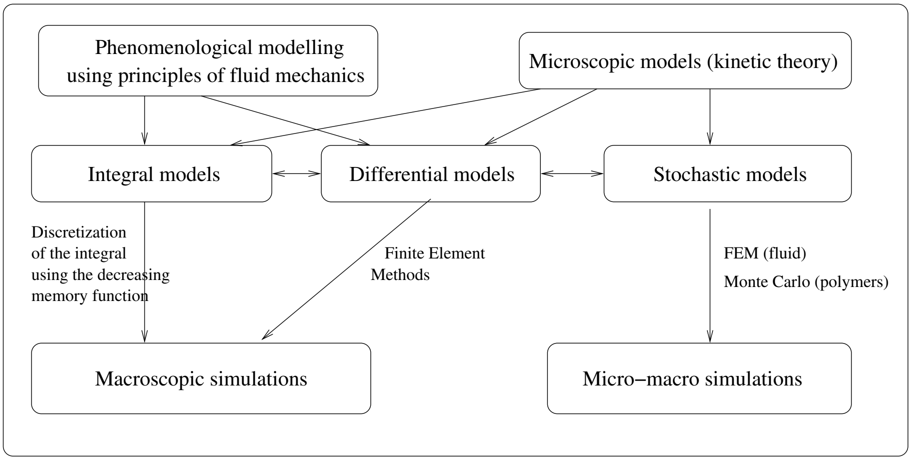

Phenomenological modelling using principles of fluid mechanics
Microscopic models (kinetic theory)
Integral models
Differential models
Stochastic models
Discretization of the integral using the decreasing memory function
Finite Element Methods
FEM (fluid)
Monte Carlo (polymers)
Macroscopic simulations
Micro−macro simulations

Fig. 11. Macro-macro and micro-macro models for complex fluids.

Practice confirms this. It indeed appears that simulations with micro-macro models generally compare better to experiments (see R. Keunings [71, 72]). However, for complex flows and general non-Newtonian fluids, it is still difficult to agree quantitatively with the experiments. In short, it remains a lot to do from the modelling viewpoint, but it is generally admitted that the micro-macro approach is the most promising way to improve the models.

From the numerical point of view, the *major* drawback of the micro-macro approach is its computational cost. The introduction of an additional field to describe the configuration of the microstructure in the fluid implies additional computations and additional memory storage.

For example, for the micro-macro models introduced above in their stochastic form (98), the discretization by a CONNFFESSIT approach requires the storage at each node of the mesh of an ensemble of configurations $(\boldsymbol{X}_t^{i,M})_{1\leq i\leq M}$ of the polymer chains. Even though the SDEs associated to each configuration, and at various node of the mesh can be solved in parallel on each time step, the computational cost remains very high. The micro-macro approach is currently not sufficiently efficient to be used in commercial codes for industrial purposes.

In view of the arguments above, it seems natural to try and design some numerical methods that couple the macro-macro and the micro-macro approaches. The macro-macro model is used where the flow is simple, and the detailed micro-macro model is used elsewhere. The idea of adaptive modelling based on modelling error *a posteriori* analysis (see J.T. Oden and K.S. Vemaganti [100], J.T. Oden and S. Prudhomme [99] or M. Braack and A. Ern [19] has been recently adapted in this context in a preliminary work by A. Ern and T. Lelièvre [40].

We mentioned above the problems raised by the discretization of macro-macro and micro-macro models. It seems that in complex flows, numerical methods based on the micro-macro approach are more robust than those based on the macro-macro approach (see A.P.G. Van Heel [119, p.38], J.C. Bonvin [18, p.115] or C. Chauvière [24]). However, this is not yet well understood mathematically. In addition, the HWNP still limits the range of applicability of the computations, even with micro-macro models. The main interest of micro-macro approaches as compared to macro-macro approaches lies at the modelling level. It may become the method of choice for a backroom strategy. The approach allows to test and validate purely macroscopic models, to supply such models with adequate and reliable boundary conditions, etc..., even if, in the state of the art technology, it does not allow to perform simulations for actual real-world applications, owing to its extremely computationally demanding nature.

Fokker-Planck *vs* SDE formulation

To conclude this section, we would like to discuss the advantages and drawbacks of the two numerical approaches introduced above for the micro-macro approach: that based on the deterministic formulation (50) and that based on the stochastic formulation (98).

The conclusions of this comparison (see Sect. 5 and also A. Lozinski and C. Chauvière [93]) are actually very general: when it is possible to use the deterministic approach (discretization of the Fokker-Planck PDE), it is much more efficient than the stochastic approach (Monte Carlo methods to approximate the expectation). The main reason for that is that the convergence of a Monte Carlo method is slower than that of a deterministic approximation method.

The following question is then: what are the limits of the Fokker-Planck approach? As we mentioned above, designing a numerical method that satisfies the natural requirements of non-negativity and normalization of ψ is not an easy task. In the FENE case, proper variational formulations are to be employed, which take into account the boundary conditions on ψ. In practice, it is observed that the stability of numerical schemes deteriorates when $\nabla \boldsymbol{u}$ becomes too large. But there is another (more fundamental) limitation to the deterministic approach. We mentioned above that the dumbbell model may be actually too crude to describe correctly the polymer chain configuration in some specific situations. It might be better, then, to use a chain of beads and springs. For such a model, the stochastic approach and the associated discretization can both be generalized straightforwardly. However, the deterministic approach is much more problematic. The Fokker-Planck equation becomes a high-dimensional PDE, and the discretization is very difficult. We mentioned above some numerical methods to deal with such PDEs (the sparse-tensor product approach, the reduced approximation basis approach) but they are still limited to a relatively small number of springs, and are much more difficult to implement than Monte Carlo methods.

A summary of the comparison of the various approaches to model complex fluids is given in Table 1.

	MACRO	MICRO-MACRO	
modelling capabilities	low	high	
current utilization	industry	laboratories	
		discretization by Monte Carlo	discretization of Fokker-Planck
computational cost	low	high	moderate
computational bottleneck	HWNP	variance, HWNP	dimension, HWNP

Table 1. Summary of the characteristics of macro-macro and micro-macro approaches for the simulation of complex fluids.

5 Numerical simulation of a test case: the Couette flow

5.1 Setting of the problem

We consider in this section the simple situation of a start-up Couette flow (see Fig. 12). The fluid flows between two parallel planes. Such a model is typically obtained considering a flow in a rheometer, between two cylinders, and taking the limit of large radii for both the inner and the outer cylinders (see Fig. 1). At initial time, the fluid is at rest. The lower plane ($y = 0$, modelling the inner cylinder of the rheometer) is then shifted with a velocity $V(t)$, which, for simplicity, will be set to a constant value V (sinusoidal velocities may also be applied):

$$V(t) = V.$$

On the other hand, the upper plane ($y = L$, modelling the outer cylinder of the rheometer) is kept fixed. Such a setting is called a *start-up flow*, and because it is confined between two parallel plane, a *Couette flow*.

We denote by x and y the horizontal and vertical axes, respectively. The flow is assumed invariant in the direction perpendicular to (x,y).

The polymeric fluid filling in the space between the planes obeys equations (13), which we reproduce here for convenience in their nondimensional form:

$$\begin{cases} \mathrm{Re}\left(\dfrac{\partial \boldsymbol{u}}{\partial t} + (\boldsymbol{u}\cdot\nabla)\boldsymbol{u}\right) - (1-\varepsilon)\Delta\boldsymbol{u} + \nabla p - \operatorname{div}\boldsymbol{\tau}_p = \boldsymbol{f}, \\ \operatorname{div}\boldsymbol{u} = 0. \end{cases} \tag{99}$$

For Couette flow, we have $\boldsymbol{f} = 0$.

It is natural to assume that the flow is *laminar*, that is, the velocity writes $\boldsymbol{u} = u_x(t,x,y)\,\boldsymbol{e}_x$, where $\boldsymbol{e}_x$ is the unitary vector along the x-axis. The incompressibility constraint (8) immediately implies that $\boldsymbol{u} = u_x(t,y)\,\boldsymbol{e}_x$. We now denote:

$$\boldsymbol{u} = u(t,y)\,\boldsymbol{e}_x. \tag{100}$$

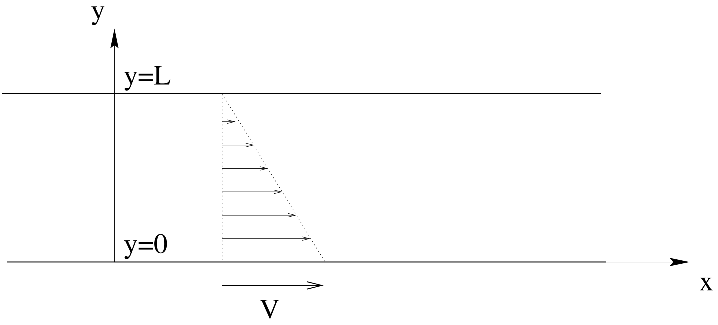


Fig. 12. Couette flow.

In the Newtonian case ($\boldsymbol{\tau}_p \quad 0$), it can be easily shown that a natural assumption on the pressure leads to

$$\begin{cases} \mathrm{Re}\,\dfrac{\partial u}{\partial t}(t,y) \quad (1-\varepsilon)\dfrac{\partial^2 u}{\partial y^2}(t,y), \\ u(0,y) \quad 0, \\ u(t,0) \quad V, \\ u(t,L) \quad 0. \end{cases} \tag{101}$$

Let us now consider the case of a non-Newtonian fluid modelled but the Hookean dumbbell model. We will treat this model as a multiscale model, even if we know from Sect. 4.2 that it is equivalent to the purely macroscopic Oldroyd-B model. Our purpose is indeed to illustrate the numerical approach for such multiscale models, and the Hookean dumbbell model is a nice setting for the exposition. For other models, the situation is more intricate, but *at least* all the difficulties of the Hookean dumbbell model are present.

In full generality, the Fokker-Planck version of the multiscale system describing the flow for the Hookean dumbbell model reads (again in a non-dimensional form), we recall:

$$\begin{cases} \mathrm{Re}\left(\dfrac{\partial \boldsymbol{u}}{\partial t} + (\boldsymbol{u}\cdot\nabla)\,\boldsymbol{u}\right) - (1-\varepsilon)\Delta\boldsymbol{u} + \nabla p - \mathrm{div}\,\boldsymbol{\tau}_p \quad 0, \\ \mathrm{div}\,\boldsymbol{u} \quad 0, \\ \boldsymbol{\tau}_p(t,x,y) \quad \dfrac{\varepsilon}{\mathrm{We}}\left(\displaystyle\int (\boldsymbol{r}\otimes\boldsymbol{r})\,\psi(t,x,y,\boldsymbol{r})\,d\boldsymbol{r} - \mathrm{Id}\right), \\ \dfrac{\partial\psi}{\partial t}(t,x,y,\boldsymbol{r}) + \boldsymbol{u}(t,x,y)\cdot\nabla_{x,y}\psi(t,x,y,\boldsymbol{r}) \\ \qquad -\mathrm{div}_{\boldsymbol{r}}\left(\left(\nabla_{x,y}\boldsymbol{u}(t,x,y)\,\boldsymbol{r} - \dfrac{1}{2\mathrm{We}}\boldsymbol{r}\right)\psi(t,x,y,\boldsymbol{r})\right) + \dfrac{1}{2\mathrm{We}}\Delta_{\boldsymbol{r}}\psi(t,x,y,\boldsymbol{r}), \end{cases} \tag{102}$$

supplied with

$$\begin{cases} \boldsymbol{u}(0,x,y) \quad \boldsymbol{0}, \\ \boldsymbol{u}(t,x,y \quad 0) \quad V\boldsymbol{e}_x, \, \forall t > 0, \\ \boldsymbol{u}(t,x,y \quad L) \quad \boldsymbol{0}, \quad \forall t > 0. \end{cases} \tag{103}$$

Owing to the specific Couette setting, and the assumptions that originate from it (notably (100)), the above general system simplifies into the much simpler one:

$$\begin{cases} \mathrm{Re}\dfrac{\partial u}{\partial t}(t,y) \quad (1-\varepsilon)\dfrac{\partial^2 u}{\partial y^2}(t,y) + \dfrac{\partial \tau}{\partial y}(t,y), \\ \tau(t,y) \quad \dfrac{\varepsilon}{\mathrm{We}}\displaystyle\int_{\mathbb{R}^2} PQ\,\psi(t,y,P,Q)\,dP\,dQ, \\ \dfrac{\partial \psi}{\partial t}(t,y,P,Q) \quad -\dfrac{\partial}{\partial P}\left(\left(\dfrac{\partial u}{\partial y}(t,y)Q - \dfrac{1}{2\mathrm{We}}P\right)\,\psi(t,y,P,Q)\right) \\ \qquad +\dfrac{\partial}{\partial Q}\left(\dfrac{1}{2\mathrm{We}}Q\;\psi(t,y,P,Q)\right) + \dfrac{1}{2\mathrm{We}}\left(\dfrac{\partial^2}{\partial P^2} + \dfrac{\partial^2}{\partial Q^2}\right)\psi(t,y,P,Q), \end{cases} \tag{104}$$

where P and Q are the two components of the end-to-end vector $\boldsymbol{r}$, along the x and y axes respectively. In the above system, $\tau(t,y)$ denotes the xy entry of the tensor $\boldsymbol{\tau}_p$. Actually, the pressure field, and the other entries of the stress tensor may be then deduced, independently.

Let us emphasize at this stage the tremendous simplifications that the Couette model allows for. Owing to the simple geometric setting and the fact that the flow is assumed laminar, the divergence-free constraint (8) is fulfilled by construction of the velocity field and can be eliminated from the system. In addition, the transport terms $(\boldsymbol{u}\cdot\nabla)\boldsymbol{u}$ and $(\boldsymbol{u}\cdot\nabla)\psi$ cancel out, again because of geometrical considerations. This explains the extremely simple form of the equation of conservation of momentum in this context, which indeed reduces to a simple one-dimensional heat equation. This set of simplifications is specific to the Couette flow. Substantial difficulties arise otherwise.

We now describe the numerical approach for (104). To begin with, we present the (simple) finite element discretization of the macroscopic equation. Then we turn to the numerical approach employed for the Fokker-Planck equation. The variant using a stochastic differential equation then follows.

5.2 Discretization of the macroscopic equation

Let us consider the stress $\tau(t,y)$ is known, and perform the variational formulation of the equation in (104) determining the velocity

$$\mathrm{Re}\frac{\partial u}{\partial t}(t,y) \quad (1-\varepsilon)\frac{\partial^2 u}{\partial y^2}(t,y) + \frac{\partial \tau}{\partial y}(t,y)$$

with a view, next, to discretize it using finite elements. Our formulation is

$$\begin{cases} \text{Search for } u : [0,T] \longrightarrow H^1_1(0,L) \text{ such that} \\ \forall v \in H^1_0(0,L), \quad \operatorname{Re} \dfrac{d}{dt}(u(t),v)_{L^2} = -(1-\varepsilon)\left(\dfrac{\partial u}{\partial y}(t), \dfrac{\partial v}{\partial y}\right)_{L^2} - \left(\tau(t), \dfrac{\partial v}{\partial y}\right)_{L^2}, \\ u(0,y) = 0, \end{cases} \tag{105}$$

where we have denoted

$$H^1_0(0,L) = \left\{v \in H^1(0,L), \quad v(0) = 0, \quad v(L) = 0\right\}$$

and

$$H^1_1(0,L) = \left\{v \in H^1(0,L), \quad v(0) = 1, \quad v(L) = 0\right\}.$$

As regards the discretization, we introduce the shape functions for P1 finite elements (for the velocity)

$$\varphi_i(y) = \begin{cases} 1 & \text{when} y = \frac{i}{N}, \\ \text{affine on} & \left[\dfrac{i-1}{N}, \dfrac{i}{N}\right] \text{ and } \left[\dfrac{i}{N}, \dfrac{i+1}{N}\right], \\ 0 & \text{when} y \in \left[0, \dfrac{i-1}{N}\right] \cup \left[\dfrac{i+1}{N}, 1\right], \end{cases} \tag{106}$$

(for $0 \leq i \leq N$), with the obvious adaptations when $i = 0$ and $i = N$, and the shape functions for P0 finite elements (for the stress)

$$\chi_i(y) = \begin{cases} 1 \text{ when} y \in \left[\dfrac{i-1}{N}, \dfrac{i}{N}\right), \\ 0 \text{ otherwise}, \end{cases} \tag{107}$$

(for $1 \leq i \leq N$), both on a regular mesh over $[0,L]$, with meshsize $h = \Delta y = \dfrac{1}{N}$. The approximations for τ and u then read

$$\tau^h(t,y) = \sum_{i=1}^{N} (\tau^h)_i(t)\chi_i(y), \tag{108}$$

$$u^h(t,y) = \sum_{i=1}^{N-1} (u^h)_i(t)\varphi_i(y) + V\,\varphi_N(y),$$

respectively. Note indeed, that, because of the boundary condition, we have for all $t > 0$, $(u^h)_0(t) = 0$ and $(u^h)_N(t) = V$.

It remains to discretize in time, which we do using a backward Euler scheme for the viscous term. The fully discrete formulation is thus

$$\left\{\begin{array}{l}
\text{Solve for } (u^h)^n_j \text{ for } j \quad 1,\dots,N-1 \text{ and for } n \geq 0 \\
\text{such that } (u^h)^0_j \equiv 0 \text{ and } \forall i \quad 1,\dots,N-1, \\
\mathrm{Re}\left(\dfrac{\sum\limits_{j\ 1}^{N-1} (u^h)^{n+1}_j \varphi_j - \sum\limits_{j\ 1}^{N-1} (u^h)^n_j \varphi_j}{\Delta t}, \varphi_i\right)_{L^2} \\
\qquad (1-\varepsilon)\left(\dfrac{\partial}{\partial y}\left(\sum\limits_{j\ 1}^{N-1} (u^h)^{n+1}_j \varphi_j + V\varphi_N\right), \dfrac{\partial}{\partial y}\varphi_i\right)_{L^2} - \left((\tau^h)^n, \dfrac{\partial}{\partial y}\varphi_i\right)_{L^2}
\end{array}\right. \tag{109}$$

where $(\tau^h)^n$ denotes the approximation of τ^h at time t^n.

In algebraic terms, this writes

$$\mathrm{Re}\, M\frac{U^{n+1}-U^n}{\Delta t} \quad -(1-\varepsilon)AU^{n+1} - GS^n + B, \tag{110}$$

where

$$U^n \quad \left[(u^h)^n_1,\dots,(u^h)^n_{N-1}\right]^T$$

is the unknown,

$$S^n \quad \left[(\tau^h)^n_1,\dots,(\tau^h)^n_N\right]^T,$$

and G is a matrix with (i,j)-entry

$$G_{i,j} \quad \int_0^L \frac{\partial\varphi_i}{\partial y}\chi_j\, dy. \tag{111}$$

The vector $B \quad -(1-\varepsilon)V\left[0,\dots,0,\int_0^L \frac{\partial\varphi_N}{\partial y}\frac{\partial\varphi_{N-1}}{\partial y}\,dy\right]^T$ is associated with the Dirichlet boundary condition. The matrices M and A respectively denote the matrices of mass and rigidity of the P1 finite elements:

$$M_{i,j} \quad \int_0^L \varphi_i\,\varphi_j\,dy, \tag{112}$$

$$A_{i,j} \quad \int_0^L \frac{\partial\varphi_i}{\partial y}\frac{\partial\varphi_j}{\partial y}\,dy. \tag{113}$$

5.3 Microscopic problem: the deterministic approach

We now turn to the discretization of the Fokker-Planck equation in (104), that is

$$\begin{aligned}\frac{\partial\psi}{\partial t}(t,y,P,Q) \quad & -\frac{\partial}{\partial P}\left(\left(\frac{\partial u}{\partial y}(t,y)Q-\frac{1}{2\mathrm{We}}P\right)\ \psi(t,y,P,Q)\right) \\ & +\frac{\partial}{\partial Q}\left(\frac{1}{2\mathrm{We}}Q\,\psi(t,y,P,Q)\right)+\frac{1}{2\mathrm{We}}\left(\frac{\partial^2}{\partial P^2}+\frac{\partial^2}{\partial Q^2}\right)\psi(t,y,P,Q).\end{aligned} \tag{114}$$

Since y is only a parameter, we omit to mention the explicit dependence of ψ upon this parameter throughout this paragraph.

We introduce the equilibrium solution of (114) (*i.e.* the steady state solution of (114) for $u \quad 0$), namely

$$\psi_\infty(P,Q) \quad \frac{1}{2\pi}\exp\left(-\frac{P^2+Q^2}{2}\right). \tag{115}$$

We next change the unknown function in (114) setting

$$\varphi \quad \frac{\psi}{\psi_\infty} \tag{116}$$

and rewrite (114) as

$$\begin{aligned}\psi_\infty\frac{\partial\varphi}{\partial t}(t,P,Q) \quad & -\frac{\partial}{\partial P}\left(\frac{\partial u}{\partial y}Q\,\psi_\infty\ \varphi\right) \\ & +\frac{1}{2\mathrm{We}}\frac{\partial}{\partial P}\left(\psi_\infty\frac{\partial}{\partial P}\varphi\right)+\frac{1}{2\mathrm{We}}\frac{\partial}{\partial Q}\left(\psi_\infty\frac{\partial}{\partial Q}\varphi\right)\end{aligned} \tag{117}$$

which is readily semi-discretized in time as

$$\begin{aligned}\psi_\infty\frac{\varphi_{n+1}-\varphi_n}{\Delta t} \quad & -\frac{\partial}{\partial P}\left(\frac{\partial u}{\partial y}Q\,\psi_\infty\ \varphi_n\right) \\ & +\frac{1}{2\mathrm{We}}\frac{\partial}{\partial P}\left(\psi_\infty\frac{\partial}{\partial P}\varphi_{n+1}\right)+\frac{1}{2\mathrm{We}}\frac{\partial}{\partial Q}\left(\psi_\infty\frac{\partial}{\partial Q}\varphi_{n+1}\right).\end{aligned} \tag{118}$$

A variational formulation of (118) on an appropriate functional space $\mathcal{V}$ (see for example B. Jourdain, C. Le Bris, T. Lelièvre and F. Otto [65, Appendix B]) is then:

$$\begin{cases}\text{Solve for } \varphi_n\in\mathcal{V} \text{ for } n\geq 0 \text{ such that } \forall\theta\in\mathcal{V}, \\ \displaystyle\int\frac{\varphi_{n+1}-\varphi_n}{\Delta t}\,\theta\,\psi_\infty \quad \int\frac{\partial u}{\partial y}Q\frac{\partial\theta}{\partial P}\varphi_n\psi_\infty \\ \displaystyle\qquad\qquad -\frac{1}{2\mathrm{We}}\int\frac{\partial\theta}{\partial P}\frac{\partial\varphi_{n+1}}{\partial P}\psi_\infty-\frac{1}{2\mathrm{We}}\int\frac{\partial\theta}{\partial Q}\frac{\partial\varphi_{n+1}}{\partial Q}\psi_\infty, \\ \varphi_0 \quad 1.\end{cases} \tag{119}$$

The most appropriate basis to use for the Galerkin basis in (119) is a basis consisting of (tensor products of) Hermite polynomials H_i:

$$\chi_{i,j}(P,Q) \quad H_i(P)H_j(Q), \tag{120}$$

where

$$H_0(P) = 1, \quad H_1(P) = P, \quad H_2(P) = \frac{1}{\sqrt{2}}(P^2 - 1). \tag{121}$$

Indeed, since

$$\frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} H_i(P) H_j(P) \exp\left(-P^2/2\right) dP = \delta_{ij}, \tag{122}$$

and since the Gaussian function is precisely the stationary solution to the equation under consideration, the basis of Hermite polynomials is well adapted to the problem under consideration. In particular, the mass matrix related to the discretization of $\int \frac{\varphi_{n+1} - \varphi_n}{\Delta t}\, \theta\, \psi_\infty$ in (119) is the identity matrix. The matrix associated with the discretization of the diffusion terms $\int \frac{\partial \theta}{\partial P} \frac{\partial \varphi_{n+1}}{\partial P} \psi_\infty + \int \frac{\partial \theta}{\partial Q} \frac{\partial \varphi_{n+1}}{\partial Q} \psi_\infty$ in (119) is diagonal. In addition, the use of such a spectral basis allows to circumvent the practical difficulty related to the fact that the equation is posed on the whole space.

5.4 Microscopic problem: the stochastic approach

Instead of using the Fokker-Planck equation viewpoint, we may alternatively introduce the couple of stochastic differential equations

$$\begin{cases} dP(t,y) = \left(\dfrac{\partial u}{\partial y}(t,y) Q(t) - \dfrac{1}{2\mathrm{We}} P(t,y)\right) dt + \dfrac{1}{\sqrt{\mathrm{We}}}\, dV_t, \\ dQ(t) = -\dfrac{1}{2\mathrm{We}} Q(t) dt + \dfrac{1}{\sqrt{\mathrm{We}}}\, dW_t, \end{cases} \tag{123}$$

where V_t and W_t are two independent one-dimensional Brownian motions, and next evaluate the stress with

$$\tau(t,y) = \frac{\varepsilon}{\mathrm{We}} \int_{\mathbb{R}^2} P Q\, \psi(t,y,P,Q)\, dP\, dQ = \frac{\varepsilon}{\mathrm{We}} \mathbb{E}(P(t,y)Q(t)). \tag{124}$$

Note that in this simple geometry and for Hookean dumbbells, $Q(t)$ does not depend on y.

In order to solve (123), we supply it with initial conditions homogeneous in y, and use a forward Euler scheme:

$$\begin{cases} P_i^{n+1} = \Delta t \dfrac{U_i^{n+1} - U_{i-1}^{n+1}}{\Delta y} Q^n + \left(1 - \dfrac{\Delta t}{2\mathrm{We}}\right) P_i^n + \sqrt{\dfrac{\Delta t}{\mathrm{We}}} \Delta V_i^n, \\ Q^{n+1} = \left(1 - \dfrac{\Delta t}{2\mathrm{We}}\right) Q^n + \sqrt{\dfrac{\Delta t}{\mathrm{We}}} \Delta W^n, \end{cases} \tag{125}$$

for $1 \le i \le N$, where ΔV_i^n and ΔW^n are standard normal random variables. The stress is then given by

$$(\tau^h)_i^{n+1} = \frac{\varepsilon}{\mathrm{We}} \mathbb{E}(P_i^{n+1} Q^{n+1}). \tag{126}$$

Following the standard Monte Carlo method, (126) is approximated replacing the expectation value by an empirical mean. A supposedly large number K of realizations of the random variables P_i^n and Q^n is generated: (for $1 \le i \le N$)

$$P_{i,k}^{n+1} \quad \Delta t \frac{U_i^{n+1} - U_{i-1}^{n+1}}{\Delta y} Q_k{}^n + \left(1 - \frac{\Delta t}{2\mathrm{We}}\right) P_{i,k}^n + \sqrt{\frac{\Delta t}{\mathrm{We}}} V_{i,k}^n, \tag{127}$$

$$Q_k^{n+1} \quad \left(1 - \frac{\Delta t}{2\mathrm{We}}\right) Q_k^n + \sqrt{\frac{\Delta t}{\mathrm{We}}} W_k^n, \tag{128}$$

for $1 \leq k \leq K$, and

$$(\tau^h)_i^{n+1} \quad \frac{\varepsilon}{\mathrm{We}} \frac{1}{K} \sum_{k \quad 1}^{K} P_{i,k}^{n+1} Q_k^{n+1} \tag{129}$$

is computed. For the evolution (127)–(128), the initial conditions P_i^0 and Q^0 are chosen as standard normal random variables, since the fluid is assumed at rest at initial time.

This discretization is the CONNFFESSIT approach mentioned above, implemented in a simple case.

A crucial remark is the following. Since the stress $(\tau^h)_i^{n+1}$ is an empirical mean (129), it is thus *also a random variable*. It follows that the macroscopic velocity itself, which solves the fully discretized version of (109) is a random variable. On the contrary, in the limit $K \rightarrow \infty$, the stress and the velocity are deterministic quantities (since the expectation value (126) is a deterministic quantity).

Consequently, when one speaks of computing the velocity or the stress using the stochastic approach, it implies performing a *collection* of simulations, and averaging over the results.

Immediately, this brings into the picture variance issues. Let us briefly explain in the present context how the noise inherently present in the numerical simulation may be somewhat reduced. This is the famous *variance reduction* problem.

A basic approach consists in correlating the trajectories P_i from one index i to another one. For this purpose, we first take as initial conditions for P_i standard normal random variables $P_{i,k}^0 \quad P_k^0$ that do not depend on i, and second use Brownian motions V_k^n, uniform in i: $V_{i,k}^n \quad V_k^n$. Equation (127) is thus replaced with

$$P_{i,k}^{n+1} \quad \Delta t \frac{U_i^{n+1} - U_{i-1}^{n+1}}{\Delta y} Q_k{}^n + \left(1 - \frac{\Delta t}{2\mathrm{We}}\right) P_{i,k}^n + \sqrt{\frac{\Delta t}{\mathrm{We}}} V_k^n. \tag{130}$$

It is observed that this technique reduces the variance on the velocity u. In addition, it provides an empirical mean that is less oscillatory w.r.t. the space variable y than that obtained from the original approach (see Sect. 6.3 below for more details).

Another method, with a large spectrum of applications, is that of *control variate*. The bottom line is to avoid computing $\mathbb{E}(PQ)$ directly, and to rather compute each of the terms of the sum

$$\mathbb{E}(PQ) \quad \mathbb{E}(\tilde{P}\tilde{Q}) + \mathbb{E}(PQ - \tilde{P}\tilde{Q})$$

where $\tilde{P}$ et $\tilde{Q}$ are two processes such that

- $\mathbb{E}(\tilde{P}\tilde{Q})$ is easy to compute or approximate, analytically or numerically,
- $\tilde{P}\tilde{Q}$ is close enough to PQ so that $\mathrm{Var}(PQ-\tilde{P}\tilde{Q}) \ll \mathrm{Var}(PQ)$.

The two extreme situations are

- $\tilde{P} = \tilde{Q} = 0$, that is, $\mathbb{E}(\tilde{P}\tilde{Q})$ is very easy to compute but no variance reduction is attained,
- $\tilde{P} = P$ and $\tilde{Q} = Q$, so that $\mathrm{Var}(PQ-\tilde{P}\tilde{Q}) = 0$ but then $\mathbb{E}(\tilde{P}\tilde{Q})$ is no easier to compute than $\mathbb{E}(PQ)$!

Somewhat in the style of preconditioners for the resolution of algebraic systems, some compromise has to be found. In the specific case under consideration, an efficient choice consists in defining $(\tilde{P},\tilde{Q})(t)$ as the solution to the same stochastic differential equations (123) for zero velocity and $(\tilde{P},\tilde{Q})(0) = (P,Q)(0)$ ($(\tilde{P},\tilde{Q})(t)$ remains at equilibrium):

$$\begin{aligned} d\tilde{P}(t) &= -\frac{1}{2\mathrm{We}}\tilde{P}(t)dt + \frac{1}{\sqrt{\mathrm{We}}}\,dV_t,\\ d\tilde{Q}(t) &= -\frac{1}{2\mathrm{We}}\tilde{Q}(t)dt + \frac{1}{\sqrt{\mathrm{We}}}\,dW_t. \end{aligned}$$

Clearly, both $\tilde{Q}$ and Q satisfy the same equation, and $\tilde{P}$ does not depend on y. On the other hand, $\mathbb{E}(\tilde{P}\tilde{Q}) = 0$ since $\tilde{P}$ and $\tilde{Q}$ are independent (since they are at initial time), and both of zero mean (arguing on the above stochastic differential equations). In order to simulate $\mathbb{E}(PQ-\tilde{P}\tilde{Q})$, the forward Euler scheme is employed: for each n, we set $\tilde{Q}_k^n = Q_k^n$ and

$$\tilde{P}_{i,k}^{n+1} = \left(1-\frac{\Delta t}{2\mathrm{We}}\right)\tilde{P}_{i,k}^n + \sqrt{\frac{\Delta t}{We}}V_{i,k}^n. \tag{131}$$

Of course, in order for an effective variance reduction to be reached, the same Gaussian variables $V_{i,k}^n$ are to be used for simulating both $\tilde{P}$ and P. If independent random variables were used for simulating $\tilde{P}$ and P, $\tilde{P}$ and P would be independent random variables and thus $\mathrm{Var}(P-\tilde{P}) = \mathrm{Var}(P)+\mathrm{Var}(\tilde{P}) > \mathrm{Var}(P)$.

The simulation of $(\tau^h)_i^{n+1}$ consists in solving

$$\begin{aligned} (\tau^h)_i^{n+1} &= \frac{\varepsilon}{\mathrm{We}}\mathbb{E}(PQ),\\ &= \frac{\varepsilon}{\mathrm{We}}(\mathbb{E}(\tilde{P}\tilde{Q})+\mathbb{E}(PQ-\tilde{P}\tilde{Q})),\\ &= \frac{\varepsilon}{\mathrm{We}}(0+\mathbb{E}(PQ-\tilde{P}\tilde{Q})),\\ &\approx \frac{\varepsilon}{\mathrm{We}}\frac{1}{K}\sum_{k=1}^{K}(P_{i,k}^{n+1}Q_k^{n+1}-\tilde{P}_{i,k}^{n+1}\tilde{Q}_k^{n+1}),\\ &\approx \frac{\varepsilon}{\mathrm{We}}\frac{1}{K}\sum_{k=1}^{K}((P_{i,k}^{n+1}-\tilde{P}_{i,k}^{n+1})Q_k^{n+1}), \end{aligned} \tag{132}$$

instead of (129).

Summarizing the above, the computation performed at time t^n, knowing $((u^h)^n$, $(\tau^h)^n)$, in order to advance forward in time Δt, is:

(1) Knowing all $(\tau^h)_i^n$ for all intervals indexed by i, these values are used in the macroscopic equation (110) to obtain the velocity values U_i^{n+1} $(1 \le i \le N-1)$.
(2) On each space interval with length Δy,
  (2.1) An ensemble of K realizations of the random variables $V_{i,k}^n$ and W_k^n $(1 \le k \le K)$ are simulated ; If variance reduction by control variate is used, the random variables $\tilde{P}_{i,k}$ are updated following (131);
  (2.2) Using the values U_i^{n+1} $(1 \le i \le N-1)$ in the schemes (127)–(128) discretizing the SDEs (123), the values $P_{i,k}^{n+1}$ and Q_k^{n+1} are obtained;
  (2.3) By computing the empirical mean (129) over the K realizations, the stress $(\tau^h)_i^{n+1}$ is obtained at the next timestep.

5.5 Extension to the FENE model

In the FENE model, the SDE that has to be discretized is

$$d\boldsymbol{X}_t + \boldsymbol{u}.\nabla\boldsymbol{X}_t\,dt \quad \nabla\boldsymbol{u}\boldsymbol{X}_t\,dt - \frac{1}{2\mathrm{We}}\frac{\boldsymbol{X}_t}{1-\|\boldsymbol{X}_t\|^2/b}\,dt + \frac{1}{\sqrt{\mathrm{We}}}d\boldsymbol{W}_t. \tag{133}$$

In the specific geometric setting of this section, denoting $\boldsymbol{X}_t \quad (P(t),Q(t))$ and $\boldsymbol{W}_t \quad (V_t,W_t)$, (133) writes:

$$\begin{cases} dP(t,y) \quad \left(\dfrac{\partial u}{\partial y}(t,y)Q(t,y) - \dfrac{1}{2\mathrm{We}}\dfrac{P(t,y)}{1-(P(t,y)^2+Q(t,y)^2)/b}\right)dt \\ \qquad\qquad + \dfrac{1}{\sqrt{\mathrm{We}}}dV_t, \\ dQ(t,y) \quad -\dfrac{1}{2\mathrm{We}}\dfrac{Q(t,y)}{1-(P(t,y)^2+Q(t,y)^2)/b}dt + \dfrac{1}{\sqrt{\mathrm{We}}}dW_t. \end{cases} \tag{134}$$

In contrast to the Hookean dumbbell case, notice that Q is now also depending on the space variable y.

Let us now discuss how to discretize this SDE, and what type of control variate technique may be employed to reduce the variance.

Compared to the Hookean dumbbell case, an additional difficulty of the discretization of (133) is the singularity of the force when $\|\boldsymbol{X}_t\|^2$ goes to b. It can be shown (see B. Jourdain and T. Lelièvre [66]) that, at the continuous level, the stochastic process $\boldsymbol{X}_t$ does not hit the boundary of $\mathscr{B}(0,\sqrt{b})$ in finite time, provided $b > 2$. Notice that without the Brownian term, it would be clear that $\boldsymbol{X}_t$ remains inside $\mathscr{B}(0,\sqrt{b})$ but this fact is not so clear in the SDE case, and actually requires an assumption on b. When discretizing (133), one is interested in imposing also this property for the discrete process $\boldsymbol{X}^n$. A naïve Euler scheme such as (127)–(128) does not satisfy this property. One option is to simply reject draws such that $\|\boldsymbol{X}^{n+1}\|^2 > b$. An alternative option has been proposed by H.C. Öttinger [102, p. 218-221]. It consists in treating implicitly the force term, and it can be shown that it yields a discrete process $\boldsymbol{X}^n$ with actual values in $\mathscr{B}(0,\sqrt{b})$. Let us write this scheme for the SDE (133) without the advection term $\boldsymbol{u}.\nabla\boldsymbol{X}_t\,dt$:

$$\begin{cases} \overline{\boldsymbol{X}^{n+1}} \quad \boldsymbol{X}^n + \nabla\boldsymbol{u}^n\boldsymbol{X}^n\Delta t - \dfrac{1}{2\mathrm{We}}\dfrac{\boldsymbol{X}^n}{1-\|\boldsymbol{X}^n\|^2/b}\Delta t + \sqrt{\dfrac{\Delta t}{\mathrm{We}}}\boldsymbol{G}^n, \\ \left(1+\dfrac{1}{4\mathrm{We}}\dfrac{\Delta t}{1-\|\boldsymbol{X}^{n+1}\|^2/b}\right)\boldsymbol{X}^{n+1} \quad \boldsymbol{X}^n \\ \qquad +\dfrac{1}{2}\left(\nabla\boldsymbol{u}^n\boldsymbol{X}^n + \nabla\boldsymbol{u}^{n+1}\overline{\boldsymbol{X}^{n+1}} - \dfrac{1}{2\mathrm{We}}\dfrac{\boldsymbol{X}^n}{1-\|\boldsymbol{X}^n\|^2/b}\right)\Delta t + \sqrt{\dfrac{\Delta t}{\mathrm{We}}}\boldsymbol{G}^n, \end{cases} \tag{135}$$

where $\boldsymbol{G}^n$ are i.i.d. Gaussian variables with covariance matrix Id.

We next consider the question of variance reduction by control variate. As mentioned above, the idea is to compute the stress tensor as

$$\boldsymbol{\tau}_p \quad \frac{\varepsilon}{\mathrm{We}}\left(\mathbb{E}\left(\frac{\boldsymbol{X}_t\otimes\boldsymbol{X}_t}{1-\|\boldsymbol{X}_t\|^2/b} - \tilde{\boldsymbol{X}}_t\otimes\tilde{\boldsymbol{F}}(\tilde{\boldsymbol{X}}_t)\right) + \mathbb{E}\left(\tilde{\boldsymbol{X}}_t\otimes\tilde{\boldsymbol{F}}(\tilde{\boldsymbol{X}}_t)\right)\right),$$

where $\tilde{\boldsymbol{X}}_t$ is a suitable chosen stochastic process, and $\tilde{\boldsymbol{F}}$ an adequate force (for example $\tilde{\boldsymbol{F}} \quad \boldsymbol{F}$) such that the variance of the term in the first expectation,

$$\mathbb{E}\left(\frac{\boldsymbol{X}_t\otimes\boldsymbol{X}_t}{1-\|\boldsymbol{X}_t\|^2/b} - \tilde{\boldsymbol{X}}_t\otimes\tilde{\boldsymbol{F}}(\tilde{\boldsymbol{X}}_t)\right),$$

is as small as possible, and the computation of the second expectation $\mathbb{E}\left(\tilde{\boldsymbol{X}}_t\otimes\tilde{\boldsymbol{F}}(\tilde{\boldsymbol{X}}_t)\right)$ is easy. For the variance of the first term to be small, $\tilde{\boldsymbol{X}}_t$ needs to be as close as possible to $\boldsymbol{X}_t$ (in stochastic terms, $\tilde{\boldsymbol{X}}_t$ needs to be *coupled* to $\boldsymbol{X}_t$). In particular, one requires that $\boldsymbol{X}_0 \quad \tilde{\boldsymbol{X}}_0$ and the Brownian motion driving $\boldsymbol{X}_t$ is the same as the one driving $\tilde{\boldsymbol{X}}_t$.

Then two types of control variate are classically used (see J. Bonvin and M. Picasso [16]). As in the previous section for Hookean dumbbells, $\tilde{\boldsymbol{X}}_t$ can be the process "at equilibrium". It consists in computing $\tilde{\boldsymbol{X}}_t$ as the solution to the same SDE as $\boldsymbol{X}_t$ (and thus $\tilde{\boldsymbol{F}} \quad \boldsymbol{F}$) without the term $\nabla\boldsymbol{u}\boldsymbol{X}_t\,dt$. If $\boldsymbol{X}_0 \quad \tilde{\boldsymbol{X}}_0$ is distributed according to the invariant law of the SDE, then the law of $\tilde{\boldsymbol{X}}_t$ does not depend on time and thus

$$\mathbb{E}\left(\frac{\tilde{\boldsymbol{X}}_t\otimes\tilde{\boldsymbol{X}}_t}{1-\|\tilde{\boldsymbol{X}}_t\|^2/b}\right) \quad \mathbb{E}\left(\frac{\tilde{\boldsymbol{X}}_0\otimes\tilde{\boldsymbol{X}}_0}{1-\|\tilde{\boldsymbol{X}}_0\|^2/b}\right)$$

which can be analytically computed. This method typically works when the system remains close to equilibrium.

When the system goes out of equilibrium, another idea is to use a *closure approximation* to obtain a model which is close to the FENE model, but which has a macroscopic equivalent so that the second term $\mathbb{E}\left(\tilde{\boldsymbol{X}}_t\otimes\tilde{\boldsymbol{F}}(\tilde{\boldsymbol{X}}_t)\right)$ can be computed by discretizing a PDE (which is very cheap compared to the Monte Carlo method).

For example, one can take the Hookean dumbbell model $(\tilde{\boldsymbol{F}}(\tilde{\boldsymbol{X}}_t) \quad \tilde{\boldsymbol{X}}_t)$ and compute $\mathbb{E}\left(\tilde{\boldsymbol{X}}_t \otimes \tilde{\boldsymbol{X}}_t\right)$ by solving the PDE for the Oldroyd-B model. One can also choose the FENE-P model $(\tilde{\boldsymbol{F}}(\tilde{\boldsymbol{X}}_t) \quad \frac{\tilde{\boldsymbol{X}}_t \otimes \tilde{\boldsymbol{X}}_t}{1-\mathbb{E}\|\tilde{\boldsymbol{X}}_t\|^2/b})$ and compute $\mathbb{E}\left(\tilde{\boldsymbol{X}}_t \otimes \tilde{\boldsymbol{F}}(\tilde{\boldsymbol{X}}_t)\right)$ by solving the associated PDE (28). Closure relations are thus important not only to obtain macroscopic models with microscopic interpretation, but also to build efficient variance reduction methods. For closure relations for the FENE model, we refer to Q. Du, C. Liu and P. Yu [37, 32].

5.6 MATLAB codes

In this section, we give the MATLAB codes[3] for the computation of the velocity and the stress in a Couette flow for the Hookean dumbbell model (start-up of shear flow). We recall that this model is equivalent to the Oldroyd-B model. We thus have three formulations of the problem:

- The macro-macro formulation:

$$\begin{cases} \mathrm{Re}\dfrac{\partial u}{\partial t}(t,y) - (1-\varepsilon)\dfrac{\partial^2 u}{\partial y^2}(t,y) \quad \dfrac{\partial \tau}{\partial y}(t,y), \\ \frac{\partial \tau}{\partial t} + \frac{1}{\mathrm{We}}\tau \quad \frac{\varepsilon}{\mathrm{We}}\frac{\partial u}{\partial y}. \end{cases} \tag{136}$$

- The micro-macro formulation with the SDEs:

$$\begin{cases} \mathrm{Re}\dfrac{\partial u}{\partial t}(t,y) - (1-\varepsilon)\dfrac{\partial^2 u}{\partial y^2}(t,y) \quad \dfrac{\partial \tau}{\partial y}(t,y), \\ \tau(t,y) \quad \dfrac{\varepsilon}{\mathrm{We}}\mathbb{E}(X_t(y)Y_t), \\ dX_t(y) \quad \dfrac{\partial u}{\partial y}(t,y)Y_t\,dt - \dfrac{1}{2\mathrm{We}}X_t(y)\,dt + \dfrac{1}{\sqrt{\mathrm{We}}}dV_t, \\ dY_t \quad -\dfrac{1}{2\mathrm{We}}Y_t\,dt + \dfrac{1}{\sqrt{\mathrm{We}}}dW_t. \end{cases} \tag{137}$$

- The micro-macro formulation with the Fokker-Planck equation:

$$\begin{cases} \mathrm{Re}\dfrac{\partial u}{\partial t}(t,y) - (1-\varepsilon)\dfrac{\partial^2 u}{\partial y^2}(t,y) \quad \dfrac{\partial \tau}{\partial y}(t,y), \\ \tau(t,y) \quad \dfrac{\varepsilon}{\mathrm{We}}\displaystyle\int XYp(t,y,X,Y)dXdY, \\ \dfrac{\partial p}{\partial t} \quad -\mathrm{div}_{(X,Y)}\left(\left((\dfrac{\partial u}{\partial y}Y,0) - (X,Y)\dfrac{1}{2\mathrm{We}}\right)p\right) + \dfrac{1}{2\mathrm{We}}\Delta_{(X,Y)}p. \end{cases} \tag{138}$$

We now insert the MATLAB source `Couette_Oldroyd_B.m` for the discretization of (136).

[3] The codes are available at the following address: `http://hal.inria.fr/inria-00165171`

```
clear all;

% Physical parameters
Re=0.1;Eps=0.9;We=0.5;
v=1.;
T=1.; % Maximal time

% Discretization
% Space
I=100;
dx=1/I;mesh=[0:dx:1];
% Time
N=100;
dt=T/N;

% Matrices
D1=diag(ones(1,I-1),-1);D1=D1(2:I,:);D1=[D1,zeros(I-1,1)];
D2=diag(ones(1,I-1));D2=[zeros(I-1,1),D2,zeros(I-1,1)];
D3=diag(ones(1,I-1),+1);D3=D3(1:(I-1),:);D3=[zeros(I-1,1),D3];
% Mass matrix
M=(1/6)*D1+(2/3)*D2+(1/6)*D3;
M=M.*dx;M=sparse(M);
MM=M(:,2:I);
% Stiffness matrix
A=(-1)*D1+2*D2+(-1)*D3;
A=A./dx;A=sparse(A);
AA=A(:,2:I);
BB=Re*MM./dt+(1-Eps)*AA;

% Vectors
u=zeros(I+1,1); % Initial velocity
tau=zeros(I,1); % Initial stress: \E(PQ)=0 at t=0
gradtau=zeros(I-1,1);
CLL=zeros(I+1,1);

% Time iterations
for t=dt:dt:T,
  uold=u;
  gradtau=tau(2:I)-tau(1:(I-1));
  if ((t/T)<0.1)
        CLL(1)=v*10*(t/T);
  else
         CLL(1)=v ;
  end;
  CL=(Re*M./dt+(1-Eps)*A)*CLL;
  F=(Re*M./dt)*u-CL+(Eps/We)*gradtau;
  u(2:I)=BB\F;
  if ((t/T)<0.1)
        u(1)=v*10*(t/T);
  else
         u(1)=v;
  end;
  for l=1:I
        tau(l)=(1-dt/We).*tau(l)+(dt/dx)*(u(l+1)-u(l));
    % tau(l)=(1-dt/We).*tau(l)+dt/dx*(uold(l+1)-uold(l));
  end;
  % Drawings
  plot(mesh',u,mesh',[(Eps/We)*tau;(Eps/We)*tau(I)]);
  axis([0 1 -1 1.2]);
  drawnow;
end;
legend('velocity','stress');
```

Exercise 1. Compare numerically and theoretically the stability of the two time-discretizations:

$$\begin{cases} \dfrac{\mathrm{Re}}{\delta t}(u_{n+1}(y)-u_n(y))-(1-\varepsilon)\dfrac{\partial^2 u_{n+1}}{\partial y^2}(y) \quad \dfrac{\partial \tau_n}{\partial y}(y), \\ \dfrac{1}{\delta t}(\tau_{n+1}(y)-\tau_n(y))+\dfrac{1}{\mathrm{We}}\tau_{n+1}(y) \quad \dfrac{\varepsilon}{\mathrm{We}}\dfrac{\partial u_{n+1}}{\partial y}, \end{cases} \tag{139}$$

and

$$\begin{cases} \dfrac{\mathrm{Re}}{\delta t}(u_{n+1}(y)-u_n(y))-(1-\varepsilon)\dfrac{\partial^2 u_{n+1}}{\partial y^2}(y) \quad \dfrac{\partial \tau_n}{\partial y}(y), \\ \dfrac{1}{\delta t}(\tau_{n+1}(y)-\tau_n(y))+\dfrac{1}{\mathrm{We}}\tau_{n+1}(y) \quad \dfrac{\varepsilon}{\mathrm{We}}\dfrac{\partial u_n}{\partial y}, \end{cases} \tag{140}$$

for zero Dirichlet boundary conditions on u_n.

Hint: For the numerics, choose a sufficiently large timestep. For the numerical analysis, consider the quantity $E_n \quad \mathrm{Re}\int_0^1 |u_n|^2(y)\,dy+\frac{\mathrm{We}}{\varepsilon}\int_0^1 |\tau_n|^2(y)\,dy$ and prove that $E_{n+1}\le E_n$, for a sufficiently small timestep for the scheme (139). Can you prove a similar result for the scheme (140) ? How to modify these schemes to obtain a stable scheme whatever the timestep ?

Below is the MATLAB source Couette_MC_VarReduc.m for the discretization of (137).

```
clear all;

% Physical parameters
Re=0.1;Eps=0.9;We=0.5;
v=1.;
T=1; % Maximal time

% Numerical parameters
% Space
I=100;
dx=1/I;mesh=[0:dx:1];
% Time
N=100;
dt=T/N;
% Number of polymers per cell (Monte Carlo)
J=1000;

% Matrices
D1=diag(ones(1,I-1),-1);D1=D1(2:I,:);D1=[D1,zeros(I-1,1)];
D2=diag(ones(1,I-1));D2=[zeros(I-1,1),D2,zeros(I-1,1)];
D3=diag(ones(1,I-1),+1);D3=D3(1:(I-1),:);D3=[zeros(I-1,1),D3];
% Mass matrix
M=(1/6)*D1+(2/3)*D2+(1/6)*D3;
M=M.*dx;M=sparse(M);
MM=M(:,2:I);
% Stiffness matrix
A=(-1)*D1+2*D2+(-1)*D3;
A=A./dx;A=sparse(A);
AA=A(:,2:I);
BB=Re*MM./dt+(1-Eps)*AA;

% Vectors
u=zeros(I+1,1); % Initial velocity
Y=zeros(J,1);X=zeros(J,I);
X_var_controle=zeros(J,1); % Control variate
Y=randn(size(Y));
% Initial condition not depending on the space variable
```

```
X=randn(J,1)*ones(1,I);
X_var_controle=X(:,1);
tau=zeros(I,1);
gradtau=zeros(I-1,1);
CLL=zeros(I+1,1);

% Time iterations
for t=dt:dt:T,
  for l=1:I,
        tau(l)=sum(Y.*(X(:,l)-X_var_controle))/J;
  end;
  tau=(Eps/We)*tau;
  gradtau=tau(2:I)-tau(1:(I-1));
  if ((t/T)<0.1)
        CLL(1)=v*10*(t/T);
  else
         CLL(1)=v ;
  end;
  CL=(Re*M./dt+(1-Eps)*A)*CLL;
  F=(Re*M./dt)*u-CL+gradtau;
  u(2:I)=BB\F;
  if ((t/T)<0.1)
        u(1)=v*10*(t/T);
  else
         u(1)=v;
  end;
  % Y, X and X_var_controle
  r=randn(J,1);
  for l=1:I,
        X(:,l)=(1-dt/(2*We))*X(:,l)+(dt/dx)*(u(l+1)-u(l))*Y+sqrt(dt/We)*r;
  end;
  X_var_controle=(1-dt/(2*We))*X_var_controle+sqrt(dt/We)*r;
  Y=(1-dt/(2*We))*Y+sqrt(dt/We)*randn(J,1);
  % Drawings
  plot(mesh',u,mesh',[tau;tau(I)]);
  axis([0 1 -1 1.2]);
  drawnow;
end;
legend('velocity','stress');
```

Exercise 2. Investigate numerically the influence of the number of dumbbells in each cell. Compare the results *with* and *without* variance reduction. Modify the program to use Brownian motions V_t for X_t which are independent from one cell to another (again with and without variance reduction). Discuss the results (see Sect. 6.3 below).

Exercise 3. Modify the program to treat FENE dumbbells. You can use either an Euler scheme to discretize the SDE and a rejection step, or the scheme (135). Program a variance reduction using the FENE-P model for the control variate.

The MATLAB source `Couette_FP.m` for the discretization of (138) follows.

```
clear all;

%%%% This file contains some integrals of Hermite polynomials
run Ortho_HD_normalise_20

%%%% Physical parameters
d=2; % dimension of the ambiant space
n=1; % number of springs
% Warning: Only d=2 and n=1 are implemented here
```

```
T=1; % Maximal time
Re=0.1;Eps=0.9;We=0.5;v=1.;

%%%% Discretization
% Space
I_esp=100; % number of spacesteps
dx=1/I_esp;mesh=0:dx:1;
% Time
N=100; % number of timesteps
dt=T/N; % timstep

l_max=2; % Maximal degree of Hermite polynomials
% Discretisation for q: FULL TENSOR PRODUCT
dim=(l_max+1)*(l_max+1);
disp('Dimension of the Galerkin basis for Fokker-Planck:');disp(dim);

% To get the tensorial index as a function of the absolute index
% 0 \leq nu(1) \leq l_max
get_nu=@(i) [ floor((i-1)/(l_max+1)), i-1-floor((i-1)/(l_max+1))*(l_max+1)];

% To get the absolute index as a function of the tensorial index
% 1 \leq i \leq dim
get_i=@(nu) 1+nu(1)*(l_max+1)+nu(2);

% Matrix S
D1=diag(ones(1,n*d),-d);D1=D1((d+1):(n+1)*d,:);
D2=diag(ones(1,n*d),d);D2=D2(1:n*d,:);
S=-D1+D2;
% Matrix D
D=S*S';
% Here, D=2 Id

%%%% Operators
disp('Computing matrices...');
%%%% Operators for Fokker-Planck
M=zeros(dim,dim);
G_de_base=zeros(dim,dim);
A=zeros(dim,dim);
% \int_X  (1/dt q^{n+1}) r \omega
% since int_P_P= Id, this is only Id
M=eye(dim,dim);
% G =  Nabla_u : \int_X  ( \nabla_X r \otimes X ) q \omega
% G depends on the timestep
% G=nabla_u*G_de_base where nabla_u is the off-diagonal component
% of the matrix \nabla u
for i=1:dim % r_i
  for j=1:dim % q_j
    % +1 : to get the indices of Ortho_HD_normalise.m
    nu_i=get_nu(i)+1;
    nu_j=get_nu(j)+1;
    G_de_base(i,j)=int_DP_P(nu_i(1),nu_j(1))*int_P_X_P(nu_i(2),nu_j(2));
  end
end
% A =  D : \int_X  ( \nabla_X q \otimes \nabla_X r) \omega
% Here, D=2 Id
% D(1,1) * \int \partial_{X_1}P_{i}(x) \partial_{X_1}P_{j}(x) \omega
for i=1:dim % r_i
    for j=1:dim % q_j
        nu_i=get_nu(i)+1;
        nu_j=get_nu(j)+1;
        A(i,j)=A(i,j)+D(1,1)*int_DP_DP(nu_i(1),nu_j(1))...
               *int_P_P(nu_i(2),nu_j(2));
    end
end
% D(2,2) * \int \partial_{X_2}P_{i}(x) \partial_{X_2}P_{j}(x) \omega
for i=1:dim % r_i
    for j=1:dim % q_j
```

```
        nu_i=get_nu(i)+1;
        nu_j=get_nu(j)+1;
        A(i,j)=A(i,j)+D(2,2)*int_P_P(nu_i(1),nu_j(1))...
               *int_DP_DP(nu_i(2),nu_j(2));
    end
end
% Computation of \int X_1 X_2 P_{i}(x) \omega
% This vector is useful to compute tau
int_X_X_q_1_2=zeros(dim,1);
for i=1:dim
  nu_i=get_nu(i)+1;
  int_X_X_q_1_2(i)=int_X_P(nu_i(1))*int_X_P(nu_i(2));
end;
%%%% Operators for the velocity
D1=diag(ones(1,I_esp-1),-1);D1=D1(2:I_esp,:);D1=[D1,zeros(I_esp-1,1)];
D2=diag(ones(1,I_esp-1));D2=[zeros(I_esp-1,1),D2,zeros(I_esp-1,1)];
D3=diag(ones(1,I_esp-1),+1);D3=D3(1:(I_esp-1),:);D3=[zeros(I_esp-1,1),D3];
% Mass matrix
M_esp=(1/6)*D1+(2/3)*D2+(1/6)*D3;
M_esp=M_esp.*dx;
M_esp=sparse(M_esp);
MM_esp=M_esp(:,2:I_esp);
% Stiffness matrix
A_esp=(-1)*D1+2*D2+(-1)*D3;
A_esp=A_esp./dx;
A_esp=sparse(A_esp);
AA_esp=A_esp(:,2:I_esp);
BB_esp=Re*MM_esp./dt+(1-Eps)*AA_esp;
%%%% Vectors
% initial conditions
u=zeros(I_esp+1,1); % velocity is zero
q=zeros(dim,I_esp);
q(1,:)=ones(1,I_esp); % equilibrium at each point
tau=zeros(I_esp,1);
gradtau=zeros(I_esp-1,1);
nabla_u=0;
CLL=zeros(I_esp+1,1);

%%%% Time iterations
disp('Time iterations');
for t=dt:dt:T,
  q_old=q;
  u_old=u;
  % Computation of u
  gradtau=tau(2:I_esp)-tau(1:(I_esp-1));
  if ((t/T)<0.1)
        CLL(1)=v*10*(t/T);
  else
         CLL(1)=v ;
  end;
  CL=(Re*M_esp./dt+(1-Eps)*A_esp)*CLL;
  F=(Re*M_esp./dt)*u-CL+gradtau;
  u(2:I_esp)=BB_esp\F;
  if ((t/T)<0.1)
        u(1)=v*10*(t/T);
  else
         u(1)=v;
  end;
  % computation of tau
  for l=1:I_esp % iteration on the cells
        nabla_u=(u(l+1)-u(l))/dx;
        nabla_u_old=(u_old(l+1)-u_old(l))/dx;
        % computation of q(:,l)
        G=nabla_u*G_de_base;
        G_old=nabla_u_old*G_de_base;
        % Crank Nicholson
        M_n_p_1=(1/dt)*M - 0.5*(G-A/(4*We));
        M_n=(1/dt)*M + 0.5*(G_old-A/(4*We));
        q(:,l)=M_n_p_1\(M_n*q_old(:,l));
        % Computation of tau(l)
        % tau = \int_X  ( X  \otimes X ) q \omega
        tau(l)=(Eps/We)*(q(:,l)'*int_X_X_q_1_2);
  end;
  % Drawings
  plot(mesh',u,mesh',[tau;tau(I_esp)]);
  axis([0 1 -1 1.2]);
  drawnow;
end;
legend('velocity','stress');
```

Exercise 4. Compare the results obtained with the three formulations. Which formulation is the most efficient computationally ? Discuss the applicability of these three formulations to the following two more general settings: chain of $N > 2$ beads linked with Hookean springs, FENE dumbbell model.

6 Mathematical and numerical issues

As mentioned earlier, the present section is much more elaborate mathematically than the preceeding sections.

6.1 Overview of the main difficulties

Let us first formally summarize the difficulties raised by the mathematical analysis of systems such as (50) and (98) (for micro-macro models) or (23) (for macro-macro models).

These systems of equations include the Navier-Stokes equations, with the additional term $\operatorname{div} \boldsymbol{\tau}_p$ in the right-hand side. The equation on $\boldsymbol{\tau}_p$ is essentially a transport equation and, formally, $\boldsymbol{\tau}_p$ has at most the regularity of $\nabla \boldsymbol{u}$ (this fact will be clear in the choice of appropriate functional spaces for existence results, and of the discretization spaces for numerical methods). The term $\operatorname{div} \boldsymbol{\tau}_p$ in the right-hand side in the momentum equation is not likely to bring more regularity on $\boldsymbol{u}$. It is thus expected that the study of these coupled systems contains at least the well-known difficulties of the Navier-Stokes equations. Recall that for the (3-dimensional) Navier-Stokes equations, it is known that global-in-time weak solutions exist but the regularity, and thus the uniqueness, of such solutions is an open problem. Only local-in-time existence and uniqueness results of strong solutions are available.

In addition to the difficulties already contained in the Navier-Stokes equations (which essentially originate from the Navier term $\boldsymbol{u} \cdot \nabla \boldsymbol{u}$), the coupling with the equation on $\boldsymbol{\tau}_p$ raises other problems. First, these equations (both for macro-macro and micro-macro models) contain a transport term ($\boldsymbol{u} \cdot \nabla \boldsymbol{\tau}_p$, $\boldsymbol{u} \cdot \nabla \psi$ or $\boldsymbol{u} \cdot \nabla \boldsymbol{X}_t$) without diffusion terms (in the space variable). They are hyperbolic in nature. The regularity on the velocity $\boldsymbol{u}$ is typically not sufficient to treat this transport term by a characteristic method. Moreover, these equations involve a nonlinear multiplicative term ($\nabla \boldsymbol{u} \nabla \boldsymbol{\tau}_p$, $\operatorname{div}_{\boldsymbol{X}} (\nabla \boldsymbol{u} \boldsymbol{X} \nabla \psi)$ or $\nabla \boldsymbol{u} \boldsymbol{X}_t$). Finally, except for very simple models

(Oldroyd-B or Hookean dumbbell), the equations defining $\boldsymbol{\tau}_p$ generally contain additional non-linearities (for micro-macro model, the force $\boldsymbol{F}$ is generally non-linear and typically blows up when the length of the polymer reaches a critical value).

To summarize, the difficulties raised by mathematical analysis of these models are related to:

- *transport terms*,
- *nonlinear terms* coming either from the coupling between the equations and $(\boldsymbol{u}, p)$ and $\boldsymbol{\tau}_p$, or inherently contained in the equations defining $\boldsymbol{\tau}_p$.

These difficulties limit the state-of-the-art mathematical well-posedness analysis to *local-in-time* existence and uniqueness results. They also have many implications on the numericals methods (choice of the discretization spaces, stability of the numerical schemes, ...). Actually, the problems raised by the discretization we mentioned in Sect. 4.4 can be seen as counterparts of the difficulties raised by the mathematical analysis. Many questions are still open, and the mathematical analysis and the numerical analysis for viscoelastic fluids are very lively fields.

In the following, we provide more detailed results for macro-macro models, and, next, micro-macro models. Considering the focus of the present article, more emphasis is laid on the latter.

6.2 Macroscopic models

We refer to M. Renardy [112] or E. Fernandez-Cara, F. Guillen and R.R. Ortega [44] for a review of the mathematical analysis of macroscopic models. For the numerical methods, we refer to R. Keunings [70] F.P.T. Baaijens [6] R. Owens and T. Phillips [104]. We recall the prototypical macroscopic model, namely the Oldroyd-B model:

$$\begin{cases} \mathrm{Re}\left(\dfrac{\partial \boldsymbol{u}}{\partial t} + \boldsymbol{u}\cdot\nabla\boldsymbol{u}\right) - (1-\varepsilon)\Delta\boldsymbol{u} + \nabla p \quad \operatorname{div}\boldsymbol{\tau}_p + \boldsymbol{f}, \\ \operatorname{div}\boldsymbol{u} \quad 0, \\ \mathrm{We}\left(\dfrac{\partial \boldsymbol{\tau}_p}{\partial t} + \boldsymbol{u}\cdot\nabla\boldsymbol{\tau}_p - \nabla\boldsymbol{u}\boldsymbol{\tau}_p - \boldsymbol{\tau}_p(\nabla\boldsymbol{u})^T\right) + \boldsymbol{\tau}_p \quad \varepsilon(\nabla\boldsymbol{u} + \nabla\boldsymbol{u}^T). \end{cases} \tag{141}$$

Mathematical results

Concerning existence results for macroscopic models, four types of results can be found in the litterature:

- local-in-time results (perturbation of the initial condition),
- global-in-time results for small data (perturbation of the stationary solution),
- existence results for stationary solutions close to equilibrium solutions,
- existence results for stationary solutions close to Navier-Stokes stationary solutions.

For illustration, let us only mention the result obtained by M. Renardy in [110]. The author considers the following coupled problem, in a bounded domain $\mathscr{D}$ of $\mathbb{R}^3$:

$$\begin{cases} \rho\left(\dfrac{\partial \boldsymbol{u}}{\partial t}+\boldsymbol{u}.\nabla\boldsymbol{u}\right) \quad \operatorname{div}\boldsymbol{\tau}_p-\nabla p+\boldsymbol{f}, \\ \operatorname{div}\boldsymbol{u} \quad 0, \\ \left(\dfrac{\partial}{\partial t}+\boldsymbol{u}.\nabla\right)(\boldsymbol{\tau}_p)_{i,j} \quad \boldsymbol{A}_{i,j,k,l}(\boldsymbol{\tau}_p)\dfrac{\partial \boldsymbol{u}_k}{\partial \boldsymbol{x}_l}+\boldsymbol{g}_{i,j}(\boldsymbol{\tau}_p), \end{cases} \tag{142}$$

with summation convention on repeated indices. The fluid is inviscid ($\eta \quad 0$). This system is supplied with homogeneous Dirichlet boundary condition on the velocity $\boldsymbol{u}$, and initial conditions. The differential models introduced in Sect. 2.3 indeed enter this framework. Introduce the fourth order tensor:

$$\boldsymbol{C}_{i,j,k,l} \quad \boldsymbol{A}_{i,j,k,l}-(\boldsymbol{\tau}_p)_{i,l}\delta_{k,j}, \tag{143}$$

where δ is the Kronecker symbol. Assume the following strong ellipticity property on $\boldsymbol{C}$: $\forall\zeta,\eta\in\mathbb{R}^3$

$$\boldsymbol{C}_{i,j,k,l}(\boldsymbol{\tau}_p)\zeta_i\zeta_k\eta_j\eta_l \geq \kappa|\zeta|^2|\eta|^2 \tag{144}$$

where $\kappa>0$ is a constant not depending on $\boldsymbol{\tau}_p$. Under additional assumptions of symmetry on the tensor $\boldsymbol{A}$, of regularity and compatibility on the initial conditions, it is shown by M. Renardy in [110] that:

Theorem 1. *There exists a time $T'>0$, such that the system (142) admits a unique solution with regularity:*

$$\boldsymbol{u}\in\bigcap_{k\ \ 0}^{4}\mathscr{C}^k(\ 0,T'\ ,H^{4-k}(\mathscr{D})),\ \boldsymbol{\tau}_p\in\bigcap_{k\ \ 0}^{3}\mathscr{C}^k(\ 0,T'\ ,H^{3-k}(\mathscr{D})).$$

The works of C. Guillopé and J.C. Saut [53, 54] are also to be mentioned. Existence results for less regular solutions are obtained there for non-zero viscosity of the solvent $\eta>0$. In a series of works, E. Fernandez-Cara, F. Guillen and R.R. Ortega study the local well-posedness in Sobolev spaces (see [44] and references therein).

We also mention the work of F.-H. Lin, C. Liu and P.W. Zhang [86] where local-in-time existence and uniqueness results and global-in-time existence and uniqueness results for small data are proven for Oldroyd-like models.

The only global-in-time existence result we are aware of is the work of P.-L. Lions and N. Masmoudi [89] where an Oldroyd-like model is studied, but with the corotational convective derivative on the stress tensor rather than the upper convected derivative.

Besides, there exist many studies on the stability of viscoelastic flows, and the change of mathematical nature of the equations (transition from parabolic to hyperbolic). We refer to M. Renardy [112], R. Owens and T. Phillips [104] and references therein.

Numerical methods

Most of the numerical methods employed in practice to simulate such models are based upon a finite element discretization in space (see however R. Owens and T. Phillips [104] for spectral methods) and a finite difference discretization in time (usually Euler schemes), with a decoupled computation of $(\boldsymbol{u}, p)$ and $\boldsymbol{\tau}_p$. More precisely, at each timestep, the equation for $(\boldsymbol{u}, p)$ is first solved, given the current stress tensor $\boldsymbol{\tau}_p$. This allows to update the velocity. Next, the equation for $\boldsymbol{\tau}_p$ is solved, and the stress is updated.

We have already mentioned in Sect. 4.4 the main three difficulties raised by the discretization: (i) a compatibility condition is needed between the discretization spaces for $\boldsymbol{u}$ and for $\boldsymbol{\tau}_p$, (ii) the transport terms need to be correctly discretized, (iii) the discretization of the nonlinear terms requires special attention. Let us now briefly describe how to deal with these difficulties for macroscopic models. Notice that, as observed in Sect. 4.4, the three difficulties mentioned above are also present for the discretization of micro-macro models. Most of the methods described below are thus also useful for the discretization of micro-macro models.

Concerning difficulty (i), it actually appears that an *inf-sup condition* is required for the three discretization spaces for respectively the pressure, the velocity and the stress tensor. More precisely, in addition to the usual inf-sup condition required for the discretization spaces for the velocity and the pressure, a compatibility between the discretization space for the velocity and that for the stress tensor is required to obtain stable schemes when η is small as compared to η_p (*i.e.* when ε is close to 1). These compatibility conditions have been analyzed by J.C. Bonvin M. Picasso and R. Sternberg in [18, 17] on the three-field Stokes system:

$$\begin{cases} -\eta \Delta \boldsymbol{u} + \nabla p - \operatorname{div} \boldsymbol{\tau}_p \quad \boldsymbol{f}, \\ \operatorname{div} \boldsymbol{u} \quad 0, \\ \boldsymbol{\tau}_p - \eta_p \dot{\boldsymbol{\gamma}} \quad \boldsymbol{g}. \end{cases} \tag{145}$$

Many methods have been proposed in the literature to treat the problem:

- Use discretization spaces that satisfy an inf-sup condition. These are usually difficult to implement (see for example J.M. Marchal and M.J. Crochet [96]),
- Introduce an additional unknown to avoid this compatibility condition (see the EVSS method in R. Guénette and M. Fortin [52]),
- Use stabilization methods, like the *Galerkin Least Square* (GLS) method, which enables to use the same discretization space for the three unknown fields (see J.C. Bonvin M. Picasso and R. Sternberg in [18, 17]).

The second difficulty (ii) is raised by the discretization of the advection terms both in the equation for $\boldsymbol{u}$ and for $\boldsymbol{\tau}_p$. It is well known that naïve discretization by a finite element method leads to unstable schemes. Many techniques have been used to circumvent this problem: stabilization techniques like *Streamline Upwind Petrov-Galerkin* (SUPG) or GLS, *Discontinuous Galerkin* methods (see M. Fortin and A. Fortin [46]), or numerical characteristic method (see J.C. Bonvin [18] or the *Backward-tracking Lagrangian Particle Method* of P. Wapperom, R. Keunings and

V. Legat [121]). We refer to R. Owens and T. Phillips [104, Chap. 7] or to R. Keunings [71] for references about these methods in the context of viscoelastic fluid simulations (see also T. Min, J.Y. Yoo and H. Choi [98] for a comparison between various numerical schemes). These difficulties are prominent for high Reynolds number (which is not practically relevant in the context of viscoelastic fluid simulations) or for high Weissenberg number (which *is* relevant).

The third difficulty (iii) we mentioned concerns the discretization of the nonlinear terms. Consider the term $\nabla \boldsymbol{u} \boldsymbol{\tau}_p + \boldsymbol{\tau}_p (\nabla \boldsymbol{u})^T$ in the convective derivative of $\boldsymbol{\tau}_p$. In most of the numerical methods, this term is treated explicitly by taking its value at the former timestep. Linearizing this term by treating the velocity explicitly and the stress implicitly leads to an ill-posed problem if the Weissenberg problem is too high.

We mentioned that two of these difficulties are prominent for large Weissenberg number. It indeed appears that numerical methods become unstable in this latter regime. This is the so-called High Weissenberg Number Problem (HWNP) we already mentioned in Sect. 4.4. Many works are related to the HWNP (we refer for example to R. Owens and T. Phillips [104, Chap. 7]). The HWNP is certainly not only related to the discretization scheme. It has indeed been observed that for some geometries, the critical Weissenberg number (above which the scheme is unstable) decreases with the mesh step size (see R. Keunings [71]), which could indicate a loss of regularity for the continuous solution itself (see D. Sandri [116]). It is still an open problem to precisely characterize the HWNP, and to distinguish between instability coming from the model itself, or its discretization. For the theoretical study of the limit We → ∞, we refer to M. Renardy [112, Chap. 6].

We would like to mention the recent works [42, 43, 60] where R. Fattal, R. Kupferman and M.A. Hulsen propose a new formulation for macroscopic models based on a change of variable: instead of using $(\boldsymbol{u}, p, \boldsymbol{\tau}_p)$ as unknowns, they set the problem in terms of $(\boldsymbol{u}, p, \boldsymbol{\phi})$, where

$$\boldsymbol{\phi} \quad \ln \boldsymbol{A}$$

and $\boldsymbol{A}$ is the conformation tensor defined by:

$$\boldsymbol{A} \quad \frac{\mathrm{We}}{\varepsilon} \boldsymbol{\tau}_p + \mathrm{Id}. \tag{146}$$

This new formulation was implemented in R. Fattal, R. Kupferman and M.A. Hulsen [43, 60] and Y. Kwon [74] for various models, various geometric settings, and various numerical methods. In this alternate formulation, the numerical instability arises only for much higher a Weissenberg number. It thus seems to be a promising method to better understand the problem.

6.3 Multiscale models

Let us recall the micro-macro model we are interested in:

$$\left\{\begin{array}{l}\mathrm{Re}\left(\dfrac{\partial \boldsymbol{u}}{\partial t}+\boldsymbol{u}\cdot\nabla\boldsymbol{u}\right)-(1-\varepsilon)\Delta\boldsymbol{u}+\nabla p \quad \mathrm{div}\,\boldsymbol{\tau}_p+\boldsymbol{f},\\ \mathrm{div}\,\boldsymbol{u} \quad 0,\\ \boldsymbol{\tau}_p \quad \dfrac{\varepsilon}{\mathrm{We}}\left(\mathbb{E}(\boldsymbol{X}_t\otimes\boldsymbol{F}(\boldsymbol{X}_t))-\mathrm{Id}\right),\\ d\boldsymbol{X}_t+\boldsymbol{u}.\nabla\boldsymbol{X}_t\,dt \quad \nabla\boldsymbol{u}\boldsymbol{X}_t\,dt-\dfrac{1}{2\mathrm{We}}\boldsymbol{F}(\boldsymbol{X}_t)\,dt+\dfrac{1}{\sqrt{\mathrm{We}}}d\boldsymbol{W}_t,\end{array}\right. \tag{147}$$

with $\boldsymbol{F}(\boldsymbol{X}_t) \quad \boldsymbol{X}_t$ for Hookean dumbbells, $\boldsymbol{F}(\boldsymbol{X}_t) \quad \frac{\boldsymbol{X}_t}{1-\|\boldsymbol{X}_t\|^2/b}$ for FENE dumbbells, or $\boldsymbol{F}(\boldsymbol{X}_t) \quad \frac{\boldsymbol{X}_t}{1-\mathbb{E}(\|\boldsymbol{X}_t\|^2)/b}$ for FENE-P dumbbells. The space variable $\boldsymbol{x}$ varies in a bounded domain $\mathscr{D}\subset\mathbb{R}^d$. This system is supplied with boundary conditions on the velocity, and initial conditions on the velocity and the stochastic processes. In the following, we suppose $\varepsilon\in(0,1)$.

We recall the Fokker-Planck version of (147):

$$\left\{\begin{array}{l}\mathrm{Re}\left(\dfrac{\partial \boldsymbol{u}}{\partial t}(t,\boldsymbol{x})+\boldsymbol{u}(t,\boldsymbol{x})\cdot\nabla\boldsymbol{u}(t,\boldsymbol{x})\right)-(1-\varepsilon)\Delta\boldsymbol{u}(t,\boldsymbol{x})+\nabla p(t,\boldsymbol{x})\\ \qquad \mathrm{div}\,(\boldsymbol{\tau}_p(t,\boldsymbol{x})),\\ \mathrm{div}\,(\boldsymbol{u}(t,\boldsymbol{x})) \quad 0,\\ \boldsymbol{\tau}_p(t,\boldsymbol{x}) \quad \dfrac{\varepsilon}{\mathrm{We}}\left(\displaystyle\int_{\boldsymbol{X}}(\boldsymbol{X}\otimes\boldsymbol{F}(\boldsymbol{X}))\psi(t,\boldsymbol{x},\boldsymbol{X})\,d\boldsymbol{X}-\mathrm{Id}\right),\\ \dfrac{\partial\psi}{\partial t}(t,\boldsymbol{x},\boldsymbol{X})+\boldsymbol{u}.\nabla_{\boldsymbol{x}}\psi(t,\boldsymbol{x},\boldsymbol{X})\\ \qquad -\mathrm{div}_{\boldsymbol{X}}\left(\left(\nabla\boldsymbol{u}(t,\boldsymbol{x})\boldsymbol{X}-\dfrac{1}{2\mathrm{We}}\boldsymbol{F}(\boldsymbol{X})\right)\psi(t,\boldsymbol{x},\boldsymbol{X})\right)+\dfrac{1}{2\mathrm{We}}\Delta_{\boldsymbol{X}}\psi(t,\boldsymbol{x},\boldsymbol{X}).\end{array}\right. \tag{148}$$

There is a growing literature on the analysis of micro-macro models for polymeric fluids. The first work we are aware of is M. Renardy [111], where the micro-macro model in its Fokker-Planck formulation (50) is studied. Since this early work, many groups have studied these models, perhaps because they are prototypical for a class of multiscale models, where some parameters needed in the macroscopic equations are computed by some microscopic models (see the general formulation (52)).

Let us recall the two main difficulties we already mentioned in Sect. 6.1,

- *transport terms* ($\boldsymbol{u}\cdot\nabla\boldsymbol{u}$, $\boldsymbol{u}\cdot\nabla\boldsymbol{X}_t$ and $\boldsymbol{u}.\nabla\psi$),
- *nonlinear terms* coming either from the coupling between the equations and $(\boldsymbol{u},p)$ and $\boldsymbol{\tau}_p$ ($\nabla\boldsymbol{u}\boldsymbol{X}_t$ or $\mathrm{div}_{\boldsymbol{X}}\,(\nabla\boldsymbol{u}\boldsymbol{X}\psi)$), or inherently contained in the equations defining $\boldsymbol{\tau}_p$ (due to the non-linear entropic force $\boldsymbol{F}$).

In the next sections, we explain how these difficulties have been addressed both from the mathematical viewpoint and the numerical viewpoint (see also T. Lelièvre [82], and T. Li and P.W. Zhang [85]).

Simplifications of the equations

The system (147) is quite difficult to study as such. Two simplifications of this general setting are usually considered for preliminary arguments: homogeneous flows and shear flows.

To specifically study the microscopic equations, one can consider *homogeneous flows*. We recall that in such flows, $\nabla \boldsymbol{u}$ does not depend on the space variable, and therefore $\boldsymbol{X}_t$ (and thus $\boldsymbol{\tau}_p$) does not depend on the space variable either. A solution to (147) is then obtained by solving the SDE without the advective term. For a velocity field $\boldsymbol{u}(t,x) \quad \boldsymbol{\kappa}(t)\boldsymbol{x}$, (147) becomes:

$$\begin{cases} \boldsymbol{\tau}_p \quad \dfrac{\varepsilon}{\mathrm{We}}\left(\mathbb{E}(\boldsymbol{X}_t \otimes \boldsymbol{F}(\boldsymbol{X}_t)) - \mathrm{Id}\right), \\ d\boldsymbol{X}_t \quad \boldsymbol{\kappa}(t)\boldsymbol{X}_t\,dt - \dfrac{1}{2\mathrm{We}}\boldsymbol{F}(\boldsymbol{X}_t)\,dt + \dfrac{1}{\sqrt{\mathrm{We}}}d\boldsymbol{W}_t. \end{cases} \tag{149}$$

To keep the difficulty related to the coupling between the macroscopic equation and the microscopic equations but to eliminate the difficulties related to transport terms, many authors (see M. Laso and H.C. Öttinger [75], J.C. Bonvin and M. Picasso [16], C. Guillopé and J.C. Saut [54], B. Jourdain, C. Le Bris and T. Lelièvre [68] or W. E, T. Li and P.W. Zhang [38]) consider shear flows (see Fig. 1). In this geometry, (147) writes:

$$\begin{cases} \mathrm{Re}\,\dfrac{\partial u}{\partial t}(t,y) - (1-\varepsilon)\dfrac{\partial^2 u}{\partial y^2}(t,y) \quad \dfrac{\partial \tau}{\partial y}(t,y) + f(t,y), \\ \tau(t,y) \quad \dfrac{\varepsilon}{\mathrm{We}}\mathbb{E}(X_t(y)F_Y(\boldsymbol{X}_t(y))), \\ dX_t(y) \quad \dfrac{\partial u}{\partial y}(t,y)Y_t(y)\,dt - \dfrac{1}{2\mathrm{We}}F_X(\boldsymbol{X}_t(y))\,dt + \dfrac{1}{\sqrt{\mathrm{We}}}dV_t, \\ dY_t(y) \quad -\dfrac{1}{2\mathrm{We}}F_Y(\boldsymbol{X}_t(y))\,dt + \dfrac{1}{\sqrt{\mathrm{We}}}dW_t, \end{cases} \tag{150}$$

where $(X_t(y), Y_t(y))$ are the two components of the stochastic process $\boldsymbol{X}_t(y)$, (V_t, W_t) are two independent Brownian motions and $(F_X(\boldsymbol{X}_t), F_Y(\boldsymbol{X}_t))$ are the two components of the force $\boldsymbol{F}(\boldsymbol{X}_t)$. In this case, $y \in (0,1)$, and Dirichlet boundary conditions are assumed on the velocity at $y \quad 0$ and $y \quad 1$. The initial conditions (X_0, Y_0) are assumed to be independent from one another and independent from the Brownian motions.

Mathematical Analysis

A fundamental energy estimate

In order to understand the mathematical structure of the system (147), we first derive an *energy estimate*. Such an estimate is called an *a priori* estimate, since it is formally derived assuming sufficient regularity on the solutions for all the manipulations to hold true. These estimates are then used to prove existence and uniqueness results, and, possibly, study longtime properties of the solutions.

Multiplying the momentum equation by $\boldsymbol{u}$ and integrating in space and time, one obtains on the one hand

$$\begin{aligned} &\frac{\mathrm{Re}}{2}\int_{\mathscr{D}} |\boldsymbol{u}|^2(t,\boldsymbol{x}) + (1-\varepsilon)\int_0^t \int_{\mathscr{D}} |\nabla \boldsymbol{u}|^2(s,\boldsymbol{x}) \\ &\quad \frac{\mathrm{Re}}{2}\int_{\mathscr{D}} |\boldsymbol{u}|^2(0,\boldsymbol{x}) - \frac{\varepsilon}{\mathrm{We}}\int_0^t \int_{\mathscr{D}} \mathbb{E}(\boldsymbol{X}_s(\boldsymbol{x}) \otimes \boldsymbol{F}(\boldsymbol{X}_s(\boldsymbol{x}))) : \nabla \boldsymbol{u}(s,\boldsymbol{x}), \end{aligned} \tag{151}$$

assuming homogeneous Dirichlet boundary conditions on $\boldsymbol{u}$.

On the other hand, using Itô calculus on $\Pi(\boldsymbol{X}_t)$ (where Π is the potential of the force $\boldsymbol{F}$ of the spring), integrating in space, time and taking the expectation value, it is seen that

$$\begin{aligned}\int_{\mathscr{D}} \mathbb{E}(\Pi(\boldsymbol{X}_t(\boldsymbol{x}))) + \frac{1}{2\mathrm{We}} \int_0^t \int_{\mathscr{D}} \mathbb{E}(\|\boldsymbol{F}(\boldsymbol{X}_s(\boldsymbol{x}))\|^2) \\ \int_{\mathscr{D}} \mathbb{E}(\Pi(\boldsymbol{X}_0(\boldsymbol{x}))) + \int_0^t \int_{\mathscr{D}} \mathbb{E}(\boldsymbol{F}(\boldsymbol{X}_s(\boldsymbol{x})) \cdot \nabla \boldsymbol{u}(s,\boldsymbol{x}) \boldsymbol{X}_s(\boldsymbol{x})) \\ + \frac{1}{2\mathrm{We}} \int_0^t \int_{\mathscr{D}} \Delta \Pi(\boldsymbol{X}_s(\boldsymbol{x})).\end{aligned} \tag{152}$$

Summing up the two equalities (151) and (152), and using

$$\mathbb{E}(\boldsymbol{X}_s(\boldsymbol{x}) \otimes \boldsymbol{F}(\boldsymbol{X}_s(\boldsymbol{x}))) : \nabla \boldsymbol{u}(s,\boldsymbol{x}) \quad \mathbb{E}(\boldsymbol{F}(\boldsymbol{X}_s(\boldsymbol{x})) \cdot \nabla \boldsymbol{u}(s,\boldsymbol{x}) \boldsymbol{X}_s(\boldsymbol{x})), \tag{153}$$

the following energy estimate is obtained:

$$\begin{aligned}\frac{\mathrm{Re}}{2} \frac{d}{dt} \int_{\mathscr{D}} |\boldsymbol{u}|^2(t,\boldsymbol{x}) + (1-\varepsilon) \int_{\mathscr{D}} |\nabla \boldsymbol{u}|^2(t,\boldsymbol{x}) + \frac{\varepsilon}{\mathrm{We}} \frac{d}{dt} \int_{\mathscr{D}} \mathbb{E}(\Pi(\boldsymbol{X}_t(\boldsymbol{x}))) \\ + \frac{\varepsilon}{2\mathrm{We}^2} \int_{\mathscr{D}} \mathbb{E}(\|\boldsymbol{F}(\boldsymbol{X}_t(\boldsymbol{x}))\|^2) \quad \frac{\varepsilon}{2\mathrm{We}^2} \int_{\mathscr{D}} \Delta \Pi(\boldsymbol{X}_t(\boldsymbol{x})).\end{aligned} \tag{154}$$

Notice that this energy estimate does not help in the study of the longtime behavior since the term in the right-hand side (which comes form Itô calculus and is non-negative since Π is convex) brings energy to the system. We will return to this question below.

As said above, this energy estimate is a first step towards an existence and uniqueness result. For example, in the case of Hookean dumbbells in a shear flow, it allows to prove the following global-in-time existence and uniqueness result (see B. Jourdain, C. Le Bris and T. Lelièvre [67]):

Theorem 2. *Assuming $u_0 \in L^2_y$ and $f_{ext} \in L^1_t(L^2_y)$, the system (150) for Hookean dumbbells admits a unique solution (u,X) on $(0,T)$, $\forall T > 0$. In addition, the following estimate holds:*

$$\begin{aligned}\|u\|^2_{L^\infty_t(L^2_y)} + \|u\|^2_{L^2_t(H^1_{0,y})} + \|X_t\|^2_{L^\infty_t(L^2_y(L^2_\omega))} + \|X_t\|^2_{L^2_t(L^2_y(L^2_\omega))} \\ \leq C\left(\|X_0\|^2_{L^2_y(L^2_\omega)} + \|u_0\|^2_{L^2_y} + T + \|f_{ext}\|^2_{L^1_t(L^2_y)}\right).\end{aligned}$$

Notice that in this case, $Y_t \quad Y_0 e^{-t/2} + \int_0^t e^{\frac{s-t}{2}} dW_s$ is analytically known, so that the existence and uniqueness result only concerns (u,X). The notion of solution employed is: the equation on u is satisfied in the distribution sense and the SDE holds for almost every (y,ω). The proof relies on a variational formulation of the PDE, and follows a very classical line. It consists in (i) building a sequence of approximate solutions (by a Galerkin procedure), (ii) using the energy estimate (which indeed has then a rigorous, better than formal, meaning) to derive some bounds on this sequence from which one deduces the existence of a limit (up to the extraction of a subsequence), (iii) passing to the limit in the variational formulation of the PDE. This approach is interesting since, as is well known, it is also useful to prove the convergence of numerical methods based on variational formulations (such as finite element methods).

This setting (Hookean dumbbell in a shear flow) is actually extremely specific. A global-in-time existence and uniqueness result is obtained since the coupling term $\nabla \boldsymbol{u} \boldsymbol{X}_t$ of the original problem (147) simplifies to $\frac{\partial u}{\partial y} Y_t$ in (150), where Y_t is known independently of (u, X). In other words, this coupling term is no more nonlinear.

For FENE dumbbell, two new difficulties have to be addressed: first, the SDE contains an explosive drift term and second, even in a shear flow, the coupling term $\nabla \boldsymbol{u} \boldsymbol{X}_t$ is genuinely nonlinear.

The FENE SDE

In this paragraph, we consider the FENE SDE in a given homogeneous flow. As we mentioned earlier, the FENE force has been introduced to prevent the length of the dumbbell from exceeding the maximal length of the polymer. What can be actually proven is the following (see B. Jourdain and T. Lelièvre [66]):

Proposition 1. *Let us consider the SDE in* (149) *for FENE force:* $\boldsymbol{F}(\boldsymbol{X})$ $\frac{\boldsymbol{X}}{1-\|\boldsymbol{X}\|^2/b}$.

- *For* $\boldsymbol{\kappa} \in L^1_{\mathrm{loc}}(\mathbb{R}_+)$ *and* $b > 0$, *this SDE admits a strong solution with values in B* $\mathscr{B}(0, \sqrt{b})$, *which is unique in the class of solutions with values in B* $\mathscr{B}(0, \sqrt{b})$.
- *Assume* $\boldsymbol{\kappa} \in L^2(\mathbb{R}_+)$. *If* $b \geq 2$, *then the solution does not touch the boundary of B in finite time. If* $0 < b < 2$, *The solution touches (a.s.) the boundary of B in finite time.*
- *Take* $\boldsymbol{\kappa} \equiv 0$ *(for simplicity) and* $0 < b < 2$. *It is possible to build two different stochastic processes satisfying the SDE.*

In practice, b is typically larger than 10, so that the SDE has indeed a unique strong solution.

The FENE model in a Couette flow

As mentioned above, for the FENE model in the Couette flow, the coupling term $\frac{\partial u}{\partial y} Y_t$ is indeed nonlinear since Y_t depends on X_t (through the force term $F_Y(\boldsymbol{X}_t)$) and thus on u. This nonlinearity implies additional difficulties in the existence result, and the *a priori* estimate we derived above does not provide enough regularity on the velocity to pass to the limit in the nonlinear term $\frac{\partial u}{\partial y}(t, y) Y_t$.

The question is then: for a given regularity of u (say $u \in L^\infty_t(L^2_y) \cap L^2_t(H^1_{0,y})$ if we consider the first energy estimate), what is the regularity of τ ? Formally, owing to the presence of the nonlinear term $\nabla \boldsymbol{u} \boldsymbol{X}_t$ in the SDE, τ has the regularity of $\exp(\int_0^t \frac{\partial u}{\partial y})$ which may be very irregular if one only assumes $u \in\in L^\infty_t(L^2_y) \cap L^2_t(H^1_{0,y})$.

One way to address this difficulty is to derive additional *a priori* regularity on the velocity. This can be performed by multiplying the equation for u in (150) by $-\frac{\partial^2 u}{\partial y^2}$ and using Girsanov theorem to explicitly obtain the dependency of τ in terms of u:

$$\begin{aligned}\tau(t,y) \quad & \mathbb{E}\left(\frac{X_t(y)Y_t(y)}{1-\frac{(X_t(y))^2+(Y_t(y))^2}{b}}\right),\\ & \mathbb{E}\left(\left(\frac{\tilde{X}_t\tilde{Y}_t}{1-\frac{\tilde{X}_t^2+\tilde{Y}_t^2}{b}}\right)\mathscr{E}\left(\frac{1}{\sqrt{\mathrm{We}}}\int_0^{\bullet}\frac{\partial u}{\partial y}(y)\tilde{Y}_s\,dV_s\right)_T\right),\end{aligned} \tag{155}$$

where $\widetilde{\boldsymbol{X}}_t \quad (\tilde{X}_t,\tilde{Y}_t)$ is the stochastic process satisfying the FENE SDE with $\frac{\partial u}{\partial y} \quad 0$:

$$d\widetilde{\boldsymbol{X}}_t \quad -\frac{1}{2\mathrm{We}}\frac{\widetilde{\boldsymbol{X}}_t}{1-\|\widetilde{\boldsymbol{X}}_t\|^2/b}dt+\frac{1}{\sqrt{\mathrm{We}}}d\boldsymbol{W}_t,$$

and $\mathscr{E}$ is the exponential martingale:

$$\mathscr{E}\left(\frac{1}{\sqrt{\mathrm{We}}}\int_0^{\bullet}\frac{\partial u}{\partial y}\tilde{Y}_s\,dV_s\right)_t \quad \exp\left(\frac{1}{\sqrt{\mathrm{We}}}\int_0^t\frac{\partial u}{\partial y}\tilde{Y}_s\,dV_s-\frac{1}{2\mathrm{We}}\int_0^t\left(\frac{\partial u}{\partial y}\tilde{Y}_s\right)^2ds\right).$$

Owing to the exponential dependency of τ on u in (155), this additional *a priori* estimate yields bounds on u in $L^\infty_t(H^1_{0,y})\cap L^2_t(H^2_y)$-norm but only locally in time.

The following local-in-time existence and uniqueness result can then be proven (see B. Jourdain, C. Le Bris and T. Lelièvre [68]):

Theorem 3. *Under the assumptions $b>6$, $f_{ext}\in L^2_t(L^2_y)$ and $u_0\in H^1_y$, $\exists T>0$ (depending on the data) s.t. the system admits a unique solution (u,X,Y) on $0,T)$. This solution is such that $u\in L^\infty_t(H^1_{0,y})\cap L^2_t(H^2_y)$. In addition, we have:*

- $\mathbb{P}(\exists t>0,((X^y_t)^2+(Y^y_t)^2) \quad b) \quad 0,$
- (X^y_t,Y^y_t) *is adapted with respect to the filtration $\mathscr{F}^{V,W}_t$ associated with the Brownian motions.*

For a similar result in a more general setting (3-dimensional flow) and forces with polynomial growth, we refer to W. E, T. Li and P.W. Zhang [39]. The authors prove a local-in-time existence and uniqueness result in high Sobolev spaces. We also refer to A. Bonito, Ph. Clément and M. Picasso [15] for existence results for Hookean dumbbells, neglecting the advection terms. When the velocity field is not regular enough, it is difficult to give a sense to the transport term in the SDE (which is actually a Stochastic *Partial* Differential Equation). We refer to C. Le Bris and P.-L. Lions [78, 79].

Longtime behavior

As we mentioned above, the *a priori* estimate (154) cannot be used to understand the longtime behavior of the system because of the non-negative term $\frac{\varepsilon}{2\mathrm{We}^2}\int_{\mathscr{D}}\Delta\Pi(\boldsymbol{X}_t(\boldsymbol{x}))$ in the right-hand side. It actually appears that eliminating this term requires to add an entropy term to the energy. To study the longtime behavior, the appropriate viewpoint is to consider the *free energy* rather than the energy.

To introduce the entropy, one needs to consider the probability density functional of the stochastic process $\boldsymbol{X}_t$, and thus the system (148) coupling the momentum equation with the Fokker-Planck equation introduced in Sect. 3.2. Let us assume zero Dirichlet boundary condition on the velocity $\boldsymbol{u}$. The expected stationary state (equilibrium) is

$$\begin{aligned} \boldsymbol{u}(\infty,\boldsymbol{x}) &\quad 0, \\ \psi(\infty,\boldsymbol{x},\boldsymbol{X}) &\quad \psi_{\text{eq}}(\boldsymbol{X}) \quad C\exp(-\Pi(\boldsymbol{X})), \end{aligned}$$

where C is a normalization factor. Using entropy estimates (see C. Ané et al. [4], F. Malrieu [95], A. Arnold, P. Markowich, G. Toscani and A. Unterreiter [5]), exponential convergence to equilibrium may be shown (see B. Jourdain, C. Le Bris, T. Lelièvre and F. Otto [64, 65]). Let us explain this with more details.

The first derivative of the kinetic energy

$$E(t) \quad \frac{\text{Re}}{2}\int_{\mathcal{D}} |\boldsymbol{u}|^2(t,\boldsymbol{x}) \tag{156}$$

writes (as in (151))

$$\frac{dE}{dt} \quad -(1-\varepsilon)\int_{\mathcal{D}} |\nabla\boldsymbol{u}|^2(t,\boldsymbol{x}) - \frac{\varepsilon}{\text{We}}\int_{\mathcal{D}}\int_{\mathbb{R}^d} (\boldsymbol{X}\otimes\nabla\Pi(\boldsymbol{X})) : \nabla\boldsymbol{u}(t,\boldsymbol{x})\psi(t,\boldsymbol{x},\boldsymbol{X}).$$

The *entropy*

$$\begin{aligned} H(t) &\quad \int_{\mathcal{D}}\int_{\mathbb{R}^d} \Pi(\boldsymbol{X})\psi(t,\boldsymbol{x},\boldsymbol{X}) + \int_{\mathcal{D}}\int_{\mathbb{R}^d} \psi(t,\boldsymbol{x},\boldsymbol{X})\ln(\psi(t,\boldsymbol{x},\boldsymbol{X})) - |\mathcal{D}|\ln C, \\ &\quad \int_{\mathcal{D}}\int_{\mathbb{R}^d} \psi(t,\boldsymbol{x},\boldsymbol{X})\ln\left(\frac{\psi(t,\boldsymbol{x},\boldsymbol{X})}{\psi_{\text{eq}}(\boldsymbol{X})}\right) \end{aligned} \tag{157}$$

is next introduced. Notice that $H(t) \geq 0$ (since $x\ln(x) \geq x-1$). Using (153) and $\operatorname{div}\boldsymbol{u}\quad 0$, a simple computation shows:

$$\begin{aligned} \frac{dH}{dt} &\quad -\frac{1}{2\text{We}}\int_{\mathcal{D}}\int_{\mathbb{R}^d} \psi(t,\boldsymbol{x},\boldsymbol{X})\left|\nabla_{\boldsymbol{X}}\ln\left(\frac{\psi(t,\boldsymbol{x},\boldsymbol{X})}{\psi_{\text{eq}}(\boldsymbol{X})}\right)\right|^2 \\ &\quad +\int_{\mathcal{D}}\int_{\mathbb{R}^d} (\boldsymbol{X}\otimes\nabla\Pi(\boldsymbol{X})) : \nabla\boldsymbol{u}(t,\boldsymbol{x})\psi(t,\boldsymbol{x},\boldsymbol{X}). \end{aligned}$$

Thus, the free energy $F(t) \quad E(t) + \dfrac{\varepsilon}{\text{We}}H(t)$ (a non-negative quantity) satisfies:

$$\frac{dF}{dt} \quad -(1-\varepsilon)\int_{\mathcal{D}} |\nabla\boldsymbol{u}|^2(t,\boldsymbol{x}) - \frac{\varepsilon}{2\text{We}^2}\int_{\mathcal{D}}\int_{\mathbb{R}^d} \psi(t,\boldsymbol{x},\boldsymbol{X})\left|\nabla\boldsymbol{X}\ln\left(\frac{\psi(t,\boldsymbol{x},\boldsymbol{X})}{\psi_{\text{eq}}(\boldsymbol{X})}\right)\right|^2. \tag{158}$$

Comparing with (154), we observe that the introduction of the entropy allows to eliminate the right-hand side. In particular, (158) shows that the only stationary state is $\boldsymbol{u}\quad 0$ et $\psi\quad\psi_{\text{eq}}$. Moreover, using a Poincaré inequality: for all$\boldsymbol{u} \in H^1_0(\mathcal{D})$,

$$\int |\boldsymbol{u}|^2 \leq C\int |\nabla\boldsymbol{u}|^2$$

and the Logarithmic Sobolev inequality: for all probability density functional ψ,

$$\int \psi \ln\left(\frac{\psi}{\psi_{\mathrm{eq}}}\right) \leq C \int \psi \left|\nabla \ln\left(\frac{\psi}{\psi_{\mathrm{eq}}}\right)\right|^2, \tag{159}$$

exponential convergence to zero for F (and thus for $\boldsymbol{u}$ in $L^2_{\boldsymbol{x}}$-norm) is obtained from (158). The Logarithmic Sobolev inequality (159) holds for $\psi_{\mathrm{eq}}(\boldsymbol{X}) \quad C\exp(-\Pi(\boldsymbol{X}))$ if Π is α-convex for example (which is the case for Hookean and FENE dumbbells). The Csiszar-Kullback inequality (see C. Ané et al. [4]) then shows that ψ converges to ψ_{eq} exponentially fast in $L^2_{\boldsymbol{x}}(L^1_{\boldsymbol{X}})$-norm.

For generalizations of these computations to non-homogeneous boundary conditions on $\boldsymbol{u}$ (and thus $\boldsymbol{u}(\infty,\boldsymbol{x}) \,/\, 0$), we refer to B. Jourdain, C. Le Bris, T. Lelièvre and F. Otto [65].

We would like also to mention that these estimates on the micro-macro system can be used as a guideline to derive new estimates on related macro-macro models (see D. Hu and T. Lelièvre [59]).

Remark 9 (On the choice of the entropy). If one considers the Fokker-Planck equation with $\boldsymbol{u} \quad 0$, it is well-known (see A. Arnold, P. Markowich, G. Toscani and A. Unterreiter [5]) that exponential convergence to equilibrium can be obtained using more general entropy functions of the form

$$H(t) \quad \int_{\mathscr{D}} \int_{\mathbb{R}^d} h\left(\frac{\psi}{\psi_{\mathrm{eq}}}\right) \psi_{\mathrm{eq}}$$

where $h : \mathbb{R} \longrightarrow \mathbb{R}^*_+$ is a convex $\mathscr{C}^2$ function, such that $h(1) \quad 0$. However, it seems that to derive the entropy estimate (158) on the coupled system (150), it is necessary to choose the "physical entropy" corresponding to the choice $h(x) \quad x\ln(x) - (x-1)$.

Remark 10 (On the assumptions on the force $\boldsymbol{F}$*).* Recall that we assumed that $\boldsymbol{F}$ $\nabla\Pi$, where Π is a radial convex function. Let us briefly discuss the assumptions on $\boldsymbol{F}$ we used so far.

- The fact that $\boldsymbol{F}$ can be written as the gradient of a potential Π is important to obtain a simple analytical expression for ψ_{eq}.
- The fact that Π is radial is a very important assumption to ensure the symmetry of the stress tensor.
- The convexity assumption on Π is important in the analysis of the SDEs (in particular for uniqueness of strong solutions).
- The α-convexity of the potential Π has been used to obtain the Logarithmic Sobolev inequality (159).

Existence results on the coupled problem with the Fokker-Planck PDE

Many authors have obtained existence and uniqueness results for the micro-macro problem (148), that is the coupled model involving the Fokker-Planck equation.

For local existence and uniqueness results, we refer to M. Renardy [111], T. Li, H. Zhang and P.W. Zhang [83] (polynomial forces) and to H. Zhang and P.W. Zhang [123] (FENE force with $b > 76$). In a recent work by N. Masmoudi [97], a local in time existence result is obtained for the FENE model without any assumption on b (using reflecting boundary conditions). The author also shows global in time existence result for initial data close to equilibrium (see also F.-H. Lin, C. Liu and P.W. Zhang [87] for a similar result under the assumption $b > 12$).

Global existence results have also been obtained for closely related problems:

- *Existence results for a regularized version*: In J.W. Barrett, C. Schwab and E. Süli [7, 8], a global existence result is obtained for (148) (Hookean and FENE force) using a regularization of some terms, which allows for more regular solutions. More precisely, the velocity $\boldsymbol{u}$ in the Fokker Planck equation is replaced by a smoothed velocity, and the same smoothing operator is used on the stress tensor $\boldsymbol{\tau}_p$ in the right-hand side of the momentum equations. See also L. Zhang, H. Zhang and P.W. Zhang [124].
- *Existence results with a corotational derivative*: In J.W. Barrett, C. Schwab and E. Süli [7, 8] (again with some regularizations) and P.-L. Lions and N. Masmoudi [90, 97] (without any regularizations), the authors obtain global-in-time existence results replacing $\nabla\boldsymbol{u}$ in the Fokker-Planck equation by $\frac{\nabla\boldsymbol{u}-\nabla\boldsymbol{u}^T}{2}$ (which is similar to considering the corotational derivative of $\boldsymbol{\tau}_p$ instead of the upper convected derivative in differential macro-macro models). More precisely, in [90], a global-in-time existence result of weak solutions is obtained in dimension 2 and 3, while in [97], it is proved that in dimension 2, there exists a unique global-in-time strong solution. A related recent result by F.-H. Lin, P. Zhang and Z. Zhang is [88].

We would like also to mention the related works [27, 28, 31] (existence results for coupled Navier-Stokes Fokker-Planck micro-macro models) by P. Constantin, C. Fefferman, N. Masmoudi and E.S. Titi, and also the work of C. Le Bris and P.-L. Lions [78, 79] about existence and uniqueness of solutions to Fokker-Planck type equations with irregular coefficients.

Numerical methods

In this section, we review the literature for the numerical analysis of methods to discretize (98). For the discretization of the micro-macro problem in the Fokker-Planck version, we refer to Sect. 4.4.

The idea of coupling a Finite Element Method for discretization in space and a stochastic method (Monte Carlo to approximate the expectation and Euler scheme on the SDE) has been first proposed by M. Laso and H.C. Öttinger [75]. Such a method is called *Calculation Of Non-Newtonian Flow: Finite Elements and Stochastic SImulation Technique* (CONNFFESSIT). At first, Lagrangian methods were used on the SDE, and *independent Brownian motions on each trajectories* (see M. Laso and H.C. Öttinger [76]). The algorithm then consists in: (i) computing $(\boldsymbol{u}, p)$, (ii) computing the trajectories of the fluid particles carrying the dumbbells (characteristic method), (iii) integrating the SDEs along these trajectories and (iv) computing the stress tensor $\boldsymbol{\tau}_p$ by local empirical means in each finite element. This Lagrangian approach is the most natural one since it is naturally obtained from the derivation of the model (see Sect. 4.2). However, owing to the term $\operatorname{div} \boldsymbol{\tau}_p$, numerical results are very noisy in space when using independent Brownian motions on each trajectory. Moreover, such an approach requires to maintain a sufficiently large number of dumbbells per cell of the mesh, which is not easy to satisfy (there is a need to add some dumbbells and to destroy others during the simulation).

The idea then came up to use the Eulerian version of the SDE, and introducing fields of end-to-end vectors: $\boldsymbol{X}_t(\boldsymbol{x})$. This is the concept of *Brownian Configuration Field* introduced by M.A. Hulsen, A.P.G. van Heel and B.H.A.A. van den Brule in [61]. In this Eulerian description, the most natural and simple choice is to use *the same Brownian motion* at each position in space. This reduces the noise in space and the variance of the velocity (but not the variance of the stress, see below and the work [63] by B. Jourdain, C. Le Bris and T. Lelièvre). The discretization of the transport term can then be done using a *Discontinuous Galerkin* method (see M.A. Hulsen, A.P.G. van Heel and B.H.A.A. van den Brule [61]), the characteristic methods (see J.C. Bonvin [18] or the *Backward-Tracking Lagrangian Particle Method* of P. Wapperom, R. Keunings and V. Legat [121]), or classical finite element methods with stabilization.

Let us recall how the CONNFFESSIT method writes in a shear flow (see Sect. 5.4). In this special case, both the Lagrangian and the Eulerian approaches lead to the same discretization: for given u_h^n, $X_{h,n}^k$ and $Y_{h,n}^k$, compute $u_h^{n+1} \in V_h$ such that for all $v \in V_h$,

$$\left\{\begin{array}{ll} \dfrac{\mathrm{Re}}{\delta t}\displaystyle\int_y (u_h^{n+1} - u_h^n)v & -(1-\varepsilon)\displaystyle\int_y \frac{\partial u_h^{n+1}}{\partial y}\frac{\partial v}{\partial y} - \int_y \tau_{h,n}\frac{\partial v}{\partial y} + \int_y f v, \\ \tau_{h,n} & \dfrac{\varepsilon}{\mathrm{We}}\dfrac{1}{K}\displaystyle\sum_{k\quad 1}^{K} X_{h,n}^k F_Y(X_{h,n}^k, Y_{h,n}^k), \\ X_{h,n+1}^k - X_{h,n}^k & \left(\dfrac{\partial u_h^{n+1}}{\partial y} Y_{h,n}^k - \dfrac{1}{2\mathrm{We}} F_X(X_{h,n}^k, Y_{h,n}^k)\right)\delta t \\ & \quad + \dfrac{1}{\sqrt{\mathrm{We}}}\left(V_{h,t_{n+1}}^k - V_{h,t_n}^k\right), \\ Y_{h,n+1}^k - Y_{h,n}^k & -\dfrac{1}{2\mathrm{We}} F_Y(X_{h,n}^k, Y_{h,n}^k)\delta t + \dfrac{1}{\sqrt{\mathrm{We}}}\left(W_{h,t_{n+1}}^k - W_{h,t_n}^k\right). \end{array}\right. \tag{160}$$

The index n is the timestep and the index k is the realization number in the SDE ($1 \le k \le K$ where K is the number of dumbbells in each cell). Finally, V_h is a finite element space. We suppose in the following that $V_h \quad$ P1 is the finite element space of continuous piecewise linear functions so that $X_{h,n}$, $Y_{h,n}$ and $\tau_{h,n}$ are piecewise constant functions in space (they belong to the functional space P0). We refer to Fig. 13.

Convergence of the CONNFFESSIT method

In the CONNFFESSIT method, three numerical parameters are to be chosen: the timestep δt, the spacestep h and the number of dumbbells (or realizations) K. It is expected that the method converges in the limit $\delta t \to 0$, $h \to 0$ and $K \to \infty$.

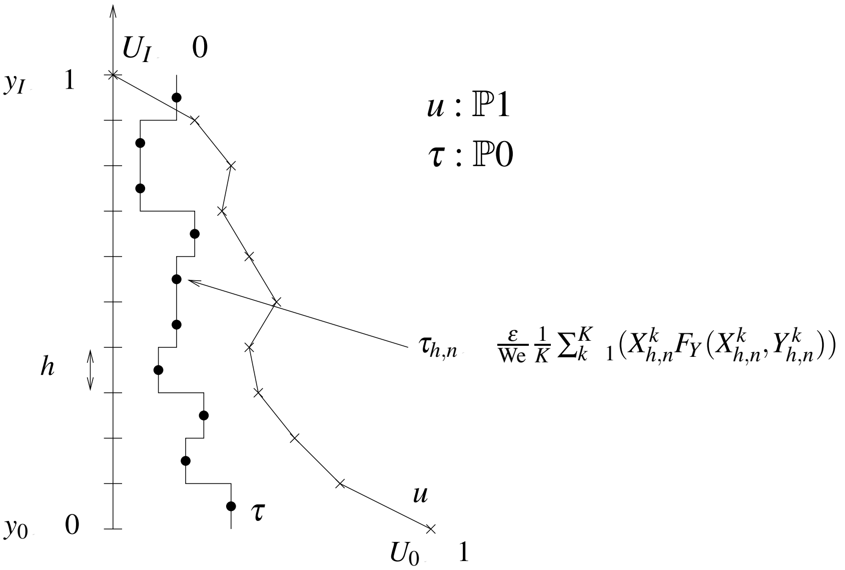

U_I 0
y_I 1
$u : \mathbb{P}1$
$\tau : \mathbb{P}0$
$\tau_{h,n}$ $\frac{\varepsilon}{\mathrm{We}}\frac{1}{K}\sum_{k\quad 1}^{K}(X_{h,n}^k F_Y(X_{h,n}^k, Y_{h,n}^k))$
h
u
τ
y_0 0
U_0 1

Fig. 13. The CONNFFESSIT method in a shear flow.

This has been proven in B. Jourdain, C. Le Bris and T. Lelièvre [67] and W. E, T. Li and P.W. Zhang [38] for Hookean dumbbells in a shear flow.

Theorem 4. *Assuming $u_0 \in H_y^2$, $f_{ext} \in L_t^1(H_y^1)$, $\frac{\partial f_{ext}}{\partial t} \in L_t^1(L_y^2)$ and $\delta t < \frac{1}{2}$, we have (for $V_h \quad \mathbb{P}1$): $\forall n < \frac{T}{\delta t}$,*

$$\left|\left|u(t_n) - \overline{u}_h^n\right|\right|_{L_y^2(L_\omega^2)} + \left|\left|\mathbb{E}(X_{t_n}Y_{t_n}) - \frac{1}{K}\sum_{k\quad 1}^{K}\overline{X}_{h,n}^k\overline{Y}_n^k\right|\right|_{L_y^1(L_\omega^1)} \leq C\left(h + \delta t + \frac{1}{\sqrt{K}}\right).$$

Remark 11. It can be shown that the convergence in space is optimal (see T. Lelièvre [81]):

$$\left|\left|u(t_n) - \overline{u}_h^n\right|\right|_{L_y^2(L_\omega^2)} \leq C\left(h^2 + \delta t + \frac{1}{\sqrt{K}}\right).$$

The main difficulties in the proof of Theorem 4 originate from the following facts:

- The velocity u_h^n is a *random variable*. The energy estimate at the continuous level cannot be directly translated into an energy estimate at the discrete level (which would yield the stability of the scheme).
- The end-to-end vectors $(\overline{X}_{h,n}^k, \overline{Y}_n^k)_{1\leq k\leq K}$ are *coupled* random variables (even though the driving Brownian motions $(V_{h,t}^k, W_{h,t}^k)_{1\leq k\leq K}$ are independent).

- The stability of the numerical scheme requires an almost sure bound on the Y_n^k :

$$\delta t \frac{1}{K} \sum_{k \quad 1}^{K} (Y_n^k)^2 < 1.$$

To prove convergence, a cut-off procedure on Y_n^k is employed:

$$\overline{Y}_{n+1}^k \quad \max(-A, \min(A, Y_{n+1}^k)) \tag{161}$$

with $0 < A < \sqrt{\frac{3}{5\delta t}}$. In Theorem 4, $\overline{u}_h^n$, $\overline{X}_n^k$ $\overline{Y}_n^k$ denotes random variables obtained by the CONNFFESSIT scheme (160) with the cutoff procedure (161). It can be checked that for sufficiently small δt or sufficiently large K, this cut-off procedure is not used.

For a result without cut-off, we refer to B. Jourdain, C. Le Bris and T. Lelièvre [67]. For an extension of these results to a more general geometry and discretization by a finite difference scheme, we refer to T. Li and P.W. Zhang [84]. For a convergence result in space and time, we refer to A. Bonito, Ph. Clément and M. Picasso [14].

Variance of the results and dependency of the Brownian motions in space

One important practical quantity when using Monte Carlo methods is the variance of the result. If the variance is too large, the numerical method is basically useless. We already mentioned above (see Sect. 5.5) variance reduction methods. It is also interesting to investigate how the variance of the results depends upon the numerical parameters. In the framework of the CONNFFESSIT method, this variance is particularly sensitive to the dependency of the Brownian motion on the space variable.

One can check (at least for regular solutions) that the dependency of the Brownian motion on the space variable does not influence the macroscopic quantities $(\boldsymbol{u}, p, \boldsymbol{\tau}_p)$ at the continuous level. This can be rigorously proved for Hookean *dumbbells* in a shear flow. It can also be checked that the convergence result of Theorem 4 is insensitive to the dependency of the Brownian motion on the space variable. However, at the discrete level, this dependency strongly influences the variance of the results. It is observed that using Brownian motions independent from one cell of the mesh to another rather than Brownian motions not depending on space increases the variance of the velocity, but reduces the variance on the stress (see P. Halin, G. Lielens, R. Keunings, and V. Legat [57], J.C. Bonvin and M. Picasso [16] and B. Jourdain, C. Le Bris and T. Lelièvre [63]).

This can be precisely analyzed for the case of Hookean dumbbells in a shear flow. It can be shown that (see B. Jourdain, C. Le Bris and T. Lelièvre [63]):

a) The variance on the velocity is minimal for a Brownian motion not depending on space.
b) Using Brownian motions independent from one cell to another is not the best method to reduce the variance on τ.
c) It is possible to reduce the variance on τ with the same computational cost as when using a Brownian motion not depending on space. It consists in using a Brownian motion alternatively multiplied by $+1$ or -1 on nearest-neighbour cells.

7 Other types of complex fluids

7.1 Liquid crystals

So far, we have only considered dilute solutions of flexible polymers. Some other polymers behave more like *rigid rods*. This introduces anisotropy in the system. Solutions of such rigid polymers are called polymeric *liquid crystals*. One of the major aspect to account for in the modelling of solutions of rod-like polymers is that the interaction of the polymers becomes important at much a lower concentration than with flexible polymers.

Modelling of liquid crystals, along with mathematical and numerical studies, is today a very lively and active field of research. The present short section does not reflect the variety of scientific enterprises dealing with liquid crystals. It is just a brief incursion in this world to see, once, the basic models. One adequate model is the *Doi model* (see M. Doi and S.F. Edwards [36] and H.C. Öttinger [102]). It describes the evolution for a configuration vector $\boldsymbol{R}_t$ by a stochastic differential equation:

$$\begin{aligned} & d\boldsymbol{R}_t + \boldsymbol{u}\cdot\nabla\boldsymbol{R}_t\,dt \\ & \quad \left(\mathrm{Id} - \frac{\boldsymbol{R}_t\otimes\boldsymbol{R}_t}{\|\boldsymbol{R}_t\|^2}\right)\left(\left(\nabla\boldsymbol{u}\boldsymbol{R}_t - \frac{1}{2}B^2\nabla V(\boldsymbol{R}_t)\right)dt + B\,d\boldsymbol{W}_t\right) \\ & \quad -\frac{d-1}{2}B^2\frac{\boldsymbol{R}_t}{\|\boldsymbol{R}_t\|^2}\,dt, \end{aligned} \tag{162}$$

where B is a positive constant and $d\quad 2$ or 3 is the dimension of the ambient space. Notice that B may also be a function $B(\boldsymbol{R}_t)$ in some models (with then an additional term involving $\nabla(B^2)$ in the drift term). Notice also that we assume that all the initial conditions $\boldsymbol{R}_0(\boldsymbol{x})$ have a fixed length L so that $\forall(t,\boldsymbol{x})$, $\|\boldsymbol{R}_t(\boldsymbol{x})\|\quad\|\boldsymbol{R}_0(\boldsymbol{x})\|\quad L.$ The potential V accounts for the mean-field interaction between the polymers. For example, the Maier-Saupe potential is:

$$V(\boldsymbol{R})\quad -\frac{1}{L^4}\mathbb{E}(\boldsymbol{R}_t\otimes\boldsymbol{R}_t):\boldsymbol{R}\otimes\boldsymbol{R}. \tag{163}$$

The stress tensor is then given by:

$$\boldsymbol{\tau}_p(t)\quad \mathbb{E}(\boldsymbol{u}_t\otimes\boldsymbol{u}_t) + \mathbb{E}\Big(\boldsymbol{u}_t\otimes\big((\mathrm{Id}-\boldsymbol{u}_t\otimes\boldsymbol{u}_t)\nabla V(\boldsymbol{u}_t)\big)\Big) - \mathrm{Id} \tag{164}$$

where $\boldsymbol{u}_t\quad \dfrac{\boldsymbol{R}_t}{L}$ is the rod orientation. We have neglected the viscous contribution in (164). The fully coupled system then consists in the first two equations of (49) with (162)–(164). Notice that the main differences with the equations seen so far in this article are the nonlinearity in the sense of MacKean due to the presence of the expectation value in the potential V and the fact that the diffusion term depends on the process $\boldsymbol{R}_t$.

For an analysis of the coupled system with the Fokker-Planck version of (162)–(164) in the special case of shear flow, we refer to H. Zhang and P.W. Zhang [122].

The longtime behavior of the Fokker-Planck equation has been studied by P. Constantin, I. Kevrekidis and E.S. Titi in [30] (see also [29]). A thorough analysis of the variety of possible steady states and their stability is studied by G. Forest, Q. Wang and R. Zhou in [45]. Some numerical methods to solve the stochastic differential equation (162) are proposed by H.C. Öttinger in [102]. On the other hand, we are not aware of any rigorous numerical analysis of numerical methods to solve this system without closure approximation.

7.2 Suspensions

We now slightly change the context. Multiscale modelling of complex fluids is very advanced for polymer flows. It is a well established scientific activity. However, it is also a growing activity for some other types of fluids, far from polymer flows. We give here the illustrative example of civil engineering fluids, with muds and clays. It is not forbidden to believe that other materials of civil engineering, like cement, will benefit a lot from multiscale modelling approaches in a near future.

For concentrated suspensions (such as muds or clays), one model available in the literature is the *Hebraud-Lequeux model* [62]. This model describes the rheology of the fluid in terms of a Fokker-Planck equation ruling the evolution in time of the probability of finding, at each point, the fluid in a given state of stress. To date, although current research is directed toward constructing multidimensional variants, the model is restricted to the one-dimensional setting, that is, the Couette flow. The stress at the point y and at time t is thus determined by one scalar variable σ:

$$\left\{\begin{array}{ll} \frac{\partial p}{\partial t}(t,y,\sigma) & -\frac{\partial u}{\partial y}(t,y)\frac{\partial p}{\partial \sigma}(t,y,\sigma)+D(p)\frac{\partial^2 p}{\partial \sigma^2}(t,y,\sigma) \\ & \quad -H(|\sigma|-1)p(t,y,\sigma)+D(p)\delta_0, \\ D(p) & \int_{|\sigma|\geq 1} p(t,y,\sigma)\,d\sigma. \end{array}\right. \tag{165}$$

In the above system, where we have on purpose omitted all physical constants, the function H denotes the Heaviside function. It aims at modelling the presence of a threshold constraint (here set to one): when the constraint is above the threshold, the stress relaxes to zero, which translates into the two last terms of the Fokker-Planck equation. The diffusion in the stress space is also influenced nonlinearly by the complete state of stress, as indicated by the definition of $D(p)$. On the other hand, the function $\frac{\partial u}{\partial y}(t,y)$ accounts for a shear rate term, here provided by the macroscopic flow. The contribution to the stress at the point y under consideration is then given by the average

$$\tau(t,y) \quad \int_{\mathbb{R}} \sigma\, p(t,y,\sigma)\,d\sigma. \tag{166}$$

The fully coupled system consisting of the Fokker-Planck equation (165), the expression (166) of the stress tensor, and the macroscopic equation for the Couette flow (first line of (104)) has been studied mathematically in a series of work by E. Cancès, I. Catto, Y. Gati and C. Le Bris [21, 22, 23].

Alternately to a direct attack of the Fokker-Planck equation (165), one might wish to simulate the associated stochastic differential equation *with jumps* that reads

$$d\sigma_t \quad \frac{\partial u}{\partial y}dt + \sqrt{2\mathbb{P}(|\sigma_t| \geq 1)} \quad dW_t - 1_{\{|\sigma_{t^-}|\geq 1\}}\sigma_{t^-}\, dN_t, \tag{167}$$

where W_t is a Brownian motion and N_t is an independent Poisson process with unit intensity. Note that, in addition to the jumps, equation (167) is nonlinear in the sense of MacKean, as the diffusion coefficient depends on the marginal law of the solution at time t.

The coupled system to simulate then reads

$$\begin{cases} \frac{\partial u}{\partial t}(t,u) - \frac{\partial^2 u}{\partial y^2}(t,y) \quad \frac{\partial \tau}{\partial y}(t,y) \\ \forall y, \begin{cases} \tau(t,y) \quad \mathbb{E}(\sigma_t(y)) \\ d\sigma_t(y) \quad \frac{\partial u}{\partial y}dt + \sqrt{2\mathbb{P}(|\sigma_t(y)| \geq 1)} \quad dW_t - 1_{\{|\sigma_{t^-}(y)|\geq 1\}}\sigma_{t^-}(y)\, dN_t, \end{cases} \end{cases} \tag{168}$$

where one should note that the stochastic differential equation has jumps.

Numerical simulations of this system have been carried out successfully (see Y. Gati [49]). For the numerical analysis of the particle approximation, we refer to M. Ben Alaya and B. Jourdain [9].

7.3 Blood flows

Blood is a complex fluid consisting of a suspension of cells in plasma. These cells are mainly *red blood cells* or erythrocytes, white blood cells or leucocytes, and platelets. Red blood cells constitute 98% of the cells in suspension. These microstructures are mostly responsible for the non-Newtonian behavior of blood. A red blood cell is a biconcave disk of diameter $8.5\mu m$ and thickness $2.5\mu m$. It consists of a highly flexible membrane which is filled with a solution (haemoglobin). The ambient flow modifies the shape of the membrane. This phenomenon allows storage and release of energy in the microstructures, like for polymeric fluids. At low shear rates, red blood cells agglomerate into long structures called *rouleaux*.

It is observed that at high shear rates (like for pulsatile flow in healthy arteries, see for example J.F. Gerbeau, M. Vidrascu and P. Frey [50] or A. Quarteroni and L. Formaggia [107]), blood behaves essentially as a Newtonian fluid. At low shear rates (in arterioles, venules, recirculatory regions in aneurysms and stenoses), blood is a non-Newtonian fluid: it exhibits shear-thining, viscoelastic and thixotropic effects. This can be interpreted as follows: in flows with high shear rates, red blood cells cannot agglomerate, and the rheology is not influenced by the microstructures, while in flows with low shear rates, red blood cells agglomerate and this influences the rheology. Notice that we here discuss simple mechanical properties, neglecting important biochemical factors (like in clot formation for example).

In [41, 103], R.G. Owens and J. Fang propose a micro-macro model for blood, which is very similar to the model presented in Sect. 4. This model applies in some sufficiently large flow domains, so that statistics on the configurations of red blood cells at each macroscopic point make sense. In other context, it may be important to consider each red blood cell as a separated entity like in the work [80] by A. Lefebvre and B. Maury.

Let us first suppose that the velocity field is given and homogeneous. The microscopic variables used to describe the microstructure (namely the red blood cells) are a vector $\boldsymbol{X}$ (similar to the end-to-end vector for polymeric fluids) and an integer $k \geq 1$ which measures the size of the aggregate the red blood cell belongs to. Consider then the non-negative function $\psi_k(t,\boldsymbol{X})$ such that $\psi_k(t,\boldsymbol{X})d\boldsymbol{X}$ is the number of red blood cells (per unit volume of fluid) belonging to an aggregate of size k having end-to-end vector between $\boldsymbol{X}$ and $\boldsymbol{X}+d\boldsymbol{X}$. We denote by $N_j \quad \frac{1}{j}\int \psi_j(t,\boldsymbol{X})d\boldsymbol{X}$ the number of aggregates of k red blood cells per unit volume.

The following Fokker-Planck equation rules the evolution of $(\psi_k(t,\boldsymbol{X}))_{k\geq 1}$:

$$\begin{aligned}\frac{\partial \psi_k}{\partial t} \quad & -\mathrm{div}_{\boldsymbol{X}}\left(\left(\nabla \boldsymbol{u}\boldsymbol{X} - \frac{2}{\zeta_k}\boldsymbol{F}(\boldsymbol{X})\right)\psi_k\right) + \frac{2kT}{\zeta_k}\Delta_{\boldsymbol{X}}\psi_k \\ & + h_k(\dot\gamma)\psi_k^{eq} - g_k(\dot\gamma)\psi_k. \end{aligned} \tag{169}$$

In Equation (169),

$$h_k(\dot\gamma) \quad \frac{a(\dot\gamma)}{2N_k^{eq}}\sum_{i\quad 1}^{k-1} N_i N_{k-i} + \frac{b(\dot\gamma)}{N_k^{eq}}\sum_{j\quad 1}^{\infty} N_{k+j}$$

is an aggregation rate coefficient and

$$g_k(\dot\gamma) \quad \frac{b(\dot\gamma)}{2}(k-1) + a(\dot\gamma)\sum_{j\quad 1}^{\infty} N_j$$

is a fragmentation rate coefficient. Both depend on the shear rate $\dot\gamma \quad \sqrt{\frac{1}{2}\dot{\boldsymbol{\gamma}}:\dot{\boldsymbol{\gamma}}}$ with $\dot{\boldsymbol{\gamma}} \quad \nabla\boldsymbol{u}+\nabla\boldsymbol{u}^T$. At equilibrium (namely for zero shear rate: $\dot\gamma \quad 0$), the number of aggregates of k red blood cells per unit volume is N_k^{eq}. An analytical expression for N_k^{eq} can be derived, in terms of $a(0)$, $b(0)$ and the total number of red blood cells per unit volume N_0 (which is a conserved quantity). The function $\psi_k^{eq} \quad Z^{-1}\exp(-\Pi)kN_k^{eq}$ describes the statistics of the red blood cells at equilibrium (Π is the potential of the force $\boldsymbol{F}$). Notice that by integrating (169) with respect to $\boldsymbol{X}$ (and dividing by k), the following Smoluchowski equation on $(N_k(t))_{k\geq 1}$ is obtained:

$$\frac{dN_k}{dt} \quad h_k(\dot\gamma)N_k^{eq} - g_k(\dot\gamma)N_k.$$

The parameters of the model are N_0, the friction coefficient ζ_k (which is typically chosen as $\zeta_k \quad k\zeta_1$) and the functions a and b which can be calibrated using experiments (see R.G. Owens and J. Fang [103, 41]).

In complex flows (for which $\nabla\boldsymbol{u}$ depends on the space variable $\boldsymbol{x}$), the functions ψ_k also depend on $\boldsymbol{x}$ and the derivative $\frac{\partial}{\partial t}$ in (169) is replaced by a convective derivative $\frac{\partial}{\partial t} + \boldsymbol{u}\cdot\nabla$. The micro model is coupled to the momentum equations through the Kramers expression for the extra stress tensor:

$$\boldsymbol{\tau} \quad \sum_{k\ \ 1}^{\infty} \boldsymbol{\tau}_k,$$

$$\boldsymbol{\tau}_k(t,\boldsymbol{x}) \quad \int \boldsymbol{F}(\boldsymbol{X}) \otimes \boldsymbol{X} \psi_k(t,\boldsymbol{x},\boldsymbol{X}) d\boldsymbol{X} - kN_k(t,\boldsymbol{x}) k_B T \mathrm{Id}.$$

Let us mention one modelling challenge: it is observed that the distribution of red blood cells is not uniform across a vessel (cell-depleted region near the vessel walls), and it is not clear how to account for this phenomenon in the micro-macro model. In the case of a Hookean force, it is possible to derive a macro-macro version of this model, which can then be further simplified (see R.G. Owens and J. Fang [41, 103]). Only this macro-macro version has been used so far in simulations for comparisons with experimental data (see again R.G. Owens and J. Fang [41, 103]).

References

1. M.P. Allen and D.J. Tildesley. *Computer simulation of liquids*. Oxford Science Publications, 1987.
2. A. Ammar, B. Mokdad, F. Chinesta, and R. Keunings. A new family of solvers for some classes of multidimensional partial differential equations encountered in kinetic theory modeling of complex fluids. *J. Non-Newtonian Fluid Mech.*, 139:153–176, 2006.
3. A. Ammar, B. Mokdad, F. Chinesta, and R. Keunings. A new family of solvers for some classes of multidimensional partial differential equations encountered in kinetic theory modeling of complex, part II: Transient simulation using space-time separated representations. *J. Non-Newtonian Fluid Mech.*, 144:98–121, 2007.
4. C. Ané, S. Blachère, D. Chafaï, P. Fougères, I. Gentil, F. Malrieu, C. Roberto, and G. Scheffer. *Sur les inégalités de Sobolev logarithmiques*. Société Mathématique de France, 2000. In French.
5. A. Arnold, P. Markowich, G. Toscani, and A. Unterreiter. On convex Sobolev inequalities and the rate of convergence to equilibrium for Fokker-Planck type equations. *Comm. Part. Diff. Eq.*, 26:43–100, 2001.
6. F.P.T. Baaijens. Mixed finite element methods for viscoelastic flow analysis: a review. *J. Non-Newtonian Fluid Mech.*, 79:361–385, 1998.
7. J.W. Barrett, C. Schwab, and E. Süli. Existence of global weak solutions for some polymeric flow models. *Math. Models and Methods in Applied Sciences*, 15(6):939–983, 2005.
8. J.W. Barrett and E. Süli. Existence of global weak solutions to kinetic models for dilute polymers. *Multiscale Model. Simul.*, 6(2):506–546, 2007.
9. M. Ben Alaya and B. Jourdain. Probabilistic approximation of a nonlinear parabolic equation occuring in rheology. *Journal of Applied Probability*, 44(2):528–546, 2007.
10. D. Bernardin. Introduction la rhéologie des fluides : approche macroscopique, 2003. Ecole de printemps, G.D.R. Matériaux vitreux, disponible `http://www.lmcp.jussieu.fr/lmcp/GDR-verres/html/Rheologi\_1.pdf`. In French.
11. R.B. Bird, R.C. Armstrong, and O. Hassager. *Dynamics of polymeric liquids*, volume 1. Wiley Interscience, 1987.
12. R.B. Bird, C.F. Curtiss, R.C. Armstrong, and O. Hassager. *Dynamics of polymeric liquids*, volume 2. Wiley Interscience, 1987.

13. R.B. Bird, P.J. Dotson, and N.L. Johnson. Polymer solution rheology based on a finitely extensible bead-spring chain model. *J. Non-Newtonian Fluid Mech.*, 7:213–235, 1980. Errata: *J. Non-Newtonian Fluid Mech.*, 8:193 (1981).
14. A. Bonito, Ph. Clément, and M. Picasso. Finite element analysis of a simplified stochastic Hookean dumbbells model arising from viscoelastic flows. *M2AN Math. Model. Numer. Anal.*, 40(4):785–814, 2006.
15. A. Bonito, Ph. Clément, and M. Picasso. Mathematical analysis of a stochastic simplified Hookean dumbbells model arising from viscoelastic flow. *J. Evol. Equ.*, 6(3):381–398, 2006.
16. J. Bonvin and M. Picasso. Variance reduction methods for CONNFFESSIT-like simulations. *J. Non-Newtonian Fluid Mech.*, 84:191–215, 1999.
17. J. Bonvin, M. Picasso, and R. Sternberg. GLS and EVSS methods for a three fields Stokes problem arising from viscoelastic flows. *Comp. Meth. Appl. Mech. Eng.*, 190:3893–3914, 2001.
18. J.C. Bonvin. *Numerical simulation of viscoelastic fluids with mesoscopic models*. PhD thesis, Ecole Polytechnique Fédérale de Lausanne, 2000. Available at `http://library.epfl.ch/theses/?nr=2249`.
19. M. Braack and A. Ern. A posteriori control of modeling errors and discretization errors. *Multiscale Model. Simul.*, 1(2):221–238, 2003.
20. H.-J. Bungartz and M. Griebel. Sparse grids. *Acta Numer.*, 13:147–269, 2004.
21. E. Cancès, I. Catto, and Y. Gati. Mathematical analysis of a nonlinear parabolic equation arising in the modelling of non-Newtonian flows. *SIAM J. Math. Anal.*, 37:60–82, 2005.
22. E. Cancès, I. Catto, Y. Gati, and C. Le Bris. A micro-macro model describing Couette flows of concentrated suspensions. *Multiscale Model. Simul.*, 4:1041–1058, 2005.
23. E. Cancès and C. Le Bris. Convergence to equilibrium of a multiscale model for suspensions. *Discrete and Continuous Dynamical Systems - Series B*, 6:449–470, 2006.
24. C. Chauvière. A new method for micro-macro simulations of viscoelastic flows. *SIAM J. Sci. Comput.*, 23(6):2123–2140, 2002.
25. C. Chauvière and A. Lozinski. Simulation of dilute polymer solutions using a Fokker-Planck equation. *Computers and fluids*, 33(5-6):687–696, 2004.
26. F. Comets and T. Meyre. *Calcul stochastique et modèles de diffusion*. Dunod, 2006. In French.
27. P. Constantin. Nonlinear Fokker-Planck Navier-Stokes systems. *Commun. Math. Sci.*, 3(4):531–544, 2005.
28. P. Constantin, C. Fefferman, A. Titi, and A. Zarnescu. Regularity of coupled two-dimensional nonlinear Fokker-Planck and Navier-Stokes systems. *Commun. Math. Phys.*, 270(3):789–811, 2007.
29. P. Constantin, I. Kevrekidis, and E.S. Titi. Asymptotic states of a Smoluchowski equation. *Archive Rational Mech. Analysis*, 174(3):365–384, 2004.
30. P. Constantin, I. Kevrekidis, and E.S. Titi. Remarks on a Smoluchowski equation. *Disc. and Cont. Dyn. Syst.*, 11(1):101–112, 2004.
31. P. Constantin and N. Masmoudi. Global well-posedness for a Smoluchowski equation coupled with Navier-Stokes equations in 2D. *Commun. Math. Phys.*, 278, 179–191, 2008.
32. P. Degond, M. Lemou, and M. Picasso. Viscoelastic fluid models derived from kinetic equations for polymers. *SIAM J. Appl. Math.*, 62(5):1501–1519, 2002.
33. P. Delaunay, A. Lozinski, and R.G. Owens. Sparse tensor-product Fokker-Planck-based methods for nonlinear bead-spring chain models of dilute polymer solutions. *CRM Proceedings and Lecture Notes, Volume 41*, 2007.

34. F. Devreux. *Matière et désordre : polymères, gels, verres*. Cours de l'Ecole Polytechnique, 2000. In French.
35. M. Doi. *Introduction to Polymer Physics*. International Series of Monographs on Physics. Oxford University Press, 1996.
36. M. Doi and S.F. Edwards. *The Theory of Polymer Dynamics*. International Series of Monographs on Physics. Oxford University Press, 1988.
37. Q. Du, C. Liu, and P. Yu. FENE dumbbell model and its several linear and nonlinear closure approximations. *Multiscale Model. Simul.*, 4(3):709–731, 2005.
38. W. E, T. Li, and P.W. Zhang. Convergence of a stochastic method for the modeling of polymeric fluids. *Acta Mathematicae Applicatae Sinica, English Series*, 18(4):529–536, 2002.
39. W. E, T. Li, and P.W. Zhang. Well-posedness for the dumbbell model of polymeric fluids. *Commun. Math. Phys.*, 248:409–427, 2004.
40. A. Ern and T. Lelièvre. Adaptive models for polymeric fluid flow simulation. *C. R. Acad. Sci. Paris, Ser. I*, 344(7):473–476, 2007.
41. J. Fang and R.G. Owens. Numerical simulations of pulsatile blood flow using a new constitutive model. *Biorheology*, 43:637–770, 2006.
42. R. Fattal and R. Kupferman. Constitutive laws for the matrix-logarithm of the conformation tensor. *J. Non-Newtonian Fluid Mech.*, 123:281–285, 2004.
43. R. Fattal and R. Kupferman. Time-dependent simulation of viscoelastic flows at high Weissenberg number using the log-conformation representation. *J. Non-Newtonian Fluid Mech.*, 126:23–37, 2005.
44. E. Fernández-Cara, F. Guillén, and R.R. Ortega. *Handbook of numerical analysis. Vol. 8: Solution of equations in $\mathbb{R}^n$ (Part 4). Techniques of scientific computing (Part 4). Numerical methods of fluids (Part 2).*, chapter Mathematical modeling and analysis of viscoelastic fluids of the Oldroyd kind, pages 543–661. Amsterdam: North Holland/ Elsevier, 2002.
45. G. Forest, Q. Wang, and R. Zhou. The flow-phase diagram of Doi-Hess theory for sheared nematic polymers II: finite shear rates. *Rheol. Acta*, 44(1):80–93, 2004.
46. M. Fortin and A. Fortin. A new approach for the FEM simulation of viscoelastic flows. *J. Non-Newtonian Fluid Mech.*, 32:295–310, 1989.
47. D. Frenkel and B. Smit. *Understanding molecular dynamics: from algorithms to applications*. Academic Press, London, 2002.
48. H. Gao and P. Klein. Numerical simulation of crack growth in an isotropic solid with randomized internal cohesive bonds. *J. Mech. Phys. Solids*, 46(2):187–218, 1998.
49. Y. Gati. *Modélisation mathématique et simulations numériques de fluides non newtoniens*. PhD thesis, Ecole Nationale des Ponts et Chaussées, 2004. Available at `http://pastel.paristech.org/883/01/these.pdf`. In French.
50. J.-F. Gerbeau, M. Vidrascu, and P. Frey. Fluid-structure interaction in blood flows on geometries coming from medical imaging. *Computers and Structure*, 83(2–3):155–165, 2005.
51. H. Giesekus. A simple constitutive equation for polymeric fluids based on the concept of deformation-dependent tensorial mobility. *J. Non-Newtonian Fluid Mech.*, 11:69–109, 1982.
52. R. Guénette and M. Fortin. A new mixed finite element method for computing viscoelastic flows. *J. Non-Newtonian Fluid Mech.*, 60:27–52, 1999.
53. C. Guillopé and J.C. Saut. Existence results for the flow of viscoelastic fluids with a differential constitutive law. *Nonlinear Analysis, Theory, Methods & Appl.*, 15(9):849–869, 1990.

54. C. Guillopé and J.C. Saut. Global existence and one-dimensional nonlinear stability of shearing motions of viscoelastic fluids of Oldroyd type. *RAIRO Math. Model. Num. Anal.*, 24(3):369–401, 1990.
55. E. Hairer, S.P. Nørsett, and G. Wanner. *Solving ordinary differential equations I.* Springer, 1992.
56. E. Hairer and G. Wanner. *Solving ordinary differential equations II.* Springer, 2002.
57. P. Halin, G. Lielens, R. Keunings, and V. Legat.The Lagrangian particle method for macroscopic and micro-macro viscoelastic flow computations. *J. Non-Newtonian Fluid Mech.*, 79:387–403, 1998.
58. D.J. Higham. An algorithmic introduction to numerical simulation of stochastic differential equations. *SIAM Review*, 43(3):525–546, 2001.
59. D. Hu and T. Lelièvre. New entropy estimates for the Oldroyd-B model, and related models. *Communications in Mathematical Sciences*, 5(4), 909–916, 2007.
60. M.A. Hulsen, R. Fattal, and R. Kupferman. Flow of viscoelastic fluids past a cylinder at high Weissenberg number: stabilized simulations using matrix logarithms. *Journal of Non-Newtonian Fluid Mechanics*, 127(1):27–39, 2005.
61. M.A. Hulsen, A.P.G. van Heel, and B.H.A.A. van den Brule. Simulation of viscoelastic flows using Brownian configuration fields. *J. Non-Newtonian Fluid Mech.*, 70:79–101, 1997.
62. P. Hébraud and F. Lequeux. Mode-coupling theory for the pasty rheology of soft glassy materials. *Phys. Rev. Lett.*, 81:2934–2937, 1998.
63. B. Jourdain, C. Le Bris, and T. Lelièvre. On a variance reduction technique for micro-macro simulations of polymeric fluids. *J. Non-Newtonian Fluid Mech.*, 122:91–106, 2004.
64. B. Jourdain, C. Le Bris, and T. Lelièvre. An elementary argument regarding the long-time behaviour of the solution to a stochastic differential equation. *Annals of Craiova University, Mathematics and Computer Science series*, 32:1–9, 2005.
65. B. Jourdain, C. Le Bris, T. Lelièvre, and F. Otto. Long-time asymptotics of a multiscale model for polymeric fluid flows. *Archive for Rational Mechanics and Analysis*, 181(1):97–148, 2006.
66. B. Jourdain and T. Lelièvre. Mathematical analysis of a stochastic differential equation arising in the micro-macro modelling of polymeric fluids. In I.M. Davies, N. Jacob, A. Truman, O. Hassan, K. Morgan, and N.P. Weatherill, editors, *Probabilistic Methods in Fluids Proceedings of the Swansea 2002 Workshop*, pages 205–223. World Scientific, 2003.
67. B. Jourdain, T. Lelièvre, and C. Le Bris. Numerical analysis of micro-macro simulations of polymeric fluid flows: a simple case. *Math. Models and Methods in Applied Sciences*, 12(9):1205–1243, 2002.
68. B. Jourdain, T. Lelièvre, and C. Le Bris. Existence of solution for a micro-macro model of polymeric fluid: the FENE model. *Journal of Functional Analysis*, 209:162–193, 2004.
69. I. Karatzas and S.E. Shreve. *Brownian motion and stochastic calculus.* Springer-Verlag, 1988.
70. R. Keunings. *Fundamentals of Computer Modeling for Polymer Processing*, chapter Simulation of viscoelastic fluid flow, pages 402–470. Hanser, 1989.
71. R. Keunings. A survey of computational rheology. In D.M. Binding et al., editor, *Proc. 13th Int. Congr. on Rheology*, pages 7–14. British Society of Rheology, 2000.

72. R. Keunings. Micro-macro methods for the multiscale simulation of viscoelastic flows using molecular models of kinetic theory. In D.M. Binding and K. Walters, editors, *Rheology Reviews 2004*. British Society of Rheology, 2004.
73. P.E. Kloeden and E. Platen. *Numerical Solution of Stochastic Differential Equations*, volume 23 of *Applications of Mathematics*. Springer, 1992.
74. Y. Kwon. Finite element analysis of planar 4:1 contraction flow with the tensor-logarithmic formulation of differential constitutive equations. *Korea-Australia Rheology Journal*, 16(4):183–191, 2004.
75. M. Laso and H.C. Öttinger. Calculation of viscoelastic flow using molecular models : The CONNFFESSIT approach. *J. Non-Newtonian Fluid Mech.*, 47:1–20, 1993.
76. M. Laso, M. Picasso, and H.C. Öttinger. Two-dimensional, time-dependent viscoelastic flow calculations using CONNFFESSIT. *AIChE J.*, 43:877–892, 1997.
77. C. Le Bris. *Systèmes multiéchelles: modélisation et simulation*, volume 47 of *Mathématiques et Applications*. Springer, 2005. In French.
78. C. Le Bris and P. L. Lions. Renormalized solutions to some transport equations with partially $W^{1,1}$ velocities and applications. *Annali di Matematica pura ed applicata*, 183:97–130, 2004.
79. C. Le Bris and P. L. Lions. Existence and uniqueness of solutions to Fokker-Planck type equations with irregular coefficients. *Comm. Part. Diff. Eq.*, 33(7), 1272–1317, 2008.
80. A. Lefebvre and B. Maury. Apparent viscosity of a mixture of a Newtonian fluid and interacting particles. *Comptes Rendus Académie des Sciences, Mécanique*, 333(12):923–933, 2005.
81. T. Lelièvre. Optimal error estimate for the CONNFFESSIT approach in a simple case. *Computers and Fluids*, 33:815–820, 2004.
82. T. Lelièvre. *Problèmes mathématiques et numériques posés par la simulation d'écoulement de fluides polymériques*. PhD thesis, Ecole Nationale des Ponts et Chaussées, 2004. Available at `http://cermics.enpc.fr/~lelievre/rapports/these.pdf`. In French.
83. T. Li, H. Zhang, and P.W. Zhang. Local existence for the dumbbell model of polymeric fluids. *Comm. Part. Diff. Eq.*, 29(5-6):903–923, 2004.
84. T. Li and P.W. Zhang. Convergence analysis of BCF method for Hookean dumbbell model with finite difference scheme. *SIAM MMS*, 5(1):205–234, 2006.
85. T. Li and P.W. Zhang. Mathematical analysis of multi-scale models of complex fluids. *Comm. Math. Sci.*, 5(1):1–51, 2007.
86. F.-H. Lin, C. Liu, and P.W. Zhang. On hydrodynamics of viscoelastic fluids. *Comm. Pure Appl. Math.*, 58(11):1437–1471, 2005.
87. F.-H. Lin, C. Liu, and P.W. Zhang. On a micro-macro model for polymeric fluids near equilibrium. *Comm. Pure Appl. Math.*, 60(6):838–866, 2007.
88. F. Lin, P. Zhang and Z. Zhang. On the global existence of smooth solution to the 2-D FENE dumbbell model, *Comm. Math. Phys.*, 277, 531–553, 2008.
89. P.L. Lions and N. Masmoudi. Global solutions for some Oldroyd models of non-Newtonian flows. *Chin. Ann. Math., Ser. B*, 21(2):131–146, 2000.
90. P.L. Lions and N. Masmoudi. Global existence of weak solutions to micro-macro models. *C. R. Math. Acad. Sci.*, 345(1):15–20, 2007.
91. A.S. Lodge. *Elastic Liquids*. Academic Press, 1964.
92. A. Lozinski. *Spectral methods for kinetic theory models of viscoelastic fluids*. PhD thesis, Ecole Polytechnique Fédérale de Lausanne, 2003. Available at `http://library.epfl.ch/theses/?nr=2860`.
93. A. Lozinski and C. Chauvière. A fast solver for Fokker-Planck equation applied to viscoelastic flows calculations. *J. Comp. Phys.*, 189(2):607–625, 2003.

94. L. Machiels, Y. Maday, and A.T. Patera. Output bounds for reduced-order approximations of elliptic partial differential equations. *Comput. Methods Appl. Mech. Engrg.*, 190(26-27):3413–3426, 2001.
95. F. Malrieu. *Inégalités de Sobolev logarithmiques pour des problèmes d'évolution non linéaires*. PhD thesis, Université Paul Sabatier, 2001.
96. J.M. Marchal and M.J. Crochet. A new mixed finite element for calculating viscoelastic flows. *J. Non-Newtonian Fluid Mech.*, 26:77–114, 1987.
97. N. Masmoudi. Well-posedness for the FENE dumbbell model of polymeric flows. *Communications on Pure and Applied Mathematics,* 61(12), 1685–1714, 2008.
98. T. Min, J.Y. Yoo, and H. Choi. Effect of spatial discretization schemes on numerical solutions of viscoelastic fluid flows. *J. Non-Newtonian Fluid Mech.*, 100:27–47, 2001.
99. J.T. Oden and S. Prudhomme. Estimation of modeling error in computational mechanics. *J. Comput. Phys.*, 182:496–515, 2002.
100. J.T. Oden and K.S. Vemaganti. Estimation of local modeling error and goal-oriented adaptive modeling of heterogeneous materials. i. error estimates and adaptive algorithms. *J. Comput. Phys.*, 164:22–47, 2000.
101. B. Øksendal. *Stochastic differential equations. An introduction with applications*. Springer, 2003.
102. H.C. Öttinger. *Stochastic Processes in Polymeric Fluids*. Springer, 1995.
103. R.G. Owens. A new microstructure-based constitutive model for human blood. *J. Non-Newtonian Fluid Mech.*, 140:57–70, 2006.
104. R.G. Owens and T.N. Phillips. *Computational rheology*. Imperial College Press / World Scientific, 2002.
105. A. Peterlin. Hydrodynamics of macromolecules in a velocity field with longitudinal gradient. *J. Polym. Sci. B*, 4:287–291, 1966.
106. N. Phan-Thien and R.I. Tanner. A new constitutive equation derived from network theory. *J. Non-Newtonian Fluid Mech.*, 2:353–365, 1977.
107. A. Quarteroni and L. Formaggia. *Mathematical modelling and numerical simulation of the cardiovascular system.*, volume 12 of *Handbook of Numerical Analysis*, Chap. 1, pages 3–127. Elsevier, 2004. G. Ciarlet Ed. N. Ayache guest Ed.
108. S. Reese. Meso-macro modelling of fibre-reinforced rubber-like composites exhibiting large elastoplastic deformation. *International Journal of Solids and Structures*, 40(4):951–980, 2003.
109. S. Reese. A micromechanically motivated material model for the thermo-viscoelastic material behaviour of rubber-like polymers. *International Journal of Plasticity*, 19(7):909–940, 2003.
110. M. Renardy. Local existence of solutions of the Dirichlet initial-boundary value problem for incompressible hypoelastic materials. *SIAM J. Math. Anal.*, 21(6):1369–1385, 1990.
111. M. Renardy. An existence theorem for model equations resulting from kinetic theories of polymer solutions. *SIAM J. Math. Anal.*, 22:313–327, 1991.
112. M. Renardy. *Mathematical analysis of viscoelastic flows*. SIAM, 2000.
113. D. Revuz and M. Yor. *Continuous martingales and Brownian motion*. Springer-Verlag, 1994.
114. L.C.G. Rogers and D. Williams. *Diffusions, Markov Processes, and Martingales, Volume 1: Foundations*. Cambridge University Press, 2000.
115. L.C.G. Rogers and D. Williams. *Diffusions, Markov Processes, and Martingales, Volume 2: Itô calculus*. Cambridge University Press, 2000.
116. D. Sandri. Non integrable extra stress tensor solution for a flow in a bounded domain of an Oldroyd fluid. *Acta Mech.*, 135(1-2):95–99, 1999.

117. D. Stroock and S.R.S. Varadhan. *Multidimensional diffusion processes*. Springer, 1979.
118. J.K.C. Suen, Y.L. Joo, and R.C. Armstrong. Molecular orientation effects in viscoelasticity. *Annu. Rev. Fluid Mech.*, 34:417–444, 2002.
119. A.P.G. Van Heel. *Simulation of viscoelastic fluids: from microscopic models to macroscopic complex flows*. PhD thesis, Delft University of Technology, 2000.
120. T. von Petersdorff and C. Schwab. Numerical solution of parabolic equations in high dimensions. *M2AN Math. Model. Numer. Anal.*, 38(1):93–127, 2004.
121. P. Wapperom, R. Keunings, and V. Legat. The Backward-tracking Lagrangian Particle Method for transient viscoelastic flows. *J. Non-Newtonian Fluid Mech.*, 91:273–295, 2000.
122. H. Zhang and P.W. Zhang. A theoretical and numerical study for the rod-like model of a polymeric fluid. *Journal of Computational Mathematics*, 22(2):319–330, 2004.
123. H. Zhang and P.W. Zhang. Local existence for the FENE-dumbbell model of polymeric fluids. *Archive for Rational Mechanics and Analysis*, 2:373–400, 2006.
124. L. Zhang, H. Zhang, and P. Zhang. Global existence of weak solutions to the regularized Hookean dumbbell model. *Commun. Math. Sci.*, 6(1), 83–124, 2008.

Fast Algorithms for Boundary Integral Equations

Lexing Ying

Department of Mathematics, University of Texas, Austin, TX 78712, USA,
lexing@math.utexas.edu

Summary. This article reviews several fast algorithms for boundary integral equations. After a brief introduction of the boundary integral equations for the Laplace and Helmholtz equations, we discuss in order the fast multipole method and its kernel independent variant, the hierarchical matrix framework, the wavelet based method, the high frequency fast multipole method, and the recently proposed multidirectional algorithm.

Keywords. Boundary integral equations, Laplace equation, Helmholtz equation, fast algorithms, fast multipole method, hierarchical matrices, wavelets, multiscale methods, multidirectional methods.
AMS subject classifications. 45A05, 65R20.

1 Outline

Many physical problems can be formulated as partial differential equations (PDEs) on certain geometric domains. For some of them, the PDEs can be reformulated using the so-called boundary integral equations (BIEs). These are integral equations which only involve quantities on the domain boundary. Some advantages of working with the BIEs are automatic treatments of boundary condition at infinity, better condition numbers, and fewer numbers of unknowns in the numerical solution. On the other hand, one of the major difficulties of the BIEs is that the resulting linear systems are dense, which is in direct contrast to the sparse systems of the PDEs. For large scale problems, direct solution of these dense linear systems becomes extremely time-consuming. Hence, how to solve these dense linear systems efficiently has become one of the central questions. Many methods have been developed in the last twenty years to address this question. In this article, we review some of these results.

We start in Sect. 2 with a brief introduction of the boundary integral formulation with the Laplace and Helmholtz equations as our examples. A major difference between these two equations is that the kernel of the Laplace equation is non-oscillatory while the one of the Helmholtz equation is oscillatory. For the non-oscillatory kernels, we discuss the fast multipole method (FMM) in Sect. 3 and its kernel independent variant in Sect. 4, the hierarchical matrices frame in Sect. 5, and the wavelet based methods in Sect. 6. For the oscillatory kernels, we review the high frequency fast multipole method (HF-FMM) in Sect. 7 and the recently developed multidirectional method in Sect. 8.

The purpose of this article is to provide an introduction to these methods for advanced undergraduate and graduate students. Therefore, our discussion mainly focuses on algorithmic ideas rather than theoretical estimates. For the same reason, we mostly refer only to the original papers of these methods and keep the size of the reference list to a minimum. Many important results are not discussed here due to various limitations and we apologize for that.

2 Boundary Integral Formulation

Many linear partial differential equation problems have boundary integral equation formulations. In this section, we focus on two of the most important examples and demonstrate how to transform the PDE formulations into the BIE formulations. Our discussion mostly follows the presentation in [11, 18, 20]. We denote $\sqrt{-1}$ with i and assume that all functions are sufficiently smooth.

2.1 Laplace equation

Let D be a bounded domain with smooth boundary in $\mathbb{R}^d (d \quad 2,3)$. n is the exterior normal to D. The Laplace equation on D with Dirichlet boundary condition is

$$-\Delta u \quad 0 \quad \text{in}\, D \tag{1}$$

$$u \quad f \quad \text{on}\, \partial D \tag{2}$$

where f is defined on ∂D. The geometry of the problem is shown in Fig. 1. We seek to represent $u(x)$ for $x \in D$ in an integral form which uses only quantities on the boundary ∂D.

The Green's function for the Laplace equation is

$$G(x,y) \quad \begin{cases} \frac{1}{2\pi} \ln \frac{1}{|x-y|} & (d \quad 2) \\ \frac{1}{4\pi} \frac{1}{|x-y|} & (d \quad 3) \end{cases} \tag{3}$$

Some of the important properties of $G(x,y)$ are

- $G(x,y)$ is symmetric in x and y,
- $G(x,y)$ is non-oscillatory, and
- $-\Delta_x G(x,y) \quad \delta_y(x)$ and $-\Delta_y G(x,y) \quad \delta_x(y)$

where Δ_x and Δ_y take the derivatives with respect x and y, respectively, and δ_x is the Dirac function located at x. The following theorem is a simple consequence of Stokes' theorem.

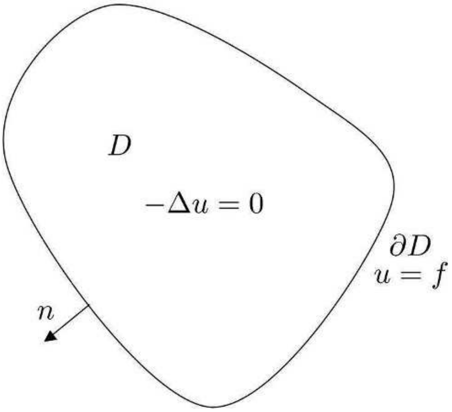


Fig. 1. Domain of the Dirichlet boundary value problem of the Laplace equation.

Theorem 1. *Let u and v to be two sufficiently smooth functions on $\bar{D}$. Then*

$$\int_D (u\Delta v - v\Delta u)dx \qquad \int_{\partial D}\left(u\frac{\partial v(y)}{\partial n(y)} - v\frac{\partial u(y)}{\partial n(y)}\right) ds(y).$$

A simple application of the previous theorem gives the following result.

Theorem 2. *Let u be a sufficiently smooth function on $\bar{D}$ such that $-\Delta u \quad 0$ in D. For any x in D,*

$$u(x) \qquad \int_{\partial D}\left(\frac{\partial u(y)}{\partial n(y)}G(x,y) - u(y)\frac{\partial G(x,y)}{\partial n(y)}\right) ds(y).$$

Proof. Pick a small ball B at x that is contained in D (see Fig. 2). From the last theorem, we have

$$\begin{aligned} &\int_{D\backslash B}(u(y)\Delta G(x,y) - G(x,y)\Delta u(y))ds(y) \\ &\int_{\partial(D\backslash B)}\left(u(y)\tfrac{\partial G(x,y)}{\partial n(y)} - G(x,y)\tfrac{\partial u(y)}{\partial n(y)}\right) ds(y). \end{aligned}$$

Since $-\Delta u(y) \quad 0$ and $-\Delta G(x,y) \quad 0$ for $y \in D \setminus B$, the left hand side is equal to zero. Therefore,

$$\begin{aligned} &\int_{\partial D}\left(u(y)\tfrac{\partial G(x,y)}{\partial n(y)} - G(x,y)\tfrac{\partial u(y)}{\partial n(y)}\right) ds(y) \\ -&\int_{\partial B}\left(u(y)\tfrac{\partial G(x,y)}{\partial n(y)} - G(x,y)\tfrac{\partial u(y)}{\partial n(y)}\right) ds(y) \end{aligned}$$

where n points towards x on ∂B. Now let the radius of the ball B go to zero. The first term of the right hand side goes to $-u(x)$ while the second term approaches 0.

From the last theorem, we see that $u(x)$ for x in D can be represented as a sum of two boundary integrals. In the boundary integral formulation, we seek to represent $u(x)$ using only one of them. This degree of freedom gives rise to the following two approaches.

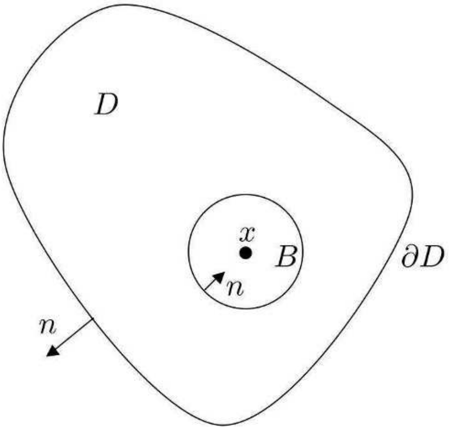


Fig. 2. Proof of Theorem 2.

Method 1

We represent $u(x)$ for $x \in D$ using the integral that contains $G(x,y)$

$$u(x) \quad \int_{\partial D} \varphi(y) G(x,y) ds(y) \tag{4}$$

where φ is an unknown density on ∂D. This formulation is called the *single layer form* and φ is often called the *single layer density*. One can show that any sufficiently nice $u(x)$ can be represented using the single layer form (see [20] for details). Letting x approach $z \in \partial D$, we get

$$f(z) \quad u(z) \quad \int_{\partial D} \varphi(y) G(z,y) ds(y),$$

which is an integral equation that involves only boundary quantities φ and f. Therefore, the steps to solve the Laplace equation using the single layer form are:

1. Find $\varphi(z)$ on ∂D such that

$$f(z) \quad \int_{\partial D} \varphi(y) G(z,y) ds(y). \tag{5}$$

   This equation is a *Fredholm equation of the first kind* (see [20]).

2. For x in D, compute $u(x)$ by

$$u(x) \quad \int_{\partial D} \varphi(y) G(x,y) ds(y). \tag{6}$$

Method 2

We can also represent $u(x)$ for $x \in D$ using the integral that contains $\frac{\partial G(x,y)}{\partial n(y)}$

$$u(x) = -\int_{\partial D} \varphi(y) \frac{\partial G(x,y)}{\partial n(y)} ds(y) \tag{7}$$

where φ is again an unknown density on ∂D. This formulation is called the *double layer form* and φ is the *double layer density*. In fact, the double layer form is capable of representing any sufficiently nice $u(x)$ in D [20]. If we now let x approach $z \in \partial D$, we obtain the following equation on the boundary:

$$f(z) = u(z) = \frac{1}{2}\varphi(z) - \int_{\partial D} \frac{\partial G(z,y)}{\partial n(y)} \varphi(y) ds(y).$$

The extra $\frac{1}{2}\varphi(z)$ term comes up because the integral (7) is not uniformly integrable near $z \in \partial D$. Hence, one cannot simply exchange the limit and integral signs. Since the boundary ∂D is smooth, the integral operator with the kernel $\frac{\partial G(z,y)}{\partial n(y)}$ is a compact operator. The steps to solve the Laplace equation using the double layer form are:

1. Find $\varphi(z)$ on ∂D such that

$$f(z) = \frac{1}{2}\varphi(z) - \int_{\partial D} \frac{\partial G(z,y)}{\partial n(y)} \varphi(y) ds(y). \tag{8}$$

   This equation is a *Fredholm equation of the second kind.*

2. For x in D, compute $u(x)$ with

$$u(x) = -\int_{\partial D} \frac{\partial G(x,y)}{\partial n(y)} \varphi(y) ds(y). \tag{9}$$

Between these two approaches, we often prefer to work with the double layer form (Method 2). The main reason is that the Fredholm equation of the second kind has a much better condition number, thus dramatically reducing the number of iterations required in a typical iterative solver.

2.2 Helmholtz equation

We now turn to the Helmholtz equation. Let D be a bounded domain with smooth boundary in $\mathbb{R}^d (d = 2,3)$ and n be the exterior normal to D. The unbounded Helmholtz equation on $\mathbb{R}^d \setminus \bar{D}$ $(d = 2,3)$ with Dirichlet boundary condition describes the scattering field of a sound soft object:

$$-\Delta u - k^2 u = 0 \quad \text{in } \mathbb{R}^d \setminus \bar{D} \tag{10}$$

$$u(x) = -u^{inc}(x) \quad \text{for } x \in \partial D \tag{11}$$

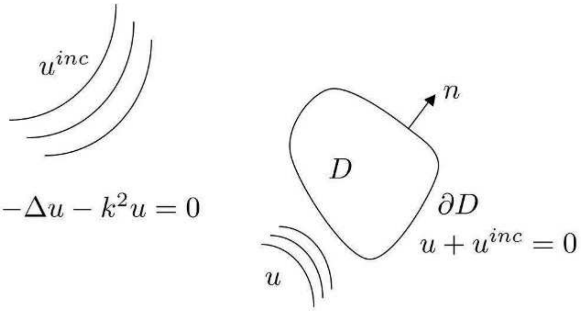


Fig. 3. Domain of the Dirichlet boundary value problem of the Helmholtz equation.

$$\lim_{r\to\infty} r^{\frac{d-1}{2}} \left(\frac{\partial u}{\partial r} - \mathrm{i}ku \right) \quad 0 \tag{12}$$

where k is the wave number, u^{inc} is the incoming field and u is the scattering field. The last equation is called the *Sommerfeld radiation condition* which guarantees that the scattering field propagates to infinity. The geometry of the problem is described in Fig. 3. Our goal is again to represent $u(x)$ for $x \in \mathbb{R}^d \setminus \bar{D}$ in an integral form which uses quantities defined on the boundary ∂D.

The Green's function of the Helmholtz equation is

$$G(x,y) \quad \begin{cases} \frac{\mathrm{i}}{4} H_0^1(k|x-y|) & (d \quad 2) \\ \frac{1}{4\pi} \frac{\exp(\mathrm{i}k|x-y|)}{|x-y|} & (d \quad 3) \end{cases} \tag{13}$$

Some of the important properties of $G(x,y)$ are

- $G(x,y)$ is symmetric,
- $G(x,y)$ is oscillatory,
- $(-\Delta_x - k^2)G(x,y) \quad \delta_y(x)$ and $(-\Delta_y - k^2)G(x,y) \quad \delta_x(y)$.

Theorem 3. *Let C be a bounded domain with smooth boundary. Suppose that u is sufficiently smooth in $\bar{C}$ and satisfies $(-\Delta - k^2)u \quad 0$ in C. Then for any x in C*

$$u(x) \quad \int_{\partial C} \left(\frac{\partial u(y)}{\partial n(y)} G(x,y) - u(y) \frac{\partial G(x,y)}{\partial n(y)} \right) ds(y).$$

Proof. Pick a small ball B centered at x. Then we have

$$\begin{aligned} &\int_{C\setminus B} (u(y)\Delta G(x,y) - G(x,y)\Delta u(y))dy \\ &\int_{\partial(C\setminus B)} \left(u(y) \frac{\partial G(x,y)}{\partial n(y)} - G(x,y) \frac{\partial u(y)}{\partial n(y)} \right) ds(y). \end{aligned}$$

The left hand side is equal to

$$\int_{C\setminus B} (u \cdot (\Delta G + k^2 G) - G(\Delta u + k^2 u))dy \quad 0.$$

The rest of the proof is the same as the one of Theorem 2.

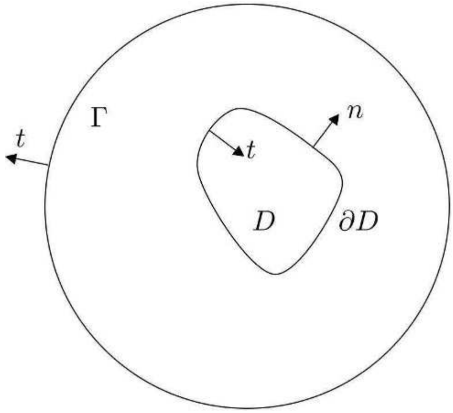


Fig. 4. Proof of Theorem 4.

The above theorem addresses a bounded domain C. However, what we are really interested in is the unbounded domain $\mathbb{R}^d \setminus \bar{D}$.

Theorem 4. *Suppose that u is sufficiently smooth and satisfies $(-\Delta - k^2)u \quad 0$ in $\mathbb{R}^d \setminus D$. Then for any x in $\mathbb{R}^d \setminus D$,*

$$u(x) \quad \int_{\partial D} \left(\frac{\partial u(y)}{\partial n(y)} G(x,y) - u(y) \frac{\partial G(x,y)}{\partial n(y)} \right) ds(y).$$

Proof. Pick a large ball Γ that contains D. Consider the domain $\Gamma \setminus \bar{D}$ (see Fig. 4). Let t be the exterior normal direction of $\Gamma \setminus \bar{D}$. From the previous theorem, we have

$$\begin{aligned} u(x) \quad & \int_{\partial \Gamma} \left(\frac{\partial u(y)}{\partial t} G(x,y) - u(y) \frac{\partial G(x,y)}{\partial t} \right) ds(y) + \\ & \int_{\partial D} \left(\frac{\partial u(y)}{\partial t} G(x,y) - u(y) \frac{\partial G(x,y)}{\partial t} \right) ds(y). \end{aligned}$$

Using the Sommerfeld condition at infinity, one can show that the integral over $\partial \Gamma$ goes to zero as one pushes the radius of Γ to infinity [11]. Noting that $t \quad -n$ on ∂D, we have

$$u(x) \quad \int_{\partial D} \left(u(y) \frac{\partial G(x,y)}{\partial n(y)} - \frac{\partial u(y)}{\partial n(y)} G(x,y) \right) ds(y).$$

From the last theorem, we see that $u(x)$ for x in $\mathbb{R}^d \setminus \bar{D}$ can be represented as a sum of two integrals. In the boundary integral formulation of the Helmholtz equation, one option is to represent $u(x)$ by the *double layer form*:

$$u(x) \quad \int_{\partial D} \frac{\partial G(x,y)}{\partial n(y)} \varphi(y) ds(y)$$

Different from the double layer form of the Laplace equation, the double layer form of the Helmholtz equation is not capable of representing arbitrary field $u(x)$ for $x \in \mathbb{R}^d \setminus \bar{D}$. If k is one of the internal resonant numbers such that the internal Neumann problem with zero boundary condition has non-trivial solution, then this double layer form is singular (see [11]). In practice, we use

$$u(x) \quad \int_{\partial D} \left(\frac{\partial G(x,y)}{\partial n(y)} - \mathrm{i}\eta G(x,y) \right) \varphi(y) ds(y).$$

where η is a real number (for example, $\eta \quad k$). As we let x approach z on ∂D, we get

$$-u^{inc}(z) \quad u(z) \quad \frac{1}{2}\varphi(z) + \int_{\partial D} \left(\frac{\partial G(z,y)}{\partial n(y)} - \mathrm{i}\eta G(z,y) \right) \varphi(y) ds(y)$$

where the extra term $\frac{1}{2}\varphi(z)$ is due to the fact that the integral is improper at $z \in \partial D$.

The steps to solve the Helmholtz equation using this double layer form are:

1. Find a function $\varphi(z)$ on ∂D such that

$$-u^{inc}(z) \quad \frac{1}{2}\varphi(z) + \int_{\partial D} \left(\frac{\partial G(z,y)}{\partial n(y)} - \mathrm{i}\eta G(z,y) \right) \varphi(y) ds(y). \tag{14}$$

2. For point x in $\mathbb{R}^3 \setminus D$, compute $u(x)$ with

$$u(x) \quad \int_{\partial D} \left(\frac{\partial G(x,y)}{\partial n(y)} - \mathrm{i}\eta G(x,y) \right) \varphi(y) ds(y). \tag{15}$$

We have seen the derivations of the BIEs for the interior Laplace Dirichlet boundary value problem and the exterior Helmholtz Dirichlet boundary value problem. Though both cases use the Green's functions of the underlying equation and the Stokes' theorem, the derivation for the Helmholtz equation is complicated by the existence of the internal resonant numbers. For other elliptic boundary value problems, the derivations of the BIE formulations often differ from case to case.

2.3 Discretization

In both BIEs discussed so far, we need to solve a problem of the following form: find $\varphi(x)$ on ∂D such that

$$f(x) \quad \varphi(x) + \int_{\partial D} K(x,y)\varphi(y)ds(y), \quad i.e., \quad f \quad (I+K)\varphi$$

or

$$f(x) \quad \int_{\partial D} K(x,y)\varphi(y)ds(y), \quad i.e., \quad f \quad K\varphi.$$

where $K(x,y)$ is either the Green's function or its derivative of the underlying PDE. In order to solve these equations numerically, we often use one of the following three discretization methods: the Nyström method, the collocation method, and the Galerkin method. Let us discuss these methods briefly using the Fredholm equation of the second kind.

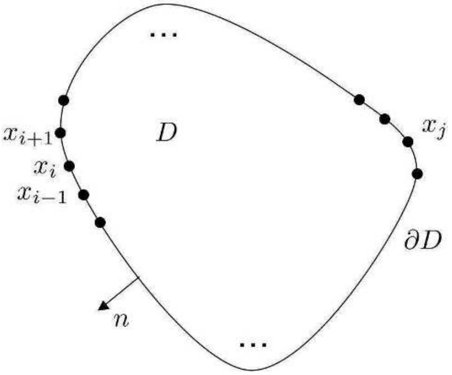


Fig. 5. Nyström method

Nyström method

The idea of the Nyström method is to approximate integral operators with quadrature operators. The steps are:

1. Approximate the integral operator $(K\varphi)(x) :\quad \int K(x,y)\varphi(y)dy$ with the quadrature operator
$$(K_N\varphi)(x) :\quad \sum_{j\quad 1}^{N} K(x,x_j)\lambda_j\varphi(x_j)$$
where $\{x_j\}$ are the quadrature points and $\{\lambda_j\}$ are the quadrature weights (see Fig. 5). Here we make the assumption that $\{\lambda_j\}$ are independent of x. In practice, $\{\lambda_j\}$ often depend on x when x_j is in the neighborhood of x if the kernel $K(x,y)$ has a singularity at $x \quad y$.
2. Find $\varphi(x)$ such that $\varphi + K_N\varphi \quad f$. We write down the equation at $\{x_i\}$:
$$\varphi_i + \sum_{j\quad 1}^{n} K(x_i,x_j)\lambda_j\varphi_j \quad f_i, \quad i \quad 1,\cdots,N \tag{16}$$
and solve for $\{\varphi_i\}$. Here $f_i \quad f(x_i)$.
3. The value of $\varphi(x)$ at $x \in \partial D$ is computed using
$$\varphi(x) \quad f(x) - \sum_{j\quad 1}^{n} K(x,x_j)\lambda_j\varphi_j. \tag{17}$$

Collocation method

The idea of the collocation method is to use subspace approximation. The steps are:

1. Approximate $\varphi(x)$ by $\sum_{j\quad 1}^{N} c_j\varphi_j(x)$ where $\{\varphi_j(x)\}$ are basis functions on ∂D. Let $\{x_j\}$ be a set of points on ∂D (see Fig. 6), where x_j is often the center of $supp(\varphi_j)$.

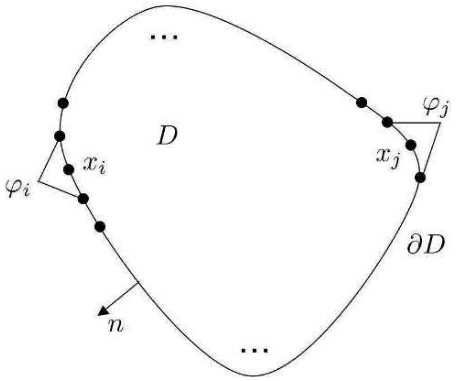


Fig. 6. Collocation and Galerkin methods

2. Find $\{c_j\}$ such that $\varphi + K\varphi \quad f$ is satisfied at $\{x_j\}$, i.e.,

$$\sum_{j \quad 1}^{N} c_j \varphi_j(x_i) + (K(\sum_{j \quad 1}^{N} c_j \varphi_j))(x_i) \quad f(x_i), \quad i \quad 1, \cdots, N \tag{18}$$

Galerkin method

The idea of the Galerkin method is to use space approximation with orthogonalization. The steps are:

1. Approximate $\varphi(x)$ by $\sum_{j \quad 1}^{N} c_j \varphi_j(x)$ where $\{\varphi_j(x)\}$ are often localized basis functions on ∂D.
2. Find $\{c_j\}$ such that $\varphi + K\varphi - f$ is orthogonal to all the subspace generated by $\varphi\ (x)\}$.

$$\langle \varphi_i, \sum_{j \quad 1}^{N} c_j \varphi_j + K(\sum_{j \quad 1}^{N} c_j \varphi_j) - f \rangle \quad 0, \quad i \quad 1, \cdots, N \tag{19}$$

2.4 Iterative solution

The following discussion is in the setting of the Nyström method. The situations for the other methods are similar. In the matrix form, the linear system that one needs to solve is

$$(\boldsymbol{I} + \boldsymbol{K\Lambda})\boldsymbol{\varphi} \quad \boldsymbol{f}$$

where $\boldsymbol{I}$ is the identity matrix, $\boldsymbol{K}$ is the matrix with entries $K(x_i, x_j)$, $\boldsymbol{\Lambda}$ is the diagonal matrix with the diagonal entries equal to $\{\lambda_j\}$, $\boldsymbol{\varphi}$ is the vector of $\{\varphi_j\}$, and $\boldsymbol{f}$ is the vector of $\{f_j\}$.

Since $\boldsymbol{K}$ is a dense matrix, the direct solution of this equation takes $O(N^3)$ steps. For large N, this becomes extremely time-consuming and solving this system directly is not feasible. Therefore, we need to resort to iterative solvers.

Since the integral operator K is compact, its eigenvalues decay to zero. This is also true for the discretized version, the matrix $\boldsymbol{K}$. Therefore, the condition number of $\boldsymbol{I}+\boldsymbol{K\Lambda}$ is small and independent of the number of quadrature points N. As a result, the number of iterations is also independent of N. In each iteration, one computes $\boldsymbol{K\psi}$ for a given vector $\boldsymbol{\psi}$. Since $\boldsymbol{K}$ is dense, a naive implementation of this matrix-vector multiplication takes $O(N^2)$ steps, which can be still quite expensive for large values of N. How to compute the product $\boldsymbol{K\psi}$ is the question that we will address in the following sections.

Before we move on, let us compare the PDE and BIE formulations. For the PDE formulations, a numerical solution often requires $O((1/h)^d)$ unknowns for a given discretization size h. Special care is necessary for unbounded exterior problems. Since the resulting linear system is sparse, each iteration of the iterative solver is quite fast though the number of iterations might be large. Finally, the PDE formulations work for domains with arbitrary geometry and problems with variable coefficients.

For the BIE formulations, a numerical solution involves only $O((1/h)^{d-1})$ unknowns on the domain boundary for a given discretization size h. No special care is needed for exterior domains. The resulting system is always dense, so fast algorithms are necessary for efficient iterative solutions of the BIE formulations. As we have seen already, the Green's functions are fundamental in deriving the integral equations. Since the Green's functions are often unknown for problems with variable coefficients, most applications of the BIE formulations are for problems with constant coefficient.

3 Fast Multipole Method

In each step of the iterative solution of a BIE formulation, we face the following problem. Given a set of charges $\{f_i, 1 \le i \le N\}$ located at points $\{p_i, 1 \le i \le N\}$ (see Fig. 7) and the Green's function $G(x,y)$ of the underlying equation, we want to compute at each p_i the potential

$$u_i \quad \sum_{j \quad 1}^{N} G(p_i, p_j) f_j. \tag{20}$$

As we pointed earlier, a naive algorithm takes $O(N^2)$ steps, which can be quite expensive for large values of N. In this section, we introduce the fast multipole method by Greengard and Rokhlin [15, 16] for the Green's function of the Laplace equation. This remarkable algorithm reduces the complexity from $O(N^2)$ to $O(N)$ for any fixed accuracy ε.

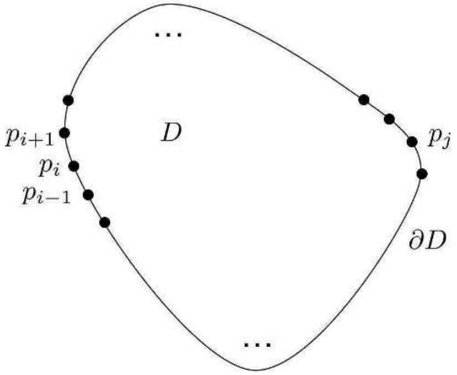


Fig. 7. Distribution of quadrature points $\{p_i\}$ on the boundary of the domain D.

3.1 Geometric part

Two sets A and B are said to be *well-separated* if the distance between A and B are greater than their diameters. Let us consider the interaction from a set of points $\{y_j\}$ in B to a set of points $\{x_i\}$ in A, where both $\{y_j\}$ and $\{x_i\}$ are subsets of $\{p_i\}$. The geometry is shown in Fig. 8.

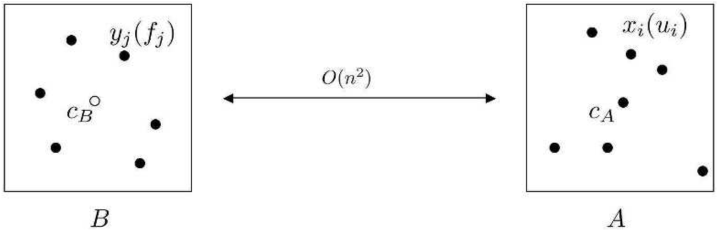


Fig. 8. Two boxes A and B are well-separated. Direct computation takes $O(N^2)$ steps.

Suppose that $\{f_j\}$ are the charges at $\{y_j\}$. Let us consider the following approximation for the potential u_i at each x_i

$$u_i \approx u(c_A) \quad \sum_j G(c_A,y_j)f_j \approx G(c_A,c_B)\sum_j f_j. \tag{21}$$

This approximation is quite accurate when A and B are far away from each other and is in fact used quite often in computational astrophysics to compute the interaction between distant galaxies. However, for two sets A and B which are merely well-separated (the distance between them is comparable to their diameters), this approximation introduces significant error. Let us not worry too much about the accuracy

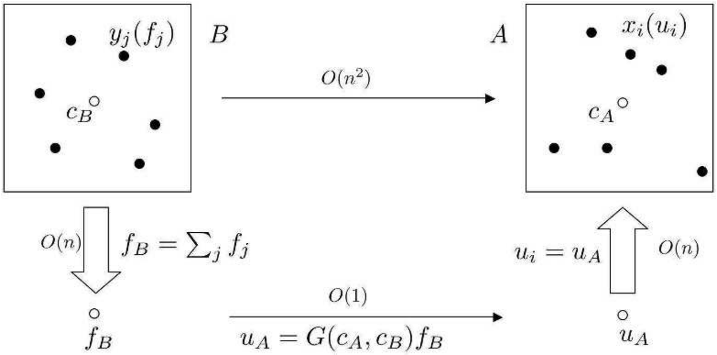


Fig. 9. The three steps of the approximate procedure. The total number operations is $O(N)$.

at this moment and we will come back to this point later. A geometric description of this approximation is given in Fig. 9.

We have introduced two representations in this simple approximation:

- f_B, the *far field representation* of B that allows one to approximately reproduce in the far field of B the potential generated by the source charges inside B.
- u_A, the *local field representation* of A that allows one to approximately reproduce inside A the potential generated by the source charges in the far field of A.

The computation of u_A from f_B

$$u_A \quad G(c_A, c_B) f_B$$

is called a *far-to-local* translation. Assuming both $\{y_j\}$ and $\{x_i\}$ contain $O(n)$ points, the naive direct computation of the interaction takes $O(n^2)$ steps. The proposed approximation is much more efficient:

- $f_B \quad \sum_j f_j$ takes $O(n)$ steps.
- $u_A \quad G(c_A, c_B) f_B$ takes $O(1)$ steps.
- $u_i \quad u_A$ for all $x_i \in A$ takes $O(n)$ steps as well.

Hence, the complexity of this three step procedure is $O(n)$. Viewing the interaction between A and B in a matrix form, we see that this interaction is approximately low rank if A and B are well-separated. In fact, in the above approximation, a rank-1 approximation is used.

However, in the problem we want to address, all the points $\{p_i\}$ are mixed together and each p_i is both a source and a target. Therefore, one cannot apply the above procedure directly. The solution is to use an adaptive tree structure, namely the octree in 3D or the quadtree in 2D (see Fig. 10). We first choose a box that contains all the points $\{p_i\}$. Starting from this top level box, each box of the quadtree is recursively partitioned unless the number of points inside it is less than a prescribed constant (in practice this number can vary from 50 to 200). Assuming that the points $\{p_i\}$ are distributed quite uniformly on ∂D, the number of levels of the quadtree is

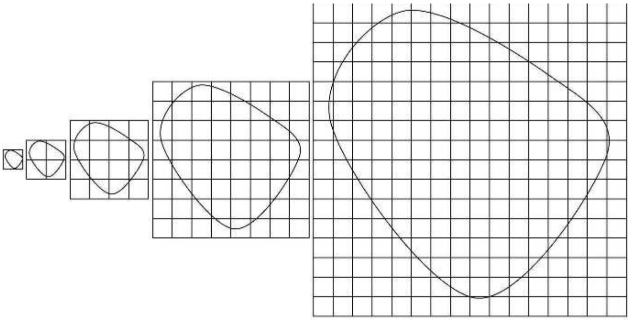


Fig. 10. The quadtree generated from the domain in Fig. 7. Different levels of the quadtree are shown from left to right.

$O(\log N)$. For a given box B in the quadtree, all the adjacent boxes are said to be in the *near field* while the rest are in the *far field*. The *interaction list* of B contains the boxes on the same level that are *in B's far field but not the far field of B's parent*. It is not difficult to see that the size of the interaction list is always $O(1)$.

No computation is necessary at the zeroth and the first levels. At the second level (see Fig. 11), each box B has $O(1)$ well-separated boxes (e.g. A). These boxes are colored in gray and in B's interaction list. The interaction between B and each box in its interaction list can be approximated using the three step procedure described above. The same computation is repeated over all the boxes on this level.

To address the interaction between B and its adjacent boxes, we go to the next level (see Fig. 12). Suppose that B' is a child of B. Since the interaction between B' and B's far field has already been taken care of in the previous level, we only need

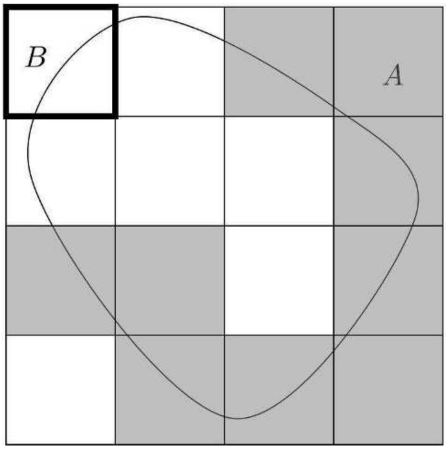


Fig. 11. Computation at the second level.

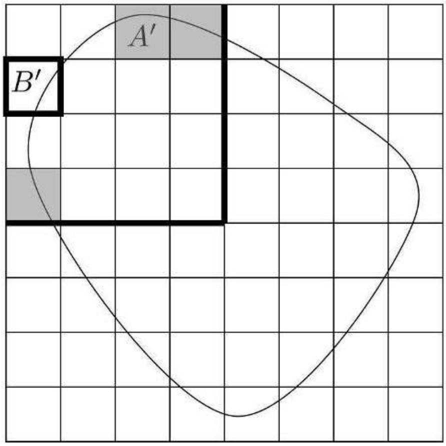


Fig. 12. Computation at the third level.

to address the interaction between B' and the boxes in B''s interaction list (e.g. A'). These boxes are also colored in gray and the interaction between B' and each one of them can be approximated again using the three step procedure described above.

To address the interaction between B' and its adjacent boxes, we again go to the next level (see Fig. 13). B'' (a child of B') has $O(1)$ boxes in its interaction list. The interaction between B'' and each one of them (e.g. A'') is once again computed using the three step procedure described above. Suppose now that B'' is also a leaf box. We then need to address the interaction between B'' and its adjacent boxes. Since the number of points in each leaf box is quite small, we simply use the direct computation for this.

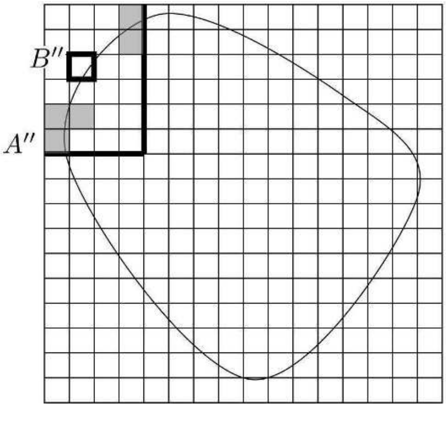


Fig. 13. Computation at the fourth (last) level.

The algorithm is summarized as follows:

1. At each level, for each box B, compute $f_B \quad \sum_{p_j \in B} f_j$.
2. At each level, for each pair A and B in each other's interaction list, add $G(c_A, c_B) f_B$ to u_A. This is the far-to-local translation.
3. At each level, for each box A, add u_A to u_j for each $p_j \in A$.
4. At the final level, for each leaf box B, compute the interaction with its adjacent boxes directly.

The complexity of this algorithm is $O(N \log N)$ based on the following considerations.

1. Each point belongs to one box in each of $O(\log N)$ levels. The complexity of the first step is $O(N \log N)$.
2. There are $O(N)$ boxes in the octree. Each box has $O(1)$ boxes in the interaction list. Since each far-to-local translation takes $O(1)$ operations, the complexity of the second step is $O(N)$.
3. Each point belongs to one box in each of $O(\log N)$ levels. The complexity is $O(N \log N)$.
4. There are $O(N)$ leaf boxes in total. Each one has $O(1)$ neighbors. Since each leaf box contains only $O(1)$ points, the direct computation costs $O(N)$ steps.

As we have mentioned earlier, the goal is $O(N)$. Can we do better? The answer is yes. Let us take a look at a box B and its children $B_1, \cdots, B_4$. Based on the definition of f_B, we have

$$f_B \quad \sum_{p_j \in B} f_j \quad \sum_{p_j \in B_1} f_j + \sum_{p_j \in B_2} f_j + \sum_{p_j \in B_3} f_j + \sum_{p_j \in B_4} f_j \quad f_{B_1} + f_{B_2} + f_{B_3} + f_{B_4}.$$

Therefore, once $\{f_{B_i}\}$ are all known, f_B can be computed using only $O(1)$ operations. This step is called a *far-to-far translation*. The dependence between f_B and $\{f_{B_i}\}$ suggests that we traverse the quadtree bottom-up during the construction of the far field representations.

Similarly, instead of putting u_A to each of its points, it is sufficient to add u_A to $\{u_{A_i}\}$ where $\{A_i\}$ are the children of A. The reason is that u_{A_i} will eventually be added to the individual points. This step of adding u_A to $\{u_{A_i}\}$ obviously takes $O(1)$ operations as well and it is called a *local-to-local translation*. Since $\{u_{A_i}\}$ now depend on u_A, we need to traverse the octree top-down during the computation of the local field representations.

Combining the far-to-far and local-to-local translations with the above algorithm, we have the complete the description of the geometric structure of the FMM.

1. Bottom-up traversal of the octree. At each level, for each box B,
   - if leaf, compute f_B from the points in B,
   - if non-leaf, compute f_B from the far field representations of its children.
2. At each level, for each pair A and B in each other's interaction list, add $G(c_A, c_B) f_B$ to u_A.
3. Top-down traversal of the octree. At each level, for each box A,

  - if leaf, add u_A to u_j for each point p_j in A,
  - if non-leaf, add u_A to the local field representations of its children.
4. At the final level, for each leaf box B, compute the interaction with its adjacent boxes directly.

Compared with the previous version, the only changes are made in the first and the third steps, while the second and the fourth steps remain the same. Let us estimate its complexity. It is obvious that we perform one far-to-far translation and one local-to-local translation to each of the $O(N)$ boxes in the octree. Since each of the far-to-far and local-to-local translations takes only $O(1)$ operations, the complexity of the first and the third steps is clearly $O(N)$. Therefore, the overall complexity of the algorithm is $O(N)$.

3.2 Analytic part

In the discussion of the geometric part of the FMM, we did not worry too much about the accuracy. In fact, simply taking the far field representation $f_B \quad \sum_{p_j \in B} f_j$ and the local field representation $u_A \quad G(c_A, c_B) f_B$ gives very low accuracy. Next, we discuss the analytic part of the FMM, which provides efficient representations and translations that achieve any prescribed accuracy ε. In fact one can view the $f_B \quad \sum_{p_j \in B} f_j$ to be the zeroth moment of the charge distribution $\{f_j\}$ at $\{p_j\}$ in B. The idea behind the analytic part of the FMM is simply to utilize the higher order moments and represent them compactly using the property of the underlying PDE.

2D case

In the two dimensional case, we can regard $\{p_i\}$ to be points in the complex plane. Up to a constant,

$$G(x,y) \quad \ln|x-y| \quad \mathrm{Re}(\ln(x-y))$$

for $x, y \in \mathbb{C}$. Therefore, we will regard the kernel to be $G(x,y) \quad \ln(x-y)$ and throw away the imaginary part at the end of the computation.

Far field representation

Suppose that $\{y_j\}$ are source points inside a box (see Fig. 14) and $\{f_j\}$ are charges located at $\{y_j\}$.

Since

$$G(x,y) \quad \ln(x-y) \quad \ln x + \ln\left(1 - \frac{y}{x}\right) \quad \ln x + \sum_{k\ 1}^{\infty} \left(-\frac{1}{k}\right) \frac{y^k}{x^k},$$

we have for any x in the far field of this box

$$u(x) \quad \sum_j G(x, y_j) f_j \quad \left(\sum_j f_j\right) \ln x + \sum_{k\ 1}^{p} \left(-\frac{1}{k} \sum_j y_j^k f_j\right) \frac{1}{x^k} + O(\varepsilon)$$

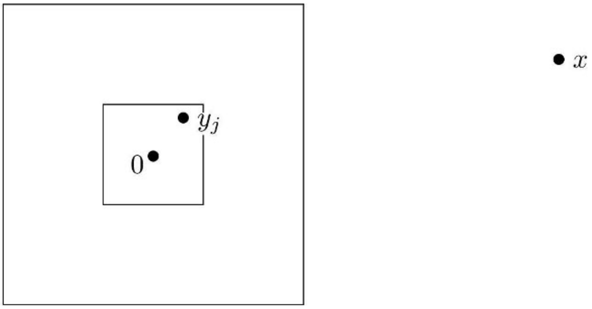


Fig. 14. Far field representation.

where $p \quad O(\log(1/\varepsilon))$ because $|y_j/x| < \sqrt{2}/3$. We define the far field representation to be the coefficients $\{a_k, 0 \leq k \leq p\}$ given by

$$a_0 \quad \sum_j f_j \quad \text{and} \quad a_k \quad -\frac{1}{k}\sum_j y_j^k f_j \quad (1 \leq k \leq p). \tag{22}$$

It is obvious that from $\{a_k\}$ we can approximate the potential for any point x in the far field efficiently within accuracy $O(\varepsilon)$. This representation clearly has complexity $O(\log(1/\varepsilon))$ and is also named the *multipole expansion*.

Local field representation

Suppose that $\{y_j\}$ are source points in the far field of a box (see Fig. 15) and $\{f_j\}$ are charges located at $\{y_j\}$.

From the Taylor expansion of the kernel

$$G(x,y) \quad \ln(x-y) \quad \ln(-y) + \ln\left(1 - \frac{x}{y}\right) \quad \ln(-y) + \sum_{k \quad 1}^{\infty}\left(-\frac{1}{k}\right)\frac{x^k}{y^k},$$

we have, for any x inside the box,

$$u(x) \quad \sum_j G(x,y_j)f_j \quad \sum_j \ln(-y_j)f_j + \sum_{k \quad 1}^{p}\left(-\frac{1}{k}\sum_j \frac{f_j}{y_j^k}\right)x^k + O(\varepsilon)$$

where $p \quad O(\log(1/\varepsilon))$ because $|x/y_j| < \sqrt{2}/3$. We define the local field representation to be the coefficient $\{a_k, 0 \leq k \leq p\}$ given by

$$a_0 \quad \sum_j \ln(-y_j)f_j \quad \text{and} \quad a_k \quad -\frac{1}{k}\sum_j \frac{f_j}{y_j^k} \quad (1 \leq k \leq p). \tag{23}$$

Based on $\{a_k\}$, we can approximate the potential for any point x inside the box efficiently within accuracy $O(\varepsilon)$. This representation has complexity $O(\log(1/\varepsilon))$ and is also named the *local expansion*.

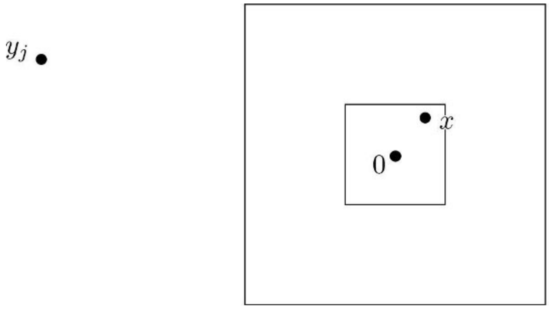


Fig. 15. Local field representation.

Far-to-far translation

Let us now consider the far-to-far translation which transforms the far field representation of a child box B' to the far field representation of its parent box B (see Fig. 16). We assume that B' is centered at a point z_0 while B is centered the origin. Suppose that the far field representation of child B' is $\{a_k, 0 \leq k \leq p\}$, i.e.,

$$u(z) \quad a_0 \ln(z - z_0) + \sum_{k\ \ 1}^{p} a_k \frac{1}{(z - z_0)^k} + O(\varepsilon)$$

for any z in the far field of B'. The far field representation $\{b_l, 0 \leq l \leq p\}$ of B is given by

$$b_0 \quad a_0 \quad \text{and} \quad b_l \quad -\frac{a_0 z_0^l}{l} + \sum_{k\ \ 1}^{l} a_k z_0^{l-k} \binom{l-1}{k-1} \quad (1 \leq l \leq p) \tag{24}$$

and for any z in the far field of B

$$u(z) \quad b_0 \ln z + \sum_{l\ \ 1}^{p} b_l \frac{1}{z^l} + O(\varepsilon).$$

From the definition of $\{b_l\}$, it is clear that each far-to-far translation takes $O(p^2)$ $O(\log^2(1/\varepsilon))$ operations.

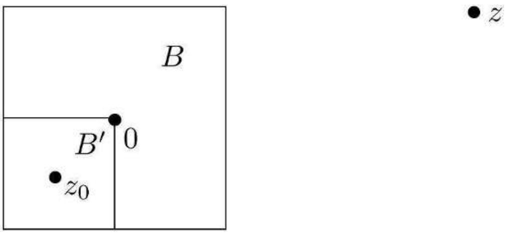


Fig. 16. Far-to-far translation.

Far-to-local translation

The far-to-local translation transforms the far field representation of a box B to the local field representation of a box A in B's interaction list. We assume that B is centered at z_0 while A is centered at the origin (see Fig. 17). Suppose that the far field representation of B is $\{a_k, 0 \le k \le p\}$, i.e.,

Fig. 17. Far-to-local translation.

$$u(z) \quad a_0 \ln(z - z_0) + \sum_{k\ \ 1}^{p} a_k \frac{1}{(z-z_0)^k} + O(\varepsilon)$$

for any z in the far field of B. The local field representation $\{b_l, 0 \le l \le p\}$ of A is given by

$$b_0 \quad a_0 \ln(-z_0) + \textstyle\sum_{k\ \ 1}^{p} (-1)^k \frac{a_k}{z_0^k} \quad \text{and}$$
$$b_l \quad -\frac{a_0}{l z_0^l} + \frac{1}{z_0^l} \textstyle\sum_{k\ \ 1}^{p} \frac{a_k}{z_0^k} \binom{l+k-1}{k-1} (-1)^k \quad (1 \le l \le p).$$

and for any z in A

$$u(z) \quad \sum_{l\ \ 0}^{p} b_l z^l + O(\varepsilon).$$

It is clear that each far-to-local translation takes $O(p^2) \quad O(\log^2(1/\varepsilon))$ operations as well.

Local-to-local translation

The local-to-local translation transforms the local field representation of a parent box A to the local field representation of its child A'. We assume that the center of A is z_0 while the center of A' is the origin (see Fig. 18). Suppose that the local field representation of A is $\{a_k, 0 \le k \le p\}$, i.e.,

$$u(z) \quad \sum_{k\ \ 0}^{p} a_k (z - z_0)^k + O(\varepsilon)$$

for any z in A. Then the local field representation $\{b_l, 0 \le l \le p\}$ at A' is given by

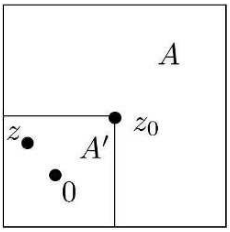


Fig. 18. Local-to-local translation.

$$b_l \quad \sum_{k\ \ l}^{n} a_k \binom{k}{l} (-z_0)^{k-l} \quad (0 \le l \le p)$$

and for any z in A'

$$u(z) \quad \sum_{l\ \ 0}^{p} b_l z^l + O(\varepsilon).$$

The complexity of a local-to-local translation is again $O(p^2) \quad O(\log^2(1/\varepsilon))$. To summarize the 2D case, both the far and local field representations are of size $O(p)$ $O(\log(1/\varepsilon))$ for a prescribed accuracy ε. All three translations are of complexity $O(p^2) \quad O(\log^2(1/\varepsilon))$. Therefore, the complexity of the FMM algorithm based on these representations and translations is $O(N)$ where the constant depends on ε in a logarithmic way.

3D case

Up to a constant, the 3D Green's function of the Laplace equation is

$$G(x,y) \quad \frac{1}{|x-y|}.$$

For two points $x \quad (r,\theta,\varphi)$ and $x' \quad (r',\theta',\varphi')$ in spherical coordinates, we have an important identity

$$\frac{1}{|x-x'|} \quad \sum_{n\ \ 0}^{\infty} \sum_{m\ \ -n}^{n} (r')^n Y_n^{-m}(\theta',\varphi') \frac{1}{r^{n+1}} Y_n^m(\theta,\varphi)$$

for $r \ge r'$.

Far field representation

Suppose that $\{y_j \quad (r_j,\theta_j,\varphi_j)\}$ are source points with charges $\{f_j\}$ inside a box centered at the origin. Let us consider the potential generated by $\{y_j\}$ at a point $x \quad (r,\theta,\varphi)$ in the far field (see Fig. 14). Using the given identity, we get

$$u(x) \quad \sum_j G(x,y_j) f_j \quad \sum_{n\ 0}^{p} \sum_{m\ -n}^{n} \left(\sum_j f_j r_j^n Y_n^{-m}(\theta_j,\varphi_j) \right) \frac{1}{r^{n+1}} Y_n^m(\theta,\varphi) + O(\varepsilon)$$

where $p \quad \log(1/\varepsilon)$ because $|y_j/x| < \sqrt{3}/3$. We define the far field representation to be the coefficients $\{\alpha_n^m, 0 \le n \le p, -n \le m \le n\}$ given by

$$\alpha_n^m \quad \sum_j f_j r_j^n Y_n^{-m}(s_j).$$

From these coefficients $\{\alpha_n^m\}$, one can approximate $u(x)$ for any x in the far field efficiently.

Local field representation

Suppose that $\{y_j \quad (r_j,\theta_j,\varphi_j)\}$ are source points with charges $\{f_j\}$ in the far field of a box. Let us consider the potential generated by $\{y_j\}$ at a point x inside the box. We assume that the box is centered at the origin (see Fig. 15). Following the above identity, we have at x

$$u(x) \quad \sum_j G(x,y_j) \quad \sum_{n\ 0}^{p} \sum_{m\ -n}^{n} \left(\sum_j f_j \frac{1}{r_k^{n+1}} Y_n^m(\theta_j,\varphi_j) \right) r^n Y_n^m(\theta,\varphi) + O(\varepsilon)$$

where $p \quad \log(1/\varepsilon)$ because $|x/y_j| < \sqrt{3}/3$. We define the local field representation to be the coefficients $\{\beta_n^m, 0 \le n \le p, -n \le m \le n\}$ given by

$$\beta_n^m \quad \sum_j f_j \frac{1}{r_k^{n+1}} Y_n^m(s_j).$$

It is clear that, from these coefficients $\{\beta_n^m\}$, one can approximate $u(x)$ for any x inside the box efficiently.

Far-to-far, far-to-local and local-to-local translations

Similar to the 2D case, we have explicit formulas for the three translations. The derivation of these formulas depend heavily on special function theories. We point to [16] for the details.

Since both the far field and local field representations have $O(p^2)$ coefficients, a naive implementation of these translations requires $O(p^4)$ operations, which is quite large even for moderate values of p. If we take a look at the FMM closely, we discover that the most time-consuming step is to perform the far-to-local translations. This is due to the fact that for each box B there can be as many as $6^3 - 3^3 \quad 189$ boxes in its interaction list. For each of these boxes, a far-to-local translation is required. Therefore, computing the far-to-local translations with a much lower complexity is imperative for the success of a 3D FMM implementation.

In [9], Cheng et al. introduce highly efficient ways for computing these translations. For the far-to-far and local-to-local translations, a "point and shoot" method is used to reduce the complexity from $O(p^4)$ to $O(p^3)$. Let us consider for example the far-to-far translation between a child box B' and its parent B. The main idea is that if the z axes of the spherical coordinate systems at B' and B coincided, the transformation from the far field representation of B' to the ones of B would be computed in $O(p^3)$ steps. Therefore, the far-to-far translation is partitioned into three steps:

- "Rotate" the coordinate system at B' so that the z axis points to the center of B. The far field representation at B' is transformed accordingly. This step takes $O(p^3)$ operations.
- Perform the far-to-far translation from B' to B in the rotated coordinate system. This step takes $O(p^3)$ operation as well.
- Finally, "rotate" the coordinate system at B back to the original configuration and transform the far field representation at B accordingly. This step takes $O(p^3)$ operations as well.

For the far-to-local translation, the main idea is to use the plane wave (exponential) expansion, which diagonalizes the far-to-local translation. Given two well-separated boxes A and B, the steps are

- Transform the far field representation to six plane wave expansions, one for each of the six directions $\pm x, \pm y, \pm z$. This step has $O(p^3)$ complexity.
- Depending on the location of A, use one of the six plane wave expansions to compute the far-to-local translation from B to A. After this step, the local field representation at A is stored in the plane wave form. Since the plane wave expansion diagonalizes the far-to-local translation, the complexity of this step is $O(p^2)$.
- Transform the plane wave expansions at A back to the local field representation. Notice that at A there are also six plane wave expansions for six different directions. This step takes $O(p^3)$ operations as well.

Since the first step is independent of the target box A, one only needs to perform it once for each box B. The same is true for the last step as it is independent of the source box B. On the other hand, the second step, which can be called as many as 189 times for each box, is relatively cheap as its complexity is $O(p^2)$.

4 Kernel Independent Fast Multipole Method

The FMM introduced in the previous section is highly efficient yet quite technical. As we have seen, both the representations and translations in the 3D case depend heavily on the results from special functions and their derivations are far from trivial. The Laplace equation is only one of the elliptic PDEs with non-oscillatory kernels: other examples include the Stokes equations, the Navier equation, the Yukawa equation and so on. Deriving expansions and translations for the kernels of these equations one by one can be a tedious task. In this section, we introduce the kernel independent fast multipole method which addresses all these kernels in a unified framework [27]. Some of the ideas in this framework appeared earlier in [1, 4].

The geometric part of the kernel independent fast multipole method is exactly the same as the standard FMM. Hence, our discussion focuses only on the analytic part. We will start with the 2D case and then comment on the difference for the 3D case.

4.1 2D case

Far field representation

Let us consider a simple physics experiment first (see Fig. 19). Suppose that B is a box with radius r and that we have a set of charges $\{f_j\}$ at $\{y_j\}$ inside B. These charges generate a non-zero potential in the far field. Let us now put a metal circle of radius $\sqrt{2}r$ around these charges and connect this metal circle to the ground. As a result, a charge distribution would appear on this metal circle to cancel out the potential field generated by the charges inside the box. Due to the linearity of the problem, we see that the potential field due to the charges inside the box can be reproduced by the charge distribution on the circle if we flip its sign. This experiment shows that the volume charges inside the box can be replaced with an equivalent surface charge distribution on the circle if one is only interested in the potential in the far field.

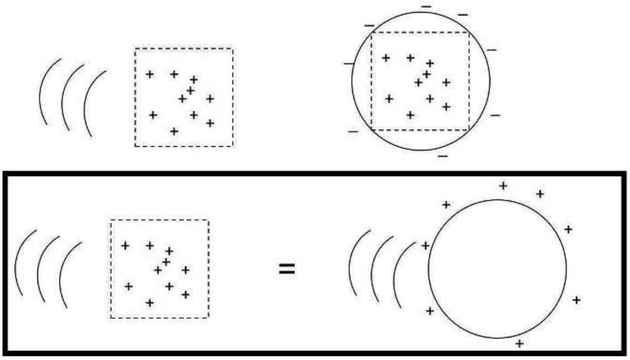


Fig. 19. The existence of an equivalent charge distribution.

A natural question to ask is, given a prescribed accuracy ε, how many degrees of freedom one needs to describe the equivalent charge distribution. Let us recall that the far field representation is only needed for the far field. It is well-known that the potential generated by the high frequency modes of the charge distribution on the circle dies out very quickly in the far field: it decays like $(\sqrt{2}/3)^n$ for the nth mode. As a result, we only need to capture the low frequency modes of the charge distribution. Our solution is to place $O(\log 1/\varepsilon)$ equally spaced points $\{y_k^{B,F}\}_k$ on the circle. The equivalent charges $\{f_k^{B,F}\}_k$ supported at these points are used as the far field representation (see Fig. 20).

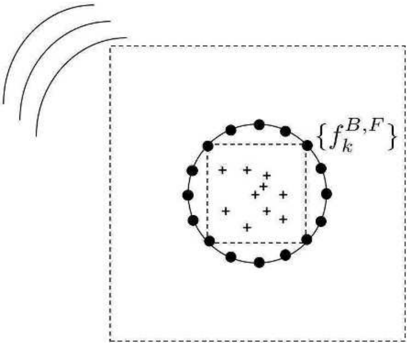


Fig. 20. The equivalent charges $\{f_k^{B,F}\}_k$ of the box B.

The next question is how to construct the equivalent charges $\{f_k^{B,F}\}_k$. One of the solutions is to pick a large circle of radius $(4-\sqrt{2})r$, the exterior of which contains the far field of B. If the potential fields generated by the source charges and $\{f_k^{B,F}\}_k$ are identical on this circle, then they have to match in the far field as well due to the uniqueness of the exterior problem of the Laplace equation. Based on this observation, the procedure of constructing $\{f_k^{B,F}\}_k$ consists of two steps (see Fig. 21).

- Pick $O(\log(1/\varepsilon))$ equally spaced locations $\{x_k^{B,F}\}_k$ on the large circle. Use kernel evaluation to compute the potentials $\{u_k^{B,F}\}_k$ at these locations generated by the charges inside B.
- Invert the interaction matrix between $\{y_k^{B,F}\}_k$ and $\{x_k^{B,F}\}_k$ to find $\{f_k^{B,F}\}_k$ so that they generate the potentials $\{u_k^{B,F}\}_k$. This inversion problem might be ill-posed, so one might need to regularize it with Tikhonov regularization [20].

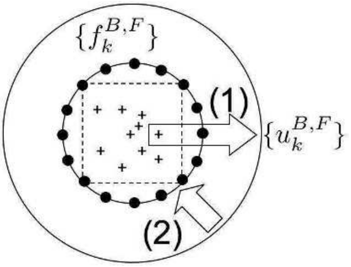


Fig. 21. The construction of the equivalent charges $\{f_k^{B,F}\}_k$.

Local field representation

Suppose that A is a box with radius r and that we have a set of charges $\{f_j\}$ located at points $\{y_j\}$ in the far field of A. To represent the potential field generated by these charges inside A, we first put a circle of radius $\sqrt{2}r$ around A (see the following figure). Let us call the potential on the circle the check potential field. From the uniqueness property of the interior problem of the Laplace equation, we know that, if we are able to capture the check potential field, we then can construct the potential everywhere in the box.

Similar to the case of the equivalent charge distribution, the next question is how many degrees of freedom we need to represent the check potential field. Since the potential is generated by points in the far field, it is quite smooth on the circle as the high frequency modes die out very quickly. Therefore, we only need a few samples to capture the check potential field. We put $O(\log(1/\varepsilon))$ samples $\{x_k^{A,L}\}_k$ on the circle. The potentials $\{u_k^{A,L}\}_k$ at these locations are taken to be the local field representation.

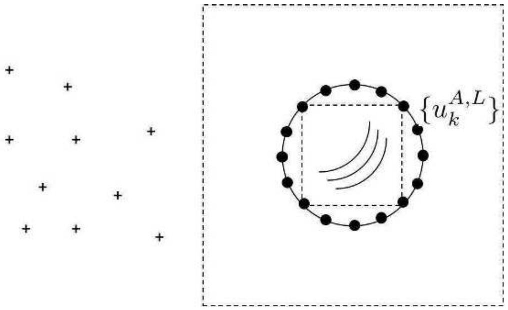


Fig. 22. The check potentials $\{u_k^{A,L}\}_k$ of the box B.

In order to reconstruct the potential inside the box A from the check potentials $\{u_k^{A,L}\}_k$, we first take a look at the example in Fig. 23. As before, the charges in the far field of A produce a potential field inside A. Let us now put a large metal circle of radius $(4-\sqrt{2})r$ around the box A and connect it to the ground. As a result, a charge distribution will appear on the large circle to cancel the potential field generated by the far field charges inside A. Again due to the linearity of the problem, we conclude that the potential due to the charges in the far field can be reproduced by the charge distribution on the large circle if we flip the sign of the surface charge distribution. This experiment shows that, if one can find the appropriate surface charge distribution on the large circle, the potential inside the box A can then be reconstructed.

Motivated by this example, we propose the following procedure to compute the potential inside A given the check potentials $\{u_k^{A,L}\}_k$ (see Fig. 24).

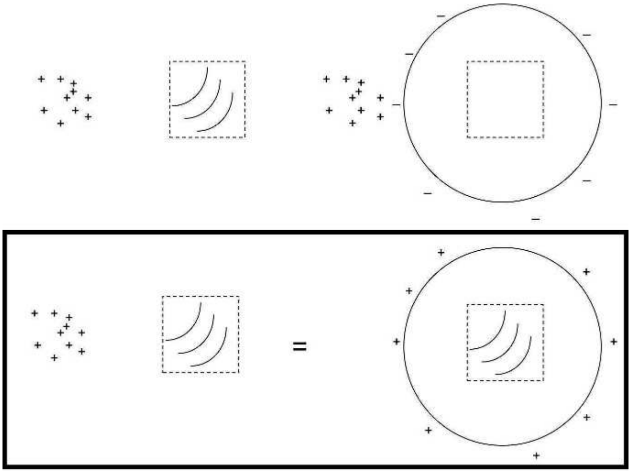


Fig. 23. The existence of equivalent charges for the local field representation.

- Pick $O(\log(1/\varepsilon))$ points $\{y_k^{A,L}\}_k$ on the large ring. Invert the interaction matrix between $\{y_k^{A,L}\}_k$ and $\{x_k^{A,L}\}_k$ to find the charges $\{f_k^{A,L}\}_k$ that produce $\{u^{A,L}\}_k$. This inversion might be ill-posed, so one might need to regularize it with Tikhonov regularization.
- Use the kernel evaluation to compute the potential inside A using the charges $\{f_k^{A,L}\}_k$.

To summarize, we have used the equivalent charges as the far field representation and the check potentials as the local field representation. Now let us consider the three translations of the kernel independent FMM.

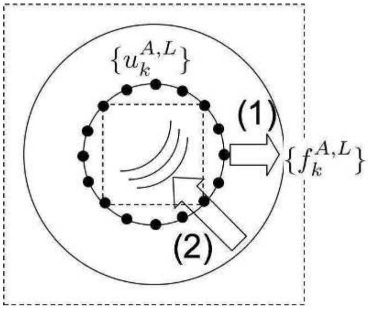


Fig. 24. The evaluation of the local field from the check potentials $\{u_k^{A,L}\}_k$.

Far-to-far translation

Given the equivalent charges of a child box B', the far-to-far translation computes the equivalent charges of the parent box B. The situation is similar to the construction of the equivalent charges that is described before if one is willing to consider the equivalent charges of B' as the source charges inside B. The steps of this translation are:

- Use the equivalent charges $\{f_k^{B',F}\}_k$ as source charges to evaluate the potential $\{u_k^{B,F}\}_k$ at $\{x_k^{B,F}\}_k$ (see Fig. 25).
- Invert the interaction between $\{y_k^{B,F}\}_k$ and $\{x_k^{B,F}\}_k$ to find the equivalent charges $\{f_k^{B,F}\}_k$. This step might again be ill-posed, so Tikhonov regularization might be needed.

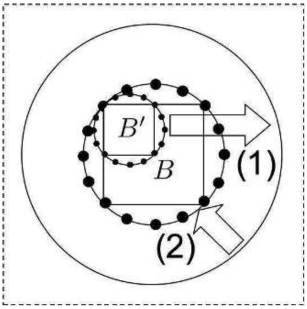


Fig. 25. Far-to-far translation.

Far-to-local translation

Given the equivalent charges of a box B, the far-to-local translation transforms them to the check potentials of a box A in B's interaction list (see the following figure). This translation is particularly simple for the kernel independent FMM. It consists of only a single step:

- Evaluate the potential $\{u_k^{A,L}\}_k$ using the equivalent charges $\{f_k^{B,F}\}_k$ (see Fig. 26).

Local-to-local translation

Given the check potentials of a parent box A, the local-to-local translation transforms them to the check potentials of its child box A'. The steps of the local-to-local translation are:

- Invert the interaction between $\{y_k^{A,L}\}_k$ and $\{x_k^{A,L}\}_k$ to find the equivalent charges $\{f_k^{A,L}\}_k$ that produce the check potentials $\{u_k^{A,L}\}_k$ (see Fig. 27). Tikhonov regularization is invoked whenever necessary.

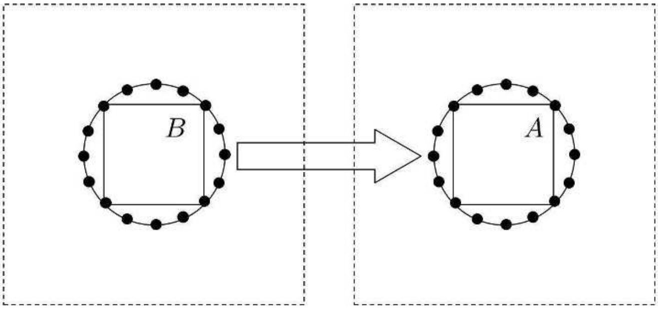


Fig. 26. Far-to-local translation.

- Check potentials $\{u_k^{A',L}\}_k$ are then computed using kernel evaluation with $\{f_k^{A,L}\}_k$ as the source charges.

Since the matrices used in the far-to-far and local-to-local translations only depend on the size of the boxes, their inversions can be precomputed and stored. Therefore, the kernel independent FMM algorithm only uses matrix vector multiplications and kernel evaluations. This general framework works well not only for PDE kernels such as the Green's functions of the Laplace equation, the Stokes equations, the Navier equation and Yukawa equation, but also for various radial basis functions after a slight modification.

4.2 3D case

In 3D, we need $O(p^2)$ $O(\log^2 1/\varepsilon)$ points to represent the equivalent charge distribution and the check potential field. If we put these points on a sphere, the three translations would require $O(p^4)$ operations. This poses the same problem we faced in the discussion of the 3D FMM algorithm.

In order to reduce this complexity, we choose to replace the sphere with the boundary of a box. This box is further discretized with a Cartesian grid and both

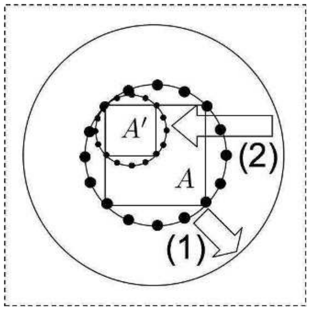


Fig. 27. Local-to-local translation.

the equivalent charges and the check potentials are located at the boundary points of the Cartesian grid. The main advantage of choosing the Cartesian grid is that the far-to-local translation, which is the most frequently used step, becomes a discrete convolution operator since the Green's function of the underlying PDE is translation invariant. This discrete convolution can be accelerated using the standard FFT techniques, and the resulting complexity of these translation operators are reduced to $O(p^3 \log p)$.

5 Hierarchical Matrices

Let us recall the computational problem that we face in each step of the iterative solution. Given a set of charges $\{f_i, 1 \le i \le N\}$ located at points $\{p_i, 1 \le i \le N\}$ (see Fig. 28) and the Green's function $G(x,y)$ of the Laplace equation, we want to compute at each p_i the potential

$$u_i \quad \sum_{j \quad 1}^{N} G(p_i, p_j) f_j.$$

From the discussion above, we know that, if two sets A and B are well-separated, the interaction interaction $G(x,y)$ for $x \in A$ and $y \in B$ is approximately low rank. The hierarchical matrix framework puts this observation into an algebraic form. Our presentation in this section is far from complete, and we refer to [6] for a comprehensive treatment.

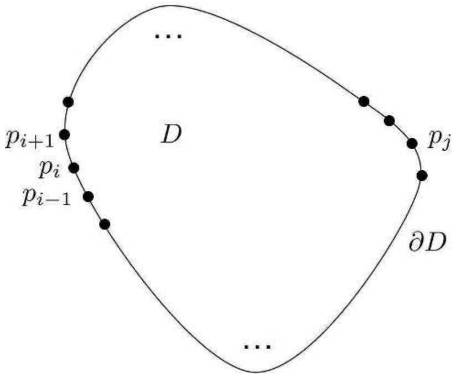


Fig. 28. Distribution of quadrature points $\{p_i\}$ on the boundary of the domain D.

5.1 Construction

Let us consider the following simple example where the domain D is a 2D disk. The boundary ∂D is subdivided into a hierarchical structure such that each internal node has two children and each leaf contains $O(1)$ points in $\{p_i\}$.

At the beginning level, ∂D is partitioned into with 4 large segments (see the following figure). Some pairs (e.g., A and B) are well-separated. Suppose the points $\{p_i\}$ are ordered according to their positions on the circle. As a result, the interaction from B to A corresponds to a subblock of the full interaction matrix $\boldsymbol{G}$ $(G(p_i,p_j))_{1\le i,j\le N}$. Since the interaction between B and A is approximately low rank, this subblock can be represented in a low rank compressed form. All the subblocks on this level which have low rank compressed forms are colored in gray (see Fig. 29).

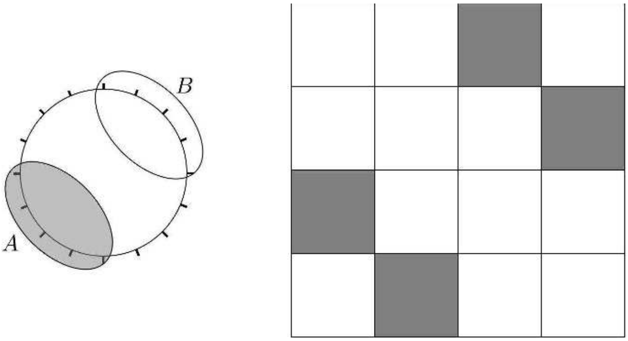


Fig. 29. Left: two well-separated parts on the second level of the hierarchical partition. In the matrix form, the interaction between them corresponds to a off-diagonal subblock. Right: all the blocks that correspond to well-separated parts on this level.

In order to consider the interaction between B and its neighbors, we go down to the next level. Suppose B' is a child of B. Similar to the case of the FMM, the interaction between B' and B's far field has already been taken care of in the previous level. We now need to consider the interaction between B' and the segments that are in the far field of B' but not the far field of B. There are only $O(1)$ segments in this region (colored in gray as well) and A' is one of them. As B' and A' are now well-separated, the interaction from B' to A' is approximately low rank. Therefore, the subblock that corresponds to this interaction can be stored in a low rank compressed form. All the subblocks on this level which have low rank compressed forms are again colored in gray (see Fig. 30).

We go down one level further to address the interaction between B' and its neighbors. For the same reason, B'' (a child of B') has only $O(1)$ segments in its far field but not in B''s far field. The subblocks that correspond to the interaction between B'' and these segments can be stored in low rank compressed forms (see Fig. 31). Suppose now that B'' is also a leaf segment. Since the interaction between B'' and its adjacent segments are not necessarily low rank, the subblocks corresponding to these interactions are stored densely. Noticing that each leaf segment contains only $O(1)$ points, the part of $\boldsymbol{G}$ that requires dense storage is $O(N)$.

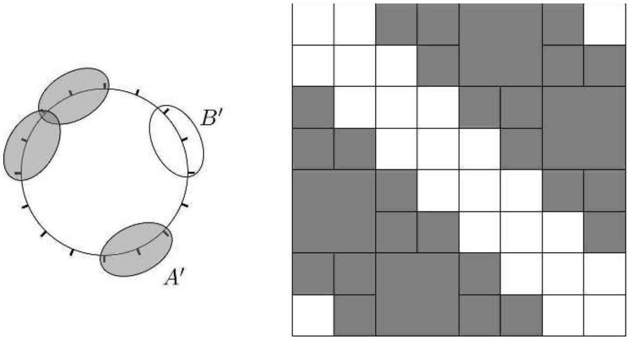


Fig. 30. At the third level.

From this simple example, we see that the hierarchical matrix framework is a way to partition the full interaction matrix into subblocks based on a hierarchical subdivision of the points. The off-diagonal blocks of a hierarchical matrix are compressed in low rank forms, while the diagonal and the next-to-diagonal blocks are stored densely.

A natural question at this point is which low rank compressed form one should use to represent the off-diagonal blocks. A first answer is to construct of these off-diagonal blocks first and then perform the truncated singular value decomposition (SVD) to compress them. The resulting form gives the best compression for a prescribed accuracy ε as the singular value decomposition is optimal in compressing matrices. However, there are two major disadvantages. First, the SVD usually requires one to construct the off-diagonal blocks first, which costs at least $O(N^2)$ operations. Second, since the singular vectors resulted from the SVD are not directly related to the vectors of the subblocks of $\boldsymbol{G}$, storing these vectors requires a lot of memory space. To overcome these two problems, we resort to several other methods.

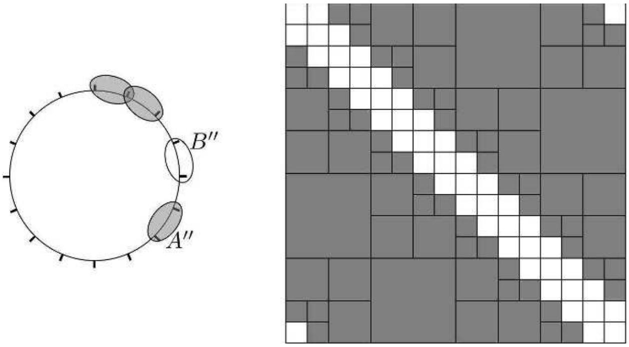


Fig. 31. At the fourth level.

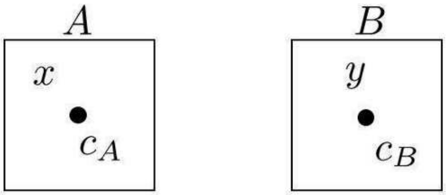


Fig. 32. Taylor expansion approach for constructing the low rank representations between two separated parts A and B.

Taylor expansion

Suppose that c_A and c_B are to be the center of segments A and B respectively (see Fig. 32). From the truncated Taylor expansion, we have

$$G(x,y) \quad \sum_{|\alpha|<p} \frac{1}{\alpha!} \partial_x^{\alpha} G(c_A,y)(x-c_A)^{\alpha} + O(\varepsilon) \tag{25}$$

where α is the multi-index and $p \quad O(\log(1/\varepsilon))$. In this derivation, we used the facts that

$$\partial_{\alpha} G(c_A,y) \approx \frac{1}{|y-c_A|^{|\alpha|}} \quad \text{and} \quad \frac{|x-c_A|}{|y-c_A|} \leq \sqrt{2}/3.$$

This factorization provides us with a compressed form of rank $O(p^d)$.

There are two disadvantages of this approach. First, for complicated kernels, $\partial_x^{\alpha} G(x,y)$ is not easy to obtain. Even for the fundamental solution of the Laplace equation, this is far from trivial for large α. Second, Taylor expansion does not exploit the special structure of the kernel. Therefore, the resulting expansion has $O(p^d)$ terms where $d \quad 2,3$ is the dimension of the problem. This is quite wasteful comparing to the $O(p^{d-1})$ coefficients used in the FMM.

Tensor-product interpolation

In this approach, we pick a Cartesian grid to cover one of the domain (say A). The Cartesian grid is a tensor product of d one dimensional grids, each of which contains pth order Chebyshev points on an closed interval where $p \quad O(\log(1/\varepsilon))$. Suppose that $\{a_i\}$ are these grid points. Since $G(x,y)$ is smooth for $x \in A$ and $y \in B$ when A and B are well-separated, we have

$$G(x,y) \quad \sum_i G(a_i,y) L_i^A(x) + O(\varepsilon) \tag{26}$$

where $\{L_i^A(x)\}$ are d-dimensional Lagrange interpolants of the grid $\{a_i\}$ over A. Similarly, we pick a grid to cover B instead of A. Let $\{b_j\}$ be the grid points. For the same reason, we have

$$G(x,y) \quad \sum_j G(x,b_j) L_j^B(y) + O(\varepsilon) \tag{27}$$

where $\{L_j^B(y)\}$ are d-dimensional Lagrange interpolants of the grid $\{b_j\}$ over B. One can also choose to cover both A and B with Cartesian grids. In this case, we have

$$G(x,y) \quad \sum_i \sum_j L_i^A(x) G(a_i,b_j) L_j^B(y) + O(\varepsilon). \tag{28}$$

This tensor-product interpolation approach (illustrated in Fig. 33) is quite general since it only utilizes the kernel evaluation. However, similar to the Taylor expansion approach, it uses $O(p^d)$ terms, which is more than necessary.

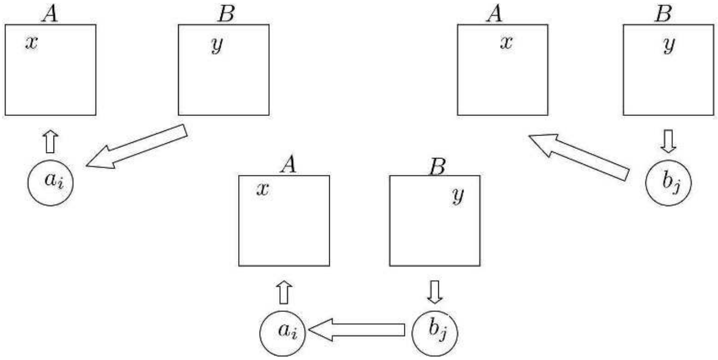


Fig. 33. Tensor-product interpolation approach for constructing the low rank representations between two separated parts A and B.

Pseudo-skeleton or cross approximation

Suppose $\{x_i, i \in I\}$ and $\{y_j, j \in J\}$ to be the point sets in A and B respectively. In our setting, they are subsets of $\{p_i\}$. We will use $\boldsymbol{G}_{I,J}$ to denote the subblock of $\boldsymbol{G}$ that corresponds to the interaction from B to A. Since the matrix $\boldsymbol{G}_{I,J}$ is approximately low rank, there exist a few columns of $\boldsymbol{G}_{I,J}$ which span its column space. Similarly, there exist a few rows of G which span its row space as well. The idea behind pseudo-skeleton approximation (or cross approximation) is to find these columns and rows and use them in the compressed representation. Suppose these columns correspond to the points $\{y_j, j \in L \subset J\}$ while these rows correspond to the points $\{x_i, i \in K \subset I\}$. The pseudo-skeleton approximation [14] is a factorization of the matrix $\boldsymbol{G}_{I,J}$ in the following form:

$$\boldsymbol{G}_{I,J} \quad \boldsymbol{G}_{I,L} \boldsymbol{M} \boldsymbol{G}_{K,J} + O(\varepsilon) \tag{29}$$

where $\boldsymbol{G}_{I,L}$ is the submatrix of $\boldsymbol{G}_{I,J}$ that contains the columns of points $\{y_j, j \in L\}$ while $\boldsymbol{G}_{K,J}$ is the submatrix of $\boldsymbol{G}_{I,J}$ that contains the rows of points $\{x_i, i \in K\}$ (see Fig. 34).

Several methods have been proposed to construct such a pseudo-skeleton approximation for $\boldsymbol{G}_{I,J}$. Approaches for selecting the sets $\{x_i, i \in K\}$ and $\{y_j, j \in L\}$ include greedy methods, adaptive methods, and random sampling techniques (see [6] for details). The middle matrix $\boldsymbol{M}$ is often computed using least square techniques.

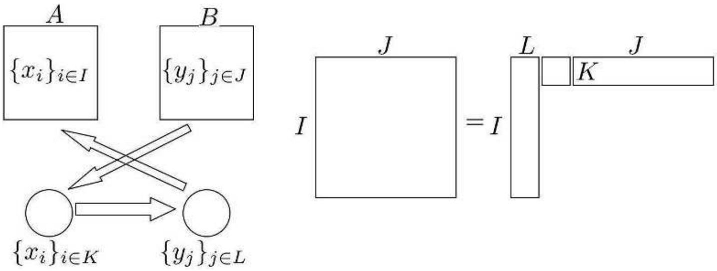


Fig. 34. Pseudo-skeleton approach for constructing the low rank representations between two separated parts A and B.

5.2 Hierarchical matrix arithmetics

Since the hierarchical matrix framework is an algebraic approach, it is possible to define a matrix arithmetics (addition, multiplication and inversion) for the hierarchical matrices. Here, we discuss these operations briefly.

Addition

Given two hierarchical matrices A and B with the same hierarchical structure, we seek a hierarchical matrix C such that $C \approx A+B$. Since both A and B have the same structure, we can perform the addition block by block. Suppose that P, Q and R are the blocks of A, B and C at the same location. There are two cases to consider. In the first case, P and Q are both dense blocks. Then R is simply the sum of P and Q. In the second case, both P and Q are stored in the compressed form. Let us further assume that $P \quad P_1P_2^t$ and $Q \quad Q_1Q_2^t$ where P_1, P_2, Q_1 and Q_2 are tall matrices. Then we have

$$R \quad \left(P_1\ Q_1\right)\left(P_2\ Q_2\right)^t.$$

The matrices $\left(P_1\ Q_1\right)$ and $\left(P_2\ Q_2\right)^t$ are further compressed using the pivoted QR factorization and the SVD.

Multiplication

Given two hierarchical matrices A and B with the same hierarchical structure, we seek a hierarchical matrix C such that $C \approx AB$. The multiplication algorithm for hierarchical matrices is similar to the one for block matrices. The basic step is to multiply two block matrices P and Q. There are four different cases to consider. In the first case, both P and Q are stored in the low rank compressed form, say $P \quad P_1P_2^t$ and $Q \quad Q_1Q_2^t$. Then

$$PQ \quad P_1(P_2^tQ_1)Q_2^t.$$

In the second case, P is in the low rank form $P \quad P_1P_2^t$ while Q is still in a hierarchical form. Without loss of generality, we assume

$$Q = \begin{pmatrix} Q_1 & Q_2 \\ Q_3 & Q_4 \end{pmatrix}.$$

By splitting P_2^t into

$$P_2^t = \begin{pmatrix} P_{2,1}^t & P_{2,2}^t \end{pmatrix},$$

we can compute PQ as

$$PQ = P_1 \left(\begin{pmatrix} P_{2,1}^t & P_{2,2}^t \end{pmatrix} \begin{pmatrix} Q_1 & Q_2 \\ Q_3 & Q_4 \end{pmatrix} \right).$$

where the multiplication in the parentheses is carried out recursively.

In the third case, P is in the hierarchical form while Q is in the low rank form $Q = Q_1 Q_2^t$. Let us assume that

$$P = \begin{pmatrix} P_1 & P_2 \\ P_3 & P_4 \end{pmatrix}.$$

By splitting Q_1 into

$$Q_1 = \begin{pmatrix} Q_{1,1} \\ Q_{1,2} \end{pmatrix},$$

we can compute PQ as

$$PQ = \left(\begin{pmatrix} P_1 & P_2 \\ P_3 & P_4 \end{pmatrix} \begin{pmatrix} Q_{1,1} \\ Q_{1,2} \end{pmatrix} \right) Q_2^t.$$

where the multiplication in the parentheses is carried out recursively.

In the last case, both P and Q is in the hierarchical form. We then resort to the block matrix multiplication algorithm and reduce its multiplication to the first three cases.

Inversion

Given a hierarchical matrix A, we want to compute a hierarchical matrix C such that $C \approx A^{-1}$. The solution is simply to apply the block matrix version of the LU factorization. Regular matrix addition and multiplication operations are now replaced with the ones of the hierarchical matrices described above.

6 Wavelet Based Methods

In this section, we consider yet another approach to the same problem discussed in the previous sections. Given a set of charges $\{f_j, 1 \le i \le N\}$ located at points $\{p_i, 1 \le i \le N\}$ (see the following figure) and the Green's function $G(x,y)$ of the Laplace equation, we want to compute at each p_i

$$u_i = \sum_{j=1}^{N} G(p_i, p_j) f_j.$$

Our discussion in this section focuses on the 2D case. Suppose that the boundary ∂D is parameterized by a periodic function $g(s)$ for $s \in \;0,1\;$. The matrix $\boldsymbol{G}$ with entries $G(p_i, p_j)$ can be viewed as an adequately sampled image of the continuous periodic 2D function $G(g(s), g(t))$. This image is smooth except at the diagonal where $s \quad t$ (see Fig. 35).

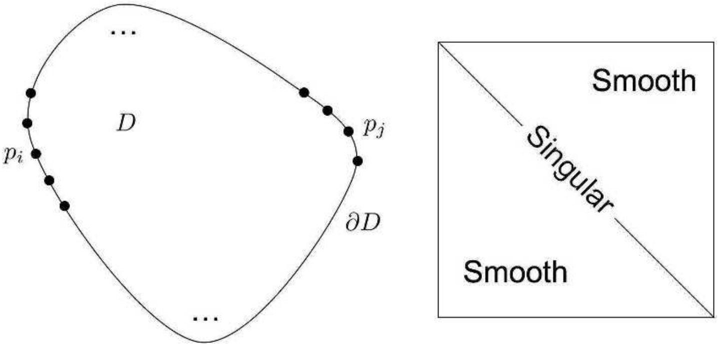


Fig. 35. Left: distribution of quadrature points $\{p_i\}$ on the boundary of the domain D. Right: a qualitative description of the 2D function $G(g(s), g(t))$.

Our plan is to find the best way to compress this matrix and then see whether the compression can help our computational problem. The presentation of this section follows [3].

6.1 Wavelet compression

One good way to compress such an image is to use 2D wavelets. Let us start our discussion with 1D wavelets on a unit periodic interval. The following discussion about the wavelets are quite brief and we refer the readers to [12, 21] for detailed exposition. Suppose that j is the index of the level and k is the spatial index. The scaling functions of the wavelet analysis are scaled and shifted copies of a mother scaling function $\varphi(x)$:

$$\{\varphi_{j,k}(x) : \quad 2^{-j/2}\varphi(2^{-j}x - k)\}_{-\infty < j \le 0, 0 \le k < 2^{-j}}$$

Similarly, the wavelet functions are scaled and shifted versions of a mother wavelet function $\psi(x)$:

$$\{\psi_{j,k}(x) : \quad 2^{-j/2}\psi(2^{-j}x - k)\}_{-\infty < j \le 0, 0 \le k < 2^{-j}}.$$

Let us assume that our wavelets are orthogonal and compactly supported. Therefore, a wavelet or a scaling function on the jth scale has a support of size $O(2^j)$. Due to the orthogonality condition, the scaling function at the 0th level $\varphi_{0,0}$ and the wavelets

$\{\psi_{j,k}\}_{-\infty<j\leq 0,0\leq k<2^{-j}}$ form an orthogonal basis of $L^2(\ 0,1\ )$. We further assume that our wavelets have M vanishing moments, i.e.,

$$\int \psi(x)x^m dx \quad 0, \quad m \quad 0,1,\cdots,M-1$$

The 2D wavelets are built from the 1D wavelets using the tensor-product construction. The 2D wavelet orthobasis contains the following functions

$$\varphi_{0,(0,0)}(s,t): \quad \varphi_{0,0}(s)\varphi_{0,0}(t),$$
$$\{\psi^1_{j,(k^1,k^2)}(s,t): \quad \varphi_{j,k^1}(s)\psi_{j,k^2}(t)\}_{-\infty<j<0,0\leq k^1,k^2<2^{-j}},$$
$$\{\psi^2_{j,(k^1,k^2)}(s,t): \quad \psi_{j,k^1}(s)\varphi_{j,k^2}(t)\}_{-\infty<j<0,0\leq k^1,k^2<2^{-j}},$$
$$\{\psi^3_{j,(k^1,k^2)}(s,t): \quad \psi_{j,k^1}(s)\psi_{j,k^2}(t)\}_{-\infty<j<0,0\leq k^1,k^2<2^{-j}}.$$

We now commit the *wavelet crime*: instead of studying how the discrete image $\boldsymbol{G}$ is compressed by the discrete 2D wavelet transform, we study wavelet coefficients of the continuous function $G(g(s),g(t))$ associated with the first N^2 elements of the orthobasis:

$$\varphi_{0,(0,0)},$$
$$\{\psi^1_{j,(k^1,k^2)}\}_{-\log_2 N+1\leq j\leq 0,0\leq k<2^{-j}},$$
$$\{\psi^2_{j,(k^1,k^2)}\}_{-\log_2 N+1\leq j\leq 0,0\leq k<2^{-j}},$$
$$\{\psi^3_{j,(k^1,k^2)}\}_{-\log_2 N+1\leq j\leq 0,0\leq k<2^{-j}}$$

Since the singularity is only on the diagonal, we only need to focus on its neighborhood. Near the diagonal, the kernel $G(g(s),g(t)) \quad \ln|g(s)-g(t)|$ has the same singularity behavior as $\ln|s-t|$. Therefore, we can simply take the 2D function $G(g(s),g(t))$ to be $\ln|s-t|$.

Let us first estimate the coefficients of the first kind of wavelets $\{\psi^1_{j,(k^1,k^2)}\}$. Suppose $\psi^1_{j,(k^1,k^2)} \quad \varphi_{j,k^1}(s)\psi_{j,k^2}(t)$ and the two components $\varphi_{j,k^1}(\cdot)$ and $\psi_{j,k^2}(\cdot)$ have non-overlapping supports. Since our wavelets have M vanishing moments, we get

$$\begin{aligned}
&\iint \psi^1_{j,(k^1,k^2)}(s,t)\ln|s-t|dsdt\\
&\int\int \varphi_{j,k^1}(s)\psi_{j,k^2}(t)\ln|s-t|dsdt\\
\leq &\int |\varphi_{j,k^1}(s)|\left(\int \psi_{j,k^2}(t)\ln|s-t|dt\right)ds\\
\leq &\int |\varphi_{j,k^1}(s)|\left(\max_{s\in supp(\varphi_{j,k^1}),t\in supp(\psi_{j,k^2})}\frac{2^{jM}}{|s-t|^M}\cdot\int \psi_{j,k^2}(t)dt\right)ds\\
\leq &\, C\cdot \max_{s\in supp(\varphi_{j,k^1}),t\in supp(\psi_{j,k^2})}\frac{2^{j(M+1)}}{|s-t|^M}\\
\leq &\, C\cdot \max_{s\in supp(\varphi_{j,k^1}),t\in supp(\psi_{j,k^2})}\frac{2^{jM}}{|s-t|^M}
\end{aligned}$$

for some constant C. Here we use the fact that $j \leq 0$. For a given accuracy ε, we set B to be $(1/\varepsilon)^{1/M}$. Suppose that the support of $\varphi_{j,k^1}(s)$ and $\psi_{j,k^2}(t)$ are separated so that $\min_{s\in supp(\varphi_{j,k^1}),t\in supp(\psi_{j,k^2})} |s-t| \geq B\cdot 2^j$, then

$$\iint \psi^1_{j,(k^1,k^2)}(s,t)\ln(|s-t|)dsdt \quad O(B^{-M}) \quad O(\varepsilon),$$

which is negligible.

Now let us count the number of non-negligible coefficients. For a fix level j and a fixed index k^2, the number of indices k^1 such that

$$\min_{s\in supp(\varphi_{j,k^1}),t\in supp(\psi_{j,k^2})} |s-t| \leq B\cdot 2^j$$

is $O(B)$. Therefore, for a fixed j and a fixed k^2, there are at most $O(B)$ wavelets $\{\psi^1_{j,(k^1,k^2)}(s,t)\}$ whose inner products with the kernel are greater than ε. The same argument works for the other two kinds of wavelets $\{\psi^2_{j,(k^1,k^2)}(s,t)\}$ and $\{\psi^3_{j,(k^1,k^2)}(s,t)\}$ because they all contain 1D wavelets (either in the variable s or in t). As a result, there are $O(3\cdot 2^{-j}\cdot B)$ non-negligible coefficients on each level j. Summing this over all $\log_2 N$ levels, we have in total

$$\sum_{j \quad -\log_2 N+1}^{0} O(3\cdot 2^{-j}\cdot B) \quad O(B\cdot N) \quad O(N)$$

non-negligible coefficients.

In order to compute the $O(N)$ non-negligible coefficients, we first notice that each wavelet or scaling function can be represented as the sum of a small number of scaling functions of the next level. Therefore, all we need to compute is the inner product of a function with a scaling function. Let us consider the 1D case to illustrate the idea.

$$\begin{aligned}\int f(x)\varphi_{j,k}(x)dx \quad & 2^{-j/2}\int f(x)\varphi(2^{-j}x-k+1)dx \\ & 2^{-j/2}\int f(x+2^j(k-1))\varphi(2^{-j}x)dx\end{aligned}$$

Let us assume that, for some τ_M, our scaling functions satisfy

$$\int \varphi(x+\tau_M)x^m dx \quad 0, \quad m \quad 1,\cdots,M-1 \quad \text{and} \quad \int \varphi(x)dx \quad 1.$$

Scaling functions with these properties have been constructed in [3]. With the help of these properties, we have

$$2^{-j/2}\int f(x+2^j(k-1))\varphi(2^{-j}x)dx \approx 2^{j/2}f(2^j(k-1+\tau_M))+O(2^{j(M+1/2)})$$

where the last equation uses only one point quadrature of f.

To summarize, we now have the following approximation

$$G(g(s),g(t)) \quad \sum_{i \quad 1}^{O(N)} c_i \eta_i(s,t) + O(\varepsilon) \tag{30}$$

where $\eta_i(s,t) \quad \eta_i^1(s)\eta_i^2(t)$ is a 2D wavelet with non-negligible coefficient c_i for each i. This approximation is called the *non-standard form* of Beylkin, Coifman and Rokhlin. If we plot the supports of the wavelets with non-negligible coefficients, the figure looks very much like the one of the hierarchical matrices (see Fig. 36).

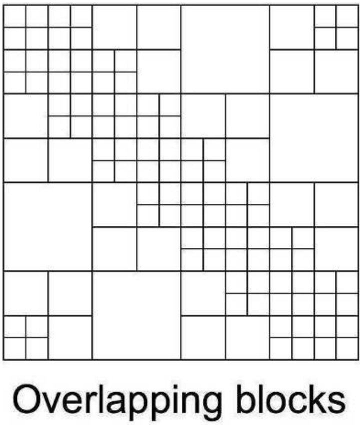


Fig. 36. The supports of the non-negligible terms in the non-standard form.

6.2 Fast matrix vector multiplication

We have so far focused on the compression the kernel matrix with the 2D wavelet basis. Let us discuss why this gives us a fast matrix vector multiplication algorithm. Using the non-standard form of the kernel matrix, we have

$$\begin{aligned} \int G(g(s),g(t))f(t)dt &\approx \sum_{i \quad 1}^{O(N)} \int c_i \eta_i(s,t) f(t)dt \\ & \quad \sum_{i \quad 1}^{O(N)} \int c_i \eta_i^1(s)\eta_i^2(t) f(t)dt \\ & \quad \sum_{i \quad 1}^{O(N)} \eta_i^1(s) \cdot \left(c_i \int \eta_i^2(t) f(t)dt \right) \end{aligned}$$

where $\{\eta_i^1(s)\}$ and $\{\eta_i^1(t)\}$ are either wavelets or scaling functions in 1D. We recall that a fast wavelet transform produces in its intermediate steps the inner products of the input function with all the scaling functions and wavelets. Therefore, the terms $\{\int \eta_i^2(t)f(t)dt\}$ for all i can be computed using a single fast wavelet transform by keeping all intermediate results.

Based on this simple observation, the wavelet based fast matrix multiplication algorithm has the following steps:

- Compute $\{\alpha_i \quad \int \eta_i^2(t)f(t)dt\}$ using a fast wavelet transform by keeping all intermediate results. The complexity of this step is $O(N)$.
- Compute $\{\beta_i \quad c_i\alpha_i\}$. This step takes only $O(N)$ operations.
- Synthesize $\sum_i \eta_i^1(s)\beta_i$ using an *extended* fast wavelet transform. This transform is *extended* in the sense that it includes not only the wavelet coefficients but also the scaling function coefficients since some of $\{\eta_i^1(s)\}$ are scaling functions. The complexity of this step is again $O(N)$.

As a result, the total complexity is $O(N)$, which is the same as the FMM. Before we end this section, let us summarize the main ideas behind the wavelet based method

- View the interaction matrix as an image. Since the singularity is only along the diagonal, a good compression with $O(N)$ non-negligible coefficients can be achieved using 2D wavelets.
- The tensor-product construction of the 2D wavelets allows one to use 1D fast wavelet transform to compute matrix vector multiplication in optimal time $O(N)$.

7 High Frequency FMM for the Helmholtz Kernel

In the rest two sections of this article, we discuss the computation of the oscillatory kernel of the Helmholtz equation in the 3D case.

$$-\Delta u - k^2 u \quad 0 \quad \text{in } \mathbb{R}^d \setminus \bar{D}.$$

Let us first rescale the geometry so that k is equal to 1. The equation then becomes

$$-\Delta u - u \quad 0 \quad \text{in } \mathbb{R}^d \setminus \bar{D}.$$

We face the following problem in each step of the iterative solver. Given a set of charges $\{f_i, 1 \leq i \leq N\}$ located at points $\{p_i, 1 \leq i \leq N\}$ and the Green's function

$$G(x,y) \quad h_0(|x-y|) \quad \frac{\exp(\mathrm{i}|x-y|)}{\mathrm{i}|x-y|}$$

of the Helmholtz kernel (up to a constant factor), we want to compute at each p_i

$$u_i \quad \sum_j G(p_i, p_j) f_j \tag{31}$$

(see Fig. 37).

In this section, we present the high frequency FMM (HF-FMM) by Rokhlin et al. [25, 26, 8] that calculates all $\{u_i\}$ in $O(N \log N)$ time. Suppose that the size of

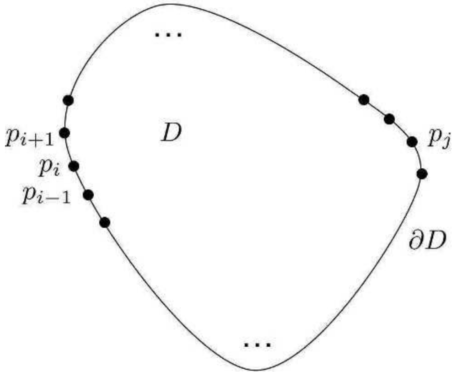


Fig. 37. Distribution of quadrature points $\{p_i\}$ on the boundary of the domain D.

the object is K wavelengths. Since one usually uses a constant number of points per wavelength in most of the scattering applications, $N \quad O(K^2)$.

For two points $x \quad (r, \theta, \varphi)$ and $x' \quad (r', \theta', \varphi')$ in spherical coordinates, we have the following important identity:

$$G(x,x') \quad h_0(|x-x'|) \quad \sum_{n\ 0}^{\infty} \sum_{m\ -n}^{n} Y_n^{-m}(\theta', \varphi') j_n(r') Y_n^m(\theta, \varphi) h_n(r)$$

when $r > r'$.

Far field representation

Suppose that a set of charges $\{f_j\}$ are located at $\{y_j \quad (r_j, \theta_j, \varphi_j)\}$ inside a box centered at the origin. Let us consider the potential generated by $\{y_j\}$ at a point $x \quad (r, \theta, \varphi)$ in the far field (see Fig. 38). Using the identity just mentioned, we have

$$u(x) \quad \sum_j G(x, y_j) f_j \quad \sum_{n\ 0}^{p} \sum_{m\ -n}^{n} \left(\sum_j f_j Y_n^{-m}(\theta_j, \varphi_j) j_n(r_j) \right) Y_n^m(\theta, \varphi) h_n(r) + \cdots$$

where p controls the number of terms to keep in the expansion and we will come back to it later. The far field representation is defined to be the coefficients $\{\alpha_n^m\}$ given by

$$\alpha_n^m \quad \sum_j f_j Y_n^{-m}(\theta_j, \varphi_j) j_n(r_j). \tag{32}$$

This representation is also called the *h-expansion* (see [8]).

Local field representation

Suppose that a set of charges $\{f_j\}$ are located at $\{y_j \quad (r_j, \theta_j, \varphi_j)\}$ in the far field of a box. Let us consider the potential generated by $\{y_j\}$ at a point $x \quad (r, \theta, \varphi)$ inside

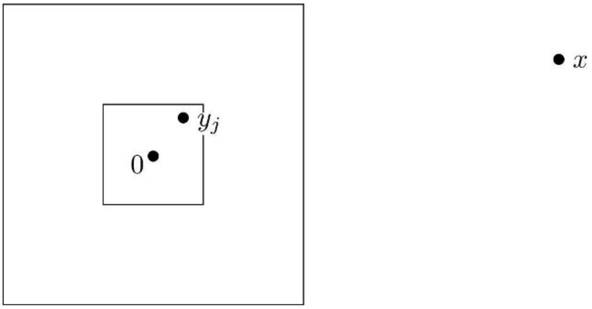


Fig. 38. Far field representation.

the box (see Fig. 39). We assume again that the center of the box is at the origin. From the identity given above, we have

$$u(x) \quad \sum_j G(x,y_j) f_j \quad \sum_{n \ 0}^{p} \sum_{m \ -n}^{n} \left(\sum_j f_j Y_n^m(\theta_j,\varphi_j) h_n(r_j) \right) Y_n^{-m}(\theta,\varphi) j_n(r) + \cdots .$$

The local field representation is defined to be $\{\beta_n^m\}$ given by

$$\beta_n^m \quad \sum_j f_j Y_n^m(\theta_j,\varphi_j) h_n(r_j). \tag{33}$$

This representation is called the *j-expansion*.

The first question we need to address is what is the value of p, i.e., how many terms to keep in these two expansions for a prescribed accuracy ε. For a box with radius R, the n-th term of the h-expansion

$$u(x) \quad \sum_{n \ 0}^{\infty} \sum_{m \ -n}^{n} \left(\sum_j f_j Y_n^{-m}(\theta_j,\varphi_j) j_n(r_j) \right) Y_n^m(\theta,\varphi) h_n(r)$$

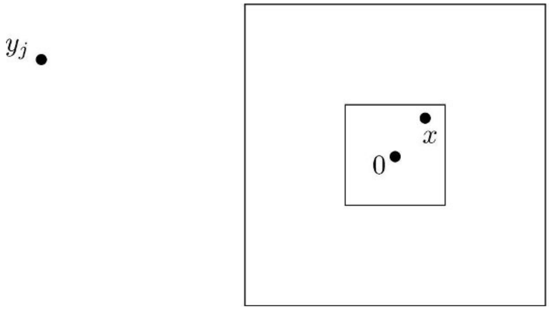


Fig. 39. Local field representation.

behaves like $h_n(3R)j_n(\sqrt{3}R)$. This product only decays when $n \geq 3R$. In order to get an accurate expansion, we are forced to choose $p \quad O(R)$ and keep all terms for $n \quad 0,1,\cdots,p,-n \leq m \leq n$. Therefore, the number of coefficients in h-expansion is $O(R^2)$. The same is also true for the j-expansion. Let us recall that the point set $\{p_i\}$ is distributed on the boundary surface ∂D. It is not difficult to see that there are $O(R^2)$ points in a box with radius R as well. This means that, from the information theoretical point of view, there is no compression at all when one transforms the charges $\{f_j\}$ to the h-expansion coefficients $\{\alpha_n^m\}$.

When the radius of the box R is $O(1)$, the h-expansion and the j-expansion both have complexity $O(1)$. Therefore, it is still reasonable to use them as the far field and local field representations. The far-to-far, far-to-local, and local-to-local translations are very similar to the case of the 3D Laplace kernel:

- Far-to-far and local-to-local translations. The "point and shoot" approach is used. The complexity is cut down from $O(p^4)$ to $O(p^3)$ if the number of terms in the h-expansion is $O(p^2)$.
- Far-to-local translation. The plane wave (exponential) expansion is used to diagonalize the far-to-local translation. The translation between the h-expansion (or the j-expansion) and the plane wave expansion takes $O(p^3)$ operations, while each far-to-local translation in the plane wave expansion uses only $O(p^2)$ steps.

For large boxes, for example when $R \quad O(K)$, the situation is drastically different. The number of terms in the h-expansion or the j-expansion is equal to $O(R^2)$ $O(K^2)$. As a result, the complexity of the three translations are $O(R^3) \quad O(K^3)$ $O(N^{3/2})$, which is already higher than the $O(N\log N)$ complexity that we aim for. The solution to this problem, the so-called high frequency fast multipole method (HF-FMM), is to represent the h expansion and j expansion in a form such that the far-to-far, far-to-local, and local-to-local translations are all diagonalized.

Far field signature

For the h-expansion, we transform its coefficients $\{\alpha_n^m\}$ to

$$f(\theta,\varphi): \quad \sum_{n\ 0}^{p}\sum_{m\ -n}^{n} \alpha_n^m(-1)^{n+1}Y_n^m(\theta,\varphi).$$

This function is in fact the *far field signature* of the potential of

$$u(x) \quad \sum_{n\ 0}^{p}\sum_{m\ -n}^{n} \alpha_n^m Y_n^m(\theta,\varphi)h_n(r)$$

where $x \quad (r,\theta,\varphi)$. Similar, for j-expansion, we transform its coefficients $\{\beta_n^m\}$ to

$$g(\theta,\varphi): \quad \sum_{n\ 0}^{p}\sum_{m\ -n}^{n} \beta_n^m(-1)^{n+1}Y_n^m(\theta,\varphi).$$

This function, which is also called the far field signature, can be viewed as a source distribution on a unit sphere which reproduces the potential inside the box if one pushes the radius of the sphere to infinity and rescales the source distribution appropriately. All three translations are diagonalized in the far field signatures $f(\theta,\varphi)$ and $g(\theta,\varphi)$. We refer to [8, 10, 24] for the formulas of these translations.

In the HF-FMM algorithm, the octree is divided into the low and high frequency regimes. In a typical case, the low frequency regime contains all the boxes with radius < 1, while the boxes in the high frequency regime have radius ≥ 1. The h-expansion and the j-expansion serve as the far field and local field representations in the low frequency regime while the far field signatures $f(\theta,\varphi)$ and $g(\theta,\varphi)$ are the representations in the high frequency regime. A switch of the representations appears at the boxes with radius ≈ 1.

We would like to comment that the far field signatures $f(\theta,\varphi)$ and $g(\theta,\varphi)$ cannot be used for the low frequency box with radius $r < 1$. The reason is that the far-to-local translation of the far field signatures involves extremely large numbers when the box is too small. Therefore, given a fixed precision of calculation and a prescribed accuracy ε, one can only use the far field signatures on sufficiently large boxes in order to avoid numerical instability. The overall structure of the HF-FMM algorithm is shown in Fig. 40.

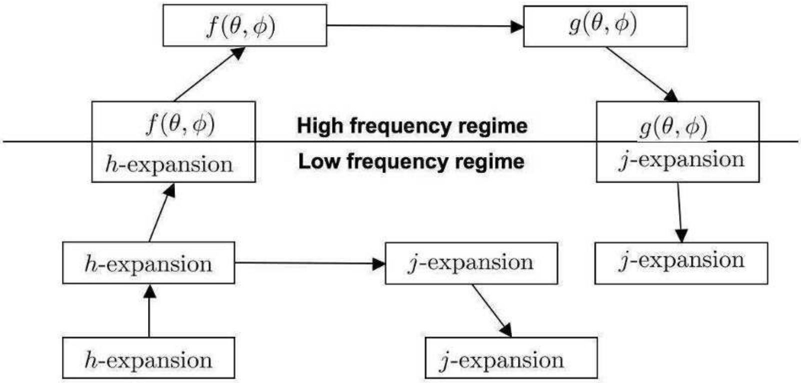


Fig. 40. The overall structure of the HF-FMM algorithm.

Most of the computation of HF-FMM is devoted to the high frequency regime, while the computation in the low frequency regime is similar to the one of the 3D Laplace kernel. Since the point set $\{p_i\}$ is sampled from the two-dimensional boundary ∂D, there are $O((K/r)^2)$ boxes with a fixed radius r. For each of them, the computation involves the far-to-far, far-to-local and local-to-local translations. Since all these translations are diagonalized, each of them has complexity $O(r^2)$. Therefore, the number of operations spent on the boxes with radius r is $O((K/r)^2)\cdot O(r^2)$ $O(K^2)$. Summing this over all $\log K$ level, we conclude that the complexity of the HF-FMM is

$$O(K^2 \log K) \quad O(N \log N).$$

8 Multidirectional Method

Through our discussion of the fast algorithms for the Laplace equation, we see that the interaction between a domain B and its far field has a low separation rank which is almost independent of the size of B. This low rank property has played a fundamental role in the fast multipole method, its kernel independent variant, and the hierarchical matrix framework.

As we have seen from the previous section, the situation is quite opposite for the Helmholtz equation. Suppose that B is a domain such that its radius is much larger than the wavelength. The interaction between B and its far field (see Fig. 41) through the Helmholtz kernel

$$G(x,y) \quad h_0(|x-y|) \quad \frac{\exp(\mathrm{i}|x-y|)}{\mathrm{i}|x-y|}$$

is not low rank anymore. In fact, the rank is proportional to the square of the radius of B. A natural question to ask is that whether it is possible to recover the low rank property in the setting of the Helmholtz kernel. The answer is positive and it is the motivation behind a multidirectional algorithm developed recently in [13].

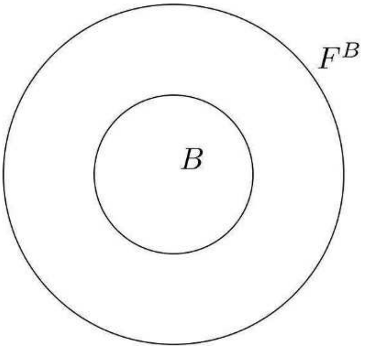


Fig. 41. The interaction between B and its far field F^B is not low rank for the Helmholtz kernel.

8.1 Analysis

Directional parabolic separation

Let us start by considering the geometric configuration in Fig. 42. Suppose B is a domain with radius r. The wedge A, which has an opening angle $O(1/r)$, is at a distance r^2 or greater from B. Whenever two sets A and B follow this geometric configuration, we say that they satisfy the *directional parabolic separation condition.*

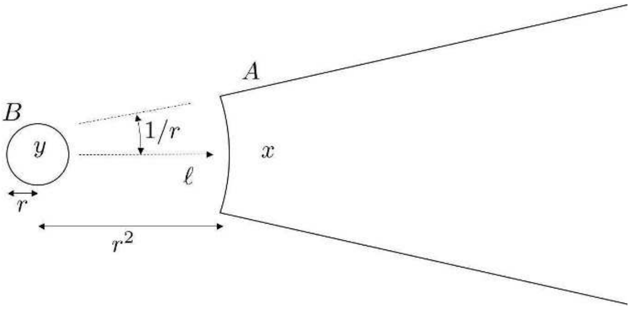


Fig. 42. Directional parabolic separated condition.

Theorem 5. *Suppose that B and A satisfy the directional parabolic separation condition. Then there exist an integer $T(\varepsilon)$ and two sets of functions $\{\alpha_i(x), 1 \le i \le T(\varepsilon)\}$ and $\{\beta_i(y), 1 \le i \le T(\varepsilon)\}$ such that, for any $x \in A$ and $y \in B$*

$$\left| \frac{\exp(\mathrm{i}|x-y|)}{\mathrm{i}|x-y|} - \sum_{i \ \ 1}^{T(\varepsilon)} \alpha_i(x)\beta_i(y) \right| < \varepsilon$$

where the number of terms $T(\varepsilon)$ of the expansion is independent of the radius of B.

The main idea behind this theorem is quite simple and it is not difficult to see why it can work. In the wedge A, the radiation generated by the points in B looks almost like a plane wave since the opening angle of the wedge A is inversely proportional to the radius of B. After one factors out the plane wave (which itself has a rank-1 separated representation), the rest of the interaction is smooth and hence has an approximate low rank separated representation.

The construction of $\{\alpha_i(x)\}$ and $\{\beta_i(y)\}$ is similar to the pseudo-skeleton approach discussed in the hierarchical matrix framework. In practice, the following randomized procedure works quite well.

- Randomly sample the set B to find positions $\{b_q\}$ such that the functions $\{G(x,b_q)\}_q$ span the space of the functions $\{G(x,y)\}_y$ within a prescribed accuracy ε.
- Randomly sample the set A to find positions $\{a_p\}$ such that the functions $\{G(a_p,y)\}_p$ span the space of the functions $\{G(x,y)\}_x$ within a prescribed accuracy ε.
- Find the matrix $D \quad (d_{qp})$ such that

$$\left| \frac{e^{\mathrm{i}|x-y|}}{\mathrm{i}|x-y|} - \sum_q \frac{e^{\mathrm{i}|x-b_q|}}{\mathrm{i}|x-b_q|} \sum_p d_{qp} \frac{e^{\mathrm{i}|a_p-y|}}{\mathrm{i}|a_p-y|} \right| \quad O(\varepsilon).$$

The first two steps use the pivoted QR factorizations, while the last step can be viewed as a least square problem.

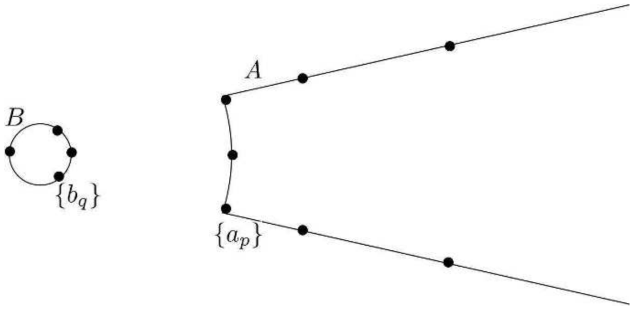


Fig. 43. Directional equivalent charges.

Directional equivalent charges

Suppose that B and A satisfy the directional parabolic separation condition. Let $\{y_j\}$ be a set of points in B. We consider the potential at $x \in A$ generated by the charges $\{f_j\}$ at $\{y_j\}$ (see Fig. 43).

Using the low rank representation generated above, we have

$$\left| \sum_i \frac{e^{\mathrm{i}|x-y_i|}}{\mathrm{i}|x-y_i|} f_i - \sum_q \frac{e^{\mathrm{i}|x-b_q|}}{\mathrm{i}|x-b_q|} \sum_p d_{qp} \sum_i \frac{e^{\mathrm{i}|a_p-y_i|}}{\mathrm{i}|a_p-y_i|} f_i \right| \quad O(\varepsilon).$$

This equation suggests that, by placing charges

$$\left\{ \sum_p d_{qp} \sum_i \frac{e^{\mathrm{i}|a_p-y_i|}}{\mathrm{i}|a_p-y_i|} f_i \right\}$$

at the points $\{b_q\}$, one can reproduce the potential at x accurately. We call these charges the *directional equivalent charges* of B in the direction of A. The above formula also provides a way to compute the directional equivalent charges from the source charges $\{f_j\}$:

- Evaluate the potentials at $\{a_p\}$ generated by $\{f_j\}$.
- Multiply the potentials with the matrix $D \quad (d_{qp})$ to obtain the directional equivalent charges.

Directional check potentials

Now let us reverse the roles of B and A (see Fig. 44). Suppose that $\{y_j\}$ are a set of points in the B. We consider the potential at $x \in A$ generated by the charges $\{f_j\}$ at $\{y_j\}$. Using the low rank representation of the kernel, we have

$$\left| \sum_i \frac{e^{\mathrm{i}|x-y_i|}}{\mathrm{i}|x-y_i|} f_i - \sum_q \frac{e^{\mathrm{i}|x-b_q|}}{\mathrm{i}|x-b_q|} \sum_p d_{qp} \sum_i \frac{e^{\mathrm{i}|a_p-y_i|}}{\mathrm{i}|a_p-y_i|} f_i \right| \quad O(\varepsilon).$$

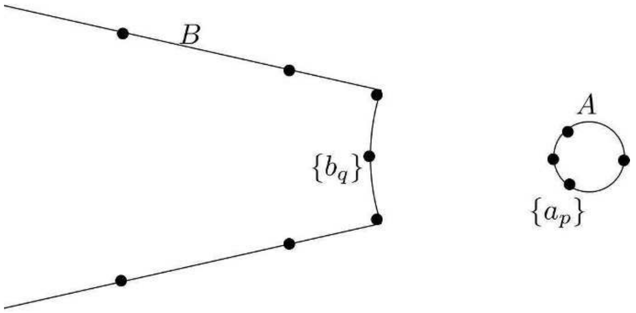


Fig. 44. Directional check potentials.

This equation shows that from the potentials

$$\left\{ \sum_i \frac{e^{\mathrm{i}|a_p - y_i|}}{\mathrm{i}|a_p - y_i|} f_i \right\}$$

at $\{a_p\}$ we can reconstruct the potential at any point $x \in A$ efficiently and accurately. The steps are:

- Multiply these potentials with the matrix $D \quad (d_{qp})$.
- Use the result as the charges at $\{b_q\}$ to compute the potential at x.

We call these potentials the *directional check potentials* of A in the direction of B.

8.2 Algorithmic description

Geometric part

Similar to the HF-FMM, the octree is divided into the low frequency regime (where the width of the box is < 1) and the high frequency regime (where the width of the box is ≥ 1). However, the definition of the far field region is much more complicated:

- For a box B with width $r < 1$ in the low frequency regime, the far field F^B contains all the well-separated boxes.
- For a box B with width $r \geq 1$ in the high frequency regime, the far field F^B contains the boxes which are at least r^2 away. The interaction list of B contains the boxes which are in the far field of B but not in the far field of the parent of B. All these boxes belong to a shell with radius from r^2 to $4r^2$. The far field is further partitioned into $O(r^2)$ wedges $\{W^{B,\ell}\}$ indexed by $\{\ell\}$, each with an opening angle of size $O(1/r)$ (see Fig. 45).

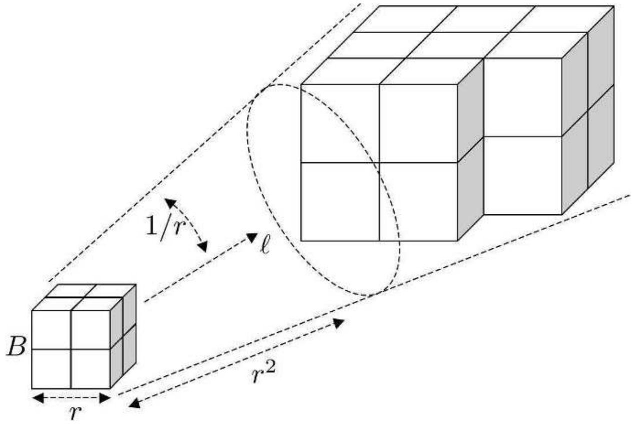


Fig. 45. The far field of B is partitioned into multiple wedges, each with an opening angle of size $O(1/r)$.

Far field and local field representations

For a box B with width $r < 1$ in the low frequency regime, the far field representation is the (non-directional) equivalent charges $\{f_k^{B,F}\}_k$ of the kernel independent FMM. From the previous discussion, we know that its complexity is $O(1)$.

For a box B with width $r \geq 1$ in the high frequency regime, the far field representation consists of the directional equivalent charges $\{f_k^{B,F,\ell}\}_k$ of all $O(r^2)$ wedges $\{W^{B,\ell}\}$. In order to compute the potential at a point x in the far field, we need to use the charges $\{f_k^{B,F,\ell}\}_k$ associated with the wedge $W^{B,\ell}$ that x belongs to. As we use $O(1)$ directional equivalent charges for each direction, the complexity of the far field representation is $O(r^2)$.

For a box A with width $r < 1$ in the low frequency regime, the local field representation is the (non-directional) check potentials $\{u_k^{A,L}\}_k$ of the kernel independent FMM. Its complexity is $O(1)$.

For a box A with width $r \geq 1$ in the high frequency regime, the local field representation consists of the directional check potentials $\{u_k^{A,L,\ell}\}_k$ of all $O(r^2)$ wedges $\{W^{A,\ell}\}$. For a point x in A, in order to compute the potential at x generated by the source charges in wedge $W^{A,\ell}$, we need to use the check potentials $\{u_k^{A,L,\ell}\}_k$. Since the directional check potentials for each direction contain $O(1)$ coefficients, the complexity of the local field representation is $O(r^2)$.

Far-to-far, far-to-local, and local-to-local translations

The translations in the low frequency regime are exactly the same as the ones of the kernel independent FMM. Therefore, we only discuss these translations in the high frequency regime.

The far-to-local translation is quite simple. Consider two boxes A and B in each other's interaction list. Suppose A is in $W^{B,\ell}$ and B is in $W^{A,\ell'}$. The far-to-local translation from B to A simply evaluates $\{u_k^{A,L,\ell'}\}_k$ using $\{f_k^{B,F,\ell}\}_k$.

For the far-to-far translation, we construct the directional equivalent charges of a parent box B from its child box B'. Let us consider the wedges $\{W^{B,\ell}\}$ one by one. An important observation, which is clear from the following figure, is that $W^{B,\ell}$ is contained in a wedge $W^{B',\ell'}$ of its child B'. Therefore, to construct $\{f_k^{B,F,\ell}\}_k$ at B, we can simply regard $\{f_k^{B',F,\ell'}\}_k$ as the source charges (see Fig. 46).

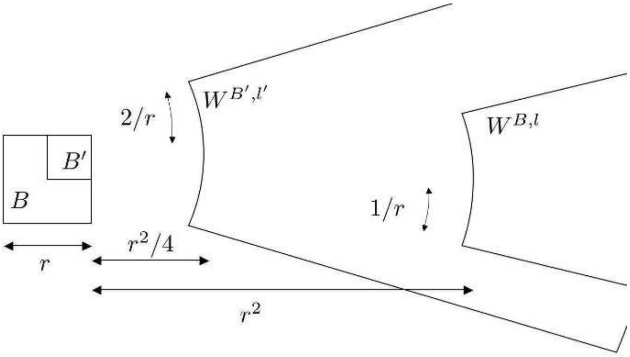


Fig. 46. Far-to-local translation between B and B'.

As a result, the steps of a far-to-local translation in the high frequency regime are:

- Use the directional equivalent charges $\{f_k^{B',F,\ell'}\}_k$ of B' as the source charges to compute the potentials at locations $\{a_p\}$ of the box B.
- Multiplication with the matrix (d_{qp}) to obtain $\{f_k^{B,F,\ell}\}_k$.

The local-to-local translation is implemented in a similar way. The main components of this algorithm is illustrated in Fig. 47.

Let us now discuss the computational complexity of this multidirectional algorithm. For a box of width r, most of the computation is devoted to the three translations.

- There are $O(r^2)$ far-to-far translations, one for each wedge. Since each far-to-far translation takes $O(1)$ operations, the complexity is $O(r^2)$.
- There are $O(r^2)$ local-to-local translations, again one for each wedge. Since each local-to-local translation takes also $O(1)$ operation, the complexity is again $O(r^2)$.
- Let us count the number of far-to-local translations for a box B. All the boxes in B's interaction list belong to a shell with radius between r^2 and $4r^2$. It is clear that there are $O(r^2)$ boxes in this shell since the points are sampled from the

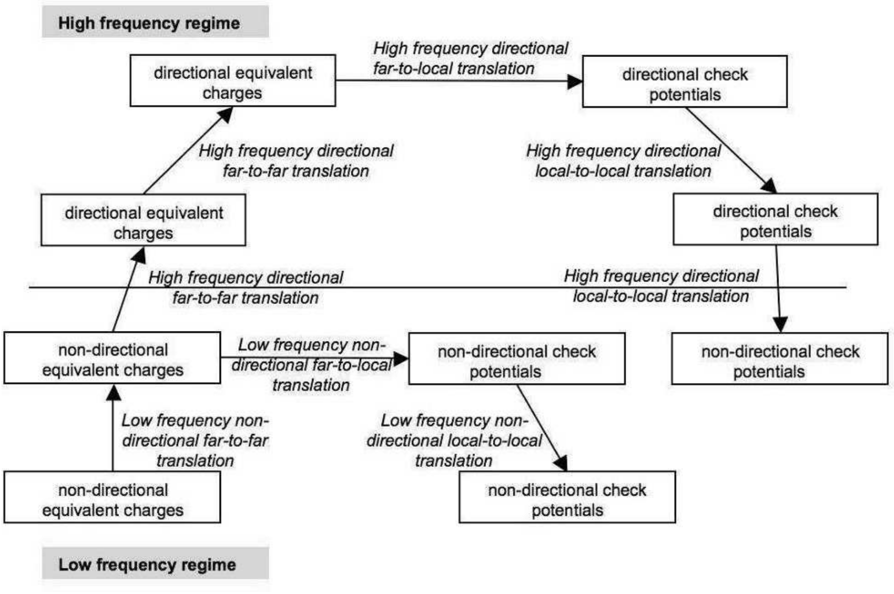

High frequency regime
directional equivalent charges
High frequency directional far-to-local translation
directional check potentials
High frequency directional far-to-far translation
High frequency directional local-to-local translation
directional equivalent charges
directional check potentials
High frequency directional far-to-far translation
High frequency directional local-to-local translation
non-directional equivalent charges
Low frequency non-directional far-to-local translation
non-directional check potentials
non-directional check potentials
Low frequency non-directional far-to-far translation
Low frequency non-directional local-to-local translation
non-directional equivalent charges
non-directional check potentials
Low frequency regime

Fig. 47. The overall structure of the multidirectional algorithm.

surface boundary ∂D. Since each far-to-local translation takes $O(1)$ operations, the complexity is also $O(r^2)$.

For a given size r, there are $O(K^2/r^2)$ boxes of this size. Therefore, the number of steps spent on each level is $O(K^2/r^2) \cdot O(r^2) \quad O(K^2)$. Finally, summing over all $O(\log K)$ levels, we conclude that the complexity of this multidirectional algorithm is

$$O(K^2 \log K) \quad O(N \log N),$$

which is the same as the complexity of the HF-FMM algorithm.

9 Concluding Remarks

This paper discussed several fast algorithms for boundary integral equations. In the case of non-oscillatory kernels, we reviewed the fast multipole method (FMM) and its kernel independent variant, the hierarchical matrix framework, and the wavelet-based method. In each of these methods, we exploit the fact that the interaction between two well-separated regions is approximately low rank. For the oscillatory kernels, we discussed the high frequency fast multipole method (HF-FMM) and the recently proposed multidirectional algorithm. The HF-FMM used the far field signature to diagonalize the well-separated interaction, while the multidirectional algorithm decomposes the interaction in a directional way using the so-called directional parabolic separation condition.

Our choice of the methods is quite personal. Many other efficient algorithms were left out, such as the panel-clustering method [19], the FFT-based methods [4, 5, 7, 23], the local Fourier basis method [2], and the direct solver method [22]. Furthermore, we omitted the important field of time-domain integral equations, which often provide efficient ways to solve linear parabolic and hyperbolic PDEs. We point the readers to [10, 17] for fast algorithms for these boundary integral equations.

Acknowledgement. The author thanks the organizers of the 2007 Summer School on Multiscale Modeling and Simulation in Science for their hospitality and the students at this same school for their comments and suggestions on the manuscript. The author particularly acknowledges Stefan Engblom and Per Lötstedt for reading an early version of this article and correcting numerous mistakes and misprints. The author is supported by an Alfred P. Sloan Research Fellowship and a startup grant from the University of Texas at Austin.

10 Exercises

Exercise 1. Solve the Laplace equation

$$-\Delta u \quad 0 \quad \text{in}\, D$$

$$u \quad f \quad \text{on}\, \partial D$$

on the domain $D \quad \{(x_1,x_2) : x_1^2 + 4x_2^2 < 1\}$ using the second kind integral equation

$$f(z) \quad \frac{1}{2}\varphi(z) - \int_{\partial D} \frac{\partial G(z,y)}{\partial n(y)} \varphi(y) ds(y).$$

Let us parameterize the boundary ∂D as

$$x_1 \quad \cos(\theta) \quad x_2 \quad 0.5\sin(\theta), \quad \theta \in \; 0, 2\pi \, .$$

You can use the trapezoidal rule to numerically approximate the integral $\int_{\partial D} \frac{\partial G(z,y)}{\partial n(y)} \varphi(y) dy$. For this problem, $\lim_{y \to z} \frac{\partial G(z,y)}{\partial n(y)}$ exists and can be expressed in terms of the curvature of ∂D at z.

- Solve for $\varphi(z)$ when the boundary condition is $f(\theta) \quad \cos(4\theta)$ in the angular parameterization. Let the number of quadrature points N be 128.
- Please plot the solution $u(x)$ for x in $\{(x_1,x_2) : x_1^2 + 4x_2^2 < 0.9^2\}$.

Exercise 2. Solve the Helmholtz equation

$$-\Delta u - k^2 u \quad 0 \quad \text{in}\, \mathbb{R}^d \setminus \bar{D}$$

$$u(x) \quad -u^{inc}(x) \quad \text{for}\, x \in \partial D$$

$$\lim_{r \to \infty} r^{\frac{d-1}{2}} \left(\frac{\partial u}{\partial r} - \mathrm{i} k u \right) \quad 0$$

on the domain $\mathbb{R}^2 \setminus \bar{D}$ where $D \quad \{(x_1,x_2) : x_1^2+4x_2^2 < 1\}$ using the second kind integral equation

$$-u^{inc}(z) \quad \frac{1}{2}\varphi(z)+\int_{\partial D}\left(\frac{\partial G(z,y)}{\partial n(y)}\right)\varphi(y)ds(y).$$

where η is set to be zero. Let us use the same parameterization for ∂D as the previous problem and again the trapezoidal rule to discretize the integral. The following formula will be useful for the computation of $\frac{\partial G(z,y)}{\partial n(y)}$

$$\frac{d}{dr}H_0^n(r) \quad \frac{nH_n^1(r)}{r}-H_{n+1}^1(r).$$

The limit $\lim_{y\to z}\frac{\partial G(z,y)}{\partial n(y)}$ exists as well in this case and is equal to the one of the previous example.

- Choose $k \quad 64$ and $N \quad 8k$. Given $u^{inc}(x) \quad \exp(ik(x\cdot d))$ with $d \quad (1,0)$, please solve for $\varphi(z)$.
- Pleases plot the scattering field $u(x)$ for x in $\{(x_1,x_2) : x_1^2+4x_2^2 > 1.1^2\}$.

Exercise 3. Let us consider the wavelet based method. The boundary is a circle parameterized by $g(s) \quad (\cos(2\pi s),\sin(2\pi s))$ for $s \in 0,1$. Take the kernel to be the Green's function of the Laplace equation:

$$\ln|g(s)-g(t)|.$$

- Please discretize the kernel with N points. For the diagonal, simply put 0. This gives you an $N\times N$ image.
- Compress this image with 2D Daubechies wavelets.
- Compare, for different values of N and ε, how many wavelet coefficients are greater than ε.

References

1. C. R. Anderson. An implementation of the fast multipole method without multipoles. *SIAM J. Sci. Statist. Comput.*, 13(4):923–947, 1992.
2. A. Averbuch, E. Braverman, R. Coifman, M. Israeli, and A. Sidi. Efficient computation of oscillatory integrals via adaptive multiscale local Fourier bases. *Appl. Comput. Harmon. Anal.*, 9(1):19–53, 2000.
3. G. Beylkin, R. Coifman, and V. Rokhlin. Fast wavelet transforms and numerical algorithms. I. *Comm. Pure Appl. Math.*, 44(2):141–183, 1991.
4. E. Bleszynski, M. Bleszynski, and T. Jaroszewicz. AIM: Adaptive integral method for solving large-scale electromagnetic scattering and radiation problems. *Radio Science*, 31:1225–1252, 1996.
5. N. Bojarski. K-space formulation of the electromagnetic scattering problems. Technical report, Air Force Avionic Lab. Technical Report AFAL-TR-71-75, 1971.

6. S. Börm, L. Grasedyck, and W. Hackbusch. Hierarchical matrices. Technical Report 21, Max-Planck-Institut für Mathematik in den Naturwissenschaften, Leipzig, 2003.
7. O. P. Bruno and L. A. Kunyansky. A fast, high-order algorithm for the solution of surface scattering problems: basic implementation, tests, and applications. *J. Comput. Phys.*, 169(1):80–110, 2001.
8. H. Cheng, W. Y. Crutchfield, Z. Gimbutas, L. F. Greengard, J. F. Ethridge, J. Huang, V. Rokhlin, N. Yarvin, and J. Zhao. A wideband fast multipole method for the Helmholtz equation in three dimensions. *J. Comput. Phys.*, 216(1):300–325, 2006.
9. H. Cheng, L. Greengard, and V. Rokhlin. A fast adaptive multipole algorithm in three dimensions. *J. Comput. Phys.*, 155(2):468–498, 1999.
10. W. Chew, E. Michielssen, J. M. Song, and J. M. Jin, editors. *Fast and efficient algorithms in computational electromagnetics*. Artech House, Inc., Norwood, MA, USA, 2001.
11. D. Colton and R. Kress. *Inverse acoustic and electromagnetic scattering theory*, volume 93 of *Applied Mathematical Sciences*. Springer-Verlag, Berlin, second edition, 1998.
12. I. Daubechies. *Ten lectures on wavelets*, volume 61 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1992.
13. B. Engquist and L. Ying. Fast directional multilevel algorithms for oscillatory kernels. *SIAM J. Sci. Comput.*, 29(4):1710–1737 (electronic), 2007.
14. S. A. Goreinov, E. E. Tyrtyshnikov, and N. L. Zamarashkin. A theory of pseudoskeleton approximations. *Linear Algebra Appl.*, 261:1–21, 1997.
15. L. Greengard. *The rapid evaluation of potential fields in particle systems*. ACM Distinguished Dissertations. MIT Press, Cambridge, MA, 1988.
16. L. Greengard and V. Rokhlin. A fast algorithm for particle simulations. *J. Comput. Phys.*, 73(2):325–348, 1987.
17. L. Greengard and J. Strain. A fast algorithm for the evaluation of heat potentials. *Comm. Pure Appl. Math.*, 43(8):949–963, 1990.
18. W. Hackbusch. *Integral equations*, volume 120 of *International Series of Numerical Mathematics*. Birkhäuser Verlag, Basel, 1995. Theory and numerical treatment, Translated and revised by the author from the 1989 German original.
19. W. Hackbusch and Z. P. Nowak. On the fast matrix multiplication in the boundary element method by panel clustering. *Numer. Math.*, 54(4):463–491, 1989.
20. R. Kress. *Linear integral equations*, volume 82 of *Applied Mathematical Sciences*. Springer-Verlag, New York, second edition, 1999.
21. S. Mallat. *A wavelet tour of signal processing*. Academic Press Inc., San Diego, CA, 1998.
22. P. G. Martinsson and V. Rokhlin. A fast direct solver for boundary integral equations in two dimensions. *J. Comput. Phys.*, 205(1):1–23, 2005.
23. K. Nabors, F. Korsmeyer, F. Leighton, and J. K. White. Preconditioned, adaptive, multipole-accelerated iterative methods for three-dimensional first-kind integral equations of potential theory. *SIAM Journal on Scientific and Statistical Computing*, 15:713–735, 1994.
24. N. Nishimura. Fast multipole accelerated boundary integral equation methods. *Applied Mechanics Reviews*, 55(4):299–324, 2002.
25. V. Rokhlin. Rapid solution of integral equations of scattering theory in two dimensions. *J. Comput. Phys.*, 86(2):414–439, 1990.
26. V. Rokhlin. Diagonal forms of translation operators for the Helmholtz equation in three dimensions. *Appl. Comput. Harmon. Anal.*, 1(1):82–93, 1993.
27. L. Ying, G. Biros, and D. Zorin. A kernel-independent adaptive fast multipole algorithm in two and three dimensions. *J. Comput. Phys.*, 196(2):591–626, 2004.

Wavelets and Wavelet Based Numerical Homogenization

Olof Runborg

Department of Numerical Analysis, KTH, 100 44 Stockholm, Sweden,
olofr@nada.kth.se

1 Introduction

Wavelets is a tool for describing functions on different *scales* or *level of detail*. In mathematical terms, wavelets are functions that form a basis for $L^2(\mathbb{R})$ with special properties; the basis functions are spatially localized and correspond to different scale levels. Finding the representation of a function in this basis amounts to making a *multiresolution decomposition* of the function. Such a wavelet representation lends itself naturally to analyzing the fine and coarse scales as well as the localization properties of a function. Wavelets have been used in many applications, from image and signal analysis to numerical methods for partial differential equations (PDEs). In this tutorial we first go through the basic wavelet theory and then show a more specific application where wavelets are used for numerical homogenization. We will mostly give references to the original sources of ideas presented. There are also a large number of books and review articles that cover the topic of wavelets, where the interested reader can find further information, e.g. [25, 51, 48, 7, 39, 26, 23], just to mention a few.

2 Wavelets

Wavelet bases decompose a function into parts that describe it on different scales. Starting from an example, we show in this section how such bases can be constructed in a systematic way via multiresolution analysis; in particular we present in a little more detail the construction of Daubechies's famous compactly supported wavelets. We also discuss wavelets' approximation properties and time frequency localization, as well as the fast wavelet transform. The focus is on the basic theory and ideas of first generation wavelets. Later extensions will be mentioned but not elaborated on.

2.1 An example

We begin by discussing multiresolution decomposition of a function in an informal way via an example illustrated in Fig. 1. The starting point is the function at the top

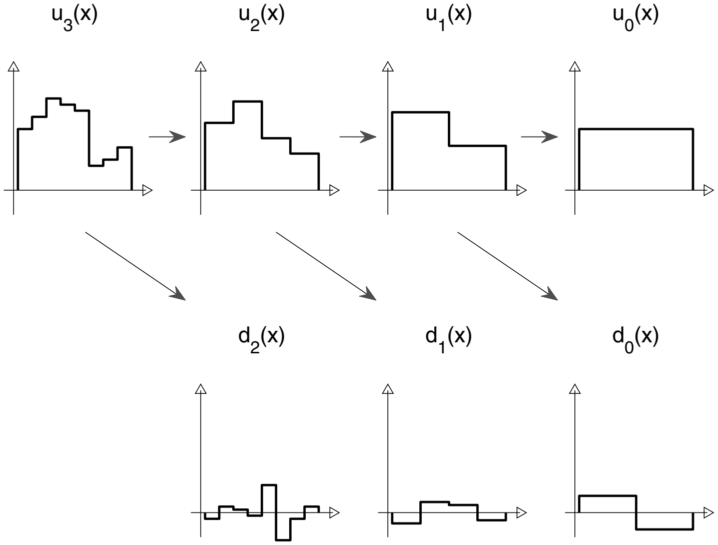


Fig. 1. Multiscale decomposition of a function.

left in Fig. 1 plotted in the interval $0,1$. It is a piecewise constant function where each piece has the same length. We call this function $u_3(x)$ with 3 indicating that the length of the pieces is $1/8 \quad 2^{-3}$. We can now construct a new piecewise constant function by taking the average of adjacent pieces in $u_3(x)$. This is the function in the second subfigure. We call it $u_2(x)$ since the length of the pieces is now doubled, $1/4 \quad 2^{-2}$. Below in the second row is the difference between $u_3(x)$ and $u_2(x)$ which we call $d_2(x)$. Hence,

$$u_3(x) \quad u_2(x) + d_2(x).$$

Clearly, this process can be continued iteratively, as indicated in the figure. We construct $u_1(x)$ by taking local averages of $u_2(x)$ and let $d_1(x) \quad u_2(x) - u_1(x)$, etc. In the end we only have $u_0(x)$, the total average of $u_3(x)$, and the differences such that

$$u_3(x) \quad u_0(x) + d_0(x) + d_1(x) + d_2(x).$$

The differences can be interpreted precisely as the information contents of the function on different scale levels.

This kind of multiresolution decomposition is useful in many applications, and at the heart of the theory of wavelets. Some of the advantages of decomposing a function in this way is:

- The approximations $u_j(x)$ give an explicit description of how the original function looks like if viewed on different scales. The larger j, the more details of the function are present. The differences $d_j(x)$ contain the parts of the function that belong to the different scales. This is often of great use in understanding and analyzing phenomena.

- The size of the $d_j(x)$ functions rapidly go to zero as j increases, if the original function is smooth. One can therefore approximate the function with just a few of the $d_j(x)$.
- The differences $d_j(x)$ indicate where the original function is non-smooth. This is used in many applications from edge detection in image analysis to mesh refinement algorithms in the numerical solution of PDEs.

2.2 The Haar wavelet and scaling spaces

We will now discuss the example above and put it in a systematic framework by considering the spaces that $u_j(x)$ and $d_j(x)$ belong to, and the basis functions that span those spaces. We note first that the initial function $u_3(x)$ can be written as a linear combination of piecewise constant functions $\{\phi_{3,k}\}$ indicated in the top left frame of Fig. 2,

$$u_3(x) \quad \sum_k u_{3,k}\phi_{3,k}(x),$$

for some coefficients $\{u_{3,k}\}$. The functions $\phi_{3,k}(x)$ are the basis functions for the space of all piecewise constant functions with pieces of length 2^{-3}. We call this space V_3. Reasoning in the same way for the other u_j functions, we have that $u_j(x) \in V_j$, the space of piecewise constant functions with piece length 2^{-j}, spanned by $\phi_{j,k}(x)$ as shown in the first row of Fig. 2. The spaces V_j are called *scaling spaces* and in general they contain the functions "viewed on scale j", where more detailed[1] functions belong to spaces with larger j. For practical and technical reasons we will henceforth restrict ourselves to piecewise constant functions that are also in $L^2(\mathbb{R})$, the space of square integrable functions.

The difference functions $d_j(x)$ can be treated in a similar fashion, but with different basis functions $\psi_{j,k}(x)$. These are indicated in the second row of Fig. 2. Thus

$$d_j(x) \quad \sum_k d_{j,k}\psi_{j,k}(x),$$

for some coefficients $\{d_{j,k}\}$. The space spanned by $\{\psi_{j,k}\}$ will be called W_j. The functions $\{\psi_{j,k}\}$ are called *wavelets* and W_j are the *wavelet spaces* or *detail spaces*.

We can now make a number of observations about these spaces and basis functions.

First, the basis functions for the V_j spaces are actually all translated and dilated versions of one function. The same is true for the W_j spaces. In fact, we can write

$$\phi_{j,k}(x) \quad 2^{j/2}\phi(2^j x - k), \qquad \psi_{j,k}(x) \quad 2^{j/2}\psi(2^j x - k),$$

where

$$\phi(x) \quad \begin{cases} 1, & \text{if } 0 \le x \le 1, \\ 0, & \text{otherwise,} \end{cases} \qquad \psi(x) \quad \begin{cases} 1, & \text{if } 0 \le x \le 1/2, \\ -1, & \text{if } 1/2 < x \le 1, \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

[1] Note that this is the opposite convention compared to the tutorial by Lexing Ying in this volume, where more detailed functions belong to spaces with *smaller j*.

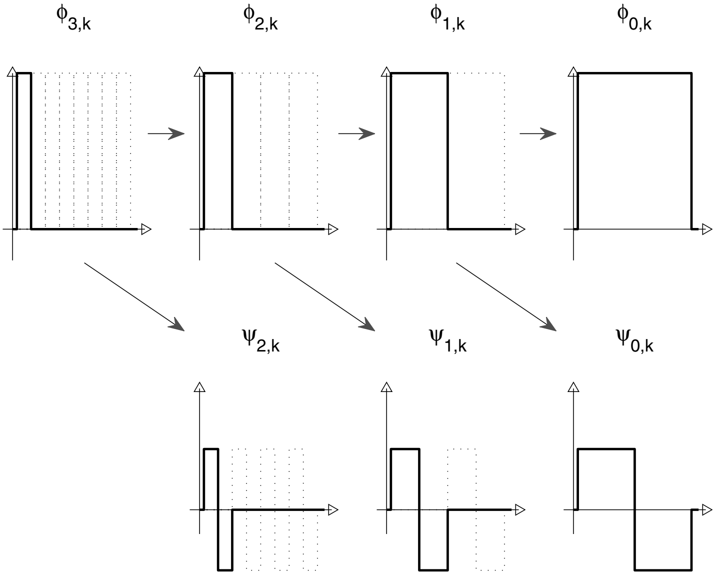


Fig. 2. Basis functions for V_j (above) and W_j (below).

The prefactor $2^{j/2}$ is somewhat arbitrary and chosen here to normalize the L^2 norm of the functions to unity. In wavelet theory $\phi(x)$ is called the *shape function* and $\psi(x)$ is called the *mother wavelet* of the wavelet system. They are shown in the top left frame of Fig. 3. We thus have that

$$\{\phi_{j,k};\ k \in \mathbb{Z}\} \text{ is an orthonormal basis for } V_j, \tag{2}$$

$$\{\psi_{j,k};\ k \in \mathbb{Z}\} \text{ is an orthonormal basis for } W_j. \tag{3}$$

The second observation is that each function u_{j+1} in V_{j+1} can be uniquely decomposed as a sum $u_j + d_j$ where $u_j \in V_j$ and $d_j \in W_j$. What is more, u_j and d_j are *orthogonal* in $L^2(\mathbb{R})$. This follows since the basis functions $\phi_{j,k}$ and $\psi_{j,k}$ are orthogonal in $L^2(\mathbb{R})$ for fixed j; it is easily verified that

$$\langle \phi_{j,k}, \psi_{j,k'} \rangle \quad \int \phi_{j,k}(x) \psi_{j,k'}(x) \quad 0, \qquad \forall j,k,k'.$$

This means that V_{j+1} is a direct sum of V_j and W_j, which are also orthogonal. Hence

$$V_{j+1} \quad V_j \oplus W_j, \qquad V_j \perp W_j. \tag{4}$$

The wavelet space W_j is thus the “difference” between two successive scaling spaces $V_j \subset V_{j+1}$. More precisely W_j is the *orthogonal complement* of V_j in V_{j+1} and in the decomposition of u_{j+1} above, u_j and d_j are the orthogonal projections of u_{j+1} onto V_j and W_j respectively.

We can draw a couple of conclusions from these observations. From (2), (3) and (4) it follows that

$$\{\psi_{j,k}\}_{k\in\mathbb{Z}} \cup \{\phi_{j,k}\}_{k\in\mathbb{Z}} \text{ is an orthonormal basis for } V_{j+1}.$$

Furthermore, iterating on (4), we see that, for any $j_0 < j$,

$$V_{j+1} \quad W_j \oplus W_{j-1} \oplus \cdots \oplus W_{j_0+1} \oplus W_{j_0} \oplus V_{j_0}.$$

If we formally let j tend to infinity in this sum it will contain functions with increasingly fine scales, and if we also let j_0 go to minus infinity it will be a sum over all wavelet spaces. In fact, one can show rigorously that

$$\overline{\bigoplus_{j\in\mathbb{Z}} W_j} \quad L^2(\mathbb{R}). \tag{5}$$

Moreover,

$$\{\psi_{j,k};\ j,k \in \mathbb{Z}\} \text{ is an orthonormal basis for } L^2(\mathbb{R}). \tag{6}$$

Hence, if we interpret each W_j as a space containing functions that only have one specific scale, $L^2(\mathbb{R})$, containing functions with all scales, is an infinite sum of the W_j spaces.

A third observation is that the scaling function has integral one,

$$\int \phi(x)dx \quad 1, \tag{7}$$

and the mother wavelet has zero mean,

$$\int \psi(x)dx \quad 0. \tag{8}$$

This turns out to be important basic requirements for all scaling and wavelet functions to guarantee basic regularity and approximation properties.

The construction above was introduced by Alfred Haar in the early 20th century [34], and it is now known as the Haar wavelet system. From a numerical point of view it has some nice properties, like the compact support and orthogonality of the basis functions, but also some drawbacks, mainly the fact that the basis functions are discontinuous. It took until the 1980s before a practical generalization of Haar's construction was made. Smoother wavelets with better numerical properties were then discovered. This was the start of the modern treatment of wavelets.

2.3 Multiresolution analysis

In the previous section we identified some of the core properties of the Haar wavelet system and described them in terms of functions spaces and orthonormal basis sets. We can now generalize the ideas illustrated in the initial example above. These generalizations will lead up to new types of wavelets with better properties than the simple Haar wavelets. To do this we use the concept of an (orthogonal) *Multiresolution Analysis* (MRA), first introduced by Meyer [47] and Mallat [45]. An MRA is a family of closed function spaces V_j and a real valued shape function $\phi(x)$ with the following properties:

(a) $\cdots \subset V_j \subset V_{j+1} \subset \cdots \subset L^2(\mathbb{R})$,

(b) $f(x) \in V_j \Leftrightarrow f(2x) \in V_{j+1}$,

(c) $\{\phi(x-k)\}_{k\in\mathbb{Z}}$ is an orthonormal basis for V_0,

(d) $\overline{\cup V_j} \quad L^2(\mathbb{R})$ and $\cap V_j \quad \{0\}$.

In general, it follows from (a) and (b) that V_j contains functions with finer and finer details when j increases. Property (c) provides the basis functions and (d) is a technical requirement to ensure completeness of the system.

These four properties are in fact all that is needed to form a wavelet system of the same type as in Sect. 2.2; the remaining parts of the wavelet systems can be constructed from $\phi(x)$ and the V_j spaces in the MRA as follows.

Like in the Haar case, we define $\phi_{j,k}(x) : \quad 2^{j/2}\phi(2^j\cancel{x} \quad k)$. Together (b) and (c) implies (2), that $\{\phi_{j,k}\}_{k\in\mathbb{Z}}$ is an orthonormal basis for V_j (see Exercise 1). One can also show that ϕ (or $-\phi$) satisfy (7). The wavelet spaces W_j are defined as the orthogonal complement of V_j in V_{j+1}. Then (4) and (5) follow using (d). Furthermore, there exists a mother wavelet $\psi(x)$ and it can be constructed in a fairly explicit way. Since $V_0 \subset V_1$ and $\{\phi_{1,k}\}$ is a basis for V_1 we can write

$$\phi(x) \quad \sum_k h_k \phi_{1,k}(x) \quad \sqrt{2}\sum_k h_k \phi(2x-k), \tag{9}$$

for some $\{h_k\}$ values. The scaling function thus satisfies a *refinement equation*. The $\{h_k\}$ values are called the (low pass) *filter coefficients* for the wavelet system. From these we compute the (high pass) filter coefficients $\{g_k\}$ by the *alternating flip* rule,

$$g_k \quad (-1)^k h_{1-k}, \tag{10}$$

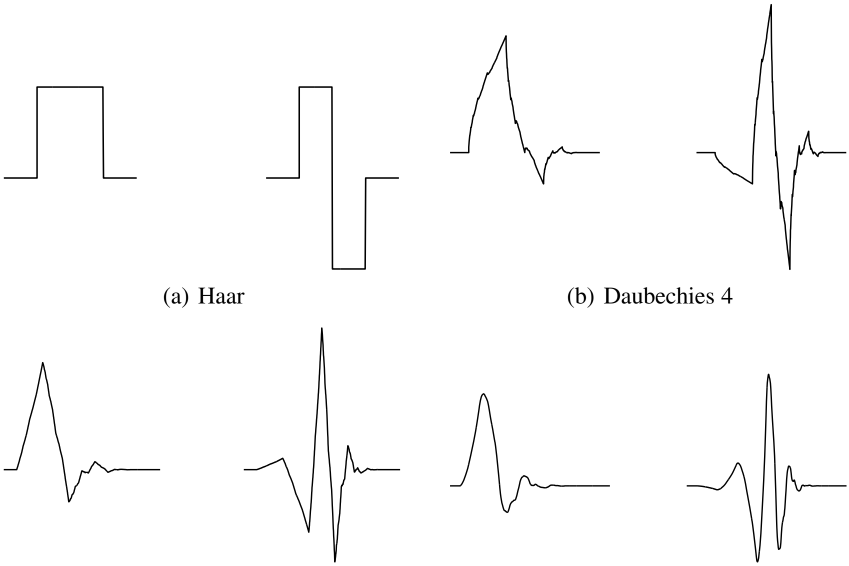
and define the mother wavelet as

$$\psi(x) : \quad \sum_k g_k \phi_{1,k}(x) \quad \sqrt{2}\sum_k (-1)^k h_{1-k}\phi(2x-k). \tag{11}$$

(Justification of this choice is given in Exercises 2 and 3 below.) Then (8) holds. Setting $\psi_{j,k} : \quad 2^{j/2}\psi(2^jx-k)$ also (3) holds and by (d) we also get (6). For a detailed description of MRA we refer the reader to the book by Daubechies [25].

The simplest MRA is the one built on the Haar basis that was introduced in the example above. In that case (a) is satisfied since the piecewise constant functions with piece length 2^{-j} is a subset of the piecewise constant functions with piece length $2^{-(j+1)}$. Moreover, (b) is true by construction and the shape function in (1) clearly satisfies (c). Finally, (d) holds since piecewise constant functions are dense in $L^2(\mathbb{R})$. The filter coefficients for Haar are

$$\{h_k\} \quad \left\{\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right\}, \qquad \{g_k\} \quad \left\{\frac{1}{\sqrt{2}}, \frac{-1}{\sqrt{2}}\right\}.$$



(a) Haar

(b) Daubechies 4

(c) Daubechies 6

(d) Daubechies 8

Fig. 3. Scaling functions $\phi(x)$ (left) and mother wavelets $\psi(x)$ (right) for the first families of Daubechies orthogonal compactly supported wavelets [24].

There are, however, many other examples. In fact, most wavelet systems fit into the MRA framework and it provides a good basis for constructing new wavelets. Although the first smooth wavelets were constructed earlier by Strömberg [52] and later by Meyer [46], Battle [5] and Lemarié [43], it was with the introduction of MRA that the construction of wavelets took off. New wavelets were designed to have desirable properties like orthogonality, compact support, smoothness, symmetry and vanishing moments (see Sect. 2.5). Not all those properties can be obtained at the same time, but it is for instance possible to have smooth orthonormal compactly supported wavelets, as shown by Daubechies [24]. These wavelets, pictured in Fig. 3, are the natural generalizations of the Haar wavelets; their shapes are, however, rather more unconventional. In Sect. 2.7 we will show how they can be derived. A multitude of other wavelets have been constructed since. We refer to [7] for more details and examples. Let us furthermore mention the early work on continuous wavelet transforms by Grossman and Morlet [33] who also coined the term "wavelet." Filter coefficients $\{h_k\}$ and $\{g_k\}$ for many wavelets are available in Matlab packages like Wavelet Toolbox, WaveLab and Wavekit; see also the classic book [25] for the coefficients of the earlier wavelets.

The basic wavelet theory that we focus on in this tutorial can be generalized in many ways. The orthogonality restriction in the MRA can be relaxed and wavelet systems where the basis functions are not orthogonal can be built. One important example is *bi-orthogonal* wavelets [18] in which some orthogonality properties are maintained but not all. This gives more freedom in the construction and it is for example possible to build compactly supported *symmetric* wavelets in this way, which is not possible for strictly orthogonal systems. In *semiorthogonal* wavelets [16] even more orthogonality properties are kept. *Wavelet packages* have an improved frequency localization compared to standard wavelets, obtained by tiling the time-frequency plane (see Sect. 2.6) in a way better adapted to the function at hand [21]. In *multiwavelets* several scaling functions and mother wavelets are used to build a wavelet system [1, 31]. For higher dimensions a number of wavelet like systems has been constructed where, the basis functions are indexed not just by localization and scale, but also orientation; *curvelets* [13] and *ridgelets* [12] are examples of this approach. A significant generalization of the basic wavelet theory are the *second generation* wavelets put forth by Sweldens [53]. The core ideas of multiresolution analysis are kept but the dilation/translation structure and the corresponding dyadic intervals, are replaced by general nested subsets that partition the space. This allows for the construction of wavelets in a broader range of settings, for instance on irregular grids (e.g. triangulations) and on surfaces in 3D. Wavelets of this type (and also first generation wavelets) can be built naturally using Swelden's systematic *lifting* algorithm.

Remark 1. In more than one dimension, the MRA is extended by considering tensor product spaces. In two dimensions, we set

$$\boldsymbol{V}_j \quad V_j \otimes V_j, \qquad \ldots \subset \boldsymbol{V}_j \subset \boldsymbol{V}_{j+1} \subset \ldots \subset L^2(\mathbb{R}^2) \qquad j \in \mathbb{Z},$$

with V_j being the one-dimensional spaces introduced above. As before, we let the wavelet spaces $\boldsymbol{W}_j$ be the orthogonal complements of $\boldsymbol{V}_j$ in $\boldsymbol{V}_{j+1}$, so that $\boldsymbol{V}_{j+1} \quad \boldsymbol{V}_j \oplus \boldsymbol{W}_j$ and $\boldsymbol{V}_j \perp \boldsymbol{W}_j$. In this case $\boldsymbol{W}_j$ is composed of three parts,

$$\boldsymbol{W}_j \quad (W_j \otimes W_j) \oplus (V_j \otimes W_j) \oplus (W_j \otimes V_j),$$

where, the W_j spaces are those of the one-dimensional case, given by (4). Similar tensor product extensions can be made also for higher dimensions.

Exercise 1. Suppose that V_j and $\phi(x)$ are the scaling spaces and scaling function for a MRA. Use property (b) and (c) of the MRA to show that $\{\phi_{j,k}(x) \quad 2^{j/2}\phi(2^j x - k)\}$ with $k \in \mathbb{Z}$ is an orthonormal basis for V_j.

2.4 Filter coefficients

The filter coefficients encode and generalize the idea of taking local averages (via $\{h_k\}$) and differences, (via $\{g_k\}$) in each refinement step, described for the case of Haar wavelets in the introduction. The $\{h_k\}$ coefficients can actually be used to directly characterize the wavelet system. In principle they define the shape function via (9) and then the mother wavelet via (10) and (11). Not every $\{h_k\}$ sequence is allowed, however. By (9) a basic requirement is that $\{h_k\}$ is square summable, i.e. belong to ℓ^2. In fact $\sum_k |h_k|^2 \quad 1$ since $||\phi||_{L^2} \quad 1$. In order to define reasonable shape functions and mother wavelets one needs to restrict $\{h_k\}$ further. Here we assume that $|h_k|$ decays sufficiently fast as $|k| \to \infty$ to allow the manipulations we do. We then derive two necessary conditions. First, after integrating (9) we get that

$$\int \phi(x)dx = \sqrt{2}\sum_k h_k \int \phi(2x-k)dx = \frac{1}{\sqrt{2}}\int \phi(x)dx \sum_k h_k,$$

and that

$$\begin{aligned}\int \phi(x)\phi(x-k)dx &= 2\sum_{m,n} h_m h_n \int \phi(2x-m)\phi(2x-2k-n)dx \\ &= \sum_{m,n} h_m h_n \int \phi(x)\phi(x+m-2k-n)dx.\end{aligned}$$

Together with (2) and (7) we get the necessary conditions on $\{h_k\}$

$$\sum_k h_k = \sqrt{2}, \tag{12}$$

$$\sum_n h_{n+2k} h_n = \delta_k. \tag{13}$$

An alternative characterization of the filter coefficients is given by introducing the Fourier transform of $\{h_k\}$, in this context also known as the *generating function* of $\{h_k\}$,

$$\hat{h}(\xi) := \frac{1}{\sqrt{2}}\sum_k h_k e^{ik\xi}. \tag{14}$$

The requirement (12) is then simply written as

$$\hat{h}(0) = 1. \tag{15}$$

To translate the requirement (13) we note that

$$\begin{aligned}|\hat{h}(\xi)|^2 + |\hat{h}(\xi+\pi)|^2 &= \frac{1}{2}\sum_{n,k}\left(h_n e^{in\xi} h_k e^{-ik\xi} + h_n e^{in(\xi+\pi)} h_k e^{-ik(\xi+\pi)}\right) \\ &= \frac{1}{2}\sum_{n,k} h_n h_k e^{i(n-k)\xi}\left(1 + e^{i(n-k)\pi}\right) \\ &= \frac{1}{2}\sum_{n,k} h_n h_{n+k} e^{-ik\xi}\left(1 + (-1)^k\right) \\ &= \sum_{n,k} h_n h_{n+2k} e^{-i2k\xi} = \sum_k e^{-i2k\xi}\left(\sum_n h_n h_{n+2k}\right).\end{aligned}$$

It follows that (13) is equivalent to

$$|\hat{h}(\xi)|^2 + |\hat{h}(\xi+\pi)|^2 = 1. \tag{16}$$

Thus to find new orthogonal wavelet systems one must first find $\{h_k\}$ sequences satisfying (12) and (13), or alternatively find a 2π-periodic function $\hat{h}(\xi) \in L^2(\ 0,2\pi\ )$ satisfying (15) and (16). To define an MRA $\{h_k\}$ also need to satisfy some technical requirements that guarantees convergence. For instance, a sufficient condition is that there exist $\varepsilon > 0$ such that

$$\sum_k |h_k||k|^{\varepsilon} < \infty, \qquad \text{and} \qquad \hat{h}(\xi)\ /\ 0, \quad \forall |\xi| \leq \frac{\pi}{2}. \tag{17}$$

In this case we can construct an MRA from $\{h_k\}$, defining $\phi(x)$ from (9) and V_j as the closure of the linear span of the corresponding $\{\phi_{j,k}\}$, see [24, 45]. As before, the $\{g_k\}$ coefficients are defined by (10) and $\psi(x)$ by (11).

Exercise 2. Show that if $\psi_{j,k}(x)$ is orthogonal to $\phi_{j,k'}(x)$ for all j,k,k' a necessary condition for the $\{g_k\}$ coefficients is that

$$\sum_n h_{n+2k} g_n \quad 0, \qquad \forall k.$$

Also show that with the choice $g_n \quad (-1)^n h_{1-n}$ this condition is satisfied. (Hint: Let $a_n \quad h_{n+2k} g_n$ and note that then $a_{1-n-2k} \quad -a_n$.)

Exercise 3. Show that (7) and (8) implies that

$$\sum_k g_k \quad 0.$$

Moreover, show that this is satisfied if (10) holds and if $\{h_k\}$ satisfy (12) and (13). (Hint: Define $\hat{g}(\xi)$ from $\{g_k\}$ in the same was as $\hat{h}(\xi)$ was defined from $\{h_k\}$. Relate $\hat{g}(\xi)$ to $\hat{h}(\xi)$. Then use (15) and (16).)

2.5 Approximation properties

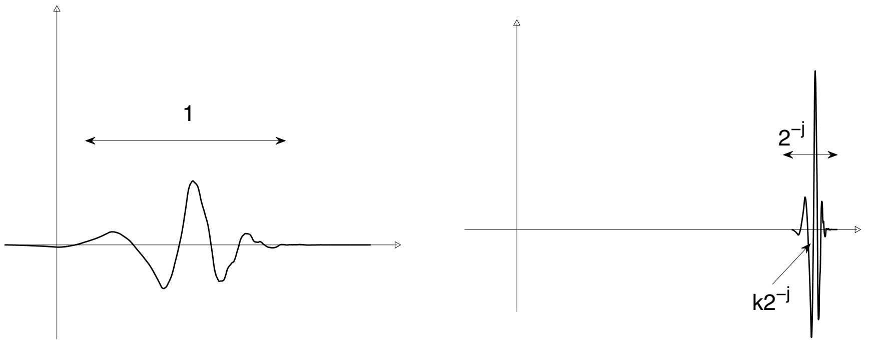
We have already mentioned the good approximation properties of wavelets. Let us now explain this in more detail and make the statement more precise. The basic approximation mechanism relies on the mother wavelet having *vanishing moments*. This is a generalization of (8) defined as follows.

Definition 1. *The mother wavelet ψ has M vanishing moments if*

$$\int \psi(x) x^m dx \quad 0, \qquad m \quad 0,\ldots,M-1. \tag{18}$$

We note that this implies that also $\psi_{j,k}(x)$ has M vanishing moments (see Exercise 4).

Another important property is the space localization of the wavelets. By this we mean that most of the energy of the wavelet concentrates in a small domain. This follows automatically from the construction if the mother wavelet decays rapidly at infinity, $|\psi(x)| \to 0$ as $|x| \to \infty$. For instance, if $\psi(x)$ has compact support in $\ -a,a\ $



(a) $\psi(x)$ (b) $\psi_{j,k}(x)$

Fig. 4. Examples of translations and dilations of the Daubechies 8 mother wavelet $\psi(x)$. For $\psi_{j,k}(x)$ the localization in space changes to around $k2^{-j}$ and the support narrows to size 2^{-j}.

then from the translation and dilation, $\psi_{j,k} \quad 2^{j/2}\psi(2^j x - k)$ has compact support in $\ k2^{-j} - 2^{-j}a,\ k2^{-j} + 2^{-j}a\ $; see Fig. 4. We note here that compactly supported wavelet systems with arbitrary many vanishing moments do indeed exist [24]. We will sketch how these can be derived below in Sect. 2.7.

To see the importance of the vanishing moment and space localization properties we start from a function $u(x)$ which has the wavelet expansion

$$u(x) \quad \sum_{j,k} u_{j,k}\psi_{j,k}(x),$$

where we suppose that the mother wavelet has compact support in $\ -a, a\ $. Let us focus on an individual coefficient $u_{j,k}$. We denote by Ω the (compact) support of $\psi_{j,k}$. Then,

$$u_{j,k} \quad \int_\Omega u(x)\psi_{j,k}(x)dx.$$

Suppose the restriction of $u(x)$ to Ω is in $C^p(\Omega)$, i.e. in the support of $\psi_{j,k}(x)$ we assume that $u(x)$ is smooth with p continuous derivatives. We can then Taylor expand $u(x)$ around a point $x_0 \in \Omega$. Let $\tilde{M} \quad \min(p, M)$. We get

$$u_{j,k} \quad \int_\Omega \sum_{m \quad 0}^{\tilde{M}-1} u^{(m)}(x_0)\frac{(x-x_0)^m}{m!}\psi_{j,k}(x)dx + \int_\Omega R(x)\frac{(x-x_0)^{\tilde{M}}}{\tilde{M}!}\psi_{j,k}(x)dx,$$

where $R(x)$ is the Taylor remainder term, satisfying

$$\sup_{x\in\Omega}|R(x)| \leq \sup_{x\in\Omega}|u^{(\tilde{M})}(x)|.$$

Since $\tilde{M} \leq M$ the first term is zero because of the vanishing moments of $\psi_{j,k}(x)$. Moreover, noting that by the Cauchy-Schwarz inequality,

$$\int_{\Omega} |\psi_{j,k}(x)| dx \leq \left(\int_{\Omega} 1^2 dx \times \int_{\Omega} |\psi_{j,k}(x)|^2 dx \right)^{1/2} \quad |\Omega|^{1/2},$$

we obtain

$$|u_{j,k}| \leq \frac{|\Omega|^{\tilde{M}}}{\tilde{M}!} \sup_{x\in\Omega} \left| u^{(\tilde{M})}(x) \right| \int_{\Omega} |\psi_{j,k}(x)| dx \leq \frac{|\Omega|^{\tilde{M}+1/2}}{\tilde{M}!} \sup_{x\in\Omega} \left| u^{(\tilde{M})}(x) \right| .$$

From the discussion above we see that $|\Omega| \quad 2^{-j+1}a$ and we finally have

$$|u_{j,k}| \leq c 2^{-j(\min(p,M)+1/2)} \sup_{x\in\Omega} \left| u^{(\min(p,M))}(x) \right| ,$$

for some constant c that is independent of j.

From this we can conclude that as long as $u(x)$ is smooth in the support of $\psi_{j,k}$, i.e. p is large, then the wavelet coefficients $u_{j,k}$ decay rapidly with j when the number of vanishing moments is large. The coefficients $u_{j,k}$ may, however, be large when p is small, hence at the points where $u(x)$ is non-smooth. For piecewise smooth functions this only happens in a few places. Thus most fine-scale coefficients are very small and can be neglected. One only needs to keep those where $u(x)$ changes sharply or is discontinuous. In this sense wavelets are good at approximating also piecewise smooth functions, in contrast to Fourier bases which exhibit large Gibbs type errors when the approximated function is non-smooth. This a reason for the success of wavelets in compression of e.g. images which can be modeled as two-dimensional piecewise smooth functions.

Exercise 4. Show that if the mother wavelet $\psi(x)$ has M vanishing moments satisfying (18), then the same is true for every wavelet basis function $\psi_{j,k}(x) \quad 2^{j/2}\psi(2^j x - k)$.

2.6 Time-frequency analysis

We will here discuss another particular property of wavelet bases, namely their localization in time and frequency. We are interested in what parts of the time-frequency (t,ω)-plane[2] that is covered by the basis functions, hence the essential support of $|v(t)\hat{v}(\omega)|$ for the function v. Let us consider a function

$$u(t) \quad \sum_n u_n v_n(t),$$

and compare representations with three different types of basis functions $\{v_n(t)\}$. The basis functions and their essential time-frequency localizations are illustrated in Fig. 6.

[2] Since we interpret the independent variable as time here, we denote it by t instead of x.

1. Time representation
   In this case the basis functions are the set of translated (approximate) delta functions, $v_n(t) \approx \delta(t - t_n)$ with Fourier transform $\hat{v}_n(\omega) \approx e^{-i\omega t_n}$. Thus the support of $|v_n(t)\hat{v}_n(\omega)|$ is concentrated around $t = t_n$ and essentially spread out over all frequencies. We can interpret this as total localization in space and no localization in frequency, cf. the left frames of Fig. 6.
2. Frequency representation
   In this case we use Fourier basis, $v_n(t) = e^{i\omega_n t}$, with Fourier transform $\hat{v}_n(\omega) = \delta(\omega - \omega_n)$. This is the opposite situation to the time representation case, and the support of $|v_n(t)\hat{v}_n(\omega)|$ is concentrated around $\omega = \omega_n$ and essentially spread out over all time, i.e. total localization in frequency and no localization in time, as indicated by the middle frames of Fig. 6.
3. Wavelet representation
   Here we have $v_n(t) = \psi_{j,k}(t)$. As already discussed in Sect. 2.5 the support of the wavelet function $\psi_{j,k}(t)$ is centered around $k2^{-j}$ and has width proportional to 2^{-j}. The Fourier transform of the wavelets is

   $$\hat{\psi}_{j,k}(\omega) = \int 2^{j/2}\psi(2^j t - k)e^{-i\omega t}dt = 2^{-j/2}e^{-i2^{-j}\omega k}\hat{\psi}\left(2^{-j}\omega\right).$$

   We note furthermore that $\hat{\psi}^{(p)}(\omega) = \widehat{(it)^p\psi(t)}$ and if the mother wavelet has M vanishing moments, then

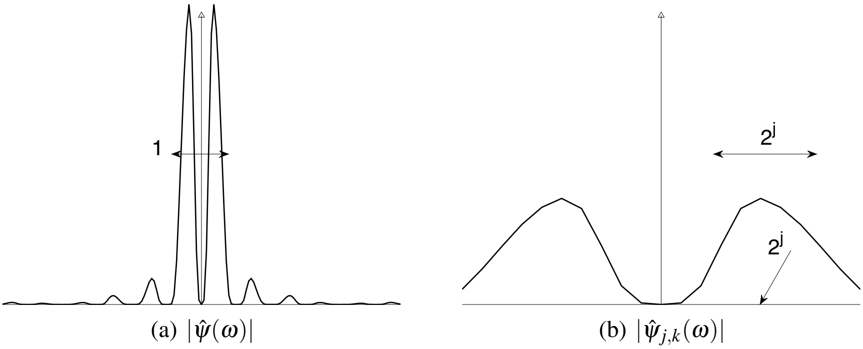
   $$\hat{\psi}^{(p)}(0) = \int (it)^p\psi(t)dt = 0, \qquad p = 0,\ldots,M-1. \tag{19}$$

   Hence, $\hat{\psi}(\omega)$ thus has an M-th order zero at $\omega = 0$. Together this means that the support of $|\hat{\psi}_{j,k}(\omega)|$ is essentially centered around $|\omega| = 2^j$ in an interval of size 2^j, see Fig. 5. The wavelets are thus localized both in time and frequency, with the essential support of $|\psi_{j,k}(t)\hat{\psi}_{j,k}(\omega)|$ centered at $(t,\omega) \sim (k2^{-j}, 2^j)$; see the right frames of Fig. 6.

The three different representations thus divide the time-frequency plane in quite different ways. One way to interpret wavelets is as a limited musical score, where each basis function, like a musical note, corresponds to a particular localization both in time and space. Figure 7 shows an example decomposition of a function using the different representations. The localized wave package at the left is reflected in the size of the wavelet coefficients.

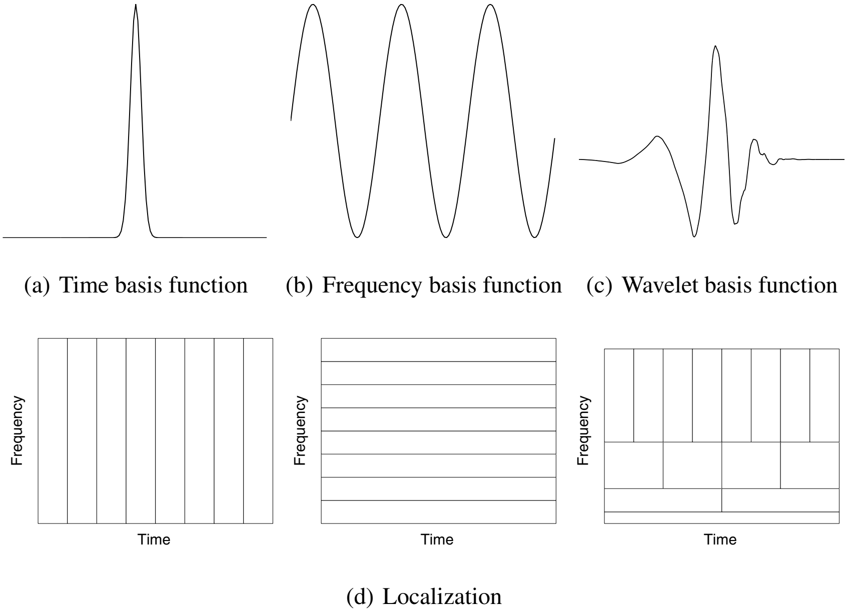
2.7 Orthogonal compactly supported wavelets

In this section we show how smooth orthogonal compactly supported wavelets with many vanishing moments can be constructed. These were first discovered by Daubechies [24]. We essentially follow her derivation, which is done directly from the generating function (14) that characterizes the wavelet system. We are thus looking for a $\hat{h}(\xi)$ that satisfies (15) and (16), as well as (17).



(a) $|\hat{\psi}(\omega)|$ (b) $|\hat{\psi}_{j,k}(\omega)|$

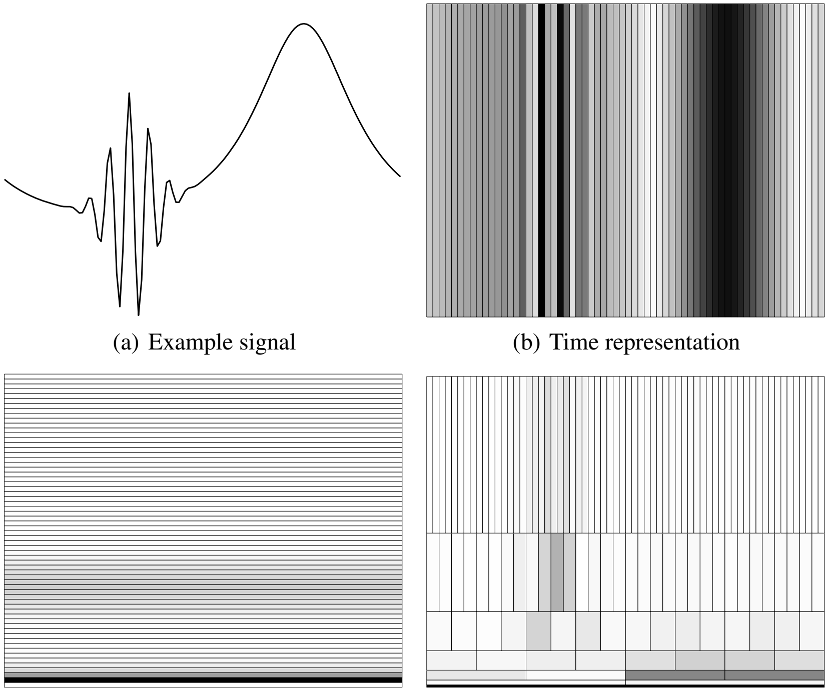
Fig. 5. Examples of the effect of translations and dilations on the Fourier transform of the Daubechies 8 mother wavelet $\hat{\psi}(\omega)$. For $\hat{\psi}_{j,k}(\omega)$ the localization in frequency changes to around 2^j and the support increases to size 2^j.



(a) Time basis function (b) Frequency basis function (c) Wavelet basis function

(d) Localization

Fig. 6. Basis functions and time-frequency localizations



(a) Example signal

(b) Time representation

(c) Fourier representation

(d) Wavelet representation

Fig. 7. Different representations of the example signal in (a): Localization in time (b), frequency (c), time and frequency (d). Gray levels indicate the size of the corresponding coefficient.

Compactly supported wavelets are equivalent to finite length or finite impulse response (FIR) filters, where all but a finite number of $\{g_k\}$ and $\{h_k\}$ coefficients are zero. Moreover, we note that if $\hat{h}(\xi)$ satisfies (15) and (16), then so does $e^{im\xi}\hat{h}(\xi)$, for any real m. This means that we need to find $\hat{h}(\xi)$ of the form

$$\hat{h}(\xi) = \frac{1}{\sqrt{2}}\sum_{k=0}^{N} h_k e^{ik\xi} = \frac{1}{\sqrt{2}}\sum_{k=0}^{N} h_k z^k =: P(z), \qquad z = z(\xi) = e^{i\xi},$$

where $P(z)$ is a *polynomial* of degree N (to be determined) with real coefficients.

By taking the Fourier transform of (11) we obtain

$$\hat{\psi}(2\xi) = \frac{1}{\sqrt{2}}\sum_{k}(-1)^k h_{1-k} e^{ik\xi}\hat{\phi}(\xi) = e^{-i(\xi+\pi)}\hat{h}(\xi+\pi)\hat{\phi}(\xi) = -\frac{P(-z)\hat{\phi}(\xi)}{z}. \tag{20}$$

When $\psi(x)$ has M vanishing moments (19) holds. It follows (see Exercise 5) that $P(z)$ has an M-th order zero at $z = -1$ and we can write

$$P(z) = \left(\frac{1+z}{2}\right)^M Q(z), \tag{21}$$

for some polynomial $Q(z)$ of degree $N-M$ with real coefficients. Conditions (15) and (16) for $\hat{h}(\xi)$ is then equivalent to the following conditions for $Q(z)$:

$$Q(1) = 1, \tag{22}$$

$$\left|\frac{1+z}{2}\right|^{2M} |Q(z)|^2 + \left|\frac{1-z}{2}\right|^{2M} |Q(-z)|^2 = 1, \qquad \forall |z| = 1. \tag{23}$$

Let $c_{k,n}$ be the binomial coefficients $\binom{n}{k}$ and define the n-th degree real polynomial $T_n(y)$ as follows:

$$\begin{aligned}
1 &= (y+1-y)^{2n+1} = \sum_{k=0}^{2n+1} c_{k,2n+1} y^{2n+1-k}(1-y)^k \\
&= y^{n+1}\sum_{k=0}^{n} c_{k,2n+1} y^{n-k}(1-y)^k + (1-y)^{n+1}\sum_{k=0}^{n} c_{k+n+1,2n+1} y^{n-k}(1-y)^k \\
&= y^{n+1}\sum_{k=0}^{n} c_{k,2n+1} y^{n-k}(1-y)^k + (1-y)^{n+1}\sum_{k=0}^{n} c_{k,2n+1}(1-y)^{n-k} y^k \\
&=: y^{n+1} T_n(1-y) + (1-y)^{n+1} T_n(y).
\end{aligned}$$

Here we used the symmetry $c_{k,2n+1} = c_{2n+1-k,2n+1}$. We note that

$$T_n(0) = c_{0,2n+1} = 1, \quad \text{and} \quad T_n(y) > 0, \quad y \in [0,1], \tag{24}$$

since all coefficients $c_{k,n}$ are positive. Moreover,

$$\left|\frac{1+z}{2}\right|^2 + \left|\frac{1-z}{2}\right|^2 = 1, \qquad \forall |z| = 1.$$

Therefore, if we can find a polynomial $Q(z)$ with real coefficients such that

$$|Q(z)|^2 = T_{M-1}\left(\left|\frac{1-z}{2}\right|^2\right), \qquad \forall |z| = 1,$$

it will satisfy (22) and (23) above. Let us now prove that this is always possible. Denote by $S_n(z)$ the n-th Chebyshev polynomial, which satisfies the trigonometric identity $S_n(\cos\theta) = \cos(n\theta)$. Since $\{S_k(z)\}$ are linearly independent, there are coefficients $\{\alpha_k\}$ such that

$$\begin{aligned}
T_n\left(\left|\frac{1-z}{2}\right|^2\right) &= T_n\left(\frac{1-\cos\xi}{2}\right) = \sum_{k=0}^{n} \alpha_k S_k(\cos\xi) = \sum_{k=0}^{n} \alpha_k \cos(k\xi) \\
&= \frac{1}{2}\sum_{k=0}^{n} \alpha_k \left(e^{ik\xi} + e^{-ik\xi}\right) = \frac{e^{-in\xi}}{2}\sum_{k=0}^{n} \alpha_k \left(z^{n+k} + z^{n-k}\right) =: e^{-in\xi}\tilde{T}_n(z),
\end{aligned}$$

when $|z| = 1$. This defines the $2n$-th degree polynomial $\tilde{T}_n(z)$ with real coefficients. A key property of $\tilde{T}_n(z)$ follows from (24),

$$T_n\left(\left|\frac{1-z}{2}\right|^2\right) = \left|T_n\left(\left|\frac{1-z}{2}\right|^2\right)\right| = |\tilde{T}_n(z)|, \qquad \forall |z| = 1. \tag{25}$$

We thus need to find a $Q(z)$ that is essentially the square root of $\tilde{T}_{M-1}(z)$.

Whenever $z \neq 0$ we have the identity

$$\tilde{T}_n(1/z) = \frac{1}{2}\sum_{k=0}^{n} \alpha_k \left(z^{-n-k} + z^{-n+k}\right) = \frac{z^{-2n}}{2}\sum_{k=0}^{n} \alpha_k \left(z^{n-k} + z^{n+k}\right) = z^{-2n}\tilde{T}_n(z).$$

There is no root at the origin[3] and therefore, if z_* is a root of $\tilde{T}_n$, then so is $1/z_*$. Moreover, since $\{\alpha_k\}$ are real, $\tilde{T}_n(\bar{z}) = \overline{\tilde{T}_n(z)}$, so that if z_* is a root, then also $\bar{z}_*$ is a root. Thus, roots come in groups of four, $\{z_*,\ 1/z_*,\ \bar{z}_*,\ 1/\bar{z}_*\}$ if z_* is complex, and in pairs $\{z_*,\ 1/z_*\}$ if z_* is real. Let $\{z_k\}$ define the n_c groups of complex roots and $\{r_k\}$ the n_r pairs of real roots, with $4n_c + 2n_r = 2n$. Then

$$\tilde{T}_n(z) = \frac{\alpha_n}{2}\left(\prod_{k=1}^{n_r}(z-r_k)(z-1/r_k)\right)\left(\prod_{k=1}^{n_c}(z-z_k)(z-\bar{z}_k)(z-1/z_k)(z-1/\bar{z}_k)\right).$$

We now observe that when $|z| = 1$, then $|z - 1/\bar{z}_*| = |z - z_*|/|z_*|$. Using (25) we get

$$T_n\left(\left|\frac{1-z}{2}\right|^2\right) = \frac{|\alpha_n|}{2}\left(\prod_{k=1}^{n_r}\frac{|z-r_k|^2}{|r_k|}\right)\left(\prod_{k=1}^{n_c}\frac{|z-z_k|^2|z-\bar{z}_k|^2}{|z_k|^2}\right) = |Q(z)|^2,$$

for all $|z| = 1$, where

$$Q(z) = q_0\left(\prod_{k=1}^{n_r}(z-r_k)\right)\left(\prod_{k=1}^{n_c}(z-z_k)(z-\bar{z}_k)\right),$$

and

$$q_0 = \pm\left[\frac{|\alpha_n|}{2}\left(\prod_{k=1}^{n_r}\frac{1}{|r_k|}\right)\right]^{1/2}\left(\prod_{k=1}^{n_c}\frac{1}{|z_k|}\right),$$

with the sign chosen such that $Q(1) = +1$. Finally, since

$$(z-z_k)(z-\bar{z}_k) = z^2 - 2z\,\Re z_k + |z_k|^2,$$

the coefficients of $Q(z)$ are real. The degree is $n_r + 2n_c = n$. Applying the construction to $\tilde{T}_{M-1}(z)$ we get the desired $Q(z)$. The degree of the polynomial $P(z)$ in (21) is $N = 2M-1$, and thus the number of non-zero filter coefficients is $2M$.

[3] It can be verified that $T_n(y)$ is indeed of precisely degree n, and not less. Furthermore, since the degree of $S_k(z)$ is precisely k it follows from the construction that $0 \neq \alpha_n = 2\tilde{T}_n(0)$.

It remains to verify the technical condition (17). The left sum is trivially bounded since there are only N non-zero h_k. The right condition translates to $P(z) \;/\; 0$ for $|z| \quad 1$ and $\Re z \geq 0$. By (21) it is enough that this holds for $|Q(z)|^2$. This in turn means that we need $T_{M-1}(y) \;/\; 0$ for $y \in \; 0, 1/2\,$ which is ensured by (24).

In conclusion, for any M we can explicitly construct a $Q(z)$ satisfying (22) and (23) by following the steps above. This gives us the desired $P(z)$ and then $\hat{h}(\xi)$. In fact, there are many possible $Q(z)$ since the choice of z_k and r_k within the root groups is arbitrary. These lead to different wavelet systems. We end this section with a couple of examples.

Example 1. When $M \quad 1$ we have simply that $T_0 \quad \tilde{T}_0 \quad Q \equiv 1$. Then

$$\hat{h}(\xi) \quad \frac{1+e^{i\xi}}{2} \quad \Rightarrow \quad h_0 \quad h_1 \quad 1/\sqrt{2},$$

which gives the Haar wavelets.

Example 2. When $M \quad 2$ we get $T_1(y) \quad 1+2y$ and

$$\tilde{T}_1(z) \quad z\left(2-\frac{z+1/z}{2}\right) \quad -\frac{1}{2}+2z-\frac{1}{2}z^2 \quad -\frac{1}{2}(z-2-\sqrt{3})(z-2+\sqrt{3}).$$

Setting $Q(z) \quad q_0(z-2-\sqrt{3}) \quad : a+bz$, we find that

$$a \quad -q_0(2+\sqrt{3}) \quad \frac{1+\sqrt{3}}{2}, \quad b \quad q_0 \quad \frac{1-\sqrt{3}}{2}, \qquad q_0 \quad \frac{-1}{\sqrt{2(2+\sqrt{3})}}.$$

This leads to

$$\hat{h}(\xi) \quad \left(\frac{1+e^{i\xi}}{2}\right)\left(a+be^{i\xi}\right) \quad \frac{a}{4}+\frac{2a+b}{4}e^{i\xi}+\frac{2b+a}{4}e^{i2\xi}+\frac{b}{4}e^{i3\xi},$$

or

$$h_0 \quad \frac{a}{2\sqrt{2}}, \quad h_1 \quad \frac{2a+b}{2\sqrt{2}}, \quad h_2 \quad \frac{2b+a}{2\sqrt{2}}, \quad h_3 \quad \frac{b}{2\sqrt{2}}.$$

These are the filter coefficients for the Daubechies 4 wavelets pictured in the top right frame of Fig. 3.

Exercise 5. Show that (19) and (20) implies (21).
Hint: First show that (19) and (20) implies $\hat{h}^{(p)}(\pi) \quad 0$ for $p \quad 0,\dots,M-1$. Then verify that when $p > 0$,

$$\hat{h}^{(p)}(\xi) \quad i^p \sum_{k \quad 1}^{p} d_{k,p} z^k P^{(k)}(z),$$

for some coefficients $d_{k,p}$ with $d_{p,p} \;/\; 0$. Conclude from this that (21) must hold.

2.8 Wavelet transforms

We now go through how the decomposition of a function is done in the discrete case using the filter coefficients. In numerical computations we would like to avoid having to evaluate the complicated functions $\phi_{j,k}(x)$ and $\psi_{j,k}(x)$. This is very often possible.

For sufficiently fine detail level we can approximate the scaling coefficients by sample values. In fact, if $u(x) \in C^1(\mathbb{R})$ and $\phi(x)$ has compact support, then one can show from (7) that

$$u_{j,k} \quad \int u(x)\phi_{j,k}(x)dx \quad 2^{-j/2}u\left(k2^{-j}\right)+O\left(2^{-3j/2}\right). \tag{26}$$

In applications one typically identifies $2^{j/2}u_{j,k}$ precisely with the sample values.

Suppose now that we have approximated a function in this way and have

$$u(x) \quad \sum_k u_{j+1,k}\phi_{j+1,k}(x) \in V_{j+1}.$$

We want to decompose u into its coarse scale part in V_j and its fine scale part in W_j,

$$u(x) \quad \underbrace{\sum_k u^{\mathrm{c}}_{j,k}\phi_{j,k}(x)}_{\in V_j}+\underbrace{\sum_k u^{\mathrm{f}}_{j,k}\psi_{j,k}(x)}_{\in W_j}. \tag{27}$$

We thus want to find the coefficients $\{u^{\mathrm{c}}_{j,k}\}$ and $\{u^{\mathrm{f}}_{j,k}\}$, where the superscripts signify ”coarse” and ”fine” scale, respectively. We have

$$\begin{aligned} u^{\mathrm{c}}_{j,k} \quad & \int u(x)\phi_{j,k}(x)dx \quad 2^{j/2}\int u(x)\phi(2^jx-k)dx \\ & \sum_\ell h_\ell 2^{(j+1)/2}\int u(x)\phi(2^{j+1}x-2k-\ell)dx \\ & \sum_\ell h_\ell\langle u,\phi_{j+1,\ell+2k}\rangle \quad \sum_\ell h_\ell u_{j+1,\ell+2k} \quad \sum_\ell h_{\ell-2k}u_{j+1,\ell}. \end{aligned}$$

We get a similar expression for $u^{\mathrm{f}}_{j,k}$ with h_k replaced by g_k,

$$u^{\mathrm{c}}_{j,k} \quad \sum_\ell h_{\ell-2k}u_{j+1,\ell}, \qquad u^{\mathrm{f}}_{j,k} \quad \sum_\ell g_{\ell-2k}u_{j+1,\ell}. \tag{28}$$

Hence, the transform $\{u_{j+1,k}\} \to \{u^{\mathrm{c}}_{j,k}\} \cup \{u^{\mathrm{f}}_{j,k}\}$, i.e. the decomposition (27), can be done by only using the filter coefficients h_k and g_k. We do not need to involve the functions $\phi_{j,k}(x)$ and $\psi_{j,k}(x)$. Moreover, if the wavelets are compactly supported, there are only a finite number of non-zero $\{h_k\}$ and $\{g_k\}$ coefficients so the transform can be done very quickly with just a few multiplications and additions per coefficient, the number being independent of j.

Numerically we can only consider a finite description. The standard setting is that we assume $u(x) \in V_{j+1}$ and that we have $N \quad 2^{j+1}$ scaling coefficients that correspond to sample values in a compact interval, say $0,1$. From this we want to decompose $u(x)$ into $N-1$ wavelet coefficients for W_j and one scaling coefficient for V_0, still corresponding to the interval $0,1$. We consider the Haar case as an example. Let $U_{j+1}\quad\{u_{j+1,k}\}$ be the vector of length 2^{j+1} containing the scaling coefficients for V_{j+1}. Similarly, we set $U_j^{\mathrm{f}}\quad\{u_{j,k}^{\mathrm{f}}\}$ and $U_j^{\mathrm{c}}\quad\{u_{j,k}^{\mathrm{c}}\}$. With a slight abuse of notation we will write that the coefficient vector U_{j+1} belongs to V_{j+1}, etc. noting that in this restricted setting, in fact, both V_j and W_j are isomorphic to $\mathbb{R}^j$. We can then write (28) in matrix form,

$$\mathscr{W}_j U_{j+1}\quad\begin{pmatrix}U_j^{\mathrm{f}}\\U_j^{\mathrm{c}}\end{pmatrix},\qquad U_{j+1}\in V_{j+1}\quad U_j^{\mathrm{f}}\in W_j,\quad U_j^{\mathrm{c}}\in V_j,\tag{29}$$

where

$$\mathscr{W}_j\quad\frac{1}{\sqrt{2}}\begin{pmatrix}1&-1&0&\cdots&&\\0&0&1&-1&0&\cdots\\\vdots&\vdots&&\ddots&\ddots&\\0&0&\cdots&0&1&-1\\1&1&0&\cdots&&\\0&0&1&1&0&\cdots\\\vdots&\vdots&&\ddots&\ddots&\\0&0&\cdots&0&1&1\end{pmatrix}\in\mathbb{R}^{2^{j+1}\times 2^{j+1}}.\tag{30}$$

With a simple matrix vector multiplication we can thus extract the coarse and fine part of $u(x)$. The matrix $\mathscr{W}_j$ is called the wavelet transform for level j. We note that it is *sparse* and *orthonormal*, $\mathscr{W}_j^T\mathscr{W}_j\quad I$. The multiplication is therefore stable and costs $\mathscr{O}(2^{j+1})$ operations.

From this starting point we can now decompose $u(x)$ into a full wavelet representation similar to the example in Sect. 2.1,

$$u(x)\quad\underbrace{u_0(x)}_{\in V_0}+\sum_{j'\ \ 0}^{j}\underbrace{d_{j'}(x)}_{\in W_{j'}}\quad\underbrace{\sum_k u_{0,k}^{\mathrm{c}}\phi_{0,k}(x)}_{\in V_0}+\sum_{j'\ \ 0}^{j}\underbrace{\sum_k u_{j',k}^{\mathrm{f}}\psi_{j',k}(x)}_{\in W_{j'}}.$$

After having extracted U_j^{c} and U_j^{f} from U_{j+1} using $\mathscr{W}_j$ we can continue hierarchically to extract $U_{j-1}^{\mathrm{f}}\quad\{u_{j-1,k}^{\mathrm{f}}\}$ and $U_{j-1}^{\mathrm{c}}\quad\{u_{j-1,k}^{\mathrm{c}}\}$ from U_j^{c} using $\mathscr{W}_{j-1}$, etc. The computational cost of this operation is $\mathscr{O}(2^j)$. Iterating on this we have

$$\begin{tikzcd}
U_{j+1} \arrow[r] \arrow[dr] & U_j^{\mathrm{f}} & & & & \\
& U_j^{\mathrm{c}} \arrow[r] \arrow[dr] & U_{j-1}^{\mathrm{f}} & & & \\
& & U_{j-1}^{\mathrm{c}} \arrow[r] \arrow[dr] & U_{j-2}^{\mathrm{f}} & & \\
& & & U_{j-2}^{\mathrm{c}} \arrow[r] \arrow[dr] & \cdots & \\
& & & & \cdots \arrow[r] \arrow[dr] & U_0^{\mathrm{f}} \\
& & & & & U_0^{\mathrm{c}}
\end{tikzcd}\tag{31}$$

$$\mathscr{O}(2^{j+1})\ +\ \mathscr{O}(2^j)\ +\ \mathscr{O}(2^{j-1})\ +\ \cdots\ +\mathscr{O}(1)$$

The total cost for this decomposition is given by the last line, which sums up to $\mathscr{O}(N)$ operations where $N = 2^{j+1}$ is the length of the transformed vector. The decomposition is called the *fast wavelet transform* and it has optimal complexity. (With N coefficients to compute, the complexity cannot be better.) This should be compared with the fast Fourier transform which computes Fourier coefficients from sample values at a cost of $\mathscr{O}(N \log N)$.

There is a corresponding fast inverse wavelet transform that reconstructs the original $\{u_{j+1,k}\}$ values from the wavelet coefficients. Since $\mathscr{W}_j^{-1} = \mathscr{W}_j^T$ this amounts to performing the same matrix multiplications as in the forward transform, but with $\mathscr{W}_j^T$ instead of $\mathscr{W}_j$, and in the opposite order. It has the same $\mathscr{O}(N)$ complexity.

For higher order wavelets the general form of $\mathscr{W}_j$ is

$$\mathscr{W}_j = \begin{pmatrix} \tilde{g}_0 & \tilde{g}_1 & \tilde{g}_2 & \cdots & & & \\ 0 & 0 & g_0 & g_1 & g_2 & \cdots & \\ \vdots & \vdots & & \ddots & \ddots & & \\ 0 & 0 & \cdots & g_{n-2} & g_{n-1} & g_n & \\ \tilde{h}_0 & \tilde{h}_1 & \tilde{h}_2 & \cdots & & & \\ 0 & 0 & h_0 & h_1 & h_2 & \cdots & \\ \vdots & \vdots & & \ddots & \ddots & & \\ 0 & 0 & \cdots & h_{n-2} & h_{n-1} & h_n & \end{pmatrix} \in \mathbb{R}^{2^{j+1} \times 2^{j+1}}. \tag{32}$$

The filter coefficients $\{h_k\}$ and $\{g_k\}$ are longer sequences, which creates a problem when computing the coefficients that correspond to locations close to the boundaries of the interval where $u(x)$ is given. In principle, one needs to know about $u(x)$ outside the interval in order to compute them, but $u(x)$ is by assumption only given inside the interval. There are several ways to deal with this dilemma. One can for instance set $u(x)$ to zero outside, continue $u(x)$ periodically ($u(x+1) = u(x)$) or by mirroring ($u(-x) = u(x)$ and $u(x+2) = u(x)$). One can also directly modify $\{h_k\}$ and $\{g_k\}$ for boundary coefficients [19]. In the end, for all the methods this means that the first and middle few lines of $\mathscr{W}_j$ change, indicated by the modification $g_k, h_k \to \tilde{g}_k, \tilde{h}_k$ in the equation above. The wavelet transforms (29) and (31) work in the same way with the modified $\mathscr{W}_j$ as before. When the wavelets and scaling functions are compactly supported the filter coefficients are of finite length and $\mathscr{W}_j$ is a sparse matrix; the cost to perform the matrix vector multiplication is therefore still $\mathscr{O}(2^{j+1})$ and the complexity of the forward and inverse wavelet transform is $\mathscr{O}(N)$.

Remark 2. The two-dimensional wavelet transform corresponding to the two-dimensional scaling and wavelet spaces in Remark 1 can be written as a tensor product of one-dimensional transforms,

$$\mathscr{W}_j \otimes \mathscr{W}_j. \tag{33}$$

The fast wavelet transform generalizes easily to higher dimensions based on this.

Exercise 6. Write a MATLAB program based on the *cascade algorithm* to approximate and plot the scaling function and mother wavelet for the Daubechies 4 system with two vanishing moments. This algorithm is simply the fast inverse wavelet transform applied to $u^{\mathrm{c}}_{0,k} \quad \delta_k$, $u^{\mathrm{f}}_{j,k} \quad 0$ for the scaling function and to $u^{\mathrm{c}}_{0,k} \quad 0$, $u^{\mathrm{f}}_{j,k} \quad \delta_j\delta_k$ for the mother wavelet. Since

$$\phi(x) \quad \sum_k \delta_k \phi_{0,k}(x), \qquad \psi(x) \quad \sum_j \sum_k \delta_k \delta_j \psi_{j,k}(x),$$

by (26), an inverse transform to fine enough scale J gives the good approximation

$$\phi(k2^{-J}) \approx 2^{J/2} u^{\mathrm{c}}_{J,k},$$

and similarly for $\psi(x)$. Use the filter coefficients derived in Example 2 of Sect. 2.7.

3 Wavelet based numerical homogenization

Wavelets have been used in several ways for the numerical solution of partial differential and integral equations. Many of these problems involve Calderon-Zygmund or pseudo differential operators, which can be well compressed when projected on wavelet spaces; properties of the operator kernel allow compression by arguments similar to those in Sect. 2.5. This is the basis for many fast algorithms for integral equations and boundary integral formulations of PDEs [10, 23]; for more details, see the contribution by Lexing Ying in this volume. It will also play a role in the wavelet based numerical homogenization discussed here. For the discretization of PDEs wavelet Galerkin methods are often used, in which the approximate finite element spaces and basis functions are taken to be fine scaling spaces and wavelets, [57]. The ability of wavelets to detect local regularity and singularities have made them particularly useful in adaptive schemes, where better resolution in non-smooth regions is obtained by using finer scale level of the wavelets there. This can be seen as *space refinement* rather than the usual mesh refinement. See e.g. [35, 20] for hyperbolic problems and [17] for elliptic problems.

In this final section we will be interested in using wavelets to simplify numerical simulation of PDEs where the existence of *subgrid scale* phenomena poses considerable difficulties. With subgrid scale phenomena, we mean those processes which could influence the solution on the computational grid but which have length scales shorter than the grid size. Highly oscillatory initial data may, for example, interact with fine scales in the material properties and produce coarse scale contributions to the solution.

We consider the general problem where L_ε is a linear differential operator for which ε indicates small scales in the coefficients. The solution u_ε of the differential equation

$$L_\varepsilon u_\varepsilon \quad f_\varepsilon, \tag{34}$$

will typically inherit the small scales from the operator L_ε or the data f_ε. A concrete example could be the simple model problem

$$L_\varepsilon u_\varepsilon \quad -\frac{d}{dx}\left(g_\varepsilon(x)\frac{d}{dx}\right)u_\varepsilon(x) \quad f_\varepsilon(x), \qquad 0<x<1, \tag{35}$$
$$u_\varepsilon(0) \quad u_\varepsilon(1) \quad 0,$$

where the coefficient $g_\varepsilon(x)$ and right hand side $f_\varepsilon(x)$ have a fine scale structure; it may for instance be highly oscillatory, or have a localized sharp transition. Numerical difficulties originate from the small scales in L_ε and f_ε. Let

$$L_{\varepsilon h}u_{\varepsilon h} \quad f_{\varepsilon h}. \tag{36}$$

be a discretization of (34), with the typical element size or step size h. If ε denotes a typical wave length in u_ε then h must be substantially smaller than ε in order to resolve the ε-scale in the numerical approximation. This can be costly if ε is small compared to the overall size of the computational domain.

There are a number of traditional ways to deal with this multiple scale problem. Several methods are based on physical considerations for a specific application, such as turbulence models in computational fluid dynamics, [56], and analytically derived local subcell models in computational electromagnetics, [54]. Geometrical optics or geometrical theory of diffraction approximations of high frequency wave propagation are other classical techniques to overcome the difficulty of highly oscillatory solutions, [40]. All these techniques result in new sets of approximative equations that do not contain the small scales, but which anyway attempt to take the effect of these scales into account. A more general analytical technique for achieving this goal is classical homogenization, discussed below.

If the small scales are localized, there are some direct numerical procedures which are applicable. Local mesh refinement is quite common but could be costly if the small scales are very small or distributed. There are also problems with artificial reflections in mesh size discontinuities and time step limitations for explicit techniques. Numerical shock tracking or shock fitting can also be seen as subgrid models, [2].

In the remainder of this paper we will present a wavelet based procedure for constructing subgrid models to be used on a coarse grid where the smallest scales are not resolved. The objective is to find a finite dimensional approximation of (34),

$$\bar{L}_{\varepsilon\bar{h}}\bar{u}_{\varepsilon\bar{h}} \quad \bar{f}_{\varepsilon\bar{h}},$$

that accurately reproduces the effect of subgrid scales and that in some sense is similar to a discretization of a differential equation. The discrete operator $\bar{L}_{\varepsilon\bar{h}}$ should thus resemble a discretized differential operator, and it should be designed such that $\bar{u}_{\varepsilon\bar{h}}$ is a good approximation of u_ε even if $\bar{h}$ is not small compared to ε. This goal resembles that of classical analytical homogenization, which we will now briefly discuss.

3.1 Classical homogenization

Homogenization is a well established analytical technique to approximate the effect of smaller scales onto larger scales in multiscale differential equations. The problem is often formulated as follows. Consider a set of operators L_ε indexed by the small parameter ε. Find the limit solution $\bar{u}$ and the *homogenized operator* $\bar{L}$ defined by

$$L_\varepsilon u_\varepsilon \quad f, \qquad \lim_{\varepsilon\to 0} u_\varepsilon \quad \bar{u}, \qquad \bar{L}\bar{u} \quad f, \tag{37}$$

for all f in some function class. In certain cases the convergence above and existence of the homogenized operator can be proved, [6].

For simple model problems, with coefficients that are periodic on the fine scale, exact closed form solutions can be obtained. For instance, with $g(x,y)$ positive, 1-periodic in y and bounded away from zero, we have for the one-dimensional elliptic example (35),

$$L_\varepsilon \quad -\frac{d}{dx}\left(g(x,x/\varepsilon)\frac{d}{dx}\right), \qquad \bar{L} \quad -\frac{d}{dx}\left(\hat{g}(x)\frac{d}{dx}\right), \qquad \hat{g}(x) \quad \left(\int_0^1 \frac{dy}{g(x,y)}\right)^{-1}. \tag{38}$$

With the same $\hat{g}$ we get for the hyperbolic operators,

$$L_\varepsilon \quad \frac{\partial}{\partial t} + g(x,x/\varepsilon)\frac{\partial}{\partial x}, \qquad \bar{L} \quad \frac{\partial}{\partial t} + \hat{g}(x)\frac{\partial}{\partial x}. \tag{39}$$

In higher dimensions, the solution to (37) is more complicated. For (39), even the type of the homogenized equation depends strongly on the properties of the coefficients, see [28, 37]. For the multidimensional elliptic case (35) the structure of the homogenized operator can still be written down, as long as the coefficients are periodic or stochastic. Let $G(y) : \mathbb{R}^d \mapsto \mathbb{R}^{d\times d}$ be uniformly elliptic and 1-periodic in each of its arguments. Let I_d denote the unit square in d dimensions. It can then be shown, [6], that

$$L_\varepsilon \quad -\nabla\cdot\left(G\left(\frac{x}{\varepsilon}\right)\nabla\right), \quad \bar{L} \quad -\nabla\cdot(\hat{G}\nabla), \quad \hat{G} \quad \int_{I_d} G(y) - G(y)\frac{d\chi(y)}{dy}dy, \tag{40}$$

where $d\chi/dy$ is the jacobian of the function $\chi(y) : \mathbb{R}^d \mapsto \mathbb{R}^d$, given by solving the so called *cell problem*,

$$\nabla\cdot G(y)\frac{d\chi(y)}{dy} \quad \nabla\cdot G(y), \qquad y\in I_d,$$

with periodic boundary conditions for χ.

3.2 Numerical homogenization

Classical homogenization is very useful when it is applicable. The original problem with small scales is reduced to a homogenized problem that is much easier to approximate numerically. See the left path in Fig. 8. The final discretization $\bar{L}_{\bar{h}}\bar{u}_{\bar{h}} \quad \bar{f}_{\bar{h}}$ satisfies the criteria we set up at the end of the introduction to Sect. 3; since there are no small scales in the homogenized equation (37) the size of $\bar{h}$ can be chosen independently of ε and $\bar{u} \approx u_\varepsilon$ for small ε.

$$\begin{tikzcd}
& L_\varepsilon u_\varepsilon \quad f_\varepsilon \arrow[dl] \arrow[dr] \arrow[ddr] & \\
\bar{L}\bar{u} \quad \bar{f} \arrow[d] & & L_{\varepsilon h}u_{\varepsilon h} \quad f_{\varepsilon h} \arrow[d] \\
\bar{L}_{\bar{h}}\bar{u}_{\bar{h}} \quad \bar{f}_{\bar{h}} & & \bar{L}_{\varepsilon\bar{h}}\bar{u}_{\varepsilon\bar{h}} \quad \bar{f}_{\varepsilon\bar{h}}
\end{tikzcd}$$

Fig. 8. Schematic steps in homogenization.

If analytical homogenization is not possible, one can instead numerically compute a suitable discrete operator $\bar{L}_{\varepsilon\bar{h}}$ which has the desired properties. We call such a procedure *numerical homogenization*. The numerical homogenization can be done directly as indicated by the middle path in Fig. 8, or by first discretizing the original problem and then compressing the operator $L_{\varepsilon h}$ and the data $f_{\varepsilon h}$ as indicated by the right path. In this paper we follow the latter strategy, and present wavelet based methods in order to achieve numerical homogenization. There are other similar methods based on coarsening techniques from algebraic multigrid [41, 49]. In the finite element setting the effect of the microstructure can be incorporated in a Galerkin framework [38, 36]. The great advantage of these procedures in deriving subgrid models is their generality. They can be used on any system of differential equations and do not require separation into distinct $\mathscr{O}(\varepsilon)$ and $\mathscr{O}(1)$ scales or periodic coefficients. They can also be used to test if it is physically reasonable to represent the effect of fine scales on a coarse scale grid with a local operator.

The original ideas for wavelet based numerical homogenization are from Beylkin and Brewster, [8]. See also [27, 32, 3, 15]. The technique has been applied to several problems. We refer the reader to [50] for a waveguide filter containing a fine scale structure, with examples of how to use the numerical homogenization technique to construct subgrid models, in particular 1d models from 2d models. In [44, 29] the numerically homogenized operator was used as a coarse grid operator in multigrid methods. Applications to porous media flow were considered in [55]. Extension of the procedure to nonlinear problems can be found in [9, 42]. For a survey of wavelet based numerical homogenization see [30].

3.3 Deriving the numerically homogenized operator

Let us consider a discrete approximation of a PDE in the space V_{j+1},

$$L_{j+1}U \quad F, \qquad U,F \in V_{j+1}, \qquad L_{j+1} \in \mathscr{L}(V_{j+1}). \tag{41}$$

Here $\mathscr{L}(V_{j+1})$ denotes the space of bounded linear operators from V_{j+1} to itself, which in our case is isomorphic to matrices in $\mathbb{R}^{2^{j+1}\times 2^{j+1}}$. The equation may originate from a finite difference, finite element or finite volume discretization of a given differential equation. In the Haar case U can be identified as a piecewise constant approximation of $u(x)$, the solution to the continuous problem. We seek an operator defined on a coarser grid that extracts only the coarse part of the solution. For a function in the space V_{j+1} this amounts to the part in V_j.

We start by applying the wavelet transform matrix $\mathscr{W}_j$ in (30) or (32) to (41) from the left and use the relation $\mathscr{W}_j^T\mathscr{W}_j \quad I$. We get

$$\mathscr{W}_j L_{j+1} \mathscr{W}_j^T (\mathscr{W}_j U) \quad \mathscr{W}_j F.$$

If we decompose $\mathscr{W}_j L_{j+1} \mathscr{W}_j^T$ in four equal size blocks we have

$$\begin{pmatrix} A_j & B_j \\ C_j & L_j \end{pmatrix} \begin{pmatrix} U^{\mathrm{f}} \\ U^{\mathrm{c}} \end{pmatrix} \quad \begin{pmatrix} F^{\mathrm{f}} \\ F^{\mathrm{c}} \end{pmatrix}, \qquad U^{\mathrm{f}}, F^{\mathrm{f}} \in W_j, \qquad U^{\mathrm{c}}, F^{\mathrm{c}} \in V_j. \tag{42}$$

As before, “f” means projection onto the fine scale subspace W_j, and “c” stands for projection onto the coarse scale subspace V_j. For simplicity we shall assume that F is a discretization of a smooth function such that $F^{\mathrm{f}} \quad 0$. After eliminating U^{f} in (42) via block Gaussian elimination we then obtain

$$(L_j - C_j A_j^{-1} B_j) U^{\mathrm{c}} \quad F^{\mathrm{c}}. \tag{43}$$

In this equation we let

$$\bar{L}_j : \quad L_j - C_j A_j^{-1} B_j, \qquad \bar{L}_j \in \mathscr{L}(V_j). \tag{44}$$

This operator is half the size of the original L_{j+1}, defined on twice as coarse grid. Moreover, given $\bar{L}_j$ and F^{c} we can solve (43) to get the coarse part of the solution, U^{c}, taking the influence of fine scales not present in V_j into account. For these reasons we call $\bar{L}_j$ the *numerically homogenized operator*.

We note that the block in (42) called L_j can be written $L_j \quad P_j L_{j+1} P_j$ where P_j is the orthogonal projection on V_j. It can therefore be interpreted as one type of direct discretization on the coarse scale. We can then see $C_j A_j^{-1} B_j$ as a correction term to this discretization, which includes subgrid phenomena in $\bar{L}_j$. In the elliptic case, there is a striking similarity between the classical homogenized operator in (40) and $\bar{L}_j$ in (44). Both are written as the average of the original operator minus the correction term, which is computed in a similar way for both operators. For the analytical case, a local elliptic cell problem is solved to get $G\partial_y \chi$, while in (44), a positive operator A_j defined on a subspace $W_j \subset V_{j+1}$ is inverted to obtain $C_j A_j^{-1} B_j$. The average over the terms is obtained by integration in the analytical case, and by applying P_j in the wavelet case.

The procedure to obtain $\bar{L}_j$ can be applied recursively on $\bar{L}_j$ itself to get $\bar{L}_{j-1}$ and so on,

$$\bar{L}_j \to \bar{L}_{j-1} \to \bar{L}_{j-2} \to \dots, \qquad \bar{L}_j \in \mathscr{L}(V_j). \tag{45}$$

That this is possible can easily be verified when L_{j+1} is symmetric positive definite (see Exercise 7). Moreover, an improvement in the condition number κ is often obtained. Typically for standard discretizations

$$\kappa(\bar{L}_k) < \kappa(L_{j+1}),$$

when $k \leq j$.

Higher dimensional problems can be treated in a similar way. Suppose

$$L_{j+1}U = F, \qquad U,F \in \boldsymbol{V}_{j+1}, \qquad L_{j+1} \in \mathscr{L}(\boldsymbol{V}_{j+1})$$

where $\boldsymbol{V}_{j+1}$ is the two-dimensional scaling space in Remark 1. By using the two-dimensional wavelet transform (33) one obtains the same numerically homogenized operator (44) with V_j replaced by $\boldsymbol{V}_j$. For this case the fine scale part of U can be decomposed as

$$U^{\mathrm{f}} = \begin{pmatrix} U^{\mathrm{ff}} \\ U^{\mathrm{cf}} \\ U^{\mathrm{fc}} \end{pmatrix}, \qquad U^{\mathrm{ff}} \in W_j \otimes W_j, \qquad U^{\mathrm{cf}} \in V_j \otimes W_j, \qquad U^{\mathrm{fc}} \in W_j \otimes V_j.$$

In some cases, the homogenized operator keeps important properties of the original operator. Let the forward and backward undivided differences be defined as

$$\Delta_+ u_i = u_{i+1} - u_i, \qquad \Delta_- u_i = u_i - u_{i-1}.$$

In [27] it was shown that the one-dimensional elliptic model equation $-(gu')' = f$ discretized as

$$L_{j+1}U = -\frac{1}{h^2}\Delta_+ \mathrm{diag}(g)\Delta_- U = F \tag{46}$$

will preserve its divergence form during homogenization. That is, we will get

$$\bar{L}_j = -\frac{1}{(2h)^2}\Delta_+ H_j \Delta_-, \tag{47}$$

where H_j is a strongly diagonal dominant matrix which can be interpreted as the effective material coefficient related to g. Analogously, for the first order differential operator $g(x)\frac{\partial}{\partial x}$ the discretized form $\mathrm{diag}(g)\Delta_-/h$ is preserved during homogenization,

$$\bar{L}_j = \frac{1}{2h} H_j \Delta_-. \tag{48}$$

In two dimensions, the elliptic model equation $-\nabla(g(x,y)\nabla u) = f$ can be discretized as

$$L_{j+1} = -\frac{1}{h^2}\left(\Delta_+^x G \Delta_-^x + \Delta_+^y G \Delta_-^y\right), \qquad L_{j+1}U = F.$$

Then $\bar{L}_j$ is no longer on exactly the same form as L_{j+1}. The cross-derivatives must also be included. We get

$$\bar{L}_j = -\frac{1}{(2h)^2}\left(\Delta_+^x H^{xx} \Delta_-^x + \Delta_+^y H^{yx} \Delta_-^x + \Delta_+^x H^{xy} \Delta_-^y + \Delta_+^y H^{yy} \Delta_-^y\right), \tag{49}$$

for some matrices H^{xx}, H^{yx}, H^{xy}, and H^{yy}.

Exercise 7. Show that if L_{j+1} is symmetric positive definite then so is $\bar{L}_j$. Conclude from this that the iterated homogenization (45) is possible.
Hint for the positivity: The matrix $\mathscr{W}_j^T L_{j+1} \mathscr{W}_j$ is positive definite if and only if L_{j+1} is. Then

$$\begin{pmatrix} v^T & w^T \end{pmatrix} \mathscr{W}_j^T L_{j+1} \mathscr{W}_j \begin{pmatrix} v \\ w \end{pmatrix} > 0, \qquad \forall v, w \; / \; 0.$$

Explain why A_j^{-1} exists and then take $v \quad -A_j^{-1} B_j w$.

3.4 Compact representation of projected operators

When the operator L_{j+1} is derived from a finite difference, finite element or finite volume discretization, it is sparse and of a certain structure. In one dimension it might, for instance, be tridiagonal. However, in general $\bar{L}_j$ will not be represented by a sparse matrix even if L_{j+1} is, because A_j^{-1} would typically be dense. Computing all components of $\bar{L}_j$ would be inefficient. Fortunately, $\bar{L}_j$ will be diagonal dominant in many important cases and we can then find a sparse matrix that is a close approximation of $\bar{L}_j$. If this sparse matrix is of banded form, it can be seen as a discretization of a local differential operator acting on the coarse space.

Diagonal dominance of $\bar{L}_j$

We now consider some cases where $\bar{L}_j$ will be strongly diagonal dominant. This is related to the corresponding properties of A_j.

In some simple cases the matrix A_j is in fact diagonal. Examples include integral operators L of the form

$$Lu \quad \int_0^x a(t)u(t)dt + b(x)u(x). \tag{50}$$

As before, let P_j be the orthogonal projection on V_j. Suppose L is discretized in V_{j+1} as $L_{j+1} \quad P_{j+1} L P_{j+1}$ with a in (50) replaced by $P_{j+1} a P_{j+1}$, and similar for b. When the Haar system is used, A_j is diagonal and $\bar{L}_j$ is of the same form as L_{j+1}. Let $\bar{A}_k$ be related to $\bar{L}_{k+1}$ via (42) in the same way as A_j relates to L_{j+1}. By induction, $\bar{A}_k$ is also diagonal for $k < j$. The operators in (50) turn up for instance in problems with systems of ordinary differential equations and one-dimensional elliptic equations, see [8, 32]. In these cases, an explicit recurrence relation between scale levels can be established, which permits the computation of $\bar{L}_k$ on any fixed level k as the starting level, $j+1$, tends to infinity.

For more general problems one must instead rely on the rapid decay of elements in A_j and $\bar{L}_j$ off the diagonal, which is a consequence of the wavelet spaces' good approximation properties discussed in Sect. 2.5. The decay rate of A_j for Calderon–Zygmund and pseudo-differential operators were given by Beylkin, Coifman and Rokhlin in [10]. Letting $L_{j+1} \quad P_{j+1} L P_{j+1}$, $A_j \quad \{a_{k\ell}^j\}$, $B_j \quad \{b_{k\ell}^j\}$ and $C_j \quad \{c_{k\ell}^j\}$, they show that

$$|a^j_{k\ell}| + |b^j_{k\ell}| + |c^j_{k\ell}| \leq \frac{2^{-\lambda j} C_M}{1 + |k-\ell|^{M+1}}, \qquad |k-\ell| \geq \nu, \tag{51}$$

when the wavelet system has M vanishing moments. For Calderon–Zygmund operators $\lambda \quad 0$ and $\nu \quad 2M$. For a pseudo-differential operator $\nu \quad 0$ and the symbol $\sigma(x,\xi)$ and its adjoint should both belong to the symbol class $S^{\lambda}_{1,1}$, i.e. they should satisfy the estimate

$$|\partial^{\alpha}_{\xi} \partial^{\beta}_{x} \sigma(x,\xi)| \leq C_{\alpha,\beta}(1 + |\xi|)^{\lambda - \alpha + \beta},$$

for some constants $C_{\alpha,\beta}$. For instance, in the second order elliptic case $\lambda \quad 2$. Moreover, Beylkin and Coult, [11], showed that if (51) holds with $\lambda \quad 0$ for A_j, B_j and C_j given by L_{j+1} in (42), then the same estimate also holds for $\bar{A}_{j'}$, $\bar{B}_{j'}$ and $\bar{C}_{j'}$, here given by $P_{j'}\bar{L}_{k+1}P_{j'}$ for $j' \leq k < j$ with $j+1$ being the starting homogenization level. Hence, the decay rate is preserved after homogenization.

The decay estimate in [11] for $\bar{A}_{j'}$ is uniform in k and may not be sharp for a fixed k. There is, for example, a general result by Concus, Golub and Meurant, [22], for diagonal dominant, symmetric and tridiagonal matrices. For those cases, which include A_j corresponding to the discretization in (46) of the one-dimensional elliptic operator, the inverse has exponential decay,

$$\left| \left(A_j^{-1} \right)_{k\ell} \right| \leq C\rho^{|k-\ell|}, \qquad 0 < \rho < 1.$$

This holds also when the elliptic operator has a lower order term of type $b(x)\partial_x$ discretized with upwinding, [44].

Approximating $\bar{L}_j$

We now discuss different approximation strategies for $\bar{L}_j$. A simple approach is the basic thresholding method used in [10], where small elements of $\bar{L}_j$ are set to zero. This is, however, not practical here since the location of the non-zero elements cannot be controlled, and we want to obtain a banded approximation of $\bar{L}_j$, which corresponds to a discretization of a local differential operator.

The first, and simplest, approximation method that we use is instead to set all components outside a prescribed bandwidth ν equal to zero. This is motivated by the decay of elements off the diagonal in $\bar{L}_j$. Let us define

$$\mathrm{trunc}(M,\nu)_{ij} \quad \begin{cases} M_{ij}, & \text{if } 2|i-j| \leq \nu - 1 \\ 0, & \text{otherwise.} \end{cases} \tag{52}$$

For $\nu \quad 1$ the matrix is diagonal. For $\nu \quad 3$ it is a tridiagonal and so on. We refer to it as *truncation*.

In the second approximation method, the matrix $\bar{L}_j$ is projected onto banded form in a more effective manner. The aim is that the projected matrix should give the same result as the original matrix on a given subspace, e.g. when applied to vectors representing smooth functions. Let $\{\mathbf{v}_j\}_{j\quad 1}^{\nu}$ be a set of linearly independent vectors in $\mathbb{R}^N$. Denote by $\mathscr{T}_\nu$ the subspace of $\mathbb{R}^{N\times N}$ with matrices essentially[4] of bandwidth ν. Moreover, let

$$\mathscr{L}_\nu \quad \{M \in \mathbb{R}^{N\times N} : \ \mathrm{span}\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_\nu\} \subset \mathsf{N}(M)\},$$

where $\mathsf{N}(M)$ represents the null space of M. Then

$$\mathbb{R}^{N\times N} \quad \mathscr{T}_\nu \oplus \mathscr{L}_\nu$$

and we define the *band projection* of a matrix $M \in \mathbb{R}^{N\times N}$ as the projection of M onto $\mathscr{T}_\nu$ along $\mathscr{L}_\nu$, with the notation

$$\mathrm{band}(M,\nu) \quad \mathrm{Proj}_{\mathscr{T}_\nu} M. \tag{53}$$

As a consequence,

$$M\mathbf{x} \quad \mathrm{band}(M,\nu)\mathbf{x}, \qquad \forall \mathbf{x} \in \mathrm{span}\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_\nu\}.$$

In our setting M will usually operate on vectors representing smooth functions, for instance solutions to elliptic equations, and a natural choice for $\mathbf{v}_j$ vectors are thus the first ν polynomials,

$$\mathbf{v}_j \quad \{1^{j-1}, 2^{j-1}, \ldots, N^{j-1}\}^T, \qquad j \quad 1, \ldots, \nu.$$

Smooth solutions to the homogenized problem should be well approximated by these vectors. For the case $\nu \quad 1$ we get the standard "masslumping" of a matrix, often used in the context of finite element methods.

This technique is similar to the probing technique used by Chan et al., [14]. In that case the vectors $\mathbf{v}_j$ are sums of unit vectors. Other probing techniques have been suggested by Axelsson, Pohlman and Wittum; see [4, Chap. 8]. The choice of $\mathbf{v}_j$ vectors could be optimized if there is some a priori knowledge of the homogenized solution.

The two truncation methods described above are even more efficient when applied to H_j, the effective coefficient, instead of directly to the homogenized operator $\bar{L}_j$. For (46, 47) we could for instance approximate

$$\bar{L}_j \approx -\frac{1}{(2h)^2} \Delta_+ \,\mathrm{trunc}(H,\nu) \Delta_- .$$

The following proposition shows that when the solution to the homogenized problem belongs to the Sobolev space $H^1(\mathbb{R})$, the accuracy of this approach is one order higher.

[4] We must require that each row of the matrices in $\mathscr{T}_\nu$ has the same number of elements. Therefore, the first and last $\nu - 1$ rows will have additional elements located immediately to the right and left of the band, respectively.

Proposition 1. *Suppose $L \quad \Delta_+ H \Delta_-$ and $LU \quad h^2 f$. Consider the perturbed problems*

$$(L+\delta L)(U+\delta U_L) \quad h^2 f, \qquad \Delta_+(H+\delta H)\Delta_-(U+\delta U_H) \quad h^2 f.$$

For small enough perturbations, and the same constant C,

$$||\delta U_L|| \le C\frac{||\delta L||}{||L||}||U||, \qquad ||\delta U_H|| \le \frac{h}{2}C\frac{||\delta H||}{||H||}||u||_{H^1},$$

where $||\cdot||$ denotes the discrete L^2-norm and u is any H^1-function such that $U_j \quad u(jh)$.

See [3] for a proof.

Another approach proposed by Chertock and Levy [15] is to only approximate A_j^{-1}, the high frequency part of the projected operator. In this case we would approximate

$$\bar{L}_j \approx L_j - C_j \operatorname{trunc}(A_j^{-1}, \nu) B_j.$$

Numerical evidence in [15] strongly suggests that this gives a better approximation than truncating the full operator. In particular the strategy works better when truncation is done in each homogenization step, instead of only in the final step. It is also simpler than approximating H_j in higher dimensions.

Computing the full inverse of A_j is expensive, and to reduce computational cost one can compute the truncation $\operatorname{trunc}(\bar{L}_j, \nu)$ directly. By capitalizing on the nearly sparse structure of the matrices involved, it was shown in [11] that the cost can be reduced to $\mathcal{O}(N)$ operations for N unknowns and fixed accuracy. Moreover, the same homogenized operator will typically be reused multiple times, for instance with different right hand sides, or in different places of the geometry as a subgrid model. The computation of $\operatorname{band}(\bar{L}_j, \nu)$ can be based on $\operatorname{trunc}(\bar{L}_j, \mu)$, $\mu > \nu$. The additional computational cost is proportional to $(\nu^3 + \mu\nu)N$. The $\nu^3 N$ term corresponds to solving N $\nu \times \nu$ systems and $\mu\nu N$ to computing the right hand sides, see also [4].

In two dimensions truncation to simple banded form is in general not adequate, since the full operator will typically be block banded. However, both the crude truncation and the band projection generalize easily to treat block banded form instead of just banded. Let M be the tensor product of two $N \times N$ matrices. Then we define truncation as

$$\operatorname{trunc}_2(M,\nu)_{ij} \quad \begin{cases} M_{ij}, & \text{if } 2|i-j-rN| \le \nu - 1 - |2r|, \\ 0, & \text{otherwise,} \end{cases} \qquad |2r|+1 \le \nu. \tag{54}$$

This mimics the typical block structure of a discretized differential operator. For the band projection, the space $\mathcal{T}_\nu$ of banded matrices in the one-dimensional definition, is simply replaced by the space of matrices with the block banded sparsity pattern defined in (54).

In two dimensions, untangling the various H components of (49) from $\bar{L}_j$ is more complicated than finding the H in (47) and (48) for one-dimensional problems. Although, in principle, this can still be done, it may be easier to truncate A_j^{-1} than the H components for two-dimensional problems, as suggested in [15].

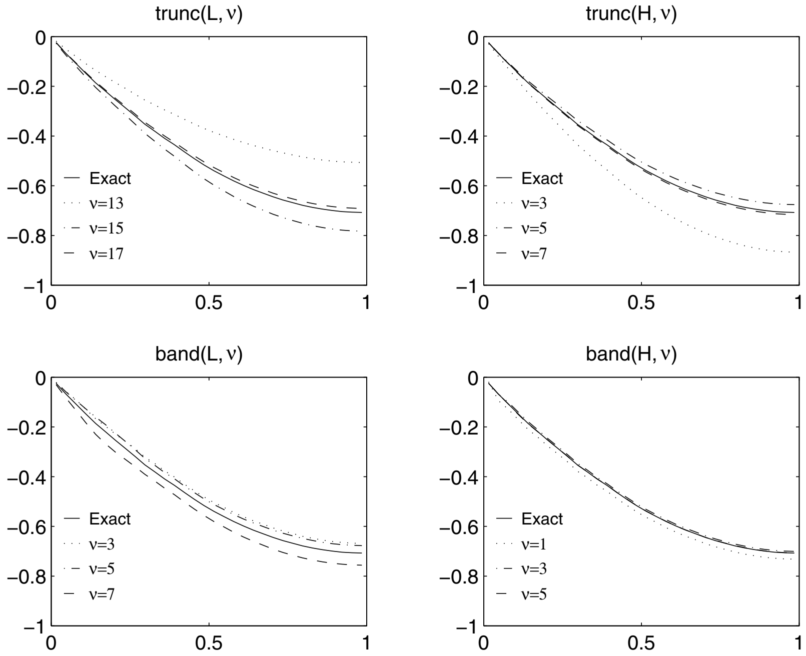


Fig. 9. Result for the elliptic model problem, $g(x)$ random, when the homogenized operator is approximated in different ways. The "exact" solution refers to the solution with the full 32×32 homogenized operator

4 Numerical examples

In this section we present numerical results for the algorithms described above. We first consider elliptic problems in one dimension and the Helmholtz equation in one and two dimensions. Uniform, central, finite volume discretization are used in the numerical experiments. The homogenization steps are done in the standard Haar basis. The computational domain is the unit interval (square in 2D). The grid size is denoted n, and the cell size $h \quad 1/n$. For many more numerical examples, see [30] and, for the approximation strategy involving A_j^{-1}, [15].

4.1 The 1D elliptic model equation

We approximate

$$-\partial_x g(x)\partial_x u \quad 1, \qquad u(0) \quad u_x(1) \quad 0,$$

where the coefficient $g(x)$ has a uniform random distribution in the interval $0.5, 1$. We take $n \quad 256$ grid points and make three homogenization steps. The coarsest level then contains 32 grid points.

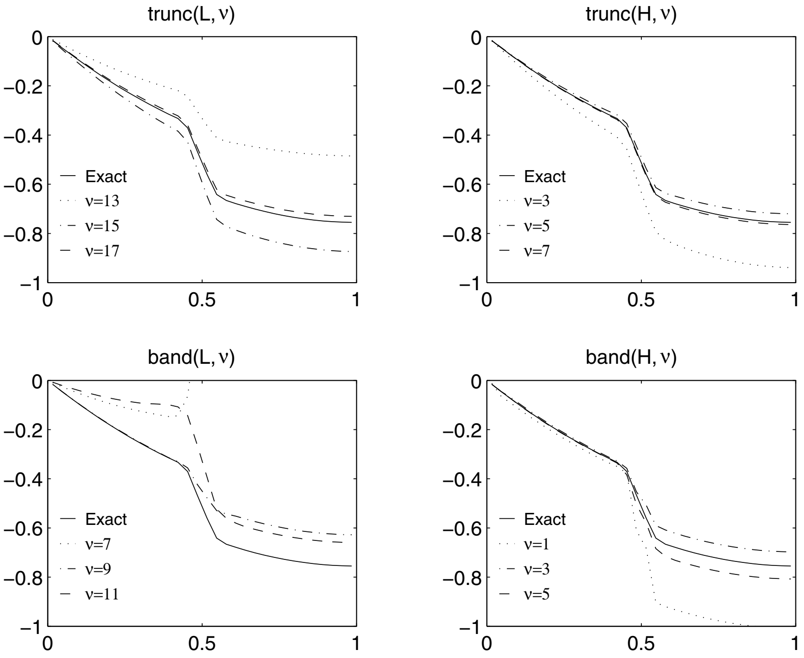


Fig. 10. Result for the elliptic model problem, $g(x)$ a slit, when the homogenized operator is approximated in different ways. The "exact" solution refers to the solution with the full 32×32 homogenized operator

In Fig. 9 different truncation strategies are compared. The exact reference solution is given by the numerically homogenized operator at the coarsest level without any truncation. This is equivalent to the projection onto the coarse scale of the solution on the finest scale. In the top two subplots we use truncation (52). In the bottom two subplots we use the band projection described in Sect. 3.4. The approximation is performed on H, see (47), and on $\bar{L}$ after all three homogenizations. We see that band projection gives a better approximation. We also see that it is more efficient to truncate H than to truncate $\bar{L}$.

Next, the coefficient in the differential equation is changed to

$$g(x) \quad \begin{cases} 1/6, & 0.45 < x < 0.55, \\ 1, & \text{otherwise.} \end{cases} \tag{55}$$

All other characteristics are kept. The result is given in Fig. 10 and it shows that the relative merits of the different methods are more or less the same. The structures of the untruncated $\bar{L}$ and H matrices are shown in Fig. 11. It should be noted that the local inhomogeneity of the full operator has spread out over a larger area, but it is still essentially local.

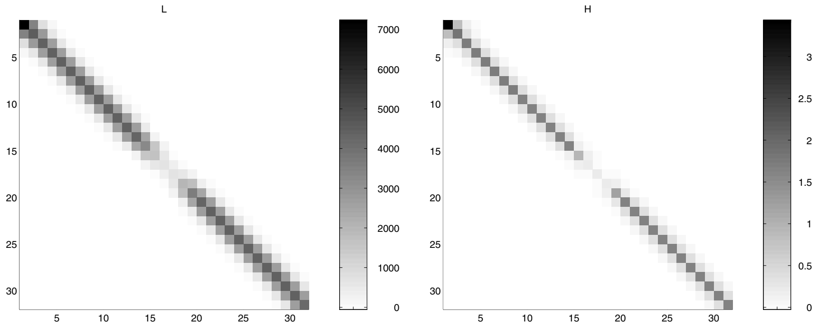


Fig. 11. Structure of the untruncated homogenized operators $\bar{L}$ (left) and H (right) for the elliptic model problem, $g(x)$ a slit. Gray level indicates absolute value of elements

4.2 The 1D Helmholtz equation

In this section we solve the Helmholtz equation

$$\partial_x g(x)\partial_x u + \omega^2 u = 0, \qquad u(0) = 1, \quad u_x(1) = 0.$$

We use $\omega = 2\pi$ and the same $g(x)$ as in (55) and again we take $n = 256$ and use three homogenizations. We get

$$\bar{L}u = (\tilde{L} - \omega^2 I)u = 0.$$

Truncation is performed on $\tilde{L}$ (or $\tilde{H}$) and not on $\bar{L}$. The result is in Fig. 12. We see that Helmholtz equation gives results similar to those of the model equation. Again band projection is more efficient than truncation and approximating H is more efficient than approximating $\bar{L}$.

4.3 The 2D Helmholtz equation

We consider the two-dimensional version of Helmholtz equation,

$$\nabla \cdot g(x,y)\nabla u + \omega^2 u = 0, \qquad (x,y) \in (0,1)^2,$$

with periodic boundary conditions in the y-direction, and at the left and right boundaries, $u(0,y) = 1$, $u_x(1,y) = 0$ respectively. This is a simple model of a plane time-harmonic wave of amplitude one entering the computational domain at the line $x = 0$, passing through a medium defined by the coefficent $g(x,y)$ and flowing out at $x = 1$. As an example we choose the $g(x,y)$ shown in Fig. 13, which represents a wall with a small hole where the incoming wave can pass through. With $\omega = 3\pi$ and $n = 48$, we obtained the results presented in Fig. 14.

The operator is homogenized following the theory for two-dimensional problems in Sect. 3.3. After one homogenization step is truncated according to (54). We show the results of truncation in Fig. 15, for various values of ν. The case $\nu = 9$ corresponds to a compression to approximately 7% of the original size. The structure of this operator is shown in Fig. 16.

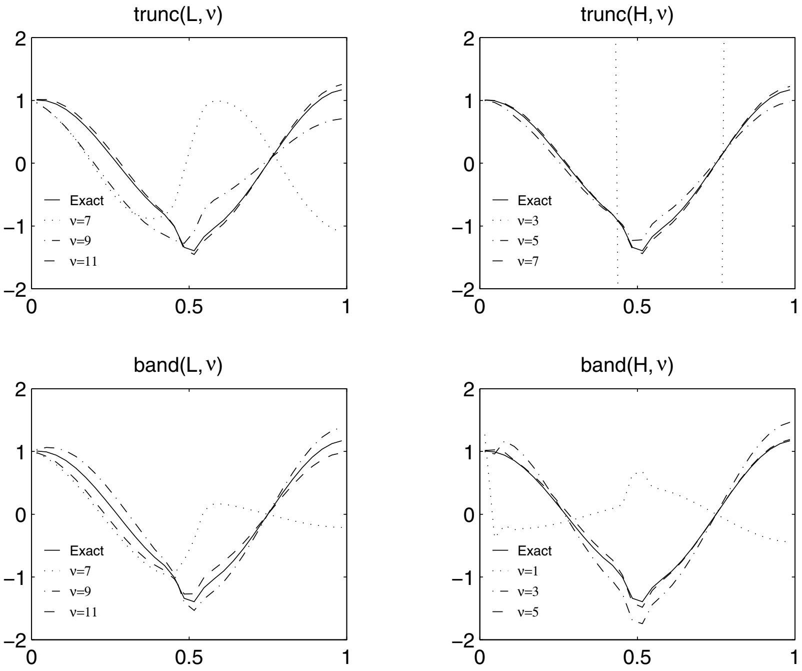
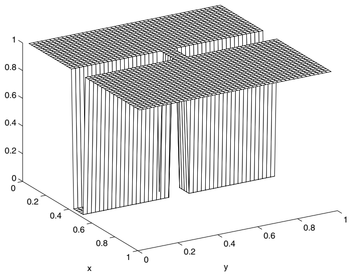


Fig. 12. Result for the Helmholtz equation, $g(x)$ a slit, when the homogenized operator is approximated in different ways. The "exact" solution refers to the solution with the full 32×32 homogenized operator



$$g(x,y) \quad \begin{cases} 10^{-4}, & 0.4 < x < 0.5 \text{ and} \\ & |y-0.5| > 0.05, \\ 1, & \text{otherwise.} \end{cases}$$

Fig. 13. The variable coefficient $g(x,y)$ used in the 2D Helmholtz example

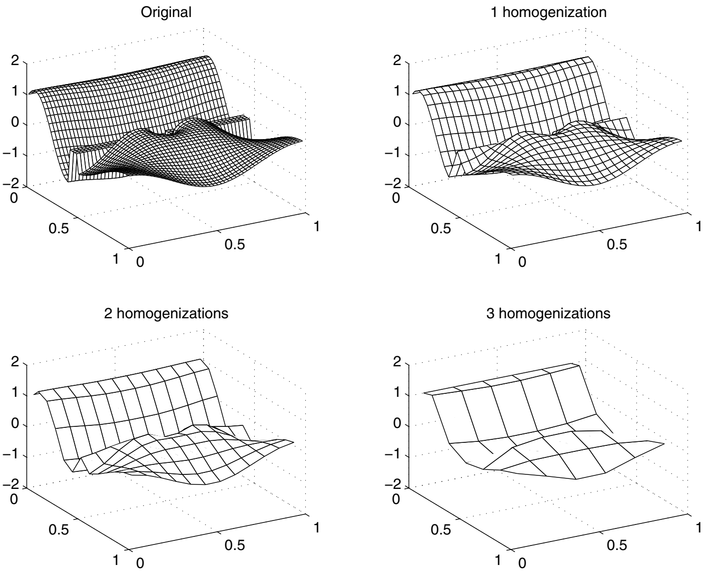


Fig. 14. Result for the 2D Helmholtz example. Solution shown for $0,\ldots,3$ homogenization steps

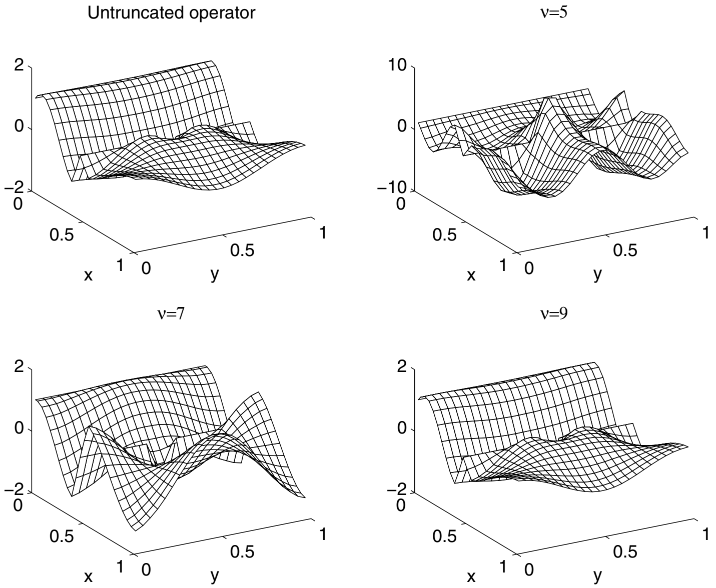


Fig. 15. Results for the 2D Helmholtz example, using the one step homogenized operator, truncated with different ν

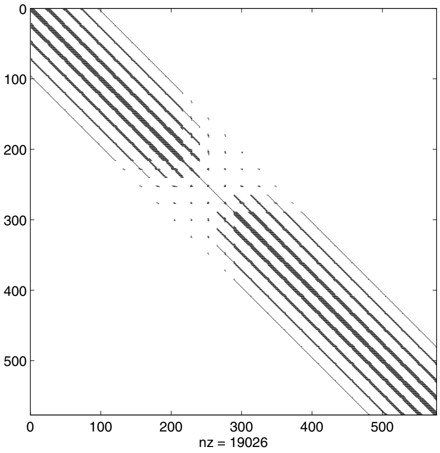


Fig. 16. Structure of the homogenized operator $\bar{L}$, after one homogenization step, for the 2D Helmholtz example. Elements larger than 0.1% of max value shown

Computer exercise

We end this tutorial with a suggested computer exercise which can be done e.g. in MATLAB. Consider the model elliptic boundary value problem in one dimension

$$-\partial_x g_\varepsilon(x)\partial_x u \quad f, \qquad u(0) \quad u_x(1) \quad 0. \tag{56}$$

Introduce the grid $\{x_k\}_{k \; 0}^{N-1}$ where $x_k \quad (k+1/2)h$ and $h \quad 1/N$. Let u_k approximate $u(x_k)$ and set $U \quad (u_0,\dots,u_{N-1})^T \in \mathbb{R}^N$, $F \quad (f(x_0),\dots,f(x_{N-1}))^T \in \mathbb{R}^N$. Then use standard central differences,

$$\frac{1}{h^2}\Delta_+ G \Delta_- U \quad F. \tag{57}$$

Here G is a diagonal matrix sampling $g_\varepsilon(x)$ in $x \quad kh$, $k \quad 0,\dots,N-1$ and $\Delta_\pm$ are the forward/backward difference operators. When approximating the boundary conditions by $u_{-1}+u_0 \quad 0$ and $u_N \quad u_{N-1}$ the difference operators have the matrix representations

$$\Delta_+ \quad \begin{pmatrix} -1 & 1 & & & \\ & -1 & 1 & & \\ & & \ddots & \ddots & \\ & & & -1 & 1 \\ & & & & -1 \end{pmatrix} \qquad \Delta_- \quad \begin{pmatrix} 2 & & & & \\ -1 & 1 & & & \\ & \ddots & \ddots & & \\ & & -1 & 1 & \\ & & & -1 & 1 \end{pmatrix}.$$

Then (57) is a second order method for (56).

For the numerical homogenization we will only use Haar wavelets. Also, we will not be concerned with computational costs, but rather with the approximation properties of the numerically homogenized operator. Let $N \quad 2^n$ for some n, sufficiently large to resolve the ε-scale in (56). In the Haar case, functions $u \in V_n$ are piecewise constant and their scaling coefficients $u_{n,k}$ satisfy $u_{n,k} = 2^{-n/2}u(x_k) \approx 2^{-n/2}u_k$. After appropriate rescaling we can therefore interpret (57) as a discretization in V_n,

$$L_n U_n = F_n, \qquad U_n, F_n \in V_n, \qquad L_n := 2^{2n}\Delta_+ G \Delta_-,$$

where U_n and F_n contain the scaling coefficients of $u(x)$ and $f(x)$ in V_n.

1. Let $g(x,y) = 0.55 + 0.45\sin(2\pi y)b(x-0.5)$, with $b(x) = \exp(-20x^2)$ and $f(x) = \sin(2\pi x) + 1/2$. Use $g_\varepsilon(x) = g(x, x/\varepsilon)$ for some small ε. Simulate the detailed equation and compare it with a) the constant coefficient equation using the arithmetic mean of $g_\varepsilon(x)$, i.e. $\bar{g} \equiv 0.55$, and b) the homogenized equation (38). (Determine, at least numerically, the homogenized coefficient $\hat{g}(x)$ using the formula in (38).)
2. Compute the numerically homogenized operator $\bar{L}_m$ using (44, 45) based on (30). Let $m < n$ be chosen so that the ε-scale averages out, i.e. $2^{-m} \geq \varepsilon$. (Make sure n is large enough though, $2^{-n} \ll \varepsilon$.) Examine the structure of $\bar{L}_m$ and verify that it its elements decay quickly away from the diagonal. Approximate $\bar{L}_m$ by truncation (52) to ν diagonals. Check how many diagonals you must keep to get an acceptable solution.
3. By (47) we can write the numerically homogenized operator on the same form as the original,
$$\bar{L}_m = 2^{2m}\Delta_+ H \Delta_-, \tag{58}$$
for some matrix H, which corresponds to the "effective material coefficient" at scale m. Compute H and verify that it is strongly diagonal dominant. Approximate H by truncation (52) and rebuild an approximation of $\bar{L}_m$ from the formula (58). How many diagonals do you need to keep now? Approximate H by band projection (53) to a single diagonal, i.e. "mass lumping," $H \approx \text{band}(H,1) = \text{diag}(H\mathbf{1})$, where $\mathbf{1}$ is the constant vector. How good is the corresponding solution? How does $\text{band}(H,1)$ compare with the original coefficient $g_\varepsilon(x)$ and with the homogenized coefficient $\hat{g}(x)$?
4. Test a few other types of coefficients:
   a) A three-scale system, e.g.
   $$g_\varepsilon(x) = \frac{1}{2}\left(g(x-0.1, x/\varepsilon_1) + g(x+0.1, x/\varepsilon_2)\right), \qquad \varepsilon_1 \ll \varepsilon_2 \ll 1.$$
   Use different $m = m_1, m_2$ to capture the behavior at different scales, i.e. $2^{-m_1} \geq \varepsilon_2 \gg 2^{-m_2} \geq \varepsilon_1$.
   b) A random coefficient,
   $$g_\varepsilon(x) = 0.1 + b(x-0.5)U(x),$$
   where $b(x)$ is as above, and $U(x)$ are uniformly distributed random numbers in the interval $[0,1]$ for each x. (Use MATLAB's `rand` command.)
   c) A localized coefficient,
   $$g_\varepsilon(x) = \begin{cases} 0, & |x-0.5| \geq \varepsilon, \\ \frac{1}{\varepsilon}, & |x-0.5| < \varepsilon. \end{cases}$$

References

1. B. Alpert. A class of bases in L^2 for the sparse representation of integral operators. *SIAM J. Math. Anal.*, 24(1):246–262, 1993.
2. J. Anderson. *Computational Fluid Dynamics, The Basics with Applications*. McGraw-Hill, 1995.
3. U. Andersson, B. Engquist, G. Ledfelt, and O. Runborg. A contribution to wavelet-based subgrid modeling. *Appl. Comput. Harmon. Anal.*, 7:151–164, 1999.
4. O. Axelsson. *Iterative Solution Methods*. Cambridge University Press, 1994.
5. G. Battle. A block spin construction of ondelettes. *Comm. Math. Phys.*, 110:601–615, 1987.
6. A. Bensoussan, J.-L. Lions, and G. Papanicolau. *Asymptotic Analysis for Periodic Structures*. North-Holland Publ. Comp., The Netherlands, 1978.
7. J. Bergh, F. Ekstedt, and M. Lindberg. *Wavelets*. Studentlitteratur, Lund, 1999.
8. G. Beylkin and M. Brewster. A multiresolution strategy for numerical homogenization. *Appl. Comput. Harmon. Anal.*, 2:327–349, 1995.
9. G. Beylkin, M. E. Brewster, and A. C. Gilbert. A multiresolution strategy for numerical homogenization of nonlinear ODEs. *Appl. Comput. Harmon. Anal.*, 5:450–486, 1998.
10. G. Beylkin, R. Coifman, and V. Rokhlin. Fast wavelet transforms and numerical algorithms I. *Comm. Pure Appl. Math.*, 44:141–183, 1991.
11. G. Beylkin and N. Coult. A multiresolution strategy for reduction of elliptic PDEs and eigenvalue problems. *Appl. Comput. Harmon. Anal.*, 5:129–155, 1998.
12. E. J. Candès and D. L. Donoho. Ridgelets: the key to higher-dimensional intermittency? *Phil. Trans. R. Soc. Lond. A.*, 357:2495–2509, 1999.
13. E. J. Candès and D. L. Donoho. Curvelets – a surprisingly effective nonadaptive representation for objects with edges. In A. Cohen, C. Rabut, and L. L. Schumaker, editors, *Curves and Surfaces*, pages 105–120. Vanderbilt Univ. Press, 2000.
14. T. Chan and T. Mathew. The interface probing technique in domain decomposition. *SIAM J. Matrix Anal. Appl.*, 13(1):212–238, January 1992.
15. A. Chertock and D. Levy. On wavelet-based numerical homogenization. *Multiscale Model. Simul.*, 3(1):65–88 (electronic), 2004/05.
16. C. K. Chui and J. Z. Wang. A cardinal spline approach to wavelets. *Proc. Amer. Math. Soc.*, 113:785–793, 1991.
17. A. Cohen, W. Dahmen, and R. A. DeVore. Adaptive wavelet methods for elliptic operator equations: Convergence rates. *Math. Comp.*, 70:27–75, 2001.
18. A. Cohen, I. Daubechies, and J. Feauveau. Bi-orthogonal bases of compactly supported wavelets. *Comm. Pure Appl. Math.*, 45:485–560, 1992.
19. A. Cohen, I. Daubechies, B. Jawerth, and P. Vial. Multiresolution analysis, wavelets and fast algorithms on an interval. *C. R. Acad. Sci. Paris Sér. I Math.*, I(316):417–421, 1993.
20. A. Cohen, S. M. Kaber, S. Müller, and M. Postel. Fully adaptive multiresolution finite volume schemes for conservation laws. *Math. Comp.*, 72(241):183–225, 2003.
21. R. R. Coifman, Y. Meyer, S. Quake, and M. V. Wickerhauser. Signal processing and compression with wave packets. In Y. Meyer, editor, *Proceedings of the International Conference on Wavelets, Marseille, 1989*. Masson, Paris, 1992.
22. C. Concus, G. H. Golub, and G. Meurant. Block preconditioning for the conjugate gradient method. *SIAM J. Sci. Stat. Comp.*, 6:220–252, 1985.
23. W. Dahmen. Wavelet and multiscale methods for operator equations. *Acta Numerica*, 6:55–228, 1997.
24. I. Daubechies. Orthonormal bases of compactly supported wavelets. *Comm. Pure Appl. Math.*, 41:909–996, 1988.

25. I. Daubechies. *Ten Lectures on Wavelets*. SIAM, 1991.
26. R. A. DeVore and B. J. Lucier. Wavelets. *Acta Numerica*, 1:1–56, 1991.
27. M. Dorobantu and B. Engquist. Wavelet-based numerical homogenization. *SIAM J. Numer. Anal.*, 35(2):540–559, April 1998.
28. W. E. Homogenization of linear and nonlinear transport equations. *Comm. Pure Appl. Math.*, 45(3):301–326, 1992.
29. B. Engquist and E. Luo. Convergence of a multigrid method for elliptic equations with highly oscillatory coefficients. *SIAM J. Numer. Anal.*, 34(6):2254–2273, 1997.
30. B. Engquist and O. Runborg. Wavelet-based numerical homogenization with applications. In T. J. Barth, T. F. Chan, and R. Haimes, editors, *Multiscale and Multiresolution Methods*, volume 20 of *Lect. Notes Comput. Sci. Eng.*, pages 97–148. Springer, Berlin, 2002.
31. J. Geronimo, D. Hardin, and P. R. Massopust. Fractal functions and wavelet expansions based on several scaling functions. *J. Approx. Theory*, 78(3):373–401, 1994.
32. A. C. Gilbert. A comparison of multiresolution and classical one-dimensional homogenization schemes. *Appl. Comput. Harmon. Anal.*, 5(1):1–35, 1998.
33. A. Grossmann and J. Morlet. Decompostion of Hardy functions into square integrable wavelets of constant shape. *SIAM J. Math. Anal.*, 15(4):723–736, 1984.
34. A. Haar. Zur Theorie der orthogonalen Funktionen-Systeme. *Math. Ann.*, 69:331–371, 1910.
35. A. Harten. Adaptive multiresolution schemes for shock computations. *J. Comput. Phys.*, 115(2):319–338, 1994.
36. T. Y. Hou and X. H. Wu. A multiscale finite element method for elliptic problems in composite materials and porous media. *J. Comput. Phys.*, 134(1):169–189, 1997.
37. T. Y. Hou and X. Xin. Homogenization of linear transport equations with oscillatory vector fields. *SIAM J. Appl. Math.*, 52(1):34–45, 1992.
38. T. J. R. Hughes. Multiscale phenomena: Green's functions, the Dirichlet-to Neumann formulation, subgrid, scale models, bubbles and the origins of stabilized methods. *Comput. Methods Appl. Mech. Engrg.*, 127:387–401, 1995.
39. B. Jawerth and W. Sweldens. An overview of wavelet based multiresolution analyses. *SIAM Rev.*, 36(3):377–412, 1994.
40. J. Keller. Geometrical theory of diffraction. *J. Opt. Soc. Amer.*, 52, 1962.
41. S. Knapek. Matrix-dependent multigrid-homogenization for diffusion problems. *SIAM J. Sci. Stat. Comp.*, 20(2):515–533, 1999.
42. J. Krishnan, O. Runborg, and I.G. Kevrekidis. Bifurcation analysis of nonlinear reaction-diffusion problems using wavelet-based reduction techniques. *Comput. Chem. Eng.*, 28:557–574, 2004.
43. P.-G. Lemarié. Une nouvelle base d'ondelettes de $L^2(\mathbb{R})$. *J. Math. Pures Appl.*, 67(3):227–236, 1988.
44. D. D. Leon. *Wavelet Operators Applied to Multigrid Methods*. PhD thesis, Department of Mathematics, UCLA, 2000.
45. S. G. Mallat. Multiresolution approximations and wavelet orthonormal bases of $L^2(\mathbb{R})$. *Trans. Amer. Math. Soc.*, 315(1):69–87, 1989.
46. Y. Meyer. Principe d'incertitude, bases Hilbertiennes at algèbres d'opérateurs. Séminaire Bourbaki 662, 1985–1986.
47. Y. Meyer. Ondelettes et fonctions splines. Séminaire équations aux dérivées partielles 6, École Polytechnique, 1986–1987.
48. Y. Meyer. *Wavelets: Algorithms and Applications*. SIAM, Philadelphia, PA, 1993.
49. N. Neuss, W. Jäger, and G. Wittum. Homogenization and multigrid. *Computing*, 66(1):1–26, 2001.

50. P.-O. Persson and O. Runborg. Simulation of a waveguide filter using wavelet-based numerical homogenization. *J. Comput. Phys.*, 166:361–382, 2001.
51. G. Strang and T. Nguyen. *Wavelets and Filter Banks*. Wellesley, Cambridge, 1996.
52. J. O. Strömberg. A modified Franklin system and higher order spline systems on $\mathbb{R}^n$ as unconditional bases for Hardy spaces. In Beckner et al., editor, *Conference on Harmonic Analysis in Honor of Antoni Zygmund*, volume II, pages 475–494, Chicago, 1981. Univ. of Chicago Press.
53. W. Sweldens. The lifting scheme: A construction of second generation wavelets. *SIAM J. Math. Anal.*, 29(2):511–546, 1997.
54. A. Taflove. *Computational Electromagnetics, The Finite-Difference Time-Domain Method*, chapter 10. Artech House, 1995.
55. C.-M. Wang. *Wavelet-Based Numerical Homogenization with Application to Flow in Porous Media*. PhD thesis, Department of Mathematics, UCLA, 2005.
56. D. Wilcox. *Turbulence Modeling for CFD*. DCW Industries, Inc., 1993.
57. J.-C. Xu and W.-C. Shann. Galerkin-wavelet methods for two-point boundary value problems. *Numer. Math.*, 63(1):123–142, 1992.

Multiscale Computations for Highly Oscillatory Problems

Gil Ariel[1,*], Björn Engquist[1,2,*], Heinz-Otto Kreiss[3], and Richard Tsai[1,*]

[1] The University of Texas at Austin, Austin, TX 78712,
{ariel,engquist,ytsai}@math.utexas.edu
[2] Department of Numerical Analysis, KTH, 100 44 Stockholm, Sweden
[3] Träskö–Storö Institute of Mathematics

Summary. We review a selection of essential techniques for constructing computational multiscale methods for highly oscillatory ODEs. Contrary to the typical approaches that attempt to enlarge the stability region for specialized problems, these lecture notes emphasize how multiscale properties of highly oscillatory systems can be characterized and approximated in a truly multiscale fashion similar to the settings of averaging and homogenization. Essential concepts such as resonance, fast-slow scale interactions, averaging, and techniques for transformations to non-stiff forms are discussed in an elementary manner so that the materials can be easily accessible to beginning graduate students in applied mathematics or computational sciences.

1 Introduction

Oscillatory systems constitute a broad and active field of scientific computations. One of the typical numerical challenges arises when the frequency of the oscillations is high compared to either the time or the spatial scale of interest. In this case, the cost for computations can typically become exceedingly expensive due to the need of sampling oscillations adequately by numerical discretizations over a relatively large domain. Several general strategies for dealing with oscillations can be found in literature, for example, asymptotic analysis [5, 22, 23], averaging [20, 29], envelope tracking [27, 28], explicit solutions to nearby oscillatory problems [25, 30, 31]. These strategies typically utilize some underlying structures, related to the oscillations, which are not oscillatory in the domain of interests. For example, the center or frequency of oscillators may vary slowly in time. Indeed, it is often the case that the quantities of interest are related to these non-oscillatory structures. Reduction in the computational costs is thus possible by avoiding direct resolution of the oscillations. Take geometrical optics [13, 21] for instance. The high frequency solution of the wave equation of the form $A(x,t)\exp(S(x,t)/\varepsilon)$ is computed via solutions of an eikonal equation for the phase S and transport equations for the amplitude A. Since eikonal and transport equations do not depend on the ε-scale oscillations, the cost of

[*] Ariel, Engquist, and Tsai are partially supported by NSF grant DMS-0714612.

computation is formally independent of the fast scale as well. These current lecture notes focus on building efficient multiscale numerical methods that only sample the fast oscillations. The sampled information is used to describe an effective time evolution for the system at longer time scales. The general approach underlying these methods come from the theory of averaging.

In these notes, we consider systems of ordinary differential equations (ODEs) which evolve on two widely separated time scales. Common examples are

1. Perturbed linear oscillations:

$$\varepsilon x' \quad Ax + \varepsilon g(x), \tag{1}$$

   where A is diagonalizable and has purely imaginary eigenvalues. The class of examples include Newton's equation of motion for perturbed harmonic oscillators

$$\varepsilon x'' \quad -\Omega^2 x + \varepsilon g(x),$$

   which are found in many applications. Here, the parameter ε, $0 < \varepsilon \ll 1$, characterizes the separation of time scales in the system.
2. Fully nonlinear oscillations induced from dissipation in the systems. Examples include Van der Pol oscillators and other relaxation oscillators.
3. Weakly coupled nonlinear oscillators that are close to a slowly varying periodic orbits. Examples include Van der Pol (with small damping) and Volterra-Lotka oscillators.

Efficient and accurate computations of oscillatory problems require significant knowledge about the underlying fast oscillations. Using either analytical or numerical methods, our general underlying principle is to model oscillations and sample their interactions. Very often, analytical methods do not yield explicit solutions, and suitable numerical methods need to be applied.

One of the current major thrusts is in developing numerical methods which allow long time computation of oscillatory solutions to Hamiltonian systems. The interest in such systems comes from molecular dynamics which attempts to simulate some underlying physics on a time scale of interest. These methods typically attempt to approximately preserve some analytical invariance of the solutions; e.g. the total energy of the system, symplectic structures, or the reversibility of the flow. Detailed reviews and further references on this active field of "geometric integration" can be found in [18] and [26].

The Verlet method and other similar geometric integrators are the methods of choice for many highly oscillatory simulations. They require, however, time steps that are shorter than the oscillatory wavelength ε and therefore cannot be used when ε is very small.

Exponential integrators allow for time steps that are longer than the oscillatory wavelength ε but they apply only to restricted classes of differential equations [18]. In a way that resembles the discussion of geometrical optics above since these methods explicitly use the exponential function to represent the leading terms in the oscillations. They work well for problems that are smooth perturbations of problems with constant coefficients.

Another general approach for dealing with multiscale phenomena computationally can be referred to as boosting [33]. The general idea is to artificially “twig” or “boost” the small parameter ε so that the stiffness of the problem is reduced. Known methods that fall into this category are Chorin’s artificial compressibility [7] and the Car-Parrinello method used in molecular dynamics [6].

This tutorial deviates from previous texts in that we do not rely or assume some specific properties or a special class of ODEs such as harmonic oscillations or Hamiltonian dynamics. Instead, the multiscale methods discussed here compute the effective behavior of the oscillatory system by integrating the oscillations numerically in *short time windows* and sampling their interactions by suitable averaging. Indeed, one of the main goals of this text is to make the ideas discussed above mathematically meaningful. Subsequent sections will define what we mean by the effective behavior of a given highly oscillatory system, describe the theory of averaging, the structure of our multiscale algorithms and its computational complexity.

The objectives of multiscale computations

One of the major challenges in problems involving multiple scales is that an accurate computations, attempting to resolve the finest scales involved in the dynamics may be computationally infeasible. In the classical numerical analysis for ODEs, the important elements are stability, consistency and ultimately convergence. In the standard theory, any stable consistent method converges to the analytical solution as the step size goes to zero. The errors depend on powers of the eigenvalues of the Jacobian of the ODE’s right hand side and the step size. However, in our multiscale setting, similar to the high frequency wave propagation or homogenization, we would like to consider the asymptotic cases when the frequency of the fastest oscillations, which is proportional to $1/\varepsilon$, tends to infinity, *before the step size is sent to zero.* Hence, we need to rethink what consistency and convergence means in the multiscale setting. One possibility is the following: let $E(t;\triangle,\varepsilon)$ denote the error of the numerical approximation at time t, using step size $\triangle$ and for problems with ε^{-1} oscillations, we consider the convergence of E for $0<\varepsilon<\varepsilon_\triangle$

$$\lim_{\triangle\longrightarrow 0}\left(\sup_{0<\varepsilon<\varepsilon_\triangle} E(t;\triangle,\varepsilon)\right).$$

In other words, with a prescribed error tolerance E, the same step size Δ can be used for small enough ε. While this notion of convergence may not be possible for the solutions of many problems, we may ask for the convergence of some functions or functionals of the solutions. Throughout the notes we discuss results from the prespective of a few key questions: what is the motivation for constructing a multiscale algorithm? What is being approximated? How does our multiscale approach differ from traditional numerical computations?

A first example, suggested by Germund Dahlquist, is the drift path of a mechanical alarm clock, moving due to fast vibrations when it is set off on a hard surface. If the drift path depends only locally in time on the fast oscillations, then it is reasonable to design a scheme that evolves the slowly changing averaged drift path by measuring the effects of the fast solutions only locally in time. Herein lies the possibility of reducing the computational complexity.

A second example is Kapitza's pendulum — a rigid pendulum whose pivot is attached to a strong periodic forcing is vibrating vertically with amplitude ε and frequency $1/\varepsilon$. When the oscillations are sufficiently fast, the pendulum swings slowly back and forth, *pointing upwards*, with a slow period that is practically independent of ε. Obviously, in the absence of the oscillatory forcing, the pendulum is only stable pointing downwards. Pyotr Kapitza (physics Nobel Laureate in 1978) used this example to illustrate a general stabilization mechanisms [17]. This, and similar simple dynamical systems are often used as example benchmark problems to study how different methods approximates highly oscillatory problems.

The following assumptions are made throughout these notes: in the fastest scale, the given system exhibits oscillations with amplitudes independent of ε, and that at a larger time scale, some slowly changing quantities can be defined by the oscillatory solutions of the system. To facilitate our discussion, we now present our model scenario described by the following two coupled systems. Consider a highly oscillatory system in $\mathbb{R}^{d_1}$ coupled with a slow system in $\mathbb{R}^{d_2}$:

$$\varepsilon x' \quad f(x,v,t)+\varepsilon g(x,v,t), \tag{2}$$

$$v' \quad h(v,x,t), \qquad x(0) \quad x_0 \in \mathbb{R}^{d_1}, \quad v(0) \quad v_0 \in \mathbb{R}^{d_2}. \tag{3}$$

We assume that x is highly oscillatory, and v is the slow quantity of interest. However, without proper information about x, v cannot be found. We are also interested in some slowly varying quantity that is being defined along the trajectories of x:

$$\beta' \quad \psi(\beta,t;x(\,\cdot\,)). \tag{4}$$

In a following section, we shall see that for very special initial conditions, the solutions of (2) may be very smooth and exhibit no oscillations. The problem of initialization, i.e. finding the suitable initial data so that the slowly varying solutions can be computed appear in meteorology. We refer the readers to the paper of Kreiss and Lorenz [25] for further reading on the theory for the slow manifolds. However, in many autonomous equations, e.g. linear equations, the only slowly varying solutions in the system are the equilibria of the system. For problems like the inverted pendulum, it is clear that the slowly varying solutions are not of interest. Then some complicated interactions between the oscillations must take place, and one must look into different strategies in order to characterize the effective influence of the oscillations in x in the evolution of v.

Our objective is to accurately compute the slowly changing quantity v in a long time scale (i.e. $0 \le t \le T$, for some constant T independent of ε). Furthermore, we wish to compute it with a cost that is at least sublinear to (ideally independent of) the cost for resolving all the fast oscillations in this time scale. In general, our objective may be achieved if fast oscillations are computed only in very short time intervals and yet the dynamics for those slowly changing quantities is consistently evolved. Figure 1 depicts two possible schematic structures for such an algorithm. In this section, we give a few examples of where such type of slowly changing quantities occur; these consist of systems with resonant modes and weakly perturbed Hamiltonians systems so that invariances are changing slowly. Furthermore, we shall see in the next section that these slowly changing quantities may be evaluated conveniently by short time averaging with suitable kernels.

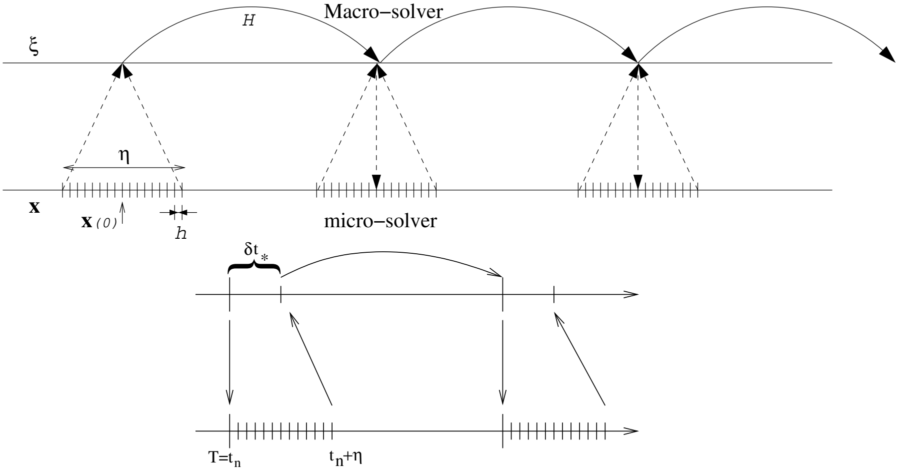


Fig. 1.

Let us briefly comment on the relation of these notes to the standard stiff ODE solvers for multiscale problems with transient solutions, [8, 19]. A typical example of such stiff problems is equation (1) where the eigenvalues of A are either negative or zero. The initial time steps are generally small enough to resolve the transient and of the type of the micro-solver in Fig. 1. After the transient, much longer time steps are possible as in the macro-solver in Fig. 1. Special properties of stiff ODE methods suppress the fast modes and only the slower modes need to be well approximated. Problems with highly oscillatory solutions are much harder to simulate since the fast modes are present for all times and may interact to give contributions to the slower modes.

1.1 Example oscillatory problems

Linear systems with imaginary eigenvalues

$$x' \quad i\lambda x, \qquad \lambda \in \mathbb{R}.$$

The solution is readily given by $x(t) \quad x(0)e^{i\lambda t}$. Note that this system is equivalent to the system in $\mathbb{R}^2$:

$$\begin{pmatrix} x \\ y \end{pmatrix}' \quad \begin{pmatrix} 0 & \lambda \\ -\lambda & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}.$$

Hamiltonian systems

Hamiltonian dynamics are defined by the partial derivatives of a Hamiltonian function $H(q,p)$ which represent the total energy of the system. Here, q is a generalized coordinate system and p the associated momentum. The equations of motion are given by

$$\begin{aligned} q' &\quad H_p(q,p), \\ p' &\quad -H_q(q,p), \end{aligned} \tag{5}$$

where H_p and H_q denote partial derivatives of H with respect to p and q, respectively. In Hamiltonian mechanics, $H(p,q) \quad \frac{1}{2}p^2 + V(q)$ and the dynamics defined in (5) yield Newton's equation of motion $q'' \quad -\nabla_q V(q)$. If $V(x)$ is a convex function then the solutions of this equation are typically oscillatory. An important class of equations of this type appear in molecular dynamics with pairwise potentials

$$H(p,q) \quad \frac{1}{2}\sum_{i\ \ 1}^{N} \frac{1}{m_i} p_i^T p_i + \frac{1}{2}\sum_{i,j\ \ 1}^{N} V_{ij}(|q_i - q_j|)$$

where p_i and q_i are components of the vectors p and q.
Notable examples are

$$V_{ij}(r) \quad \frac{-Gm_i m_j}{r} \qquad \text{(electric or gravitational potential)}$$

and

$$V_{ij}(r) \quad 4\varepsilon_{ij}\left(\left(\frac{\sigma_{ij}}{r}\right)^{12} - \left(\frac{\sigma_{ij}}{r}\right)^{6}\right), \qquad \text{(Lennard-Jones potential)}$$

for all $i \ / \ j$, etc.

Volterra–Lotka

This is a simplified model for the predator-prey problem in population dynamics. In this model, x denotes the population of a predator species while y denotes the population of a prey species

$$\begin{aligned} x' &\quad x\left(1 - \frac{y}{\nu}\right), \\ y' &\quad \frac{y}{\nu}(x-1). \end{aligned} \tag{6}$$

An example trajectory is depicted in Fig. 2.

Relaxation oscillators

The Van der Pol oscillator is another typical example of nonlinear oscillators. One version of the equation for a Van der Pol oscillator takes the form

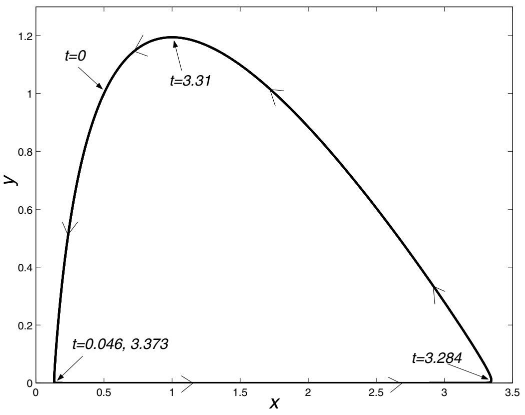


Fig. 2. The trajectory of the Volterra–Lotka oscillator (6) with $\nu = 0.01$, $x(0) = 0.5$ and $y(0) = 1$.

$$\begin{aligned} x' &= y-(x^2-1)x, \\ y' &= -x. \end{aligned}$$

This equation can be interpreted as a model of a basic RLC circuit, consisting of a resistor, inductor, and a capacitor; the state variable x corresponds to the current in the inductor and y the voltage in the capacitor. It can be shown that there is a unique periodic solution of this equation and other non-equilibrium solutions approach it as time increases. This periodic solution is called the limit cycle or the invariant manifold of the system. A general result for detecting periodic solutions for such type of systems on a plane is the Poincaré–Bendixson theorem, which says that if a compact limit set in the plane contains no equilibria, it is a closed orbit; i.e. it is a periodic trajectory of a solution.

As a second example, consider [9]

$$\begin{aligned} x' &= -1-x+8y^3, \\ y' &= \frac{1}{\nu}\left(-x+y-y^3\right), \end{aligned} \qquad (7)$$

where $0 < \nu \ll 1$. For small ν, trajectories quickly come close to the limit cycle defined by $-x+y-y^3 = 0$. The upper and lower branches of this cubic polynomial are stable up to the turning points at which $dx/dy = 0$. For any initial condition, the solution of (7) is rapidly attracted to one of the stable branches on an $\mathscr{O}(\nu)$ time scale. The trajectory then moves closely along the branch until it becomes unstable. At this point the solution is quickly attracted to the other stable branch. The trajectory is depicted in Fig. 3.

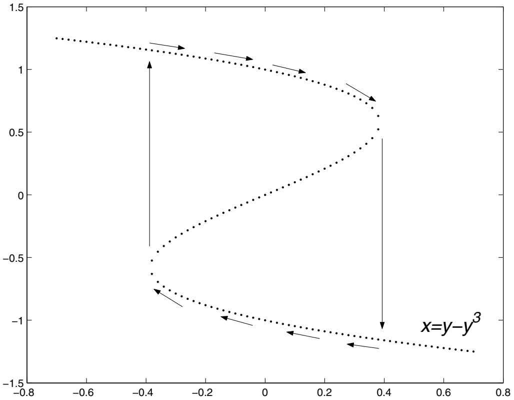


Fig. 3. The trajectory and slow manifold of the relaxation oscillator (7)

1.2 Invariance

Hamiltonian systems:

The Hamiltonian equations of motion (5) admit several invariances. First and foremost is the energy $H(p,q)$

$$\frac{d}{dt}H(p(t),q(t)) \quad H_p p' + H_q q' \quad 0$$

$$\Rightarrow H(p(t),q(t)) \quad H(p_{0,}q_0) \quad \text{const.}$$

Let us prove Liouville's theorem on volume preservation of Hamiltonian systems. Consider a smooth Hamiltonian $H(p,q)$. Let

$$\varphi_t(p_0,q_0) \quad \begin{pmatrix} p(t;p_0,q_0) \\ q(t;p_0,q_0) \end{pmatrix}.$$

Hence,

$$\frac{d}{dt}\frac{\partial\varphi_t}{\partial(p_0,q_0)} \quad \begin{pmatrix} -H_{pq} & H_{qq} \\ H_{pp} & H_{qp} \end{pmatrix} \begin{pmatrix} \frac{\partial\varphi_t}{\partial(p,q)} \\ (p(t),q(t)) \end{pmatrix}, \qquad q,p\in\mathbb{R}$$

$$\Rightarrow \frac{d}{dt}\det\frac{\partial\varphi_t}{\partial(p_0,q_0)} \quad 0. \tag{8}$$

Consider t as a parameter for the family of coordinate changes (diffeomorphisms) $\phi_t : (p_0,q_0) \mapsto (p(t;p_0,q_0),q(t,p_0,q_0))$. Then we have the following change of coordinates formula, for any fixed t,

$$\int_V f(p,q)dqdp \quad \int_U f(\phi_t(p_0,q_0))Jdp_0dq_0,$$

where $V \quad \phi_t(U)$ and

$$J: \quad \det \frac{\partial \phi_t}{\partial(p_0, q_0)}.$$

Thus, (8) implies that

$$\frac{dJ}{dt} \equiv 0$$

In particular, taking $f \equiv 1$ implies that U and V have the same volume in the phase (p,q)-space.

Volterra-Lotka oscillators

Let $I(u,v) \quad \log u - u + 2\log v - v$. Substituting (6) yields $(d/dt)I(u(t),v(t)) \quad 0$ for $t > 0$.

Relative phase between two linear oscillators

Let $u(t) \quad (\cos(t), \sin(t))$ and $v(t) \quad (\cos(t+\phi_0), \sin(t+\phi_0))$ be the solutions of some oscillators. Then

$$\xi(t) \quad u(t)\cdot v(t) \quad \cos\phi_0$$

measures the phase difference between $u(t)$ and $v(t)$ and remains constant in time.

In view of the above examples, the following questions naturally appear:

- Can one design numerical schemes so that the important invariances are preserved?
  What is the computational cost or benefit?
  How well do common numerical approximations preserve known invariances of interest and for what time scale?
- What is the importance of preserving invariances? How can this notion be quantified?
- How do small perturbations affect the invariances? For example, in the following linear system which conserves energy for $\varepsilon \quad 0$: $x'' \quad -\omega^2 x + \varepsilon\cos(\lambda t)$. How does weak periodic forcing affect the energy? At what time scale does the forcing become important? Can these effects be computed efficiently?

Some aspects of these issues and others are discussed in [18] and [26].

1.3 Resonance

Resonances among oscillations appear in many situations. For example, in pushing a child on a playground swing. It is intuitively clear that unless the swing is pushed at a frequency which is close to the natural oscillation frequency of the swing, the child will be annoyed. However, when the swing is pushed at the right frequency, the amplitude of the swing is gradually increasing. In this subsection, we review a few basic examples of resonance.

Resonance in a forced linear spring

We start with a linear spring under periodic forcing,

$$x'' \quad -\omega^2 x + \cos\lambda t.$$

Rewriting into a first order system, we obtain

$$\frac{d}{dt}\begin{pmatrix} x \\ x' \end{pmatrix} \quad \begin{pmatrix} 0 & 1 \\ -\omega^2 & 0 \end{pmatrix}\begin{pmatrix} x \\ x' \end{pmatrix} + \begin{pmatrix} 0 \\ \cos\lambda t \end{pmatrix},$$

with initial condition $\begin{pmatrix} x_0 \\ x_0' \end{pmatrix}$. One can show that the solution operator for the homogeneous problem is

$$S_t \quad \begin{pmatrix} \cos\omega t & \omega^{-1}\sin\omega t \\ -\omega\sin\omega t & \cos\omega t \end{pmatrix}$$

so that the solution for the inhomogeneous problem is

$$\begin{pmatrix} x \\ x' \end{pmatrix} \quad S_t\begin{pmatrix} x_0 \\ x_0' \end{pmatrix} + \int_0^t S_{t-s}\begin{pmatrix} 0 \\ \cos\lambda s \end{pmatrix} ds$$

$$\Rightarrow \begin{pmatrix} x \\ x' \end{pmatrix} \quad S_t\begin{pmatrix} x_0 \\ x_0' \end{pmatrix} + \int_0^t \begin{pmatrix} \omega^{-1}\sin(\omega t - \omega s)\cos\lambda s \\ \cos(\omega t - \omega s)\cos\lambda s \end{pmatrix} ds.$$

When $\lambda \quad \omega$, resonance happens. More precisely, we see that

$$x(t) \quad x_0\cos\omega t + \frac{x_0'}{\omega}\sin\omega t + \frac{1}{\omega}\int_0^t \sin(\omega t - \omega s)\cos\omega s\, ds$$

$$\Rightarrow x(t) \quad x_0\cos\omega t + \frac{x_0'}{\omega}\sin\omega t + \frac{t}{2}\sin\omega t.$$

In addition,

$$\begin{aligned}\int_0^t \sin(\omega t - \omega s)\cos\omega s\, ds \quad & \int_0^t (\sin\omega t\cos\omega s - \sin\omega s\cos\omega t)\cos\omega s\, ds \\ & \sin(\omega t)\int_0^t \cos^2(\omega s)ds - \cos(\omega t)\int_0^t \sin(\omega s)\cos(\omega s)ds \\ & \sin(\omega t)\int_0^t \frac{1}{2}(1+\cos 2\omega s)ds - \cos(\omega t)\int_0^t \frac{1}{2}\sin(2\omega s)ds \\ & \underbrace{\frac{t}{2}\sin\omega t}_{\text{result of resonance}} + \frac{1}{2}\int_0^t (\sin(\omega t)\cos(2\omega s) - \cos(\omega t)\sin(2\omega s))\, ds \\ & \underbrace{\frac{t}{2}\sin\omega t}_{\text{result of resonance}} + \underbrace{\frac{1}{2}\int_0^t \sin(\omega t - 2\omega s)ds}_{=0}.\end{aligned}$$

If $\lambda^2 \,/\, \omega^2$,we have

$$x \quad \left(x_0 - \frac{1}{\omega^2 - \lambda^2}\right)\cos\omega t + \frac{x_0'}{\omega}\sin\omega t + \frac{1}{\omega^2 - \lambda^2}\cos\lambda t.$$

Exercise 1. Compute the solution of the forced oscillation under friction: for $\mu > 0$

$$x'' \quad -2\mu x' - \omega^2 x + \cos\lambda t.$$

Show that if $\lambda \quad \omega$, the amplitude of the oscillations remains bounded and is largest when $\lambda \quad \sqrt{\omega^2 - 2\mu^2}$.

Resonance in first order systems

Consider,

$$x' \quad \frac{i}{\varepsilon}\Lambda x + f\left(x, \frac{t}{\varepsilon}\right), \qquad x(0,\varepsilon) \quad x_0,$$

where Λ is a diagonal matrix. We make the substitution:

$$x(t) \quad e^{\frac{i}{\varepsilon}\Lambda t} w(t), \qquad w(0) \quad x_0,$$

and obtain the corresponding equation for w:

$$w' \quad e^{-\frac{i}{\varepsilon}\Lambda t} f\left(e^{\frac{i}{\varepsilon}\Lambda t} w, \frac{t}{\varepsilon}\right), \qquad w(0) \quad x_0.$$

The simplest type of resonance can be obtained by taking $f(x,t/\varepsilon) \quad x$, i.e., $w' \quad w$. We see that the resonance occurs due to the linearity of f in x, and that it results in $|x|$ changing at a rate independent of ε. If $f(x,t/\varepsilon) \quad f_l(x) + \exp(it/\varepsilon)$ and one of the diagonal elements of Λ is 1, then similar to the resonance in the forced linear spring, the resonance with the forcing term contributes a linear in time growth of $|x|$.

Resonances may occur due to nonlinear interaction. Following the above example, take

$$\Lambda \quad \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix} \qquad \text{and} \qquad f(x,t) \quad \begin{pmatrix} 0 \\ -x_1^4 x_2^{-1} \end{pmatrix}.$$

Hence,

$$x \quad \begin{pmatrix} e^{it/\varepsilon} w_1 \\ e^{2it/\varepsilon} w_2 \end{pmatrix} \quad \Rightarrow$$

$$w' \quad \begin{pmatrix} e^{-it/\varepsilon} & 0 \\ 0 & e^{-2it/\varepsilon} \end{pmatrix} \begin{pmatrix} 0 \\ -e^{4it/\varepsilon} e^{-2it/\varepsilon} w_1^4 w_2^{-1} \end{pmatrix} \quad \begin{pmatrix} 0 \\ -w_1^4 w_2^{-1} \end{pmatrix}.$$

Again, due to the resonance in the system, $|x_2|$ is changing at a rate that is independent of ε.

2 Slowly varying functions of the solutions

In this section we shall study the effect of non-linear interactions. We excerpt important results from [24] and [1, 2, 14].

2.1 Problems with dominant fast linear oscillations and nonlinear interactions

We start with a number of examples.

$$x' = \frac{i\lambda}{\varepsilon}x + x^2, \qquad x(0) = x_0. \tag{9}$$

The solution of (9) can be obtained explicitly. Introducing a new variable *x by*

$$x = e^{\frac{i\lambda}{\varepsilon}t}w$$

gives us a new equation whose right hand side is bounded independent of ε

$$w' = e^{\frac{i\lambda}{\varepsilon}t}w^2, \qquad w(0) = w_0 = x_0.$$

The solution is readily given by

$$w(t) = \frac{1}{\frac{1}{w_0} + \frac{i\varepsilon}{\lambda}(e^{\frac{i\lambda}{\varepsilon}t} - 1)} = \frac{w_0}{1 + \frac{i\varepsilon}{\lambda}w_0(e^{\frac{i\lambda}{\varepsilon}t} - 1)}.$$

As a result,

$$w(t) = w_0\left(1 - \frac{i\varepsilon}{\lambda}w_0(e^{\frac{i\lambda}{\varepsilon}t} - 1)\right) + \mathscr{O}(\varepsilon^2). \tag{10}$$

Thus, the nonlinear term changes the solution only by $\mathscr{O}(\varepsilon)$ in arbitrarily long time intervals. In Sect. 3.2, we will show that $w(t)$ is close to an effective equation from averaging:

$$\bar{w}' = \bar{f}(w), \qquad \bar{f}(w) = \int_0^1 e^{it}w^2 dt = 0.$$

Hence, $\bar{w}(t) = w_0$. Indeed, we see that $|w(t) - \bar{w}(t)| = |w(t) - w_0| \le C_0\varepsilon$ for $0 \le t \le T_1$.

An alternative solution method involves a procedure which is easier to generalize. From (9) we have,

$$\begin{aligned} w(t) - w_0 &= \int_0^t e^{\frac{i\lambda}{\varepsilon}s}w^2 ds = -\frac{i\varepsilon}{\lambda}e^{\frac{i\lambda}{\varepsilon}s}w^2|_0^t + \frac{2i\varepsilon}{\lambda}\int_0^t e^{\frac{i\lambda}{\varepsilon}s}ww' ds \\ &= -\frac{i\varepsilon}{\lambda}\left(e^{\frac{i\lambda}{\varepsilon}t}w^2(t) - w_0^2\right) + \frac{2i\varepsilon}{\lambda}\int_0^t e^{i\frac{2\lambda}{\varepsilon}s}w^3 ds. \end{aligned} \tag{11}$$

Integrating by parts again yields an integral equation for $w(t)$

$$w(t) + \frac{i\varepsilon}{\lambda}e^{\frac{i\lambda}{\varepsilon}t}w^2(t) - \frac{4\varepsilon^2}{\lambda^2}e^{i\frac{2\lambda}{\varepsilon}}w^3(t) = w_0 + \frac{i\varepsilon}{\lambda}w_0^2 - \frac{4\varepsilon^2}{\lambda^2}w_0^3 + \frac{4\varepsilon^2}{\lambda^2}\int_0^t e^{i\frac{3\lambda}{\varepsilon}s}w^4(s)ds.$$

The solution $w(t)$ can then be constructed using fixed point iterations

$$w_{(k+1)}(t) = F(w_{(k)}, w_0, t) + \frac{4\varepsilon^2}{\lambda^2}\int_0^t e^{i\frac{3\lambda}{\varepsilon}s}w_{(k)}^4(s)ds, \qquad k = 0, 1, 2, \cdots, \tag{12}$$

where $w_{(0)} \quad w_0$ and

$$F(w,w_0,t) \quad w_0 - \frac{i\varepsilon}{\lambda} e^{\frac{i\lambda}{\varepsilon}t} w^2(t) + \frac{i\varepsilon}{\lambda} w_0^2 + \frac{4\varepsilon^2}{\lambda^2} e^{i\frac{2\lambda}{\varepsilon}} w^3(t) - \frac{4\varepsilon^2}{\lambda^2} w_0^3.$$

By induction, one can show that the iterations converge for $t \geq 0$, $0 \leq \varepsilon \leq \varepsilon_0$ and

$$w(t) + \frac{i\varepsilon}{\lambda} e^{\frac{i\lambda}{\varepsilon}t} w^2(t) \quad w_0 + \frac{i\varepsilon}{\lambda} w_0^2 + \mathscr{O}(\varepsilon^2).$$

From (11), we have $w(t) \quad w_0 + \mathscr{O}(\varepsilon)$. Hence,

$$w(t) \quad w_0 \left(1 - \frac{i\varepsilon}{\lambda} w_0 \left(e^{\frac{i\lambda}{\varepsilon}t} - 1\right)\right) + \mathscr{O}(\varepsilon^2). \tag{13}$$

Now, consider

$$x' \quad \frac{i\lambda_1}{\varepsilon} x + y, \qquad y' \quad \frac{i\lambda_2}{\varepsilon} y + y^2.$$

Changing variables to

$$x \quad e^{\frac{i\lambda_1}{\varepsilon}t} u, \qquad y \quad e^{\frac{i\lambda_2}{\varepsilon}t} w$$

yields

$$u' \quad e^{i(-\lambda_1+\lambda_2)t/\varepsilon} w, \quad w' \quad e^{\frac{i\lambda_2}{\varepsilon}t} w^2.$$

From (10), we can obtain an asymptotic expansion for w:

$$w \quad w_0 + \sum_{j \quad 1}^{\infty} \varepsilon^j \beta^j e^{i\frac{j\lambda_2}{\varepsilon}t}.$$

Therefore:

$$u' \quad e^{\frac{i}{\varepsilon}(\lambda_2-\lambda_1)t} w_0 + \sum_{j \quad 1}^{\infty} \varepsilon^j \beta^j e^{\frac{i}{\varepsilon}((j+1)\lambda_2-\lambda_1)t}.$$

If $\nu\lambda_2 - \lambda_1 \;/\; 0$ for all $\nu \quad 1,2,\ldots$, then

$$u(t) \quad u_0 + \mathscr{O}(\varepsilon).$$

However, if $\nu\lambda_2 \quad \lambda_1$, then resonance occurs and

$$u(t) \quad \begin{cases} u_0 + \varepsilon^{\nu-1}\beta^{\nu-1}t, & \text{if } \nu > 1, \\ u_0 + tw_0, & \text{if } \nu \quad 1. \end{cases}$$

Thus, the solution is not bounded for all time.

As a generalization, consider the system

$$x' \quad \frac{i}{\varepsilon} \Lambda x + P(x), \qquad x(0) \quad x_0, \tag{14}$$

where

$$\Lambda \quad \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_d \end{pmatrix}, \qquad \lambda_1,\cdots,\lambda_d \in \mathbb{R},$$

and $P \quad (p_1(x),\cdots,p_d(x))$ is a vector of polynomials in $x \quad (x_1,\cdots,x_d)$. Let

$$x \quad e^{\frac{i}{\varepsilon}\Lambda t}w.$$

We then have

$$w' \quad e^{-\frac{i}{\varepsilon}\Lambda t}P\left(e^{\frac{i}{\varepsilon}\Lambda t}w\right), \qquad w(0) \quad w_0 \quad x_0. \tag{15}$$

The right hand side of (15) consists of expressions of the form

$$e^{\frac{i}{\varepsilon}(\Sigma m_j\lambda_j)t}p(w), \tag{16}$$

where the m_j are integers and p is a polynomial in w. There are two possibilities.

1. $\tau \quad \sum m_j\lambda_j \quad 0$ for some terms. We call these terms the resonant modes. In this case (15) has the form

$$w' \quad Q_0(w)+Q_\varepsilon\left(\frac{t}{\varepsilon},w\right), \tag{17}$$

where Q_0 contains the terms corresponding to resonant modes, and Q_ε the remaining terms involving oscillatory exponentials. One can show that the solution of

$$\bar{w}' \quad Q_0(\bar{w}), \qquad \bar{w}(0) \quad w_0 \quad x_0, \tag{18}$$

is very close to w for a long time; i.e.

$$|w(t)-\bar{w}(t)| \le C_0\varepsilon, \qquad 0 \le t \le T_1, \tag{19}$$

and in general $w(t)$ does not stay close to the initial value w_0.

2. $\tau \quad \sum m_j\lambda_j \; / \; 0$ for all terms. No resonance occurs in the system. The solution stays close to the initial value:

$$w(t) \quad w_0+\mathscr{O}(\varepsilon). \tag{20}$$

We remark here that the term

$$f\left(\frac{t}{\varepsilon},w\right) \quad e^{-\frac{i}{\varepsilon}\Lambda t}P\left(e^{\frac{i}{\varepsilon}\Lambda t}w\right)$$

is in general not strictly periodic, even though it is composed of many highly oscillatory terms. Nonetheless, the self averaging effect of the highly oscillatory terms can be observed using integration by parts:

$$\begin{aligned} w(t) \quad & w_0 + \sum_\tau \int_0^t e^{\frac{i\tau}{\varepsilon}\xi} p_\tau(w)d\xi \\ & w_0 - i\varepsilon \sum_\tau \frac{1}{\tau} e^{\frac{i\tau}{\varepsilon}\xi} p_\tau(w)|_0^t + i\varepsilon \sum_\tau \frac{1}{\tau} \int_0^t e^{\frac{i\tau}{\varepsilon}\xi} \frac{\partial p_\tau}{\partial p} w' d\xi \\ & w_0 - i\varepsilon \sum_\tau \frac{1}{\tau} e^{\frac{i\tau}{\varepsilon}\xi} p_\tau(w)|_0^t + i\varepsilon \sum_\tau \frac{1}{\tau} \int_0^t e^{\frac{i\tau}{\varepsilon}\xi} \tilde{p}_\tau(w)' d\xi. \end{aligned} \tag{21}$$

The integrals in (21) are over terms of type (16) and we can therefore repeat the above arguments. If some of the terms are not of exponential type, then they will, in general, be of order $\mathscr{O}(\varepsilon t)$. For $\varepsilon t \ll 1$ we can replace (21) by

$$\tilde{w}(t) \quad w_0 - i\varepsilon \sum_\tau \frac{1}{\tau} e^{\frac{i\tau}{\varepsilon}\xi} p_\tau(\tilde{w})|_0^t,$$

i.e.,

$$\tilde{w}(t) \quad w_0 + \mathscr{O}(\varepsilon) \qquad \text{for}\, \varepsilon t \ll 1.$$

A more accurate result is

$$\tilde{w}(t) \quad w_0 - i\varepsilon \sum_\tau \frac{1}{\tau} \left(e^{\frac{i\tau}{\varepsilon}\xi} - 1\right) p_\tau(y_0) + i\varepsilon \tilde{p}_0(y_0)t + \mathscr{O}(\varepsilon^2 t^2).$$

If all the terms are of exponential type, then we can use integration by parts to reduce them at least to order $\mathscr{O}(\varepsilon^2 t)$. We obtain the following theorem.

Theorem 1. *Assume that for all integers α_j the linear combinations*

$$\sum \alpha_j \lambda_j \ / \ 0.$$

Then

$$\tilde{w} \quad w_0 + \mathscr{O}(\varepsilon)$$

in time intervals $0 \le t \le T$. $T \quad \mathscr{O}(\varepsilon^{-p})$ for any p.

There are no difficulties in generalizing the result and techniques to more general equations

$$x' \quad \frac{1}{\varepsilon} \Lambda(t)x + P(x,t).$$

Here $\Lambda(t)$ is slowly varying and $P(x,t)$ is a polynomial in x with slowly varying coefficients in time.

Remark 1. We see that without the presence of resonance, the highly oscillatory solution x of system (14) stays closely to

$$e^{i\frac{\Lambda}{\varepsilon}t} x_0,$$

for a very long time. Regarding to our ultimate goal of developing efficient algorithms, we may conclude that if no resonance occurs in the system, no computation is needed, since $e^{i\frac{\Lambda}{\varepsilon}t} x_0$ is already a good approximation to the solution.

However, if resonance occurs, then the "envelop" of the oscillations in the solution, $x(t)$ changes non-trivially. In this case, efficient algorithms can be devised from solving the initial value problem of equation (17):

$$w' = Q_0(w) + Q_\varepsilon\left(\frac{t}{\varepsilon}, w\right), \qquad w(0) = x_0.$$

As we showed above, one may even simply drop the oscillatory term and solve an equation that is completely independent of the fast scale:

$$w' = Q_0(w), \qquad w(0) = x_0$$

and still obtain accurate approximations. We refer the readers to the paper of Scheid [31] for an interesting algorithm that explores this special structure of the right hand side. In a later part of these notes, we shall first show that the term Q_ε can be easily "averaged" out without even using its explicit form. This may prove to be very useful for designing our multiscale algorithms for more complicated systems.

Exercise 2. Show that the fixed-point iterations defined in (12) converge for arbitrary time intervals. Follow the steps:

1. For $0 \le t < T$, the difference $e_k(t) := w_{(k)}(t) - w_{(k-1)}(t)$ converge to 0 as k approaches infinity.
2. Show that $w_{(k)}$ is uniformly bounded for $k = 1, 2, \ldots$.
3. Establish the estimate in (13).
4. Arguments in the previous steps can be repeated to extend the solution to larger time intervals.

2.2 Slowly varying solutions

Consider

$$\begin{aligned} \varepsilon x' &= (A(t) + \varepsilon B(x,t))x + F(t), \\ x(0) &= x_0, \end{aligned} \tag{22}$$

where $0 < \varepsilon \ll 1$ and $A(t)$, $F(t)$ satisfy the same conditions as in the linear case, i.e. A, A^{-1}, F and their derivatives are of order one. Also, for x of order one, B and its derivatives with respect to x and t are of order one, and

$$|B| \le C|x|,$$

for some constant $C > 0$. Formally, taking $\varepsilon = 0$, yields the leading order equation $Ax + f = 0$. Denoting the solution $\Phi_0 = -A^{-1}F$ we substitute

$$x = \Phi_0(t) + x_1,$$

and obtain by Taylor expansion

$$\begin{aligned}\varepsilon x_1' &\quad (A(t)+\varepsilon B(x_1+\Phi_0,t))\,(x_1+\Phi_0)+F(t)-\varepsilon\Phi_0'\\ &\quad (A_1(t)+\varepsilon B_1(x_1,t))\,x_1+\varepsilon F_1(t),\end{aligned}$$

where B_1 has the same properties as B and

$$A_1(t) \quad A(t)+\mathscr{O}(\varepsilon), \qquad F_1 \quad B(\Phi_0,t)\Phi_0-\Phi_0'.$$

Thus the new system is of the same form as the original one with the forcing function reduced to $\mathscr{O}(\varepsilon)$. Repeat the process p times yields

$$\begin{aligned}x &\quad \sum_{\nu\ 0}^{p-1}\varepsilon^{\nu}\Phi_{\nu}+x_p,\\ \varepsilon x_p' &\quad (A_p(t)+\varepsilon B_p(x_p,t))x_p+\varepsilon^p F_p, \quad A_p \quad A+\mathscr{O}(\varepsilon),\\ x_p(0) &\quad x(0)-\sum_{\nu\ 0}^{p-1}\varepsilon^{\nu}\Phi_{\nu}(0).\end{aligned} \tag{23}$$

Therefore, we have

Theorem 2. *The solution of (22) has p derivatives bounded independently of ε if and only if*

$$x(0) \quad \sum_{\nu\ 0}^{p-1}\varepsilon^{\nu}\Phi_{\nu}(0)+\mathscr{O}(\varepsilon^p),$$

i.e. $x(0)$ is, except for terms of order $\mathscr{O}(\varepsilon^p)$, uniquely determined.

If F and all its derivates vanish at $t \quad 0$, then the initial condition

$$x(0) \quad 0$$

defines a solution for which any number of derivatives are bounded independent of ε. We can construct such a solution even if F and its derivatives do not vanish at t=0, provided we can extend F smoothly to negative t. If the solution operator of the linearized problem,

$$\varepsilon v_p' \quad A_p(t)v,$$

is bounded, then $x_p \quad \mathscr{O}(\varepsilon^{p-j})$ in time intervals of length $\mathscr{O}(\varepsilon^{-j})$.

Generalizing (22), consider

$$\begin{aligned}\varepsilon x' &\quad (A(t)+\varepsilon C(v,t)+\varepsilon B(v,x,t))x+F(t),\\ v' &\quad D(v,x,t)x+G(v,t).\end{aligned} \tag{24}$$

Here A, A^{-1}, B, C, D, F, G and their derivatives with respect to x, v, t are of order $\mathscr{O}(1)$, if x, v, t are of order $\mathscr{O}(1)$. Following the same reasoning as before, substitute

$$x \quad -A^{-1}(t)F(v,t)+x_1$$

to obtain a system of the same form with F replaced by εF_1. Repeating the process p times yields

$$\begin{aligned} \varepsilon x_p' &\quad (A(t)+\varepsilon C_p(v,t)+\varepsilon B_p(v,x_p,t))x_p+\varepsilon^p F_p(t), \\ v' &\quad D_p(v,x_p,t)x_p+G_p(v,t). \end{aligned} \tag{25}$$

We conclude the following:

Theorem 3. *The solution of (24) has p time derivative which are bounded independently of ε if we choose*

$$x_p(0) \quad \mathscr{O}(\varepsilon^p),$$

i.e., $x(0)$ is, except for terms of order $\mathscr{O}(\varepsilon^p)$, uniquely determined by $v(0)$.

Generalizations. More generally, we can consider systems

$$\varepsilon w' \quad h(w,t).$$

If there is a solution $w(t)$ with $w'(t) \quad \mathscr{O}(1)$, then $h(w(t),t) \quad \mathscr{O}(\varepsilon)$. This suggests the existence of a C^∞-function $\phi(t)$ with

$$h(\phi(t),t) \quad 0, \qquad t \geq 0.$$

Introducing the new variable

$$\tilde{w} \quad w-\phi,$$

one obtains

$$\begin{aligned} \varepsilon \tilde{w}' &\quad h(\tilde{w}+\phi,t)-h(\phi,t)-\varepsilon\phi'(t) \\ &\quad (M(t)+N(\tilde{w},t))\,\tilde{w}-\varepsilon\phi'(t) \end{aligned}$$

where

$$M(t) \quad h_w(\phi(t),t), \qquad |N(\tilde{w},t)| \leq \text{const.}\ |\tilde{w}|.$$

If we further assume that

$$\tilde{w}(0) \quad w(0)-\phi(0) \quad \varepsilon z_0, \qquad z_0 \quad \mathscr{O}(1),$$

then we can rescale the equation for $\tilde{w}$ by introducing a new variable, $z \quad \varepsilon^{-1}\tilde{w}$. One obtains for $z(t)$

$$\varepsilon z' \quad (M(t)+\varepsilon Z(z,t))\,z-\phi'(t).$$

Since we are interested in highly oscillatory problems, assume that $M(t)$ has m purely imaginary eigenvalues which are independent of t. Denote

$$\kappa_j \quad i\mu_j, |\mu_j| \geq \delta > 0, \qquad j \quad 1,\ldots,m,$$

and n eigenvalues

$$\kappa_{m+1} \quad \ldots \quad \kappa_{m+n} \quad 0.$$

Without loss of generality, assume

$$M \quad \begin{pmatrix} A & 0 \\ 0 & 0 \end{pmatrix}, \qquad |A^{-1}| \quad \mathscr{O}(1),$$

where A is an $m \times m$ matrix with eigenvalues κ_j, $j \quad 1,\ldots,m$. If we partition z accordingly,

$$z \quad \begin{pmatrix} x \\ v \end{pmatrix},$$

then we obtain

$$\begin{aligned} \varepsilon x' & \quad \left(A + \varepsilon Z^I(v,x,t)\right)x - \left(\phi'\right)^I, \\ v' & \quad Z^{II}(v,x,t) - \frac{1}{\varepsilon}\left(\phi'\right)^{II}. \end{aligned}$$

If $(\phi')^{II} \quad \mathscr{O}(\varepsilon)$, then the resulting system has the form (24).

2.3 Interaction between the fast and the slow scales

Continuing our discussion and ignoring the terms that are higher order in ε, we consider the following model equation:

$$\begin{aligned} \varepsilon x' & \quad (A(t) + \varepsilon C(v,t) + \varepsilon B(v,x,t))x, \\ v' & \quad D(v,x,t)x + G(v,t), \\ x(0) & \quad x_0, \quad v(0) \quad v_0. \end{aligned} \tag{26}$$

We obtain the slow solution v_s, if we set $x_0 \quad 0$, i.e.,

$$v_S' \quad G(v_S,t), \qquad v_S(0) \quad v_0, \qquad x \equiv 0. \tag{27}$$

Let us make the following assumption.

Assumption 1. The solution operators S_1,S_2 of

$$v_L' \quad \frac{\partial G}{\partial v}(v_S)v_L$$

and

$$\varepsilon x_L' \quad (A(t) + \varepsilon C(v_S,t))x_L,$$

respectively, are uniformly bounded.

Here, the solution operator $S_1(t,\, s)$ for $t > s$ maps $\mathrm{V}_L(s)$ to $\mathrm{V}_L(t)$, and $\mathrm{S}_2(t,\, s)$ acts the same way for X_L.

If $x_0 \;/\; 0$, then the slow solution will be perturbed and we want to estimate $v - v_S$. We start with a rather crude estimate. We assume that x_0 is small and want to show

$$|v - v_S| \quad \mathscr{O}(|x_0|^2 t + \varepsilon|x_0|)$$

in time intervals $0 \leq t \leq T$ with $T \ll |x_0|^{-1}$. We linearize (26) around $v \quad v_s$ and $x \quad 0$. Let $v \quad v_S + v_L$, $x \quad x_L$, then the linearized equations have the form

$$\begin{aligned} \varepsilon x_L' &\quad (A(t)+\varepsilon C(v_S,t))x_L, \\ v_L' &\quad \tilde{D}(v_S,t)x_L + \frac{\partial G}{\partial v}(v_S)v_L, \quad \tilde{D} \quad D(v_S,0,t), \\ x_L(0) &\quad x_0,\ v_L(0) \quad v_0. \end{aligned}$$

By assumption

$$|x_L| \leq \text{const.}|x_0|.$$

Duhamel's principle and integration by parts gives us

$$\begin{aligned} v_L(t) &\quad \int_0^t S_1(t,\xi)\tilde{D}x_L d\xi \\ &\quad \varepsilon\int_0^t S_1(t,\xi)\tilde{D}(A+\varepsilon C)^{-1}x_L' d\xi \\ &\quad \varepsilon S_1(t,\xi)\tilde{D}(A+\varepsilon C)^{-1}x_L|_0^t - \varepsilon\int_0^t \frac{\partial}{\partial\xi}(S_1(t,\xi)\tilde{D}(A+\varepsilon C)^{-1})x_L' d\xi. \end{aligned}$$

The last integral can be treated in the same way. Therefore, Assumption 1 gives us, for any p,

$$|v_L(t)| \leq \text{const.}\left(\varepsilon|x_0| + \mathcal{O}(\varepsilon^p t)\right).$$

Assume now that A is constant, has distinct purely imaginary eigenvalues and that B,D are polynomials in x. Our goal is to give conditions such that our estimate will be improved to

$$|v - v_S| \quad \mathcal{O}(\varepsilon x_0) \text{ in time intervals } 0 \leq t \leq T, T \ll (\varepsilon|x_0|)^{-1}. \tag{28}$$

Without restriction we can assume that the system has the simplified form

$$\varepsilon x' \quad (i\Lambda + \varepsilon\Lambda_1(t) + \varepsilon B(x,t))x, \tag{29}$$

$$v' \quad D(x,t)x + G(v,t), \tag{30}$$

where Λ, Λ_1 are diagonal matrices and $\Lambda_1 + \Lambda_1^* \leq 0$. We introduce new variables

$$x \quad e^{\frac{i}{\varepsilon}\Lambda t}z.$$

Then (29) becomes

$$z' \quad \Lambda_1(t)z + \tilde{B}z, \tag{31}$$

where

$$\tilde{B} \quad e^{-\frac{i}{\varepsilon}\Lambda t}B\left(e^{\frac{i}{\varepsilon}\Lambda t}z,t\right)e^{\frac{i}{\varepsilon}\Lambda t}.$$

We split

$$\tilde{B} \quad B_1 + B_2,$$

where B_1 is a polynomial in z without exponentials, and all terms of B_2 contain exponentials. B_2 produces a $\mathcal{O}(\varepsilon|x_0|^2 t)$-change of z and, therefore, we neglect it. Thus, we can simplify (31) to

$$z' \quad \Lambda_1 z + B_1(z,t)z.$$

If $B_1 \ / \ 0$, then we can in general not expect that $|z| \leq K|x_0|$ holds in time intervals $T \gg |x_0|^{-1}$. We proved that the solution of (29) is of the form

$$x(t) \quad e^{\frac{i}{\varepsilon}\Lambda t} z(t), \tag{32}$$

where $z(t)$ is varying slowly. We introduce (32) into (30). We can also split

$$D(x,t)x \quad D_1 + D_\varepsilon, \tag{33}$$

where D_1 does not contain any exponentials and all terms of D_ε contain exponentials. Observe that D_1 is quadratic in z, i.e. $D_1 \quad \mathscr{O}(|x_0|^2)$. We can further deduce that $\frac{d}{dt}D_1(x(t),t)$ is independent of ε, while $\frac{d}{dt}D_\varepsilon \sim \mathscr{O}(\varepsilon^{-1})$. We have the following important conclusion:

- If $D_1 \ / \ 0$, then in general
$$|v - v_S| \quad \mathscr{O}(|x_0|^2 t).$$
- If $D_1 \quad 0$, then
$$|v - v_S| \quad \mathscr{O}(\varepsilon|x_0|),$$
and (28) holds.

We should look at the above result together with what we obtained in Sect. 2.1, in particular, the estimate (19) for the case when resonance occurs in the equation for x, and (20) for the case without resonance in the system.

2.4 Slow variables and slow observables

Consider the following ODE system

$$\begin{aligned} x' &\quad -\varepsilon^{-1}y + x, \quad x(0) \quad 1,\\ y' &\quad \varepsilon^{-1}x + y, \qquad y(0) \quad 0. \end{aligned} \tag{34}$$

The solution of this linear system is $(x(t),y(t)) \quad (e^t \cos\varepsilon^{-1}t, e^t \sin\varepsilon^{-1}t)$ whose trajectory forms a slowly expanding spiral: i.e. the solution rotates around the origin with a fast frequency $2\pi/\varepsilon$ and the distance to the origin grows in time by e^t. Although both $x(t)$ and $y(t)$ change on the ε time scale, the system can be decomposed into “fast and slow constituents”: a fast rotational phase and a slowly changing amplitude. Denoting $\xi \quad x^2 + y^2$, we have

$$\xi' \quad \frac{d}{dt}\xi(x(t),y(t)) \quad 2xx' + 2yy' \quad 2x^2 + 2y^2 \quad 2\xi.$$

Three important points call attention:

- ξ' is bounded independent of ε. Accordingly, we refer to the function $\xi(x,y)$ as a slow variable for (34).

- ξ' can be written into a function of the slow variable ξ only. Accordingly, we say that the equation for the slow variable ξ is closed.
- Suppose that there is another (slow) function $\zeta(x,y)$ such that $\frac{d}{dt}\zeta(x(t),y(t))$ $\zeta_x x' + \zeta_y y'$ is bounded independent of ε. Then, away from the origin, $\nabla\zeta(x,y)$ is parallel to $\nabla\xi(x,y)$. Otherwise, at every point away from the origin, $\nabla\zeta$ and $\nabla\xi$ form a local basis for the two dimensional vector space. Consequently, the velocity field $\Phi(x,y;\varepsilon): \quad (-\varepsilon^{-1}y+x, \varepsilon^{-1}x+y)$ can be written as a linear combination of $\nabla\zeta$ and $\nabla\xi$; we write $\Phi \quad a\nabla\xi + b\nabla\zeta$. From the hypotheses on the slowness of ξ and ζ, $\Phi\cdot\nabla\xi$ and $\phi\cdot\nabla\zeta$ are both bounded independent of ε, implying that the coefficients a and b are also bounded. However, this leads to Φ being bounded which contradicts with the given equation (34).

The three observations described above are essential for building the multiscale numerical methods introduced in the next section.

More generally, consider the ODE systems

$$x' \quad \varepsilon^{-1}f(x)+g(x), \qquad x(0) \quad x_0, \tag{35}$$

where $x\in\mathbb{R}^d$. We assume that for $0<\varepsilon<\varepsilon_0$, and for any x_0 in a region $\mathscr{A}\subset\mathbb{R}^d$, the unique solution of (35), denoted $x(t;\varepsilon,x_0)$, exists in $t\in\ 0,T$ and stays in some bounded region D. For brevity, we will omit the explicit dependence of the solution on ε and x_0 whenever it is clear from context.

Definition 1. *Let U be a nonempty open subset of $\mathscr{A}$. A smooth function $\xi:\mathbb{R}^d\mapsto\mathbb{R}$ is said to be slow with respect to (35) in U, if there exists a constant C such that*

$$\max_{x_0\in U,\ t\in\ 0,T}\left|\frac{d}{dt}\xi(x(t;\varepsilon,x_0))\right|\le C.$$

Otherwise, $\xi(x)$ is said to be fast. Similarly, we say that a quantity or constant is of order one if it is bounded independent of ε. It is also no problem generalizing this notion to time dependent slow variables $\xi(x,t)$.

Loosely speaking, $\xi(x)$ being slow means that, to leading order in ε, the quantity $\xi(x(t))$ is evolving on a time scale that is ε independent for all trajectories emanating from a macroscopic domain (radius does not shrink with ε).

Following Definition 1, for systems of the form

$$x' \quad f\left(\frac{t}{\varepsilon},x\right), \qquad f \text{ bounded},$$

each scalar component of the state variables x is considered a slow variable. Indeed, for those functions f that are periodic in the first argument, we know from our previous discussion that $x(t)$ stays very close to its initial value for all $0\le t\le T$. Furthermore, in the case of resonance discussed in Sect. 1.3, $x(t)$ drifts away from the initial value in an average distance that is growing linearly in time. For integrable Hamiltonian systems, the action variables are the slow variables for the systems.

As another example, in Sect. 2.3 function $D_1(y,t)$ may be considered as a slow variable for (29).

At this point, it is natural to ask how many slow variables exist for a given highly oscillatory system? The obvious answer is infinitely many, since any constant multiplication of a found slow variable yield another new one. A more reasonable question is to ask what the dimension of the set of all slow variables is. In preparation to answering this question, we need to make concrete a few more concepts. In the following sections this question is answered for some specific cases.

Definition 2. *Let* $\alpha_1, \cdots, \alpha_k : \mathscr{A} \subset \mathbb{R}^n \mapsto \mathbb{R}$ *be k smooth functions,* $k \leq n$*.* $\alpha_1(x), \ldots, \alpha_k(x)$ *are called functionally independent if the Jacobian has full rank; i.e.*

$$\text{rank}\left(\frac{\partial(\alpha_1, \cdots, \alpha_k)}{\partial x}\right) \qquad k.$$

Let $\alpha(x) \quad (\alpha_1(x), \ldots, \alpha_k(x))^T$ *be a vector containing k functionally independent components.*

When coupled with system (35), $\alpha(x)$ *is called a maximal vector of functionally independent slow variables if, for any other vector of size* ν *whose components are functionally independent, then* $k \geq \nu$*.*

Our objective is to use an appropriate set of slow variables together with some other smooth functions to provide a new coordinate system for a subset of the state space of system (35). Such a coordinate system separates the slow behavior from the fast oscillations and provides a way to approximate a large class of slow behavior of (35). See Fig. 4 for an illustration; locally near the trajectory, the space is decomposed into three special directions, $\nabla\phi$ defines the fast direction, and the two slow variables ξ_1 and ξ_2 help gauging the slow behavior of a highly oscillatory system.

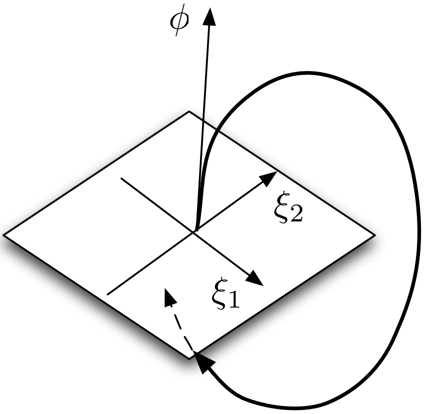


Fig. 4. Illustration of a slow chart. The slow variables ξ_1 and ξ_2 provide a local coordinate system near a trajectory.

Definition 3. *Let $\xi \quad (\xi_1(x),\dots,\xi_k(x))$ denote a maximal vector of functionally independent slow variables with respect to (35) in $\mathscr{A}$, and $\phi : \mathscr{A} \subset \mathbb{R}^d \mapsto \mathbb{R}^{d-k}$ be some smooth functions. If the Jacobian matrix $\partial(\xi,\phi)/\partial x$ is nonsingular in $\mathscr{A}$, one obtains a local coordinate systems, i.e., a chart of the states space. We refer to such a chart as a slow chart for $\mathscr{A}$ with respect to the ODE (35).*

In other words, a slow chart is a local coordinate system in which a maximal number of coordinates are slow with respect to (35).

Lemma 1. *Let (ξ,ϕ) denote a slow chart for $\mathscr{A} \subset \mathbb{R}^d$ and $\alpha(x) : \mathscr{A} \to \mathbb{R}$ a slow variable. Then, there exists a function $\tilde{\alpha}(\xi) : \mathbb{R}^k \to \mathbb{R}$ such that $\alpha(x) \quad \tilde{\alpha}(\xi(x))$.*

Proof. Otherwise, $\alpha(x)$ is a new slow variable that is functionally independent of the coordinates of ξ, in contradiction to the maximal assumption.

Another type of slow behavior can be observed through integrals of the trajectory, referred to as slow observables.

Definition 4. *A bounded functional $\beta : C^1(\mathscr{A}\times \ 0,T\ ) \cap L^1(\mathscr{A}\times \ 0,T\ ) \mapsto \mathbb{R}$ is called a (global) slow observable if*

$$\beta(t) \quad \int_0^t \tilde{\beta}(x(\tau;\varepsilon,x_0),\tau)d\tau.$$

Differentiation with respect to time shows that global observables are slow.

From the discussion in Sect. 4.2, we deduce that with an appropriate choice of kernel and η, local averages of the form

$$\beta(t) \quad \int_{\infty}^{+\infty} \frac{1}{\eta} K\left(\frac{t-\tau}{\eta}\right) \tilde{\beta}(x(\tau;\varepsilon,x_0),\tau)d\tau,$$

can also be slow. We refer to these as local observables.

We observe that along the trajectory passing through y_0, a slow variable defines a slow changing quantity ϑ. We first consider the unperturbed equation

$$\varepsilon y' \quad f(y,t),$$

and a slow variable α.

$$\frac{d}{dt}\alpha(y(t)) \quad \nabla\alpha|_{y(t)} \cdot y'(t) \quad \frac{1}{\varepsilon}\nabla\alpha|_{y(t)} \cdot f \quad : \phi_{\alpha,f}(t;y_0). \tag{36}$$

Notice that since α is a slow variable, $|\phi_{\alpha,f}(t;y_0)| \le C_1$. If this bound is valid for $0 < \varepsilon \le \varepsilon_0$, then $\nabla\alpha \cdot f \quad 0$ for a neighborhood of $y(t)$. Now consider

$$\varepsilon \tilde{y}' \quad f_\varepsilon(\tilde{y},t) + \varepsilon g(\tilde{y},t).$$

We may directly consider integrating a slow observable $\vartheta(t)$ satisfying

$$\frac{d}{dt}\vartheta \quad \phi_{\alpha,f_\varepsilon}(t;y_0), \qquad \vartheta(0) \quad \vartheta_0.$$

Notice that

$$\phi_{f_\varepsilon} \quad \frac{1}{\varepsilon}\nabla\alpha|_{\tilde{y}(t)}\cdot f(\tilde{y}(t),t)+\nabla\alpha|_{y(t)}\cdot g(\tilde{y}(t),t) \quad \nabla\alpha|_{y(t)}\cdot g(\tilde{y}(t),t).$$

So $\vartheta(t)$ is slowly varying in $\mathscr{O}(1)$ time scale.

2.5 Building slow variables by parametrizing time

The time variable may be used to create slow variables that couple different oscillators if we can locally use the coordinates of the state space to parametrize time so that time is treated as a dependent variable. Consider the equations of the form $\varepsilon y' \quad f(y,t)$ and assume that there exists a function τ, independent of ε, such that $\varepsilon\tau(y(t)) \quad t$. The function τ by its definition is not slow since

$$\frac{d}{dt}\tau(y(t)) \quad \frac{t}{\varepsilon}.$$

However, if we have $\varepsilon\tilde{\tau}(z(t)) \quad t$ for the solutions, $z(t)$, of another oscillatory problem, then the function $\theta(y,z): \quad \tau(y)-\tilde{\tau}(z)$ is a slow variable since

$$\frac{d}{dt}\theta(y(t),z(t))\equiv 0.$$

The existence of inverse functions depend on the monotonicity in time of any coordinate of the trajectories. For oscillatory problems, the monotonicity cannot hold globally. In many problems, even though the inverse function τ does not exist globally, its derivative can be defined globally. In this case, we may employ (36) to integrate a slow quantity. For example, the derivative of $\arctan(z)$ is defined on the whole real line. Similarly, on the complex plane, the derivative of the arg function is defined everywhere except at the origin. In the latter case, (36) can be regarded defining a continuous $\theta(t)$ on the Riemann sheet.

One advantage of using time as a slow variable is in defining relative phase between two planar oscillators. Consider

$$\varepsilon z_k' \quad i\lambda_k z_k, \qquad k \quad 1,2.$$

We formally define

$$\alpha(z_1,z_2): \quad \arg(z_1)-\arg(z_2)$$

and obtain the equation for the slow observable. Through this approach, we can define and integrate the slowly changing relative phase between two oscillators.

2.6 Effective closure

Let $U(t)\in\mathbb{R}^n$ and $V(t)\in\mathbb{R}^m$ be two smooth functions. Assume that, for all $0\le t\le T$, both $U(t)$ and $V(t)$ are bounded above by C_0 and that

$$\frac{dU}{dt} \quad G(U)+\varepsilon H(U,V,t),$$

for some bounded smooth function $H:\mathbb{R}^n\times\mathbb{R}^m\times\mathbb{R}^+\mapsto \; -C_1,C_1$. We say that the dynamics of U is effectively closed. This means that for $0\le t\le T$ and ε sufficiently small, one can ignore the influence of $V(t)$ and compute instead

$$\frac{d\tilde{U}}{dt} \quad G(\tilde{U}), \qquad \tilde{U}(0) \quad U(0),$$

as an approximation of $U(t)$; i.e.

$$|U(t)-\tilde{U}(t)|\le C_1\varepsilon.$$

In the spiral example (34), the equation for the single slow variable $\xi \quad x^2+y^2$ is effectively closed. The following gives an example of slow variables whose dynamics along the trajectories are not effectively closed. In the complex plane, consider the system

$$\begin{aligned} x' &\quad \frac{i}{\varepsilon}x+x^*y, \\ y' &\quad \frac{2i}{\varepsilon}y. \end{aligned} \tag{37}$$

Here the x^* denotes the complex conjugate of x. Evidently, $\xi_1: \quad xx^*$ and $\xi_2: \quad yy^*$ are two slow variables. However, the differential equation for ξ_1 along the non-equilibrium trajectories of (37) is given by $\xi_1' \quad 2\mathrm{Re}((x^*)^2y)$, which cannot be described in terms of ξ_1 and ξ_2 alone. Hence, the equation for ξ_1 is not effectively closed. In fact, it is easily verified that $\xi_3 \quad (x^*)^2y$ is also a slow variable and that $(\xi_1,\xi_2,\xi_2,\arg x)$ is a slow chart.

Later, we will see that in many oscillatory systems, the effective equations for the slow coordinates in a slow chart are effectively closed.

3 Averaging

One of the most important analytic tools for studying highly oscillatory systems are averaging methods, see e.g. [5, 20, 29] . In this section we present a few key results and discuss them using simplified examples.

3.1 Time averaging and integration by parts

We start with a simple example. Let $a(t)$ be a C^1 function whose derivative is bounded on the real line. Consider the integral

$$I(t) \quad \int_0^t \cos\left(\frac{s}{\varepsilon}\right)a(s)ds \quad \varepsilon\sin\left(\frac{t}{\varepsilon}\right)a(t)-\varepsilon\int_0^t \sin\left(\frac{s}{\varepsilon}\right)a'(s)ds.$$

Then $|I(t)| \leq C_0(1+t)\varepsilon$ for some constant C_0 coming from the maximum value of a and a' in $0,t_{\ulcorner}$. Let $a(t)$ be a λ-periodic function in C^p, $p \geq 1$; and let $a_\infty :$ $\max_{0\leq t\leq\lambda} |a(t)|$. If $a(t)$ has zero average, $\int_0^\lambda a(s)ds \quad 0$, then for all $T \geq 0$,

$$\left|\int_0^T a(t)dt\right| \leq \lambda a_\infty.$$

We define a particular anti-derivatives of $a(t)$ as follows

$$a^{0}{}_{\ulcorner}(t) \quad a(t) \text{ and } a^{k}{}_{\ulcorner}(t) \quad \int_0^t a^{k-1}{}_{\ulcorner}(s)ds + c_k, \quad k \quad 1,2,3,\ldots \tag{38}$$

where the constant c_k is chosen such that $\int_0^\lambda a^{k}{}_{\ulcorner}(s)ds \quad 0$. As a result, $a^{k}{}_{\ulcorner}(t)$ are also λ-periodic since

$$a^{k+1}{}_{\ulcorner}(t+\lambda) - a^{k+1}{}_{\ulcorner}(t) \quad \int_t^{t+\lambda} a^{k}{}_{\ulcorner}(s)ds \quad \int_0^\lambda a^{k}{}_{\ulcorner}(s)ds \quad 0, \ k \quad 1,2,3,\cdots. \tag{39}$$

Consequently, all anti-derivatives are uniformly bounded:

$$|a^{k}{}_{\ulcorner}(t)| \leq \lambda a_\infty, \qquad \forall t. \tag{40}$$

If f is differentiable, we can perform integration by parts

$$\int_0^T a\left(\frac{s}{\varepsilon}\right) f(s)ds \quad \left[\varepsilon a^{1}{}_{\ulcorner}\left(\frac{s}{\varepsilon}\right) f(s)\right]_{s\ \ 0}^{T} - \varepsilon \int_0^T a^{1}{}_{\ulcorner}\left(\frac{s}{\varepsilon}\right) f^{(1)}(s)ds,$$

where $f^{(k)}$ is the k-th derivative of f. The process can be repeated depending on the differentiability of f.

- If $f(T) \quad f(0) \quad 0$, and $f \in C^p$, then
$$\left|\int_0^T a\left(\frac{s}{\varepsilon}\right) f(s)ds\right| \leq \sup_{0\leq t\leq T} |f^{(p)}(t)| a_\infty \cdot \varepsilon^p.$$
- If f is in C^∞, we can further obtain a formal asymptotic expansion approximation for the integral
$$\int_0^T a\left(\frac{s}{\varepsilon}\right) f(s)ds \quad \sum_k \left[(-1)^{k-1}\varepsilon^k a^{k}{}_{\ulcorner}\left(\frac{s}{\varepsilon}\right) f^{(k-1)}(s)\right]_{s\ \ 0}^{T}.$$
- If $\bar{a} : \quad \int_0^\lambda a(\xi)d\xi \ / \ 0$, then
$$I_\varepsilon : \quad \int_0^T a\left(\frac{s}{\varepsilon}\right) f(s)ds \longrightarrow \bar{I} : \quad \bar{a}\left(\int_0^T f(s)ds\right) \text{ as } \varepsilon \to 0. \tag{41}$$
- Similar averaging results can be obtained for functions $a(t)$ which are not necessarily periodic but whose anti-derivatives are nonetheless bounded.

Exercise 3. Prove (41).

3.2 How does averaging in an oscillatory system appear?

Consider

$$x' \quad f\left(\frac{t}{\varepsilon},x\right), \qquad x(0) \quad x_0, \tag{42}$$

where $f(t,x)$ is Lipschitz in both t and x with constant L and is λ-periodic, $f(t+\lambda,x) \quad f(t,x)$. In addition, consider

$$y' \quad \bar{f}(y), \quad y(0) \quad y_0, \qquad \text{where} \quad \bar{f}(x) \quad \frac{1}{\lambda}\int_0^{\lambda} f(t,x)dt. \tag{43}$$

We call (43) the averaged, or effective equation derived from (42). The following calculation shows that $|x_\varepsilon(t)-y(t)| \le C_1\varepsilon$ for a long time which is independent of ε. Observe that

$$\begin{aligned} x(t)-x_0 \quad & \int_0^t f\left(\frac{\tau}{\varepsilon},x(\tau)\right)d\tau \\ & \int_{t_M}^t f\left(\frac{\tau}{\varepsilon},x(\tau)\right)d\tau + \sum_{j\ 0}^{M-1}\int_{j\varepsilon\lambda}^{(j+1)\varepsilon\lambda} f\left(\frac{\tau}{\varepsilon},x(\tau)\right)d\tau, \end{aligned}$$

where $t-t_M < \varepsilon\lambda$. In each interval $t_j \quad j\varepsilon\lambda \le t \le t_{j+1} \quad (j+1)\varepsilon\lambda$,

$$\begin{aligned} \int_{j\varepsilon\lambda}^{(j+1)\varepsilon\lambda} f\left(\frac{\tau}{\varepsilon},x(\tau)\right)d\tau \quad & \int_{j\varepsilon p}^{(j+1)\varepsilon\lambda} f\left(\frac{\tau}{\varepsilon},x(t_j)\right)+\mathscr{O}(\varepsilon)d\tau \\ & \int_{j\varepsilon\lambda}^{(j+1)\varepsilon\lambda} f\left(\frac{\tau}{\varepsilon},x(t_j)\right)d\tau+\mathscr{O}(\varepsilon^2) \\ & \varepsilon p\cdot\frac{1}{\lambda}\int_0^{\lambda} f(s,x_j)ds+\mathscr{O}(\varepsilon^2) \\ & \varepsilon\lambda\overline{f}(x_j)+\mathscr{O}(\varepsilon^2). \end{aligned}$$

Hence,

$$\begin{aligned} x(t)-x_0 \quad & \int_0^t f\left(\frac{\tau}{\varepsilon},x(\tau)\right)d\tau \\ & \int_{t_M}^t f\left(\frac{\tau}{\varepsilon},x(\tau)\right)d\tau + \sum_{j\ 0}^{M-1}\int_{j\varepsilon\lambda}^{(j+1)\varepsilon\lambda} f\left(\frac{\tau}{\varepsilon},x(\tau)\right)d\tau \\ & \int_0^t \overline{f}(x(\tau))d\tau+\mathscr{O}(\varepsilon). \end{aligned}$$

Now, since

$$y(t)-x_0 \quad \int_0^t \bar{f}(y(\tau))d\tau,$$

by Gronwall's lemma we have

$$|x(t)-y(t)| \le L\int_0^t |x(\tau)-y(\tau)|d\tau + C\varepsilon.$$

This result can be generalized as follows. Let

$$x' \quad \sum_{j \quad 1}^{M} f_j\left(\frac{t}{\varepsilon}, x\right), \qquad x(0) \quad x_0,$$

where f_j is λ_j-periodic in the time. Then, a calculation similar to the above shows that the solution of

$$y' \quad \bar{f}(y), \qquad y(0) \quad x_0,$$

with

$$\bar{f}(x) \quad \sum_j \frac{1}{\lambda_j} \int_0^{\lambda_j} f_j(\tau, x) d\tau,$$

is close to $x(t)$ on a time segment.

Theorem 4. *For $t \in \; 0, T \lceil$, $T < \infty$ and independent of ε (assume $x(t), y(t)$ exist in such interval)*

$$|x(t) - y(t)| \le C_1 \varepsilon.$$

Note that $y' \quad \bar{f}(y)$ is independent of ε. While $x_\varepsilon(t)$ is highly oscillatory, there are no ε-scale oscillations in $y(t)$. We conclude that the cost of integrating the averaged equation is independent of ε and is in general much more efficient than computing x_ε. If we just pick an arbitrary t^*, $z' \quad f(t^*, z)$, $z(0) \quad x_0$ in general we can not expect that $x(t) \quad z(t) + \mathscr{O}(\varepsilon)$.

Averaging over oscillations may appear in many different ways and should be handled with caution. The following problem presents a case in homogenization, in which harmonic averages are derived as parameters for an effective equation.

Exercise 4. In the following problem, high frequency oscillations in a_ε interact with those in $\frac{d}{dx}u_\varepsilon$ and creates low frequency behavior of $u_\varepsilon(x)$:

$$\begin{cases} \frac{d}{dx}\left(a_\varepsilon(x)\frac{d}{dx}u_\varepsilon\right) \quad f(x), & 0 < x < 1, \\ u_\varepsilon(0) \quad u_\varepsilon(1) \quad 0, \\ a_\varepsilon(x) \quad a(\frac{x}{\varepsilon}) > 0. \end{cases}$$

We derive an effective equation for $u_\varepsilon(x)$ by performing the following steps.

1. Integrate the equation with respect to x and show that

$$\begin{cases} a_\varepsilon \frac{du_\varepsilon}{dx} \quad \int_0^x f(\xi)d\xi + C, \\ u_\varepsilon(x) \quad \int_0^x (a_\varepsilon(\xi))^{-1} F(\xi) d\xi, \quad \text{where } F(\xi) \quad \int_0^\xi f(\eta)d\eta + C. \end{cases}$$

   Determine C from boundary conditions.
2. Show that

$$\lim_{\varepsilon \to 0} \int_0^x a\left(\frac{\xi}{\varepsilon}\right)^{-1} F(\xi) d\xi \quad \int_0^1 a(y)^{-1} dy \int_0^x F(\xi) d\xi, \quad F \in C \; 0, 1 \lceil .$$

3. Show that

$$u_\varepsilon \longrightarrow \bar{u} \quad A^{-1}\int_0^x \left(\int_0^\xi F(\eta)d\eta + C\right)d\xi \text{ as } \varepsilon \longrightarrow 0,$$

where

$$A \quad \frac{1}{\int_0^1 a(y)^{-1}dy}$$

We conclude that $\bar{u}(x)$ satisfies the effective equation:

$$A\frac{d^2\bar{u}}{dx^2} \quad f(x), \qquad 0<x<1, \qquad \bar{u}(0) \quad \bar{u}(1) \quad 0$$

In summary, we present the following facts about averaging, whose proof can be found, for example, in [20] and [29].

Theorem 5. *Let $x,y,x_0 \in D \subset \mathbb{R}^n$, $\varepsilon \in (0,\varepsilon_0$. Suppose*

1. *f,g, and $|\nabla f|$ are bounded by M which is independent of ε.*
2. *g is Lipschitz in a bounded domain D.*
3. *$f(t,x)$ is λ-periodic in t, λ independent of ε.*

Then, the solution of

$$x' \quad f\left(\frac{t}{\varepsilon},x\right) + \varepsilon g\left(\frac{t}{\varepsilon},x,\varepsilon\right), \qquad x(0) \quad x_0 \tag{44}$$

is close to the solution of the averaged equation

$$y' \quad \bar{f}(x), \qquad y(0) \quad x_0, \qquad \bar{f}(y) \quad \frac{1}{\lambda}\int_0^\lambda f(t,y)dt$$

on a time scale of order one. More precisely, for all $t \in 0,T$, $T<\infty$ independent of ε,

$$|x(t)-y(t)| \le C\varepsilon T e^{\varepsilon L t},$$

where $C>0$ and L denotes a Lipschitz constant for $\bar{f}$.

Moreover, equation (44) can be written in the form [20]

$$x' \quad \bar{f}(x) + \varepsilon f_1\left(\frac{t}{\varepsilon},x,\varepsilon\right), \qquad x(0) \quad x_0, \tag{45}$$

where $f_1(t,x,\varepsilon)$ is λ-periodic in t and $f_1 \to 0$ as $\varepsilon \to 0$.

3.3 Effective closure in coupled oscillators

Given the system (35), and in a neighborhood of the trajectory starting from x_0, let (ξ, ϕ) be a slow chart in which ϕ is a fast angular coordinate on the unit circle $\mathbb{S}^1$, i.e., $0 < C_1/\varepsilon < \phi' < C_2/\varepsilon$. Then, by these hypotheses, we know that

$$\begin{array}{ll} \xi' & g_I(\xi,\phi), \\ \phi' & \varepsilon^{-1} g_{II}(\xi,\phi), \end{array} \tag{46}$$

where $C_1 < g_{II}(\xi,\phi) < C_2$. Applying the averaging result (45), Equation (46) can be rewritten as

$$\begin{array}{lll} \xi' & \int g_I(\xi,\phi)d\phi + \varepsilon g_{III}(\xi,\phi) & \bar{g}_I(\xi) + \varepsilon g_{III}(\xi,\phi), \\ \phi' & \varepsilon^{-1} g_{II}(\xi,\phi). & \end{array}$$

Hence, the equation for ξ is effectively closed.

4 Computational considerations

In this section we will describe a few computational methods which gain efficiency by taking into account some of the special properties of the system discussed in previous sections. We will mostly be concerned with equations of the form

$$x_\varepsilon' \quad g\left(\frac{t}{\varepsilon}, x_\varepsilon\right), \qquad x(0) \quad x_0,$$

where $g(t,x)$ is λ-periodic, and its averaged form

$$\bar{x}' \quad \bar{g}(\bar{x}), \qquad \bar{x}(0) \quad x_0.$$

By the averaging principle, we have that

$$|x_\varepsilon(t) - x(\bar{t})| \le C\varepsilon, \qquad 0 \le t \le T.$$

4.1 Stability and efficiency

Suppose uniform time stepping is used in the computations.[4] The typical local truncation error of a p'th-order method is $\mathscr{O}((\triangle t L)^p)$, where L is a uniform bound for the $p+1$ derivative of the right hand side. Applied to the two equations above, the error varies tremendously. For x_ε, the error term is

$$E_1 \quad \mathscr{O}\left(\left[\frac{\triangle t}{\varepsilon}\right]^p\right),$$

[4] With oscillatory systems, variable time step algorithms are not as advantageous in improving efficiency as in stiff, dissipative systems.

while for $\bar{x}$, the truncation error is

$$E_2 \quad \mathcal{O}(\triangle t^p).$$

Thus, in order for the solution to be reasonably accurate, the step size $\triangle t$ has to be small compared to ε. In addition, typical explicit non-multiscale numerical schemes suffer from linear instabilities when using step sizes that are too large compared to the Lipschitz constant of the right hand size. This constraint restricts the step size of such a method to be of order ε On the other hand, the efficiency of solving an ODE to time T using step size $\triangle t$ is $\mathcal{O}(T/\triangle t)$. Hence, it is clear that it is usually much more efficient to solve the averaged equation for $\bar{x}$ than the original one for x_ε.

In the following we will develop and discuss some of the tools and ideas required to construct a multiscale algorithm that solves the averaged equation without actually deriving it. Instead, the idea of the Heterogenous Multiscale Method is to approximate the averaged equation on the fly using short time integration of the equation for x_ε.

4.2 Averaging kernels

In many numerical calculations involving oscillations with different frequencies, the right hand side may not be strictly periodic. As an example see (15). For this reason, as well as for efficiency considerations, it is convenient to average using some general purpose kernels. In the previous section, we see the need to compute the average of $f(t,x)$ over a period in t

$$\bar{f}(x) : \quad \frac{1}{\lambda}\int_0^\lambda f(\tau,x)d\tau.$$

In this section, we show that $\bar{f}(x)$ can be accurately and efficiently approximated by averaging with respect to a compactly supported kernel whose support is larger, but independent of λ. For simplicity, we shall ignore the x dependence in f.

We will use $\mathbb{K}^{p,q}$ to denote the function space for kernels discussed in this paper.

Definition 5. *Let $K^{p,q}(I)$ denote the space of normalized functions with support in I, q continuous derivatives and p vanishing moments, i.e., $K \in \mathbb{K}^{p,q}(I)$ if $K \in C_c^q(\mathbb{R})$, $supp(K) \quad I$, and*

$$\int_{\mathbb{R}} K(t)t^r dt \quad \begin{cases} 1, & r \quad 0; \\ 0, & 1 \le r \le p. \end{cases}$$

Furthermore, we will use $K_\eta(t)$ to denote a scaling of K as

$$K_\eta(t) : \quad \frac{1}{\eta}K\left(\frac{t}{\eta}\right).$$

For shorthand, we will also use $K^{p,q}$ to denote a function in $\mathbb{K}^{p,q}(\ -1,1\ )$.

Most of the numerical examples in this manuscript are obtained using the exponential kernel $K^{\exp} \in \mathbb{K}^{1,\infty}(\ -1,1\ )$:

$$K^{\exp}(t) \quad C_0\chi_{-1,1_{\ulcorner}}(t)\exp(5/(t^2-1)) \tag{47}$$

Here, $\chi_{-1,1_{\ulcorner}}$ is the characteristic function of the interval [–1, 1] and C_0 is a normalization constant such that $||K^{\,\exp}||_{L^1(\mathbb{R})} \quad 1$. A second commonly used kernel is

$$K^{\cos}(t) \quad \frac{1}{2}\,\chi_{-1,1_{\ulcorner}}(t)(1+\cos(\pi t)) \in K^{1,1}(I).$$

For convenience, we write $f(t) \quad \bar{f}+g(t)$ where

$$\bar{f} \quad \frac{1}{\lambda}\int_0^{\lambda} f(s)ds.$$

Hence, $g(t)$ is λ-periodic with zero average.

The following analysis shows that the convolution $K_\eta * f$ well approximates the average $\bar{f}$,

$$\begin{aligned}\int_{\mathbb{R}}\frac{1}{\eta}K\left(\frac{t-s}{\eta}\right)f(\frac{s}{\varepsilon})ds \quad & \int_{t-\eta}^{t+\eta}\frac{1}{\eta}K\left(\frac{t-s}{\eta}\right)\left(\bar{f}+g\left(\frac{s}{\varepsilon}\right)\right)ds\\ & \bar{f}\int_{t-\eta}^{t+\eta}\frac{1}{\eta}K\left(\frac{t-s}{\eta}\right)ds+\frac{1}{\eta}\int_{t-\eta}^{t+\eta}K\left(\frac{t-s}{\eta}\right)g\left(\frac{s}{\varepsilon}\right)ds\\ & \bar{f}+\frac{1}{\eta}\int_{t-\eta}^{t+\eta}K\left(\frac{t-s}{\eta}\right)g\left(\frac{s}{\varepsilon}\right)ds.\end{aligned}$$

Integrating by parts, we have

$$\begin{aligned}&\frac{1}{\eta}\int_{t-\eta}^{t+\eta}K\left(\frac{t-s}{\eta}\right)g\left(\frac{s}{\varepsilon}\right)ds\\ &\qquad \frac{\varepsilon}{\eta}K\left(\frac{t-s}{\eta}\right)G\left(\frac{s}{\varepsilon}\right)|_{s\ \ t-\eta}^{t+\eta}-\frac{\varepsilon}{\eta^2}\int_{t-\eta}^{t+\eta}K'\left(\frac{t-s}{\eta}\right)G\left(\frac{s}{\varepsilon}\right)ds\\ &\qquad -\frac{\varepsilon}{\eta^2}\int_{t-\eta}^{t+\eta}K'\left(\frac{t-s}{\eta}\right)G\left(\frac{s}{\varepsilon}\right)ds,\end{aligned}$$

where G is the anti-derivative of g given by (38). Hence,

$$\left|\frac{1}{\eta}\int_{t-\eta}^{t+\eta}K\left(\frac{t-s}{\eta}\right)g\left(\frac{s}{\varepsilon}\right)ds\right| \le \frac{\varepsilon}{\eta}||K'||_\infty||G||_\infty.$$

Since g is periodic and bounded, its anti-derivative is also a bounded function. For example, taking $\eta \quad \sqrt{\varepsilon}$, $\bar{f}$ is approximated to order $\sqrt{\varepsilon}$. Repeating this process q times yields

$$\left|\int K_\eta(t-s)f(s)ds-\bar{f}\right| \le C_{K,g}\left(\frac{\varepsilon}{\eta}\right)^q. \tag{48}$$

For convenience, we shall denote $K_\eta * f(t)$ by $<f(t)>_\eta$

4.3 What does a multiscale algorithm approximate?

Loosely speaking, our goal is to construct an algorithm that approximates the slow behaviour of a highly oscillatory ODE system. An important observation is that the slow behavior of a system can be a result of internal mutual cancellation of the oscillations. This, for example, is the case with resonances. Hence, it may not be clear what these slow aspects are. For this reason, we take a wide approach and require that our algorithm approximates all variables and observables which are slow with respect to the ODE.

How is this possible? We now prove that an algorithm which approximates the slow coordinates in a slow chart (ξ,ϕ) approximates all slow variables and observables.

Slow variables: Let $\alpha(x)$ denote a slow variable. From Lemma 1 we have that $\alpha(x) \quad \tilde{\alpha}(\xi(x))$ for some function $\tilde{\alpha}$. Therefore, values of $\alpha(x)$ depend only on ξ. Furthermore, it is not necessary to know $\tilde{\alpha}$, for suppose $\xi \quad \xi(x(t))$ at some time t. Then, $\alpha(x(t)) \quad \tilde{\alpha}(\xi) \quad \tilde{\alpha}(\xi(x(t))))$. In other words, all points x which correspond to the same ξ yield the same value for $\alpha(x)$.

Slow observables – global time averages: We observe that for any smooth functions $\alpha(x,t)$, we have that $\bar{\alpha}(t) \quad \int_0^t \alpha(x(s),s)ds$ is slow since $|(d/dt)\bar{\alpha}(t)| \quad |\alpha(x(t),t)|$, which is bounded independent of ε. In ODE form, we have

$$\bar{\alpha}' \quad \alpha(x,t)$$

which complies to the form required by the averaging theorem. Therefore, $\bar{\alpha}$ can be integrated as a passive variable at the macroscopic level. In other words, it can be approximated by

$$\bar{\alpha}(t) \quad \int_0^t < \frac{d}{ds}\alpha(x(s),s) >_\eta ds$$

Slow observables – local time averages: Consider time averages of the form $< \alpha(x(s)) >_\eta$. Since (ξ,ϕ) is a chart, we have that $\alpha(x(s)) \quad \tilde{\alpha}(\xi,\phi)$ for some function $\tilde{\alpha}$. However, as proven in Sect. 4.2, convolution with kernels approximates averaging with respect to the fast angular phase ϕ. Here,

$$< \alpha(x(s)) >_\eta \quad < \tilde{\alpha}(\xi,\phi) >_\eta \quad \int \tilde{\alpha}(\xi,\phi)d\phi + \text{error} \quad \bar{\alpha}(\xi(t)) + \text{error},$$

where the error is evaluated in Sect. 4.2. Hence, a consistent approximation of ξ implies a consistent approximation of $< \alpha(x(s)) >_\eta$. Moreover, in Sect. 5 we will show that the explicit form of $\tilde{\alpha}$ or $\bar{\alpha}$ are not required since all local time averages can be calculated as a by product of micro-solver steps in the algorithm.

4.4 Boosting methods

In the context of averaging, the idea of boosting is particularly simple. Consider, for example, the averaging Theorem 5 which states that, with functions $f(t,x)$, which are 1-periodic in time, the solution of

$$x' \quad f\left(\frac{t}{\varepsilon},x\right), \qquad x(0) \quad x_0 \tag{49}$$

and

$$y' \quad \bar{f}(y), \qquad y(0) \quad x_0, \qquad \bar{f}(y) \quad \int_0^1 f(s,y)ds, \tag{50}$$

are close to order ε:

$$|x(t)-y(t)| < C\varepsilon,$$

on a time scale of order one. Suppose we are interested in solving (49) with a prescribed accuracy Δ which is small, but not as small than ε, i.e., $\varepsilon \ll \Delta \ll 1$. Consider the modified equation

$$z' \quad f\left(\frac{t}{\Delta},z\right), \qquad z(0) \quad x_0. \tag{51}$$

Following the same averaging argument, $z(t)$ is close, to order Δ, to the averaged equation (50). Hence, by the triangle inequality

$$|z(t)-x(t)| \leq |z(t)-y(t)| + |y(t)-x(t)| < C(\varepsilon+\Delta) < 2C\Delta. \tag{52}$$

Solving the boosted equation (51) instead of the original one, (49) introduces an error which is of order Δ. On the other hand, the stiffness of the equation is much reduced. The discussion in Sect. 4.1 shows that the efficiency of solving the boosted equation (51) is $\mathscr{O}(\Delta^{-1})$, which can be a considerable improvement over the $\mathscr{O}(\varepsilon^{-1})$ required to solve (49). Moreover, (51) has the exact same form as (49) and preserves the same invariance. For example, if the original system is Hamiltonian, that the boosted version is also Hamiltonian.

Despite their simplicity, boosting suffers from two major drawbacks. The first is related to the nature of the asymptotic expansion used to obtain the averaged equation. Similar to expanding functions in power series, the asymptotic expansion in the averaging Theorem 5 has a "radius of convergence". This implies that the averaged equation may provide a poor approximation for (49) if ε is not small enough. In other words, the proximity between $x(t)$ and $y(t)$ "kicks in" at some value ε_0, which is usually unknown. Hence, the error estimate (52) fails if $\Delta > \varepsilon_0$.

Another drawback is that the efficiency of the method is bound to be $\mathscr{O}(\Delta^{-1})$, no matter what the order of the integrator is. This is not the case with HMM, as will be discussed in the following section. Nonetheless, boosting serves as an important benchmark to test and evaluate the efficiency of our algorithm.

5 Heterogeneous Multiscale Methods

The Heterogeneous Multiscale Method (HMM) is a general framework for systems evolving on multiple, well separated time scales. We will focus on problems with two time scales which are referred to as slow/fast, or macro/micro scales. An HMM consists of two components: a macro-solver, integrating a generally unknown averaged equation, and a micro-solver, approximating the averaged equation using short time integration of the full ODE system.

5.1 A vanilla HMM example

Consider, for example, equations of the form

$$x' \quad f\left(\frac{t}{\varepsilon},x\right), \qquad x(0) \quad x_0,$$

where $x \in \mathbb{R}^d$ and $f(t,x)$ is a smooth function which is 1-periodic in t. We rewrite the system as an homogeneous equation on $\mathbb{R}^d \times \; 0,1$,

$$\begin{array}{llll} x' & f(\phi,x), & x(0) & x_0, \\ \phi' & \varepsilon^{-1}, & s(\phi) & 0, \end{array} \tag{53}$$

where ϕ is an angular variable defined on the quotient space $\mathbb{R}/\; 0,1$. The latter space is isomorphic to the unit circle $\mathbb{S}^1$. By Definition 1, it is clear that all the coordinates in x are slow with respect to (53) while ϕ is fast. Hence, (x,ϕ) is a slow chart for (53). Furthermore, from Sect. 3, the solution for x is close (to order ε) to an effective equation which is effectively closed:

$$\bar{x}' \quad \bar{f}(\bar{x}), \qquad \bar{x}(0) \quad x_0,$$

where

$$\bar{f}(z) \quad \int_0^1 f(\tau,z)d\tau.$$

Earlier we saw that it is much favorable to solve for $\bar{x}$ rather than for x. However, the averaged forcing $\bar{f}(\cdot)$ is usually unknown. For this reason, following Sect. 4.2 we approximate $\bar{f}(\cdot)$ as $\langle f(t,x(t))\rangle_\eta$. Applying a forward Euler scheme for x with a macroscopic step size H implies taking $x_{n+1} \quad x_n + H\langle f(t,x(t))\rangle_\eta$. This is summarized in the following algorithm. Let x_n denote our approximation of (53) at time $t_n \quad nH$.

1. $n \quad 0$
2. Micro-simulation: approximate (53) numerically in a reduced time segment $t_n - \eta$, $t_n + \eta$ with step size h and x_0 replaced by x_n Denote the solution $x^n(t)$.
3. Force evaluation: calculate $F_n \quad \langle f(\frac{\cdot}{\varepsilon}), x^n(\cdot)\rangle_\eta$.
4. Macro-step (forward Euler example): take $x_{n+1} \quad x_n + HF_n$.
5. $n \quad n+1$. Repeat steps 2–4 to time T.

The efficiency of the algorithm is $\mathcal{O}(T\eta/Hh)$. It is further analyzed in Sect. 5.4.

5.2 Systematically constructing heterogeneous multiscale methods

Consider stiff ordinary differential equations (ODEs) of the form

$$\frac{du}{dt} \quad f_\varepsilon(u,t), \tag{54}$$

where $u : (0,T) \mapsto \mathbb{R}^d$, and a subset of the eigenvalues of $\partial f_\varepsilon/\partial u$ are inversely proportional to a small positive parameter ε. When ε is very small, the complexity of numerically solving such systems becomes prohibitively high. However, in many situations, one is interested only in a set of quantities U that are derived from the solution of the given stiff system (54), and typically, these quantities change slowly in time; i.e. both *U and dU/dt are bounded independent of ε*. For example, U could be the averaged kinetic energy of a particle system u.

Our objective is to construct and analyze ODE solvers that integrate the system

$$\frac{d}{dt}U = F(U,D), \tag{55}$$

where D is the data that can be computed by local solution of (54). U is called the slow (macroscopic) variable that is also some function or functional of u; i.e. $U = U(u,t)$.

If F is well-defined and has a convenient explicit mathematical expression, then there is no need to solve the stiff system (5 4) — one only needs to solve (55). In many situations, the dependence of F on U is not explicitly available. Our proposed strategy involves setting up a formal numerical discretization for (55), and evaluates F from short time history of u with properly chosen initial condition.

We will follow the framework of E and Engquist [11] in constructing efficient multiscale methods. In this framework, one assumes a macroscopic model

$$\Phi(U,D) = 0, \qquad U \in \Omega_{(M)} \tag{56}$$

which may not be explicitly given, but can be evaluated from a given microscopic model,

$$\varphi(u,d) = 0, \qquad u \in \Omega_{(m)} \tag{57}$$

where u are the microscopic variables. $D = D(u)$ and $d = d(U)$ denote the set of data or auxiliary conditions that further couple the macro- and microscopic models. Model (56) is formally discretized at a macroscopic scale, and the adopted numerical scheme dictates when the necessary information $D(u)$ should be acquired from solving (57), locally on the microscopic scale with auxiliary conditions $d(U)$. As part of $d(U)$ and $D(u)$, the macro- and microscopic variables are related by reconstruction and compression operators:

$$\mathscr{R}(U,D_R) = u, \qquad \mathscr{Q}(u) = U, \qquad \mathscr{Q}(\mathscr{R}(U,D_R)) = U,$$

where D_R are the needed data that can be evaluated from u. Errors of this type of schemes generally take the structure [10, 14]

$$\text{Error} = E_H + E_h,$$

where E_H is the error of the macroscopic model (56), and E_h contains the errors from solving (57) and the passing of information through $\mathscr{R}$ and $\mathscr{Q}$. This approach has been used in a number of applications, such as contact line problems, epitaxial growth, thermal expansions, and combustion. See the review article [12].

Figure 1 shows two typical structures of such ODE solvers. An ODE solver for U lies on the upper axis and constructs approximations of U at the grid points depicted there. The fine meshes on the lower axis depict the very short evolutions of (54) with initial values determined by $\mathscr{R}(U(t_n))$. The reconstruction operator then takes each time evolution of u and evaluates F and U. The algorithms in [10, 16, 15], and [28] are also of a similar structure. As a simple example, the forward Euler scheme applied to (55) would appear to be

$$U_{n+1} \quad U_n + H \cdot \tilde{F}(U_n), \tag{58}$$

where $\tilde{F}$ contains the passage of $\mathscr{Q}\Phi_t\mathscr{R}(U_n)$ — reconstruction $\mathscr{R}$, evolution Φ_t, and compression $\mathscr{Q}$, and H is the step size. If each evolution of the full scale system (54) is reasonably short, the overall complexity of such type of solvers would be smaller than solving the stiff system (54) for all time. The vanilla HMM presented in Sect. 5.1 uses the identity operator for both $\mathscr{Q}$ and $\mathscr{R}$.

Essential questions that need to be resolved for such a scheme include:

- With the system for u, and a choice of $U(u)$, is F well-defined by the procedure defined above? If not, how can it be properly defined?
- What are $\mathscr{R}$ and $\mathscr{Q}$?
- How long should each evolution be computed?
- What does consistency mean?
- What about stability and convergence?

For a fixed given $\varepsilon > 0$, all well known methods will converge as the step-size $H \to 0$, and there is no difference between stiff and non-stiff problems. In [11], convergence for stiff problems ($\varepsilon \ll H$) is defined by the following error:

$$E(H) \quad \max_{0 \le t_n \le T} \Big(\sup_{0<\varepsilon<\varepsilon_0(H)} |U(t_n) - U_n| \Big). \tag{59}$$

Here, $\varepsilon_0(H)$ is a positive function of H, serving as an upper bound for the range of ε, and $U(t_n)$ and U_n denote respectively the analytical solution and the discrete solution at $t_n \quad nH$. With this notion, it is clear that a sensible method has to utilize the slowly varying property of U and generate accurate approximation with a complexity that is sublinear in ε^{-1}.

The problems we are interested in can be described as follows. A full scale system (54) written in the unknown variable u is given, and the oscillations in u have frequency of order ε^{-1}. We shall also call this system the fine scale system. It is assumed that the fine scale system describes the full behavior of the problem. We want to compute the effective behavior of the given full scale system using a number of slowly changing quantities, $(U, \mathscr{V})$. These slowly changing quantities generally defined as functions or functionals of u, and their governing equations may have no explicit analytical formula. Our approach is to discretize the effective equations for $(U, \mathscr{V})$ formally and use numerical solutions of u to extract the missing information needed to evaluate the formal discretization of the governing equations.

Notation 1. Let $u(t;\alpha)$ denote the solution of the initial value problem:

$$\frac{du}{dt} \quad f_\varepsilon(u,t), \qquad u(t_*) \quad \alpha, \tag{60}$$

for some arbitrary initial condition at time t_*.

The notation $u(t)$ or u will be reserved for the solution of the same ODE, Equation (60), for $t > 0$ with the given initial condition u_0.

Definition 6. *Let $\mathscr{G}(\cdot,t)$ be a functional of u. The initial data α is said to be consistent with u under $\mathscr{G}$ if $\mathscr{G}(u(\,\cdot\,;\alpha),t) = \mathscr{G}(u,t)+\mathscr{O}(\varepsilon^r)$, for some $r>0$.*

In Kapitza's pendulum problem, the pivot of a rigid pendulum with length l is attached to a strong periodic forcing, vibrating vertically with period ε. The system has one degree of freedom, and can be described by the angle, θ, between the pendulum arm and the upward vertical direction:

$$l\theta'' = \left(g+\frac{1}{\varepsilon}\sin\left(2\pi\frac{t}{\varepsilon}\right)\right)\sin(\theta), \tag{61}$$

with initial conditions $\theta(0) = \theta_0, \theta'(0) = \omega_0$. With large ε, the only stable equilibria are $\theta_0 = n\pi$, corresponding to the pendulum pointing downward. When ε is sufficiently small and both θ_0 and ω_0 are close to 0, the pendulum will oscillate slowly back and forth, *pointing upward,* with displacement $\theta < \theta_{\max}$. The set up of the pendulum and an example solution are depicted in Fig. 5. The period of the slow oscillation is, to leading order in ε, bounded independent of the forcing period ε. On top of the slow motion around the stable $\theta = 0$ configuration, the trajectory of θ exhibits fast oscillations with amplitude and period proportional to ε.

In [32], the second order equation is written as a first order system using $u = (\theta,\omega)$, where ω is the derivative of θ. The slow variable $U = (\Theta,\Omega)$ consists of the weak limit of the angle θ and its derivative $\dot{\theta}$, and the effective force for U can be adequately approximated by the time averaging of the right hand side of (61). However, *the reconstruction operator $\mathscr{R}$ can no longer be the identity operator.* The initial values of u at t_n for each fine scale evolution should be carefully constructed such that the averages of θ matches with Θ in order to keep the correct resonance between the terms $\sin(2\pi t/\varepsilon)$ and $\sin(\theta)$. To this end, the reconstruction operator must carry a correction term when setting up ω at t_n :

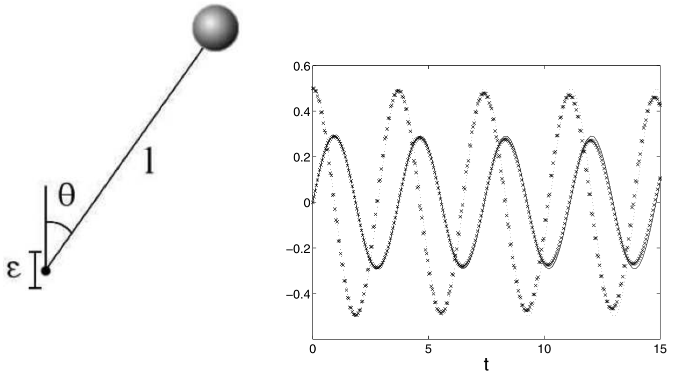
$$\omega_n^0 = \Omega_n - \frac{1}{\varepsilon}\int_{t_n-\varepsilon/2}^{t_n+\varepsilon/2}\int_{t_n}^{t} a_\varepsilon\left(\frac{s}{\varepsilon}\right)\sin(\theta_n(s))dsdt.$$

Consistency of the described multiscale solver to this type of system is thus established.

Problem 1 (Closure). Given V which consists of a set of slow variables or functionals of u, determine the set of extended variables $U = (V,W) : [0,T] \mapsto \mathbb{R}^d\times\mathbb{R}^s$, whose components are functions or functionals of u, so that there exists a function $\mathscr{F}$ independent of ε such that the solution $U(t)$ of the ODE

$$\frac{d}{dt}U = \frac{d}{dt}\begin{pmatrix}V\\W\end{pmatrix} = \mathscr{F}(U,t)+\mathscr{O}(\varepsilon), \tag{62}$$

is equivalent to its evaluation using the whole scale solution u, i.e,



(a) (b)

Fig. 5. (a) Kapitza's pendulum; (b) The slow scale solutions to equation (61); Three orders of magnitude ($\varepsilon \quad 10^{-6}$) separate the period of the slow oscillation apparent in the graphs from the fast oscillation.

$$U(t) \quad U(u,t) \quad \begin{pmatrix} V(u,t) \\ W(u,t) \end{pmatrix},$$

where $U(t)$ denotes the ODE solution, and $U(u,t)$ the functional evaluation using $u(t)$, and similarly for V and W.

One of our strategies is to look for algebraic functions α and ψ when constructing $W(u,t)$. When these functions are composed of $u(t)$ and viewed as functions of time, $\alpha_\varepsilon(t) \quad \alpha(u(t))$ and $\psi_\varepsilon(t) \quad \psi(u(t))$, they should satisfy the following conditions:

1. α and ψ are linear combinations of some simple functions of u;
2. $d\alpha_\varepsilon/dt$ is bounded independent of ε;
3. $d^\nu \langle \psi_\varepsilon \rangle /dt^\nu > \delta > 0$ for some $0 \le \nu$ and for some δ independent of ε;
4. $\alpha_\varepsilon(t)$ converges pointwise to a smooth function $\bar{\alpha}(t)$, and ψ_ε weakly to a continuous function Ψ.

Here, $\langle \psi_\varepsilon \rangle$ denotes a moving average with respect to a kernel, as described in Sect.4.2. These approaches are motivated by the analysis of resonance, the averaging methods, see e.g. [25, 5, 1, 3], and our previous work on Kapitza's pendulum and a few other model problems. Another interesting point of view makes use of the idea of Young measures [4].

In practice, we do not have $u(t)$, since we do not solve the stiff equation for a long time interval independent of ε. However, the solution U to the closure problem defines an equivalence class for the initial conditions for u. As long as an initial data is selected such that it is consistent to $u(t)$ with respect to $U(t)$, $U(t)$ is properly evolved. Instead, our strategy is to compute the solution $u(\cdot;a)$ for a duration that vanishes with ε, starting from a specified time and using some initial values a. Once $U(t)$ is approximated, we can approximate dU/dt numerically without explicitly evaluating $\mathscr{F}$. Naturally, this initial value α should be consistent with u with respect to the functional $U(u,t)$:

Problem 2 (Reintialization/reconstruction). Given a functional $U(u,t)$ and its value U_n at t_n that specifies a set of constraints.

$$\text{Find } a_n \in \mathbb{R}^d \text{ such that } U(u(\cdot;a_n),t_n) \quad U_n + \mathscr{O}(\varepsilon^p) \text{ for some } p > 0.$$

Note that unlike the common constraints of conserved integrals in the computations of Hamiltonian systems, we consider constraints, such as those specified by the components of U, that can be slowly varying in time.

In summary, our multiscale method is outlined as follow: Assuming $U(t_j)$ is known at $t \quad t_j$,

1. Find a_j that solves $U(u(\cdot;a_j),t_j) \simeq U(t_j)$ (Reinitialization);
2. Solve the given stiff equation and obtain $u(t;a_j)$ for $t_j \leq t \leq t_j + \eta_\varepsilon$ (Microscale solution);
3. Evaluate $\mathscr{F}(t_j) \quad dU/dt$ at t_j using $u(t;a_j)$;
4. Use $\mathscr{F}(t_j)$ and $U(t_j)$ to get U at $t_j + \Delta t$. (Macroscale solution)

Note that Δt should be independent of ε and η_ε vanish with ε.

In the following examples the slow behavior is approximated using functions only (the slow chart) and not functionals.

5.3 Example: an expanding spiral

Consider the system (34) describing the expanding spiral

$$\begin{array}{ll} x' \quad -\varepsilon^{-1}y + x, & x(0) \quad 1, \\ y' \quad \varepsilon^{-1}x + y, & y(0) \quad 0. \end{array} \tag{63}$$

Previously, in Sect. 2.4, it was shown that $(\xi,\phi) \quad (x^2+y^2, \tan^{-1}(y/x))$ is a slow chart for (63). The time evolution of the only slow variable ξ takes the form

$$\xi' \quad \langle \xi' \rangle_\eta + \mathscr{O}(\varepsilon) \quad \langle 2xx' + 2yy' \rangle_\eta . \tag{64}$$

This motivates the following multiscale algorithm for approximating $\xi'(t)$. For simplicity, we apply a macroscopic forward Euler solver with step size H. Any consistent and stable integrator can be used as micro-solver. We denote $t_n \quad nH$ and by x_n, y_n and ξ_n our approximation for $x(t_n)$, $y(t_n)$ and $\xi(t_n)$, respectively. Note that x_n and y_n do not have to be close to $x(t_n)$ and $y(t_n)$. We only require that the slow variable is approximated. The algorithm is depicted in Fig. 6.

1. Initial conditions: $(x(0),y(0)) \quad (x_0,y_0)$, $n \quad 0$.
2. Micro-simulation: Solve (63) in $t_n - \eta/2, t_n + \eta/2$ with initial conditions
$$(x(t_n),y(t_n)) \quad (x_n,y_n).$$

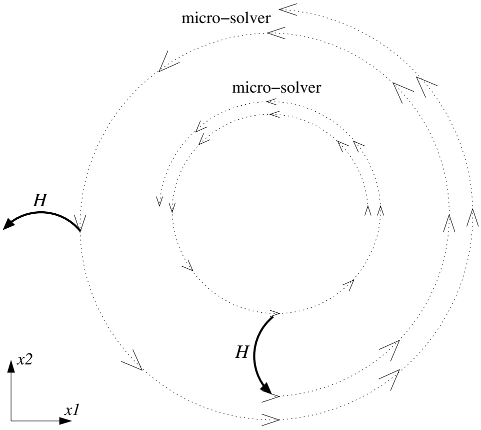


Fig. 6. Two macroscopic steps for the HMM algorithm in the expanding spiral example (63).

3. Force estimation: Approximate ξ' by $\Delta\xi_n \quad \langle 2xx' + 2yy' \rangle_\eta$. The step involves convoluting $2xx' + 2yy'$ with an averaging kernel as discussed in Sect. 4.2.
4. Macro-step (forward Euler): $\xi_{n+1} \quad \xi_n + H\Delta\xi_n$.
5. Reconstruction (second order accurate): $(x_{n+1}, y_{n+1}) \quad (x_n, y_n) + HF_n$, where F_n is the least squares solution of the linear system

$$F_n \cdot \nabla\xi(x_n, y_n) \quad \Delta\xi_n$$

6. $n \quad n+1$. Repeat steps 2–5 to time T.

5.4 HMM using slow charts

Suppose an ODE system of the form (35) admits a slow chart (ξ, ϕ), where ξ $(\xi^1, \dots, \xi^k) \in \mathbb{R}^k$ are slow and $\phi \in \mathbb{S}^1$ is fast. In the next section we will see that many highly oscillatory systems indeed admit a slow chart of that form. Then, the algorithm suggested in the previous section can be easily generalized as follows. As before, for simplicity we concentrate on the forward Euler case. Higher order methods are considered in [1]. Approximated quantities at the n'th macroscopic time step are denoted by a subscript n.

1. Initial conditions: $x(0) \quad x_0$, $n \quad 0$.
2. Micro-simulation: Solve (35) in $t_n - \eta/2, t_n + \eta/2$ with initial conditions $x(t_n) \quad x_n$
3. Force estimation: Approximate ξ' by $\Delta\xi_n \quad \langle \nabla\xi \cdot x' \rangle_\eta$ using convolution with an averaging kernel.
4. Macro-step (forward Euler): $\xi_{n+1} \quad \xi_n + H\Delta\xi_n$.

5. Reconstruction (second order accurate): $x_{n+1} \quad x_n + HF_n$, where F_n is the least squares solution of the linear system
$$F_n \cdot \nabla\xi(x_n) \quad \Delta\xi_n. \tag{65}$$
6. $n \quad n+1$. Repeat steps 2–5 to time T.

Complexity

In this section we analyze the accuracy of the suggested method outlined above. Each step of the approximations preformed in our algorithm introduces a numerical error. In order to optimize performance, the different sources of errors are balanced to a fixed prescribed accuracy Δ. We show how the different parameters: ε, η, h and H scale with Δ in order to have a global accuracy of order Δ. Note that the maximal possible accuracy is $\Delta \quad \varepsilon$, since this is the error introduced by simulating the averaged equation rather than the original one. We also study the Δ dependence of the complexity of the algorithm.

We begin with estimating the error in our evaluation of the averaged force $\Delta\xi_n$. There are several sources of errors:

- Global error in each micro-simulation. Using an m'th order method with step size h the global error is $\eta h^m/\varepsilon^{m+1}$.
- Quadrature error in $K'_\eta * \xi$: Using a quadrature formula of degree r the error is $\eta h^m/\varepsilon(m+1)$. However, due to the regularity of the kernel used $K \in C^q$, the integrand is smooth and periodic. Hence, the coefficients of its Fourier decomposition decay very fast. As a result, it is advantageous to use the trapezoidal rule, which is exact for $e^{2\pi ikx}$, $k \in \mathbb{N}$. This implies that the quadrature error is typically very small and can be neglected.
- Approximating $\Delta\xi_n$ by $\langle\nabla\xi \cdot x'\rangle_\eta$: Using a kernel $K \in \mathbb{K}^{p,q}$ the error is the larger between η^p and $(\varepsilon/\eta)^q/\eta$. Note that we are losing one order of η compared to (48) since $\Delta\xi_n$ is found through integration by parts (cf. Sect. 5.4). The above two bounds to the averaging error are equal if $\eta^{p+q+1} \quad \varepsilon^q$, where, for large η, the term η^p dominates, while for small η the other. Since we would like to optimize our complexity, it is always preferable to work in the latter regime. Hence, we can take the averaging error to be $(\varepsilon/\eta)^q/\eta$.

Balancing all terms yields the optimal scaling of the simulation parameters with Δ.

The global accuracy of integrating the original full ODE to time $T \quad \mathcal{O}(1)$ using a macro-solver of order s with step size H is, assuming errors are accumulative,
$$E \le D \max\left\{H^s, \frac{\eta h^m}{\varepsilon^{m+1}}, \frac{\varepsilon^q}{\eta^{q+1}}\right\}, \tag{66}$$
For some $D > 0$. For short hand we drop the constant in all following expressions. Balancing the different sources of errors to a prescribed accuracy Δ yields

$$\begin{aligned} \eta &\quad \varepsilon^{\frac{q}{q+1}} \Delta^{-\frac{1}{q+1}}, \\ H &\quad \Delta^{\frac{1}{s}}, \\ h &\quad \varepsilon^{1+\frac{1}{m(q+1)}} \Delta^{\frac{s+1}{sm}+\frac{1}{m(q+1)}}. \end{aligned} \tag{67}$$

The complexity is then

$$C \quad \frac{\eta}{h}\frac{T}{H} \quad \varepsilon^{-\frac{m+1}{m(q+1)}} \Delta^{-\frac{1}{s}-\frac{s+1}{sm}-\frac{m+1}{m(q+1)}}. \tag{68}$$

With a smooth kernel we can consider the $q \longrightarrow \infty$ limit. In this case the complexity estimate is reduced to

$$C(q \longrightarrow \infty) \quad \Delta^{-\frac{1}{s}-\frac{s+1}{sm}}. \tag{69}$$

Figure 7 depicts the relative error of the HMM approximation compared to the analytical solution of the expanding spiral example (34). The kernel was constructed from polynomials to have exactly two continuous derivatives and a single vanishing moments, i.e., $q \quad 2$ and $p \quad 1$. Fourth order Runge–Kutta schemes were used for both the micro- and the macro-solvers. The simulation parameters are chosen to balance all errors as discussed above.

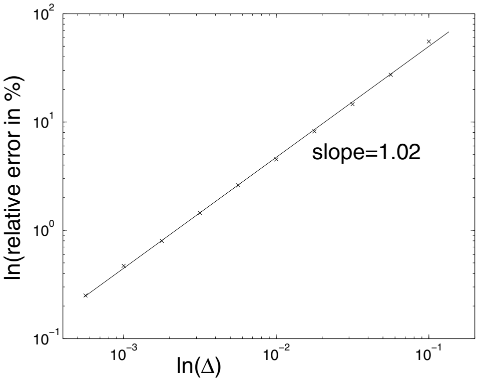


Fig. 7. A log-log plot of the relative error of the HMM approximation to a linear ODE compared to the exact solution: $E \quad \max_{t_n \in 0,T} 100 \times |\xi_{\text{HMM}}(t_n) - \xi_{\text{exact}}(t_n)|/|\xi_{\text{exact}}(t_n)|$, as a function of Δ.

From the parameter scaling (67) it is clear that the step size of the macro-solver, H, does not depend on the stiffness ε, but only on the prescribed accuracy Δ. Our algorithm is therefore multiscale is the sense that it converges uniformly for all $\varepsilon < \varepsilon_0$ [11]. More precisely, denote the sample times of the macro-solver by $t_0 \quad 0,\ldots,t_N$ T and the corresponding numerical approximations for x by $x_0,\ldots,x_N$. The exact solution is denoted $x(t)$. We have that, for any variable $\alpha(x)$ that is slow with respect to $x(t)$

$$\lim_{H\to 0} \sup_{k \quad 0,\ldots,N} \sup_{\varepsilon<\varepsilon_0} |\alpha(x(t_k)) - \alpha(x_k)| \longrightarrow 0. \tag{70}$$

Note that the order of the limits is important.

5.5 Almost linear oscillators in resonance

Consider an ODE system of the form

$$x' = \frac{1}{\varepsilon}Ax + f(x), \tag{71}$$

where, $x \in \mathbb{R}^d$ and A is an $d \times d$ real diagonalizable matrix with purely imaginary eigenvalues $\pm i\omega_1, \ldots, \pm i\omega_r$, $2r \le d$. In addition, we assume that all oscillatory modes are in resonance. This implies that the ratio of every pair of frequencies is rational, i.e., for all $i, j = 1 \ldots r$, there exist integers m_{ij} and n_{ij}, such that $m_{ij}\omega_i = n_{ij}\omega_j$.

Theorem 6. *There exists a slow chart (ξ, ϕ) in $\mathbb{R}^d \setminus \{\xi_i \neq 0, \forall i\}$ for (71) such that all the coordinates of ξ are polynomial in x and $\phi \in \mathbb{S}^1$.*

The theorem is proven in [1]. As an example, consider (71) with

$$A = \begin{pmatrix} & 1 & & \\ -1 & & & \\ & & & 2 \\ & & -2 & \end{pmatrix}.$$

Changing variables so that A is diagonalized yields the complex system

$$z = \frac{1}{\varepsilon}\begin{pmatrix} i & & & \\ & -i & & \\ & & 2i & \\ & & & -2i \end{pmatrix} z + f(x)$$

where $z = (z_1, z_1^*, z_2, z_2^*)^T$ and z^* denotes the complex conjugate of z. It is easily verified that the following are slow variables

$$\begin{aligned} \xi_1 &= z_1 z_1^*, \\ \xi_2 &= z_2 z_2^*, \\ \xi_3 &= z_1^2 z_2^*. \end{aligned}$$

Transforming back to the original coordinates $x = (x_1, v_1, x_2, v_2)$ the slow variables become the real polynomials

$$\begin{aligned} \xi_1 &= x_1^2 + v_1^2, \\ \xi_2 &= x_2^2 + v_2^2, \\ \xi_3 &= x_1 x_2^2 + 2v_1 x_2 v_2 - x_1 v_2^2. \end{aligned}$$

The first two variables, ξ_1 and ξ_2 correspond to the square of the amplitude of the two harmonic oscillators described by (x_1, v_1) and (x_2, v_2), respectively. The third variable, ξ_3, corresponds to the relative propagation of phase in the two oscillators. It is slow because, to leading order in ε, the phase of (x_2, v_2) increases twice as fast as that of (x_1, v_1).

Exercise 5. Verify that $\nabla\xi_1$, $\nabla\xi_2$, and $\nabla\xi_3$ are not linearly dependent in any region in $\mathbb{R}^4 \setminus Q$ where Q is the zeros of ξ_1, ξ_2, and ξ_3.

Fully nonlinear oscillators

Dealing with non-linear oscillators is more complicated than linear ones. However, the slow behavior of weakly coupled systems of oscillators such as Van der Pol, relaxation and Volterra-Lotka can still be described using some generalization of amplitude and relative phase. This is beyond the scope of these notes. We refer to [2] for further reading.

6 Computational exercises

Computer exercise 1. Let $u = (x,y,z)$ and

$$f_\varepsilon(x,y,z) = \begin{pmatrix} a & \frac{1}{\varepsilon} & 0 \\ -\frac{1}{\varepsilon} & b & 0 \\ 0 & 0 & -\frac{1}{10} \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ x^2 + cy^2 \end{pmatrix}. \tag{72}$$

The equation for u is

$$u' = f_\varepsilon(u), \qquad u(0) = (1,0,1).$$

Take $\varepsilon = 10^{-4}$, $a = b = 0$ and $c = 1$. Find approximations for $z(t)$ in $0 < t \leq 1$ using the following schemes and compare with the analytical solution. Plot the trajectories of your approximations of $x(t)$ and $y(t)$ on the xy-plane, and the graph $z(t)$ as a function of time. Explain what you observe in each case.

(a) Forward Euler using $\triangle t = \varepsilon/50$.
(b) Backward Euler for x and y and Forward Euler for z, using $\triangle t = 0.1$.
(c) Verlet method or Midpoint rule for x and y, and Forward Euler for z, using $\triangle t = \varepsilon/50$.
(d) Solve this problem by the HMM–FE–fe method (see below), with $\mathcal{Q} = \mathcal{R} = I$ (see Sect. 5.2). $h = \varepsilon/50$, $H = 0.1$, and $hM = 2 \cdot 10^{-3}$.
(e) Derive linear stability criteria on H for HMM–FE–fe, assuming that $h = c_0\varepsilon$.
(f) Let $a = b = 1$ in the system defined above. Solve it by the same HMM–FE–fe scheme with the same parameters as in (d). Does this scheme correctly approximate the behavior of z in the time interval $0 < t \leq 1$? Explain.

HMM–FE–fe scheme for $u' = f_\varepsilon(u)$.

- Macroscale with Forward Euler (FE)

$$U^{n+1} = U^n + HF^n, \qquad U^0 = \mathcal{Q}(u_0)$$

- Microscale with Forward Euler (fe)

$$u_{k+1}^n = u_k^n + hf_\varepsilon(u_k^n), \qquad k = 0, \pm 1, \cdots, \pm M,$$
$$u_0^n = \mathscr{R}(U^n).$$

- Averaging

$$F^n := \frac{1}{2M}\sum_{k=-M}^{M} K^{\cos}\left(\frac{k}{2M}\right) f_\varepsilon(u_k^n),$$
$$K^{\cos}(t) = \frac{1}{2}\chi_{[-1,1]}(t)\,(1+\cos(\pi t)),$$
$$\chi_{[-1,1]}(x) = \begin{cases} 1, & -1 \le x \le 1, \\ 0, & \text{otherwise.} \end{cases}$$

Computer exercise 2. Following the previous problem, define the slow variable

$$\xi(x,y) = x^2 + y^2 \text{ and } \xi(t) := x^2(t) + y^2(t),$$

where $x(t)$ and $y(t)$ are defined in (72).

(a) Show that $d\xi/dt$ can be approximated by averaging:

$$\left|\frac{d\xi}{dt}(t_n) - \int_{-\infty}^{\infty} -\frac{d}{dt}K^{\cos}\left(\frac{t_n - t}{2Mh}\right)\left(x^2(t) + y^2(t)\right)dt\right| \le C\eta^p.$$

Find p.

(b) Modify your previous HMM–FE–fe code to HMM–FE–rk4 (see below) as follows and determine if the dynamics of z is accurately approximated by this new scheme. Plot your approximations as in the previous problem. Explain your findings.

(c) Do the same thing as in the previous problem, but with $c = 0$. Does your multiscale algorithm work? Why?

Constrained HMM–FE–rk4 scheme for $u' = f_\varepsilon(u)$.

- Macroscale with Forward Euler

$$U^{n+1} = U^n + HF^n, \qquad U^0 = \mathscr{Q}(u_0).$$

- Microscale with Runge–Kutta 4 (rk4)

$$u_{k+1}^n = \text{rk4}(u_k^n, h), \qquad k = 0, \pm 1, \cdots, \pm M,$$
$$u_0^n = \mathscr{R}(U^n).$$

Here rk4 is an explicit Runge–Kutta 4 routine using step size h.

$$\text{rk4}(y,h) = y + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4),$$

$$k_1 = hf_\varepsilon(y), \quad k_2 = hf_\varepsilon(y + \frac{1}{2}k_1), \quad k_3 = hf_\varepsilon(y + \frac{1}{2}k_2), \quad k_4 = hf_\varepsilon(y + k_3).$$

- Averaging

$$dz^n : \quad \frac{1}{2M}\sum_{k\quad -M}^{M} K^{\cos}\left(\frac{k}{2M}\right)\left(x_k^n\cdot x_k^n + c y_k^n\cdot y_k^n - \frac{z_k^n}{10}\right).$$

$$d\xi^n : \quad \frac{1}{2M}\sum_{k\quad -M}^{M} G\left(\frac{k}{2M}\right)(x_k^n\cdot x_k^n + y_k^n\cdot y_k^n),$$

  where $G(\frac{k}{2M}) : \quad \frac{-1}{2Mh}\frac{d}{dt}K^{\cos}(\frac{t}{2Mh})$.
- Evaluate effective force
  Find a unit vector dX^n such that

$$d\xi^n \quad \nabla_{x,y}\xi\,|_{x_k^n,y_k^n}\cdot dX^n.$$

$$F^n : \quad \begin{pmatrix} dX^n \\ dz^n \end{pmatrix}.$$

Computer exercise 3. Consider the inverted pendulum equation:

$$l\theta'' \quad \left(g+\frac{1}{\varepsilon}\sin\left(2\pi\frac{t}{\varepsilon}\right)\right)\sin(\theta). \tag{73}$$

Let $\omega \quad \theta'$, rewrite it into a system of first order equations for (θ,ω). Let $\Omega_{n+\frac{1}{2}}$ denote the averaged macroscopic angular momentum at time $(n+\frac{1}{2})H$ and Θ_n be the averaged macroscopic angle. Compute the inverted pendulum solutions by using the parameters $\varepsilon \quad 10^{-6}$, $(\Theta_0,\Omega_0) \quad (0.0,-0.4)$, $g \quad 0.1$, $l \quad 0.05$. Experiment with $\eta \quad 10\varepsilon$ and 30ε.
This problem is analyzed in [32].

HMM for the inverted pendulum problem.

- Macroscale with Verlet
  Given $U^n \quad (\Theta^n,\Omega^n)$, for $n \quad 0,1,2,ldots$

$$\begin{aligned} \Omega^{n+\frac{1}{2}} &\quad \Omega^n + \frac{H}{2}\cdot\tilde{F}^n, \\ \Theta^{n+1} &\quad \Theta^n + H\cdot\Omega^{n+\frac{1}{2}}, \\ \Omega^{n+1} &\quad \Omega^{n+\frac{1}{2}} + \frac{H}{2}\cdot\tilde{F}^{n+1}, \end{aligned}$$

  Here, $\tilde{F}\ \theta^n,\omega^n$ denotes the averaged force using the solutions whose values at $t_n \quad nH$ are (θ^n,ω^n).
- Microscale evolution
  Solve $l\theta'' \quad (g+\frac{1}{\varepsilon}\sin(2\pi\frac{t}{\varepsilon}))\sin(\theta)$ for $t_{n-\eta}\le t\le t_{n+\eta}$ with the "reconstructed initial"

$$\theta(t_n) \quad \Theta^n,$$

$$\omega(t_n) \quad \theta'(t_n) \quad \mathscr{R}(\Theta^n,\Omega^n): \quad \Omega^n - \sin(\Theta^n)\frac{\cos\left(2\pi\frac{t_n}{\varepsilon}\right)}{2\pi l}.$$

$$\left(\omega(t_n) \approx \Omega^n - \left\langle \int_{t_n}^{t} a_\varepsilon\left(\frac{s}{\varepsilon}\right)\sin(\theta(s))ds\right\rangle\right)$$

- Average
  Using the solution computed in the microscale evolutions around t_n. Evaluate

$$\tilde{F}^n \quad \left\langle (g+\frac{1}{\varepsilon}\sin\left(2\pi\frac{t}{\varepsilon}\right))\sin(\theta)\right\rangle_\eta \quad K_\eta * f(t_n),$$

where

$$K_\eta(t) \quad \frac{422.11}{\eta}\exp\left[5\left(\frac{4t^2}{\eta^2}-1\right)^{-1}\right],$$

and

$$f(t) \quad \left(g+\frac{1}{\varepsilon}\sin\left(2\pi\frac{t}{\varepsilon}\right)\right)\sin(\theta(t)).$$

Use the Trapezoidal rule to approximate the above convolutions.

Computer exercise 4. The following is a well studied system taken from the theory of stellar orbits in a galaxy

$$\begin{array}{cc} r_1''+a^2r_1 & \varepsilon r_2^2 \\ r_2''+b^2r_2 & 2\varepsilon r_1 r_2 \end{array}.$$

Rewrite the above equation into the standard form (35).

(a) To see how resonances occur, change into polar coordinates and take $a \quad \pm 2b$.
(b) Let $a \quad 2$ and $b \quad 1$. Find a maximal slow chart.
(c) Apply the HMM algorithm described in Sect. 5.4 to approximate the slow behaviour of the system.

References

1. G. Ariel, B. Engquist, and Y.-H. Tsai. A multiscale method for highly oscillatory ordinary differential equations with resonance. *Math. Comp.*, 2008. To appear.
2. G. Ariel, B. Engquist, and Y.-H. Tsai. Numerical multiscale methods for coupled oscillators. *Multiscale Model. Simul.*, 2008. Accepted.
3. V.I. Arnol'd. *Mathematical methods of classical mechanics*. New York, Springer-Verlag, 2 edition, 1989.
4. Z Artstein, J. Linshiz, and E.S. Titi. Young measure approach to computing slowly advancing fast oscillations. *Multiscale Model. Simul.*, 6(4):1085–1097, 2007.
5. N. N. Bogoliubov and Yu. A. Mitropolski. *Asymptotic Methods in the Theory of Nonlinear Oscillations*. Gordon and Breach, New York, 1961.
6. R. Car and M. Parrinello. Unified approach for molecular dynamics and density functional theory. *Phys. Rev. Lett.*, 55(22):2471–2475, 1985.

7. A.J. Chorin. A numerical method for solving incompressible viscous flow problems. *J. Comp. Phys.*, 2:12–26, 1967.
8. G. Dahlquist. A special stability problem for linear multistep methods. *Nordisk Tidskr. Informations-Behandling*, 3:27–43, 1963.
9. G. Dahlquist, L. Edsberg, G. Skollermo, and G. Söderlind. Are the numerical methods and software satisfactory for chemical kinetics? In *Numerical Integration of Differential Equations and Large Linear Systems*, volume 968 of *Lecture Notes in Math.*, pages 149–164. Springer-Verlag, 1982.
10. W. E. Analysis of the heterogeneous multiscale method for ordinary differential equations. *Commun. Math. Sci.*, 1(3):423–436, 2003.
11. W. E and B. Engquist. The heterogeneous multiscale methods. *Commun. Math. Sci.*, 1(1):87–132, 2003.
12. W. E, B. Engquist, X. Li, W. Ren, and E. Vanden-Eijnden. Heterogeneous multiscale methods: A review. *Commun. Comput. Phys.*, 2(3):367–450, 2007.
13. B. Engquist and O. Runborg. Computational high frequency wave propagation. *Acta Numerica*, 12:181–266, 2003.
14. B. Engquist and Y.-H. Tsai. Heterogeneous multiscale methods for stiff ordinary differential equations. *Math. Comp.*, 74(252):1707–1742, 2003.
15. C. W. Gear and I. G. Kevrekidis. Projective methods for stiff differential equations: problems with gaps in their eigenvalue spectrum. *SIAM J. Sci. Comput.*, 24(4):1091–1106 (electronic), 2003.
16. C. W. Gear and I. G. Kevrekidis. Constraint-defined manifolds: a legacy code approach to low-dimensional computation. *J. Sci. Comput.*, 25(1-2):17–28, 2005.
17. D. Ter Haar, editor. *Collected Papers of P.L. Kapitza*, volume II. Pergamon Press, 1965.
18. E. Hairer, C. Lubich, and G. Wanner. *Geometric numerical integration*, volume 31 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 2002. Structure-preserving algorithms for ordinary differential equations.
19. E. Hairer and G. Wanner. *Solving ordinary differential equations. II*, volume 14 of *Springer Series in Computational Mathematics*. Springer-Verlag, 1996.
20. J. Hale. *Ordinary differential equations*. New York, Wiley-Interscience, 1969.
21. J. B. Keller. Geometrical theory of diffraction. *J. Opt. Soc. Amer.*, 52:116–130, 1962.
22. J. Kevorkian and J. D. Cole. *Perturbation Methods in Applied Mathematics*, volume 34 of *Applied Mathematical Sciences*. Springer-Verlag, New York, Heidelberg, Berlin, 1980.
23. J. Kevorkian and J. D. Cole. *Multiple Scale and Singular Perturbation Methods*, volume 114 of *Applied Mathematical Sciences*. Springer-Verlag, New York, Berlin, Heidelberg, 1996.
24. H.-O. Kreiss. Problems with different time scales. *Acta Numerica*, 1:101–139, 1992.
25. H.-O. Kreiss and J. Lorenz. Manifolds of slow solutions for highly oscillatory problems. *Indiana University Mathematics Journal*, 42(4):1169–1191, 1993.
26. B. Leimkuhler and S. Reich. *Simulating Hamiltonian dynamics*, volume 14 of *Cambridge Monographs on Applied and Computational Mathematics*. Cambridge University Press, 2004.
27. L. R. Petzold. An efficient numerical method for highly oscillatory ordinary differential equations. *SIAM J. Numer. Anal.*, 18(3):455–479, 2003.
28. L. R. Petzold, O. J. Laurent, and Y. Jeng. Numerical solution of highly oscillatory ordinary differential equations. *Acta Numerica*, 6:437–483, 1997.
29. J. A. Sanders and F. Verhulst. *Averaging Methods in Nonlinear Dynamical Systems*, volume 59 of *Applied Mathematical Sciences*. Springer-Verlag, New York, Berlin, Heidelberg, Tokyo, 1985.

30. R. E. Scheid. The accurate numerical solution of highly oscillatory ordinary differential equations. *Math. Comp.*, 41(164):487–509, 1983.
31. R. E. Scheid. Difference methods for problems with different time scales. *Math. Comp.*, 44(169):81–92, 1985.
32. R. Sharp, Y.-H. Tsai, and B. Engquist. Multiple time scale numerical methods for the inverted pendulum problem. In B. Engquist, P. Lötstedt, and O. Runborg, editors, *Multiscale methods in science and engineering*, volume 44 of *Lect. Notes Comput. Sci. Eng.*, pages 241–261. Springer, Berlin, 2005.
33. E. Vanden-Eijnden. On HMM-like integrators and projective integration methods for systems with multiple time scales. *Commun. Math. Sci.*, 5(2):495–505, 2007.

Quantum Mechanics/Classical Mechanics Modeling of Biological Systems

Håkan W. Hugosson and Hans Ågren

Department of Theoretical Chemistry, School of Biotechnology, Royal Institute of
Technology, SE-100 44 Stockholm, Sweden,
hakan@theochem.kth.se,
agren@theochem.kth.se

1 From Waves to Particles, from Static to Dynamic and from Single Molecules to Biological Systems

The objective of the project was to give an overview and hands-on insight into modern biotechnological modeling in and across several scales (and dimensions), from Quantum Mechanics to Classical Mechanics, from Molecular Mechanics to Molecular Dynamics, and from Single Molecules to Biological Systems. Here we give a starting tour of using the classical Molecular Dynamics software package AMBER [1], and the quantum mechanical Car-Parrinello package CPMD (www.cpmd.org) [2]. Introductions to some useful visualization software software specialised for biological systems was also given (Visual Molecular Dynamics) [3].

The main focus of this project is to apply multi-scale, here the so-called QM/MM molecular dynamics, methods to a biological system. In current biophysical modeling two main branches exist; classical force-field MD and static QM calculations. Though both methods are regularly and successfully applied to biological systems, many intrinsic restrictions exist, some of which are greatly reduced or solved by using multi-scale QM/MM MD. In classical MD, the by-far most common approach, the potential energy surface is parametrized through a fitting to empirical and/or theoretical data. They are therefore intrinsically restricted to situations where no significant changes of the electronic structure occur, but for example chemical reactions can only be treated adequately by a quantum mechanical description. Further restrictions include that the force field parameters are pre-defined and do not change when the system changes, either from conformational changes or from changes in the local environment. Secondly, if one wishes to model e.g. metallic centers or more uncommon organic molecules, no parameters may be available and a sometimes arduous and non-trivial parametrization is necessary.

The investigation of our physical surroundings by making quantum mechanical electronic structure calculations is a rapidly growing field of research. Lately electronic structure calculations have breached into biology, the science of life itself. This approach has proven to be very powerful and has modeled chemical reaction pathways also in complicated enzymes. The reach of QM calculations is further enhanced in first-principles molecular dynamics (FP-MD). Here a MD sampling of the conformational space is made, giving the system studied more degrees of freedom, thereby making the simulations less biased upon choices of e.g. initial conditions and reaction coordinates. Since crystal structures are inherently not the natural biological conditions/environment, and like NMR structures, are time averages over structures, this increased independence can be vital for a correct investigation. Also, in contrast to calculations which provide a map of the potential surface at the zero temperature limit, FP-MD allows inclusion of finite temperature effects. Finally, a multi-scale QM/MM method allows the accurate modeling of systems larger than 100-200 atoms. Since this is the case for virtually all biological systems, especially if one includes the solvent, this extension is vital in biophysics/biochemistry.

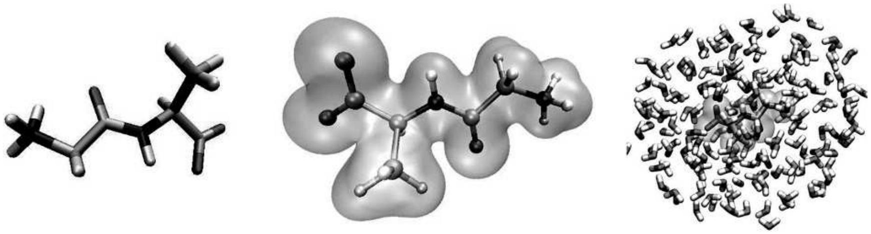


Fig. 1. A graphical illustration of how the dipeptide is modeled first in a classical force-field (MM) description (ball and springs), then in a quantum mechanical (wave function) picture and, finally, in a QM/MM fashion, with the QM dipeptide interacting with MM water [4]

2 The Exercises

The excercises were divided into four separate but joined parts, which were completed over 3 days of studies. The students were supplied with a written instruction manual, reference literature and sample input files. Finally, a project report in the form of a group presentation and a 10 page report was made and presented at the workshop by the students.

2.1 From Quantum Mechanics to Molecular Dynamics

In the first part of the project, simulations for a single small dipeptide using either quantum mechanics or classical (force field - ff) mechanics to describe the intra and inter molecular interactions, were set up . The simulations are performed at 0K and in vacuum, using either the AMBER suite of programs (classical mechanics) or the CPMD suite of programs (for the QM simulations). The dipeptide was chosen since it is small enough to allow for full QM modeling (very computer resource intensive), while also retaining some of the main characteristics of biological systems, e.g. high flexibility and some protein functional groups.

To illustrate the versatility of the AMBER suite, the module ”ANTECHAMBER” and the so-called Generalized Amber Force Field (GAFF) parameters were used. Starting from a .pdb (Protein Data Base) file of the dipeptide (alanine-glycine), the .pdb-format being the standard for reported biological structures, and using the AMBER ”LEAP” program, all the necessary input files for the AMBER minimisation and dynamics program ”Sander” (the MD driver) can be created in a near-to automated fashion. Using GAFF will generate force field parameters for most organic molecules, and even if less well parameterized than the specialized force field parameters available for proteins and nucleic acids, they are often well performing. Using Sander we then performed a simple geometry optimization of the system. From the same .pdb file of the dipeptide we also created the necessary input for a QM simulation using the CPMD software package. To illustrate the difficulty in finding the minimal energy structure, we suggested the students to also perform the geometry optimization using an annealing algorithm, i.e. a short molecular dynamics run was performed where the temperature of the system was gradually reduced to zero (the total energy of the annealed structure is lower).

The second part introduced the dimension of time (in the form of temperature and movement) into the simulations. Into the Sander input files for the geometry optimization we now added keywords to perform molecular dynamics (MD) simulations at 300K, still in vacuum. The energy optimized structure, and restart files, from Part 1 were used as starting points. Into the CPMD input file we also added the necessary keywords to perform first temperature scaling MD (at 300K) but also Nose-Hoover thermostat MD. We began with a “Quench-Annealing” run (i.e. calculating the optimized wave function for the starting geometry and then also reducing the beginning temperature of the system to zero), to relax bonds and remove excess energy from our system.

2.2 From Single Molecules to Biological Systems

In the third part a more biologically realistic system was modeled by immersing the dipeptide in a water solution. Here it was necessary, as seen from the analysis of the prior section - for reasons of available computer resources and accessible time spans, to equilibrate the surrounding water system first using classical MD. Returning to the LEAP program the dipeptide was setup with a surrounding water box. The dynamics were once again visualized using VMD, noting the evolution of geometries (as in the previous section), also the difference when performing the MD with an initial restraint on the solute molecule.

In a summation of prior experiences and skills from this project; multi-scale quantum mechanics/molecular mechanics (QM/MM) simulations were setup and executed. This allowed a reliable quantum mechanical simulation of a solute dipeptide and a more expedient handling of the (less critical) molecular water solvent. Using the equilibrated systems from the previous part we transported the resulting structures into a QM/MM addition to the CPMD software package. As an extension, possibly setting up a prototype system, an illustration of how one may partition also a bonded system into a QM and a MM region using so-called linker atoms may be given.

3 Conclusion

The objective of the project was to give an overview and hands-on insight into modern biotechnological modeling in and across several scales (and dimensions), from Quantum Mechanics to Classical Mechanics, from Molecular Mechanics to Molecular Dynamics, and from Single Molecules to Biological Systems. A very step-wise approach to teaching these methods was taken. This had the advantage that the students, coming from very diverse backgrounds in computational science, could start at the level most appropriate for each one and then progress onwards from there.

References

1. D.A. Case, D.A. Pearlman, P.A. Kollman, and coworkers, AMBER 6, University of California,(1999).
2. R. Car, M. Parrinello, Phys. Rev. Lett. **55**, 2471 (1985); J. Hutter, A. Alavi, T. Deutsch, M. Bernasconi, S. Goedecker, D. Marx, M. Tuckerman, and M. Parrinello, MPI für Festkörperforschung and IBM Zürich Research Laboratory (1995-1999); P. Carloni and U. Rothlisberger, *Theoretical Biochemistry- Processes and Properties of Biological Systems*, L. A. Eriksson (Ed), Elsevier Science, Amsterdam (2001).
3. Humphrey, W., Dalke, A. and Schulten, K., “VMD - Visual Molecular Dynamics,” J. Molec. Graphics, 1996, vol. 14, pp. 33-38.
4. Illustration taken from group project report by E. Brandt, E. Erdtman, A. Hellander, T. Murtola, K. Musa and S. Khatri, Summer School for Multi Scale Simulation and Modeling in Science 2007.

Multiple Scales in Solid State Physics

Erik Koch and Eva Pavarini

Institut für Festkörperforschung, Forschungszentrum Jülich, D-52425 Jülich, Germany,
E.Koch@fz-juelich.de, E.Pavarini@fz-juelich.de

The quest for an accurate simulations of the physical world, most vividly expressed in the vision of *Laplace's daemon* [1], is almost as old as quantitative science. Naturally, such a simulation requires the knowledge of all the relevant physical laws, i.e., a *Theory of Everything*. For the phenomena involving scales larger than an atomic nucleus and smaller than a star, or, equivalently, for processes at ordinary energies, it is known. This *Theory of almost Everything* is the combination of Newtonian gravity, Maxwell's theory of electrodynamics, Boltzmann's statistical mechanics, and quantum mechanics [2, 3]. Consequently, already shortly after the formulation of the Schrödinger equation, Dirac remarked that the theory behind atomic and solid-state physics, as well as chemistry are completely known [4]. The fundamental equation to be solved for describing the properties of atoms, molecules, or solids is the innocently looking eigenvalue problem

$$H|\Psi\rangle \quad E|\Psi\rangle \tag{1}$$

where the Hamiltonian for a set of atomic nuclei and their electrons is given by

$$H \quad -\frac{1}{2m}\sum_j \nabla_j^2 - \sum_\alpha \frac{1}{2M_\alpha}\nabla_\alpha^2 - \sum_{\alpha,j}\frac{Z_\alpha e^2}{|r_j - R_\alpha|} + \sum_{j<k}\frac{e^2}{|r_j - r_k|} + \sum_{\alpha<\beta}\frac{Z_\alpha Z_\beta e^2}{|R_\alpha - R_\beta|}\,.$$

Here Z_α and M_α are the atomic number and mass of the α^{th} nucleus, R_α is its location, e and m are the charge and mass of the electron, and r_j is the location of the j^{th} electron. This equation, augmented by gravitational potentials, and including relativistic corrections as the microscopic basis of magnetism, account for the phenomena of our everyday experience. In addition it accounts for quite counterintuitive phenomena, the most spectacular perhaps being macroscopic quantum states like superconductivity, or the entangled states that enable quantum computing.

The recipe for simulating, e.g., a superconductor is then straightforward: Simply specify the types of atoms in the system, write down the Hamiltonian for the corresponding nuclear charges and electrons and find the ground-state of the system by finding the lowest eigenvalue of the Schrödinger equation (1). While simple in principle, in practice such an approach is not feasible. To understand why, let us consider a single atom of iron. Having 26 electrons, its wavefunction $\Psi(\mathbf{r}_1,\mathbf{r}_2,\ldots,\mathbf{r}_{26})$ is a function of 78 coordinates. What does it take to store such a wave function? If we are content with recording Ψ at merely ten values for each coordinate, we would have to store 10^{78} values. Storing these on DVDs with a capacity of 10 GB, we would need more than 10^{68} DVDs. With a weight of 10 grammes per DVD, this would correspond to 10^{66} kg of DVDs – more than the mass of the visible universe. Just for comparison, the mass of the earth is a minute $5 \cdot 10^{24}$ kg. Thus there is not enough matter in the visible universe for storing even the crudest of representations of the wave function of a single iron-atom. This complexity of the wave function is the essence of the *many-body problem*.

But would it really be desirable to know the full wave function of a solid, even if it was possible? On the one hand, yes, because from the wave function we could readily calculate all expectation values. Thus we would be able to make reliable predictions of material properties. The physics would, however, be buried in the masses of data. In some sense the situation would be like that of the cartographers in Lois Borges' short story [5]:

> ON EXACTITUDE IN SCIENCE ... In that empire, the craft of cartography attained such perfection that the map of a single province covered the space of an entire city, and the map of the empire itself an entire province. In the course of time, these extensive maps were found somehow wanting, and so the college of cartographers evolved a map of the empire that was of the same scale as the empire and that coincided with it point for point. Less attentive to the study of cartography, succeeding generations came to judge a map of such magnitude cumbersome, and, not without irreverence, they abandoned it to the rigours of sun and rain. In the western deserts, tattered fragments of the map are still to be found, sheltering an occasional beast or beggar; in the whole nation, no other relic is left of the discipline of geography.
>
> From *Travels of Praiseworthy Men* (1658) by J.A. Suarez Miranda

What we really expect from a good simulation, besides reliability, is insights, e.g., into the *mechanism* of giant magneto-resistance. Only such understanding will give guidance when we try to optimize materials, e.g., when we try to find materials with ever higher magneto-resistance, which can be used in the reading heads of hard-disks of increasing capacity.

Only by discovering such mechanism, we can understand nature without always having to start from first principles. This approach is based on the concept of *emergence*, which views science as a hierarchy of structures [6]: high energy physics deals with the interactions among elementary particles, at lower energies they condense into bound states, the subject of nuclear physics. At the energy-scales of everyday-life we enter the realm of chemical bonds, studied in chemistry and condensed matter physics. At even lower energy scales we finally encounter bizarre macroscopic quantum effects, like superconductivity, studied in low-temperature physics. At each level in this hierarchy entirely new properties emerge, which are largely independent of the details on the previous level: The chemical bond, e.g., is a concept that can hardly be understood as a subtle consequence of the field equations of particle physics. Indeed, for studying the chemical bond, the arrangement of the quarks in the atomic nuclei is largely irrelevant. This is why we can base our investigations of solids on the *effective* Theory of Everything (1).

A practical approach to approximately solve equation (1) without having to deal with the full complexity of the many-body problem is *density-functional theory* [7]. Actual simulations are based on the picture of individual electrons filling atomic- or molecular-orbitals, or extended states in solids. The approach has proved extremely successful in describing chemical bonding, recognized with the 1999 Nobel Prize in chemistry.

By the nature of the approximations used, density-functional calculations are, however, largely confined to materials, where the picture of individual electrons is adequate. This model of weakly interacting quasi-particles is *Fermi-liquid theory*. There is, however, a remarkable variety of *strongly correlated* materials for which this standard model of electronic structure theory breaks down. The hallmark of these materials is that some of their electrons are neither perfectly localized, nor fully itinerant. These electrons, because of Coulomb repulsion, can no longer be considered individually. The resulting behavior presents some of the deepest intellectual challenges in physics. At the same time interest in these materials is fuelled by astounding possibilities for technological applications. Prominent examples of strongly correlated materials are the transition-metal oxides, including the high-temperature superconductors, and molecular crystals, including low-dimensional organic conductors and superconductors [8].

When dealing with strongly correlated electrons we have to face the many-body problem. As we have seen this presents enormous practical problems, so only approximate solutions, like dynamical mean-field theory [9], that maps the infinite lattice to an impurity problem which has to be solved self-consistently, are possible. While enormously reducing the cost of the simulation, such non-perturbative calculations are still limited to quite simple model-Hamiltonians [10]. It is therefore crucial to construct models that are as small as possible, while still capturing the essential chemistry of the real material.

The practical approach to studying the intriguing interplay of lattice structure, spin-, charge-, and orbital-ordering as well as superconductivity and magnetism in strongly correlated materials works in two steps. In the first step, ab-initio calculations, based on density-functional theory, are used to obtain the kinetic-energy (one-electron) part of the Hamiltonian. Next, the high-energy states are integrated out, only the low-energy partially filled (d or f) bands are retained, and a basis of first-principles Wannier functions is constructed. These Wannier functions (Fig. 1), by construction, carry the information on the lattice and the chemistry; furthermore, they are localized, so that the Coulomb repulsion is very short range in this basis. In the second step, the material-specific few-bands many-body Hamiltonians constructed by means of these Wannier functions are solved by many-body methods, such as the dynamical mean-field approximation. This two steps approach has been used very successfully, e.g. to account for the metal-insulator transition in $3d^n$ transition-metal oxides [11]. Still, even low-energy few-bands models can be solved nowadays only thanks to high performance computers [12]. The task of solving the full many-body

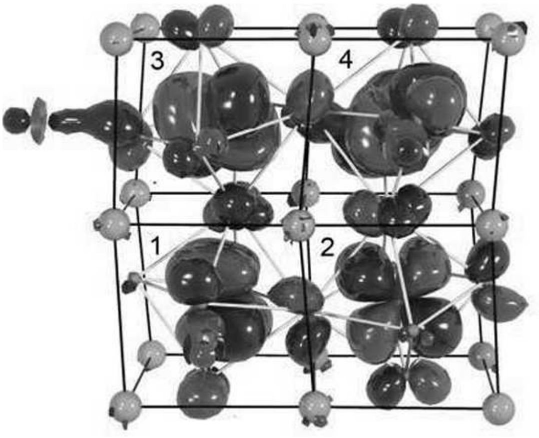


Fig. 1. Wannier functions and orbital order in monoclinic $LaVO_3$.

problem in a realistic setting will remain the main challenge in condensed matter for years to come. Bridging the high and low energy electronic degrees of freedom is not only one of the deepest problem in contemporary physics but should also provide a wealth of exciting materials for novel technologies.

References

1. Laplace, P.S.: Theorie Analytique des Probabilités, Courier, Paris (1820)
2. Ceperley, D.M.: Microscopic simulations in physics, Reviews of Modern Physics **71**, S438–S443 (1999)
3. Laughlin, R.B. and Pines, D.: The Theory of Everything, Proceedings of the National Academy of Sciences (USA) **97**, 28–31 (2000)
4. Dirac, P.M.A.: Quantum Mechanics of Many-Electron Systems, Proceedings of the Royal Society (London) **A 123**, 714–733 (1929)
5. Borges, J.L.: A Universal History of Infamy, Penguin, London, 1975.
6. Anderson, P.W.: More is Different, Science **177**, 393–396 (1972)
7. Kohn, W.: Nobel Lecture: Electronic structure of matter – wave functions and density functionals, Reviews of Modern Physics **71**, 1253–1266 (1999)
8. Osborne, I.S.: Electronic Cooperation, Science (Special Issue: Correlated Electron Systems) **288**, 461 (2000)
9. Georges, A., Kotliar, G., Krauth, W., and Rozenberg, M.J.: Dynamical mean-field theory of strongly correlated fermion systems and the limit of infinite dimensions, Reviews of Modern Physics **68**, 13–125 (1996)
10. Koch, E.: Electron correlations. In: Blügel, S., Gompper, G., Koch, E., Müller-Krumbhaar, H., Spatschek, R., and Winkler, R.G. (eds) Computational Condensed Matter Physics, FZ-Jülich (2006)
11. Pavarini, E., Biermann, S. Poteryaev, A., Lichtenstein, A.I., Georges, A., and Andersen O.K., Mott Transition and Suppression of Orbital Fluctuations in Orthorhombic $3d^1$ Perovskites, Phys. Rev. Lett. **92**, 176403 (2004); Pavarini, E., Yamasaki, A., Nuss J. and Andersen, O.K., How chemistry controls electron localization in $3d^1$ perovskites: a Wannier functions study, New J. Phys. **7** 188 (2005).
12. Dolfen, A., Pavarini, E. and Koch, E.: New Horizons for the Realistic Description of Materials with Strong Correlations, Innovatives Supercomputing in Deutschland **4**, 16 (2006)

Climate Sensitivity and Variability Examined in a Global Climate Model

Heiner Körnich and Erland Källén

Department of Meteorology, Stockholm University, SE-106 91 Stockholm, Sweden,
heiner@misu.su.se,
erland@misu.su.se

1 The climate system and its modeling

The climate system covers a large span of spatial and temporal time-scales, which range from molecular length scale and fractions of a second to global extension and thousands of years. The former are displayed in the micro physics of a cloud drop, while the latter can be found in the deep circulation of the ocean. The scale interaction gives rise to the complexity of the climate system. In the presented projects, two important aspects of the climate system are explored: its sensitivity to external forcing and its internal variability.

The climate system is regarded to consist of several components, i.e. the atmosphere, the hydrosphere (oceans), the cryosphere (ice sheets), the lithosphere (soil, rocks), and the biosphere. Owing to Earth's shape the incoming solar radiation creates an equator-to-pole temperature gradient which drives the so-called general circulation of the ocean and the atmosphere. The aim of this circulation is to diminish the induced temperature gradient by setting up a poleward heat transport. Latest estimates show that this is mainly accomplished by the atmospheric eddies [3]. The largest contribution comes from the midlatitudinal high and low pressure system which describe our daily weather. These eddies arise from the baroclinic instability of the atmosphere, which is directly related to the latitudinal temperature gradient.

Since the climate system combines multiple scales as well as different fields of science, the system is usually explored using numerical models, i.e. global climate models (GCM). Therein, a central problem is to compromise between resolved processes and computational time. As our interest lies on global spatial scales and climatological time scales of at least several decades, many processes of small and fast scales cannot be resolved in the model, but are parameterized, giving rise to a large number of parameters. The choice of the *right* value leaves some freedom to the model designer. However, it might affect the internal dynamics and the sensitivity to external forces greatly.

In order to achieve a quantitative understanding of the climate system, Earth's global energy balance [1] is reconsidered, where absorbed solar radiation and outgoing terrestrial radiation are balanced:

$$\frac{S_0}{4}(1-\alpha_p) \quad \varepsilon\sigma T_s^4. \tag{1}$$

The solar constant S_0 measures the incoming solar radiation in Wm^{-2}, which is distributed over Earth's surface, thus divided by four. The planetary albedo α_p describes how much solar radiation is reflected into space by clouds, ice or others. The terrestrial temperature seen from space relates to a mid-atmospheric level owing to the radiative properties of the atmosphere. The usage of the global surface temperature T_s in (1) requires the parameter ε which relates the surface temperature directly to the effective emissivity temperature of the climate system. The direct relation follows from the vertical heat exchange controlled by convection and latent heat release.

The model used in the presented studies is the *Planet Simulator* [2] which is developed by the Department of Meteorology at the University of Hamburg. The Planet Simulator belongs to the family of Earth system models with intermediate complexity. The atmospheric component solves the primitive equations in terms of spherical harmonics, where, in the presented results, a triangular truncation is applied at a total wave number of 21. Vertically, the model applies finite differences on 10 levels from the surface to about 16 kilometers. In order to model Earth's climate, a realistic representation of the weather systems is vital. Thus, the focus of the Planet Simulator lies on the atmosphere, while the other components of ocean, land, ice and vegetation are strongly simplified. The model package is freely available, portable, parallel, and provides a graphical user interface. The students installed the package either on their own laptop, workstations at their home institutions, or on a Linux cluster at the National Supercomputer Centre (NSC) in Linköping.

2 Project: Climate sensitivity

The first project assessed the climate sensitivity to prescribed changes of the atmospheric carbon dioxide (CO2) concentration. This problem relates to the topic of anthropogenically induced climate change. Here, the focus lay on the global-scale temperature response while simulating numerically smaller-scale feedback processes of the climate system, e.g. clouds or ice cover.

The students conducted 9 climate scenarios with different CO2 concentrations ranging from half of today's concentration to 4.5 times it. The model was integrated 50 years in order to reach a stable climate state. The last 20 years were then averaged as the response to the given forcing. The model produces for increasing CO2 a temperature increase which is comparable to the results of the latest IPCC report.

The left panel of Fig. 1 shows the effect of the CO2-content on the albedo. The largest impact stems here from the surface albedo, and thus from the ice-albedo feedback (not shown). According to (1), the reduced albedo at higher CO2-concentrations amplifies the temperature increase. The change in the atmospheric albedo due to clouds is of minor importance.

Due to the uncertainty of future economic scenarios, Paul Constantine suggested to describe the future CO2 concentration as β-distributed between 200 and 1000 ppm

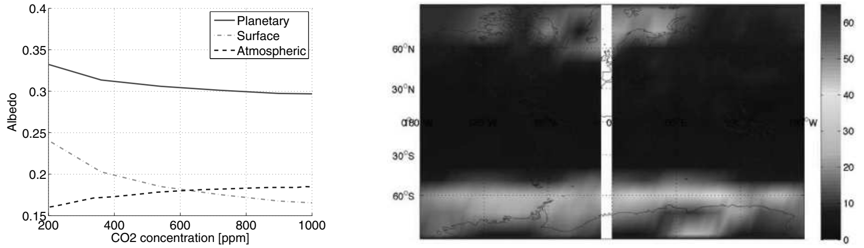


Fig. 1. Left: Planetary (solid), atmospheric (dashed) surface (dash-dotted) albedo as a function of CO2 concentration. Reproduced by courtesy of Oscar Björnham, Emanuel Rubensson, Sara Zahedi. Right: Variance of the temperature response for the β-distributed CO2-concentration as function of geographical longitude and latitude. Reproduced by courtesy of Paul Constantine.

(parts per million volume). The respective temperature response is then treated as a random variable. Here, we will examine only one case, where the maximum of the CO2-distribution lies at 600 ppm. The right panel of Fig. 1 shows the variance of the temperature response. The largest variances are seen in high latitudes, probably related to the sea-ice cover. All in all, the model simulations stress a strong impact of sea-ice on the model climate.

3 Project: Climate variability

The second project studied the issue of internal climate variability. It was examined how an external change of the incoming solar radiation affects the ability of the climate system to reduce the equator-to-pole temperature difference.

Five model simulations were prepared by the students. Besides the control experiment with the present-day value of the solar constant of 1365 Wm^{-2}, the following fractions of the current solar constant were chosen: 25%, 50%, 75%, and 125%. Each simulation was integrated for 100 years, but the last one stopped after 33 years for unknown reasons. The state of the climate system was saved every 6 hours yielding a total of 200 gigabyte raw data.

The left panel of Fig. 3 shows the normalized net radiation for all experiments. Interestingly, the relative amount of poleward transported energy decreases with decreasing solar constant. The spectrum of the wind speed at an arbitrary point in mid-latitudes (right panel of Fig. 3) yields a better understanding. Only the experiments with at least 100% of present-day's solar constant yield a local maximum at time-scales of a few days. This maximum is related to the weather systems which contribute crucially to the poleward heat transport. Due to scale interactions the weather systems also induce higher variance on lower frequencies in the climate system. For too low solar constant values, the baroclinic instability seems to be absent from the atmosphere. Therefore, the latitudinal bands nearly fulfill a local radiative balance.

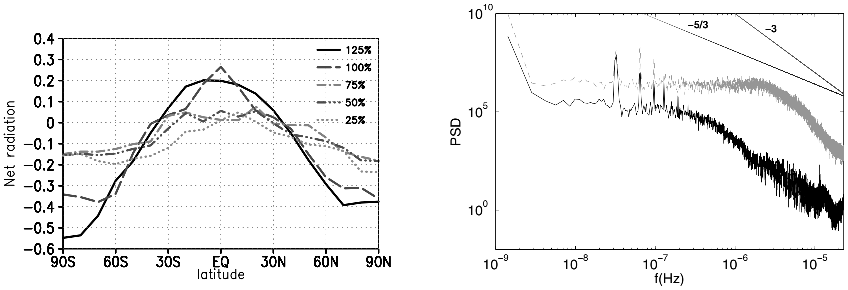


Fig. 2. Left: Difference between absorbed solar radiation and outgoing terrestrial radiation as a function of geographical latitude. Values are normalized by the global mean of the absorbed solar radiation. Right: Power spectral density of the horizontal wind speed at Austin, Texas as function of frequency in s^{-1} for the simulation with 100% (gray) and 25% (black) of present-day solar constant.

The experiments demonstrate that the baroclinic instability plays an essential role for the global temperature distribution.

Acknowledgement. For an inspiring week, the authors would like to thank the students of both groups: Aldo Benavides, Oscar Björnham, Magnar Bjørkavåg, Paul Constantine, Anna Jonsson, Robert Kakarigi, Dag Lindbo, Emanuel Rubensson, Michael Stöckli, Sofia Tapani, Khoa Tran, Anna Wramneby, and Sara Zahedi. We are most grateful to Per Lundgren at NSC for all quick support. The Planet Simulator was developed and provided by the Meteorological Institute of the University of Hamburg, working group Theoretical Meteorology.

References

1. Crafoord, C., and E. Källén: A Note on the Condition for Existence of More than One Steady-State Solution in Budyko-Sellers Type Models. J. Atmos. Sci., **35**, 1123-1125 (1978).
2. Fraedrich, K., H. Jansen, E. Kirk, U. Luksch, and F. Lunkeit: The planet simulator: Towards a user friendly model. Z. Meteorol., **14**, 299-304 (2005).
3. Trenberth, K.E., and J.M. Caron: Estimates of Meridional Atmosphere and Ocean Heat Transports. J. Climate, **14**, 3433-3443 (2001)

Coarse-scale Modeling of Flow in Gas-injection Processes for Enhanced Oil Recovery

James V. Lambers

Department of Energy Resources Engineering, Stanford University, Stanford, CA 94305-2220, USA,
lambers@stanford.edu

1 Introduction

Subsurface formations that arise in the simulation of gas-injection processes for enhanced oil recovery may exhibit geometrically complex features with complicated large-scale connectivity. They must be included in simulations of flow and transport because they can fundamentally impact simulation results. However, to reduce computational costs, simulations are generally performed on grids that are coarse compared to the given geocellular grids, so accurate upscaling is required.

In this work we are concerned with transmissibility upscaling. In the presence of full-tensor effects, multi-point flux approximations (MPFA) are desirable for the sake of accuracy, as opposed to two-point flux approximations (TPFA). However, MPFA methods add computational costs and may suffer from non-monotonicity (see [5]).

In [4], Variable Compact Multi-Point (VCMP) upscaling was introduced. This method constructs a local MPFA that accommodates full-tensor anisotropy, and guarantees a monotone pressure solution (see [3]). While it has been demonstrated that VCMP performs quite well, compared to other upscaling methods, it does not perform as well for highly channelized domains that are likely to arise in the simulation of gas injection processes.

In this paper, we consider some modifications to VCMP in order to improve its accuracy for such cases. The general design of the VCMP methods is summarized in Sect. 2. In Sect. 3, we present strategies for improving the accuracy and/or efficiency of VCMP. We discuss the results and future directions in Sect. 4.

2 Variable Compact Multi-Point (VCMP) Upscaling

In this section, we briefly review VCMP upscaling. We consider single phase, steady and incompressible flow in a heterogeneous reservoir. The governing dimensionless pressure equation, at the fine and coarse scales, is

$$\nabla \cdot (k \cdot \nabla p) \quad 0. \tag{1}$$

Here p is the pressure and k the permeability tensor, all of which are non-dimensionalized by appropriate reference values.

For simplicity, we consider a two-dimensional reservoir, and describe how VCMP upscales transmissibility to a Cartesian coarse grid. To create a MPFA, we allow the stencil to vary per cell face. Our MPFA uses a subset of the six pressure values p_j, $j \quad 1,\cdots,6$, at the six coarse cell centers nearest the face, where $j \quad 1$ and $j \quad 2$ correspond to the points that would be used in a TPFA. For each j, we let t_j denote the weight that will be assigned to point j in the flux approximation, which has the general form $f \quad -\mathbf{t}^T\mathbf{p}$, where $t \quad \begin{bmatrix} t_1 \cdots t_6 \end{bmatrix}^T$, $p \quad \begin{bmatrix} p_1 \cdots p_6 \end{bmatrix}^T$.

We solve the pressure equation on a local region of the fine grid containing the six points with two sets of generic boundary conditions. We let $\mathbf{p}^1(x,y)$ and $\mathbf{p}^2(x,y)$ be the solutions of these local problems, and p_j^i denote the value of $\mathbf{p}^i(x,y)$ at point j. For $i \quad 1,2$, we let f_i denote the coarse-scale flux across the face obtained from the local solution $\mathbf{p}^i(x,y)$. To compute the weights t_j, we solve the general optimization problem

$$\min_t \sum_{i\quad 1}^{2} \alpha_i^2 |\mathbf{t}^T\mathbf{p}^i - f_i|^2 + \sum_{j\quad 3}^{6} \beta_j^2 t_j^2, \tag{2}$$

subject to the essential linear constraints to maximize robustness. Extension to quasi-Cartesian adapted grids is discussed in [4].

3 Modifications to VCMP

In this section, we consider three strategies for improving the accuracy and efficiency of VCMP, particularly for channelized domains.

3.1 Combination with MLLG Upscaling

Local-global (LG) upscaling, introduced in [1], offers improved accuracy for permeability fields that exhibit strong global connectivity. Combined with local grid adaptation strategies, LG helps to reduce process dependency and leads to improved efficiency. This combination is known as Multi-Level Local-global upscaling, introduced in [2].

The distinction between local-global methods and local methods, such as VCMP, is that local-global methods compute global solutions of the pressure equation (1) on the coarse grid. Then, these global solutions are interpolated at points on the boundary of each extended local region. These boundary values serve as Dirichlet data for the local fine-scale solves that are used to compute upscaled transmissibilities. An iteration is used to ensure consistency between the fine and coarse scales, and because the boundary data for the local solves can account for global connectivity, greater accuracy is achieved for highly channelized domains than for local methods, including VCMP.

Like any other local method, VCMP is easily modified to use a local-global approach. However, the interpolation of the global coarse-scale solutions to the fine grid is a crucial step that must be performed carefully to ensure the same accuracy and robustness that VCMP can deliver for other domains. Whereas MLLG only uses linear interpolation, we consider both quadratic and cubic spline interpolation. Preliminary results clearly demonstrate substantial (up to 50%) improvement in the accuracy of the coarse-scale resolution of the fine-scale velocity field for channelized domains, compared to using the original, local VCMP method or MLLG with linear interpolation.

3.2 Criteria for Adaptive Mesh Refinement

MLLG includes a scheme for adaptive mesh refinement in which cells in high-flow regions are refined isotropically (for details see [2]). In addition to this criteria, a local-global version of VCMP will refine around faces for which it is unable to compute weights t_i with the proper sign based on local flow, or the computed MPFA causes the matrix for the global pressure solve to have an off-diagonal element with an incorrect sign. To determine which cells lie in high-flow regions, we compute the fluxes across each face in the coarse grid from global pressure fields, and compare them to the total flow. If the magnitude of the flux across the face is considered to be too high, then cells surrounding the face are refined.

However, this refinement may not be necessary if the high flow occurs within a channel that is nearly orthogonal to the face. We therefore use a simple channel-detection scheme in which refinement is not performed when flow across a face is nearly equal to flow across neighboring faces with the same orientation, and flow across neighboring faces with other orientations is negligible. Experimentation has demonstrated that such a channel detection scheme allows for the same accuracy to be achieved with a modest (∼8%) reduction in the number of cells.

3.3 Anisotropic Refinement

In the interest of reducing the number of cells in the coarse grid, we consider whether it is possible, in at least some cases, to refine anisotropically without sacrificing accuracy or robustness. Initial experimentation has shown that at least a small reduction in the number of cells can be achieved, without loss of accuracy, provided that

- The aspect ratio of newly created cells is limited,
- Cells in high-flow regions are still refined isotropically,
- Cells in low-flow regions, that are only refined to improve robustness, are refined anisotropically, parallel to faces that are flagged for refinement.

4 Conclusions

We have explored three avenues of improvement in the accuracy and efficiency of a proposed combination of two new methods, VCMP and MLLG, of transmissibility upscaling for coarse-scale modeling of single-phase flow in highly heterogeneous subsurface formations. In experiments with various channelized domains, all three strategies, to varying extents, yielded improved accuracy and/or efficiency in terms of reduction of the number of cells in the coarse grid or more accurate resolution of the fine-scale velocity field. In combination, these modifications should enhance performance even further.

In future work, we will consider the use of essentially non-oscillatory (ENO) interpolation schemes in computing Dirichlet boundary data for local solves, so that the monotonicity of the pressure field is not lost in the transition from the coarse scale to the fine scale. In addition, we will develop more sophisticated tests for detecting the presence and orientation of channels in order to guide adaptivity. Finally, the various adaptivity criteria include parameters that must be tuned in order to achieve optimal performance for a given permeability field; methods for automatically setting these parameters need to be developed.

In conclusion, it can be seen that substantial progress is being made toward creating a method for automatically generating coarse-scale models for gas-injection processes that are both accurate and robust for a wide variety of permeability fields and boundary conditions.

References

1. Chen, Y., Durlofsky, L.J., Gerritsen, M.G., Wen, X.H.: A coupled local-global upscaling approach for simulating flow in highly heterogeneous formations. Adv. Water Res. **26**, 1041–1060 (2003)
2. Gerritsen, M.G., Lambers, J.V.: Integration of Local-global Upscaling and Grid Adaptivity for Simulation of Subsurface Flow in Heterogeneous Formations. Comp. Geo., **12**(2), 193–208, 2008.
3. Gerritsen, M.G., Lambers, J.V., Mallison, B.T.: A Variable and Compact MPFA for Transmissibility Upscaling with Guaranteed Monotonicity. Proceedings of the 10th European Conference on the Mathematics of Oil Recovery (2006), to appear
4. Lambers, J.V., Gerritsen, M.G., Mallison, B.T.: Accurate Local Upscaling with Variable Compact Multi-point Transmissibility Calculations. Comp. Geo., Special Issue on Multiscale Methods for Flow and Transport in Heterogeneous Porous Media, to appear (2007)
5. Nordbotten, J.M., Aavatsmark, I., Eigestad, G.T.: Monotonicity of Control Volume Methods. Numer. Math. **106**(2), 255–288 (2007)

Photo-Ionization Dynamics Simulation

Garrelt Mellema

Stockholm Observatory, AlbaNova University Centre, Stockholm University, SE-106 91
Stockholm, Sweden,
garrelt@astro.su.se

1 Introduction

The radiation produced by stars (and some other objects such as accreting black holes) interacts with the gaseous matter in the universe. Since most matter is hydrogen (90% by number), often found in atomic form, the amount of radiation above 13.6 eV (1 Rydberg, the ionization energy of H) is an important property of stars. Such radiation, called extreme ultra-violet (EUV) radiation, is mostly produced by massive stars (more than $\sim$10 solar masses). The first stars in the universe were most likely such very massive stars.

EUV radiation does two things to the gas.

1. It ionizes the H atoms turning the gas into a plasma of charged particles ($\mathrm{H}^+ + \mathrm{e}^-$).
2. It heats the gas using the excess photon energy above 1 Ry.

Now obviously, there are also other elements in astrophysical gases, the next most abundant being helium (10% by number), and then C, N, O, etc.. (all at the level of $\sim 10^{-4}$ or lower). Because of its relatively high concentration, helium does also absorb some amount of photons, but this we will neglect here. The other elements have too low a concentration to absorb many photons, but are important for the radiative cooling.

The phase change in the gas due to photo-ionization of H feeds back into the dynamics because it changes the pressure. The pressure of an ideal gas is given by

$$p \quad nk_{\mathrm{B}}T, \tag{1}$$

where n is the total number of particles, k_{B} is the Boltzmann constant (1.381×10^{-16} erg/K), and T is the temperature. Photo-ionization raises the total number of particles by adding electrons to the gas (so for pure H gas the number of particles is doubled for total ionization), and by raising the temperature. The increase in pressure triggers a dynamical response from the gas, and photo-ionization leads to interesting flow patterns, see e.g. [2].

The fluid dynamic equations for photo-ionization are

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{v}) \quad 0 \tag{2}$$

$$\frac{\partial \rho \mathbf{v}}{\partial t} + \boldsymbol{\nabla} \cdot (\rho \mathbf{v} \otimes \mathbf{v}) \quad -\boldsymbol{\nabla} p \tag{3}$$

$$\frac{\partial e}{\partial t} + \nabla \cdot ((e+p)\mathbf{v}) \quad \mathscr{H} - \mathscr{C} \tag{4}$$

where ρ is the mass density, $\mathbf{v}$ is the velocity, $e \quad \frac{1}{2}\rho v^2 + p/(\gamma - 1)$ is the total energy (kinetic plus internal, γ is the adiabatic index, 5/3 in this problem), and $\mathscr{H}$ and $\mathscr{C}$ are the heating and cooling rate. Photo-ionization effects come in through $\mathscr{H}$ and p.

The number density of ionized hydrogen in the gas $n\,H^+$ is given by

$$\frac{\mathrm{d}n\,H^+}{\mathrm{d}t} \quad n\,H^0\,\Gamma - n\,H^+\,n_e \alpha(T) + n\,H^0\,n_e C(T) \tag{5}$$

where Γ is the photo-ionization rate (number of ionizing photons per second), n_e is the electron (number) density, $\alpha(T)$ is the temperature dependent recombination rate, and $C(T)$ is the temperature dependent collisional ionization rate.

The photo-ionization rate follows from the radiative transfer equation

$$\frac{\partial I_\nu}{c \partial t} + \boldsymbol{\nabla} \cdot (\boldsymbol{\Omega} I_\nu) \quad -\kappa_\nu I_\nu + j_\nu \tag{6}$$

where I_ν is the intensity of radiation in a solid angle $\boldsymbol{\Omega}$ and κ_ν and j_ν are the absorption and emission rates. We will simplify this equation by disregarding j_ν (no ionizing photons are produced outside of the star), and assuming that the star has a constant luminosity, and the size of the ionized domain is small enough that the finite speed of light does not play a role ($\frac{\partial I_\nu}{c \partial t}$=0). Then along one ray the photo-ionization rate can be written as

$$\Gamma \quad \frac{1}{4\pi r^2} \int_{\nu_0}^{\infty} \frac{L_\nu}{h\nu} a_\nu e^{-\tau_\nu(r)} \mathrm{d}\nu\,, \tag{7}$$

where r is the distance along the ray, L_ν is the stellar luminosity and τ_ν is the so-called optical depth characterizing the amount of absorption between the star and the current position

$$\tau_\nu(r) \quad \int_0^r a_\nu n\,H^0\,\mathrm{d}r\,, \tag{8}$$

where a_ν is the photo-ionization cross section of hydrogen.

Similarly the photo-ionization heating rate becomes

$$\mathscr{H} \quad \frac{n\,H^0}{4\pi r^2} \int_{\nu_0}^{\infty} h(\nu - \nu_0) \frac{L_\nu}{h\nu} a_\nu e^{-\tau_\nu(r)} \mathrm{d}\nu \tag{9}$$

The cooling rate $\mathscr{C}$ is generally a complicated function of temperature, depending on the concentrations of various atoms and ions in the gas. We will consider it given (in the form of a table). More in depth descriptions of the photo-ionization physics can be found in for example [4, 1].

2 Multi-scale Problem

Let us consider the evolution of the ionized hydrogen density. This is a stiff ODE, and physically n[H$^+$] has to lie between 0 and n[H] (so numerical under- and overshoots can produce unphysical solutions). For constant n_e (for the sake of argument), this ODE has the solution

$$n\,H^{+}_{\ulcorner}(t) \quad n\,H^{+}_{\ulcorner\,\mathrm{eq}} + (n\,H^{+}_{\ulcorner}(0) - n\,H^{+}_{\ulcorner\,\mathrm{eq}})\exp(-t/t_{\mathrm{i}}) \tag{10}$$

with

$$n\,H^{+}_{\ulcorner\,\mathrm{eq}} \quad n\,H_{\ulcorner}\frac{\Gamma+n_eC(T)}{\Gamma+n_eC(T)+n_e\alpha(T)} \tag{11}$$

$$t_{\mathrm{i}} \quad \frac{1}{\Gamma+n_eC(T)+n_e\alpha(T)} \tag{12}$$

i.e., it converges to an equilibrium solution $n\,H^{+}_{\ulcorner\,\mathrm{eq}}$ in a typical time t_{i}. This time is completely unrelated to the intrinsic time scale for the flow equations, $t_{\mathrm{hydro}} \quad \Delta r/(v+v_{\mathrm{s}})$, where $v_{\mathrm{s}}^2 \quad \gamma p/\rho$ is the sound speed. This is the first *multi-scale* problem for photo-ionization hydrodynamics. Since we are interested in the dynamics, we want to evolve our system on a time scale t_{hydro}. If $t_{\mathrm{i}} \gg t_{\mathrm{hydro}}$ this is not a problem, but for the more typical case of $t_{\mathrm{i}} \ll t_{\mathrm{hydro}}$ we have to take special measures.

This type of multi-scale problem is quite general when dealing with reactive flows, i.e. it is similar to that when chemical reactions are happening in the flow. This is typically solved with an iterative/implicit method. However, in the case of photo-ionization, the problem is more complicated. The reason is that Γ at position r depends on the optical depth τ, which in turn depends on the density of neutral hydrogen between the source and position r. Equation (8) can be rewritten as

$$\tau_\nu(r) \quad a_\nu \int_0^r n\,H^{0}_{\ulcorner}\,\mathrm{d}r \quad a_\nu N\,H^{0}_{\ulcorner}(r) \tag{13}$$

where $N\,H^{0}_{\ulcorner}(r)$ is known as the column density of neutral hydrogen between the source and point r. If during t_{hydro}, $N\,H^{0}_{\ulcorner}(r)$ changes considerably, the rate of photons arriving at position r is not constant, so Γ will be changing. This is why the photo-ionization case is more complicated than the chemical reaction case: it is a non-local effect.

There is a second multi-scale effect, this time having to do with length scales. We are solving our problem on a discretized grid with cell size Δr. This means that a cell with central position r has an optical depth

$$\Delta\tau_\nu(r) \quad a_\nu \int_{r-\frac{1}{2}\Delta r}^{r+\frac{1}{2}\Delta r} n\,H^{0}_{\ulcorner}(r)\mathrm{d}r \quad a_\nu n\,H^{0}_{\ulcorner}(r)\Delta r \tag{14}$$

If this $\Delta\tau$ is large, Γ will differ considerably between the point where the ray enters the cell, and where it leaves the cell. Using one value of Γ for the entire cell will give the wrong answer.

Our ultimate goal is to find solution(s) for these two multi-scale problems, the first having the do with the time scale, the second with the length scale.

3 Project

A method for dealing with multiscale problems was presented in [3]. The time scale problem was dealt with by using time-averaged values of the optical depth, and the length scale problem by taking a finite-volume type approach for cells. This method works well for calculating the ionization fraction, but has some problems getting the photo-ionization heating right. The reason is that a time-averaged optical depth implies the use of one value for the heating per photo-ionization reaction. In reality the efficiency of the heating changes, a process known as 'photon hardening' [4].

To study possible fixes for this problem, we will first try to modify the method from [3] to use a time-averaged value of $\exp(-\tau)$ instead of τ, since it is $\exp(-\tau)$ that enters (7) and (9).

Secondly we will develop an explicit, but efficient integrator for (5) that can be used instead of the time-averaged optical depth method from [3]. Efficiency is of the utmost importance since this problem has to be solved for every grid point, at every hydrodynamic time step. This integrator can use its own time steps, which combined have to cover the interval Δt_{hydro}. This approach is an implementation of the idea of micro time steps within macro time steps, as taught during the summer school.

References

1. M. A. Dopita and R. S. Sutherland. *Astrophysics of the diffuse universe*. Astrophysics of the diffuse universe, Berlin, New York: Springer, 2003. Astronomy and astrophysics library, ISBN 3540433627, 2003.
2. G. Mellema, S. J. Arthur, W. J. Henney, I. T. Iliev, and P. R. Shapiro. Dynamical H II Region Evolution in Turbulent Molecular Clouds. *Astrophysical Journal*, 647:397–403, August 2006.
3. G. Mellema, I. T. Iliev, M. A. Alvarez, and P. R. Shapiro. C^2-ray: A new method for photon-conserving transport of ionizing radiation. *New Astronomy*, 11:374–395, March 2006.
4. D. E. Osterbrock. *Astrophysics of gaseous nebulae and active galactic nuclei*. University Science Books, 1989, 422 p., 1989.

Time Scales in Molecular Reaction Dynamics

Yngve Öhrn and Erik Deumens

Quantum Theory Project, Departments of Chemistry and Physics, University of Florida,
Gainesville, FL 32611-8435, USA,
ohrn@qtp.ufl.edu,
deumens@qtp.ufl.edu

1 Introduction

Chemical reactions in bulk can be analyzed in terms of elementary steps that on the molecular level simply consist of molecular encounters. At the very basic level such events are described by the dynamics of participating electrons and atomic nuclei. It is generally accepted that the theory of such dynamics is contained in the time-dependent Schrödinger equation for the total reacting molecular system,

$$H\Psi \quad i\hbar\frac{\partial\Psi}{\partial t}, \tag{1}$$

where H is the quantum mechanical Hamiltonian, t the time parameter, and Ψ the wave function that describes the evolving state of the reacting system.

Because the electron dynamics is fast with typical cycle times of 10^{-17} seconds[1], and the nuclear rovibrational cycle times are orders of magnitude slower, *i.e.* $10^{-13} - 10^{-14}$ seconds, it is common to effectively separate the electron dynamics from that of the atomic nuclei. In practice this is achieved by first solving the electronic Schrödinger equation

$$H_{el}\Phi_k(x;X) \quad U_k(X)\Phi_k(x;X), \tag{2}$$

solutions of which are sought for fixed nuclear positions X. When electronic structure calculations are carried out for a large enough set of nuclear geometries so-called Born-Oppenheimer potential energy surfaces (PES's) $U_k(X)$ are obtained. A PES defines the average forces on the atomic nuclei and one can use classical mechanics, semiclassical methods or quantum mechanics to study the nuclear dynamics on such potential energy surfaces and when the electronic state changes during the reaction it can be depicted as as surface hopping.

For chemical reactions where the electron dynamics and its coupling to the nuclear degrees of freedom become important, such as may be the case in electron

[1] The atomic unit of time is 2.41888×10^{-17} seconds, which is the revolution time for the electron in the Bohr model as well as the quantum theory of atomic hydrogen

transfer processes, there are alternative theoretical and computational approaches. Our project in this Summer School on Multiscale Modeling and Simulations in Science employed such a methodology called Electron Nuclear Dynamics (END) theory [10, 2].

The starting point is the action

$$A \quad \int_{t_1}^{t_2} L(\psi, \psi^*)dt, \tag{3}$$

in terms of the quantum mechanical Lagrangian ($\hbar \quad 1$)

$$L \quad \langle\psi|H - i\frac{\partial}{\partial t}|\psi\rangle/\langle\psi|\psi\rangle. \tag{4}$$

The time-dependence is carried by a number of wave function parameters $q(t)$, such as average nuclear positions and momenta, and molecular orbital coefficients, etc. The principle of least action or the time-dependent variational principle $\delta A \quad 0$ yields the Euler-Lagrange equations

$$\frac{d}{dt}\frac{\partial L}{\partial \dot{q}} \quad \frac{\partial L}{\partial q}. \tag{5}$$

Should the wave function be so general that its variations can reach all parts of Hilbert space, then the Euler-Lagrange equations would become the time-dependent Schrödinger equation. However, for all problems of chemical interest the, necessarily, approximate wave function form for the molecular system will yield a set of coupled first-order differential equations in the time parameter t, which in a variational sense optimally approximates the time-dependent Schrödinger equation. The wave function parameters $q(t)$ that carry the time-dependence play the role of dynamical variables and it becomes important to choose a form of evolving state vector with parameters that are continuous and differentiable. Generalized coherent states are useful in this context [8, 2].

2 Minimal END

END theory can be viewed as a hierarchical approach to molecular processes. The various possible choices of families of molecular wave functions representing the participating electrons and atomic nuclei can be arranged in an array of increasing complexity ranging from a single determinantal description of the electrons and classical nuclei to a multi-configurational quantum representation of both electrons and nuclei [5]. The simplest level of END theory is implemented in a program package [3] that includes efficient molecular integral routines and well tested propagation algorithms to solve the system of coupled END equations.

This minimal END employs a wave function

$$|\psi(t)\rangle \quad |R(t),P(t)\rangle|z(t),R(t),P(t)\rangle, \tag{6}$$

where

$$\langle X|R(t),P(t)\rangle \quad \prod_{k} \exp -\frac{1}{2}(\frac{\mathbf{X}_k-\mathbf{R}_k}{b})^2+i\mathbf{P}_k\cdot(\mathbf{X}_k-\mathbf{R}_k)_{\lceil}, \tag{7}$$

and

$$\langle x|z(t),R(t),P(t)\rangle \quad \det\chi_i(\mathbf{x}_j), \tag{8}$$

with the spin orbitals

$$\chi_i \quad u_i+\sum_{j\quad N+1}^{K} u_j z_{ji}(t), \tag{9}$$

expanded in terms of atomic spin orbitals

$$\{u_i\}_1^K, \tag{10}$$

which in turn are expanded in a basis of traveling Gaussians,

$$(x-R_x)^l(y-R_y)^m(z-R_z)^n \exp \ -a(\mathbf{x}-\mathbf{R})^2-\frac{i}{\hbar M}\mathbf{P}\cdot(\mathbf{x}-\mathbf{R})_{\lceil}, \tag{11}$$

centered on the average nuclear positions **R** and moving with velocity $\mathbf{P}/M$.

In the narrow nuclear wave packet limit, $a \to 0$, the Lagrangian may be expressed as

$$\begin{aligned} L \quad & \sum_{i,j}\{\ P_{jl}+\frac{i}{2}(\frac{\partial \ln S}{\partial R_{jl}}-\frac{\partial \ln S}{\partial R'_{jl}})_{\lceil}\dot{R}_{jl}+\frac{i}{2}(\frac{\partial \ln S}{\partial P_{jl}}-\frac{\partial \ln S}{\partial P'_{jl}})\dot{P}_{jl}\} \\ & +\frac{i}{2}\sum_{p,h}(\frac{\partial \ln S}{\partial z_{ph}}\dot{z}_{ph}-\frac{\partial \ln S}{\partial z^*_{ph}}\dot{z}^*_{ph})-E, \end{aligned} \tag{12}$$

with $S \quad \langle z,R',P'|z,R,P\rangle$ and

$$E \quad \sum_{jl} P_{jl}^2/2M_l+\langle z,R',P'|H_{el}|z,R,P\rangle/\langle z,R',P'|z,R,P\rangle. \tag{13}$$

Here H_{el} is the electronic Hamiltonian including the nuclear-nuclear repulsion terms, P_{jl} is a Cartesian component of the momentum and M_l the mass of nucleus l. One should note that the bra depends on z^* while the ket depends on z and that the primed R and P equal their unprimed counterparts and the prime simply denotes that they belong to the bra.

The Euler-Lagrange equations

$$\frac{d}{dt}\frac{\partial L}{\partial \dot{q}} \quad \frac{\partial L}{\partial q}, \tag{14}$$

can now be formed for the dynamical variables

$$q \quad R_{jl}, P_{jl}, z_{ph}, z^*_{ph}, \tag{15}$$

and collected into a matrix equation:

$$\begin{bmatrix} i\mathbf{C} & \mathbf{0} & i\mathbf{C}_R & i\mathbf{C}_P \\ \mathbf{0} & -i\mathbf{C}^* & -i\mathbf{C}^*_R & -i\mathbf{C}^*_P \\ i\mathbf{C}^\dagger_R & -i\mathbf{C}^T_R & \mathbf{C}_{RR} & -\mathbf{I}+\mathbf{C}_{RP} \\ i\mathbf{C}^\dagger_P & -i\mathbf{C}^T_P & \mathbf{I}+\mathbf{C}_{PR} & \mathbf{C}_{PP} \end{bmatrix} \begin{bmatrix} \dot{\mathbf{z}} \\ \dot{\mathbf{z}}^* \\ \dot{\mathbf{R}} \\ \dot{\mathbf{P}} \end{bmatrix} \quad \begin{bmatrix} \partial E/\partial \mathbf{z}^* \\ \partial E/\partial \mathbf{z} \\ \partial E/\partial \mathbf{R} \\ \partial E/\partial \mathbf{P} \end{bmatrix}, \tag{16}$$

where the dynamical metric contains the elements

$$(C_{XY})_{ik;jl} \quad -2Im\frac{\partial^2 \ln S}{\partial X_{ik} \partial Y_{jl}}\Big|_{R' \quad R, P' \quad P}, \tag{17}$$

$$(C_{X_{ik}})_{ph} \quad (C_X)_{ph;ik} \quad \frac{\partial^2 \ln S}{\partial z^*_{ik} \partial X_{ik}}\Big|_{R' \quad R, P' \quad P}, \tag{18}$$

which are the nonadiabatic coupling terms, and

$$C_{ph;qg} \quad \frac{\partial^2 \ln S}{\partial z^*_{ph} \partial z_{qg}}\Big|_{R' \quad R, P' \quad P}. \tag{19}$$

In this minimal END approximation the electronic basis functions are centered on the average nuclear positions, which are dynamical variables. In the limit of classical nuclei these are conventional basis functions used in molecular electronic structure theory, and they follow the dynamically changing nuclear positions. As can be seen from the equations of motion discussed above the evolution of the nuclear positions and momenta is governed by Newton-like equations with Hellman-Feynman forces, while the electronic dynamical variables are complex molecular orbital coefficients which follow equations that look like those of the Time-Dependent-Hartree-Fock (TDHF) approximation [4]. The coupling terms in the dynamical metric are the well-known nonadiabatic terms due to the fact that the basis moves with the dynamically changing nuclear positions.

The time evolution of molecular processes in the END formalism employs a Cartesian laboratory frame of coordinates. This means that in addition to the internal dynamics overall translation and rotation of the molecular system are treated. The six extra degrees of freedom add work, but become a smaller part of the total effort as the complexity of the system grows. The advantage is that the kinetic energy terms are simple. This means that the effect of small kinetic energy terms, such as mass polarization, often neglected using internal coordinates, is included. Furthermore, the complications of having to choose different internal coordinates for product channels exhibiting different fragmentations are not present. One can treat all product channels on an equal footing in the same laboratory frame. Since the fundamental invariance laws with respect to overall translation and rotation are satisfied within END [2] it is straightforward to extract the internal dynamics at any time during the evolution.

3 Cross Sections

END trajectories for a molecular process are obtained by integrating (16) from suitable initial conditions for the reactants to a time where the products are well separated or no further change occurs in the system. In the case of a binary molecular reactive collision, minimal END, which uses classical nuclei, requires that for each trajectory the reactants are given some initial relative orientation. One of the reactant moieties is considered the target and placed stationary at the origin of the laboratory Cartesian coordinate system while the other collision partner, considered the projectile, is placed sufficiently distant so the interaction with the target is negligible. A Thouless determinant in a suitable basis is constructed for, say, the ground electronic state of the entire system. The projectile is given an impact parameter b and a momentum commensurate with the chosen collision energy E.

Each set of initial conditions leads to a particular set of product fragments and states. The final evolved state $|\psi\rangle$ may be projected against a number of possible final stationary electronic states $|f\rangle$ expressed in the same basis as that of the initial state to yield a transition probability $P_{fo}(b,E,\varphi) \quad |\langle f|\psi\rangle|^2$, which is a function of the collision energy E, the relative initial orientations, and the scattering angles (θ,φ) or impact parameter and angle (b,φ).

The classical differential cross section for a particular product channel with probability P_{fo} is

$$\frac{d\sigma_{fo}(E,\theta,\varphi)}{d\Omega} \qquad \sum_j P_{fo}(b_j,E,\varphi)\frac{b_j}{\sin\theta|d\Theta/db_j|}, \tag{20}$$

where the sum runs over all impact parameters b_j leading to the same scattering direction (θ,φ) for the fragment going to the detector. In this expression $\Theta(b)$ is the deflection function, which, for the first branch of the scattering region, satisfies $|\Theta| \quad \theta$.

For randomly oriented reactants, as is the case in gas phase reactions, trajectories for a sufficient number of initial relative orientations are used to produce an angular grid to calculate orientationally averaged cross sections [7, 1]

$$\frac{d\sigma_f(E,\theta,\varphi)}{d\Omega} \qquad \langle\frac{d\sigma_{fo}}{d\Omega}\rangle_o, \tag{21}$$

The well-known deficiencies of the classical cross section in (20) that occur for small angle scattering and at so-called rainbow angles, where $\frac{d\theta}{db_j} \quad 0$, as well as the lack of interference effects between the various trajectories in the sum, can be removed with semiclassical corrections such as the uniform Airy [6, 9] or the Schiff approximations [11].

The students in this project completed END trajectories for proton collisions with atomic and molecular hydrogen and obtained orientationally averaged cross sections. They also made movies in color of selected trajectories showing the dynamical electrons as an evolving cloud around the participating nuclei.

Acknowledgement. This work was completed with support of NSF grant 00057476.

References

1. R. Cabrera-Trujillo, J. R. Sabin, E. Deumens, and Y. Öhrn, *Calculations of Cross Sections in Electron Nuclear Dynamics*, Adv. Quantum Chem. **47** (2004), 253–274.
2. E. Deumens, A. Diz, R. Longo, and Y. Öhrn, *Time-Dependent Theoretical Treatments of The Dynamics of Electrons and Nuclei in Molecular Systems*, Rev. Mod. Phys **66** (1994), no. 3, 917–983.
3. E. Deumens, T. Helgaker, A. Diz, H. Taylor, J. Oreiro, B. Mogensen, J. A. Morales, M. Coutinho Neto, R. Cabrera-Trujillo, and D. Jacquemin, *Endyne version 5 Software for Electron Nuclear Dynamics*, Quantum Theory Project, University of Florida, Gainesville FL 32611-8435, `http://www.qtp.ufl.edu/endyne.html`, 2002.
4. E. Deumens and Y. Öhrn, *General harmonic approximation for time dependent molecular dynamics*, Int. J. Quant. Chem.: Quant. Chem. Symp. **23** (1989), 31.
5. E. Deumens and Y. Öhrn, *Complete Electron Nuclear Dynamics*, J. Phys. Chem. **105** (2001), 2660.
6. K. W. Ford and J. A. Wheeler, *Semiclassical description of scattering*, Ann. Phys. **7** (1959), 259.
7. D. Jacquemin, J. A. Morales, E. Deumens, and Y. Öhrn, *Electron nuclear dynamics of proton collisions with methane at 30 ev*, J. Chem. Phys. **107** (1997), 6146–6155.
8. J. R. Klauder and B.-S. Skagerstam, *Coherent states, applications in physics and mathematical physics*, World Scientific, Singapore, 1985.
9. J. A. Morales, A. C. Diz, E. Deumens, and Y. Öhrn, *Electron nuclear dynamics of* $H^+ +$ H_2 *collisions at* E_{lab} $30eV$, J. Chem. Phys **103** (1995), 9968–9980.
10. Y. Öhrn, E. Deumens, A. Diz, R. Longo, J. Oreiro, and H. Taylor, *Time evolution of electrons and nuclei in molecular systems*, Time-Dependent Quantum Molecular Dynamics (J. Broeckhove and L. Lathouwers, eds.), Plenum, New York, 1992, pp. 279–292.
11. L. I. Schiff, *Approximation method for high-energy potential scattering*, Phys. Rev. **103** (1956), 443.

Complex Band Structures of Spintronics Materials

Peter Zahn[1], Patrik Thunström[2], and Tomas Johnson[3]

[1] Department of Physik, Martin-Luther-Universität Halle-Wittenberg, D-06099 Halle, Germany, peter.zahn@physik.uni-halle.de
[2] Department of Physics, Theoretical Magnetism Group, Uppsala University, P O Box 530, 751 21 Uppsala, Sweden, patrik.thunstrom@fysik.uu.se
[3] Department of Mathematics, Uppsala University, P O Box 480, 751 06 Uppsala, Sweden, johnson@math.uu.se

1 Introduction

The launch of spintronics research was sparked by the discovery of the effect of Giant magnetoresistance in 1988. Exploiting the spin of electrons it led to new device principles for information processing, transmission and storage [7]. The pioneers of the field, the Frenchman Albert Fert and the German Peter Grünberg were honored with the Nobel Prize in 2007. It requires a detailed knowledge of the material properties to design new devices working on the basis of novel effects comprising the spin and charge degree of freedom. The spin dependent tunneling probability between two electrodes is determined by the properties of the states in the band gap of the insulator. These states can be described by a complex wave vector and for that they are forbidden in a bulk material. In contrast, they determine the electronic structure at interfaces and surfaces, in general in systems with broken translational invariance [2]. In experiments on various electrode materials with epitaxial MgO barriers one has observed oscillations of the transmission probability [8, 4]. First, we consider a one-dimensional model to account for the oscillations observed. Secondly, the complex bandstructures of the insulators MgO and ZnO are compared.

2 Oscillating Tunneling Probability: 1-dimensional Model

A one-dimensional model with two metallic regions connected by the central barrier region with the corresponding wave functions is considered as sketched in Fig. 1. We demand that the wavefunction of the system is continuously differentiable, which yields, after some calculations, the transmission coefficient

$$t \quad \frac{e^{-ikd}}{\cos(\kappa d) - i\frac{k^2+\kappa^2}{2k\kappa}\sin(\kappa d)} \ . \tag{1}$$

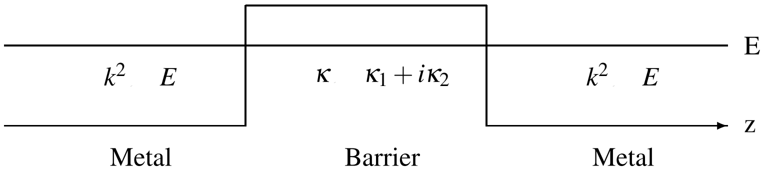


Fig. 1. 1-dimensional model system. The wave vector of the wavefunction is a real value k in the metallic electrodes, but takes complex values κ inside the barrier.

The imaginary part κ_2 of the wave vector gives rise to an exponential decay of the wavefunction while the real part κ_1 gives an oscillatory behaviour as can be seen in Fig. 2. The oscillations described by (1) are damped exponentially when d increases. The simple 1D-model can not reproduce the experimentally found oscillations.

3 Materials and Electronic Structure Calculation

ZnO is a promising candidate to obtain magnetic semiconductors by alloying with magnetic impurities [1]. Another goal is to engineer the band gap by alloying MgO and ZnO [5]. Fe/MgO/Fe tunnel junctions show a good epitaxial growth [8, 6]. This provides a high efficiency of the symmetry selection in the tunneling process. The resulting spin filter causes a large magnetoresistance [3]. The calculations of the complex bandstructure are based on density functional theory in the framework of a Korringa-Kohn-Rostoker multiple scattering Greens function formalism [9]. MgO and ZnO are considered in rocksalt structure with lattice constants of 4.052 Å and 4.216 Å, respectively. The considered cubic phase of ZnO might be stabilized in layered heterostructures like tunnel junctions. The width of the band gap is under-

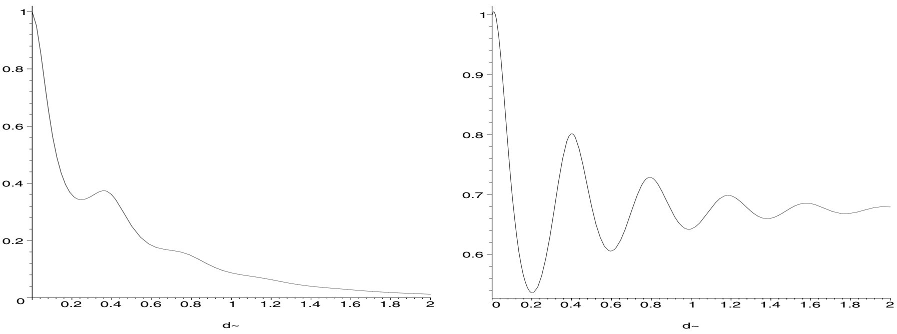


Fig. 2. Oscillatory behaviour of the transmission probability $|t|^2$ in the 1D model: $|t|^2$ (left) and $|t|^2 \exp(2d\kappa_2)$ (right) are plotted as function of barrier thickness d for k 20, $\Re(\kappa)$ κ_1 8 and $\Im(\kappa)$ κ_2 1

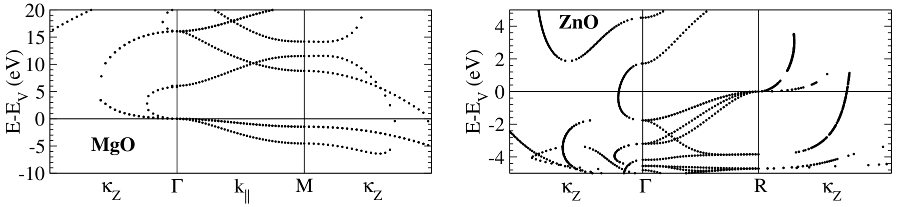


Fig. 3. Complex bandstructure of MgO and ZnO. Tetragonal symmetry is adapted to the (001) interface geometry. The bands for real wave vectors k are shown in the central panel. The points M and R correspond to the X and L point in fcc symmetry, respectively. The corresponding bands with imaginary wavevector κ_z are shown in the left and right panel.

estimated in the local density approximation with respect to the experiment, but the topology of the band structure is correctly reproduced.

4 Complex Band Structures

The bandstructures are analysed with respect to the wave vector along the (001) direction. Inspecting the band structure along a line in k-space connecting the minimum and maximum of the energy gap, a striking difference of MgO and ZnO is obvious from Fig. 3. MgO shows a direct band gap, whereas in ZnO the band gap is indirect. The imaginary part of the states available in the energy gap are responsible for the decay of the tunneling current with the barrier thickness as shown in Fig. 2, left panel. As can be seen from the left and right panels in Fig. 3, the smallest imaginary wave vectors occur in different parts of the Brillouin zone depending on the energy of the tunneling electrons. The imaginary part in dependence on the wave vector along the (100) direction and the energy in the band gap are given in Fig. 4. As expected it

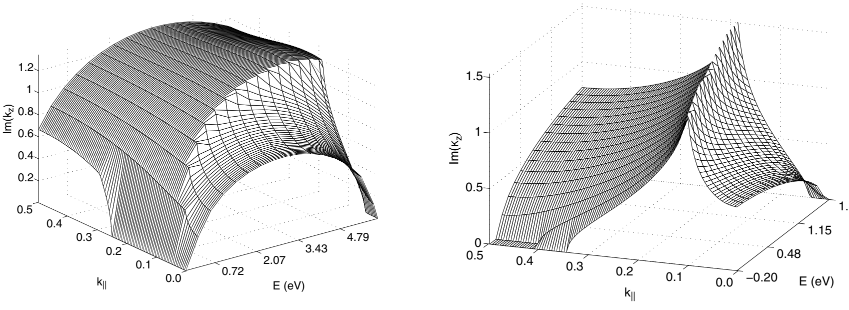


Fig. 4. Complex bandstructure of MgO (left) and ZnO (right). The smallest imaginary part κ_1 is shown in dependence on $k_{\|}$ along the $\bar{\Gamma}-\bar{X}$ line and the energy E relative to the valence band edge E_V.

is found that in MgO for energies in the bandgap the Δ_1 symmetry state dominates the transport. However, in ZnO Δ_1 states dominate only close to the band gap maximum. At lower energies the contribution around the $\bar{X}$ point grows and outweighs the Δ_1 contribution around $\bar{\Gamma}$, to completely dominate the transport near the bandgap minimum.

In summary, we have shown that the complex band structure of insulators dominates the behavior of the tunneling current. For ZnO in rocksalt structure it is found, that the position of the states carrying the tunneling current change their position in the surface Brillouin zone depending on energy inside the band gap.

References

1. K. Ando. Seeking room-temperature ferromagnetic semiconductors. *Science*, 312:1883, 2006.
2. V. Heine. On the general theory of surface states and scattering of electrons in solids. *Proc. Phys. Soc.*, 81:300, 1963.
3. Christian Heiliger, Martin Gradhand, Peter Zahn, and Ingrid Mertig. Tunneling magnetoresistance on the subnanometer scale. *Phys. Rev. Lett.*, 98:acc. June 22, 2007.
4. Rie Matsumoto, Akio Fukushima, Taro Nagahama, Yoshishige Suzuki, Koji Ando, and Shinji Yuasa. Oscillation of giant tunneling magnetoresistance with respect to tunneling barrier thickness in fully epitaxial Fe/MgO/Fe magnetic tunnel junctions. *Appl. Phys. Lett.*, 75:252506, 2007.
5. J Narayan, A K Sharma, A Kvit, C Jin, J F Muth, and O W Holland. Novel cubic znxmg1-xo epitaxial hetero structures on si (100) substrates. *Solid State Comm.*, 121:9, 2001.
6. S. S. P. Parkin, C. Kaiser, A. Panchula, P. M. Rice, B. Hughes, M. Samant, and S.-H. Yang. Giant tunnelling magnetoresistance at room temperature with MgO(100) tunnel barriers. *Nature Mat.*, 3:862, 2004.
7. S. A. Wolf, D. D. Awschalom, R. A. Buhrman, J. M. Daughton, S. von Molnár, M. L. Roukes, A. Y. Chtchelkanova, and D. M. Treger. Spintronics: A spin-based electronics vision for the future. *Science*, 294:1488, 2001.
8. S. Yuasa, T. Nagahama, A. Fukushima, Y. Suzuki, and K. Ando. Giant room-temperature magnetoresistance in single-crystal Fe/MgO/Fe magnetic tunnel junctions. *Nature Mat.*, 3:868, 2004.
9. P. Zahn, I. Mertig, R. Zeller, and P. H. Dederichs. Screened KKR with hard core potentials. *Phil. Mag. B*, 78:411, 1998.

Editorial Policy

1. Volumes in the following three categories will be published in LNCSE:

i) Research monographs
ii) Lecture and seminar notes
iii) Conference proceedings

Those considering a book which might be suitable for the series are strongly advised to contact the publisher or the series editors at an early stage.

2. Categories i) and ii). These categories will be emphasized by Lecture Notes in Computational Science and Engineering. **Submissions by interdisciplinary teams of authors are encouraged**. The goal is to report new developments – quickly, informally, and in a way that will make them accessible to non-specialists. In the evaluation of submissions timeliness of the work is an important criterion. Texts should be well-rounded, well-written and reasonably self-contained. In most cases the work will contain results of others as well as those of the author(s). In each case the author(s) should provide sufficient motivation, examples, and applications. In this respect, Ph.D. theses will usually be deemed unsuitable for the Lecture Notes series. Proposals for volumes in these categories should be submitted either to one of the series editors or to Springer-Verlag, Heidelberg, and will be refereed. A provisional judgment on the acceptability of a project can be based on partial information about the work: a detailed outline describing the contents of each chapter, the estimated length, a bibliography, and one or two sample chapters – or a first draft. A final decision whether to accept will rest on an evaluation of the completed work which should include

- at least 100 pages of text;
- a table of contents;
- an informative introduction perhaps with some historical remarks which should be accessible to readers unfamiliar with the topic treated;
- a subject index.

3. Category iii). Conference proceedings will be considered for publication provided that they are both of exceptional interest and devoted to a single topic. One (or more) expert participants will act as the scientific editor(s) of the volume. They select the papers which are suitable for inclusion and have them individually refereed as for a journal. Papers not closely related to the central topic are to be excluded. Organizers should contact Lecture Notes in Computational Science and Engineering at the planning stage.

In exceptional cases some other multi-author-volumes may be considered in this category.

4. Format. Only works in English are considered. They should be submitted in camera-ready form according to Springer-Verlag's specifications.
Electronic material can be included if appropriate. Please contact the publisher.
Technical instructions and/or LaTeX macros are available via http://www.springer.com/authors/book+authors? SGWID=0-154102-12-417900-0. The macros can also be sent on request.

General Remarks

Lecture Notes are printed by photo-offset from the master-copy delivered in camera-ready form by the authors. For this purpose Springer-Verlag provides technical instructions for the preparation of manuscripts. See also *Editorial Policy.*

Careful preparation of manuscripts will help keep production time short and ensure a satisfactory appearance of the finished book.

The following terms and conditions hold:

Categories i), ii), and iii):
Authors receive 50 free copies of their book. No royalty is paid. Commitment to publish is made by letter of intent rather than by signing a formal contract. Springer- Verlag secures the copyright for each volume.

For conference proceedings, editors receive a total of 50 free copies of their volume for distribution to the contributing authors.

All categories:
Authors are entitled to purchase further copies of their book and other Springer mathematics books for their personal use, at a discount of 33.3% directly from Springer-Verlag.

Addresses:

Lecture Notes in Computational Science and Engineering

1. D. Funaro, *Spectral Elements for Transport-Dominated Equations.*

2. H. P. Langtangen, *Computational Partial Differential Equations.* Numerical Methods and Diffpack Programming.

3. W. Hackbusch, G. Wittum (eds.), *Multigrid Methods V.*

4. P. Deuflhard, J. Hermans, B. Leimkuhler, A. E. Mark, S. Reich, R. D. Skeel (eds.), *Computational Molecular Dynamics: Challenges, Methods, Ideas.*

5. D. Kröner, M. Ohlberger, C. Rohde (eds.), *An Introduction to Recent Developments in Theory and Numerics for Conservation Laws.*

6. S. Turek, *Efficient Solvers for Incompressible Flow Problems.* An Algorithmic and Computational Approach.

7. R. von Schwerin, ***M**ulti **B**ody **S**ystem **SIM**ulation.* Numerical Methods, Algorithms, and Software.

8. H.-J. Bungartz, F. Durst, C. Zenger (eds.), *High Performance Scientific and Engineering Computing.*

9. T. J. Barth, H. Deconinck (eds.), *High-Order Methods for Computational Physics.*

10. H. P. Langtangen, A. M. Bruaset, E. Quak (eds.), *Advances in Software Tools for Scientific Computing.*

11. B. Cockburn, G. E. Karniadakis, C.-W. Shu (eds.), *Discontinuous Galerkin Methods.* Theory, Computation and Applications.

12. U. van Rienen, *Numerical Methods in Computational Electrodynamics.* Linear Systems in Practical Applications.

13. B. Engquist, L. Johnsson, M. Hammill, F. Short (eds.), *Simulation and Visualization on the Grid.*

14. E. Dick, K. Riemslagh, J. Vierendeels (eds.), *Multigrid Methods VI.*

15. A. Frommer, T. Lippert, B. Medeke, K. Schilling (eds.), *Numerical Challenges in Lattice Quantum Chromodynamics.*

16. J. Lang, *Adaptive Multilevel Solution of Nonlinear Parabolic PDE Systems.* Theory, Algorithm, and Applications.

17. B. I. Wohlmuth, *Discretization Methods and Iterative Solvers Based on Domain Decomposition.*

18. U. van Rienen, M. Günther, D. Hecht (eds.), *Scientific Computing in Electrical Engineering.*

19. I. Babuška, P. G. Ciarlet, T. Miyoshi (eds.), *Mathematical Modeling and Numerical Simulation in Continuum Mechanics.*

20. T. J. Barth, T. Chan, R. Haimes (eds.), *Multiscale and Multiresolution Methods.* Theory and Applications.

21. M. Breuer, F. Durst, C. Zenger (eds.), *High Performance Scientific and Engineering Computing.*

22. K. Urban, *Wavelets in Numerical Simulation.* Problem Adapted Construction and Applications.

23. L. F. Pavarino, A. Toselli (eds.), *Recent Developments in Domain Decomposition Methods.*

24. T. Schlick, H. H. Gan (eds.), *Computational Methods for Macromolecules: Challenges and Applications.*

25. T. J. Barth, H. Deconinck (eds.), *Error Estimation and Adaptive Discretization Methods in Computational Fluid Dynamics*.

26. M. Griebel, M. A. Schweitzer (eds.), *Meshfree Methods for Partial Differential Equations*.

27. S. Müller, *Adaptive Multiscale Schemes for Conservation Laws*.

28. C. Carstensen, S. Funken, W. Hackbusch, R. H. W. Hoppe, P. Monk (eds.), *Computational Electromagnetics*.

29. M. A. Schweitzer, *A Parallel Multilevel Partition of Unity Method for Elliptic Partial Differential Equations*.

30. T. Biegler, O. Ghattas, M. Heinkenschloss, B. van Bloemen Waanders (eds.), *Large-Scale PDE-Constrained Optimization*.

31. M. Ainsworth, P. Davies, D. Duncan, P. Martin, B. Rynne (eds.), *Topics in Computational Wave Propagation*. Direct and Inverse Problems.

32. H. Emmerich, B. Nestler, M. Schreckenberg (eds.), *Interface and Transport Dynamics.* Computational Modelling.

33. H. P. Langtangen, A. Tveito (eds.), *Advanced Topics in Computational Partial Differential Equations.* Numerical Methods and Diffpack Programming.

34. V. John, *Large Eddy Simulation of Turbulent Incompressible Flows.* Analytical and Numerical Results for a Class of LES Models.

35. E. Bänsch (ed.), *Challenges in Scientific Computing - CISC 2002.*

36. B. N. Khoromskij, G. Wittum, *Numerical Solution of Elliptic Differential Equations by Reduction to the Interface.*

37. A. Iske, *Multiresolution Methods in Scattered Data Modelling.*

38. S.-I. Niculescu, K. Gu (eds.), *Advances in Time-Delay Systems.*

39. S. Attinger, P. Koumoutsakos (eds.), *Multiscale Modelling and Simulation.*

40. R. Kornhuber, R. Hoppe, J. Périaux, O. Pironneau, O. Wildlund, J. Xu (eds.), *Domain Decomposition Methods in Science and Engineering.*

41. T. Plewa, T. Linde, V.G. Weirs (eds.), *Adaptive Mesh Refinement – Theory and Applications.*

42. A. Schmidt, K.G. Siebert, *Design of Adaptive Finite Element Software.* The Finite Element Toolbox ALBERTA.

43. M. Griebel, M.A. Schweitzer (eds.), *Meshfree Methods for Partial Differential Equations II.*

44. B. Engquist, P. Lötstedt, O. Runborg (eds.), *Multiscale Methods in Science and Engineering.*

45. P. Benner, V. Mehrmann, D.C. Sorensen (eds.), *Dimension Reduction of Large-Scale Systems.*

46. D. Kressner, *Numerical Methods for General and Structured Eigenvalue Problems.*

47. A. Boriçi, A. Frommer, B. Joó, A. Kennedy, B. Pendleton (eds.), *QCD and Numerical Analysis III.*

48. F. Graziani (ed.), *Computational Methods in Transport.*

49. B. Leimkuhler, C. Chipot, R. Elber, A. Laaksonen, A. Mark, T. Schlick, C. Schütte, R. Skeel (eds.), *New Algorithms for Macromolecular Simulation.*

50. M. Bücker, G. Corliss, P. Hovland, U. Naumann, B. Norris (eds.), *Automatic Differentiation: Applications, Theory, and Implementations.*

51. A.M. Bruaset, A. Tveito (eds.), *Numerical Solution of Partial Differential Equations on Parallel Computers.*

52. K.H. Hoffmann, A. Meyer (eds.), *Parallel Algorithms and Cluster Computing.*

53. H.-J. Bungartz, M. Schäfer (eds.), *Fluid-Structure Interaction.*

54. J. Behrens, *Adaptive Atmospheric Modeling.*

55. O. Widlund, D. Keyes (eds.), *Domain Decomposition Methods in Science and Engineering XVI.*

56. S. Kassinos, C. Langer, G. Iaccarino, P. Moin (eds.), *Complex Effects in Large Eddy Simulations.*

57. M. Griebel, M.A Schweitzer (eds.), *Meshfree Methods for Partial Differential Equations III.*

58. A.N. Gorban, B. Kégl, D.C. Wunsch, A. Zinovyev (eds.), *Principal Manifolds for Data Visualization and Dimension Reduction.*

59. H. Ammari (ed.), *Modeling and Computations in Electromagnetics: A Volume Dedicated to Jean-Claude Nédélec.*

60. U. Langer, M. Discacciati, D. Keyes, O. Widlund, W. Zulehner (eds.), *Domain Decomposition Methods in Science and Engineering XVII.*

61. T. Mathew, *Domain Decomposition Methods for the Numerical Solution of Partial Differential Equations.*

62. F. Graziani (ed.), *Computational Methods in Transport: Verification and Validation.*

63. M. Bebendorf, *Hierarchical Matrices.* A Means to Efficiently Solve Elliptic Boundary Value Problems.

64. C.H. Bischof, H.M. Bücker, P. Hovland, U. Naumann, J. Utke (eds.), *Advances in Automatic Differentiation.*

65. M. Griebel, M.A. Schweitzer (eds.), *Meshfree Methods for Partial Differential Equations IV.*

66. B. Engquist, P. Lötstedt, O. Runborg (eds.), *Multiscale Modeling and Simulation in Science*.

For further information on these books please have a look at our mathematics catalogue at the following URL: `www.springer.com/series/3527`

Monographs in Computational Science and Engineering

1. J. Sundnes, G.T. Lines, X. Cai, B.F. Nielsen, K.-A. Mardal, A. Tveito, *Computing the Electrical Activity in the Heart.*

For further information on this book, please have a look at our mathematics catalogue at the following URL: `www.springer.com/series/7417`

Texts in Computational Science and Engineering

1. H. P. Langtangen, *Computational Partial Differential Equations.* Numerical Methods and Diffpack Programming. 2nd Edition

2. A. Quarteroni, F. Saleri, *Scientific Computing with MATLAB and Octave.* 2nd Edition

3. H. P. Langtangen, *Python Scripting for Computational Science*. 3rd Edition

4. H. Gardner, G. Manduchi, *Design Patterns for e-Science*.

5. M. Griebel, S. Knapek, G. Zumbusch, *Numerical Simulation in Molecular Dynamics*.

For further information on these books please have a look at our mathematics catalogue at the following URL: `www.springer.com/series/5151`