Hans Munthe-Kaas
Brynjulf Owren
Editors

ABEL SYMPOSIA 3

Mathematics and Computation, a Contemporary View

The Abel Symposium 2006

Springer

ABEL SYMPOSIA

Edited by the Norwegian Mathematical Society

Some Participants of the Abel Symposium 2006

From left: Anna-Karin Tornberg, David Bindel, Paul Tupper, Peter D. Lax, Ingrid Daubechies, Brynjulf Owren.
Photo credit: Hans Munthe-Kaas

Hans Munthe-Kaas · Brynjulf Owren
Editors

Mathematics and Computation, a Contemporary View

The Abel Symposium 2006

Proceedings of the Third Abel Symposium, Alesund, Norway, May 25–27, 2006

Springer

Editors
Hans Munthe-Kaas
University of Bergen
Department of Mathematics
Joh. Brunsgt. 12
5008 Bergen
Norway
e-mail: hans.munthe-kaas@mi.uib.no

Brynjulf Owren
Department of Mathematical Sciences
NTNU
7491 Trondheim
Norway
e-mail: bryn@math.ntnu.no

ISBN: 978-3-540-68848-8 e-ISBN: 978-3-540-68850-1
DOI: 10.1007/978-3-540-68850-1

Library of Congress Control Number: 2008932114

Mathematics Subject Classification (2000): 65-06

Cover design: Mönnich, Max

Printed on acid-free paper

9 8 7 6 5 4 3 2 1

springer.com

Preface to the Series

The Niels Henrik Abel Memorial Fund was established by the Norwegian government on January 1. 2002. The main objective is to honor the great Norwegian mathematician Niels Henrik Abel by awarding an international prize for outstanding scientific work in the field of mathematics. The prize shall contribute towards raising the status of mathematics in society and stimulate the interest for science among school children and students. In keeping with this objective the board of the Abel fund has decided to finance one or two Abel Symposia each year. The topic may be selected broadly in the area of pure and applied mathematics. The Symposia should be at the highest international level, and serve to build bridges between the national and international research communities. The Norwegian Mathematical Society is responsible for the events. It has also been decided that the contributions from these Symposia should be presented in a series of proceedings, and Springer Verlag has enthusiastically agreed to publish the series. The board of the Niels Henrik Abel Memorial Fund is confident that the series will be a valuable contribution to the mathematical literature.

Ragnar Winther
Chairman of the board of the Niels Henrik Abel Memorial Fund

Preface

The Abel Symposium 2006 focused on the intersection between computer science, computational science and mathematics. Ever since the early years of computers, applied mathematics has depended heavily upon computational methods. However, in recent years, computation has also been affecting pure mathematics in fundamental ways. Conversely, ideas and methods of pure mathematics are becoming increasingly important in computational and applied mathematics. At the core of computer science is the study of computability and complexity for discrete mathematical structures. Studying the foundations of computational mathematics raises similar questions concerning continuous mathematical structures.

There are several reasons for these developments. The exponential growth of computing power is bringing computational methods into ever new application areas. Equally important is the advance of software and programming languages, which to an increasing degree allows the representation of abstract mathematical structures in program code. Symbolic computing is putting algorithms from mathematical analysis in the hands of pure and applied mathematicians, and the combination of symbolic and numerical techniques is becoming increasingly important both in computational science and in areas of pure mathematics.

We are witnessing a development where a focus on computability, computing and algorithms is contributing towards a unification of areas of computer science, applied and pure mathematics. The basis for this conference was a belief that these developments will prevail in the twenty-first century. The Symposium brought together some of the leading international researchers working in these areas, presented a snapshot of current state of the art, and raised questions about future research directions.

The symposium took place in Ålesund, from May 25–27, 2006 and was organized by

- Ron DeVore, University of South Carolina
- Arieh Iserles, University of Cambridge

- Hans Munthe-Kaas, University of Bergen
- Peter Olver, University of Minnesota
- Brynjulf Owren, NTNU
- Nick Trefethen, University of Oxford

The scientific committee made a deliberate choice to compose a group of international invitees consisting of senior leading researchers and also brilliant young people. The participants were encouraged to be open (and perhaps even provocative) on future developments within computational science. This resulted in a meeting with open and very stimulating discussions.

Talks presented

- Doug Arnold: *Finite element exterior calculus and its applications*
- David Bindel: *Modeling resonant microsystems: toward cell phones on a chip?*
- Folkmar Bornemann: *The whence and whither of using PDEs in computer vision*
- Franco Brezzi: *Recent developments in Mimetic Finite Differences*
- Albert Cohen: *Some remarks on Compressed Sensing*
- Wolfgang Dahmen: *Adaptive multiscale methods*
- Ioana Dumitriu: *Toward accurate polynomial evaluation in rounded arithmetic: foundations for the future*
- Alan Edelman: *New Applications of Random Matrix Theory or Stochastic Eigen-analysis*
- Björn Engquist : *Heterogeneous Multi-scale Methods*
- Anna Gilbert: *Putting the "Computational" in "Computational Harmonic Analysis"*
- Leslie Greengard: *Modern algorithms and the future of mathematical software*
- Tom Hou: *The Interplay between Local Geometric Properties and the Global Regularity for the 3D Incompressible Euler Equations*
- Peter D. Lax: *The numerical solution of hyperbolic systems of conservation laws*
- Christian Lubich: *Variational approximations in quantum dynamics*
- Nilima Nigam: *The good, the bad, and the not-so-ugly: algorithms for computational scattering*
- Guillermo Sapiro: *Mathematics and computation in image processing and other high dimensional signals*
- Stephen Smale: *The mathematics of learning, from machine to human*
- Rob Stevenson: *Optimal adaptive finite element methods*
- Eitan Tadmor: *Theory and computation of entropy stability in quasilinear PDEs*
- Mike Todd: *The role of ellipsoids in optimization theory*
- Anna-Karin Tornberg: *Fluid-structure interactions: the collective dynamics of suspensions*

- Paul Tupper: *A difficult open conjecture in the analysis of molecular dynamics*
- Divakar Viswanath: *Strange attractors from Lorenz to turbulence*
- Shing-Tung Yau: *Minimization with the affine normal direction*

International participants who did not give talks:

- Ingrid Daubechies, Princeton
- Richard Falk, Rutgers
- Ernst Hairer, Geneva
- Reinout Quispel, LaTrobe

Participants from Norwegian universities and research labs:

- Petter Bjørstad, Bergen
- Elena Celledoni, Trondheim
- Snorre Christiansen, Oslo
- Michael Floater, Oslo
- Helge Holden, Trondheim
- Kenneth H.-Karlsen, Oslo
- Trond Kvamsdal, Sintef
- Anne Kværnø, Trondheim
- Hans P. Langtangen, Oslo
- Tom Lyche, Oslo
- Syvert P. Nørsett, Trondheim
- Einar Rønquist, Trondheim
- Trond Steihaug, Bergen
- Tor Sørevik, Bergen
- Xue-Cheng Tai, Bergen
- Warwick Tucker, Bergen
- Ragnar Winther, Oslo
- Antonella Zanna, Bergen

More information about the symposium may be found at this web-page:
`http://abelsymposium.no/2006`

May 23, 2008
Hans Munthe-Kaas

Bergen and Trondheim,
Brynjulf Owren

Contents

List of Contributors

Hsiao-Bing Cheng
Mathematics
Cornell University
Ithaca
New York 14853
USA

Li-Tien Cheng
Mathematics
UCSD
La Jolla, CA
USA

S. Gemmrich
Department of Mathematics
and Statistics
McGill University
805 Sherbrooke
Montreal H3A 2K6
Canada
gemmrich@math.mcgill.ca

Xianfeng Gu
Computer Science
Stony Brook University
Stony Brook
NY 11790
USA
gu@cs.sunysb.edu

Thomas Y. Hou
Applied and Comput. Math
217-50 Caltech, Pasadena
CA 91125
hou@acm.caltech.edu
and
LSEC
Academy of Mathematics
and Systems Sciences
Chinese Academy of Sciences
Beijing 100080, China

Ruo Li
Applied and Comput. Math.
Caltech, Pasadena
CA 91125
and
LMAM&School of Mathematical
Sciences, Peking University, Beijing
China
rli@acm.caltech.edu

N. Nigam
Department of Mathematics
and Statistics
McGill University
805 Sherbrooke
Montreal, QC
Canada H3A 2K6
nigam@math.mcgill.ca

O. Steinbach
Institute of Computational Mathematics, TU Graz
Steyrergasse 30, A 8010 Graz
Austria
o.steinbach@tugraz.at

Eitan Tadmor
Department of Mathematics
Center for Scientific Computation and Mathematical Modeling (CSCAMM)
and Institute for Physical Science and Technology (IPST)
University of Maryland, MD 20742
USA
tadmor@cscamm.umd.edu

P.F. Tupper
Department of Mathematics and Statistics
McGill University
805 Sherbrooke Ouest
Montreal, QC
Canada H3A 2K6
tupper@math.mcgill.ca

D. Viswanath
Department of Mathematics
University of Michigan
530 Church Street
Ann Arbor, MI 48109
USA
divakar@umich.edu

Yalin Wang
Mathematics
Harvard University
Cambridge, MA
USA

Shing-Tung Yau
Mathematics
Harvard University
Cambridge, MA
USA

Weigang Zhong
Department of Mathematics
Center for Scientific Computation and Mathematical Modeling (CSCAMM)
University of Maryland
College Park, MD 20742
wzhong@math.umd.edu

Geometric Methods in Engineering Applications

Xianfeng Gu*, Yalin Wang, Hsiao-Bing Cheng, Li-Tien Cheng, and Shing-Tung Yau

Computer Science, Stony Brook University, Stony Brook, NY 11790, USA, gu@cs.sunysb.edu

Summary. In this work, we introduce two sets of algorithms inspired by the ideas from modern geometry. One is computational conformal geometry method, including harmonic maps, holomorphic 1-forms and Ricci flow. The other one is optimization method using affine normals.

In the first part, we focus on conformal geometry. Conformal structure is a natural structure of metric surfaces. The concepts and methods from conformal geometry play important roles for real applications in scientific computing, computer graphics, computer vision and medical imaging fields.

This work systematically introduces the concepts, methods for numerically computing conformal structures inspired by conformal geometry. The algorithms are theoretically rigorous and practically efficient.

We demonstrate the algorithms by real applications, such as surface matching, global conformal parameterization, conformal brain mapping etc.

In the second part, we consider minimization of a real-valued function f over $\mathbb{R}^{n+1}$ and study the choice of the affine normal of the level set hypersurfaces of f as a direction for minimization. The affine normal vector arises in affine differential geometry when answering the question of what hypersurfaces are invariant under unimodular affine transformations. It can be computed at points of a hypersurface from local geometry or, in an alternative description, centers of gravity of slices. In the case where f is quadratic, the line passing through any chosen point parallel to its affine normal will pass through the critical point of f. We study numerical techniques for calculating affine normal directions of level set surfaces of convex f for minimization algorithms.

1 Introduction

Conformal structure is a natural geometric structure of a metric surface. It is more flexible than Riemannian metric structure and more rigid than topological structure, therefore it has advantages for many important engineering applications.

The first example is from computer graphics. Surface parameterization refers to the process to map a surface onto the planar domain, which plays

a fundamental role in graphics and visualization for the purpose of texture mapping. Surface parameterization can be reformulated as finding a special Riemannian metric with zero Gaussian curvature everywhere, namely a flat metric. If the parameterization is known, then pull back metric induced by the map is the flat metric; conversely, if a flat metric of the surface is known, then the surface can be flattened onto the plane isometrically to induce the parameterization.

The second example is from geometric modeling. Constructing manifold splines on a surface is an important issue for modeling. In order to define parameters and the knots of the spline, special atlas of the surface is required such that all local coordinate transition maps are affine. One way to construct such an atlas is as follows, first a flat metric of the surface is found, then a collection of open sets are located to cover the whole surface, finally each open set is flattened using the flat metric to form the atlas.

The third example is from medical imaging. The human brain cortex surface is highly convolved. In order to compare and register brain cortex surfaces, it is highly desirable to canonically map them to the unit sphere. This is equivalent to find a Riemannian metric on the cortex surface, such that the Gaussian curvature induced by this metric equals to one everywhere. Once such a metric is obtained, the cortex surface can be coherently glued onto the sphere piece by piece isometrically.

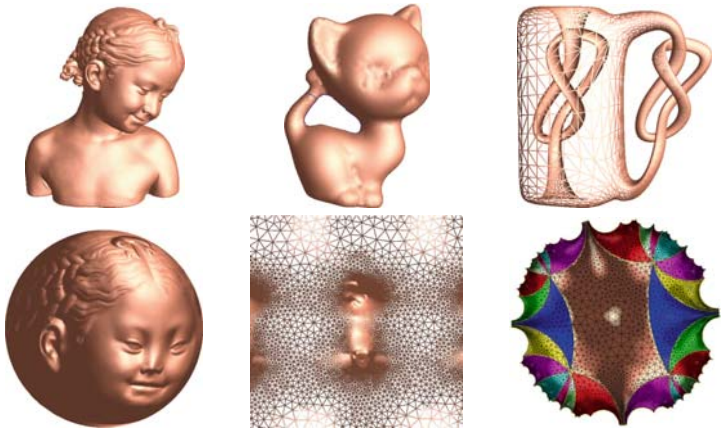
For most applications, the desired metrics should minimize the angle distortion and the area distortion. The angles measured by the new metric should be consistent with those measured by the original metric. The existence of such metrics can be summarized as Riemann uniformization theorem. Finding those metrics is equivalent to compute surface conformal structure. Therefore, it is of fundamental importance to compute conformal structures of general surfaces.

In modern geometry, conformal geometry of surfaces are studied in Riemann surface theory. Riemann surface theory is a rich and mature field, it is the intersection of many subjects, such as algebraic geometry, algebraic topology, differential geometry, complex geometry etc. This work focuses on converting theoretic results in Riemann surface theory to practical algorithms.

2 Previous Works

Much research has been done on mesh parameterization due to its usefulness in computer graphics applications. The survey of [Floater and Hormann 2005] provides excellent reviews on various kinds of mesh parameterization techniques. Here, we briefly discuss the previous work on the conformal mesh parameterization.

Several researches on conformal mesh parameterization tried to discretize the nature of the conformality such that any intersection angle at any point on a given manifold is preserved on the parameterized one at the corresponding



Spherical Euclidean Hyperbolic

point. Floater [Floater 1997] introduced a mesh parameterization technique based on convex combinations. For each vertex, its 1-ring stencil is parameterized into a local parameterization space while preserving angles, and then the convex combination of the vertex is computed in the local parameterization spaces. The overall parameterization is obtained by solving a sparse linear system. [Sheffer and de Sturler 2001] presented a constrained minimization approach, so called angle-based flattening (ABF), such that the variation between the set of angles of an original mesh and one of 2D flatten version is minimized. In order to obtain a valid and flipping-free parameterization, several angular and geometric constraints are incorporated with the minimization process. Lately, they improved the performance of ABF by using an advanced numerical approach and a hierarchical technique [Sheffer et al. 2005].

Recently, much research has been incorporated with the theories of differential geometry. [Levy et al. 2002] applied the Cauchy–Riemann equation for mesh parameterization and provided successful results on the constrained 2D parameterizations with free boundaries. [Desbrun et al. 2002] minimized the Dirichlet energy defined on triangle meshes for computing conformal parameterization. It has been noted that the approach of [Desbrun et al. 2002] has the same expressional power with [Levy et al. 2002]. Gu and Yau [Gu and Yau 2003] computed the conformal structure using the Hodge theory. A flat metric of the given surface is induced by computing the holomorphic 1-form with a genus-related number of singularities and used for obtaining a globally smooth parameterization. [Gortler et al. 2005] used discrete 1-forms for mesh parameterization. Their approach provided an interesting result in mesh parameterization with several holes, but they cannot control the curvatures on the boundaries. Ray et al. [Ray et al. 2005] used the holomorphic 1-form to follow up the principle curvatures on manifolds and computed a quad-dominated parameterization from arbitrary models.

Kharevych et al. [Kharevych et al. 2005] applied the theory of circle patterns from [Bobenko and Springborn 2004] to globally conformal parameterizations. They obtain the uniform conformality by preserving intersection angles among the circum-circles each of which is defined from a triangle on the given mesh. In their approach, the set of angles is non-linear optimized first, and then the solution is refined with cooperating geometric constraints. They provide several parameterization results, such as 2D parameterization with predefined boundary curvatures, spherical parameterization, and globally smooth parameterization of a high genus model with introduced singularity points. [Gu et al. 2005] used the discrete Ricci flow [Chow and Luo 2003] for generating manifold splines with a single extraordinary point. The Ricci flow is utilized for obtaining 2D parameterization of high-genus models in their paper.

In theory, the Ricci flow [Chow and Luo 2003] and the variations with circle patterns [Bobenko and Springborn 2004] have the same mathematical power. However, because of the simplicity of the implementation, we adopt the Ricci flow as a mathematical tool for the parameterization process.

In contrast to all previous approaches, the parameterization spaces in our interests are not only the 2D spaces but also arbitrary hyperbolic spaces. As a result, we can provide novel classes of applications in this paper, such as parameterization with interior and exterior boundaries having prescribed curvatures, PolyCube-mapping, quasi-conformal cross-parameterization with high-genus surfaces, and geometry signatures.

3 Theoretic Background

In this section, we introduce the theories of conformal geometry.

3.1 Riemann Surface

Suppose S is a two dimensional topological manifold covered by a collection of open sets $\{U_\alpha\}$, $S \subset \bigcup_\alpha U_\alpha$. A homeomorphism $\phi_\alpha : U_\alpha \to \mathbb{C}$ maps U_α to the complex plane. (U_α, ϕ_α) forms a local coordinate system. Suppose two open sets U_α and U_β intersect, then each point $p \in U_\alpha \bigcap U_\beta$ has two local coordinates, the transformation between the local coordinates is defined as the *transition function*

$$\phi_{\alpha\beta} := \phi_\beta \circ \phi_\alpha^{-1} : \phi_\alpha(U_\alpha \cap U_\beta) \to \phi_\beta(U_\beta \cap U_\beta). \tag{1}$$

A complex function $f : \mathbb{C} \to \mathbb{C}$ is *holomorphic*, if its derivative exists. If f is invertible, and f^{-1} is also holomorphic, then f is called *bi-holomorphic*.

Definition 1 (Conformal Structure). *A two dimensional topological manifold S with an atlas $\{(U_\alpha, \phi_\alpha)\}$, if all transition functions $\phi_{\alpha\beta}$'s are bi-holomorphic, then the atlas is called a conformal atlas. The union of all conformal atlas is called the conformal structure of S.*

A surface with conformal structure is called a Riemann surface. All metric surfaces are Riemann surfaces.

3.2 Uniformization Metric

Suppose S is a C^2 smooth surface embedded in $\mathbb{R}^3$ with parameter (u^1, v^2). The position vector is $\mathbf{r}(u^1, u^2)$, the tangent vector is $d\mathbf{r} = \mathbf{r}_1 du^1 + \mathbf{r}_2 du^2$, where $\mathbf{r}_1, \mathbf{r}_2$ are the partial derivatives of $\mathbf{r}$ with respect to u^1, u^2 respectively. The length of the tangent vector is represented as the *first fundamental form*

$$ds^2 = \sum g_{ij} du^1 du^2 \tag{2}$$

where $g_{ij} =< \mathbf{r}_i, \mathbf{r}_j >$. The matrix (g_{ij}) is called the *Riemannian metric* matrix.

A special parameterization can be chosen to simplify the Riemannian metric, such that $g_{11} = g_{22} = e^{2\lambda}$ and $g_{12} = 0$, such parameter is called the *isothermal coordinates*. If all the local coordinates of an atlas are isothermal coordinates, then the atlas is the conformal atlas of the surface. For all orientable metric surfaces, such atlas exist, namely

Theorem 1 (Riemann Surface). *All orientable metric surfaces are Riemann surfaces.*

The *Gauss curvature* measures the deviation of a neighborhood of a point on the surface from a plane, using isothermal coordinates, the Gaussian curvature is calculated as

$$K = -\frac{2}{e^{2\lambda}} \Delta\lambda, \tag{3}$$

where Δ is the Laplace operator on the parameter domain.

Theorem 2 (Gauss–Bonnet). *Suppose a closed surface S, the Riemannian metric $\mathbf{g}$ induces the Gaussian curvature function K, then the total curvature is determined by*

$$\int_S K dA = 2\pi\chi(S), \tag{4}$$

where $\chi(S)$ is the Euler number of S.

Suppose $u : S \to \mathbf{R}$ is a function defined on the surface S, then $e^{2u}\mathbf{g}$ is another Riemannian metric on S. Given arbitrary two tangent vectors at one

point, the angle between them can be measured by $\mathbf{g}$ or $e^{2u}\mathbf{g}$, the two measurements are equal. Therefore we say $e^{2u}\mathbf{g}$ is *conformal* (or angle preserving) to $\mathbf{g}$. $(S, \mathbf{g})$ and $(S, e^{2u}\mathbf{g})$ are endowed with different Riemannian metrics but the same conformal structure.

The following Poincaré uniformization theorem postulates the existence of the conformal metric which induces constant Gaussian curvature,

Theorem 3 (Poincaré Uniformization). *Let $(S, \mathbf{g})$ be a compact 2-dimensional Riemannian manifold, then there is a metric $\bar{\mathbf{g}}$ conformal to $\mathbf{g}$ which has constant Gauss curvature.*

Such a metric is called the *uniformization metric*. According to Gauss–Bonnet Theorem 4, the sign of the constant Gauss curvature is determined by the Euler number of the surface. Therefore, all closed surfaces can be conformally mapped to three canonical surfaces, the sphere for genus zero surfaces $\chi > 0$, the plane for genus one surfaces $\chi = 0$, and the hyperbolic space for high genus surfaces $\chi < 0$.

3.3 Holomorphic 1-Forms

Holomorphic and meromorphic functions can be defined on the Riemann surface via conformal structure. Holomorphic differential forms can also be defined.

Definition 2 (holomorphic 1-form). *Suppose S is a Riemann surface with conformal atlas $\{(U_\alpha, z_\alpha\}$, where z_α is the local coordinates. Suppose a complex differential form ω is represented as*

$$\omega = f_\alpha(z_\alpha)dz_\alpha,$$

where f_α is a holomorphic function, then ω is called a holomorphic 1-form.

Holomorphic 1-forms play important roles in computing conformal structures.

A holomorphic 1-form can be interpreted as a pair of vector fields, $\omega_1 + \sqrt{-1}\omega_2$, such that the curl and divergence of ω_1, ω_2 are zeros,

$$\nabla \times \omega_i = 0, \nabla \cdot \omega_i = 0, i = 1, 2,$$

and

$$\mathbf{n} \times \omega_1 = \omega_2,$$

everywhere on the surface. Both ω_i are *harmonic 1-forms*, the following Hodge theorem clarifies the existence and uniqueness of harmonic 1-forms.

Theorem 4 (Hodge). *Each cohomologous class of 1-forms has a unique harmonic 1-form.*

3.4 Ricci Flow

In geometric analysis, *Ricci flow* is a powerful tool to compute Riemannian metric. Recently, Ricci flow is applied to prove the Poincaré conjecture. Ricci flow is the process to deform the metric $\mathbf{g}(t)$ according to its induced Gauss curvature $K(t)$, where t is the time parameter

$$\frac{dg_{ij}(t)}{dt} = -K(t)g_{ij}(t). \tag{5}$$

It is proven that the curvature evolution induced by the Ricci flow is exactly like heat diffusion on the surface

$$\frac{K(t)}{dt} = -\Delta_{\mathbf{g}(t)}K(t), \tag{6}$$

where $\Delta_{\mathbf{g}(t)}$ is the Laplace–Beltrami operator induced by the metric $\mathbf{g}(t)$. Ricci flow converges, the metric $\mathbf{g}(t)$ is conformal to the original metric at any time t. Eventually, the Gauss curvature will become constant just like the heat diffusion $K(\infty) \equiv const$, the limit metric $\mathbf{g}(\infty)$ is the *uniformization metric*.

3.5 Harmonic Maps

Suppose S_1, S_2 are metric surfaces embedded in $\mathbb{R}^3$. $\phi : S_1 \to S_2$ is a map from S_1 to S_2. The harmonic energy of the map is defined as

$$E(\phi) = \int_{S_1} < \nabla\phi, \nabla\phi > dA.$$

The critical point of the harmonic energy is called the *harmonic maps*.

The normal component of the Laplacian is

$$\Delta\phi^{\perp} =< \Delta\phi, \mathbf{n} \circ \phi > \mathbf{n},$$

If ϕ is a harmonic map, then the tangent component of Laplacian vanishes,

$$\Delta\phi = \Delta\phi^{\perp},$$

where Δ is the Laplace–Beltrami operator.

We can diffuse a map to a harmonic map by the heat flow method:

$$\frac{d\phi}{dt} = -(\Delta\phi - \Delta\phi^{\perp}).$$

4 Computational Algorithms

In practice, all surfaces are represented as simplicial complexes embedded in the Euclidean space, namely, triangular meshes. All the algorithms are discrete approximations of their continuous counter parts. We denote a mesh by M, and use v_i to denote its ith vertex, edge e_{ij} for the edge connecting v_i and v_j, and f_{ijk} for the triangle formed by v_i, v_j and v_k, which are ordered counter-clock-wisely.

If a mesh M is with boundaries, we fist convert it to a closed symmetric mesh $\bar{M}$ by the following *double covering* algorithm:

1. Make a copy mesh M' of M.
2. Reverse the orientation of M' by change the order of vertices of each face, $f_{ijk} \to f_{jik}$.
3. Glue M and M' along their boundaries to form a closed mesh $\bar{M}$.

In the following discussion, we always assume the surfaces are closed. We first introduce harmonic maps for genus zero surfaces, then holomorphic one-forms for genus one surfaces and finally Ricci flow method for high genus surfaces.

4.1 Genus Zero Surfaces: Harmonic Maps

For genus zero surfaces, the major algorithm to compute their conformal mapping is *harmonic maps*, the basic procedure is to diffuse a degree one map until the map becomes harmonic:

1. Compute the normal of each face, then compute the normal of each vertex as the average of normals of neighboring faces.
2. Set the map ϕ equals to the Gauss map,
$$\phi(v_i) = \mathbf{n}_i.$$
3. Diffuse the map by Heat flow acting on the maps
$$\phi(v_i) - = (\Delta\phi(v_i) - \Delta\phi(v_i)^{\perp})\varepsilon$$
where $\Delta\phi(v_i))^{\perp}$ is defined as
$$< \Delta\phi(v_i), \phi(v_i) > \phi(v_i).$$
4. Normalize the map by setting
$$\phi(v_i) = \frac{\phi(v_i) - \mathbf{c}}{|\phi(v_i) - \mathbf{c}|},$$
where $\mathbf{c}$ is the mass center defined as
$$\mathbf{c} = \sum_{v_i} \phi(v_i).$$
5. Repeat step 2 and 3, until $\Delta\phi(v_i))$ is very closed to $\Delta\phi(v_i))^{\perp}$.

where Δ is a discrete Laplace operator, defined as

$$\Delta\phi(v_i) = \sum_j w_{ij}(\phi(v_i) - \phi(v_i)),$$

where v_j is a vertex adjacent to v_i, w_{ij} is the edge weight

$$w_{ij} = \frac{\cot\alpha + \cot\beta)}{2},$$

α, β are the two angles against edge e_{ij}.

The harmonic map $\phi : M \to \mathbb{S}^2$ is also conformal. The conformal maps are not unique, suppose $\phi_1, \phi_2 : M \to \mathbb{S}^2$ are two conformal maps, then $\phi_1 \circ \phi_2^{-1} : \mathbb{S}^2 \to \mathbb{S}^2$ is a conformal map from sphere to itself, it must be a so-called Möbius transformation. Suppose we map the sphere to the complex plane by a stereo-graphics projection

$$(x, y, z) \to \frac{2x + 2\sqrt{-1}y}{2 - z},$$

then the Möbius transformation has the form

$$w \to \frac{aw + b}{cw + d}, ad - bc = 1, a, b, c, d \in \mathbb{C}.$$

The purpose of the normalization step is to remove Möbius ambiguity of the conformal map from M to $\mathbb{S}^2$.

For genus zero open surfaces, the conformal mapping is straight forward:

1. Double cover M' to get $\bar{M}$.
2. Conformally map the doubled surface to the unit sphere.
3. Use the sphere Möbius transformation to make the mapping symmetric.
4. Use stereographic projection to map each hemisphere to the unit disk.

The Möbius transformation on the disk is also a conformal map and with the form

$$w \to e^{i\theta}\frac{w - w_0}{1 - \bar{w}_0 w}, \tag{7}$$

where w_0 is arbitrary point inside the disk, *theta* is an angle. Figure 1 illustrates two conformal maps from the David head surface to the unit disk, which differ by a Möbius transformation.

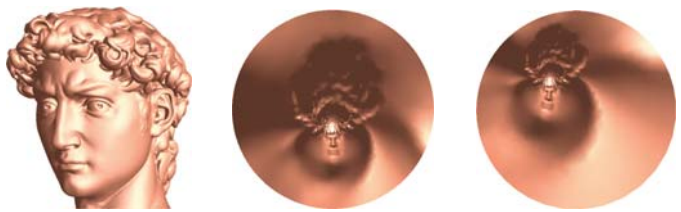


Fig. 1. According to Riemann Mapping theorem, a topological disk can be conformally mapped to the unit disk. Two such conformal maps differ by a Möbius transformation of the unit disk

4.2 Genus One Surfaces: Holomorphic One-Forms

For genus one closed surfaces, we compute the basis of holomorphic one-form group, which induces the conformal parameterization directly. A holomorphic one-form is formed by a pair of harmonic one-forms ω_1, ω_2, such that ω_2 is conjugate to ω_1.

In order to compute harmonic one-forms, we need to compute the homology basis for the surface. A homology base curve is a consecutive halfedges, which form a closed loop. First we compute a *cut graph* of the mesh, then extract a homology basis from the cut graph. Algorithm for cut graph:

1. Compute the dual mesh $\bar{M}$, each edge $e \in M$ has a unique dual edge $\bar{e} \in \bar{M}$.
2. Compute a spanning tree $\bar{T}$ of $\bar{M}$, which covers all the vertices of $\bar{M}$.
3. The cut graph is the union of all edges whose dual are not in $\bar{T}$,

$$G = \{e \in M | \bar{e} \notin \bar{T}\}.$$

Then, we can compute homology basis from G:

1. Compute a spanning tree T of G.
2. G
   $T = \{e_1, e_2, \cdots, e_n\}$.
3. $e_i \cup T$ has a unique loop, denoted as γ_i.
4. $\{\gamma_1, \gamma_2, \cdots, \gamma_n\}$ form a homology basis of M.

A harmonic one-form is represented as a linear map from the halfedge to the real number, $\omega : \{Half\ Edges\} \to \mathbb{R}$, such that

$$\begin{cases} \omega \partial f & \equiv 0 \\ \Delta\omega & \equiv 0 \\ \int_{\gamma_i} \omega & = c_i \end{cases} \tag{8}$$

where ∂ represents boundary operator, $\partial f_{ijk} = e_{ij} + e_{jk} + e_{ki}$, therefore $\omega \partial f_{ijk} = \omega(e_{ij}) + \omega(e_{jk}) + \omega(e_{ki})$; $\Delta\omega$ represents the Laplacian of ω,

$$\Delta\omega(v_i) = \sum_j w_{ij}\omega(h_{ij}),$$

h_{ij} are the half edges from v_i to v_j; $\{c_i\}$ are prescribed real numbers. It can be shown that the solution to the above equation group exists and is unique.

On each face f_{ijk} there exists a unique vector $\mathbf{t}$, such that on each edge, $\omega(h_{ij}) = < v_j - v_i, \mathbf{t} >, \omega(h_{jk}) = < v_k - v_j, \mathbf{t} >$ and $\omega(h_{ki}) = < v_i - v_k, \mathbf{t} >$. Let $\mathbf{t}' = \mathbf{n} \times \mathbf{t}$, then $\omega'(h_{ij}) = < v_j - v_i, \mathbf{t}' >$ defines another harmonic one-form, which is conjugate to ω, (ω, ω') form a holomorphic one-form.

We cut a surface M along its cut graph to get a topological disk D_M, by gluing multiple copies of D_M consistently, we can construct a finite portion of the universal covering space of M.

We then integrate a holomorphic one-form to map D_M to the plane conformally in the following way:

Fig. 2. Holomorphic 1-forms on different surfaces

1. Fix one vertex $v_0 \in D_M$, and map it to the origin $\phi(v_0) = (0, 0)$
2. For any vertex $v \in D_M$, compute the shortest path γ from v_0 to v in D_M
3. $\phi(v) = (\int_\gamma \omega, \int_\gamma \omega')$

We then visualize the holomorphic one-forms by texture mapping a checker board onto D_M using texture coordinates ϕ. Figure 2 demonstrates the holomorphic 1-forms on three different surfaces.

4.3 High Genus Surfaces: Discrete Ricci Flow

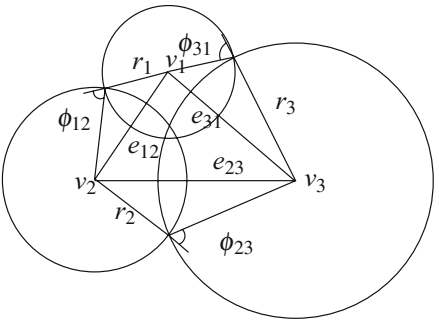
For high genus surfaces, we apply discrete Ricci flow method to compute their uniformization metric and then embed them in the hyperbolic space.

Circle Packing Metric

We associate each vertex v_i with a circle with radius γ_i. On edge e_{ij}, the two circles intersect at the angle of Φ_{ij}. The edge lengths are

$$l_{ij}^2 = \gamma_i^2 + \gamma_j^2 + 2\gamma_i\gamma_j \cos\Phi_{ij}$$

A circle packing metric is denoted by $\{\Sigma, \Phi, \Gamma\}$, where Σ is the triangulation, Φ the edge angle, Γ the vertex radii.



Two circle packing metrics $\{\Sigma, \Phi_1, \Gamma_1\}$ and $\{\Sigma, \Phi_2, \Gamma_2\}$ are conformal equivalent, if:

- The radii of circles are different, $\Gamma_1 \neq \Gamma_2$.
- The intersection angles are same, $\Phi_1 \equiv \Phi_2$.

In practice, the circle radii and intersection angles are optimized to approximate the induced Euclidean metric of the mesh as close as possible.

Poincaré Disk

According to Riemann uniformization theorem, high genus surfaces can be conformally embed in hyperbolic space. Instead of treat each triangle as an Euclidean triangle, we can treat each triangle as a hyperbolic triangle. The hyperbolic space is represented using *Poincaré disk*, which is the unit disk on the complex plane, with Riemannian metric

$$ds^2 = \frac{4dwd\bar{w}}{(1-\bar{w}w)^2}.$$

The rigid motion in Poincaré disk is Möbius transformation 7. The geodesics are circle arcs which are orthogonal to the unit circle. A hyperbolic circle in Poincaré disk with center c and radius r is also an Euclidean circle with center C and radius R, such that $\mathbf{C} = \frac{2-2\mu^2}{1-\mu^2|\mathbf{c}|^2}$ and $R^2 = |\mathbf{C}|^2 - \frac{|\mathbf{c}|^2-\mu^2}{1-\mu^2|\mathbf{c}|^2}$, $\mu = \frac{e^r-1}{e^r+1}$.

Hyperbolic Ricci Flow

Let

$$u_i = \log tanh \frac{\gamma_i}{2}, \tag{9}$$

then discrete hyperbolic Ricci flow is defined as

$$\frac{du_i}{dt} = \bar{K}_i - K_i. \tag{10}$$

In fact, discrete Ricci flow is the gradient flow of the following *hyperbolic Ricci energy*

$$f(\mathbf{u}) = \int_{\mathbf{u}_0}^{\mathbf{u}} \sum_{i=1}^{n} (\bar{K}_i - K_i) du_i, \tag{11}$$

where n is the number of edges, $\mathbf{u} = (u_1, u_2, \cdots, u_m)$, m is the number of vertices. In practice, if we set $\bar{K}_i \equiv 0$ by minimizing the Ricci energy using Newton's method, the hyperbolic uniformization metric can be computed efficiently.

Once the hyperbolic metric for a mesh is calculated, the mesh can be flattened face by face in the Poincaré disk. Determining the position of a vertex in the Poincaré disk is equivalent to finding the intersection between two hyperbolic circles, which can be converted as finding the intersection between two Euclidean circles.

5 Applications

Conformal geometry has broad applications in medical imaging, computer graphics, geometric modeling and many other fields.

5.1 Conformal Brain Mapping

Human cortex surfaces are highly convoluted, it is difficult to analyze and study them. By using conformal maps, we can map the brain surface to the canonical unit sphere and carry out all the geometric processing, analysis, measurement on the spherical domain. Because the conformal map preserves angle structure, local shapes are well preserved, it is valuable for visualization purpose. Different cortical surfaces can be automatically registered on the canonical parameter domain, it is more efficient to compare surfaces using conformal brain mapping.

Figure 3 illustrates an example of conformal brain mapping. The cortical surface is reconstructed from MRI images and converted as a triangular mesh.

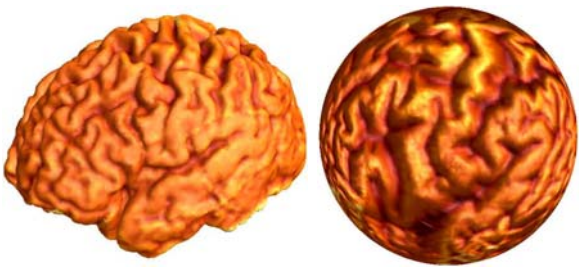


Fig. 3. Conformal brain mapping

Fig. 4. Texture mapping using conformal mapping

5.2 Global Conformal Parameterization

In computer graphics, surface parameterization plays an important role for various applications, such as texture mapping, texture synthesis.

Basically, a surface is mapped to the plane, the planar coordinates of each vertex are used as texture coordinates. It is highly desirable to reduce the distortion between the texture image and the geometric surface. Conformal mapping is useful because it is angle distortion free. Figure 4 illustrates an example for texture mapping using global conformal parameterization of a genus two surface.

5.3 Manifold Splines

In geometric modeling, conventional splines are defined on the planar domains. It is highly desirable to define splines on surfaces with arbitrary topologies directly.

In order to define splines on manifolds, one needs to compute a special atlas of the manifold, such that all chart transition maps are affine. Such kind of atlas can be easily constructed by integrating a holomorphic 1-form.

Figure 5 demonstrates one example of genus 6 surface. The holomorphic 1-form induces an affine atlas with singularities, the planar Powell–Sabin splines are defined on the atlas directly.

6 Affine Normal

Many problems in engineering field can be formulated as optimization problems. Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is a differentiable function, finding its critical points is the basic task.

Situated at any point, a natural direction to choose for minimization is the steepest descent direction. This is the direction along which the function

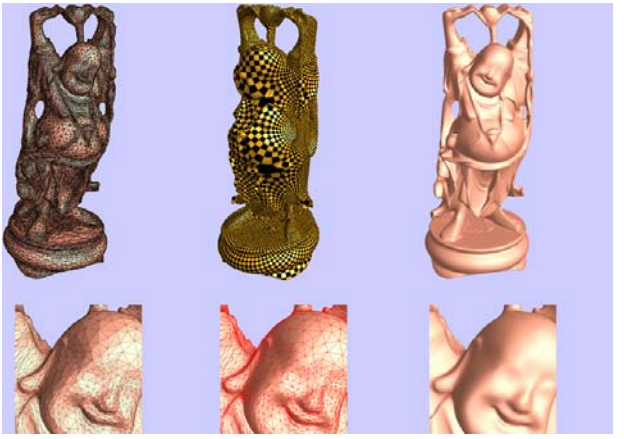


Fig. 5. Manifold splines constructed from holomorphic 1-form

is locally diminishing the most rapidly. The steepest descent direction can be computed from derivative information of f through the form $-\nabla f$. Unfortunately, while this direction is intuitively sound, it shows slow convergence.

Newton's method use quadratic approximation at the origin,

$$f(\mathbf{x}) \approx \mathbf{x}^T \nabla^2 f(0)\mathbf{x} + \nabla f^T(0)\mathbf{x} + c.$$

When the quadratic approximation is taken at a point $\mathbf{y}$, the critical point is at $\mathbf{x} = \mathbf{y} - (\nabla^2 f(\mathbf{y})^{-1} \nabla f(\mathbf{y})$, therefore, one can use $\mathbf{x}$ to replace $\mathbf{y}$ as the next guess. By iteration, the critical point can be reached. Newton's method quadratically converges. But it is expensive to compute the inverse of the second derivative matrices, the Hessian matrices.

6.1 Affine Normal

Let M be a hypersurface in $\mathbb{R}^{n+1}$, N is the normal vector field on M. Now if X and Y are vector fields on M and $D_X Y$ is the flat connection on $\mathbb{R}^{n+1}$, then we decompose

$$D_X Y = \nabla_X Y + h(X, Y)N,$$

where $\nabla_X Y$ is the tangential part of $D_X Y$ and $h(X, Y)$ the normal part, also known as the second fundamental form. Furthermore, in this case, $\nabla_X Y$ is the Levi–Civita connection of the Riemannian metric induced by $\mathbb{R}^{n+1}$ on M.

We choose an arbitrary local frame field $e_1, e_2, \cdots, e_n$ tangent to M and $det(e_1, e_2, \cdots, e_{n+1}) = 1$, we may define h as $D_X Y = \nabla_X Y + h(X, Y)e_{n+1}$, to arrive at the affine metric

$$II_{ik} = H^{-\frac{1}{n+2}} h_{ik},$$

where H is the determinant $det\{h_{ik}\}$. In this case, the affine normal field is given by ΔM, where the Laplacian is with respect to the affine metric II and M is the position vector of M.

Suppose M is a level set surface of the function f, we can derive the affine normal field as

$$H^{\frac{1}{n+2}} \begin{pmatrix} f^{ij}(-\frac{n}{n+2} f^{pq} f_{pqi} + n \frac{f_{n+1,i}}{|\nabla f|}) \\ -\frac{n}{|\nabla f|} \end{pmatrix},$$

where the coordinates x_i used are rotated so that x_{n+1} is in the normal direction. It can be shown that when the hypersurface is an ellipsoid, all affine normals point towards its center. In fact, the affine normals of the level sets of a quadratic polynomial will point toward the unique critical point, even if that critical point is unstable.

6.2 Affine Normal Descent Algorithm

Using this affine normal field, we can summarize our algorithm in the following steps, iterated to convergence:

1. Compute the affine normal direction to the level set of the function at the current approximation location.
2. Use a line search to find the minimum of the function along that direction. This location serves as the new approximation.

We call this the affine normal descent algorithm. For the quadratic minimization problem, due to the nature of the affine normal and the ellipsoidal level sets of f, the approximations of this algorithm will take the value of the exact minimum after one iteration. Thus, the affine normal and the vector $-(\nabla^2 f)^{-1}\nabla f$ used in Newton's method are parallel to each other in this case. This means we may view this algorithm as an extension of the steepest descent method, using the affine normal direction, which points at the center of ellipsoids, instead of the steepest descent direction, which points at the center of spheres. On the other hand, we may view it as a relative of Newton's method, both exact for the quadratic minimization problem but with different higher order terms.

6.3 Efficiency

In terms of computational costs, we note that the previously derived formula for the affine normal direction requires first, second, and third derivatives of f, as well as inversion of an $n \times n$ matrix of second derivatives. While it may be possible to generate other forms or approximations of the affine normal direction that simplify the inversion or diminish the need of derivative information. Instead, we consider a different viewpoint of the affine normal to bypass the need for such information. Consider a convex hypersurface, a point on that surface, and the tangent plane located there. Furthermore, consider the class of planes intersecting with the surface and parallel to the

Table 1. Five-dimensional Result: In this five-dimensional example, convergence is achieved after five iterations

j	$f(p_j)$	$\lvert p_j - p_{j-1}\rvert$
0	2.02669978966015	
1	−0.87543513107826	2.17147295853185
2	−1.16211480429203	0.13960261665982
3	−1.16232787552543	0.00003483822789
4	−1.16232787579106	0.00001466875411
5	−1.16232787579106	0.00000000001789

tangent plane. On each of these planes, we look at the center of gravity of the region enclosed by the intersection of the plane with the surface. The union of these centers of gravity forms a curve. It turns out that the one-sided tangent direction of this curve at the point of interest is the affine normal vector. Thus, an alternate approach for calculating the affine normal vector involves calculating centers of gravity, completely bypassing the need for derivative information higher than that of the first derivative which is required for tangent planes.

6.4 Experimental Results

We tested our method for several cases and measure the accuracy and efficiency. For a five-dimensional convex function, let

$$f(\mathbf{x}) = \sum_{i=1}^{5} (x_i^2 + \sin x_i).$$

Let $(-0.2, 0, 0.4, 1, -0.3)$ be the starting point. The iterations of our algorithm are shown in Table 1, the algorithm converge to the point

$$\begin{pmatrix} -0.45018354967147 \\ -0.45018354967140 \\ -0.45018354967139 \\ -0.45018354967125 \\ -0.45018354967129 \end{pmatrix}$$

From the statistics, we can see that the affine normal method is efficient and practical.

7 Conclusion

This paper introduces some algorithms inspired by geometric insights.

We first introduce a series of computational algorithms to compute conformal Riemannian metrics on surfaces, especially the uniformization metrics. The algorithms include harmonic maps, holomorphic 1-forms and surface Ricci

flow on discrete meshes. The methods are applied for various applications in computer graphics, medical imaging and geometric modeling.

In the future, we will generalize these algorithms for discrete 3-manifolds represented as tetrahedral meshes.

Second, we introduce an efficient optimization algorithm based on affine differential geometry, which reaches the critical point for quadratic functions in one step. The method is practical and efficient. In the future, we will improve the method for computing affine normals.

References

[Bobenko and Springborn 2004] Bobenko, A. I., and Springborn, B. A. 2004. Variational principles for circle patterns and koebe's theorem. *Transactions on the American Mathematical Society 356*, 659–689.

[Chow and Luo 2003] Chow, B., and Luo, F. 2003. Combinatorial ricci flow on surfaces. *Journal of Differential Geometry 63*, 1, 97–129.

[Desbrun et al. 2002] Desbrun, M., Meyer, M., and Alliez, P. 2002. Intrinsic parameterizations of surface meshes. *Computer Graphics Forum (Proc. Eurographics 2002) 21*, 3, 209–218.

[Floater and Hormann 2005] Floater, M. S., and Hormann, K. 2005. Surface parameterization: a tutorial and survey. In *Advances in Multiresolution for Geometric Modelling*, N. A. Dodgson, M. S. Floater, M. A. Sabin, N. A. Dodgson, M. S. Floater, and M. A. Sabin, Eds., Mathematics and Visualization. Springer, 157–186.

[Floater 1997] Floater, M. S. 1997. Parametrization and smooth approximation of surface triangulations. *Computer Aided Geometric Design 14*, 3, 231–250.

[Gortler et al. 2005] Gortler, S. J., Gotsman, C., and Thurston, D. 2005. Discrete one-forms on meshes and applications to 3D mesh parameterization. *Computer Aided Geometric Design*, to appear.

[Gu and Yau 2003] Gu, X., and Yau, S.-T. 2003. Global conformal surface parameterization. *Symposium on Geometry Processing*, 127–137.

[Gu et al. 2005] Gu, X., He, Y., Jin, M., Luo, F., Qin, H., and Yau, S.-T. 2005. Manifold splines with single extraordinary point. *submitted for publication*.

[Kharevych et al. 2005] Kharevych, L., Springborn, B., and Schröder, P. 2005. Discrete conformal mappings via circle patterns. *submitted for publication*.

[Levy et al. 2002] Levy, B., Petitjean, S., Ray, N., and Maillot, J. 2002. Least squares conformal maps for automatic texture atlas generation. *SIGGRAPH 2002*, 362–371.

[Ray et al. 2005] Ray, N., Li, W. C., Levy, B., Sheffer, A., and Alliez, P. 2005. Periodic global parameterization. *submitted for publication*.

[Sheffer and de Sturler 2001] Sheffer, A., and de Sturler, E. 2001. Parameterization of faced surfaces for meshing using angle based flattening. *Engineering with Computers 17*, 3, 326–337.

[Sheffer et al. 2005] Sheffer, A., Lévy, B., Mogilnitsky, M., and Bogomyakov, A. 2005. ABF++: Fast and robust angle based flattening. *ACM Transactions on Graphics 24*, 2, 311–330.

Boundary Integral Equations for the Laplace–Beltrami Operator

S. Gemmrich, N. Nigam*, and O. Steinbach

Department of Mathematics and Statistics, McGill University, 805 Sherbrooke, Montreal, QC, Canada H3A 2K6, nigam@math.mcgill.ca

1 Introduction and Motivation

We present a boundary integral method, and an accompanying boundary element discretization, for solving boundary-value problems for the Laplace–Beltrami operator on the surface of the unit sphere $\mathcal{S}$ in $\mathbb{R}^3$. We consider a closed curve $\mathcal{C}$ on $\mathcal{S}$ which divides $\mathcal{S}$ into two parts $\mathcal{S}_1$ and $\mathcal{S}_2$. In particular, $\mathcal{C} = \partial\mathcal{S}_1$ is the boundary curve of $\mathcal{S}_1$. We are interested in solving a boundary value problem for the Laplace–Beltrami operator in $\mathcal{S}_2$, with boundary data prescribed on $\mathcal{C}$.

We shall begin by describing a physical problem of interest. Then, we derive an integral representation formula for solutions of the Laplace–Beltrami operator on the sphere, and introduce the single and double layer potentials [5, 8]. We investigate their jump properties, and use these to derive an integral equation for the solution of a Dirichlet problem. A variational strategy is presented, along with some numerical experiments validating our ideas. To the best of our knowledge, the discretizations of these integral equations have not been studied before. We believe they present an elegant and natural solution strategy for boundary value problems on the sphere.

This work is motivated in part by recent investigations into the motion of point vortices on spheres, specifically in bounded regions with walls on the sphere. Kidambi and Newton [4] considered such a problem, assuming the bounded sub-surface of the sphere lent itself to the method of images. Crowdy, in a series of papers [1, 2, 3], has also investigated the motion of vortices on spheres. In [3], he uses conformal mapping onto the complex plane to study the motion of a vortex on a sphere with walls. We shall study a closely related model problem, for which the methods of [3, 4] would be applicable. However, we propose an integral-equation method instead which is valid for *any* bounded sub-region $\mathcal{S}_2$, provided the curve $\mathcal{C}$ is sufficiently smooth. This technique will be valid even where the method of images is not, and which does not involve explicit knowledge of conformal mappings between the stereographically-projected subregion of interest, and the upper half of the complex plane.

1.1 Point Vortex Motion on a Sphere with Walls

The underlying physical phenomenon considered in [3] is the motion of a point vortex in an incompressible fluid on the surface of the unit sphere, $\mathcal{S}$. There is a bounded solid region, denoted $\mathcal{S}_1 \subseteq \mathcal{S}$, with a simply connected boundary, $\mathcal{C}$. No fluid can penetrate into $\mathcal{S}_1$. Let $\mathcal{S}_2$ be the surface of the sphere excluding $\mathcal{S}_1 \cup \mathcal{C}$, see Fig. 1.

A point on the sphere $\mathcal{S}$ will be described in terms of the spherical angles,

$$x(\varphi, \theta) = \begin{pmatrix} \cos\varphi \sin\theta \\ \sin\varphi \sin\theta \\ \cos\theta \end{pmatrix} \in \mathcal{S}, \quad \varphi \in [0, 2\pi), \theta \in [0, \pi].$$

We consider a point vortex of strength κ located at a point $x_0 \in \mathcal{S}_2$. The flow motion is assumed irrotational, except for the point vorticity associated with the vortex. This assumption needs some justification, which we will discuss below. The incompressible nature of the fluid allows us to prescribe a stream function, $\Psi(x_0, x)$, for the fluid velocity. That is, the velocity field satisfies

$$u = \nabla\Psi \times \mathbf{e}_r.$$

Here $\mathbf{e}_r$ is the unit radial vector to the surface. The vorticity is then defined as

$$\boldsymbol{\omega} = \omega \mathbf{e}_r := \nabla \times u.$$

If the fluid motion is irrotational except at x_0, then $\omega = 0$ except at that point. We insist that the boundary $\mathcal{C}$ be a streamline of the motion. Without loss of generality, we can set the streamline constant to zero. The function Ψ is really the Green's function for the Laplace–Beltrami operator on the subsurface $\mathcal{S}_2$ of the sphere:

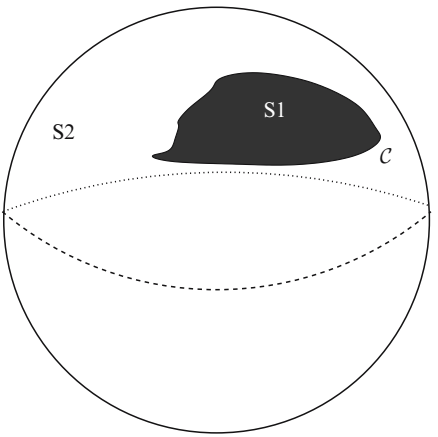


Fig. 1. The unit sphere, $\mathcal{S}$, with an impenetrable island $\mathcal{S}_1$ on the surface. $\mathcal{C}$ is the boundary of the island

$$-\triangle_{\mathcal{S}}\,\Psi(x_0,x)=\kappa\delta(|x-x_0|), \qquad \forall x\in\mathcal{S}_2, \tag{1a}$$

$$\Psi(x_0,x)=0, \qquad \forall x\in\mathcal{C}. \tag{1b}$$

We recall that, in spherical coordinates, $\triangle_{\mathcal{S}}$ is defined as

$$\Delta_S u(x) \;=\; \left[\frac{1}{\sin^2\theta}\frac{\partial^2}{\partial\varphi^2}+\frac{1}{\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial}{\partial\theta}\right)\right]u(x(\varphi,\theta)). \tag{2}$$

The assumption that the fluid motion is irrotational can be justified by noting that since the fluid is incompressible, it can perfectly slip at Σ. This allows us to prescribe a circulation at $\mathcal{C}$, so that the Gauss constraint for the vorticity,

$$\int_{S_2}\omega\,d\sigma=0,$$

is satisfied.

Analogously to finding Green's functions in the plane, we can find Ψ in terms of the fundamental singularity U of the Laplace–Beltrami operator on the entire surface of the sphere, and a smooth function $v_{x_0}(x)$. That is,

$$\Psi(x,x_0)=U(x,x_0)+v_{x_0}(x) \tag{3}$$

where v_{x_0} solves

$$-\triangle_{\mathcal{S}}v_{x_0}(x)=\frac{\kappa}{4\pi}, \qquad \forall x\in\Omega, \qquad v_{x_0}(x)=-U(x_0,x), \qquad x\in\mathcal{C}. \tag{4}$$

We can interpret the fundamental singularity, U, as the stream function for a point vortex of strength κ on the sphere without boundaries. We denote the fundamental singularity, with $\kappa=1$, as U^* henceforth. Note that from [3], we get that

$$U(x_a,x)=-\kappa\log\left|\frac{(z-z_a)(\overline{z-z_a})}{(1+|z|^2)(1+|z_a|^2)}\right|$$

where we've stated this in terms of the mapped points z, z_a in the complex plane:

$$z=\cot(\theta/2)e^{i\phi}, \qquad x=(\theta,\phi)$$

The fundamental singularity U satisfies the partial differential equation:

$$-\triangle_{\mathcal{S}}U=\kappa\left(\delta(|x-x_0|)-\frac{1}{4\pi}\right), \tag{5a}$$

and the Gauss condition for the vorticity, $\omega=-\triangle_{\mathcal{S}}\,U$:

$$\int_S\omega\,ds=0. \tag{5b}$$

The implication of (5a) is that there is a "sea" of uniform vorticity, $\frac{1}{4\pi}$, in which the point vortex at x_0 must be embedded if moving on the whole

sphere. We cannot find a distribution $\tilde{U}$ which satisfies $-\triangle_S \tilde{U} = \delta|x - x_0|$ on the sphere, and which simultaneously satisfies the Gauss constraint. In order to satisfy the constraint and simultaneously have an irrotational flow, we must either counterbalance the point vortex at x_0 by another vortex on the sphere, or have the entire fluid moving with a uniform background vorticity. This feature of the fundamental singularity will appear again in the next section, and will require us to impose a side constraint on the solution density, when employing integral equations. This is reminiscent of similar problems arising in the solution of potential problems in unbounded regions of the plane.

At this juncture, we could use the Green's function Ψ to study the motion of the point vortex, which is governed in the stereographic coordinates z by

$$\frac{\partial z_0}{\partial t} = \frac{-i}{2}(1 + |z_0|^2)\frac{\partial v_{z_0}}{\partial z}|_{z=z_0}.$$

The solutions of this evolution equation are described in terms of level sets of the smooth part, $v_{x_0}(x_0) = \textit{constant}$. Such an investigation is performed in the papers by Kidambi and Newton [4], and Crowdy [3]. Instead, we shall study a closely related mathematical model problem.

Consider the Dirichlet boundary value problem in $\mathcal{S}_2$ for the Laplace–Beltrami operator:
Find a smooth u such that for given Dirichlet data g

$$\Delta_{\mathcal{S}} u(x) = 0 \quad \text{for } x \in \mathcal{S}_{\in}, \tag{6a}$$

$$u(x) = g(x) \quad \text{for } x \in \mathcal{C} \tag{6b}$$

We wish to solve (6) by reformulating the boundary value problem as an integral equation. As usual, the process of reformulation is not unique; we shall be employing a layer ansatz, and solving an integral equation of the first kind for the unknown density. We note that we could equivalently have chosen to study the Neumann or Robin problem for the system. We could use the Green's function for $\mathcal{S}_2$, Ψ, to solve this Dirichlet problem. The methods suggested in [4] and [3] would also be applicable for our model problem, with some caveats: the method of Kidambi and Newton relies on the ability to use the method of images, while Crowdy's work requires knowledge of a conformal map from the stereographically-projected $\mathcal{S}_2$ into the upper half plane or the unit circle. Instead, we propose an integral equation method which is valid for *any* bounded sub-region $\mathcal{S}_2$, provided the curve $\mathcal{C}$ is sufficiently smooth, and without conformally mapping to the plane. Additionally, integral equations allow us to solve problems with lower regularity properties, a feature we shall explore in upcoming work.

If $\mathcal{S}_1$ were to degenerate, ie, if the interior of $\mathcal{C}$ had zero area, we would need to add extra conditions to satisfy the Gauss constraint. Mathematically, we would be dealing with the screen problem, and anticipate singular behaviour on the corners of the screen. On the surface of the entire sphere without walls, we must either embed the point vortex in a fluid of uniform vorticity (hence no longer irrotational), or counter-balance it by another point vortex.

2 An Integral Representation Formula on the Sphere

We begin by reminding the reader of some vectorial identities on the sphere. Let $\mathbf{e}_r, \mathbf{e}_\theta, \mathbf{e}_\varphi$ be the usual unit vectors in spherical coordinates. Recall that we can define the surface gradient of a scalar f on $\mathcal{S}$ as

$$\nabla_{\mathcal{S}} f(x) \;=\; \frac{1}{\sin\theta}\frac{\partial f}{\partial \varphi}\,\mathbf{e}_\varphi + \frac{\partial f}{\partial \theta}\,\mathbf{e}_\theta.$$

In the same way we introduce the surface divergence for a vector-valued function $\mathbf{V}$ on the sphere as

$$\mathrm{div}_{\mathcal{S}}\mathbf{V}(x) \;=\; \frac{1}{\sin\theta}\left(\frac{\partial}{\partial \varphi}V_\varphi(\varphi,\theta) + \frac{\partial}{\partial \theta}(\sin\theta\, V_\theta(\varphi,\theta))\right).$$

We easily see the identity:

$$\Delta_{\mathcal{S}} u(x) \;=\; \mathrm{div}_{\mathcal{S}}\nabla_{\mathcal{S}} u(x).$$

We introduce the vectorial surface rotation for a scalar field f on the sphere:

$$\underline{\mathrm{curl}}_{\mathcal{S}} f(x) \;=\; -\frac{\partial f}{\partial \theta}\,\mathbf{e}_\varphi + \frac{1}{\sin\theta}\frac{\partial f}{\partial \varphi}\,\mathbf{e}_\theta$$

and the (scalar) surface rotation of a vector field $\mathbf{V}$ as

$$\mathrm{curl}_{\mathcal{S}}\mathbf{V}(x) \;=\; \frac{1}{\sin\theta}\left(-\frac{\partial}{\partial \varphi}V_\theta(\varphi,\theta) + \frac{\partial}{\partial \theta}(\sin\theta\, V_\varphi(\varphi,\theta))\right).$$

We then obtain another vectorial identity for the Laplace–Beltrami operator:

$$\Delta_{\mathcal{S}} u(x) \;=\; -\mathrm{curl}_{\mathcal{S}}\underline{\mathrm{curl}}_{\mathcal{S}} u(x) \quad \text{for } x \in \mathcal{S}.$$

We shall be using a variational setting for most of this paper; to this end, we introduce the inner product

$$\langle u, v\rangle_{L_2(\mathcal{S})} \;=\; \int_{\mathcal{S}} u(x)v(x)d\sigma_x \;=\; \int_0^{2\pi}\int_0^{\pi} u(x(\varphi,\theta))v(x(\varphi,\theta))\sin\theta\, d\theta d\varphi.$$

We shall now derive the Green's identiy. We find, by integration by parts,

$$\langle -\Delta_{\mathcal{S}} u, v\rangle_{L_2(\mathcal{S})} \;=\; a_{\mathcal{S}}(u,v) \;=\; a_{\mathcal{S}}(v,u) \;=\; \langle u, -\Delta_{\mathcal{S}} v\rangle_{L_2(\mathcal{S})}$$

where we have introduced the symmetric bilinear form

$$\begin{aligned} a_{\mathcal{S}}(u,v) &:= \int_0^{2\pi}\int_0^{\pi}\left[\frac{1}{\sin\theta}\frac{\partial}{\partial\varphi}u(\varphi,\theta)\frac{\partial}{\partial\varphi}v(\varphi,\theta) + \sin\theta\frac{\partial}{\partial\theta}v(\varphi,\theta)\frac{\partial}{\partial\theta}u(\varphi,\theta)\right]d\theta d\varphi \\ &= \int_{\mathcal{S}} \nabla_S u(x)\cdot\nabla_S v(x)\, d\sigma_x. \end{aligned}$$

Stoke's theorem for the positively oriented curve $\mathcal{C}$ and region $\mathcal{S}_2$ may be written as

$$\int_{\mathcal{S}_2} \mathrm{curl}_{\mathcal{S}} \mathbf{V}(x) d\sigma_x \;=\; \int_{\mathcal{C}} \mathbf{V}(x) \cdot \mathbf{t}(x)\, ds_x.$$

Here, $\mathbf{t}$ is the unit tangent vector to $\mathcal{C}$. We note that a similar identity holds for the region $\mathcal{S}_1$, with care taken with the orientation of the tangent. Now, setting $\mathbf{V} = v(x)\mathbf{W}(x)$ and applying the product rule we get

$$\int_{\mathcal{S}_2} \underline{\mathrm{curl}}_{\mathcal{S}} v(x) \cdot \mathbf{W}(x)\, d\sigma_x \;=\; -\int_{\mathcal{C}} v(x)[\mathbf{W}(x) \cdot \mathbf{t}(x)] ds_x + \int_{\mathcal{S}_2} v(x) \mathrm{curl}_{\mathcal{S}} \mathbf{W}(x) d\sigma_x.$$

With $\mathbf{W}(x) = \underline{\mathrm{curl}}_{\mathcal{S}} u(x)$ we finally obtain Green's first formula for the Laplace–Beltrami operator,

$$-\int_{\mathcal{S}_2} \underline{\mathrm{curl}}_{\mathcal{S}} v(x) \cdot \underline{\mathrm{curl}}_{\mathcal{S}} u(x)\, d\sigma_x = \int_{\mathcal{C}} v(x)[\underline{\mathrm{curl}}_{\mathcal{S}} u(x) \cdot \underline{t}(x)] ds_x + \int_{\mathcal{S}_2} v(x) \Delta_{\mathcal{S}} u(x) d\sigma_x. \tag{7}$$

Note that the left hand side of (7) coincides with the bilinear form $a_{\mathcal{S}}(u, v)$, with the role of $\mathcal{S}$ being played by $\mathcal{S}_2$.

2.1 Fundamental Solution and a Representation Formula

Proposition 1. [6, 7] *The fundamental solution of the Laplace–Beltrami operator $\Delta_{\mathcal{S}}$ as defined in* (2) *is given by*

$$U^*(x, x_0) = -\frac{1}{4\pi} \log |1 - (x, x_0)| \tag{8}$$

$$= -\frac{1}{4\pi} \log \left[1 - \cos(\varphi - \varphi_0) \sin\theta \sin\theta_0 - \cos\theta \cos\theta_0\right]. \tag{9}$$

In particular,

$$\Delta_{\mathcal{S}} U^*(x), x_0) \;=\; \frac{1}{4\pi} \tag{10}$$

for $x = x(\varphi, \theta)$, $x_0 = x(\varphi_0, \theta_0) \in \mathcal{S}$ with $(\varphi, \theta) \neq (\varphi_0, \theta_0)$.

Remark 1. For $x, x_0 \in \mathcal{S}$ we have

$$|x - x_0|^2 = |x|^2 - 2(x, x_0) + |x_0|^2 = 2[1 - (x, x_0)].$$

Hence we obtain

$$-\frac{1}{2\pi} \log |x - x_0| \;=\; -\frac{1}{4\pi} \left[\log[1 - (x, x_0)] + \log 2\right].$$

In particular, the fundamental solution of the three-dimensional Laplace–Beltrami operator corresponds to the fundamental solution of the two-dimensional Laplace operator.

The first Green's identity can be used to derive a representation formula for smooth functions defined on the sphere.

Proposition 2. *Every sufficiently smooth function u on $\mathcal{S}_2$ ($u \in \mathcal{C}^2(\mathcal{S}_2) \cap \mathcal{C}^1(\bar{\mathcal{S}}_2)$) satisfies the following representation formula*

$$\begin{aligned}
&\frac{1}{4\pi}\int_{\mathcal{S}_2} u(x)d\sigma_x - \int_{\mathcal{S}_2} U^*(x,x_0)\Delta_{\mathcal{S}}u(x)d\sigma_x \\
&-\int_{\mathcal{C}} U^*(x,x_0)\,\underline{curl}_{\mathcal{S}}u(x)\cdot\mathbf{t}(x)ds_x + \int_{\mathcal{C}} u(x)\,\underline{curl}_{\mathcal{S}}U^*(x,x_0)\cdot\mathbf{t}(x)ds_x \\
&= \begin{cases} u(x_0) & \text{if } x_0 \in \mathcal{S}_1, \\ 0 & \text{if } x_0 \in \mathcal{S}\setminus\bar{\mathcal{S}}_1. \end{cases}
\end{aligned} \tag{11}$$

Proof. We obtain Green's second formula by interchanging the roles of u and v in (7), adding the two identities and using the symmetry of the left hand side.

$$\begin{aligned}
&\int_{\mathcal{S}_2} u(x)\Delta_{\mathcal{S}}v(x) - v(x)\Delta_{\mathcal{S}}u(x)\,d\sigma_x \\
&= \int_{\mathcal{C}} [v(x)\,\underline{\mathrm{curl}}_{\mathcal{S}}u(x) - u(x)\,\underline{\mathrm{curl}}_{\mathcal{S}}v(x)]\cdot\mathbf{t}(x)ds_x.
\end{aligned} \tag{12}$$

We define the ε-neighbourhood of x_0 on $\mathcal{S}$, $B_\varepsilon(x_0) := \{y \in \mathcal{S} : |y - x_0| > \varepsilon\}$ and set $\mathcal{S}_{2,\varepsilon} := \mathcal{S}\setminus B_\varepsilon(x_0)$. The second Green's formula for $\mathcal{S}_{2,\varepsilon}$ with $v(x) = U^*(x,x_0)$ yields:

$$\begin{aligned}
&\frac{1}{4\pi}\int_{\mathcal{S}_{2,\varepsilon}} u(x)\,d\sigma_x - \int_{\mathcal{S}_{2,\varepsilon}} U^*(x,x_0)\,\Delta_{\mathcal{S}}u(x)\,d\sigma_x \\
&= \int_{\mathcal{C}} [U^*(x,x_0)\,\underline{\mathrm{curl}}_{\mathcal{S}}u(x) - u(x)\,\underline{\mathrm{curl}}_{\mathcal{S}}U^*(x,x_0)]\cdot\mathbf{t}(x)ds_x \\
&+ \int_{\partial B_\varepsilon(x_0)} [U^*(x,x_0)\,\underline{\mathrm{curl}}_{\mathcal{S}}u(x) - u(x)\,\underline{\mathrm{curl}}_{\mathcal{S}}U^*(x,x_0)]\cdot\mathbf{t}(x)ds_x
\end{aligned} \tag{13}$$

First we observe that

$$\left|\int_{\mathcal{S}_2} U^*(x,x_0)\,\Delta_{\mathcal{S}}u(x)\,d\sigma_x\right| \le ||\Delta_{\mathcal{S}}\,u||_{L^\infty(\mathcal{S})}\int_{\mathcal{S}_2}|U^*(x,x_0)|\,d\sigma_x \le M,$$

and hence:

$$\lim_{\varepsilon\to 0}\int_{\mathcal{S}_{2,\varepsilon}} U^*(x,x_0)\,\Delta_{\mathcal{S}}u(x)\,d\sigma_x = \int_{\mathcal{S}_2} U^*(x,x_0)\,\Delta_{\mathcal{S}}u(x)\,d\sigma_x.$$

Furthermore, we can estimate the integral

$$\left| \int\limits_{\partial B_\varepsilon} U^*(x,x_0)\, \underline{\mathrm{curl}}_{\mathcal{S}} u(x) \cdot \mathbf{t}(x)\, ds_x \right| \leq ||\underline{\mathrm{curl}}_{\mathcal{S}} u||_{L^\infty} \int\limits_{\partial B_\varepsilon} |U^*(x,x_0)|\, ds_x,$$

By changing coordinates, without loss of generality x_0 can be taken to be the north pole of the sphere, i.e. $x_0 = (0,0,1)^\top$. The curve ∂B_ε is then fully described by the latitude θ_ε, say. According to the cosine law we have $\cos\theta_\varepsilon = 1 - \frac{\varepsilon^2}{2}$ and thus

$$\begin{aligned} \int\limits_{\partial B_\varepsilon} |U^*(x,x_0)|\, ds_x &= \frac{1}{4\pi} \int_{2\pi}^{0} \log|1-\cos\theta_\varepsilon|\, \sin\theta_\varepsilon\, d\varphi \\ &= -\frac{1}{2}\varepsilon\sqrt{(1-\frac{\varepsilon^2}{4})}\, \log\frac{\varepsilon^2}{2} \longrightarrow 0 \quad (\text{as } \varepsilon \to 0). \end{aligned}$$

To analyse the second contribution along ∂B_ε, we again assume x_0 to be the northpole. We then compute

$$\underline{\mathrm{curl}}_{\mathcal{S}} U^*(x,x_0) = \frac{\sin\theta_\varepsilon}{4\pi(1-\cos\theta_\varepsilon)} \mathbf{e}_\varphi.$$

Since the line element on the surface of the sphere is given by

$$\mathbf{t}(x(\varphi,\theta)) \cdot ds_{x(\varphi,\theta)} = d\theta\, \mathbf{e}_\theta + \sin\theta d\varphi\, \mathbf{e}_\varphi,$$

we deduce that (note the orientation of ∂B_ε):

$$-\int\limits_{\partial B_\varepsilon} u(x)[\underline{\mathrm{curl}}_{\mathcal{S}} U^*(x,x_0)\cdot\mathbf{t}(x)]ds_x = -\frac{\sin^2\theta_\varepsilon}{4\pi(1-\cos\theta_\varepsilon)} \int_{2\pi}^{0} u(\varphi,\theta_\varepsilon)\, d\varphi,$$

which in the limit as $\varepsilon \longrightarrow 0$ tends to

$$\frac{u(x_0)}{2} \lim_{\varepsilon\to 0} \frac{\varepsilon^2\,(1-\frac{\varepsilon^2}{4})}{\frac{\varepsilon^2}{2}} = u(x_0).$$

Taking the limit as $\epsilon \to 0$ in (13) proves the result. □

At this juncture, we draw the reader's attention to the term $\int_{\mathcal{S}_2} u(x)\, d\sigma_x$ in the representation formula (11). If u satisfied $\triangle_{\mathcal{S}} u = 0$ in $\mathcal{S}_2$, the familiar integral representation formula for the Laplacian in 2-D would not involve such a term; indeed, the left hand side of the representation formula consists of line integrals only if u satisfies the side constraint, $\int_{\mathcal{S}_2} u\, d\sigma_x = 0$. This is linked to the Gauss constraint.

3 Layer Potentials and Boundary Integral Operators

Having derived an integral representation formula for solutions of the Laplace–Beltrami problem in $\mathcal{S}_2$ in the previous section, we are now in a position to reformulate the boundary value problem as an integral equation.

3.1 Single and Double Layer Potentials

Following the integral representation derived in Proposition 2 we define the following two layer potentials:

- The *single layer potential* with sufficiently smooth density function σ:

$$(\widetilde{V}\sigma)(x) := \int_{\mathcal{C}} U^*(x,y)\,\sigma(y)\,ds_y \quad \text{for } x \notin \mathcal{C}$$

- And the *double layer potential* with sufficiently smooth density function μ:

$$(\widetilde{W}\mu)(x) := \int_{\mathcal{C}} \mu(y)\,[\underline{\mathrm{curl}}_{\mathcal{S}} U^*(x,y)\cdot \mathbf{t}(y)]\,ds_y \quad \text{for } x \notin \mathcal{C}$$

By Proposition 2, every solution to the homogeneous Laplace–Beltrami equation can be written as the sum of a single and a double layer potential modulo a constant. This is the starting point for the so-called direct boundary integral approach. However, for the purpose of this paper we follow the layer ansatz based on the following observation.

For $x \notin \mathcal{C}$, the single layer potential satisfies:

$$\begin{aligned}\Delta_{\mathcal{S}}(\widetilde{V}\sigma)(x) &= \Delta_{\mathcal{S}} \int_{\mathcal{C}} U^*(x,y)\,\sigma(y)\,ds_y = \int_{\mathcal{C}} \Delta_{\mathcal{S}} U^*(x,y)\,\sigma(y)\,ds_y \\ &= \frac{1}{4\pi}\int_{\mathcal{C}} \sigma(y)\,ds_y = 0 \qquad (14a)\end{aligned}$$

$$\text{under the constraint } \int_{\mathcal{C}} \sigma(y)\,ds_y = 0. \qquad (14b)$$

Hence, we may find the general solution of the Dirichlet boundary value problem (6) as

$$u(x) = \int_{\mathcal{C}} U^*(x,y)\,\sigma(y)\,ds_y + p, \qquad (15)$$

where $p \in \mathbb{R}$ is some Lagrange multiplier related to the constraint (14b). Similarly, the double layer potential satisfies the Laplace–Beltrami equation for $x \notin \mathcal{C}$:

$$\begin{aligned}\Delta_{\mathcal{S}}(\widetilde{W}\mu)(x) &= \Delta_{\mathcal{S}} \int_{\mathcal{C}} \mu(y)\,[\underline{\mathrm{curl}}_{\mathcal{S}} U^*(x,y)\cdot \mathbf{t}(y)]\,ds_y \\ &= \int_{\mathcal{C}} \mu(y)[\Delta_{\mathcal{S}}\,\underline{\mathrm{curl}}_{\mathcal{S}}\, U^*(x,y)\cdot \mathbf{t}(y)]\,ds_y \\ &= \int_{\mathcal{C}} \mu(y)[\underline{\mathrm{curl}}_{\mathcal{S}}\,\Delta_{\mathcal{S}} U^*(x,y)\cdot \mathbf{t}(y)]\,ds_y \\ &= 0,\end{aligned}$$

without any further constraints on the density μ. We might thus also try to look for the solution to (6) in the form of a double layer.

3.2 Jump Relations for $\widetilde{V}$ and $\widetilde{W}$

In the previous section, we have only defined the layer potentials for x away from the boundary curve. However, in order to align the operators with the given Dirichlet data along $\mathcal{C}$, we need to investigate their behavior in the limit as x approaches $\mathcal{C}$. Similarly, if one is interested in solving the Neumann problem in which the tangential component of the vectorial surface rotation is prescribed along $\mathcal{C}$, one has to investigate the limit features of this quantity for the layer potentials. In both cases, there will be certain jump relations across the curve $\mathcal{C}$. For the purpose of this paper however, we will restrict ourselves to the Dirichlet case. First, consider the single layer potential with density σ for $\tilde{x} \notin \mathcal{C}$:

$$\begin{aligned}(\widetilde{V}\sigma)(\tilde{x}) &= \int_{\mathcal{C}} U^*(\tilde{x},x)\,\sigma(x)\,ds_x \\ &= -\frac{1}{4\pi}\int_{\mathcal{C}} \log[1-\langle \tilde{x}\,,\,x\rangle]\,\sigma(x)\,ds_x \qquad (16)\end{aligned}$$

The following lemma describes the limit behavior of the single layer potential.

Lemma 1. *For $x_0 \in \mathcal{C}$ we have:*

$$(V\sigma)(x_0) := \lim_{\mathcal{S}\ni\tilde{x}\to x_0} (\widetilde{V}\sigma)(\tilde{x}) = \int_{\mathcal{C}} U^*(x_0,y)\,\sigma(y)\,ds_y$$

as a weakly singular line integral and hence $(\widetilde{V}\sigma)$ is continuous across $\mathcal{C}$.

Proof. Fix an arbitrary $\varepsilon > 0$. Let $x_0 \in \mathcal{C}$ be fixed, and $\tilde{x} \in \mathcal{S}$ satisfy $|\tilde{x}-x_0|<\varepsilon$. Introduce the notation

$$\mathcal{C}_{\varepsilon,\le} := \{y \in \mathcal{C}, |y-x_0| \le \varepsilon\}, \qquad \mathcal{C}_{\varepsilon,>} := \{y \in \mathcal{C}, |y-x_0| > \varepsilon\}.$$

Then, if we define

$$I_\varepsilon(\tilde{x}) := \int_{\mathcal{C}} U^*(\tilde{x}, y)\, \sigma(y)\, ds_y - \int_{\mathcal{C}_{\varepsilon,>}} U^*(x_0, y)\, \sigma(y)\, ds_y,$$

we can easily show

$$I_\varepsilon = \int_{\mathcal{C}_{\varepsilon,>}} [U^*(\tilde{x}, y) \, - \, U^*(x_0, y)] \, \sigma(y)\, ds_y + \int_{\mathcal{C}_{\varepsilon,\leq}} U^*(\tilde{x}, y)\, \sigma(y)\, ds_y. \quad (17)$$

The first integral in (17) vanishes in the limit as $\tilde{x} \longrightarrow x_0$, i.e.

$$\lim_{\tilde{x}\to x_0} \int_{\mathcal{C}_{\varepsilon,>}} [U^*(\tilde{x}, y) \, - \, U^*(x_0, y)] \, \sigma(y)\, ds_y = 0.$$

The second term in (17) we can bound in terms of the density σ:

$$\left| \int_{\mathcal{C}_{\varepsilon,\leq}} U^*(\tilde{x}, y)\, \sigma(y)\, ds_y \right| \leq \|\sigma\|_{L_\infty(\mathcal{C})} \int_{\mathcal{C}_{\varepsilon,\leq}} |U^*(\tilde{x}, y)| \, ds_y.$$

To finish the proof, note that we can estimate

$$\int_{\mathcal{C}_{\varepsilon,\leq}} |U^*(\tilde{x}, y)| \, ds_y \leq \int_{\substack{y\,\in\,\mathcal{C} \\ |y-\tilde{x}|\,\leq\, 2\varepsilon}} |U^*(\tilde{x}, y)| \, ds_y \overset{\tilde{x}\to x_0}{\longrightarrow} \int_{\substack{y\,\in\,\mathcal{C} \\ |y-x_0|\,\leq\, 2\varepsilon}} |U^*(x_0, y)| \, ds_y \overset{\varepsilon\to 0}{\longrightarrow} 0.$$

Putting these estimates together, we see that $\lim\limits_{\varepsilon\to 0} \lim\limits_{\tilde{x}\to x} I_\varepsilon(\tilde{x}) = 0$, which proves the assertion. □

The case of the double layer potential is slightly more involved, since the limit process reveals a hidden delta function. To see this, consider the double layer potential with density μ for $x \notin \mathcal{C}$:

$$\begin{aligned}(\widetilde{W}\mu)(x) &= \int_{\mathcal{C}} \mu(\tilde{x}) \, [\underline{\mathrm{curl}}_{\mathcal{S}} U^*(x, \tilde{x}) \cdot \mathbf{t}(\tilde{x})] \, ds_{\tilde{x}} \\ &= \frac{1}{4\pi} \int_{\mathcal{C}} \mu(\tilde{x}) \, \frac{1}{A(x,\tilde{x})} \left[\begin{pmatrix} -\cos(\varphi - \widetilde{\varphi}) \cos\theta \sin\widetilde{\theta} + \sin\theta \cos\widetilde{\theta} \\ -\sin\widetilde{\theta} \sin(\varphi - \widetilde{\varphi}) \end{pmatrix} \cdot \mathbf{t}(\tilde{x}) \right] ds_{\tilde{x}} \end{aligned} \quad (18)$$

$A(x, \tilde{x}) = 1 - \cos(\varphi - \widetilde{\varphi}) \sin\theta \sin\widetilde{\theta} - \cos\theta \cos\widetilde{\theta}$. Where $x = x(\varphi, \theta)$ and $\tilde{x} = x(\widetilde{\varphi}, \widetilde{\theta})$

Lemma 2. *For $x_0 \in \mathcal{C}$ we have:*

$$(\gamma_0^{\mathcal{S}_2}\widetilde{W}\mu)(x_0) := \lim_{\mathcal{S}_2 \ni x \to x_0} (\widetilde{W}\mu)(x)$$
$$= (K\mu)(x_0) + \left(1 - \frac{\alpha(x_0)}{2\pi}\right)\mu(x_0),$$

where $\alpha(x_0)$ represents the interior (with respect to $\mathcal{S}_2$) angle of $\mathcal{C}$ at x_0. For a smooth curve, $\alpha = \pi$. The operator $(K\mu)$ is given by the following integral expression:

$$(K\mu)(x_0) = \lim_{\varepsilon \to 0} \quad (K_\varepsilon \mu)(x_0)$$
$$= \lim_{\varepsilon \to 0} \int\limits_{|\tilde{x} - x_0| \geq \varepsilon} \mu(\tilde{x})\ [\underline{curl}_{\mathcal{S}}\, U^*(x, \tilde{x}) \cdot \mathbf{t}(\tilde{x})]\ ds_{\tilde{x}}. \tag{19}$$

Hence the double layer potential satisfies:

$$\left[(\gamma_0 \widetilde{W}\mu)\right]_{\mathcal{C}} := (\gamma_0^{\mathcal{S}_2}\widetilde{W}\mu) + (\gamma_0^{\mathcal{S}_1}\widetilde{W}\mu) = \mu, \tag{20}$$

where we tacitely assumed the orientation of the tangential vector $\mathbf{t}$ along $\mathcal{C}$ to be in accordance with the orientation of $\mathcal{S}_2$ in the sense of Stoke's theorem.

Proof. Given $\varepsilon > 0$, let $x \in \mathcal{S}_2$ with $\|x - x_0\| < \varepsilon$.We introduce the notation

$$\mathcal{C}_{\varepsilon,<} := \{\tilde{x} \in \mathcal{C}, |\tilde{x} - x_0| < \varepsilon\}, \qquad \mathcal{C}_{\varepsilon,\geq} := \{\tilde{x} \in \mathcal{C}, |\tilde{x} - x_0| \geq \varepsilon\}$$

Then,

$$(W\mu)(x) - (K_\varepsilon \mu)(x_0) = \int_{\mathcal{C}_{\varepsilon,\geq}} \mu(\tilde{x})\, [\underline{\mathrm{curl}}_{\mathcal{S}}\, U^*(x, \tilde{x}) - \underline{\mathrm{curl}}_{\mathcal{S}}\, U^*(x_0, \tilde{x})] \cdot \mathbf{t}(\tilde{x})\, ds_{\tilde{x}}$$
$$\int_{\mathcal{C}_{\varepsilon,<}} \mu(\tilde{x})\, \underline{\mathrm{curl}}_{\mathcal{S}}\, U^*(x, \tilde{x}) \cdot \mathbf{t}(\tilde{x})\, ds_{\tilde{x}}$$

and the first integral again vanishes as x approaches x_0, i.e.

$$\left| \int_{\mathcal{C}_{\varepsilon,\geq}} \mu(\tilde{x})\, [\underline{\mathrm{curl}}_{\mathcal{S}}\, U^*(x, \tilde{x}) - \underline{\mathrm{curl}}_{\mathcal{S}}\, U^*(x_0, \tilde{x})] \cdot \mathbf{t}(\tilde{x})\, ds_{\tilde{x}} \right| \overset{x \to x_0}{\longrightarrow} 0.$$

The second term can be rewritten as follows:

$$\int_{\mathcal{C}_{\varepsilon,<}} \mu(\tilde{x})\, \underline{\mathrm{curl}}_{\mathcal{S}}\, U^*(x, \tilde{x}) \cdot \mathbf{t}(\tilde{x})\, ds_{\tilde{x}}$$
$$= \int_{\mathcal{C}_{\varepsilon,<}} [\mu(\tilde{x}) - \mu(x_0)]\, \underline{\mathrm{curl}}_{\mathcal{S}}\, U^*(x, \tilde{x}) \cdot \mathbf{t}(\tilde{x})\, ds_{\tilde{x}}$$
$$+ \mu(x_0) \int_{\mathcal{C}_{\varepsilon,<}} \underline{\mathrm{curl}}_{\mathcal{S}}\, U^*(x, \tilde{x}) \cdot \mathbf{t}(\tilde{x})\, ds_{\tilde{x}}. \tag{21}$$

For the first integral on the right hand side of (21) we have the estimate

$$\begin{aligned}
&\left| \int_{\mathcal{C}_{\varepsilon,<}} [\mu(\tilde{x}) - \mu(x_0)] \, \underline{\mathrm{curl}}_{\mathcal{S}} \, U^*(x,\tilde{x}) \cdot \mathbf{t}(\tilde{x}) \, ds_{\tilde{x}} \right| \\
&\leq \sup_{\mathcal{C}_{\varepsilon,<}} |\mu(\tilde{x}) - \mu(x_0)| \int_{\mathcal{C}_{\varepsilon,<}} |\underline{\mathrm{curl}}_{\mathcal{S}} \, U^*(x,\tilde{x}) \cdot \mathbf{t}(\tilde{x})| \, ds_{\tilde{x}} \\
&\leq \qquad M \cdot \text{length } (\mathcal{C}_\varepsilon) \cdot \sup_{\mathcal{C}_{\varepsilon,<}} |\mu(\tilde{x}) - \mu(x_0)|
\end{aligned}$$

for some constant M, and hence the integral vanishes in the limit as $\varepsilon \longrightarrow 0$. For the second integral in (21), we define $\Omega_\varepsilon(x) := \{\tilde{x} \in \mathcal{S}_2 : \ |x - \tilde{x}| < \varepsilon\}$ to see

$$\begin{aligned}
\mu(x_0) \int_{\mathcal{C}_{\varepsilon,<}} \underline{\mathrm{curl}}_{\mathcal{S}} \, U^*(x,\tilde{x}) \cdot \mathbf{t}(\tilde{x}) \, ds_{\tilde{x}} &= \mu(x_0) \int_{\partial\Omega_\varepsilon(x_0)} \underline{\mathrm{curl}}_{\mathcal{S}} \, U^*(x,\tilde{x}) \cdot \mathbf{t}(\tilde{x}) \, ds_{\tilde{x}} \\
&\quad - \mu(x_0) \int_{\substack{\tilde{x} \in \mathcal{S}_2 \\ |\tilde{x} - x_0| = \varepsilon}} \underline{\mathrm{curl}}_{\mathcal{S}} \, U^*(x,\tilde{x}) \cdot \mathbf{t}(\tilde{x}) \, ds_{\tilde{x}}
\end{aligned}$$

Using the representation formula with $u \equiv 1$ we get

$$= \mu(x_0) \left(1 - \frac{1}{4\pi} \int_{\Omega_\varepsilon} d\sigma_{\tilde{x}} \right) - \mu(x_0) \int_{\substack{\tilde{x} \in \mathcal{S}_2 \\ |\tilde{x} - x_0| = \varepsilon}} \underline{\mathrm{curl}}_{\mathcal{S}} \, U^*(x,\tilde{x}) \cdot \mathbf{t}(\tilde{x}) \, ds_{\tilde{x}}$$

and without loss of generality we compute the remaining integral with respect to the northpole to find for all x_0:

$$\lim_{\varepsilon \to 0} \int_{\substack{\tilde{x} \in \mathcal{C} \\ |\tilde{x} - x_0| = \varepsilon}} \underline{\mathrm{curl}}_{\mathcal{S}} \, U^*(x,\tilde{x}) \cdot \mathbf{t}(\tilde{x}) \, ds_{\tilde{x}} = \frac{\alpha(x_0)}{2\pi}.$$

Putting the parts together we see that

$$\lim_{\varepsilon \to 0} \lim_{x \to x_0} \left((\widetilde{W}\mu)(x) - (K_\varepsilon \mu)(x_0) \right) = \left(1 - \frac{\alpha(x_0)}{2\pi} \right) \mu(x_0).$$

□

4 A BIE Strategy for Solving the Dirichlet Problem

With the single and double layer potentials defined as in the previous section, we are now in a position to reformulate the Dirichlet problem for the Laplace–Beltrami operator. For the purposes of this paper, we assume

sufficient smoothness of the data g and the curve $\mathcal{C}$ such that all the operators are well-defined; these assumptions can be relaxed, and the precise regularity and smoothness assumptions necessary are a subject of a forthcoming work. In what follows, however, we assume the curve $\mathcal{C}$ is at least C^2, and that the boundary data is smooth. We shall present a numerical example where we allow g to be Lipschitz.

Recall that we wish to find a smooth function u such that

$$-\triangle_{\mathcal{S}} u = 0 \text{ in } \mathcal{S}_2, \text{ and } u = g \text{ on } \mathcal{C}. \tag{22}$$

We seek a solution of this equation in terms of a layer ansatz. That is, we wish to find a density, σ or μ, so that either

$$u := \widetilde{V}\sigma \quad \text{or} \quad u := \widetilde{W}\mu \tag{23}$$

solves 22.

Lemma 3. *If the density σ solves the boundary integral equation*

$$V\sigma = g, \text{ on } \mathcal{C}, \text{ and also } \int_{\mathcal{C}} \sigma ds = 0, \tag{24}$$

then the function $u := \widetilde{V}\sigma$ solves (22).If the density μ solves

$$(\frac{1}{2}I + K)\mu = g, \qquad on \mathcal{C}, \tag{25}$$

then the solution of (22) is given by the double layer potential, $u := \widetilde{W}\mu$.

The proof is immediate from the previous section.

4.1 An Indirect Integral Equation Formulation

For concreteness, we describe in some detail a variational strategy to solve a boundary integral equation, whose solution then can be used to solve (22). We seek a solution u of the Laplace–Beltrami operator with prescribed boundary values on $\mathcal{C}$. The solution is assumed to be of the form

$$u = \widetilde{V}\sigma + p$$

where the density satisfies the integral equation

$$V\sigma + p = g \text{ on } \mathcal{C}$$

along with the constraint

$$\int_{\mathcal{C}} \sigma ds = 0.$$

We can write the weak formulation of this problem in saddle-point form as: *Find* $\sigma \in \mathcal{H}$ *and a multiplier* $p \in \mathbb{R}$ *such that*

$$\langle V\sigma, \chi\rangle + p\langle 1, \chi\rangle = \langle g, \chi\rangle, \tag{26a}$$

$$q\langle 1, \sigma\rangle = 0, \tag{26b}$$

for any test function $\chi \in \mathcal{H}$ and any real constant q. Under the present assumptions on smoothness, we set $\mathcal{H} = C(\mathcal{C})$, and $\langle \cdot, \cdot \rangle$ is simply the L^2- inner product along $\mathcal{C}$; we will describe the appropriate Sobolev spaces in which to naturally seek σ in subsequent work. The angle brackets will then represent the duality pairings in L^2.

The discretization strategy is now standard. Let τ_h be a partition of $\mathcal{C}$, with sub-interval size $h > 0$. We approximate $\mathcal{H}$ by a finite-dimensional space, S_h, which is parametrized by the meshsize h; as $h \to 0$, the approximation error $\inf_{v_h \in S_h} \|u - v_h\|_{\mathcal{H}} \to 0$ for all $u \in \mathcal{H}$. We then study the discrete Galerkin problem:
Find $\sigma_h \in S_h$, $p_h \in \mathbb{R}$, *such that for all* $(\chi_h, q_h) \in S_h \times \mathbb{R}$,

$$\langle V\sigma_h, \chi_h\rangle + p_h\langle 1, \chi_h\rangle ds = \langle g, \chi_h\rangle, \tag{27a}$$

$$q_h\langle 1, \sigma_h\rangle = 0, \tag{27b}$$

We shall provide an error analysis of this system in a subsequent paper, based on the correct choices of Sobolev spaces for the densities σ and approximation spaces S_h; at present, we present numerical experiments to validate the boundary element strategy.

4.2 Numerical Experiments

In what follows, we choose $\mathcal{C}$ to be the equator of the sphere, which is described by the latitude $\theta = \pi/2$. We solve the Laplace–Beltrami equation in the southern hemisphere $\mathcal{S}_2$. To do this we prescribe Dirichlet data g on $\mathcal{C}$ and solve the discrete Galerkin system (27). The partitions $\tau_h = \cup_{i=1}^{N} \Omega_i$ of $\mathcal{C}$ are chosen to consist of uniform sub-intervals of size h. The approximations are sought in the space of piece wise constant functions, i.e.

$$S_h := \{\chi \,|\, \chi(\varphi) = c_i \text{ for } \varphi \in \Omega_i\}.$$

In the specific case of the southern hemisphere, a Green's function for the problem is known and we can write down the solution in closed form as follows:

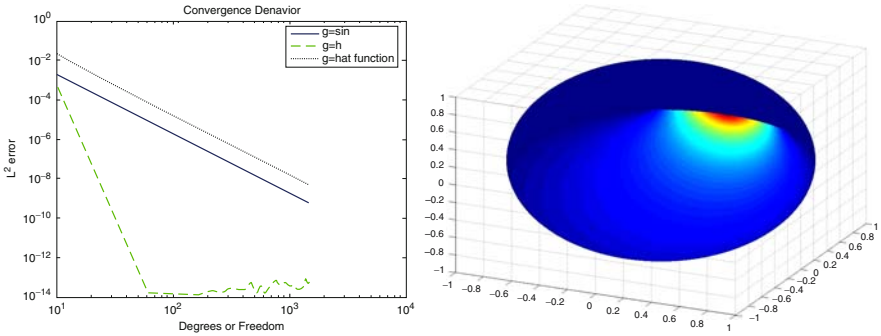
$$u(\varphi, \theta) = \frac{1}{4\pi} \int_0^{2\pi} \frac{\cos(\theta)}{-1 + \cos(\varphi - \varphi_0)\sin(\theta)}\, g(\varphi_0)\, d\varphi_0, \tag{28}$$

for $\theta > \frac{\pi}{2}$ and $\varphi \in [0, 2\pi]$. This expression serves as a reference for our computed solution.

We report the convergence behavior of the method in terms of the L^2 error of the computed solution along the latitude $\theta = 3/4\,\pi$. In Table 1, we report

Table 1. L^2 error of BEM solution, measured along $\theta = 3/4\,\pi$

DoF	Convergence for $g = \sin(\varphi)$		Convergence for $g = h(\varphi)$	
	L^2 error	Ratio	L^2 error	Ratio
20	2.41e−4	–	0.023	–
40	2.99e−5	8.06	2.57e−4	89.49
80	3.73e−6	8.02	3.06e−5	8.40
160	4.66e−7	8.00	3.73e−6	8.20
320	5.83e−8	7.99	4.61e−7	8.09
640	7.28e−9	8.01	4.73e−8	8.05



(a) (b)

Fig. 2. a Convergence behavior for different choices of g and **b** View on solution from below the south pole for hat-shaped data, $h(\varphi)$

this L^2-error versus the number of unknowns, for two choices of Dirichlet data: $g = \sin(\varphi)$, and $g = h(\varphi)$, where h is a hat-shaped function,

$$h(\varphi) := \begin{cases} 17\left(1 - \frac{6}{\pi}|\varphi|\right), & |\varphi| \leq \frac{\pi}{6} \\ 0 & \textit{otherwise.} \end{cases}$$

We note that halving the mesh-size reduces the L^2 error by a factor of 8 in both cases. Figure 2a shows the convergence behavior in terms of the above mentioned error versus the number of unknowns. In Fig. 2b, we see the actual solution of the boundary value problem, corresponding to the piece wise Dirichlet data $g = h(\varphi)$.

5 Conclusion

We have presented a boundary integral formulation, and associated Galerkin discretization strategy, for a boundary value problem for the Laplace–Beltrami

operator on the unit sphere in $\mathbb{R}^3$. Numerical experiments verify the applicability of the idea, and a rigorous error analysis will be presented in future work.

Acknowledgements

The authors wish to thank the organizers of the Abel Symposium 2006. NN was supported by NSERC and FQRNT.

References

1. D. Crowdy and M. Cloke: "Analytical solutions for distributed multipolar vortex equilibria on a sphere", Phys. Fluids, v. 15, n. 22, (2003).
2. D Crowdy: "Stuart vortices on a sphere", J. Fluid. Mech., v. 498, n. 381, (2004).
3. D. Crowdy: Point vortex motion on the surface of a sphere with impenetrable boundaries. Physics of Fluids, March 2006 Phys. Fluids 18, 036602 (2006).
4. R. Kidambi and P. K. Newton, "Point vortex motion on a sphere with solid boundaries", Physics of Fluids, 12, 581, 2000.
5. R. L. Duduchava, D. Mitrea, M. Mitrea: Differential operators and boundary value problems on hypersurfaces. Math. Nachr. 279 (2006) 996–1023.
6. W. J. Firey: The determination of convex bodies from their mean radius of curvature functions. Mathematika 14 (1967) 1–13.
7. J. L. Martinez–Morales: Generalized Legendre series and the fundamental solution of the Laplacian on the n–sphere. Analysis Mathematica 31 (2005) 131–150.
8. J.-C. Nédélec: Acoustic and Electromagnetic Equations. Integral Representations for Harmonic Problems. Springer, New York, 2001.

Numerical Study of Nearly Singular Solutions of the 3-D Incompressible Euler Equations

Thomas Y. Hou* and Ruo Li

Applied and Computational Mathematics, 217-50, Caltech, Pasadena, CA 91125, USA, hou@acm.caltech.edu, and LSEC, Academy of Mathematics and Systems Sciences, Chinese Academy of Sciences, Beijing 100080, China

Summary. In this paper, we perform a careful numerical study of nearly singular solutions of the 3D incompressible Euler equations with smooth initial data. We consider the interaction of two perturbed antiparallel vortex tubes which was previously investigated by Kerr in [16, 19]. In our numerical study, we use both the pseudo-spectral method with the 2/3 dealiasing rule and the pseudo-spectral method with a high order Fourier smoothing. Moreover, we perform a careful resolution study with grid points as large as $1{,}536 \times 1{,}024 \times 3{,}072$ to demonstrate the convergence of both numerical methods. Our computational results show that the maximum vorticity does not grow faster than doubly exponential in time while the velocity field remains bounded up to $T = 19$, beyond the singularity time $T = 18.7$ reported by Kerr in [16, 19]. The local geometric regularity of vortex lines near the region of maximum vorticity seems to play an important role in depleting the nonlinear vortex stretching dynamically.

1 Introduction

The question of whether the solution of the 3D incompressible Euler equations can develop a finite time singularity from a smooth initial condition is one of the most challenging problems. A major difficulty in obtaining the global regularity of the 3D Euler equations is due to the presence of the vortex stretching, which is formally quadratic in vorticity. There have been many computational efforts in searching for finite time singularities of the 3D Euler and Navier–Stokes equations, see e.g. [5, 23, 20, 12, 24, 16, 4, 2, 10, 22, 11, 19]. Of particular interest is the numerical study of the interaction of two perturbed antiparallel vortex tubes by Kerr [16, 19], in which a finite time blowup of the 3D Euler equations was reported. There has been a lot of interests in studying the interaction of two perturbed antiparallel vortex tubes in the late 1980s and early 1990s because of the vortex reconnection phenomena observed for the Navier–Stokes equations. While most studies indicated only exponential growth in the maximum vorticity [23, 1, 3, 20, 21, 24], the work of Kerr and

Hussain in [20] suggested a finite time blow-up in the infinite Reynolds number limit, which motivated Kerr's Euler computations mentioned above.

There has been some interesting development in the theoretical understanding of the 3D incompressible Euler equations. It has been shown that the local geometric regularity of vortex lines can play an important role in depleting nonlinear vortex stretching [6, 7, 8, 9]. In particular, the recent results obtained by Deng et al. [8, 9] show that geometric regularity of vortex lines, even in an extremely localized region containing the maximum vorticity, can lead to depletion of nonlinear vortex stretching, thus avoiding finite time singularity formation of the 3D Euler equations.

In a recent paper [13] (see also [14, 15]), we have performed well-resolved computations of the 3D incompressible Euler equations using the same initial condition as the one used by Kerr in [16]. In our computations, we use a pseudo-spectral method with a very high order Fourier smoothing to discretise the 3D incompressible Euler equations. The time integration is performed using the classical fourth order Runge–Kutta method with adaptive time stepping to satisfy the CFL stability condition. We use up to $1{,}536 \times 1{,}024 \times 3{,}072$ space resolution to resolve the nearly singular behavior of the 3D Euler equations. Our computational results demonstrate that the maximum vorticity does not grow faster than doubly exponential in time, up to $t = 19$, beyond the singularity time $t = 18.7$ predicted by Kerr's computations [16, 19]. Moreover, we show that the velocity field, the enstrophy, and enstrophy production rate remain bounded throughout the computations. This is in contrast to Kerr's computations in which the vorticity blows up like $O((T-t)^{-1})$ and the velocity field blows up like $O((T-t)^{-1/2})$. The vortex lines near the region of the maximum vorticity are found to be relatively smooth. With the velocity field being bounded, the non-blowup result of Deng et al. [8, 9] can be applied, which implies that there is no blowup of the Euler equations up to $T = 19$. The local geometric regularity of the vortex lines near the region of maximum vorticity seems to play an important role in the dynamic depletion of vortex stretching.

The purpose of this paper is to perform a systematic convergence study using two different numerical methods to further validate the computational results obtained in [13] (see also [14]). These two methods are the pseudo-spectral method with the 2/3 dealiasing rule and the pseudo-spectral method with a high order Fourier smoothing. For the 3D Euler equations with periodic boundary conditions, the pseudo-spectral method with the 2/3 dealiasing rule has been used widely in the computational fluid dynamics community. This method has the advantage of removing the aliasing errors completely. On the other hand, when the solution is nearly singular, the decay of the Fourier spectrum is very slow. The abrupt cut-off of the last 1/3 of its Fourier modes could generate significant oscillations due to the Gibbs phenomenon. In our computational study, we find that the pseudo-spectral method with a high order Fourier smoothing can alleviate this difficulty by applying a smooth cut-off at high frequency modes. Moreover, we find that by using a high order smoothing,

we can retain more effective Fourier modes than the 2/3 dealiasing rule. This gives a better convergence property. To demonstrate the convergence of both methods, we perform a careful resolution study, both in the physical space and spectrum space. Our extensive convergence study shows that both numerical methods converge to the same solution under mesh refinement. Moreover, we show that the pseudo-spectral method with a high order Fourier smoothing offers better accuracy than the pseudo-spectral method with the 2/3 dealiasing rule.

To understand the differences between our computational results and those obtained by Kerr in [16], we need to make some comparison between Kerr's computations [16] and our computations. In Kerr's computations, he used a pseudo-spectral discretization with the 2/3 dealiasing rule in the x and y directions, and a Chebyshev discretization in the z-direction with resolution of order $512 \times 256 \times 192$. In order to prepare the initial data that can be used for the Chebyshev polynomials, Kerr performed some interpolation and used extra filtering. As noted by Kerr [16] (see the top paragraph of page 1729), "An effect of the initial filter upon the vorticity contours at $t = 0$ is a long tail in Fig. 2a" (see also Fig. 2 of this paper). Such "a long tail" seems to be a numerical artifact. In comparison, since we use pseudo-spectral approximations in all three directions, there is no need to perform interpolation or use extra filtering as was done in [16]. Our initial vorticity contours are essentially symmetric (see Fig. 1). There is no such "a long tail" in our initial vorticity contours.

A more important difference between Kerr's computations and our computations is the difference between his numerical resolution and ours. From the numerical results presented at $t = 15$ and $t = 17$ in [16], one can observe noticeable oscillations in the vorticity contours (see Fig. 4 of [16] or Fig. 22 of this paper). By $t = 17$, the two vortex tubes have effectively turned into two thin vortex sheets which roll up at the left edge (see Figs. 24 and 25 of

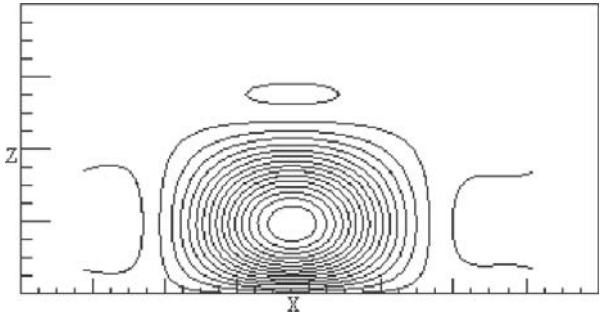


Fig. 1. The axial vorticity (the second component of vorticity) contours of the initial value on the symmetry plane. The vertical axis is the z-axis, and the horizontal axis is the x-axis

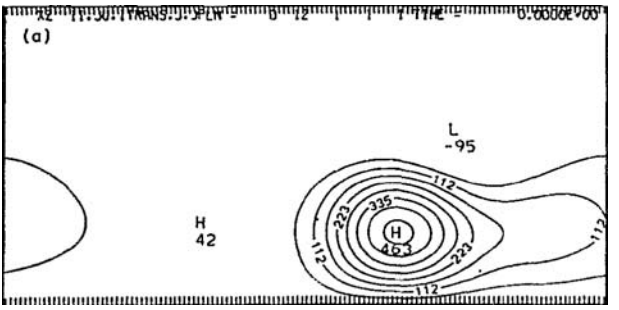


Fig. 2. Kerr's axial vorticity contours of the initial value on the symmetry plane. The vertical axis is the z-axis, and the horizontal axis is the x-axis. This is Fig. 2a of [16]

this paper). The rolled up portion of the vortex sheet travels backward in time and moves away from the dividing plane (the x–y plane). With only 192 Chebyshev grid points along the z-direction in Kerr's computations, there are not enough grid points to resolve the rolled up portion of the vortex sheet, which is some distance away from the dividing plane. The lack of resolution along the z-direction plus the Gibbs phenomenon due to the use of the 2/3 dealiasing rule in the x and y directions may contribute to the oscillations observed in Kerr's computations. In comparison, we have 3,072 grid points along the z-direction, which provide about 16 grid points across the singular layer at $t = 18$, and about eight grid points at $t = 19$ [13]. It is also worth mentioning that Kerr has only about 100 effective Fourier modes in the x and y directions (see Fig. 18 of [16]), while we have about 1,300 effective Fourier modes in $|k|$ (see Figs. 11 and 12 of this paper). The difference between our resolutions is quite significant.

It is worth noting that the computations for $t \leq 17$, which Kerr used as the primary evidence for a singularity, is still far from the predicted singularity time, $T = 18.7$. With the asymptotic scaling parameter being $T - t = 1.7$, the error in the singularity fitting could be of order one. In order to justify the predicted asymptotic behavior of vorticity and velocity blowup, one needs to perform well-resolved computations much closer to the predicted singularity time. As our computations demonstrate, the alleged singularity scaling, $\|\boldsymbol{\omega}\|_\infty \approx c/(T-t)$, does not persist in time (here $\boldsymbol{\omega}$ is vorticity). If we take $T = 18.7$, as suggested in [16], the scaling constant, c, does not remain constant as $t \to T$. In fact, we find that c rapidly decays to zero as $t \to T$ (see Fig. 20 of this paper).

The rest of this paper is organized as follows. We describe the set-up of the problem in Sect. 2. In Sect. 3, we perform a systematic convergence study of the two numerical methods. We describe our numerical results in detail and compare them with the previous results obtained in [16, 19] in Sect. 4. Some concluding remarks are made in Sect. 5.

2 The Set-Up of the Problem

The 3D incompressible Euler equations in the vorticity stream function formulation are given as follows:

$$\boldsymbol{\omega}_t + (\mathbf{u} \cdot \nabla)\boldsymbol{\omega} = \nabla \mathbf{u} \cdot \boldsymbol{\omega}, \tag{1}$$

$$-\triangle \boldsymbol{\psi} = \boldsymbol{\omega}, \quad \mathbf{u} = \nabla \times \boldsymbol{\psi}, \tag{2}$$

with initial condition $\boldsymbol{\omega} \mid_{t=0} = \boldsymbol{\omega}_0$, where $\mathbf{u}$ is velocity, $\boldsymbol{\omega}$ is vorticity, and $\boldsymbol{\psi}$ is stream function. Vorticity is related to velocity by $\boldsymbol{\omega} = \nabla \times \mathbf{u}$. The incompressibility implies that

$$\nabla \cdot \mathbf{u} = \nabla \cdot \boldsymbol{\omega} = \nabla \cdot \boldsymbol{\psi} = 0.$$

We consider periodic boundary conditions with period 4π in all three directions.

We study the interaction of two perturbed antiparallel vortex tubes using the same initial condition as that of Kerr (see Sect. III of [16]). Following [16], we call the x–y plane as the "dividing plane" and the x–z plane as the "symmetry plane". There is one vortex tube above and below the dividing plane respectively. The term "antiparallel" refers to the antisymmetry of the vorticity with respect to the dividing plane in the following sense: $\boldsymbol{\omega}(x, y, z) = -\boldsymbol{\omega}(x, y, -z)$. Moreover, with respect to the symmetry plane, the vorticity is symmetric in its y component and antisymmetric in its x and z components. Thus we have $\omega_x(x, y, z) = -\omega_x(x, -y, z)$, $\omega_y(x, y, z) = \omega_y(x, -y, z)$ and $\omega_z(x, y, z) = -\omega_z(x, -y, z)$. Here ω_x, ω_y, ω_z are the x, y, and z components of vorticity respectively. These symmetries allow us to compute only one quarter of the whole periodic cell.

A complete description of the initial condition is also given in [13]. There are a few misprints in the analytic expression of the initial condition given in [16]. In our computations, we use the corrected version of Kerr's initial condition by comparing with Kerr's Fortran subroutine which was kindly provided to us by him. A list of corrections to these misprints is given in the Appendix of [13].

We should point out that due to the difference between Kerr's discretization strategies and ours in solving the 3D Euler equations, there is some noticeable difference between the discrete initial condition generated by Kerr's discretization and the one generated by our pseudo-spectral discretization. In [16], Kerr interpolated the initial condition from the uniform grid to the Chebyshev grid along the z-direction and applied extra filtering. This interpolation and extra filtering, which were not provided explicitly in [16], seem to introduce some numerical artifact to Kerr's discrete initial condition. According to [16] (see the top paragraph of page 1729), "An effect of the initial filter upon the vorticity contours at $t = 0$ is a long tail in Fig. 2a". Since our computations are performed on a uniform grid using the pseudo-spectral approximations in all three directions, we do not need to use any interpolation

To demonstrate this slight difference between Kerr's discrete initial condition and ours, we plot the initial vorticity contours along the symmetry plane in Fig. 1 using our spectral discretization in all three directions. As we can see, the initial vorticity contours in Fig. 1 are essentially symmetric. This is in contrast to the apparent asymmetry in Kerr's initial vorticity contours as illustrated by Fig. 2, which is Fig. 2a of [16]. We also present the 3D plot of the vortex tubes at $t = 0$ and $t = 6$ respectively in Fig. 3. We can see that the two initial vortex tubes are essentially symmetric. By time $t = 6$, there is already a significant flattening near the center of the tubes.

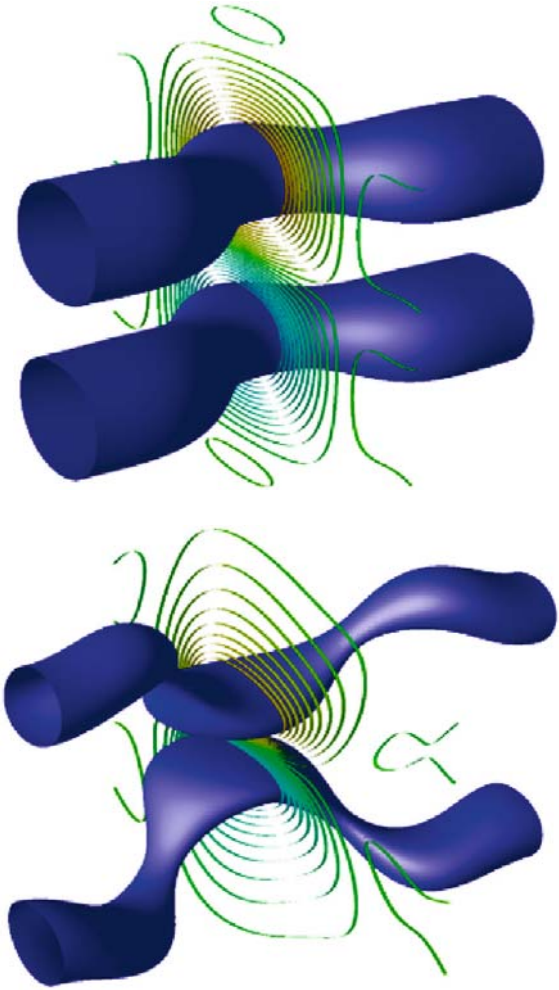


Fig. 3. The 3D view of the vortex tube for $t = 0$ and $t = 6$. The tube is the isosurface at 60% of the maximum vorticity. The ribbons on the symmetry plane are the contours at other different values

We exploit the symmetry properties of the solution in our computations, and perform our computations on only a quarter of the whole domain. Since the solution appears to be most singular in the z direction, we allocate twice as many grid points along the z direction than along the x direction. The solution is least singular in the y direction. We allocate the smallest resolution in the y direction to reduce the computation cost. In our computations, two typical ratios in the resolution along the x, y and z directions are $3:2:6$ and $4:3:8$. Our computations were carried out on the PC cluster LSSC-II in the Institute of Computational Mathematics and Scientific/Engineering Computing of Chinese Academy of Sciences and the Shenteng 6800 cluster in the Super Computing Center of Chinese Academy of Sciences. The maximal memory consumption in our computations is about 120 GB.

3 Convergence Study of the Two Numerical Methods

We use two numerical methods to compute the 3D Euler equations. The first method is the pseudo-spectral method with the 2/3 dealiasing rule. The second method is the pseudo-spectral method with a high order Fourier smoothing. The only difference between these two methods is in the way we perform the cut-off of the high frequency Fourier modes to control the aliasing error. If $\widehat{v}_k$ is the discrete Fourier transform of v, then we approximate the derivative of v along the x_j direction, v_{x_j}, by taking the discrete inverse Fourier transform of $ik_j\rho(2k_j/N_j)\widehat{v}_k$, where $k=(k_1,k_2,k_3)$ and ρ is a high frequency Fourier cut-off function. Here k_j is the wave number ($|k_j|\leqslant N_j/2$) along the x_j direction and N_j is the total number of grid points along the x_j direction. For the pseudo-spectral method with the 2/3 dealiasing rule, the cut-off function ρ is chosen such that $\rho(x)=1$ if $|x|\leq 2/3$, and $\rho(x)=0$ if $2/3<|x|\leq 1$. For the pseudo-spectral method with a high order smoothing, we choose the cut-off function ρ to be a smooth function of the form $\rho(x)\equiv\exp(-\alpha|x|^m)$ with $\alpha=36$ and $m=36$. The time integration is performed using the classical fourth order Runge–Kutta method. Adaptive time stepping is used to satisfy the CFL stability condition with CFL number equal to $\pi/4$. We use a sequence of resolutions: $768\times512\times1{,}536$, $1{,}024\times768\times2{,}048$, and $1{,}536\times1{,}024\times3{,}072$, to demonstrate the convergence of our numerical computations.

3.1 Comparison of the Two Methods

It is interesting to make some comparison of the two spectral methods we use. First of all, both methods are of spectral accuracy. The pseudo-spectral method with the 2/3 dealiasing rule has been widely used in the computational fluid dynamics community. It has the advantage of removing the aliasing error completely. On the other hand, when the solution is nearly singular, the Fourier spectrum typically decays very slowly. By cutting off the last 1/3 of the high frequency modes along each direction abruptly, this can introduce

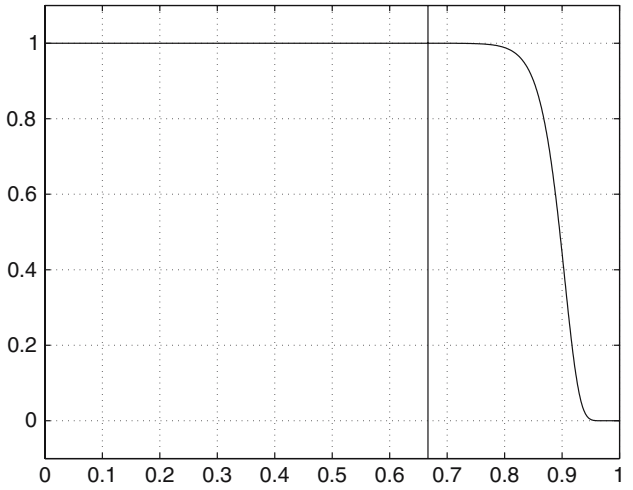


Fig. 4. The profile of the Fourier smoothing, $\exp(-36(x)^{36})$, as a function of x. The *vertical line* corresponds to the cut-off mode using the 2/3 dealiasing rule. We can see that using this Fourier smoothing we keep about 12–15% more modes than those using the 2/3 dealiasing rule

oscillations in the physical solution due to the Gibbs phenomenon. In this paper, we will provide solid numerical evidences to demonstrate this effect. On the other hand, the pseudo-spectral method with the high order Fourier smoothing is designed to keep the majority of the Fourier modes unchanged and remove the very high modes to avoid the aliasing error, see Fig. 4 for the profile of $\rho(x)$. We choose α to be 36 to guarantee that $\rho(2k_j/N_j)$ reaches the level of the round-off error ($O(10^{-16})$) at the highest modes. The order of smoothing, m, is chosen to be 36 to optimize the accuracy of the spectral approximation, while still keeping the aliasing error under control. As we can see from Fig. 4, the effective modes in our computations are about 12–15% more than those using the standard 2/3 dealiasing rule. Retaining part of the effective high frequency Fourier modes beyond the traditional 2/3 cut-off position is a special feature of the second method.

To compare the performance of the two methods, we perform a careful convergence study for the two methods. In Fig. 5, we compare the Fourier spectra of the enstrophy obtained by using the pseudo-spectral method with the 2/3 dealiasing rule with those obtained by the pseudo-spectral method with the high order smoothing. For a fixed resolution $768 \times 512 \times 1{,}536$, we can see that the Fourier spectra obtained by the pseudo-spectral method with the high order smoothing retains more effective Fourier modes than those obtained by the spectral method with the 2/3 dealiasing rule. This can be seen by comparing the results with the corresponding computations using a higher resolution $1{,}024 \times 768 \times 2{,}048$. Moreover, the pseudo-spectral method

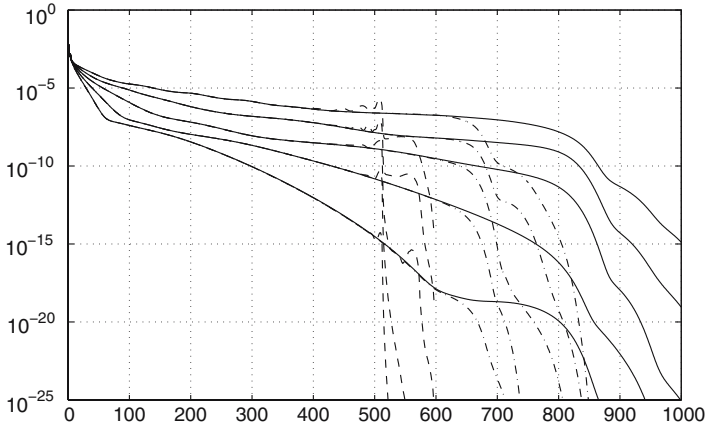


Fig. 5. The enstrophy spectra versus wave numbers. We compare the enstrophy spectra obtained using the high order Fourier smoothing method with those using the 2/3 dealiasing rule. The *dashed lines* and *dashed–dotted lines* are the enstrophy spectra with the resolution $768 \times 512 \times 1{,}536$ using the 2/3 dealiasing rule and the Fourier smoothing, respectively. The *solid lines* are the enstrophy spectra with resolution $1{,}024 \times 768 \times 2{,}048$ obtained using the high order Fourier smoothing. The times for the spectra lines are at $t = 15, 16, 17, 18, 19$ respectively

with the high order Fourier smoothing does not give the spurious oscillations in the Fourier spectra which are present in the computations using the 2/3 dealiasing rule near the 2/3 cut-off point.

We perform further comparison of the two methods using the same resolution. In Fig. 6, we plot the energy spectra computed by the two methods using resolution $768 \times 512 \times 1{,}536$. We can see that there is almost no difference in the Fourier spectra generated by the two methods in early times, $t = 8, 10$, when the solution is still relatively smooth. The difference begins to show near the cut-off point when the Fourier spectra raise above the round-off error level starting from $t = 12$. We can see that the spectra computed by the pseudo-spectral method with the 2/3 dealiasing rule introduces noticeable oscillations near the 2/3 cut-off point. The spectra computed by the pseudo-spectral method with the high order smoothing, on the other hand, extend smoothly beyond the 2/3 cut-off point. As we see from Fig. 5, a significant portion of those Fourier modes beyond the 2/3 cut-off position are still accurate. In the next subsection, we will demonstrate by a careful resolution study that the pseudo-spectral method with the high order smoothing indeed offers better accuracy than the pseudo-spectral method with the 2/3 dealiasing rule.

Similar comparison can be made in the physical space for the velocity field and the vorticity. In Fig. 7, we compare the maximum velocity as a function of time computed by the two methods using resolution $768 \times 512 \times 1{,}536$. The two solutions are almost indistinguishable. In Fig. 8, we plot the maximum vorticity as a function of time. The two solutions agree very well up to $t = 18$.

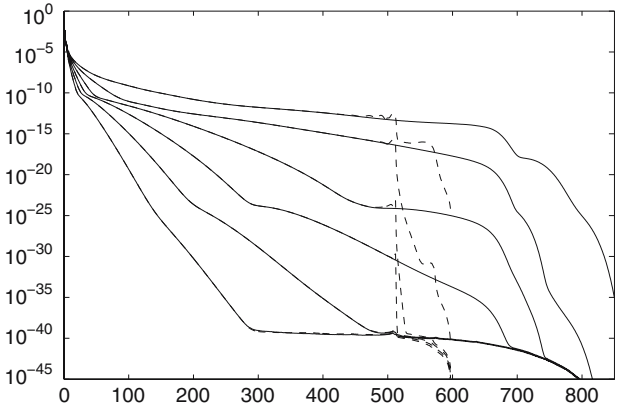


Fig. 6. The energy spectra versus wave numbers. We compare the energy spectra obtained using the high order Fourier smoothing method with those using the 2/3 dealiasing rule. The *dashed lines* and *solid lines* are the energy spectra with the resolution $768 \times 512 \times 1{,}536$ using the 2/3 dealiasing rule and the Fourier smoothing, respectively. The times for the spectra lines are at $t = 8, 10, 12, 14, 16, 18$ respectively

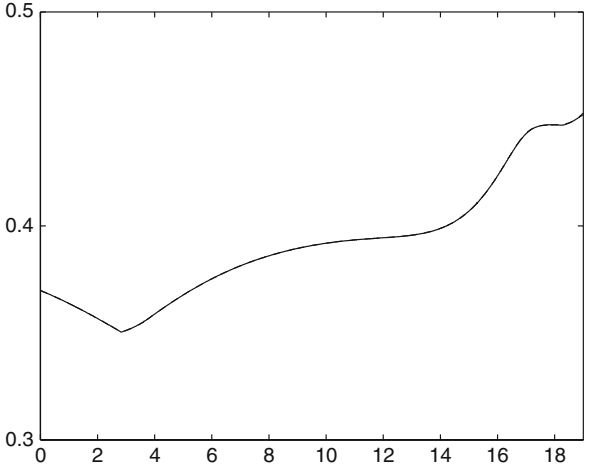


Fig. 7. Comparison of maximum velocity as a function of time computed by two methods. The *solid line* represents the solution obtained by the pseudo-spectral method with the high order smoothing, and the *dashed line* represents the solution obtained by the pseudo-spectral method with the 2/3 dealiasing rule. The resolution is $768 \times 512 \times 1{,}536$ for both methods

The solution obtained by the pseudo-spectral method with the 2/3 dealiasing rule grows slower from $t = 18$ to $t = 19$. To understand why the two solutions start to deviate from each other toward the end, we examine the contour plot of the axial vorticity in Figs. 9 and 10. As we can see, the vorticity computed

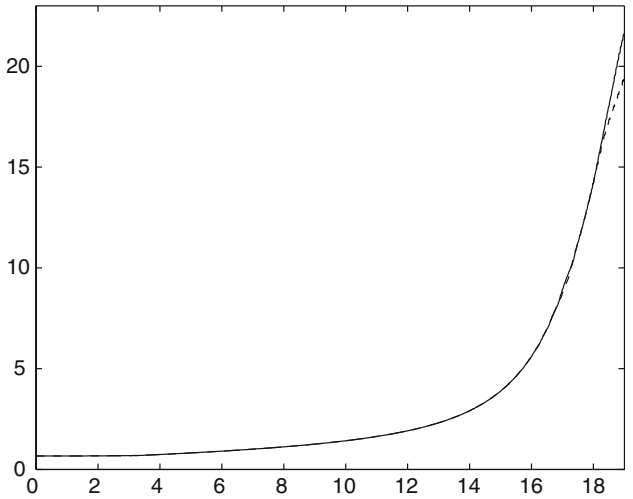


Fig. 8. Comparison of maximum vorticity as a function of time computed by two methods. The *solid line* represents the solution obtained by the pseudo-spectral method with the high order smoothing, and the *dashed line* represents the solution obtained by the pseudo-spectral method with the 2/3 dealiasing rule. The resolution is $768 \times 512 \times 1{,}536$ for both methods

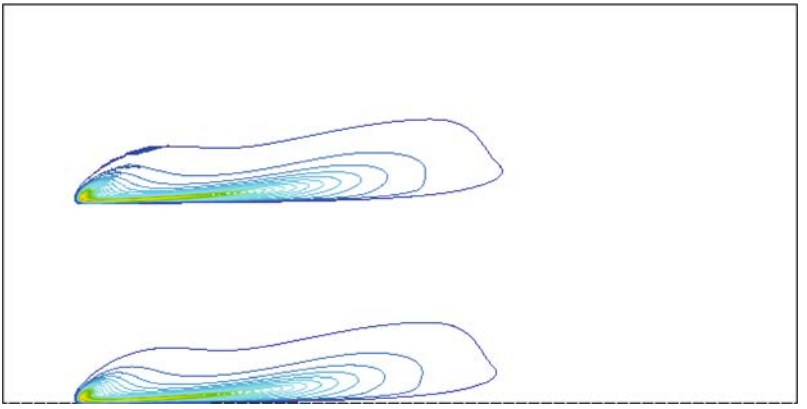


Fig. 9. Comparison of axial vorticity contours at $t = 17$ computed by two methods. The picture on the *top* is the solution obtained by the pseudo-spectral method with the 2/3 dealiasing rule, which is shifted by a distance of π in z direction, and the picture on the *bottom* is the solution obtained by the pseudo-spectral method with the high order smoothing. The resolution is $768 \times 512 \times 1{,}536$ for both methods. The box is the whole x–z computational domain $[-2\pi, 2\pi] \times [0, 2\pi]$

by the pseudo-spectral method with the 2/3 dealiasing rule already develops small oscillations at $t = 17$. The oscillations grow bigger by $t = 18$. We note that the oscillations in the axial vorticity contours concentrate near the

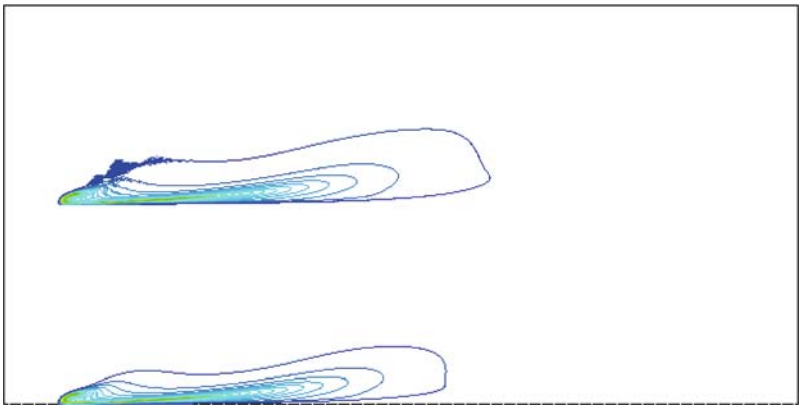


Fig. 10. Comparison of axial vorticity contours at $t = 18$ computed by two methods. This figure has the same layout as Fig. 9. The *top picture* uses the 2/3 dealiasing rule, while the *bottom picture* uses the high order smoothing. The resolution is $768 \times 512 \times 1{,}536$ for both methods

region where the magnitude of vorticity is close to zero. On the other hand, the solution computed by the spectral method with the high order smoothing is still quite smooth.

3.2 Resolution Study for the Two Methods

In this subsection, we perform a resolution study for the two numerical methods using a sequence of resolutions. For the pseudo-spectral method with the high order smoothing, we use the resolutions $768 \times 512 \times 1{,}536$, $1{,}024 \times 768 \times 2{,}048$, and $1{,}536 \times 1{,}024 \times 3{,}072$ respectively. Except for the computation on the largest resolution $1{,}536 \times 1{,}024 \times 3{,}072$, all computations are carried out from $t = 0$ to $t = 19$. The computation on the final resolution $1{,}536 \times 1{,}024 \times 3{,}072$ is started from $t = 10$ with the initial condition given by the computation with the resolution $1{,}024 \times 768 \times 2{,}048$. For the pseudo-spectral method with the 2/3 dealiasing rule, we use the resolutions $512 \times 384 \times 1{,}024$, $768 \times 512 \times 1{,}536$ and $1{,}024 \times 768 \times 2{,}048$ respectively. The computations using the first two resolutions are carried out from $t = 0$ to $t = 19$ while the computation on the largest resolution $1{,}024 \times 768 \times 2{,}048$ is started at $t = 15$ with the initial condition given by the computation with resolution $512 \times 512 \times 1{,}024$.

First, we perform a convergence study of the enstrophy and energy spectra for the pseudo-spectral method with the high order smoothing at later times (from $t = 16$ to $t = 19$) using two largest resolutions $1{,}024 \times 768 \times 2{,}048$, and $1{,}536 \times 1{,}024 \times 3{,}072$. The results are given in Figs. 11 and 12 respectively. They clearly demonstrate the spectral convergence of the spectral method with the high order smoothing.

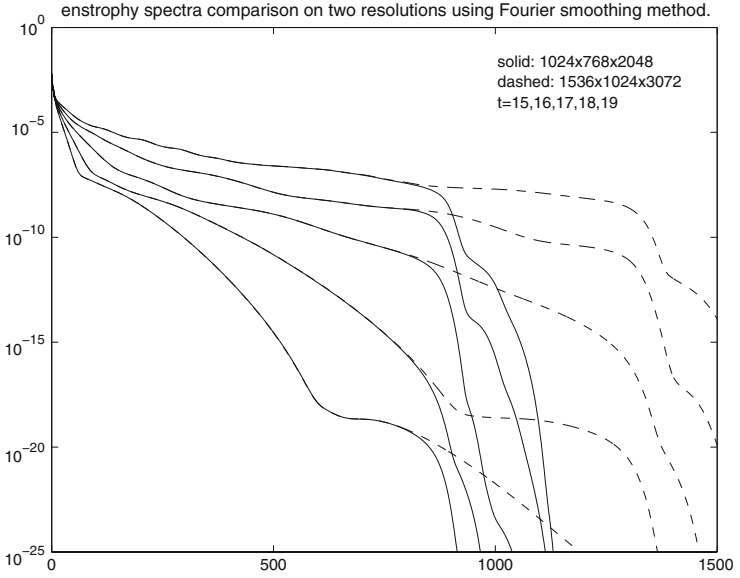


Fig. 11. Convergence study for enstrophy spectra obtained by the pseudo-spectral method with high order smoothing using different resolutions. The *dashed lines* and the *solid lines* are the enstrophy spectra on resolution $1{,}536 \times 1{,}024 \times 3{,}072$ and $1{,}024 \times 768 \times 2{,}048$, respectively. The times for the lines from *bottom* to *top* are $t = 15, 16, 17, 18, 19$

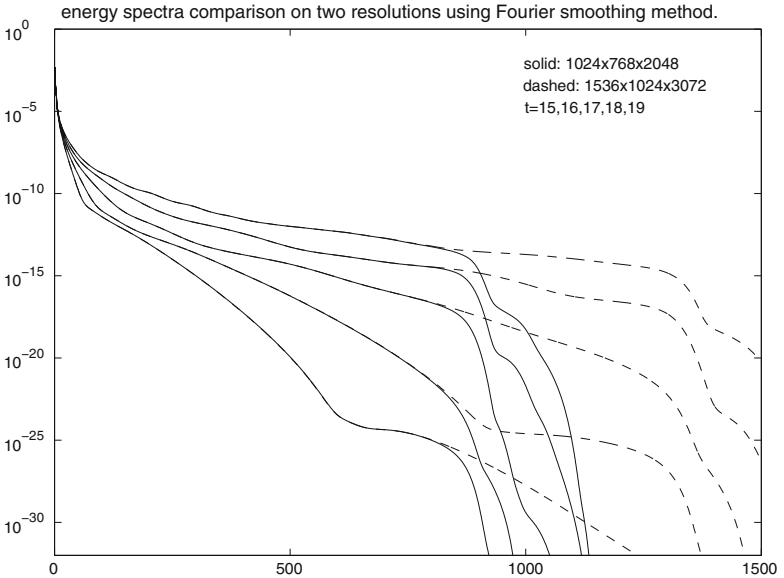


Fig. 12. Convergence study for energy spectra obtained by the pseudo-spectral method with high order smoothing using different resolutions. The *dashed lines* and the *solid lines* are the energy spectra on resolution $1{,}536 \times 1{,}024 \times 3{,}072$ and $1{,}024 \times 768 \times 2{,}048$, respectively. The times for the lines from *bottom* to *top* are $t = 15, 16, 17, 18, 19$

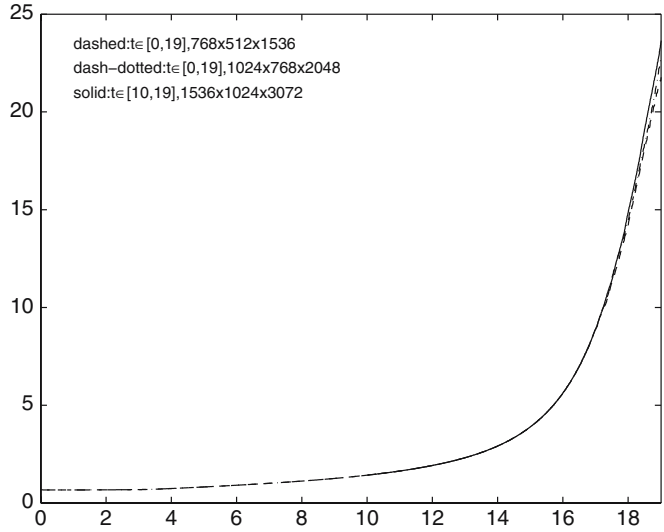


Fig. 13. The maximum vorticity $\|\boldsymbol{\omega}\|_\infty$ in time computed by the pseudo-spectral method with high order smoothing using different resolutions

To further demonstrate the accuracy of our computations, we compare the maximum vorticity obtained by the pseudo-spectral method with the high order smoothing for three different resolutions: $768\times512\times1{,}536$, $1{,}024\times768\times2{,}048$, and $1{,}536\times1{,}024\times3{,}072$ respectively. The result is plotted in Fig. 13. Two conclusions can be made from this resolution study. First, by comparing Fig. 13 with Fig. 8, we can see that the pseudo-spectral method with the high order smoothing is indeed more accurate than the pseudo-spectral method with the 2/3 dealiasing rule for a given resolution. Secondly, the resolution $768\times512\times1{,}536$ is not good enough to resolve the nearly singular solution at later times. However, we observe that the difference of the numerical solution obtained by the resolution $1{,}024\times768\times2{,}048$ is very close to that obtained by the resolution $1{,}536\times1{,}024\times3{,}072$. This indicates that the vorticity is reasonably well-resolved by our largest resolution $1{,}536\times1{,}024\times3{,}072$.

We have also performed a similar resolution study for the maximum velocity in Fig. 14. The solutions obtained by the two largest resolutions are almost indistinguishable, which suggests that the velocity is well-resolved by our largest resolution $1{,}536\times1{,}024\times3{,}072$.

Next, we perform a similar resolution study for the pseudo-spectral method with the 2/3 dealiasing rule. The results are very similar to the ones we have obtained for the pseudo-spectral method with the high order smoothing. Here we just present a few representative results. In Fig. 15, we plot the enstrophy spectra for a sequence of times from $t = 8$ to $t = 18$ using different resolutions. The resolutions we use here are $512\times384\times1{,}024$, $786\times512\times1{,}536$, and $1{,}024\times768\times2{,}048$ respectively. If we compare the Fourier spectra at $t = 18$

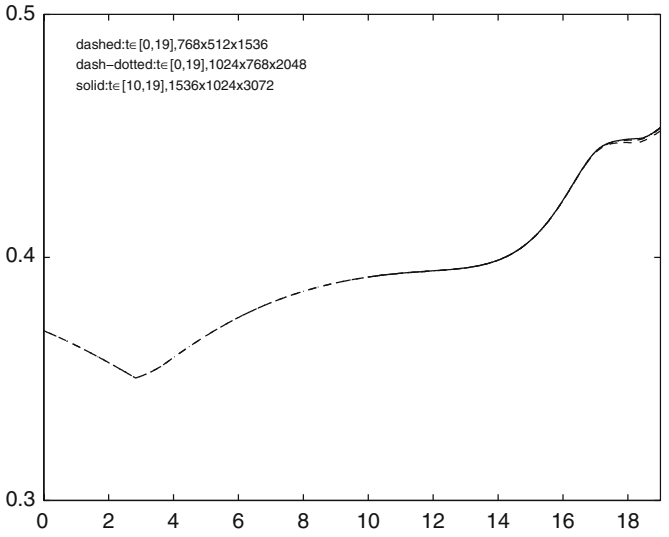


Fig. 14. Maximum velocity $\|\mathbf{u}\|_\infty$ in time computed by the pseudo-spectral method with high order smoothing using different resolutions

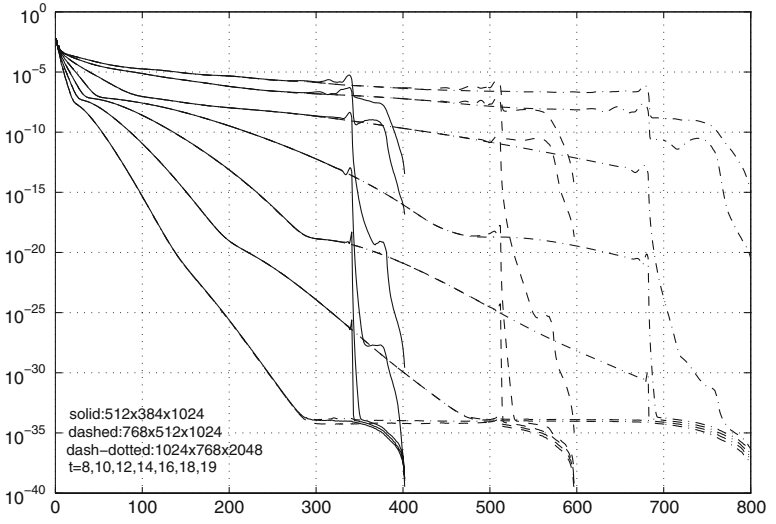


Fig. 15. Convergence study for enstrophy spectra obtained by the pseudo-spectral method with the 2/3 dealiasing rule using different resolutions. The *solid line* is computed with resolution $512 \times 384 \times 1{,}024$, the *dashed line* is computed with resolution $786 \times 512 \times 1{,}536$, and the *dashed–dotted line* is computed with resolution $1{,}024 \times 768 \times 2{,}048$. The times for the lines from *bottom* to *top* are $t = 8, 10, 12, 14, 16, 18, 19$

and $t = 19$ (the last two curves in Fig. 15), we clearly observe convergence of the enstrophy spectra as we increase our resolutions. On the other hand, the decay of the enstrophy spectra becomes very slow at later times. The oscillations near the 2/3 cut-off point become more and more pronounced as time increases. This abrupt cut-off of high frequency spectra introduces some oscillations in the vorticity contours at later times.

To demonstrate that the two numerical methods converge to the same solution when the solution is nearly singular, we compare the enstrophy spectra computed by the two numerical methods at later times using the largest resolutions that we can afford. For the pseudo-spectral method with the high order smoothing, we use resolution 1,536 × 1,024 × 3,072. For the pseudo-spectral method with the 2/3 dealiasing rule, we use resolution 1,024 × 768 × 2,048. In Fig. 16, we plot the enstrophy spectra for $t = 15, 16, 17, 18, 19$, respectively. We observe that the two methods give excellent agreement for those Fourier modes that are not affected by the high frequency cut-off. This shows that the two numerical methods converge to the same solution with spectral accuracy.

We have performed a similar convergence study for the pseudo-spectral method with the 2/3 dealiasing in the physical space for the maximum vorticity. The result is given in Fig. 17. As we can see, the computation with a higher resolution gives faster growth in the maximum vorticity. This is also

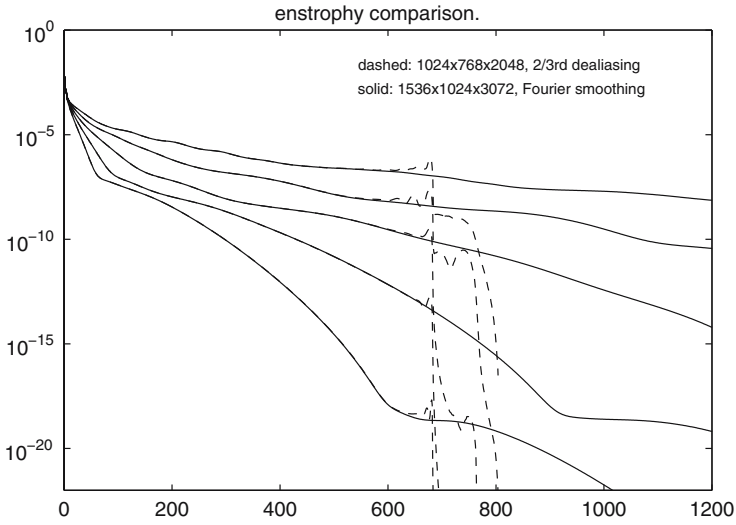


Fig. 16. The enstrophy spectra versus wave numbers. We compare the enstrophy spectra obtained using the high order Fourier smoothing method with those using the 2/3 dealiasing rule. The *dashed lines* are the enstrophy spectra using the 2/3 dealiasing rule with resolution $1{,}024 \times 768 \times 2{,}048$, and the *solid lines* are the spectra with resolution $1{,}536 \times 1{,}024 \times 3{,}072$ using the Fourier smoothing. The times for the spectra lines are at $t = 15, 16, 17, 18, 19$ respectively

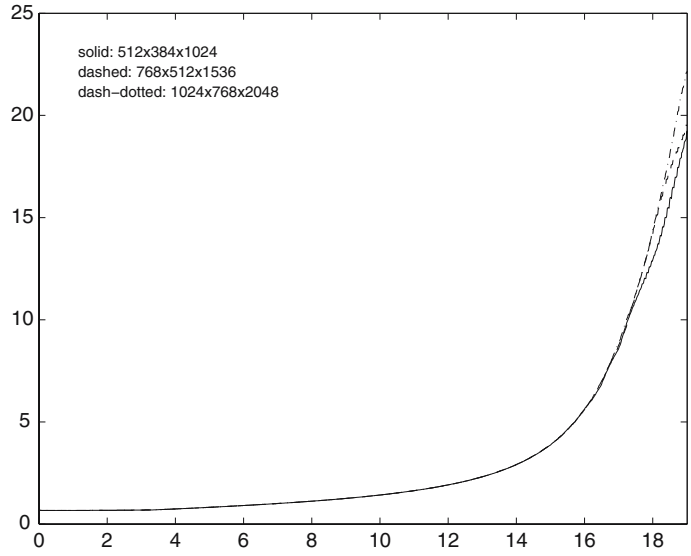


Fig. 17. The maximum vorticity $\|\boldsymbol{\omega}\|_\infty$ in time computed by the pseudo-spectral method with the 2/3 dealiasing rule using different resolutions

what we observed earlier for the pseudo-spectral method with the high order smoothing. As we will see in the next section, the maximum vorticity grows almost like doubly exponential in time. To capture this rapid dynamic growth of maximum vorticity, we must have sufficient resolution to resolve the nearly singular solution of the Euler equations at later times.

The resolution study given by Fig. 17 suggests that the maximum vorticity is reasonably resolved by resolution $768 \times 512 \times 1{,}536$ before $t = 18$. It is interesting to note that at $t = 17$, small oscillations have already appeared in the vorticity contours in the region where the magnitude of vorticity is small, see Fig. 9. Apparently, the small oscillations in the region where the vorticity is close to zero in magnitude have not yet polluted the accuracy of the maximum vorticity in a significant way. Note that there is no oscillation developed in the vorticity contours obtained by the pseudo-spectral method with the high order smoothing at this time. From Fig. 8, we know that the maximum vorticity computed by the two methods agrees reasonably well with each other before $t = 18$. This shows that the two methods can still approximate the maximum vorticity reasonably well with resolution $768 \times 512 \times 1{,}536$ before $t = 18$.

The resolution study given by Fig. 17 also suggests that the computation obtained by the pseudo-spectral method with the 2/3 dealiasing rule using resolution $768 \times 512 \times 1{,}536$ is significantly under-resolved after $t = 18$. This is also supported by the appearance of the relatively large oscillations in the vorticity contours at $t = 18$ from Fig. 10. It is interesting to note from Fig. 8 that the computational results obtained by the two methods with resolution $768 \times 512 \times 1{,}536$ begin to deviate from each other precisely around $t = 18$.

By comparing the result from Fig. 8 with that from Fig. 17, we confirm again that for a given resolution, the pseudo-spectral method with the high order smoothing gives a more accurate approximation than the pseudo-spectral method with the 2/3 dealiasing rule.

4 Analysis of Computational Results

In this section, we will present a series of numerical results to reveal the nature of the nearly singular solution of the 3D Euler equations, and compare our results with those obtained by Kerr in [16, 19]. Based on the convergence study we have performed in the previous section, we will present only those numerical results which are computed by the pseudo-spectral method with the high order smoothing using the largest resolution $1{,}536 \times 1{,}024 \times 3{,}072$.

4.1 Review of Kerr's Results

In [16], Kerr presented numerical evidence which suggested a finite time singularity of the 3D Euler equations for two perturbed antiparallel vortex tubes. He used a pseudo-spectral discretization in the x and y directions, and a Chebyshev method in the z direction with resolution of order $512 \times 256 \times 192$. His computations showed that the growth of the peak vorticity, the peak axial strain, and the enstrophy production obey $(T-t)^{-1}$ with $T = 18.9$. Kerr stated in his paper [16] (see page 1727) that his numerical results shown after $t = 17$ and up to $t = 18$ were "not part of the primary evidence for a singularity" due to the lack of sufficient numerical resolution and the presence of noise in the numerical solutions. In his recent paper [19] (see also [17, 18]), Kerr applied a high wave number filter to the data obtained in his original computations to "remove the noise that masked the structures in earlier graphics" presented in [16]. With this filtered solution, he presented some scaling analysis of the numerical solutions up to $t = 17.5$. The velocity field was shown to blow up like $O((T-t)^{-1/2})$ with T being revised to $T = 18.7$.

4.2 Maximum Vorticity Growth

From the resolution study we present in Fig. 13, we find that the maximum vorticity increases rapidly from the initial value of 0.669–23.46 at the final time $t = 19$, a factor of 35 increase from its initial value. Kerr's computations predicted a finite time singularity at $T = 18.7$. Our computations show no sign of finite time blowup of the 3D Euler equations up to $T = 19$, beyond the singularity time predicted by Kerr. We use three different resolutions, i.e. $768 \times 512 \times 1{,}536$, $1{,}024 \times 768 \times 2{,}048$, and $1{,}536 \times 1{,}024 \times 3{,}072$ respectively in our computations. As we can see, the agreement between the two successive resolutions is very good with only mild disagreement toward the end of

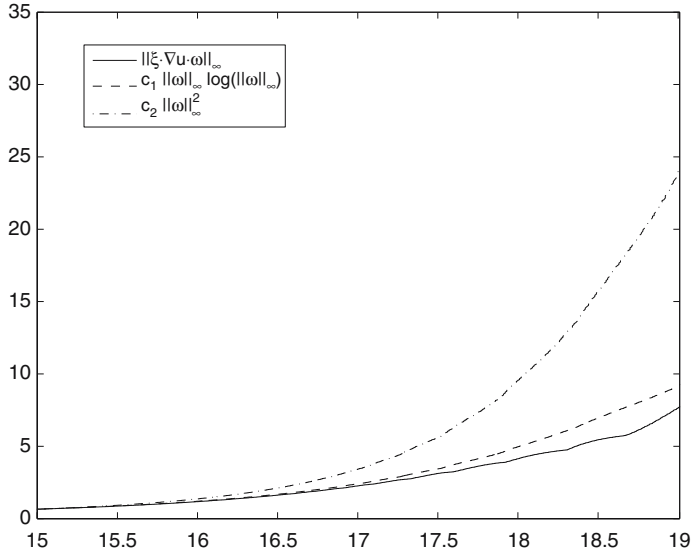


Fig. 18. Study of the vortex stretching term in time, resolution $1{,}536\times1{,}024\times3{,}072$. We take $c_1 = 1/8.128$, $c_2 = 1/23.24$ to match the same starting value for all three plots

the computations. This indicates that a very high space resolution is indeed needed to capture the rapid growth of maximum vorticity at the later stage of the computations.

In order to understand the nature of the dynamic growth in vorticity, we examine the degree of nonlinearity in the vortex stretching term. In Fig. 18, we plot the quantity, $\|\xi \cdot \nabla \mathbf{u} \cdot \boldsymbol{\omega}\|_\infty$, as a function of time, where ξ is the unit vorticity vector. If the maximum vorticity indeed blew up like $O((T-t)^{-1})$, as alleged in [16], this quantity should have been quadratic as a function of maximum vorticity. We find that there is tremendous cancellation in this vortex stretching term. It actually grows slower than $C\|\boldsymbol{\omega}\|_\infty \log(\|\boldsymbol{\omega}\|_\infty)$, see Fig. 18. It is easy to show that such weak nonlinearity in vortex stretching would imply only doubly exponential growth in the maximum vorticity. Indeed, as demonstrated by Fig. 19, the maximum vorticity does not grow faster than doubly exponential in time. In fact, the growth slows down toward the end of the computation, which indicates that there is stronger cancellation taking place in the vortex stretching term.

We remark that for vorticity that grows as rapidly as doubly exponential in time, one may be tempted to fit the maximum vorticity growth by $c/(T-t)$ for some T. Indeed, if we choose $T = 18.7$ as suggested by Kerr in [19], we find a reasonably good fit for the maximum vorticity as a function of $c/(T-t)$ for the period $15 \le t \le 17$. We plot the scaling constant c in Fig. 20. As we can see, c is close to a constant for $15 \le t \le 17$. To conclude that the 3D Euler equations indeed develop a finite time singularity, one must demonstrate that

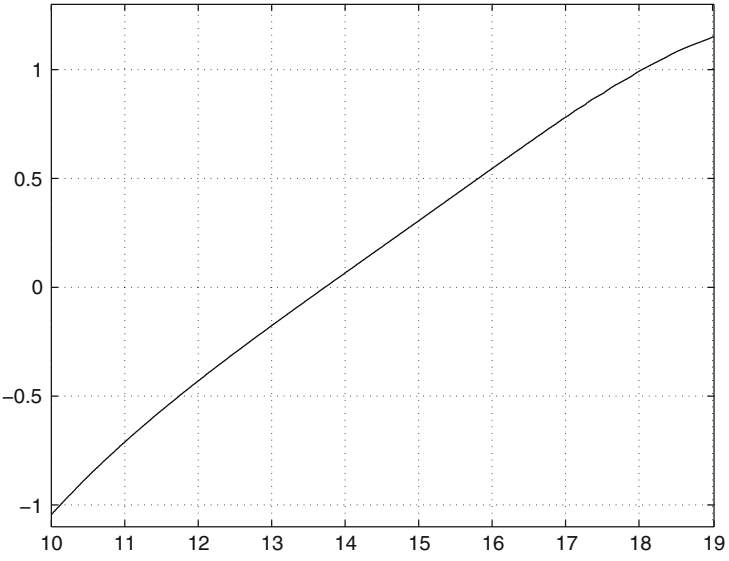


Fig. 19. The plot of $\log\log\|\omega\|_\infty$ vs time, resolution $1{,}536 \times 1{,}024 \times 3{,}072$

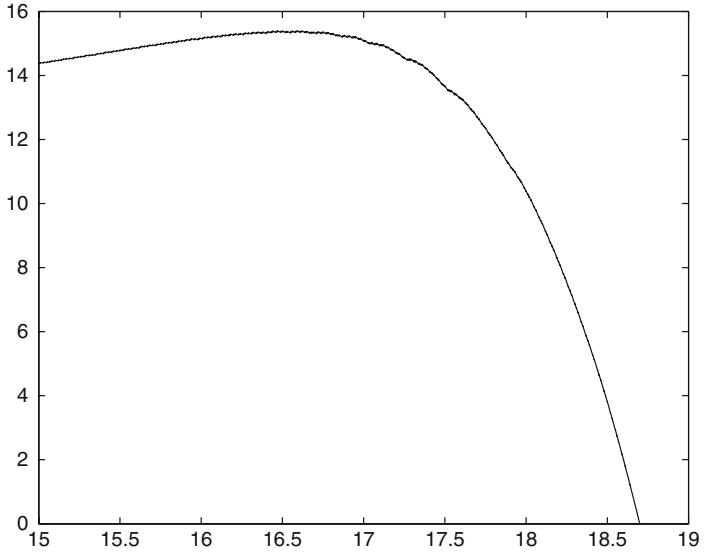


Fig. 20. Scaling constant in time for the fitting $\|\omega\|_\infty \approx c/(T-t)$, $T = 18.7$

such scaling persists as t approaches to T. As we can see from Fig. 20, the scaling constant c decreases rapidly to zero as t approaches to the alleged singularity time T. Therefore, the fitting of $\|\boldsymbol{\omega}\|_\infty \approx O((T-t)^{-1})$ is not correct asymptotically.

4.3 Velocity Profile

One of the important findings of our computations is that the velocity field is actually bounded by 1/2 up to $T = 19$. This is in contrast to Kerr's computations in which the maximum velocity was shown to blow up like $O((T-t)^{-1/2})$ [17, 19]. We plot the maximum velocity as a function of time using different resolutions in Fig. 14. The computation obtained by resolution $1{,}024 \times 768 \times 2{,}048$ and the one obtained by resolution $1{,}536 \times 1{,}024 \times 3{,}072$ are almost indistinguishable. The fact that the velocity field is bounded is significant. With the velocity field being bounded, the non-blowup theory of Deng et al. [8] can be applied, which implies non-blowup of the 3D Euler equations up to T. We refer to [13, 14] for more discussions.

4.4 Local Vorticity Structure

In this subsection, we would like to examine the local vorticity structure near the region of the maximum vorticity. To illustrate the development in the symmetry plane, we show a series of vorticity contours near the region of the maximum vorticity at late times in a manner similar to the results presented in [16]. For some reason, Kerr scaled his axial vorticity contours by a factor of 5 along the z-direction. Noticeable oscillations already develop in Kerr's axial vorticity contours at $t = 15$ and $t = 17$, see Fig. 22. To compare with Kerr's results, we scale the vorticity contours in the x–z plane by a factor of 5 in the z-direction. The results at $t = 15$ and $t = 17$ are plotted in Fig. 21. The results are in qualitative agreement with Kerr's results, except that our computations are better resolved and do not suffer from the noise and oscillations which are present in Kerr's vorticity contours.

In order to see better the dynamic development of the local vortex structure, we plot a sequence of vorticity contours on the symmetry plane at $t = 17.5, 18, 18.5$, and 19 respectively in Fig. 23. The pictures are plotted using the original length scales, without the scaling by a factor of 5 in the z direction as in Fig. 21. From these results, we can see that the vortex sheet is compressed in the z direction. It is clear that a thin layer (or a vortex sheet) is formed dynamically. The head of the vortex sheet begins to roll up around $t = 16$. Here the head of the vortex sheet refers to the region extending above the vorticity peak just behind the leading edge of the vortex sheet [16]. By the time $t = 19$, the head of the vortex sheet has traveled backward for quite a distance and away from the dividing plane. The vortex sheet has been compressed quite strongly along the z-direction. In order to resolve this nearly singular layer structure, we use 3,072 grid points along the z-direction, which gives about 16 grid points across the layer at $t = 18$ and about eight grid points across the layer at $t = 19$. In comparison, the 192 Chebyshev grid points along the z-direction in Kerr's computations would not be sufficient to resolve the rolled-up portion of the vortex sheet.

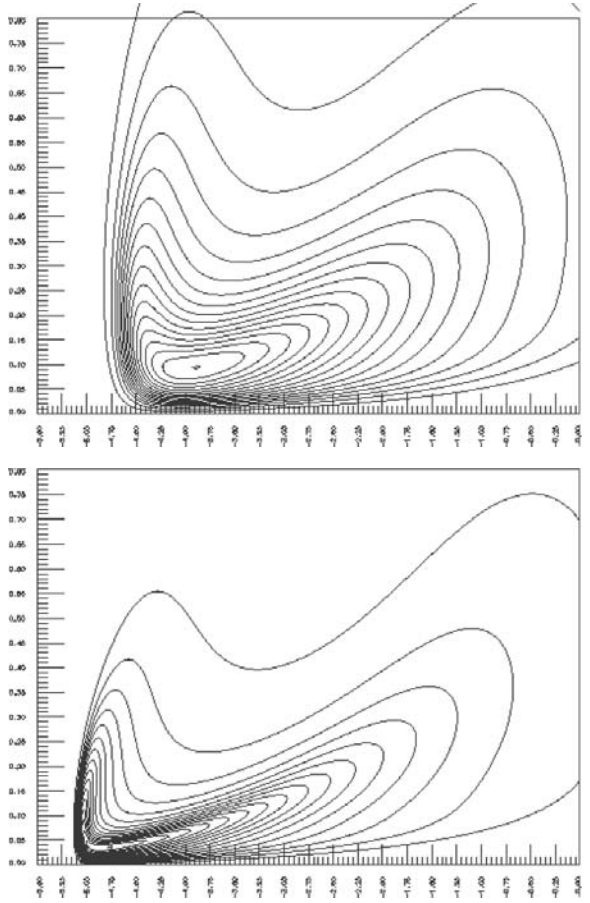


Fig. 21. The contour of axial vorticity around the maximum vorticity on the symmetry plane at $t = 15$ (*on the top*) and $t = 17$ (*on the bottom*). The vertical axis is the z-axis, and the horizontal axis is the x-axis. The figure is scaled in z direction by a factor of 5 to compare with Fig. 4 in [16]

We also plot the isosurface of vorticity near the region of the maximum vorticity in Figs. 24 and 25 to illustrate the dynamic roll-up of the vortex sheet near the region of the maximum vorticity. The isosurface of vorticity in Fig. 24 is set at $0.6 \times \|\boldsymbol{\omega}\|_\infty$. Figure 24 gives the local vorticity structure at $t = 17$. If we scale the local roll-up region on the left hand side next to the box by a factor of 4 along the z direction, as was done in [19], we would obtain a local roll-up structure which is qualitatively similar to Fig. 1 in [19]. In Fig. 25, we show the local vorticity structure for $t = 18$ and $t = 19$. In both figures, the isosurface is set at $0.5 \times \|\boldsymbol{\omega}\|_\infty$. We can see that the vortex sheets have rolled up and traveled backward in time away from the dividing plane. Moreover, we observe that the vortex lines near the region of maximum vorticity are

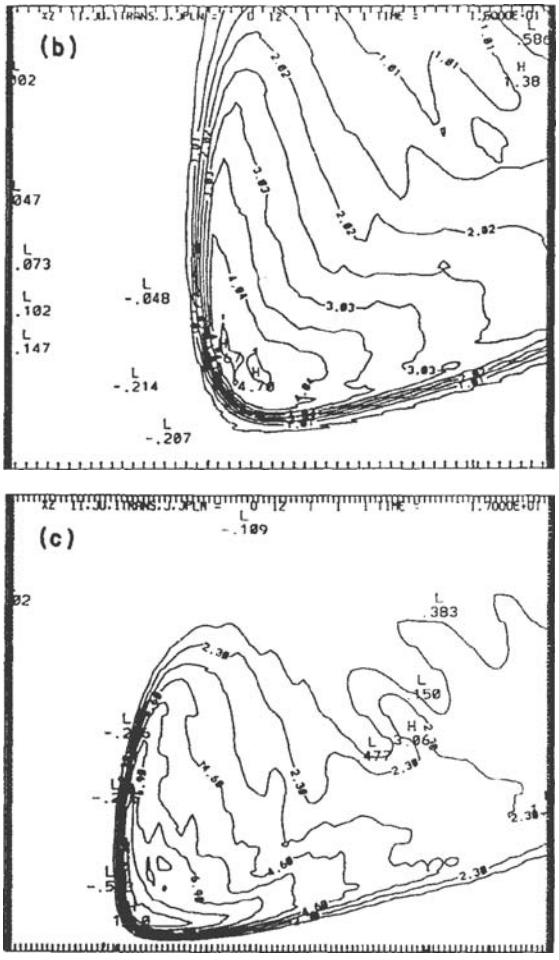


Fig. 22. Kerr's axial vorticity contours on the symmetry plane at $t = 15$ (*on the top*) and $t = 17$ (*on the bottom*). These are from Fig. 4 in [16]

relatively straight and the unit vorticity vectors seem to be quite regular. On the other hand, the inner region containing the maximum vorticity does not seem to shrink to zero at a rate of $(T-t)^{1/2}$, as predicted in [19]. The length and the width of the vortex sheet are still $O(1)$, although the thickness of the vortex sheet becomes quite small.

Another interesting question is how the vorticity vector aligns with the eigenvectors of the deformation tensor, which is defined as $M \equiv \frac{1}{2}(\nabla \mathbf{u} + \nabla^T \mathbf{u})$. In Table 1, we document the alignment information of the vorticity vector around the point of maximum vorticity with resolution $1{,}536 \times 1{,}024 \times 3{,}072$. In this table, λ_i $(i = 1, 2, 3)$ is the i-th eigenvalue of M, θ_i is the angle between

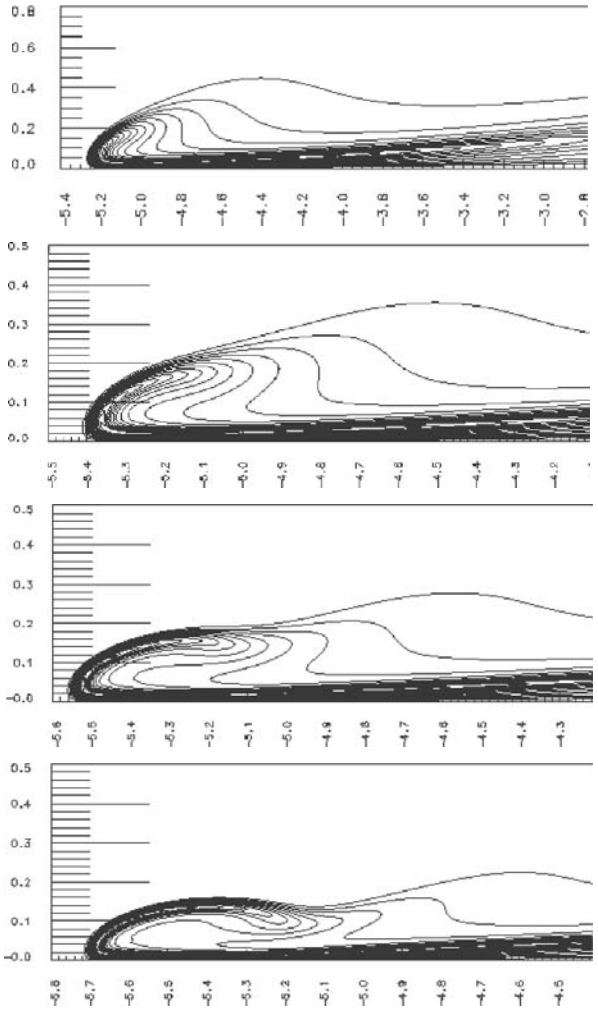


Fig. 23. The contour of axial vorticity around the maximum vorticity on the symmetry plane (the x–z plane) at $t = 17.5, 18, 18.5, 19$

the i-th eigenvector of M and the vorticity vector. One can see clearly that for $16 \leq t \leq 19$ the vorticity vector at the point of maximum vorticity is almost perfectly aligned with the second eigenvector of M. The angle between the vorticity vector and the second eigenvector is very small throughout this time interval. Note that the second eigenvalue, λ_2, is positive and is about 20 times smaller in magnitude than the largest and the smallest eigenvalues. Moreover, we observe that the magnitude of the second eigenvalue does not change much in time. This dynamic alignment of the vorticity vector with the second eigenvector of the deformation tensor is another indication that there is a dynamic depletion of vortex stretching.

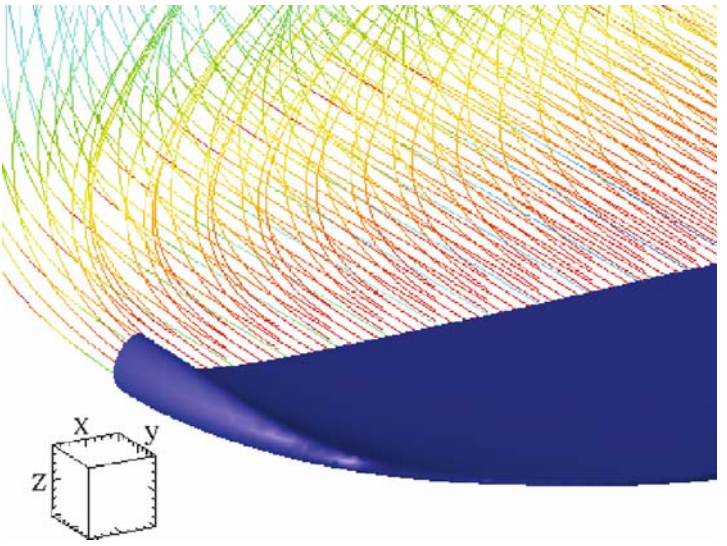


Fig. 24. The local 3D vortex structure and vortex lines around the maximum vorticity at $t = 17$. The size of the box on the *left* is 0.075^3 to demonstrate the scale of the picture. The isosurface is set at $0.6 \times \|\boldsymbol{\omega}\|_\infty$

5 Concluding Remarks

We investigate the interaction of two perturbed vortex tubes for the 3D Euler equations using Kerr's initial condition [16]. We use both the pseudo-spectral method with the standard 2/3 dealiasing rule and the pseudo-spectral method with a 36th order Fourier smoothing. We perform a careful resolution study to demonstrate the convergence of both methods. Our numerical computations demonstrate that while both methods converge to the same solution under resolution study, the pseudo-spectral method with the 36th order Fourier smoothing offers better computational accuracy for a given resolution. Moreover, we find that the pseudo-spectral method with the 36th order Fourier smoothing is more effective in reducing the numerical oscillations due to the Gibbs phenomenon while still keeping the aliasing error under control.

Our numerical study indicates that there is a very subtle dynamic depletion of vortex stretching. The maximum vorticity is shown to grow no faster than doubly exponential in time up to $T = 19$, beyond the singularity time predicted by Kerr in [16]. The velocity field is shown to be bounded throughout the computations. Vortex lines near the region of the maximum vorticity are quite regular. We provide numerical evidence that the vortex stretching term is only weakly nonlinear and is bounded by $\|\boldsymbol{\omega}\|_\infty \log(\|\boldsymbol{\omega}\|_\infty)$. This implies that there is tremendous dynamic cancellation in the nonlinear vortex stretching term. With the velocity field being bounded and the vortex lines being regular near the region of the maximum vorticity, the non-blowup

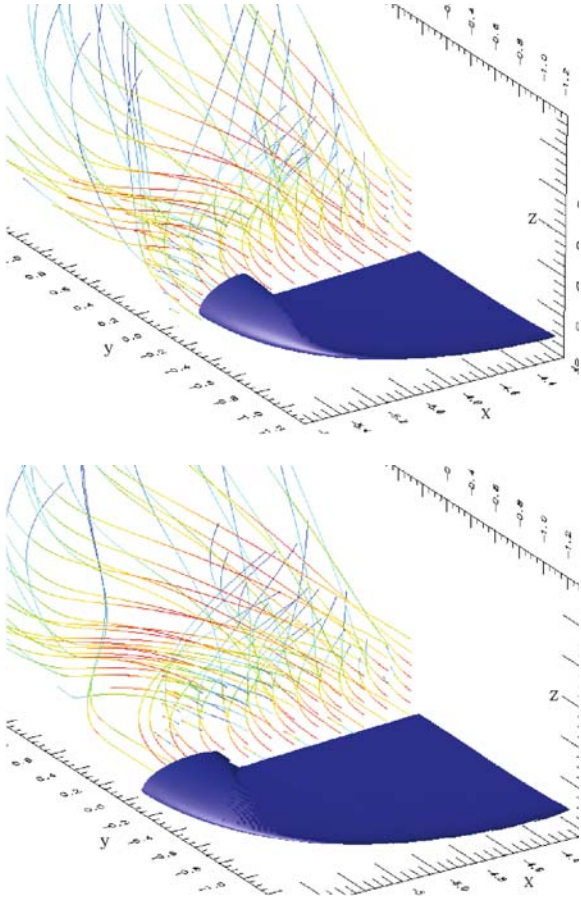


Fig. 25. The local 3D vortex structures and vortex lines around the maximum vorticity at $t = 18$ (*on the top*) and $t = 19$ (*on the bottom*). The isosurface is set at $0.5 \times \|\boldsymbol{\omega}\|_\infty$

conditions of Deng et al. [8] are satisfied. This provides a theoretical support for our computational results and sheds some light to our understanding of the dynamic depletion of vortex stretching.

Finally, we would like to mention that we have carried out a convergence study of the two numerical methods we consider in this paper for the one-dimensional Burgers equation. The Burgers equation shares some essential numerical difficulties with the 3D Euler equations that we consider here. It has the same type of quadratic nonlinearity in the convection term and it is known that it can form a shock singularity in a finite time. An important advantage of the Burgers equation is that we have an analytic solution formula which can be solved numerically up to the machine precision by using the Newton iterative method. Using this semi-analytical solution, we have computed the solution

Table 1. The alignment of the vorticity vector and the eigenvectors of M around the point of maximum vorticity with resolution $1{,}536 \times 1{,}024 \times 3{,}072$

Time	$\|\omega\|$	λ_1	θ_1	λ_2	θ_2	λ_3	θ_3
16.012295	5.628002	−1.508771	89.992936	0.206199	0.007159	1.302352	89.998852
16.515890	7.016002	−1.864394	89.995940	0.232299	0.010438	1.631355	89.990387
17.013589	8.910001	−2.322629	89.998141	0.254699	0.006815	2.066909	89.993445
17.515769	11.430017	−2.630440	89.969954	0.224305	0.085053	2.415185	89.920433
18.011609	14.890004	−3.625738	89.969613	0.257302	0.036607	3.378515	89.979590
18.516346	19.130010	−4.501348	89.966725	0.246305	0.036617	4.274913	89.984720
19.014394	23.590012	−5.477438	89.966055	0.247906	0.034472	5.258292	89.994005

Here, λ_i $(i = 1, 2, 3)$ is the i-th eigenvalue of M, θ_i is the angle between the i-th eigenvector of M and the vorticity vector. One can see that the vorticity vector is aligned very well with the second eigenvector of M

very close to the shock singularity time and documented the errors of both numerical methods using very large resolutions. The computational results we obtain on the Burgers equation completely support the convergence study of the two numerical methods for the 3D Euler equations that we present in this paper. The performance of these two numerical methods and their convergence property for the 1D Burgers equation are basically the same as those for the 3D Euler equations. More details of this study can be found in [14].

Acknowledgments

We would like to thank Prof. Lin-Bo Zhang from the Institute of Computational Mathematics in Chinese Academy of Sciences for providing us with the computing resource to perform this large scale computational project. Additional computing resource was provided by the Center of Super Computing Center of Chinese Academy of Sciences. We also thank Prof. Robert Kerr for providing us with his Fortran subroutine that generates his initial data. This work was in part supported by NSF under the NSF FRG grant DMS-0353838 and DMS-0713670. Part of this work was done while Hou visited the Academy of Systems and Mathematical Sciences of CAS in the summer of 2005 as a member of the Oversea Outstanding Research Team for Complex Systems.

References

1. C. Anderson and C. Greengard, *The vortex ring merger problem at infinite reynolds number*, Comm. Pure Appl. Maths **42** (1989), 1123.
2. O. N. Boratav and R. B. Pelz, *Direct numerical simulation of transition to turbulence from a high-symmetry initial condition*, Phys. Fluids **6** (1994), no. 8, 2757–2784.

3. O. N. Boratav, R. B. Pelz, and N. J. Zabusky, *Reconnection in orthogonally interacting vortex tubes: Direct numerical simulations and quantifications*, Phys. Fluids A **4** (1992), no. 3, 581–605.
4. R. Caflisch, *Singularity formation for complex solutions of the 3D incompressible Euler equations*, Physica D **67** (1993), 1–18.
5. A. Chorin, *The evolution of a turbulent vortex*, Commun. Math. Phys. **83** (1982), 517.
6. P. Constantin, *Geometric statistics in turbulence*, SIAM Review **36** (1994), 73.
7. P. Constantin, C. Fefferman, and A. Majda, *Geometric constraints on potentially singular solutions for the 3-D Euler equation*, Commun. in PDEs. **21** (1996), 559–571.
8. J. Deng, T. Y. Hou, and X. Yu, *Geometric properties and non-blowup of 3-D incompressible Euler flow*, Comm. in PDEs. **30** (2005), no. 1, 225–243.
9. _____, *Improved geometric conditions for non-blowup of 3D incompressible Euler equation*, Comm. in PDEs. **31** (2006), no. 2, 293–306.
10. V. M. Fernandez, N. J. Zabusky, and V. M. Gryanik, *Vortex intensification and collapse of the Lissajous-Elliptic ring: Single and multi-filament Biot-Savart simulations and visiometrics*, J. Fluid Mech. **299** (1995), 289–331.
11. R. Grauer, C. Marliani, and K. Germaschewski, *Adaptive mesh refinement for singular solutions of the incompressible Euler equations*, Phys. Rev. Lett. **80** (1998), 19.
12. R. Grauer and T. Sideris, *Numerical computation of three dimensional incompressible ideal fluids with swirl*, Phys. Rev. Lett. **67** (1991), 3511.
13. T. Y. Hou and R. Li, *Dynamic depletion of vortex stretching and non-blowup of the 3-D incompressible Euler equations*, J. Nonlinear Science. **16** (2006), no. 6, 639–664.
14. _____, *Computing nearly singular solutions using pseudo-spectral methods*, J. Comput. Phys. **226** (2007), 379–397.
15. _____, *Blowup or no blowup? The interplay between theorey and numerics*, Physica D (2008), **237** published online on Jan 25, 2008, DOI:10.1016/j.physd.2008.01.018, Vol. 237, 1937–1944.
16. R. M. Kerr, *Evidence for a singularity of the three dimensional, incompressible Euler equations*, Phys. Fluids **5** (1993), no. 7, 1725–1746.
17. _____, *Euler singularities and turbulence*, 19th ICTAM Kyoto '96 (T. Tatsumi, E. Watanabe, and T. Kambe, eds.), Elsevier Science, 1997, pp. 57–70.
18. _____, *The outer regions in singular Euler*, Fundamental problematic issues in turbulence (Birkhäuser) (Tsinober and Gyr, eds.), 1999.
19. _____, *Velocity and scaling of collapsing Euler vortices*, Phys. Fluids **17** (2005), 075103–114.
20. R. M. Kerr and F. Hussain, *Simulation of vortex reconnection*, Physica D **37** (1989), 474.
21. M. V. Melander and F. Hussain, *Cross linking of two antiparallel vortex tubes*, Phys. Fluids A (1989), 633–636.
22. R. B. Pelz, *Locally self-similar, finite-time collapse in a high-symmetry vortex filament model*, Phys. Rev. E **55** (1997), no. 2, 1617–1626.
23. A. Pumir and E. E. Siggia, *Collapsing solutions to the 3-D Euler equations*, Phys. Fluids A **2** (1990), 220–241.
24. M. J. Shelley, D. I. Meiron, and S. A. Orszag, *Dynamical aspects of vortex reconnection of perturbed anti-parallel vortex tubes*, J. Fluid Mech. **246** (1993), 613–652.

Energy-Preserving and Stable Approximations for the Two-Dimensional Shallow Water Equations

Eitan Tadmor* and Weigang Zhong

Department of Mathematics, Center for Scientific Computation and Mathematical Modeling (CSCAMM) and Institute for Physical Science and Technology (IPST), University of Maryland, MD 20742, USA, tadmor@cscamm.umd.edu

Summary. We present a systematic development of energy-stable approximations of the two-dimensional shallow water (SW) equations, which are based on the general framework of *entropy conservative* schemes introduced in [Tad03, TZ06]. No artificial numerical viscosity is involved: stability is dictated solely by eddy viscosity. In particular, in the absence of any dissipative mechanism, the resulting numerical schemes *precisely* preserve the total energy, which serves as an entropy function for the SW equations. We demonstrate the dispersive nature of such entropy conservative schemes with a series of scalar examples, interesting for their own sake. We then turn to the SW equations. Numerical experiments of the partial-dam-break problem with energy-preserving and energy stable schemes, successfully simulate the propagation of circular shock and the vortices formed on the both sides of the breach.

1 Introduction

Consider a three-dimensional domain in which the homogenous fluid flows with a free-surface under the influence of gravity. One of the widely used approaches for the description of such unsteady free-surface flows is that of shallow water. Under the shallow-water approximation that refers to the fact that a horizontal scale is in excess of the depth of the fluid, the 3D Navier–Stokes equations can be simplified to the shallow water equations with the depth-averaged continuity equation and momentum equations. Neglecting diffusion of momentum due to wind effects and Coriolis terms, we consider two-dimensional shallow water (SW) equations in the conservative form for free-surface compressible flow with flat frictionless bottom on two dimensional x–y plane,

$$\frac{\partial}{\partial t}\begin{bmatrix} h \\ uh \\ vh \end{bmatrix} + \frac{\partial}{\partial x}\begin{bmatrix} uh \\ u^2h + gh^2/2 \\ uvh \end{bmatrix} + \frac{\partial}{\partial y}\begin{bmatrix} vh \\ uvh \\ v^2h + gh^2/2 \end{bmatrix} = \zeta\frac{\partial}{\partial x}\begin{bmatrix} 0 \\ u_xh \\ v_xh \end{bmatrix} + \zeta\frac{\partial}{\partial y}\begin{bmatrix} 0 \\ u_yh \\ v_yh \end{bmatrix}. \tag{1.1}$$

Here, $h = h(x, y, t)$ is the total water depth which plays the role of density, and $(u(x, y, t), v(x, y, t))$ are the depth-averaged velocity components along x and y direction. The three equations express, respectively, conservation laws of mass and momentum in x and y direction for the shallow water flow, driven by convective fluxes on the LHS together with eddy viscous fluxes on the RHS. These fluxes involve the constant gravity acceleration $g > 0$, and $\zeta > 0$ is the eddy viscosity. By ignoring the small scale vortices in the motion, we calculate a large-scale flow motion with eddy viscosity ζ that characterizes the transport and dissipation of energy into the smaller scales of the flow.

If we turn off the eddy viscosity ($\zeta = 0$), system (1.1) is reduced to the inviscid shallow water equations,

$$\frac{\partial}{\partial t}\begin{bmatrix} h \\ uh \\ vh \end{bmatrix} + \frac{\partial}{\partial x}\begin{bmatrix} uh \\ u^2h + gh^2/2 \\ uvh \end{bmatrix} + \frac{\partial}{\partial y}\begin{bmatrix} vh \\ uvh \\ v^2h + gh^2/2 \end{bmatrix} = 0. \tag{1.2}$$

The SW equations (1.1) constitute an incompletely parabolic system, whose solutions can exhibit discontinuities associated with hydraulic jumps and bores in flows or the propagation of sharp fronts. In this paper, we are concerned with construction of energy-stable numerical methods for simulating two dimensional flows, in which initial discontinuities associated with partial-dam-break need to be evolved in time. The conservation of the total energy, $E = (gh^2 + u^2h + v^2h)/2$, guarantees that such numerical simulations of shallow water flows are nonlinearly stable and free of *artificial* numerical viscosity, which may dramatically change the profiles of the solutions in long time integration. In our computation, conservation of the total energy is enforced by utilizing *entropy conservative* fluxes which are tailored to preserve the energy, being an entropy function for the SW equations. The resulting numerical scheme is energy-stable, free of artificial numerical viscosity in the sense that energy dissipation is driven solely by the eddy viscous fluxes. In the particular case that eddy viscosity is absent, $\zeta = 0$, our scheme *precisely* preserves the total energy E.

A general framework for the construction of entropy-conservative schemes for 1D nonlinear conservation laws is introduced in Sect. 3, following [Tad03, TZ06]. We then test these entropy-conservative schemes for 1D Burgers' equation being the prototype of scalar nonlinear conservation laws in Sect. 4. In Sect. 5, we generalize the recipe for the entropy-stable approximations of two dimensional shallow water equations with the energy playing the role of entropy. The extension is carried out dimension by dimension. The algorithm along each dimension follows the same recipe outlined in the one-dimensional setup. The key ingredient behind these schemes is the construction of energy-preserving numerical fluxes. Our main results on the 2D shallow water equations are summarized in Theorem 5.1. To illustrate the performance of the new schemes, we test a two-dimensional partial-dam-break problem in Sect. 6.

The numerical results, especially those of the fine meshes, successfully simulate both the circular shock water wave propagations and the vortices formed on both sides of the breach.

2 Entropy Dissipation: The General Framework

2.1 Entropy Variables

We consider a two-dimensional hyperbolic system,

$$\frac{\partial}{\partial t}\mathbf{u} + \frac{\partial}{\partial x}\mathbf{f}(\mathbf{u}) + \frac{\partial}{\partial y}\mathbf{g}(\mathbf{u}) = 0. \tag{2.1}$$

We assume that it obeys an additional conservation law where a convex entropy function $U(\mathbf{u})$ is balanced by entropy fluxes $F(\mathbf{u})$ and $G(\mathbf{u})$,

$$\frac{\partial}{\partial t}U(\mathbf{u}) + \frac{\partial}{\partial x}F(\mathbf{u}) + \frac{\partial}{\partial y}G(\mathbf{u}) = 0. \tag{2.2}$$

Note that (2.2) holds if the entropy function $U(\mathbf{u})$ is linked to the entropy fluxes $F(\mathbf{u})$ and $G(\mathbf{u})$ through the compatibility relations,

$$U_{\mathbf{u}}^{\top}\mathbf{f}_{\mathbf{u}} = F_{\mathbf{u}}^{\top}, \quad U_{\mathbf{u}}^{\top}\mathbf{g}_{\mathbf{u}} = G_{\mathbf{u}}^{\top}. \tag{2.3}$$

In fact, multiplying (2.1) by $U_{\mathbf{u}}^{\top}$ on the left, one recovers the equivalence between (2.1) and (2.3) for all classical solutions $\mathbf{u}$'s of (2.1). These formal manipulations are valid only under the smooth region. To justify these steps in the presence of shock discontinuities, the conservation laws (2.1) are realized as appropriate vanishing viscosity limits, $\mathbf{u} = \lim_{\zeta\downarrow 0}\mathbf{u}^{\zeta}$, where $\mathbf{u}^{\zeta}$ is governed by the (possibly incompletely) parabolic system

$$\frac{\partial}{\partial t}\mathbf{u}^{\zeta} + \frac{\partial}{\partial x}\mathbf{f}(\mathbf{u}^{\zeta}) + \frac{\partial}{\partial y}\mathbf{g}(\mathbf{u}^{\zeta}) = \zeta\frac{\partial}{\partial x}\left(Q\frac{\partial}{\partial x}\mathbf{u}^{\zeta}\right) + \zeta\frac{\partial}{\partial y}\left(Q\frac{\partial}{\partial y}\mathbf{u}^{\zeta}\right), \quad \zeta \geq 0. \tag{2.4}$$

Here, $\zeta \downarrow 0$ stands for the vanishing viscosity amplitude such as the eddy viscosity coefficient in the SW equations (1.1)), and $Q = Q(\mathbf{u})$ is any *admissible* viscosity coefficient which is H-symmetric positive-definite,

$$QH = (QH)^{\top} \geq 0, \qquad H := (U_{\mathbf{uu}})^{-1}. \tag{2.5}$$

The passage from vanishing viscosity limits to weak entropy solutions of (2.1) is classical, [Lax73], and we refer to the more comprehensive recent books of e.g., [Ser99, Daf00]. Here, we shall study these limits in terms of the *entropy variables*, $\mathbf{v}(\mathbf{u}) := U_{\mathbf{u}}(\mathbf{u})$. We assume that the entropy $U(\mathbf{u})$ is convex, so that the nonlinear mapping $\mathbf{u} \mapsto \mathbf{v}$ is one-to-one. Following [God61, Moc80], we

claim that the change of variables, $\mathbf{u} = \mathbf{u}(\mathbf{v})$, puts the system (2.1) into the equivalent *symmetric form*,

$$\frac{\partial}{\partial x}\mathbf{u}(\mathbf{v}) + \frac{\partial}{\partial x}\mathbf{f}(\mathbf{u}(\mathbf{v})) + \frac{\partial}{\partial y}\mathbf{g}(\mathbf{u}(\mathbf{v})) = 0.$$

The above system is symmetric in the sense that the Jacobian matrices fluxes are,

$$\mathbf{u}_{\mathbf{v}}(\mathbf{v}) = (\mathbf{u}_{\mathbf{v}}(\mathbf{v}))^{\top}, \quad \mathbf{f}_{\mathbf{v}}(\mathbf{v}) = (\mathbf{f}_{\mathbf{v}}(\mathbf{v}))^{\top}, \text{ and } \mathbf{g}_{\mathbf{v}}(\mathbf{v}) = (\mathbf{g}_{\mathbf{v}}(\mathbf{v}))^{\top}. \tag{2.6}$$

Indeed, a straightforward computation using the compatibility relations (2.3) shows that $\mathbf{u}(\mathbf{v})$, $\mathbf{f}(\mathbf{v})$, and $\mathbf{g}(\mathbf{v})$ are, respectively, the gradients of the corresponding potential functions ϕ, ψ^x, and ψ^y,

$$\mathbf{u}(\mathbf{v}) = \phi_{\mathbf{v}}(\mathbf{v}), \quad \phi(\mathbf{v}) := \langle \mathbf{v}, \mathbf{u}(\mathbf{v})\rangle - U(\mathbf{u}(\mathbf{v})), \tag{2.7}$$

$$\mathbf{f}(\mathbf{v}) = \psi^x_{\mathbf{v}}(\mathbf{v}), \quad \psi^x(\mathbf{v}) := \langle \mathbf{v}, \mathbf{f}(\mathbf{v})\rangle - F(\mathbf{u}(\mathbf{v})), \tag{2.8}$$

$$\mathbf{g}(\mathbf{v}) = \psi^y_{\mathbf{v}}(\mathbf{v}), \quad \psi^y(\mathbf{v}) := \langle \mathbf{v}, \mathbf{g}(\mathbf{v})\rangle - G(\mathbf{u}(\mathbf{v})). \tag{2.9}$$

Hence the Jacobian matrices $H(\mathbf{v}) := \mathbf{u}_{\mathbf{v}}(\mathbf{v})$, $A^x(\mathbf{v}) := \mathbf{f}_{\mathbf{v}}(\mathbf{v})$, and $A^y(\mathbf{v}) := \mathbf{g}_{\mathbf{v}}(\mathbf{v})$ in (2.6) are symmetric, being Hessians of the potentials $\phi(\mathbf{v})$, $\psi^x(\mathbf{v})$, and $\psi^y(\mathbf{v})$. Moreover, the convexity of $U(\cdot)$ implies that H is positive definite, $H = (U_{\mathbf{uu}})^{-1} > 0$.

We now introduce the same entropy change of variables, $\mathbf{u} = \mathbf{u}(\mathbf{v})$, into the associated parabolic system (2.4), which reads

$$\frac{\partial}{\partial t}\mathbf{u}(\mathbf{v}^\zeta) + \frac{\partial}{\partial x}\mathbf{f}(\mathbf{v}^\zeta) + \frac{\partial}{\partial y}\mathbf{g}(\mathbf{v}^\zeta) = \zeta\frac{\partial}{\partial x}\left(S(\mathbf{v}^\zeta)\frac{\partial}{\partial x}\mathbf{v}^\zeta\right) + \zeta\frac{\partial}{\partial y}\left(S(\mathbf{v}^\zeta)\frac{\partial}{\partial y}\mathbf{v}^\zeta\right). \tag{2.10a}$$

By (2.6) and admissibility condition (2.5), the system (2.10a) is symmetric in the sense that the Jacobian matrices involved are all symmetric, namely, (2.6) holds and

$$S(\mathbf{u}(\mathbf{v})) = S^{\top}(\mathbf{u}(\mathbf{v})) > 0, \qquad S(\mathbf{v}) := Q(\mathbf{u}(\mathbf{v}))\mathbf{u}_{\mathbf{v}}(\mathbf{v}). \tag{2.10b}$$

Integrate (2.4) against the *entropy variable* $\mathbf{v} := U_{\mathbf{u}}$, employ the compatibility relations (2.3) and use 'differentiation by parts' on the dissipation terms on the RHS to find the following entropy balance statement,

$$\begin{aligned}\frac{\partial}{\partial t}U(\mathbf{u}^\zeta) + \frac{\partial}{\partial x}\Big(F(\mathbf{u}^\zeta) - \zeta\left\langle \mathbf{v}^\zeta, Q\mathbf{u}^\zeta_x\right\rangle\Big) + \frac{\partial}{\partial y}\Big(G(\mathbf{u}^\zeta) - \zeta\left\langle \mathbf{v}^\zeta, Q\mathbf{u}^\zeta_y\right\rangle\Big) = \\ -\zeta\left[\left\langle \mathbf{v}^\zeta_x, S(\mathbf{v}^\zeta)\mathbf{v}^\zeta_x\right\rangle + \left\langle \mathbf{v}^\zeta_y, S(\mathbf{v}^\zeta)\mathbf{v}^\zeta_y\right\rangle\right] \le 0.\end{aligned} \tag{2.11}$$

Letting $\zeta \downarrow 0$, we obtain the *entropy inequality*, [God61, Kru70, Lax71]

$$\frac{\partial}{\partial t}U(\mathbf{u}) + \frac{\partial}{\partial x}F(\mathbf{u}) + \frac{\partial}{\partial y}G(\mathbf{u}) \le 0. \tag{2.12}$$

This shows that weak solution dissipates entropy. The precise amount of entropy decay is dictated by the specific dissipation: spatial integration of (2.11) yields the entropy decay statement,

$$\frac{d}{dt}\int_y\int_x U(\mathbf{u}^\zeta)\,dxdy = -\zeta\int_y\int_x \left[\left\langle \mathbf{v}_x^\zeta, S(\mathbf{v}^\zeta)\mathbf{v}_x^\zeta\right\rangle + \left\langle \mathbf{v}_y^\zeta, S(\mathbf{v}^\zeta)\mathbf{v}_y^\zeta\right\rangle\right] dxdy \le 0. \tag{2.13}$$

2.2 The Example of the Shallow Water Equations

We consider the 2D shallow water equations (1.1) for the conservative variables $\mathbf{u} := (h, uh, vh)^\top$ where h is the water-depth and u, v are depth-averaged velocity components along x and y-direction. The total energy is given by the depth-averaged sum of the potential and kinetic energies,

$$E(\mathbf{u}) := \frac{gh^2 + u^2h + v^2h}{2}. \tag{2.14a}$$

The total energy plays the role of an entropy function for the SW equations. Straightforward computation gives us the following entropy fluxes, entropy variables and potentials.

- Entropy fluxes

$$F(\mathbf{u}) = guh^2 + \frac{u^3h + uv^2h}{2}, \quad G(\mathbf{u}) = gvh^2 + \frac{u^2vh + v^3h}{2}. \tag{2.14b}$$

- Entropy variable

$$\mathbf{v}(\mathbf{u}) = \begin{bmatrix} gh - \dfrac{u^2+v^2}{2} \\ u \\ v \end{bmatrix} \tag{2.14c}$$

  with the Jacobian matrices, $H := \mathbf{u}_\mathbf{v}$ and $H^{-1} = \mathbf{v}_\mathbf{u}$, given by

$$H = \frac{1}{g}\begin{bmatrix} 1 & u & v \\ u & c^2+u^2 & uv \\ v & uv & c^2+v^2 \end{bmatrix}, \quad H^{-1} = \frac{1}{h}\begin{bmatrix} c^2+u^2+v^2 & -u & -v \\ -u & 1 & 0 \\ -v & 0 & 1 \end{bmatrix}, \tag{2.14d}$$

  where $c := \sqrt{gh}$ is the 'sound' speed, or wave celerity.
- The potentials of the temporal and spatial fluxes $\mathbf{u}(\mathbf{v})$, $\mathbf{f}(\mathbf{u}(\mathbf{v}))$ and $\mathbf{g}(\mathbf{u}(\mathbf{v}))$ are given, respectively, by

$$\phi(\mathbf{v}) = \frac{gh^2}{2}, \quad \psi^x(\mathbf{v}) = \frac{guh^2}{2}, \quad \psi^y(\mathbf{v}) = \frac{gvh^2}{2}. \tag{2.14e}$$

The general statement of entropy balance, (2.13), amounts to

$$\frac{d}{dt}\int_y\int_x E(\mathbf{u})\,dxdy = -\zeta\int_y\int_x h(u_x^2+u_y^2+v_x^2+v_y^2)\,dxdy, \quad E(\mathbf{u}) = \frac{gh^2+u^2h+v^2h}{2}. \tag{2.15}$$

Since $h \geq 0$, we conclude that the total energy is decreasing in time, thus recovering energy stability. In fact, the expression on the RHS of (2.2) specifies the precise decay rate, which is dictated solely by the viscous fluxes through their dependence on the nonnegative eddy viscosity ζ. Our objective in this paper is to construct "faithful" approximations to the 2D shallow water equations, which precisely reproduce the energy balance (2.2).

3 Entropy Conservative Schemes: The 1D Setup

Setting $\mathbf{g} \equiv 0$ in (2.1), we consider the one-dimensional system of hyperbolic conservation laws,

$$\frac{\partial \mathbf{u}}{\partial t} + \frac{\partial}{\partial x}\mathbf{f}(\mathbf{u}) = 0, \quad x \in \mathbb{R},\ t > 0, \tag{3.1}$$

governing the N-vector of conserved variables $\mathbf{u} = [u_1, \cdots, u_N]^\top$ and balanced by the flux functions $\mathbf{f} = [f_1, \cdots, f_N]^\top$. We assume it is endowed with an entropy pair, (U, F), such that every strong solution of (3.1) satisfies the entropy equality

$$\frac{\partial}{\partial t}U(\mathbf{u}) + \frac{\partial}{\partial x}F(\mathbf{u}) = 0, \tag{3.2}$$

whereas weak solutions are sought to satisfy the entropy inequality, $U(\mathbf{u})_t + F(\mathbf{u})_x \leq 0$.

We now turn our attention to consistent approximations of (3.1), (3.2), based on semi-discrete conservative schemes of the form

$$\frac{d}{dt}\mathbf{u}_\nu(t) = -\frac{1}{\Delta x}\left(\mathbf{f}_{\nu+\frac{1}{2}} - \mathbf{f}_{\nu-\frac{1}{2}}\right). \tag{3.3}$$

Here, $\mathbf{u}_\nu(t)$ denotes the discrete solution along the equally spaced grid lines, $(x_\nu := \nu\Delta x, t)$, and $\mathbf{f}_{\nu+\frac{1}{2}}$ is the Lipschitz-continuous numerical flux which occupies a stencil of $2p$-gridvalues,

$$\mathbf{f}_{\nu+\frac{1}{2}} = \mathbf{f}(\mathbf{u}_{\nu-p+1}, \cdots, \mathbf{u}_{\nu+p}).$$

The scheme is *consistent* with the system (3.1) if $\mathbf{f}(\mathbf{u}, \mathbf{u}, \cdots, \mathbf{u}) = \mathbf{f}(\mathbf{u}),\ \ \forall \mathbf{u} \in \mathbb{R}^N$. Making the change of variables $\mathbf{u}_\nu = \mathbf{u}(\mathbf{v}_\nu)$, we obtain the equivalent form of (3.3)

$$\frac{d}{dt}\mathbf{u}(\mathbf{v}_\nu(t)) = -\frac{1}{\Delta x}\left(\mathbf{f}_{\nu+\frac{1}{2}} - \mathbf{f}_{\nu-\frac{1}{2}}\right). \tag{3.4}$$

The essential difference lies with the numerical flux, $\mathbf{f}_{\nu+\frac{1}{2}}$, which is now expressed in terms of the entropy variables,

$$\mathbf{f}_{\nu+\frac{1}{2}} = \mathbf{f}\left(\mathbf{v}_{\nu-p+1}, \cdots, \mathbf{v}_{\nu+p}\right) := \mathbf{f}\left(\mathbf{u}\left(\mathbf{v}_{\nu-p+1}\right), \cdots, \mathbf{u}\left(\mathbf{v}_{\nu+p}\right)\right),$$

consistent with the differential flux, $\mathbf{f}(\mathbf{v}, \mathbf{v}, \cdots, \mathbf{v}) = \mathbf{f}(\mathbf{v}) \equiv \mathbf{f}(\mathbf{u}(\mathbf{v}))$. The semi-discrete schemes (3.3) and (3.4) are completely identical. The entropy variables-based formula (3.4) has the advantage that it provides a natural ordering of symmetric matrices, which in turn enables us to *compare* the numerical viscosities of different schemes, consult [Tad87] for details. In particular, we will be able to utilize the so called entropy conservative discretization of [Tad03] for the convective part of the system of conservation laws (3.1), and thus recover the precise entropy balance dictated by physical dissipative terms of the underlying original systems.

The scheme (3.3) is called *entropy-conservative* if it satisfies a discrete *entropy equality*,

$$\frac{d}{dt}U\left(\mathbf{u}_\nu(t)\right) + \frac{1}{\Delta x}\left(F_{\nu+\frac{1}{2}} - F_{\nu-\frac{1}{2}}\right) = 0, \tag{3.5}$$

where $F_{\nu+\frac{1}{2}} = F\left(\mathbf{u}_{\nu-p+1}, \cdots, \mathbf{u}_{\nu+p}\right)$ is a consistent numerical entropy flux, $F(\mathbf{u}, \mathbf{u}, \cdots, \mathbf{u}) = F(\mathbf{u})$, $\forall \mathbf{u} \in \mathbb{R}^N$. Entropy conservative schemes will play an essential role in the construction of entropy stable schemes, by adding a judicious amount of physical viscosity.

The key step in the construction of entropy conservative schemes for the systems of conservation laws is the choice of *an arbitrary* piecewise-constant path in phase space. We shall use the phase space of the entropy variable $\mathbf{v}$ to connect two neighboring gridvalues, $\mathbf{v}_\nu$ and $\mathbf{v}_{\nu+1}$, at the spatial cell $[x_\nu, x_{\nu+1}]$, through the intermediate states $\{\mathbf{v}^j_{\nu+\frac{1}{2}}\}_{j=1}^N$. To this end, let $\{\mathbf{r}_j \equiv \mathbf{r}^j_{\nu+\frac{1}{2}}\}_{j=1}^N$ be an arbitrary set of N linearly independent N-vectors, and let $\{\boldsymbol{\ell}_j \equiv \boldsymbol{\ell}^j_{\nu+\frac{1}{2}}\}_{j=1}^N$ be the corresponding orthogonal set. We introduce the intermediate gridvalues, $\{\mathbf{v}^j_{\nu+\frac{1}{2}}\}_{j=1}^N$, which define a piecewise constant path in phase space across the jump $\Delta\mathbf{v}_{\nu+\frac{1}{2}} := \mathbf{v}_{\nu+1} - \mathbf{v}_\nu$,

$$\begin{cases} \mathbf{v}^1_{\nu+\frac{1}{2}} = \mathbf{v}_\nu \\ \mathbf{v}^{j+1}_{\nu+\frac{1}{2}} = \mathbf{v}^j_{\nu+\frac{1}{2}} + \left\langle \boldsymbol{\ell}_j, \Delta\mathbf{v}_{\nu+\frac{1}{2}} \right\rangle \mathbf{r}_j, \quad j = 1, 2, \cdots, N-1, . \\ \mathbf{v}^{N+1}_{\nu+\frac{1}{2}} = \mathbf{v}_{\nu+1} \end{cases} \tag{3.6}$$

Theorem 3.1 (Tadmor [Tad03, Theorem 6.1]). *Consider the system of conservation laws (3.1). Given the entropy pair (U, F), then the conservative scheme*

$$\frac{d}{dt}\mathbf{u}_\nu(t) = -\frac{1}{\Delta x_\nu}\left(\mathbf{f}^*_{\nu+\frac{1}{2}} - \mathbf{f}^*_{\nu-\frac{1}{2}}\right) \tag{3.7}$$

with a numerical flux $\mathbf{f}^*_{\nu+\frac{1}{2}}$

$$\mathbf{f}^*_{\nu+\frac{1}{2}} = \sum_{j=1}^{N} \frac{\psi\left(\mathbf{v}^{j+1}_{\nu+\frac{1}{2}}\right) - \psi\left(\mathbf{v}^{j}_{\nu+\frac{1}{2}}\right)}{\left\langle \boldsymbol{\ell}_j, \Delta\mathbf{v}_{\nu+\frac{1}{2}} \right\rangle} \boldsymbol{\ell}_j \tag{3.8}$$

is an entropy-conservative approximation, consistent with (3.1) and (3.2). Here, $\mathbf{v} = U_{\mathbf{u}}(\mathbf{u})$ *are the entropy variables associated with the entropy* U*, and* $\psi(\mathbf{v}) := \langle \mathbf{v}, \mathbf{f}(\mathbf{u}(\mathbf{v}))\rangle - F(\mathbf{u}(\mathbf{v}))$ *is the entropy potential.*

The proof is based on the fact that the entropy equality (3.5) holds if and only if $\left\langle \Delta\mathbf{v}_{\nu+\frac{1}{2}}, \mathbf{f}^*_{\nu+\frac{1}{2}} \right\rangle$ equals a *conservative difference*,

$$\left\langle \Delta\mathbf{v}_{\nu+\frac{1}{2}}, \mathbf{f}^*_{\nu+\frac{1}{2}} \right\rangle = \Delta\psi_{\nu+\frac{1}{2}}, \qquad \Delta\psi_{\nu+\frac{1}{2}} := \psi(\mathbf{v}_{\nu+1}) - \psi(\mathbf{v}_\nu). \tag{3.9}$$

Indeed, (3.9) is equivalent to (3.5),

$$\left\langle \mathbf{v}_\nu, \mathbf{f}^*_{\nu+\frac{1}{2}} - \mathbf{f}^*_{\nu-\frac{1}{2}} \right\rangle = F_{\nu+\frac{1}{2}} - F_{\nu-\frac{1}{2}}, \tag{3.10a}$$

where the numerical entropy flux $F_{\nu+\frac{1}{2}}$ is given by

$$F_{\nu+\frac{1}{2}} = \frac{1}{2}\Big[\left\langle \mathbf{v}_\nu + \mathbf{v}_{\nu+1}, \mathbf{f}^*_{\nu+\frac{1}{2}} \right\rangle - \Big(\psi(\mathbf{v}_\nu) + \psi(\mathbf{v}_{\nu+1})\Big)\Big] \tag{3.10b}$$

A straightforward manipulation of the numerical flux (3.8) confirms the desired equality (3.9),

$$\begin{aligned}
\left\langle \Delta\mathbf{v}_{\nu+\frac{1}{2}}, \mathbf{f}^*_{\nu+\frac{1}{2}} \right\rangle &= \sum_{j=1}^{N} \frac{\psi\left(\mathbf{v}^{j+1}_{\nu+\frac{1}{2}}\right) - \psi\left(\mathbf{v}^{j}_{\nu+\frac{1}{2}}\right)}{\left\langle \boldsymbol{\ell}_j, \Delta\mathbf{v}_{\nu+\frac{1}{2}} \right\rangle} \left\langle \boldsymbol{\ell}_j, \Delta\mathbf{v}_{\nu+\frac{1}{2}} \right\rangle \\
&= \sum_{j=1}^{N} \psi\left(\mathbf{v}^{j+1}_{\nu+\frac{1}{2}}\right) - \psi\left(\mathbf{v}^{j}_{\nu+\frac{1}{2}}\right) = \psi\left(\mathbf{v}^{N+1}_{\nu+\frac{1}{2}}\right) - \psi\left(\mathbf{v}^{1}_{\nu+\frac{1}{2}}\right) \\
&= \Delta\psi_{\nu+\frac{1}{2}}.
\end{aligned}$$

Although the recipe for constructing entropy-conservative fluxes in (3.8) allows an *arbitrary* choice of a path in phase space, inappropriate choices of the path may cause the computed intermediate values to lie outside the physical space, say $h < 0$. A 'physically relevant' choice is offered by a Riemann path which consists of $\{\mathbf{u}^j_{\nu+\frac{1}{2}}\}_{j=1}^N$, stationed along an (approximate) set of right eigenvectors, $\{\widehat{\mathbf{r}}_j\}$, of the Jacobian $\mathbf{f}_{\mathbf{u}}(\mathbf{u}_{\nu+\frac{1}{2}})$. Set $\mathbf{v}^j_{\nu+\frac{1}{2}} = \mathbf{v}(\mathbf{u}^j_{\nu+\frac{1}{2}})$, $j = 1, 2, \ldots, N$, and let $\boldsymbol{\ell}_j$'s be the orthogonal system to $\{\mathbf{v}^{j+1} - \mathbf{v}^j\}_{j=1}^N$. This will be our choice of a path for computing entropy stable approximations of shallow water equations in Sect. 5 below. The construction of the entropy conservative flux $\mathbf{f}^*_{\nu+\frac{1}{2}}$ follows [TZ06, Algorithm 1] which states,

Algorithm 1 *If* $\mathbf{u}_\nu = \mathbf{u}_{\nu+1}$ ***then*** $\mathbf{f}^*_{\nu+\frac{1}{2}} = \mathbf{f}(\mathbf{v}_\nu)$*;* ***else***

- *Set* $\mathbf{u}^1_{\nu+\frac{1}{2}} := \mathbf{u}_\nu$ *and compute recursively the intermediate states,*

$$\mathbf{u}^{j+1}_{\nu+\frac{1}{2}} = \mathbf{u}^j_{\nu+\frac{1}{2}} + \left\langle \widehat{\boldsymbol{\ell}}_j, \Delta \mathbf{u}_{\nu+\frac{1}{2}} \right\rangle \widehat{\mathbf{r}}_j, \quad j = 1, 2, 3. \tag{3.11}$$

  Here, $\{\widehat{\boldsymbol{\ell}}_j\}$ *and* $\{\widehat{\mathbf{r}}_j\}$ *are the left and right eigensystems of an averaged Jacobian* $\widetilde{A}_{\nu+\frac{1}{2}}$*, given by the Roe matrix,* $\widetilde{A}_{\nu+\frac{1}{2}} = \widetilde{A}(\mathbf{u}_\nu, \mathbf{u}_{\nu+1})$ *(see [Roe81]).*
- *Set* $\mathbf{r}_j := \mathbf{v}(\mathbf{u}^{j+1}_{\nu+\frac{1}{2}}) - \mathbf{v}(\mathbf{u}^j_{\nu+\frac{1}{2}})$ *and compute* $\{\boldsymbol{\ell}_j\}_{j=1}^3$ *as the corresponding orthogonal system. (Note that* $\{\mathbf{r}_j, \boldsymbol{\ell}_j\}$ *is the eigen-path in* $\mathbf{v}$*-space, corresponding to the eigen-path in* $\mathbf{u}$*-space,* $\{\widehat{\mathbf{r}}_j, \widehat{\boldsymbol{\ell}}_j\}$*.)*
- *Compute the entropy-conservative numerical flux,*

$$\mathbf{f}^*_{\nu+\frac{1}{2}} = \sum_{j=1}^{3} \frac{\psi(\mathbf{v}^{j+1}_{\nu+\frac{1}{2}}) - \psi(\mathbf{v}^j_{\nu+\frac{1}{2}})}{\left\langle \boldsymbol{\ell}_j, \Delta \mathbf{v}_{\nu+\frac{1}{2}} \right\rangle} \boldsymbol{\ell}_j. \tag{3.12}$$

4 Scalar Problems

We test our entropy stable schemes with the prototype example of inviscid Burgers' equation. Though very simple, the inviscid Burgers' equation is often used as the testing ground for numerical approximations of nonlinear conservation laws.

4.1 Entropy Conservative Schemes

We consider the inviscid Burgers' equation,

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x} f(u) = 0, \quad f(u) = \frac{1}{2} u^2. \tag{4.1}$$

Any convex function $U(u)$ serves as an entropy function for the scalar Burgers equation. The solutions of (4.1) satisfy, at the *formal* level,

$$\frac{\partial}{\partial t} U(u) + \frac{\partial}{\partial x} F(u) = 0. \tag{4.2}$$

These are additional conservation laws balanced by the corresponding entropy flux functions $F(u)$ satisfying the compatibility relation $U'f' = F'$. Spatial integration then yields the total entropy conservation (ignoring boundary contributions)

$$\int_x U(x,t)\, dx = \int_x U(x,0)\, dx. \tag{4.3}$$

We now turn to the discrete framework. Discretization in space yields the semi-discrete scheme,

$$\frac{d}{dt}u_\nu(t) + \frac{1}{\Delta x}\left(f_{\nu+\frac{1}{2}} - f_{\nu-\frac{1}{2}}\right) = 0. \tag{4.4}$$

Clearly, $\sum u_\nu(t)\Delta x$ is conserved. We seek a consistent numerical flux $f_{\nu+\frac{1}{2}}$, that is entropy conservative in the sense of satisfying the discrete analogue of (4.2),

$$\frac{d}{dt}U(u_\nu(t)) + \frac{1}{\Delta x}(F_{\nu+\frac{1}{2}} - F_{\nu-\frac{1}{2}}) = 0,$$

so that we have the additional conservation of entropy $\sum U(u_\nu(t))\Delta x$. According to Theorem 3.1, consult (3.9), such 2-point *scalar* entropy conservative fluxes are *uniquely* determined, $f_{\nu+\frac{1}{2}} = f^*_{\nu+\frac{1}{2}}$, by

$$f_{\nu+\frac{1}{2}} = f^*_{\nu+\frac{1}{2}} := \frac{\psi(u_{\nu+1}) - \psi(u_\nu)}{v(u_{\nu+1}) - v(u_\nu)}. \tag{4.5}$$

Recall that $v(u) := U'(u)$ is the entropy variable associated with the entropy pair (U, F), and $\psi(u) := v(u)f(u) - F(u)$ is the potential function of the flux $f(u(v))$. We demonstrate the constructions of above entropy conservative numerical flux with two different choices of entropy functions:

- We begin with the logarithmic entropy $U(u) = -\ln u$ together with the entropy flux $F(u) = -u$. We use the entropy variable $v(u) = -1/u$. The entropy flux potential in this case is $\psi(u) = -1/2v = u/2$. The entropy conservative numerical flux (4.5) then reads,

  $$f^*_{\nu+\frac{1}{2}} := \frac{\psi(u_{\nu+1}) - \psi(u_\nu)}{v(u_{\nu+1}) - v(u_\nu)} = \frac{1}{2}u_\nu u_{\nu+1}.$$

  This numerical flux yields the entropy conservative schemes

  $$\frac{d}{dt}u_\nu(t) = u_\nu(t)\frac{u_{\nu+1}(t) - u_{\nu-1}(t)}{2\Delta x}.$$

  This scheme was discussed by Goodman and Lax in [GL88], Hou and Lax in [HL91], and Levermore and Liu in [LL96] in their study of the *dispersive oscillations* arising in numerical solutions of the conservative schemes for the inviscid Burgers' equation.
- Next, we consider the family of entropy functions,

  $$U_p(u) = u^{2p}, \quad p = 1, 2, \cdots, \tag{4.6}$$

  with the corresponding entropy flux functions $F_p(u) = 2pu^{2p+1}/(2p+1)$. Using the entropy variable $v(u) := U'(u) = 2pu^{2p-1}$ and the potential function $\psi(u) := v(u)f(u) - F(u) = \frac{p(2p-1)}{2p+1}u^{2p+1}$, we compute the entropy conservative flux

  $$f^*_{\nu+\frac{1}{2}} := \frac{\psi(u_{\nu+1}) - \psi(u_\nu)}{v(u_{\nu+1}) - v(u_\nu)} = \frac{2p-1}{2(2p+1)} \cdot \frac{u^{2p+1}_{\nu+1} - u^{2p+1}_\nu}{u^{2p-1}_{\nu+1} - u^{2p-1}_\nu}. \tag{4.7}$$

The resulting scheme (4.4), (4.7) is entropy conservative in the sense that the discrete analogue of total entropy conservation (4.3) is satisfied,

$$\sum_{\nu} u_{\nu}^{2p}(t)\,\Delta x = \sum_{\nu} u_{\nu}^{2p}(0)\,\Delta x.$$

Thus, for each p we obtain its own U_p-entropy conservative scheme.

Remark 4.1. Although these schemes with the entropy-conservative flux (4.7) admit the dispersive oscillations shown in the numerical results of Sect. 4.3, we expect the amplitude of these oscillations to be reduced for increasing p's, as the conservation of entropies U_p,

$$\left[\sum_{\nu} u_{\nu}^{2p}(t)\,\Delta x\right]^{\frac{1}{2p}} = \left[\sum_{\nu} u_{\nu}^{2p}(0)\,\Delta x\right]^{\frac{1}{2p}} \tag{4.8}$$

approaches the maximum principle, $||u_\nu(t)||_{L^\infty} \le ||u_\nu(0)||_{L^\infty}$ (the inequality reflects the small amount of dissipation due to time discretization). Indeed, as $p \uparrow \infty$, the entropy-conservative schemes based on (4.7) approach the first-order entropy stable Engquist–Osher scheme [EO80].

4.2 Entropy Dissipation

To recover the physical relevant *entropy inequality*, that is

$$\partial_t U_p(u) + \partial_x F_p(u) \le 0,$$

one can add numerical dissipation,

$$\frac{d}{dt}u_\nu(t) + \frac{1}{\Delta x}\left(f^*_{\nu+\frac{1}{2}} - f^*_{\nu-\frac{1}{2}}\right) = \frac{\epsilon}{(\Delta x)^2}\Big(d(u_{\nu+1}) - 2d(u_\nu) + d(u_{\nu-1})\Big), \quad \epsilon > 0. \tag{4.9}$$

This serves as an approximation to the vanishing viscosity regularization

$$u_t + f(u)_x = \epsilon d(u)_{xx}, \quad d'(u) > 0,\ \epsilon > 0.$$

Sum this scheme (4.9) against the entropy variable v_ν to find

$$\frac{d}{dt}\sum_{\nu} U_p(u_\nu(t))\Delta x + \sum_{\nu} v_\nu\left(f^*_{\nu+\frac{1}{2}} - f^*_{\nu-\frac{1}{2}}\right) = \epsilon\sum_{\nu} v_\nu \frac{d(u_{\nu+1}) - 2d(u_\nu) + d(u_{\nu-1})}{\Delta x}. \tag{4.10}$$

According to (3.10a), the second term on the left of (4.10) vanishes, $\sum \left(F_{\nu+\frac{1}{2}} - F_{\nu-\frac{1}{2}}\right)\Delta x = 0$. Summation by parts on the RHS of (4.10) yields

$$\epsilon \sum_{\nu} v_{\nu} \frac{d(u_{\nu+1}) - 2d(u_{\nu}) + d(u_{\nu-1})}{\Delta x} = -\frac{\epsilon}{\Delta x} \sum_{\nu} \big(v_{\nu+1} - v_{\nu}\big) \times \big(d(u_{\nu+1}) - d(u_{\nu})\big) \le 0,$$

since $d'(v) = d'(u)u'(v) > 0$, and hence $(v_{\nu+1} - v_{\nu}) \cdot (d(u_{\nu+1}) - d(u_{\nu})) > 0$. The resulting entropy balance that follows reads,

$$\frac{d}{dt} \sum_{\nu} U_p(u_{\nu}(t)) \Delta x = -\frac{\epsilon}{\Delta x} \sum_{\nu} \Delta v_{\nu+\frac{1}{2}} \Delta d_{\nu+\frac{1}{2}} \le 0. \tag{4.11}$$

Observe that the amount of entropy dissipation on the right is completely determined by the dissipation term $\epsilon d(u)$. No artificial viscosity is introduced by the convective term. If we exclude any dissipative mechanism ($\epsilon = 0$), then we are back at the entropy conservative schemes of Sect. 4.1.

4.3 Numerical Experiments

Time Discretization

To complete the computation of a semi-discrete scheme, the semi-discrete entropy conservative scheme (4.4), (4.7) needs to be augmented with a proper time discretization. To enable a large time-stability region and maintain simplicity, the explicit three-stage third-order Runge–Kutta (RK3) method will be used, Consult [GST01] for more detail of its strong stability-preserving property,

$$\begin{cases} u^{(1)} &= u^n + \Delta t \mathcal{L}(u^n) \\ u^{(2)} &= \frac{3}{4}u^n + \frac{1}{4}u^{(1)} + \frac{1}{4}\Delta t \mathcal{L}(u^{(1)}) \\ u^{n+1} &= \frac{1}{3}u^n + \frac{2}{3}u^{(2)} + \frac{2}{3}\Delta t \mathcal{L}(u^{(2)}) \end{cases} \tag{4.12a}$$

where

$$[\mathcal{L}(u)]_{\nu} := -\frac{1}{\Delta x}(f^*_{\nu+\frac{1}{2}} - f^*_{\nu-\frac{1}{2}}). \tag{4.12b}$$

We note that this explicit RK3 time discretization produces a negligible amount of entropy dissipation. For a general framework of entropy conservative fully discrete schemes, consult [LMR02].

Continuous Initial Conditions

We first solve the inviscid Burgers equation (4.1) in the domain $x \in [0, 1]$ with initial condition, $u(0, x) = \sin(2\pi x)$ and subject to periodic boundary conditions $u(t, 1) = u(t, 0)$. In Fig. 1 we display the numerical solutions for (4.12a) and (4.12b) with the numerical flux (4.7) for different choices of p. For small values of p, the dispersive oscillations become noticeable after the shock is generated due to the absence of any dissipative mechanism in the entropy-conservative scheme. As p increases, the amplitude of the spurious dispersive oscillations decreases, which reflects the control of the increasing L^{2p}-norms in (4.8).

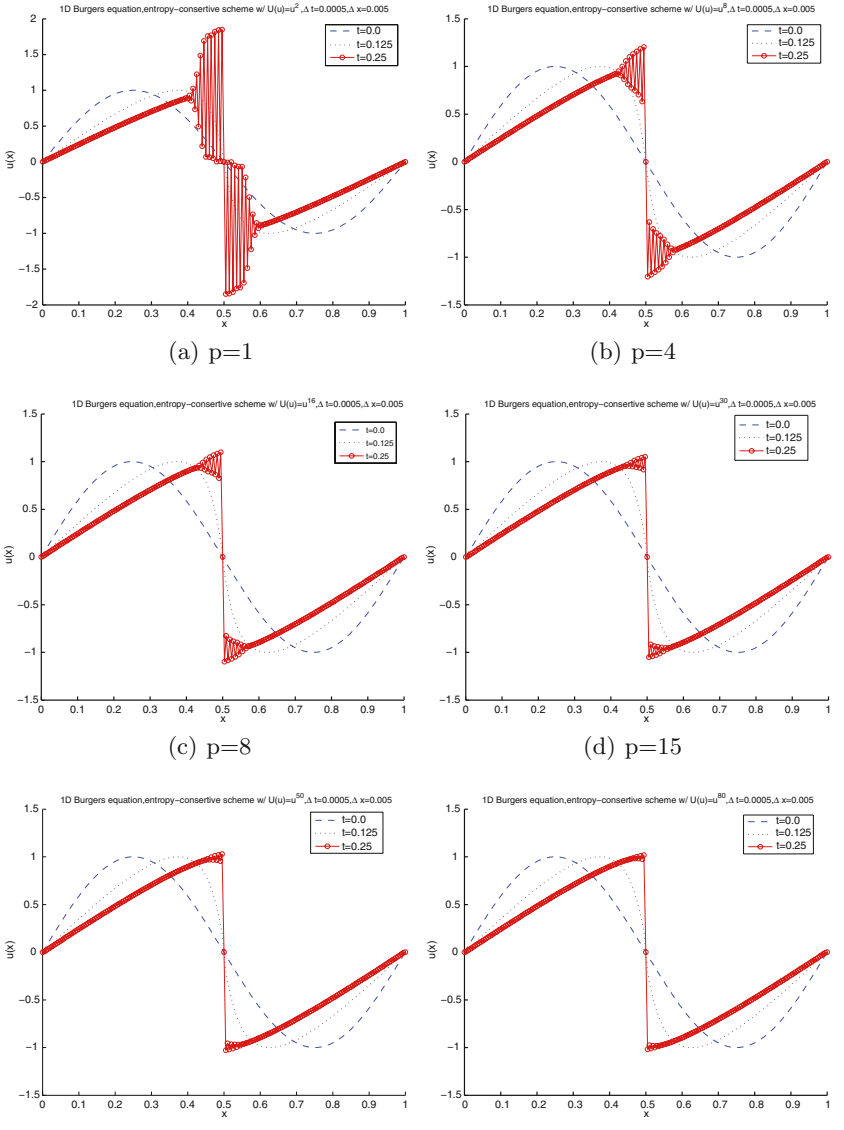


Fig. 1. 1D Burger's equation, sine initial condition, entropy-conservative schemes, 200 spatial grids, $U(u) = u^{2p}$

Discontinuous Initial Conditions

We solve the 1D inviscid Burgers equation (4.1) in the domain $x \in [0, 1]$ with the discontinuous initial condition,

$$u(0,x)=\begin{cases}2, & x\in[0,\,0.5]\\ 1, & x\in(0.5,\,1]\end{cases}$$

The boundary values are extrapolated from the interior points. Since we are only interested in the propagation of the shock wave in the computational domain [0, 1], there is interaction with the boundary values which do not vary in the time interval under consideration. In Fig. 2, we display the numerical solutions for (4.12a) and (4.12b) with the numerical flux (4.7) for different choices of p. Those solutions show the same pattern as the $\sin(2\pi x)$ initial condition. Diminishing amplitude of the dispersive oscillations demonstrates the control of the L^{2p}-norm of the solution with each p.

5 2D Shallow Water Equations

5.1 Energy Stable Schemes

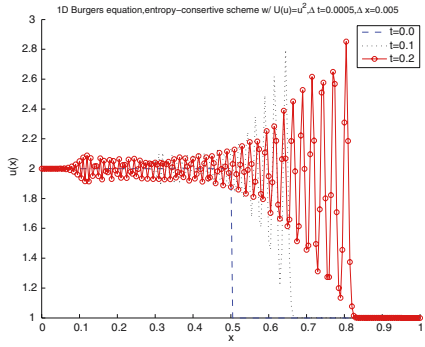
We turn to the construction of entropy/energy-stable schemes for the 2D shallow water equations,

$$\frac{\partial}{\partial t}\mathbf{u}+\frac{\partial}{\partial x}\mathbf{f}(\mathbf{u})+\frac{\partial}{\partial y}\mathbf{g}(\mathbf{u})=\zeta\frac{\partial}{\partial x}\left(h\frac{\partial}{\partial x}\mathbf{d}(\mathbf{u})\right)+\zeta\frac{\partial}{\partial y}\left(h\frac{\partial}{\partial y}\mathbf{d}(\mathbf{u})\right),\quad \mathbf{u}=\begin{bmatrix}h\\ uh\\ vh\end{bmatrix}, \tag{5.1}$$

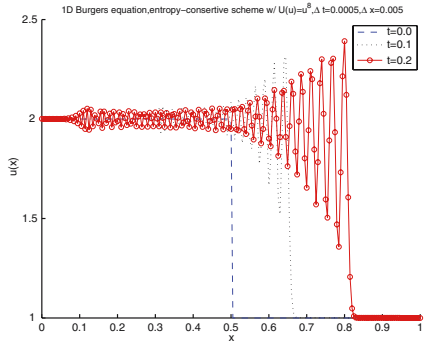
with convective fluxes $\mathbf{f}=[uh,\,u^2h+gh^2/2,\,uvh]^\top$, $\mathbf{g}=[vh,\,uvh,\,v^2h+gh^2/2]^\top$, and additional diffusive terms $\mathbf{d}=[0,\,u,\,v]^\top$.

The second-order semi-discrete entropy conservative schemes (3.7), (3.8) can be extended to two dimensional shallow water equations (5.1) in a straightforward manner. Recall that E denotes the total energy which is serving as an admissible entropy function with the corresponding entropy fluxes (F,G) associated with the two dimensional shallow water equations, $\mathbf{v}:=U_{\mathbf{v}}$ are the corresponding entropy variables (2.14c), and (ψ^x,ψ^y) are the potential pair (2.14e). We discretize the convective fluxes on the LHS using the entropy-conservative differences indicated in 1D setup dimension by dimension. For the dissipative terms on the RHS, we employ the centered differences, while the intermediate h-values are taken to be the arithmetic mean of two neighboring grid-points, $\widehat{h}_{\nu+\frac12,\mu}:=(h_{\nu+1,\mu}+h_{\nu,\mu})/2$. We then obtain the entropy stable semi-discrete schemes

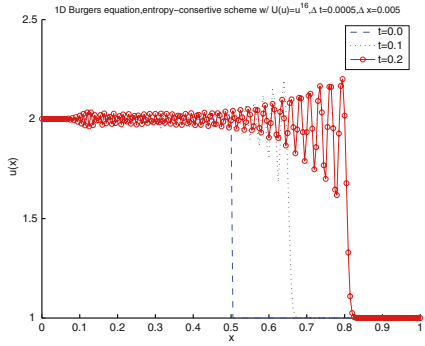
$$\begin{aligned}\frac{d}{dt}\mathbf{u}_{\nu,\,\mu}(t)&+\frac{1}{\Delta x}(\mathbf{f}^*_{\nu+\frac12,\,\mu}-\mathbf{f}^*_{\nu-\frac12,\,\mu})+\frac{1}{\Delta y}(\mathbf{g}^*_{\nu,\,\mu+\frac12}-\mathbf{g}^*_{\nu,\,\mu-\frac12})\\ &=\frac{\zeta}{\Delta x}\left(\widehat{h}_{\nu+\frac12,\,\mu}\frac{\mathbf{d}_{\nu+1,\,\mu}-\mathbf{d}_{\nu,\,\mu}}{\Delta x}-\widehat{h}_{\nu-\frac12,\,\mu}\frac{\mathbf{d}_{\nu,\,\mu}-\mathbf{d}_{\nu-1,\,\mu}}{\Delta x}\right)\\ &\quad+\frac{\zeta}{\Delta y}\left(\widehat{h}_{\nu,\,\mu+\frac12}\frac{\mathbf{d}_{\nu,\,\mu+1}-\mathbf{d}_{\nu,\,\mu}}{\Delta x}-\widehat{h}_{\nu,\,\mu-\frac12}\frac{\mathbf{d}_{\nu,\,\mu}-\mathbf{d}_{\nu,\,\mu-1}}{\Delta x}\right),\end{aligned} \tag{5.2a}$$



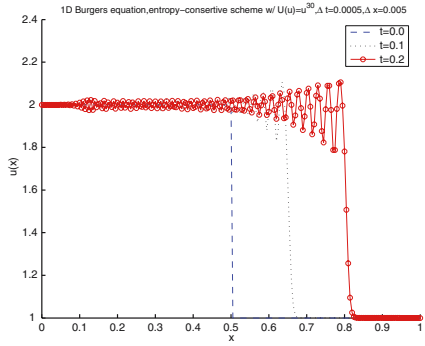
(a) p=1



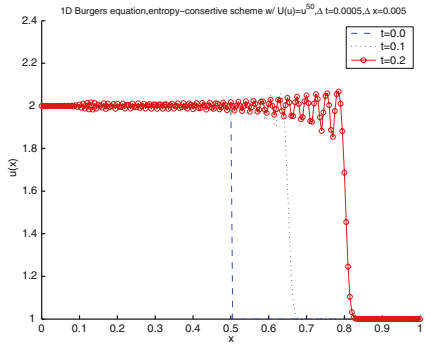
(b) p=4



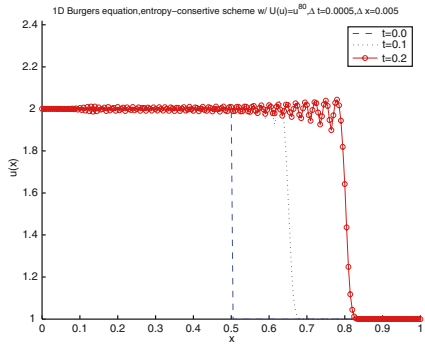
(c) p=8



(d) p=15



(e) p=25



(f) p=40

Fig. 2. 1D Burger's equation, shock initial condition, entropy-conservative schemes, 200 spatial grids, $U(u) = u^{2p}$

with the entropy-conservative fluxes $\mathbf{f}^*_{\nu+\frac{1}{2},\,\mu}$ and $\mathbf{g}^*_{\nu,\,\mu+\frac{1}{2}}$ outlined in (3.12) along x and y direction, respectively,

$$\begin{aligned}\mathbf{f}^*_{\nu+\frac{1}{2},\,\mu} &= \sum_{j=1}^{3} \frac{\psi^x\left(\mathbf{v}^{j+1}_{\nu+\frac{1}{2},\mu}\right) - \psi^x\left(\mathbf{v}^{j}_{\nu+\frac{1}{2},\mu}\right)}{\left\langle \boldsymbol{\ell}_{x_j}, \Delta\mathbf{v}_{\nu+\frac{1}{2},\mu}\right\rangle}\boldsymbol{\ell}_{x_j} \\ &= \frac{g}{2}\sum_{j=1}^{3} \frac{(h^{j+1}_{\nu+\frac{1}{2},\,\mu})^2 u^{j+1}_{\nu+\frac{1}{2},\,\mu} - (h^{j}_{\nu+\frac{1}{2},\,\mu})^2 u^{j}_{\nu+\frac{1}{2},\,\mu}}{\left\langle \boldsymbol{\ell}_{x_j},\, \Delta\mathbf{v}_{\nu+\frac{1}{2},\,\mu}\right\rangle}\,\boldsymbol{\ell}_{x_j}, \qquad (5.2b)\end{aligned}$$

$$\begin{aligned}\mathbf{g}^*_{\nu,\,\mu+\frac{1}{2}} &= \sum_{j=1}^{3} \frac{\psi^y\left(\mathbf{v}^{j+1}_{\nu,\mu+\frac{1}{2}}\right) - \psi^y\left(\mathbf{v}^{j}_{\nu,\mu+\frac{1}{2}}\right)}{\left\langle \boldsymbol{\ell}_{y_j}, \Delta\mathbf{v}_{\nu,\mu+\frac{1}{2}}\right\rangle}\boldsymbol{\ell}_{y_j} \\ &= \frac{g}{2}\sum_{j=1}^{3} \frac{(h^{j+1}_{\nu,\,\mu+\frac{1}{2}})^2 v^{j+1}_{\nu,\,\mu+\frac{1}{2}} - (h^{j}_{\nu,\,\mu+\frac{1}{2}})^2 v^{j}_{\nu,\,\mu+\frac{1}{2}}}{\left\langle \boldsymbol{\ell}_{y_j},\, \Delta\mathbf{v}_{\nu,\,\mu+\frac{1}{2}}\right\rangle}\,\boldsymbol{\ell}_{y_j}, \qquad (5.2c)\end{aligned}$$

Here, $\mathbf{u}_{\nu,\,\mu}(t)$ denotes the discrete solution at the grid point (x_ν, y_μ, t) with $x_\nu := \nu\Delta x$, $y_\mu := \mu\Delta y$, Δx and Δy being the uniform mesh sizes, and $\mathbf{d}_{\nu,\,\mu} := \mathbf{d}(\mathbf{u}_{\nu,\,\mu})$. The numerical flux $\mathbf{f}^*_{\nu+\frac{1}{2},\,\mu}$ and $\mathbf{g}^*_{\nu,\,\mu+\frac{1}{2}}$ are constructed separately along two different phase paths dictated by two sets of vectors $\{\boldsymbol{\ell}_{x_j}\}$ and $\{\boldsymbol{\ell}_{y_j}\}$. Finally, $\{u^j\}$, $\{v^j\}$, and $\{h^j\}$ are intermediate values of height and velocities along paths in the phase space. The physical relevance of the intermediate solutions along the paths needs to be maintained. To this end, we choose to work along the paths which are determined by (approximate) Riemann solvers. Specifically, we use the eigensystems of the Roe matrix in the x and y directions, [Roe81, Gla87],

$$\begin{aligned}\widetilde{A}^x &= \begin{bmatrix} 0 & 1 & 0 \\ \bar{c}^2_{\nu+\frac{1}{2},\,\mu} - \bar{u}^2_{\nu+\frac{1}{2},\,\mu} & 2\bar{u}_{\nu+\frac{1}{2},\,\mu} & 0 \\ -\bar{u}_{\nu+\frac{1}{2},\,\mu}\bar{v}_{\nu+\frac{1}{2},\,\mu} & \bar{v}_{\nu+\frac{1}{2},\,\mu} & \bar{u}_{\nu+\frac{1}{2},\,\mu} \end{bmatrix}, \\ \widetilde{A}^y &= \begin{bmatrix} 0 & 0 & 1 \\ -\bar{u}_{\nu,\,\mu+\frac{1}{2}}\bar{v}_{\nu,\,\mu+\frac{1}{2}} & \bar{v}_{\nu,\,\mu+\frac{1}{2}} & \bar{u}_{\nu,\,\mu+\frac{1}{2}} \\ \bar{c}^2_{\nu,\,\mu+\frac{1}{2}} - \bar{v}^2_{\nu,\,\mu+\frac{1}{2}} & 0 & 2\bar{v}_{\nu,\,\mu+\frac{1}{2}} \end{bmatrix}. \qquad (5.3a)\end{aligned}$$

Here $\bar{u}$, $\bar{v}$, and $\bar{c}$ are the average values of the velocities u, v and the sound speed $c := \sqrt{gh}$ at Roe-average state,

$$\bar{u} = \frac{u_R\sqrt{h_R} + u_L\sqrt{h_L}}{\sqrt{h_R}+\sqrt{h_L}}, \quad \bar{v} = \frac{v_R\sqrt{h_R} + v_L\sqrt{h_L}}{\sqrt{h_R}+\sqrt{h_L}}, \quad \bar{c} = \sqrt{\frac{g(h_R+h_L)}{2}}, \qquad (5.3b)$$

where the subscripts $(\cdot)_R$ and $(\cdot)_L$ represent two neighboring spatial grid-points. The vector sets $\{\widehat{\mathbf{r}}_{x_j}\}_{j=1}^3$ and $\{\widehat{\mathbf{r}}_{y_j}\}_{j=1}^3$ are chosen to be the right eigenvectors of the x- and y-Roe matrices (5.3a) (omitting the sub/superscripts of all averaged variables)

$$\widehat{\mathbf{r}}_{x_1} = \begin{bmatrix} 1 \\ \bar{u} - \bar{c} \\ \bar{v} \end{bmatrix}, \ \widehat{\mathbf{r}}_{x_2} = \begin{bmatrix} 0 \\ 0 \\ \bar{c} \end{bmatrix}, \ \widehat{\mathbf{r}}_{x_3} = \begin{bmatrix} 1 \\ \bar{u} + \bar{c} \\ \bar{v} \end{bmatrix}, \tag{5.3c}$$

$$\widehat{\mathbf{r}}_{y_1} = \begin{bmatrix} 1 \\ \bar{u} \\ \bar{v} - \bar{c} \end{bmatrix}, \ \widehat{\mathbf{r}}_{y_2} = \begin{bmatrix} 0 \\ -\bar{c} \\ 0 \end{bmatrix}, \ \widehat{\mathbf{r}}_{y_3} = \begin{bmatrix} 1 \\ \bar{u} \\ \bar{v} + \bar{c} \end{bmatrix}, \tag{5.3d}$$

with the corresponding left eigenvector sets $\{\widehat{\boldsymbol{\ell}}_{x_j}\}_{j=1}^3$ and $\{\widehat{\boldsymbol{\ell}}_{y_j}\}_{j=1}^3$ given by

$$\widehat{\boldsymbol{\ell}}_{x_1} = \begin{bmatrix} \dfrac{\bar{u} + \bar{c}}{2\bar{c}} \\ -\dfrac{1}{2\bar{c}} \\ 0 \end{bmatrix}, \ \widehat{\boldsymbol{\ell}}_{x_2} = \begin{bmatrix} -\dfrac{\bar{v}}{\bar{c}} \\ 0 \\ \dfrac{1}{\bar{c}} \end{bmatrix}, \ \widehat{\boldsymbol{\ell}}_{x_3} = \begin{bmatrix} \dfrac{-\bar{u} + \bar{c}}{2\bar{c}} \\ \dfrac{1}{2\bar{c}} \\ 0 \end{bmatrix}, \tag{5.3e}$$

$$\widehat{\boldsymbol{\ell}}_{y_1} = \begin{bmatrix} \dfrac{\bar{v} + \bar{c}}{2\bar{c}} \\ 0 \\ -\dfrac{1}{2\bar{c}} \end{bmatrix}, \ \widehat{\boldsymbol{\ell}}_{y_2} = \begin{bmatrix} \dfrac{\bar{u}}{\bar{c}} \\ -\dfrac{1}{\bar{c}} \\ 0 \end{bmatrix}, \ \widehat{\boldsymbol{\ell}}_{y_3} = \begin{bmatrix} \dfrac{-\bar{v} + \bar{c}}{2\bar{c}} \\ 0 \\ \dfrac{1}{2\bar{c}} \end{bmatrix}. \tag{5.3f}$$

We now are able to form the intermediate paths along x and y directions in $\mathbf{u}$-space as in (3.6): starting with $\mathbf{u}^1_{\nu+\frac{1}{2},\,\mu} = \mathbf{u}^1_{\nu,\,\mu+\frac{1}{2}} = \mathbf{u}_{\nu,\,\mu}$, we proceed with

$$\mathbf{u}^{j+1}_{\nu+\frac{1}{2},\,\mu} = \mathbf{u}^{j}_{\nu+\frac{1}{2},\,\mu} + \left\langle \widehat{\boldsymbol{\ell}}_{x_j},\, \Delta\mathbf{u}_{\nu+\frac{1}{2},\,\mu} \right\rangle \widehat{\mathbf{r}}_{x_j}, \ \ j=1,2,3, \ \ \Delta\mathbf{u}_{\nu+\frac{1}{2},\,\mu} := \mathbf{u}_{\nu+1,\,\mu} - \mathbf{u}_{\nu,\,\mu},$$
$$\mathbf{u}^{j+1}_{\nu,\,\mu+\frac{1}{2}} = \mathbf{u}^{j}_{\nu,\,\mu+\frac{1}{2}} + \left\langle \widehat{\boldsymbol{\ell}}_{y_j},\, \Delta\mathbf{u}_{\nu,\,\mu+\frac{1}{2}} \right\rangle \widehat{\mathbf{r}}_{y_j}, \ \ j=1,2,3, \ \ \Delta\mathbf{u}_{\nu,\,\mu+\frac{1}{2}} := \mathbf{u}_{\nu,\,\mu+1} - \mathbf{u}_{\nu,\,\mu}.$$

The construction of the entropy-conservative numerical fluxes $\mathbf{f}^*_{\nu+\frac{1}{2},\,\mu}$ and $\mathbf{g}^*_{\nu,\,\mu+\frac{1}{2}}$ follows the algorithm indicated in Algorithm 1.

Remark 5.1. We point out that in the case $\left\langle \widehat{\boldsymbol{\ell}}_j, \Delta\mathbf{u} \right\rangle = 0$ for certain j's in $\mathbf{u}$-space, which may cause $\langle \boldsymbol{\ell}_j, \Delta\mathbf{v} \rangle = 0$ in $\mathbf{v}$-space, hence fail Algorithm 1. Arguing along the same line as [TZ06, Remark 3.5], we compute the corresponding entropy-conservative numerical fluxes using the alternate formulas,

$$\mathbf{f}^*_{\nu+\frac{1}{2},\,\mu} = \sum_{\{j\,|\,\xi_{x_j}\neq 0\}} \frac{\psi^x(\mathbf{v}^j_{\nu+\frac{1}{2},\,\mu} + \xi_{x_j}\mathbf{r}_{x_j}) - \psi^x(\mathbf{v}^j_{\nu+\frac{1}{2},\,\mu})}{\xi_{x_j}} \boldsymbol{\ell}_{x_j}, \ \ \xi_{x_j} := \left\langle \boldsymbol{\ell}_{x_j}, \Delta\mathbf{v}_{\nu+\frac{1}{2},\,\mu} \right\rangle,$$

$$\mathbf{g}^*_{\nu,\,\mu+\frac{1}{2}} = \sum_{\{j\,|\,\xi_{y_j}\neq 0\}} \frac{\psi^y(\mathbf{v}^j_{\nu,\,\mu+\frac{1}{2}} + \xi_{y_j}\mathbf{r}_{y_j}) - \psi^y(\mathbf{v}^j_{\nu,\,\mu+\frac{1}{2}})}{\xi_{y_j}} \boldsymbol{\ell}_{y_j}, \ \ \xi_{y_j} := \left\langle \boldsymbol{\ell}_{y_j}, \Delta\mathbf{v}_{\nu,\,\mu+\frac{1}{2}} \right\rangle,$$

where the right and left eigensystems $\{\mathbf{r}_{x_j}\}_{j=1}^3$ $\{\mathbf{r}_{y_j}\}_{j=1}^3$ and $\{\boldsymbol{\ell}_{x_j}\}_{j=1}^3$ $\{\boldsymbol{\ell}_{y_j}\}_{j=1}^3$ are constructed as the precise mirror images of the Roe-paths in $\mathbf{v}$-space,

$$\mathbf{r}_j^x := [H]^{-1}_{\nu+\frac{1}{2},\,\mu}\widehat{\mathbf{r}}_{x_j}, \qquad \boldsymbol{\ell}_j^x := [H]_{\nu+\frac{1}{2},\,\mu}\widehat{\boldsymbol{\ell}}_{x_j}, \qquad j=1,2,3$$
$$\mathbf{r}_j^y := [H]^{-1}_{\nu,\,\mu+\frac{1}{2}}\widehat{\mathbf{r}}_{y_j}, \qquad \boldsymbol{\ell}_j^y := [H]_{\nu,\,\mu+\frac{1}{2}}\widehat{\boldsymbol{\ell}}_{y_j}, \qquad j=1,2,3$$

where $[H]_{\nu+\frac{1}{2},\,\mu}$ and $[H]_{\nu,\,\mu+\frac{1}{2}}$ denote the averaged symmetrizers such that $\Delta\mathbf{u}_{\nu+\frac{1}{2},\,\mu} = [H]_{\nu+\frac{1}{2},\,\mu}\Delta\mathbf{v}_{\nu+\frac{1}{2},\,\mu}$ and $\Delta\mathbf{u}_{\nu,\,\mu+\frac{1}{2}} = [H]_{\nu,\,\mu+\frac{1}{2}}\Delta\mathbf{v}_{\nu,\,\mu+\frac{1}{2}}$.

We summarize our main result on 2D shallow water equations in the following theorem.

Theorem 5.1. *Let $E = (gh^2 + u^2h + v^2h)/2$ be the total energy of the 2D shallow water equations (5.1). Then, the semi-discrete approximation (5.2a) with entropy conservative fluxes $\mathbf{f}^*_{\nu+\frac{1}{2},\,\mu}$ and $\mathbf{g}^*_{\nu,\,\mu+\frac{1}{2}}$ given in (5.2b), (5.2c), (5.3), is energy stable, and the following discrete energy balance is satisfied,*

$$\frac{d}{dt}\sum_{\nu,\,\mu} E(\mathbf{u}_{\nu,\,\mu}(t))\Delta x\Delta y = -\zeta\sum_{\nu,\,\mu}\left\{\widehat{h}_{\nu+\frac{1}{2},\,\mu}\left[\left(\frac{\Delta u_{\nu+\frac{1}{2},\,\mu}}{\Delta x}\right)^2 + \left(\frac{\Delta v_{\nu+\frac{1}{2},\,\mu}}{\Delta x}\right)^2\right]\right.$$
$$\left. + \widehat{h}_{\nu,\,\mu+\frac{1}{2}}\left[\left(\frac{\Delta u_{\nu,\,\mu+\frac{1}{2}}}{\Delta y}\right)^2 + \left(\frac{\Delta v_{\nu,\,\mu+\frac{1}{2}}}{\Delta y}\right)^2\right]\right\}\Delta x\Delta y. \quad (5.4)$$

Observe that no artificial viscosity is introduced in the sense that the energy dissipation statement (5.4) is the precise discrete analogue of the energy balance statement (2.2).

Proof. Multiply (5.2a) by $[U_{\mathbf{u}}]^\top_{\nu,\,\mu} = \mathbf{v}^\top_{\nu,\,\mu}$, and sum up all spatial cells to get the balance of the total entropy,

$$\frac{d}{dt}\sum_{\nu,\,\mu} E(\mathbf{u}_{\nu,\,\mu}(t))\Delta x\Delta y + \sum_{\nu,\,\mu}\left\langle \mathbf{v}_{\nu,\,\mu}, \mathbf{f}^*_{\nu+\frac{1}{2},\,\mu} - \mathbf{f}^*_{\nu-\frac{1}{2},\mu}\right\rangle \Delta y$$
$$+ \sum_{\nu,\,\mu}\left\langle \mathbf{v}_{\nu,\,\mu}, \mathbf{g}^*_{\nu,\,\mu+\frac{1}{2}} - \mathbf{g}^*_{\nu,\mu-\frac{1}{2}}\right\rangle \Delta x$$
$$= \zeta\sum_{\nu,\,\mu}\left\langle \mathbf{v}_{\nu,\,\mu}, \widehat{h}_{\nu+\frac{1}{2},\,\mu}\Delta\mathbf{d}_{\nu+\frac{1}{2},\,\mu} - \widehat{h}_{\nu-\frac{1}{2},\mu}\Delta\mathbf{d}_{\nu-\frac{1}{2},\mu}\right\rangle \frac{\Delta y}{\Delta x}$$
$$+ \zeta\sum_{\nu,\,\mu}\left\langle \mathbf{v}_{\nu,\,\mu}, \widehat{h}_{\nu,\,\mu+\frac{1}{2}}\Delta\mathbf{d}_{\nu,\,\mu+\frac{1}{2}} - \widehat{h}_{\nu,\mu-\frac{1}{2}}\Delta\mathbf{d}_{\nu,\mu-\frac{1}{2}}\right\rangle \frac{\Delta x}{\Delta y} \quad (5.5)$$

Since the numerical fluxes $\mathbf{f}^*_{\nu+\frac{1}{2},\,\mu}$ and $\mathbf{g}^*_{\nu,\,\mu+\frac{1}{2}}$ are chosen as the entropy conservative fluxes in x and y directions respectively, they satisfy the entropy conservative requirement (3.10a), so that their $\mathbf{v}$-moments on the left of (5.5) amount to perfect differences,

$$\left\langle \mathbf{v}_{\nu,\,\mu},\, \mathbf{f}^*_{\nu+\frac{1}{2},\,\mu} - \mathbf{f}^*_{\nu-\frac{1}{2},\,\mu}\right\rangle = F_{\nu+\frac{1}{2},\,\mu} - F_{\nu-\frac{1}{2},\,\mu}, \quad (5.6a)$$
$$\left\langle \mathbf{v}_{\nu,\,\mu},\, \mathbf{g}^*_{\nu,\,\mu+\frac{1}{2}} - \mathbf{g}^*_{\nu,\,\mu-\frac{1}{2}}\right\rangle = G_{\nu,\,\mu+\frac{1}{2}} - G_{\nu,\,\mu-\frac{1}{2}}, \quad (5.6b)$$

with consistent entropy fluxes given by (consult (3.10b)),

$$2F_{\nu+\frac{1}{2},\,\mu} = \left\langle (\mathbf{v}_{\nu,\,\mu} + \mathbf{v}_{\nu+1,\,\mu}),\, \mathbf{f}^*_{\nu+\frac{1}{2},\,\mu} \right\rangle - (\psi^x(\mathbf{v}_{\nu,\,\mu}) + \psi^x(\mathbf{v}_{\nu+1,\,\mu}))$$
$$2G_{\nu,\,\mu+\frac{1}{2}} = \left\langle (\mathbf{v}_{\nu,\,\mu} + \mathbf{v}_{\nu,\,\mu+1}),\, \mathbf{g}^*_{\nu,\,\mu+\frac{1}{2}} \right\rangle - (\psi^y(\mathbf{v}_{\nu,\,\mu}) + \psi^y(\mathbf{v}_{\nu,\,\mu+1}))\,.$$

On the other hand, summation by parts and explicit computation using the entropy variable (2.14c) on the RHS of (5.5) yield

$$\begin{aligned}
&\zeta \sum_{\nu,\,\mu} \left\langle \mathbf{v}_{\nu,\,\mu}, \widehat{h}_{\nu+\frac{1}{2},\,\mu} \Delta \mathbf{d}_{\nu+\frac{1}{2},\,\mu} \; -\widehat{h}_{\nu-\frac{1}{2},\mu} \Delta \mathbf{d}_{\nu-\frac{1}{2},\mu} \right\rangle \frac{\Delta y}{\Delta x} \\
&\qquad = -\zeta \sum_{\nu,\,\mu} \left\langle \Delta \mathbf{v}_{\nu+\frac{1}{2},\,\mu}, \widehat{h}_{\nu+\frac{1}{2},\,\mu} \Delta \mathbf{d}_{\nu+\frac{1}{2},\,\mu} \right\rangle \frac{\Delta y}{\Delta x} \\
&\qquad = -\zeta \sum_{\nu,\,\mu} \left[\frac{1}{(\Delta x)^2} \widehat{h}_{\nu+\frac{1}{2},\,\mu} \left(\left(\Delta u_{\nu+\frac{1}{2},\,\mu} \right)^2 + \left(\Delta v_{\nu+\frac{1}{2},\,\mu} \right)^2 \right) \right] \Delta x \Delta y
\end{aligned} \tag{5.7a}$$

$$\begin{aligned}
&\zeta \sum_{\nu,\,\mu} \left\langle \mathbf{v}_{\nu,\,\mu}, \widehat{h}_{\nu,\,\mu+\frac{1}{2}} \Delta \mathbf{d}_{\nu,\,\mu+\frac{1}{2}} \; -\widehat{h}_{\nu,\mu-\frac{1}{2}} \Delta \mathbf{d}_{\nu,\mu-\frac{1}{2}} \right\rangle \frac{\Delta x}{\Delta y} \\
&\qquad = -\zeta \sum_{\nu,\,\mu} \left\langle \Delta \mathbf{v}_{\nu,\,\mu+\frac{1}{2}}, \widehat{h}_{\nu,\,\mu+\frac{1}{2}} \Delta \mathbf{d}_{\nu,\,\mu+\frac{1}{2}} \right\rangle \frac{\Delta x}{\Delta y} \\
&\qquad = -\zeta \sum_{\nu,\,\mu} \left[\frac{1}{(\Delta y)^2} \widehat{h}_{\nu,\,\mu+\frac{1}{2}} \left(\left(\Delta u_{\nu,\,\mu+\frac{1}{2}} \right)^2 + \left(\Delta v_{\nu,\,\mu+\frac{1}{2}} \right)^2 \right) \right] \Delta x \Delta y
\end{aligned} \tag{5.7b}$$

By (5.6) and (5.7), the semi-discrete energy balance statement (5.4) now follows,

$$\begin{aligned}
\frac{d}{dt} \sum_{\nu,\,\mu} E(\mathbf{u}_{\nu,\,\mu}(t)) \Delta x \Delta y = -\zeta \sum_{\nu,\,\mu} \Bigg\{ & \widehat{h}_{\nu+\frac{1}{2},\,\mu} \left[\left(\frac{\Delta u_{\nu+\frac{1}{2},\,\mu}}{\Delta x} \right)^2 + \left(\frac{\Delta v_{\nu+\frac{1}{2},\,\mu}}{\Delta x} \right)^2 \right] \\
& + \widehat{h}_{\nu,\,\mu+\frac{1}{2}} \left[\left(\frac{\Delta u_{\nu,\,\mu+\frac{1}{2}}}{\Delta y} \right)^2 + \left(\frac{\Delta v_{\nu,\,\mu+\frac{1}{2}}}{\Delta y} \right)^2 \right] \Bigg\} \Delta x \Delta y. \quad \square
\end{aligned}$$

5.2 Energy Preserving Schemes

In the case that the eddy viscosity is absent, $\zeta = 0$, all the dissipation terms on the RHS of the difference scheme (5.2a) vanish,

$$\frac{d}{dt} \mathbf{u}_{\nu,\,\mu}(t) + \frac{1}{\Delta x} (\mathbf{f}^*_{\nu+\frac{1}{2},\,\mu} - \mathbf{f}^*_{\nu-\frac{1}{2},\,\mu}) + \frac{1}{\Delta y} (\mathbf{g}^*_{\nu,\,\mu+\frac{1}{2}} - \mathbf{g}^*_{\nu,\,\mu-\frac{1}{2}}) = 0. \tag{5.8}$$

The resulting scheme serves as an *energy preserving* approximation to the inviscid shallow water equations (1.2) with the discrete energy *equality*,

$$\frac{d}{dt}\sum_{\nu,\,\mu} E(\mathbf{u}_{\nu,\,\mu}(t))\Delta x\Delta y = 0.$$

Remark 5.2. We note that energy preserving semi-discrete scheme (5.2), (5.3) may allow a substantial increase of the potential *enstrophy*, $\frac{1}{2}\sum \eta^2_{\nu,\,\mu}/h_{\nu,\,\mu}$, especially for the flow over steep topography, due to spurious energy cascade into smaller scales, consult [AL77, AL81]. Here, η is the sum of the relative vorticity $v_x - u_y$ and the Coriolis parameter at that latitude. After a long term integration, a significant amount of energy is transferred into the smallest resolvable scales, where truncation error becomes relevant. It would be desirable to adapt our energy stable discretization to retain the additional conservation of enstrophy, advocated in [Ara97, AL81].

6 Numerical Experiments for 2D Shallow Water Equations

6.1 Boundary Conditions

The numerical treatment of boundaries is intended to be as physically relevant as possible. We describe two basic types of boundary conditions that are applicable to the two dimensional shallow water problems: the first type simulates a boundary at infinity or a transmissive boundary; the second type applies in the presence of solid fixed walls.

Transmissive Boundaries

These are cases in which boundaries are supposed to be transparent in the sense that waves are allowed to pass through. The inflow and outflow conditions need to be described, hence the method of characteristics in two dimension follows. The local value of the Froude number $Fr := V/\sqrt{gL}$ determines the flow regime and, accordingly, the number of boundary conditions to apply. Here V and L denote the characteristic velocity and length scales of the phenomenon, respectively. For subcritical flow, two external boundary conditions are required at inflow boundaries, whereas only one boundary condition is required at outflow boundaries. Two dimensional supercritical flow requires three inflow boundary conditions and no boundary condition at outflow boundaries where the flow is only influenced by the information coming from the interior nodes.

Reflective Boundaries

This is a particular case in which the flow is confined inside a fixed field by solid walls where we impose the reflective boundary conditions. Since our

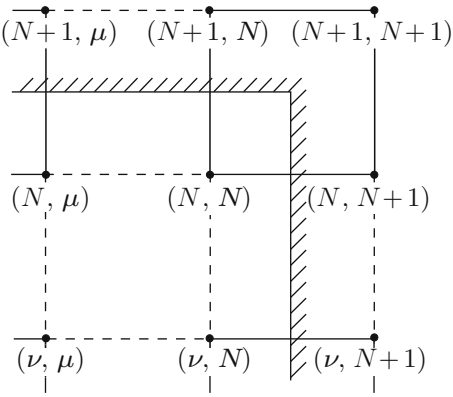


Fig. 3. Right-hand boundary

testing problems in next section are concerned with the flow in a square basin, without losing generality, we consider the computational domain in the upper-right corner with the solid boundaries along x and y-direction as shown in Fig. 3. By the three-point stencil used in our semi-discrete scheme, we try to impose the value of one computational grid point added outside boundary.

The reflection is incorporated by changing the sign of the normal component of the velocity, while the water depth is unaltered. The values at all the $(\nu,\, N+1)$ points on the right-hand side of the wall are replaced by the values at interior $(\nu,\, N)$ points and sign of the normal velocity component u is switched,

$$h_{\nu,\,N+1} = h_{\nu,\,N}, \quad u_{\nu,\,N+1} = -u_{\nu,\,N}, \quad v_{\nu,\,N+1} = v_{\nu,\,N};$$

the values at all the $(N+1,\, \mu)$ points on the top of the wall are replaced by the values at interior $(N,\, \mu)$ points and sign of the normal velocity component v is switched

$$h_{N+1,\,\mu} = h_{N,\,\mu}, \quad u_{N+1,\,\mu} = u_{N,\,\mu}, \quad v_{N+1,\,\mu} = -v_{N,\,\mu};$$

the values at all the $(N+1,\, N+1)$ point in the upper-right corner are given by

$$h_{\nu+1,\,N+1} = h_{\nu,\,\mu}, \quad u_{\nu+1,\,N+1} = -u_{\nu,\,\mu}, \quad v_{\nu+1,\,N+1} = -v_{\nu,\,\mu}.$$

6.2 Time Discretization

Similar to the time discretizations of the Burgers' equation, we integrate the entropy stable scheme (5.2) and (5.3) with the explicit three-stage Runge–Kutta method (4.12a) by its high-order accuracy, large stability region and simplicity.

$$\begin{cases} \mathbf{u}^{(1)} &= \mathbf{u}^n + \Delta t \mathcal{L}(\mathbf{u}^n) \\ \mathbf{u}^{(2)} &= \frac{3}{4}\mathbf{u}^n + \frac{1}{4}\mathbf{u}^{(1)} + \frac{1}{4}\Delta t \mathcal{L}(\mathbf{u}^{(1)}) \\ \mathbf{u}^{n+1} &= \frac{1}{3}\mathbf{u}^n + \frac{2}{3}\mathbf{u}^{(2)} + \frac{2}{3}\Delta t \mathcal{L}(\mathbf{u}^{(2)}) \end{cases} \tag{6.1a}$$

where

$$\begin{aligned} [\mathcal{L}(\mathbf{u})]_{\nu,\mu} = &-\frac{1}{\Delta x}(\mathbf{f}_{\nu+\frac{1}{2},\mu} - \mathbf{f}_{\nu-\frac{1}{2},\mu}) - \frac{1}{\Delta y}(\mathbf{g}_{\nu,\mu+\frac{1}{2}} - \mathbf{g}_{\nu,\mu-\frac{1}{2}}) \\ &+ \frac{\zeta}{\Delta x}(\widehat{h}_{\nu+\frac{1}{2},\mu}\frac{\mathbf{d}_{\nu+1,\mu} - \mathbf{d}_{\nu,\mu}}{\Delta x} - \widehat{h}_{\nu-\frac{1}{2},\mu}\frac{\mathbf{d}_{\nu,\mu} - \mathbf{d}_{\nu-1,\mu}}{\Delta x}) \\ &+ \frac{\zeta}{\Delta y}(\widehat{h}_{\nu,\mu+\frac{1}{2}}\frac{\mathbf{d}_{\nu,\mu+1} - \mathbf{d}_{\nu,\mu}}{\Delta x} - \widehat{h}_{\nu,\mu-\frac{1}{2}}\frac{\mathbf{d}_{\nu,\mu} - \mathbf{d}_{\nu,\mu-1}}{\Delta x}). \end{aligned} \tag{6.1b}$$

6.3 Numerical Results

We test our entropy-stable schemes with the two dimensional frictionless partial-dam-break problem originally studied by Fennema and Chaudhry in [FC90]. It imposes computational difficulties due to the discontinuous initial conditions. It also involves other computational issues like boundary treatments and positive-water-depth preserving solver.

As shown in Fig. 4, the simplified geometry of the problem consists of a $1,400 \times 1,400$ m^2 basin with a idealized dam in the middle. Water is limited by the fixed, solid, frictionless walls in this square basin. To prevent any damping by the source terms, a frictionless, horizontal bottom is used. All walls are assumed to be reflective. The initial water level of the dam is 10 m and the tail

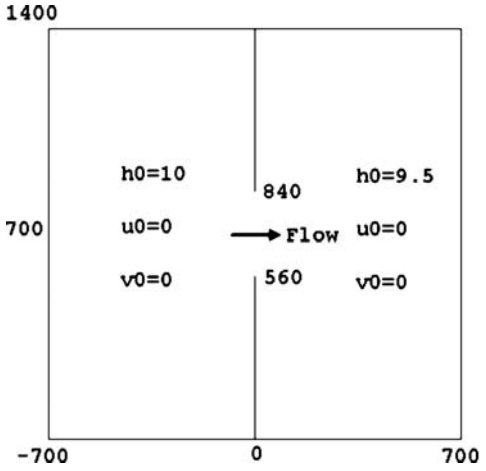


Fig. 4. Geometry configuration and initial setting of 2D Partial-Dam-Break problem

water is 9.5 m high. Central part of the dam is assumed to fail instantaneously or the gate in the middle of the dam is opened instantly. Water is released into the downstream side through a breach 280 m wide, located between $y = 560$ and $y = 840$, forming a wave that propagates while spreading laterally. A negative wave propagates upstream at the same time. For simplicity, the Coriolis force is ignored in the computation. The acceleration due to gravity is taken to be $9.8\,\mathrm{m\,s^{-2}}$. Although there is no analytical reference solution for this test problem, other numerical results of similar problems are available in [FC90, CK04].

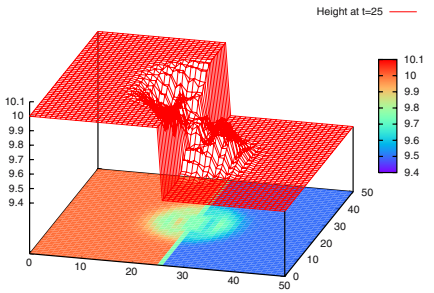
In the following figures, we display the numerical solutions for the fully discrete scheme (6.1a)–(6.1b) with the numerical fluxes (5.2b)–(5.2c). The sum of potential and kinetic energy serves as the generalized entropy function in the design of our numerical schemes,

$$E(\mathbf{u}) = \frac{gh^2 + u^2h + v^2h}{2}.$$

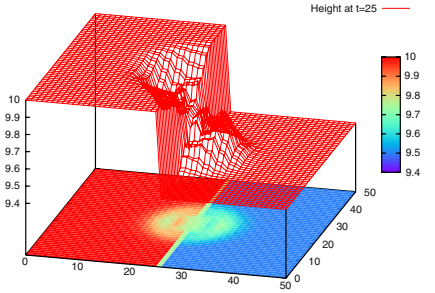
Uniform space and time grid sizes, $\Delta x = \Delta y$ and Δt are used. The computational model is run for up to 50 s after the dam broke when the water waves haven't reached the boundaries. Both inviscid and viscous cases are explored. For the viscous cases, the eddy viscosity is taken to be $10\,\mathrm{m^2 s^{-1}}$. We use different spatial resolutions for the same problem, and adjust time step according to the CFL condition.

We first solve the inviscid and viscous shallow water equations on the computational domain consisting of a 50×50 cell square grid with $\Delta x = \Delta y = 28$ m. We group our numerical results of inviscid shallow water equations along the left column of Fig. 5. For comparison, the results of viscous shallow water equations with eddy viscosity $\zeta = 10\,\mathrm{m^2 s^{-1}}$ are summarized on the right column. The first and second row of Fig. 5 depict the perspective plots of water surface profiles at t=25 s and t=50 s respectively. Remnants of the dam are represented by jumps near the middle of the plot. The vertical scale is exaggerated with respect to the horizontal scales. We observe that the numerical solutions of the water depth in Fig. 5a, c successfully simulate both the circular shock water wave propagations and the vortices formed on the both sides of the breach. The undershoots are also developed near sharp corners of the remanent dam. These steep degressions in the water surface are noticeable downstream of the breach at $t = 50$ s. Similar numerical tests were done in [CK04] by the second-order central-upwind schemes, which were originally proposed in [KT00].

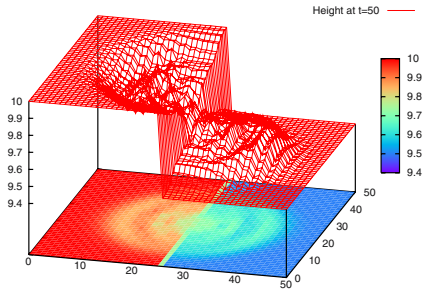
For the inviscid shallow water equations, dispersive errors of the numerical schemes, in the form of spurious oscillations in the mesh scale, are noticeable near the breach in Fig. 5a, c. For the viscous shallow water equations, as shown in Fig. 5b, d, the presence of eddy viscosity causes the oscillations to be dramatically reduced around the breach. In addition to eliminating the wiggles, the eddy viscosity terms also single out the undershoot near sharp corners of the remnants of dam without damping it.



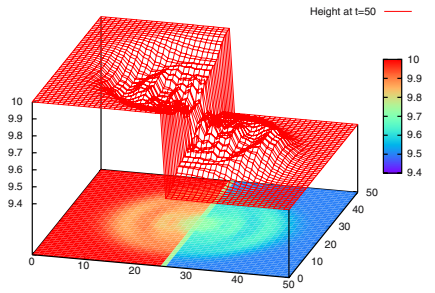
(a) Water depth at t=25 s, inviscid case



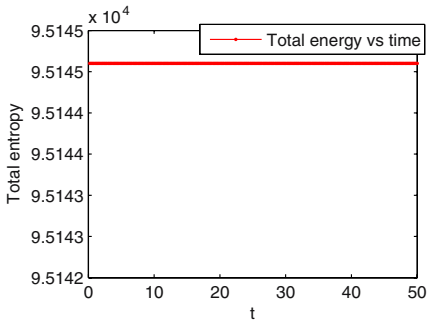
(b) Water depth at t=25 s, viscous case



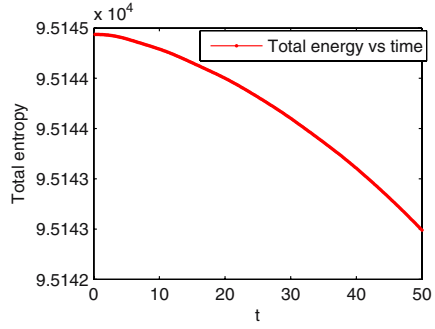
(c) Water depth at t=50 s, inviscid case



(d) Water depth at t=50 s, viscous case



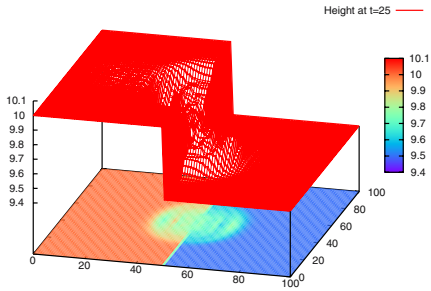
(e) Total energy v.s. time, inviscid case



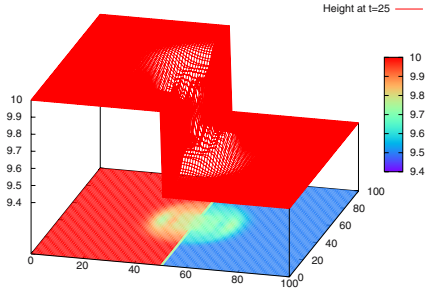
(f) Total energy v.s. time, viscous case

Fig. 5. Shallow water equations, $\zeta = 10\,\mathrm{m^2 s^{-1}}$, Dam-Break, $1{,}400 \times 1{,}400\,\mathrm{m}^2$ basin, reflective-slip boundary, $\Delta x = \Delta y = 28\,\mathrm{m}$, $\Delta t = 0.2\,\mathrm{s}$

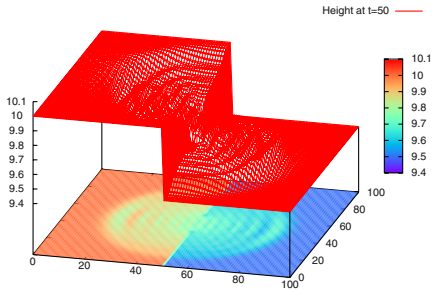
We display the total entropy scaled by 10^4 versus time in Fig. 5f. Compared with the same entropy plot of the inviscid problem in Fig. 5c, the plot of total energy in Fig. 5f reveals a $\mathcal{O}(1)$ energy decay due to the presence of eddy viscosity, while the negligible amount of energy decay introduced by RK3 time discretization for the inviscid shallow water equations is not detectable under the same scale in Fig. 5c.



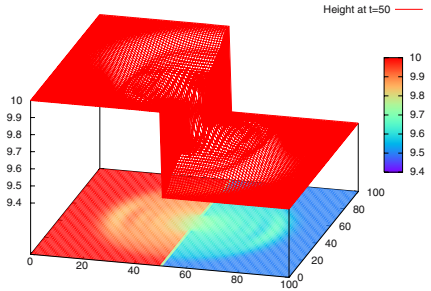
(a) Water depth at t=25 s, inviscid case



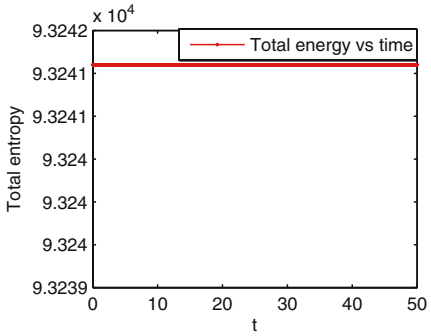
(b) Water depth at t=25 s, viscous case



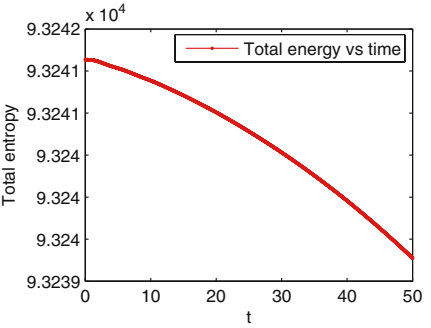
(c) Water depth at t=50 s, inviscid case



(d) Water depth at t=50 s, viscous case



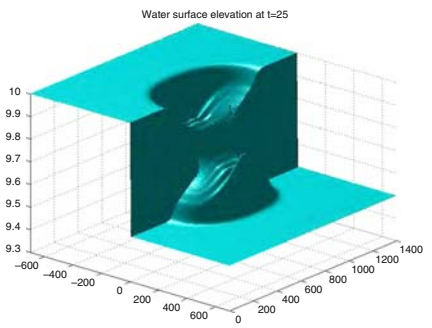
(e) Total energy v.s. time, inviscid case



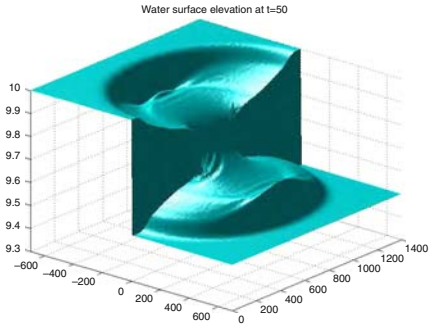
(f) Total energy v.s. time, viscous case

Fig. 6. Shallow water equations, $\zeta = 10\,\mathrm{m^2 s^{-1}}$, Dam-Break, $1{,}400 \times 1{,}400\,\mathrm{m^2}$ basin, reflective-slip boundary, $\Delta x = \Delta y = 14\,\mathrm{m}$, $\Delta t = 0.01\,\mathrm{s}$

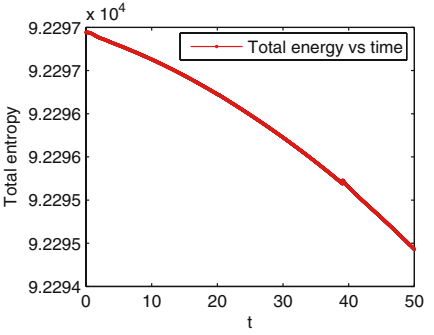
Next, in Fig. 6, we display the numerical solutions of the same problem in the refined spatial mesh with $\Delta x = \Delta y = 14\,\mathrm{m}$. Following the same pattern as in Fig. 5, Fig. 6 presents the perspective plots and total energy versus time. For the inviscid case, the profiles of the water elevation in Fig. 6a, c demonstrate smoother numerical solutions due to the decrease of the grid size, while the spurious oscillations in the mesh scale are still detectable near the breach



(a) Water depth at t=25 s, viscous case



(b) Water depth at t=50 s, viscous case



(c) Total energy v.s. time

Fig. 7. Viscous shallow water equations, $\zeta = 10\,\mathrm{m^2 s^{-1}}$, Dam-Break, $1{,}400 \times 1{,}400\,\mathrm{m^2}$ basin, reflective-slip boundary, $\Delta x = \Delta y = 7\,\mathrm{m}$, $\Delta t = 0.002\,\mathrm{s}$

because of the energy-preserving shallow water solver with the increase of the total enstrophy. For the viscous case with $\zeta = 10\,\mathrm{m^2 s^{-1}}$, Fig. 6b, d show the smoother solutions than inviscid solutions in Fig. 6a, c. The amplitude of those wiggles near the breach are significantly reduced though they are still detectable. Further refinement of the mesh from (100×100) to (200×200) generates very smooth solutions of the water depth h in Fig. 7a, b, when the oscillations are limited in the very small mesh scale.

Acknowledgments

Research was supported by NSF grants DMS04-07704, DMS07-07959 and by ONR grant N00014-91-J-1076.

References

[AL77] Akio Arakawa and Vivian R. Lamb. Computational design of the basic dynamical process of the ucla general circulation model. *Meth. Comput. Phys.*, 17:173–265, 1977.

[AL81] Akio Arakawa and Vivian R. Lamb. A potential enstrophy and energy conserving scheme for the shallow water equations. *Mont. Weat. Rev.*, 109:18–36, 1981.

[Ara97] Akio Arakawa. Computational design for long-term numerical integration of the equations of fluid motion: two-dimensional incompressible flow. I [J. Comput. Phys. **1** (1966), no. 1, 119–143]. *J. Comput. Phys.*, 135(2):101–114, 1997. With an introduction by Douglas K. Lilly, Commemoration of the 30th anniversary {of J. Comput. Phys.}.

[CK04] Alina Chertock and Alexander Kurganov. On a hybrid finite-volume-particle method. *M2AN Math. Model. Numer. Anal.*, 38(6):1071–1091, 2004.

[Daf00] Constantine M. Dafermos. *Hyperbolic conservation laws in continuum physics*, volume 325 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 2000.

[EO80] Bjorn Engquist and Stanley Osher. Stable and entropy satisfying approximations for transonic flow calculations. *Math. Comp.*, 34(149):45–75, 1980.

[FC90] Robert J. Fennema and M. Hanif Chaudhry. Explicit methods for 2d transient free-surface flows. *J. Hydraul. Eng. ASCE*, 116(8):1013–1034, 1990.

[GL88] Jonathan Goodman and Peter D. Lax. On dispersive difference schemes. I. *Comm. Pure Appl. Math.*, 41(5):591–613, 1988.

[Gla87] Paul Glaister. Difference schemes for the shallow water equations. Numerical Analysis Report 9/87, University of Reading, Department of Mathematics, 1987.

[God61] Sergei K. Godunov. An interesting class of quasi-linear systems. *Dokl. Akad. Nauk SSSR*, 139:521–523, 1961.

[GST01] Sigal Gottlieb, Chi-Wang Shu, and Eitan Tadmor. Strong stability-preserving high-order time discretization methods. *SIAM Rev.*, 43(1):89–112 (electronic), 2001.

[HL91] Thomas Y. Hou and Peter D. Lax. Dispersive approximations in fluid dynamics. *Comm. Pure Appl. Math.*, 44(1):1–40, 1991.

[Kru70] Stanislav N. Kružkov. First order quasilinear equations with several independent variables. *Mat. Sb. (N.S.)*, 81 (123):228–255, 1970.

[KT00] Alexander Kurganov and Eitan Tadmor. New high-resolution central schemes for nonlinear conservation laws and convection-diffusion equations. *J. Comput. Phys.*, 160(1):241–282, 2000.

[Lax71] Peter D. Lax. *Shock waves and entropy.* Acaemic Press, New York, 1971. Contributions to Nonlinear Functional Analysis, (E.A.Zarantonello, ed.).

[Lax73] Peter D. Lax. *Hyperbolic systems of conservation laws and the mathematical theory of shock waves.* Society for Industrial and Applied Mathematics, Philadelphia, Pa., 1973. Conference Board of the Mathematical Sciences Regional Conference Series in Applied Mathematics, No. 11.

[LL96] C. David Levermore and Jian-Guo Liu. Large oscillations arising in a dispersive numerical scheme. *Phys. D*, 99(2-3):191–216, 1996.

[LMR02] Philippe G. Lefloch, J. M. Mercier, and C. Rohde. Fully discrete, entropy conservative schemes of arbitrary order. *SIAM J. Numer. Anal.*, 40(5):1968–1992 (electronic), 2002.

[Moc80] Michael S. Mock. Systems of conservation laws of mixed type. *J. Differential Equations*, 37(1):70–88, 1980.

[Roe81] Phil L. Roe. Approximate Riemann solvers, parameter vectors, and difference schemes. *J. Comput. Phys.*, 43(2):357–372, 1981.

[Ser99] Denis Serre. *Systems of Conservation Laws, 1: Hyperbolicity, Entropies, Shock Waves*. Cambridge University Press, 1999.

[Tad87] Eitan Tadmor. The numerical viscosity of entropy stable schemes for systems of conservation laws. I. *Math. Comp.*, 49(179):91–103, 1987.

[Tad03] Eitan Tadmor. Entropy stability theory for difference approximations of nonlinear conservation laws and related time-dependent problems. *Acta Numer.*, 12:451–512, 2003.

[TZ06] Eitan Tadmor and Weigang Zhong. Entropy stable approximations of Navier-Stokes equations with no artificial numerical viscosity. *J. Hyperbolic Differ. Equ.*, 3(3):529–559, 2006.

A Conjecture about Molecular Dynamics

P.F. Tupper

Department of Mathematics and Statistics, McGill University, 805 Sherbrooke Street West, Montreal, QC, Canada H3A 2K6, tupper@math.mcgill.ca

Summary. An open problem in numerical analysis is to explain why molecular dynamics works. The difficulty is that numerical trajectories are only accurate for very short times, whereas the simulations are performed over long time intervals. It is believed that statistical information from these simulations is accurate, but no one has offered a rigourous proof of this. In order to give mathematicians a clear goal in understanding this problem, we state a precise mathematical conjecture about molecular dynamics simulation of a particular system. We believe that if the conjecture is proved, we will then understand why molecular dynamics works.

1 Introduction

Molecular dynamics is the computer simulation of a material at the atomic level. In principle the only inputs to a simulation are the characteristics of a set of particles and a description of the forces between them. An initial condition is chosen and from these first principles the evolution of the system in time is simulated using Newton's laws and a simple numerical integrator [6, 1].

Molecular dynamics is a very prevalent computational practice, as a glance at an issue of the Journal of Chemical Physics will show. It does have its limitations: the motion of only a relatively small number of particles can be simulated over a short time interval. However, most of the mesoscopic models that have been suggested to overcome these difficulties still rely on molecular dynamics as a form of calibration. It is likely that molecular dynamics will continue to be important in the future.

Given its scientific importance there is very little rigourous justification of molecular dynamics simulation. From the viewpoint of numerical analysis it is surprising that it works at all. The problem is that individual trajectories computed by molecular dynamics simulations are accurate for only small time intervals. As we will see in Sect. 3, numerical trajectories diverge rapidly from true trajectories given the step-lengths used in practice. No one disputes this fact, and no one is particularly concerned with it either. The reason is that

practitioners are never interested in particular trajectories to begin with. They are interested in ensembles of trajectories. As long as the numerical trajectories are representative of a particular ensemble of true trajectories, researchers are content. However, that this statistical information is computed accurately has yet to be rigourously demonstrated in representative cases.

The goal of this article is to present a concise mathematical conjecture that encapsulates this fundamental difficulty. We present a model system that is representative of systems commonly simulated in molecular dynamics. We present the results of numerical simulations of this system using the Störmer–Verlet method, the work-horse of molecular dynamics. In each simulation a random initial condition is generated, an approximate trajectory for the system is computed and the net displacement of one particle over the duration of the simulation is recorded. We show that even for step-sizes that are far too large to accurately compute the position of the particle, the distribution of the particle's displacement over the many initial conditions appears to be accurate. From the numerical data we conjecture a rate of convergence for this particular statistical property. We believe that if this conjectured rate of convergence (or one like it) can be rigourously established, even for this single system, then we will understand significantly better why molecular dynamics works.

The problem of explaining the accuracy of molecular dynamics simulation is well-known both in the physical sciences (for example [6, p. 81]) and in the mathematics community [12]. This latter reference is a survey of the relation between computation and statistics for initial value problems in general. There has been plenty of excellent mathematical work that has done much to explain various features of this type of simulation, but has not resolved the issue we consider here. See [13, 14, 15] for surveys.

One body of work that has addressed the statistical accuracy of under-resolved trajectories in a special case is by Stuart and co-workers. In [3, 17] they have explored some linear test systems with provable statistical properties in the limit of large numbers of particles. They are able to show that if the systems are simulated with appropriate methods the statistical features of numerical trajectories are accurate in the same limit even when the step-lengths are too large to resolve trajectories. Though these results are interesting since they are the only ones of their kind now known, for the highly non-linear problems of practical molecular dynamics very different arguments will be required.

One subproblem that has been attacked more successfully is that of the computation of ergodic averages. These are averages of functions along very long trajectories. All that numerical trajectories have to do to get these correct is sample the entire phase space evenly. This is a much weaker property than getting all statistical features correct. The most striking work on this question is by Reich [11] which establishes rapid convergence of ergodic averages for Hamiltonian systems which are uniformly hyperbolic on sets of constant energy. Unfortunately, this property has never been established for realistic

systems, and is unlikely to hold for them [9, 10]. The work [18] established similar results for systems with much weaker properties but requires radically small time steps for convergence to occur.

The contribution of this work is to precisely specify a simple problem which encapsulates all the essential difficulties of the more general problem. In Sect. 2 we present the system we will study. Section 3 shows the results of some numerical experiments on this system. There we state our conjecture based on the results. In Sect. 4 we will discuss two possible approaches to proving the conjecture. Finally, in Sect. 5 we will discuss prospects for the eventual resolution of the conjecture.

2 The System

The system consists of $n = 100$ point particles interacting on an 11.5 by 11.5 square periodic domain. We let $q \in \mathbb{T}^{2n}$ and $p \in \mathbb{R}^{2n}$ denote the positions and velocities of the particles, with $q_i \in \mathbb{T}^2, p_i \in \mathbb{R}^2$ denoting the position and velocity of particle i. The motion of the system is described by a system of Hamiltonian differential equations:

$$\frac{dq}{dt} = \frac{\partial H}{\partial p}, \quad \frac{dp}{dt} = -\frac{\partial H}{\partial q},$$

with Hamiltonian

$$H(q,p) = \frac{1}{2}\|p\|_2^2 + \sum_{i<j} V_{LJ}(\|q^i - q^j\|).$$

Here V_{LJ} denotes the famous Lennard-Jones potential. In our simulations we use a truncated version:

$$V_{LJ}(r) = \begin{cases} 4\left(\frac{1}{r^{12}} - \frac{1}{r^6}\right), & \text{if } r \le r_{\text{cutoff}}, \\ 0, & \text{otherwise.} \end{cases}$$

Figure 1 shows the positions of the particles on the periodic domain for one state of the system. Though the particles are only points, in the figure each is represented by a circle of radius 1/2.

We take our initial conditions q^0, p^0 to be randomly distributed according to the probability density function

$$Z^{-1} e^{-H(q,p)/k\mathcal{T}}, \tag{1}$$

where Z is chosen so that the function integrates to one. This is known as the canonical distribution (or ensemble) for the system at temperature $\mathcal{T}$. There is a simple physical interpretation of this distribution: if the system is weakly connected to another very large system at temperature $\mathcal{T}$, this is the distribution we will find the original system in after a long period of time. In our units $k = 1$, and we choose $\mathcal{T} = 1$.

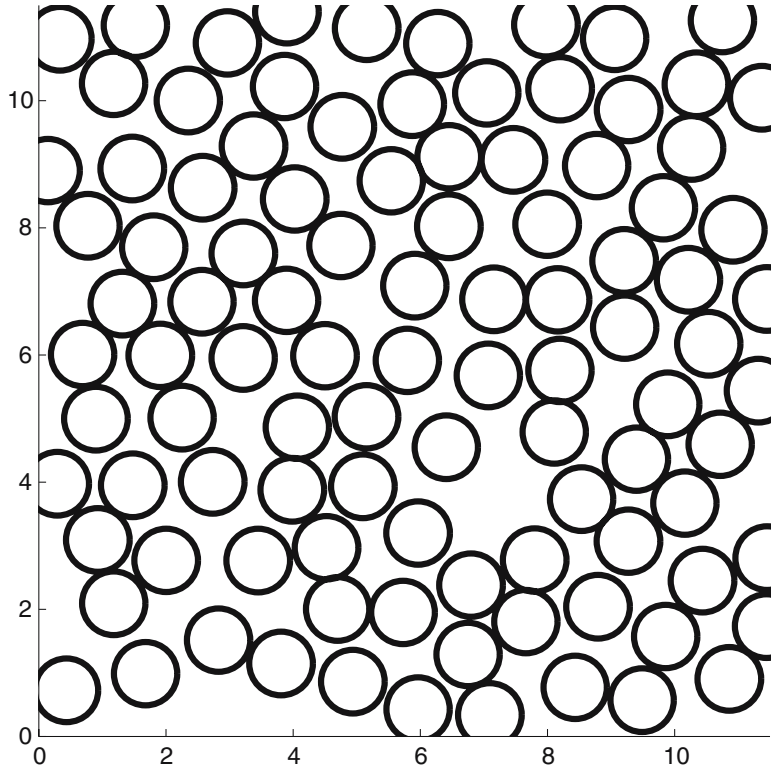


Fig. 1. The positions of the particles for a representative state of the system

There are many ways of sampling from the canonical distribution at a given temperature. For our experiments we generated initial conditions using Langevin dynamics. See [4] for an explanation of this technique and a comparison with other methods. If done correctly, the precise method of sampling from the canonical distribution will have no bearing on the results of the experiments we will present subsequently.

The numerical method we use for integrating our system is the Störmer–Verlet scheme. Given an initial q_0, p_0 and a $\Delta t > 0$ it generates a sequence of states $q_n, p_n, n \geq 0$ such that $(q_n, p_n) \approx (q(n\Delta t), p(n\Delta t))$. The version of the algorithm we use is

$$\begin{aligned} q_{n+1/2} &= q_n + p_n \Delta t/2, \\ p_{n+1} &= p_n - \Delta t \nabla V(q_{n+1/2}), \\ q_{n+1} &= q_n + p_{n+1} \Delta t/2. \end{aligned}$$

This is a second-order explicit method. It is symplectic, and as a consequence conserves phase space volume [7].

Finally we have to decide upon our step-length Δt. If Δt is too large the energy of the computed solution will increase rapidly and explode. In practice, it is observed that for small enough step lengths energy remains within a narrow band of the true energy for very long time intervals. (There is extensive theoretical justification for this phenomenon, see Sect. 4.1). Practitioners tend to pick a Δt as large as possible while still maintaining this long-term stability on their time interval of interest. For the system and initial conditions we describe here $\Delta t = 0.01$ yields good approximate energy conservation on the time interval $[0, 100]$. For our numerical experiments we will let Δt take this value and smaller. (The recommended value in [6], a standard reference, for this type of system is $\Delta t = 0.005$.)

3 The Problem

We will first examine how well trajectories are computed with $\Delta t = 0.01$. Figure 2 shows the computed x-position of one particle versus time for the same initial conditions and for a range of step-lengths. If the trajectory computed by Störmer–Verlet is accurate over the time interval $[0, 5]$, we expect that reducing the time step by a factor of a thousand would not yield a significantly different curve. However, we see that the two curves for $\Delta t = 0.01$

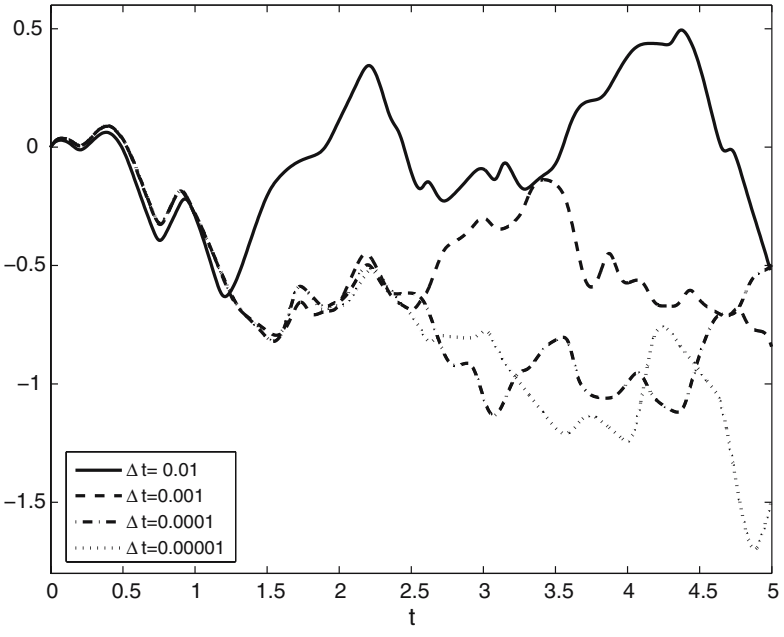


Fig. 2. Computed x-position of one particle versus time for fixed initial conditions for a range of Δt

and $\Delta t = 0.00001$ very quickly diverge. They are distinguishable to the eye almost immediately and completely diverge around 1.2 time units.

Reducing the step length to $\Delta t = 0.001$ gives a curve that agrees with the $\Delta t = 0.00001$ line longer, but still diverges around 2.5 time units. Similarly, even with $\Delta t = 0.0001$ trajectory is not accurate over the whole interval depicted.

From these numerical results, we might conjecture that reducing the step-length by a constant factor only extends the duration for which the simulation is accurate by a constant amount of time. This is consistent with theoretical results about the convergence of numerical methods for ordinary differential equations. What is surprising in this case is that the time-scale on which the trajectories are valid appears to be miniscule compared to the time-scale on which computation are actually performed. It seems that the trajectories we compute here with stepsize even as small as $\Delta t = 0.00001$ are not accurate over the whole interval $[0, 5]$ let alone over considerably longer intervals.

Fortunately we almost never care about what one particular trajectory is doing in molecular dynamics. We only care about statistical features of the trajectories when initial conditions are selected according to some probability distribution. Here we will consider the example of self-diffusion. Self-diffusion is the diffusion of one particular particle through a bath of identical particles. We can imagine somehow marking one particle at time zero and watching its motion through the system. This single-particle trajectory will depend on the positions and velocities of all the particles (including itself) at time zero. Since these are random, the trajectory of the single particle is random.

One way to measure self-diffusion is to look at the distribution of the x-coordinate of the tracer particle relative to its initial condition. To estimate this, we generate many random initial conditions, perform the simulation using the Störmer–Verlet method, and record the net displacement of the particle in the given direction. Figure 3 show the histograms of these displacements at time $T = 10$ for three different step-lengths.

In contrast to the case where we examined single trajectories, here the histograms are virtually identical for the different step-lengths. This suggests that any information we glean from the first histogram will be accurate.

To check this more carefully, we compute the variance of the total displacement at various times T for varying step-lengths. Let $R(T) = \|q_1(T) - q_1(0)\|$ denote the total displacement of the particle after time T. This is a random quantity through its dependence on the state of the system at $t = 0$. Let $R_{\Delta t}(T)$ denote this same displacement as simulated with the Störmer–Verlet method. This also is a random quantity. Now define $\langle R^2_{\Delta t}(T)\rangle$ to be the expected value of $R^2_{\Delta t}(T)$ when the initial conditions are chosen according to the canonical distribution. Let us see how this last quantity depends on Δt. We do this by generating many initial conditions from the canonical ensemble and then simulating the system for 100 time units, keeping track of the total displacement of the tracer particle.

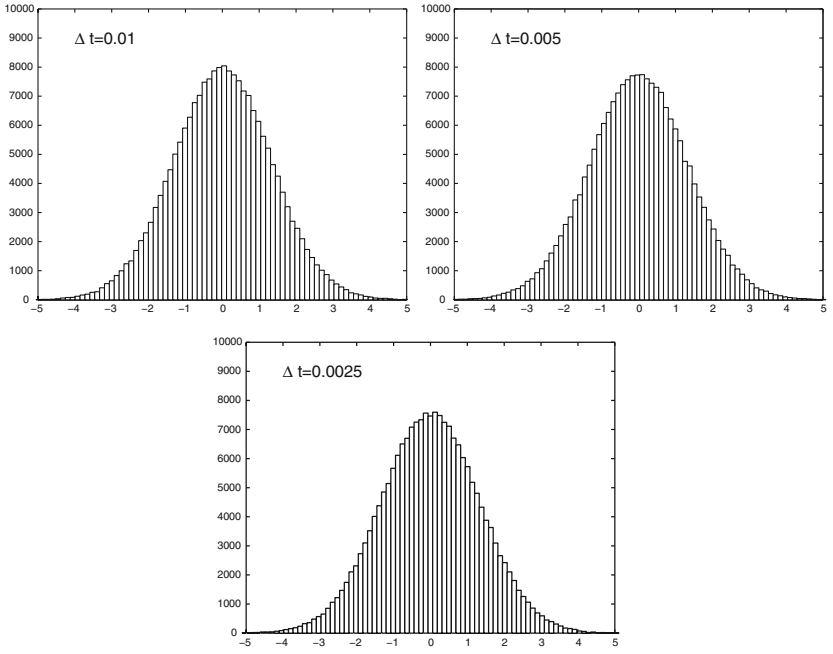


Fig. 3. Displacement in x direction of 1 particle at $T = 10$ for three different step-lengths

Figure 4 shows $\langle R^2_{\Delta t}(T)\rangle$ versus T for three choices of step length. The inset shows a subset of the data with error bars. Up to the sampling error there is no difference between the curves. As far as we can tell from this plot, the answers for $\Delta t = 0.01$ are accurate. The time-scale is much larger than the short interval we found the trajectory to be accurate over. Lest we give the impression that $\langle R^2_{\Delta t}(T)\rangle$ depends linearly on T, Fig. 5 shows the same results for a smaller time interval.

We conjecture that the reason $\langle R^2_{\Delta t}(T)\rangle$ does not appear to depend on Δt is that even for these large values of Δt it closely matches $\langle R^2(T)\rangle$. It is not clear at all what the rate of convergence of $R_{\Delta t}(T)$ to $R(T)$ is and how it depends on T. However we make the following conjecture:

Conjecture 1 *For the system described in Sect. 2 with the initial distribution given by (1) and the Störmer–Verlet integrator with time step Δt*

$$\left|\langle R^2_{\Delta t}(T)\rangle - \langle R^2(T)\rangle\right| \leq C\Delta t^2,$$

for all $T \in [0, Ae^{B/\Delta t}]$, for some constants A, B, C.

We will explain the reasons for hypothesizing this particular dependence in the next section. Here we will briefly note what dependence the classical theory

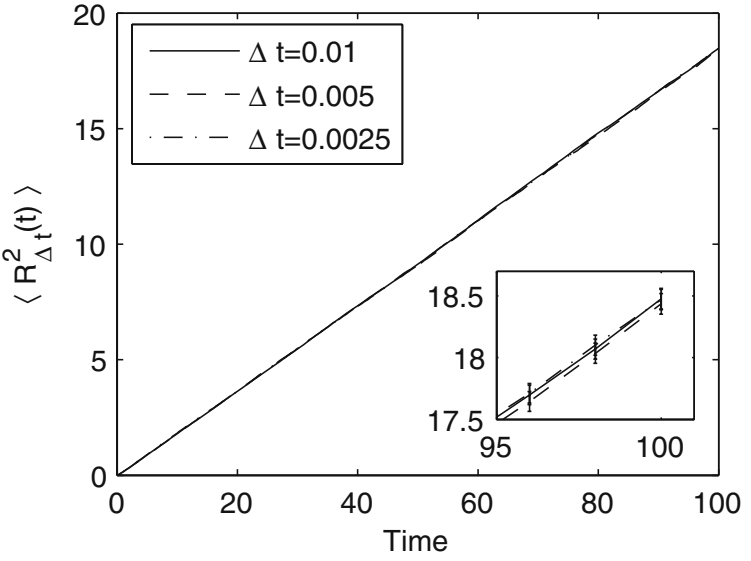


Fig. 4. Expected squared total displacement in the x direction of a single particle as a function of time for three different step-lengths

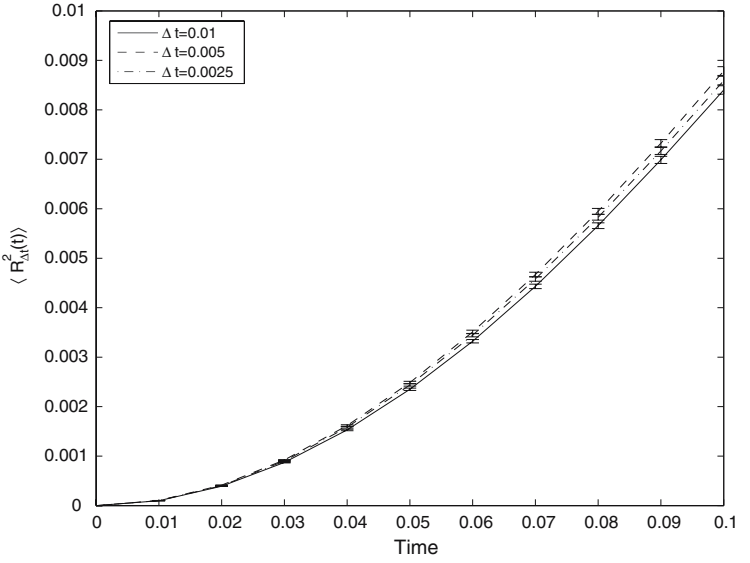


Fig. 5. Same as Fig. 4 but on a smaller time interval

of convergence for numerical ODEs gives:

$$\left|\langle R^2_{\Delta t}(T)\rangle - \langle R^2(T)\rangle\right| \leq Ce^{LT}\Delta t^2$$

for $T \in [0, E\log(F/\Delta t)]$ for sufficiently small Δt for some $C, L, E, F > 0$. (See [16, p. 239], for example.) So we need to explain why the error remains so small even for long simulations.

4 Two Approaches

We will discuss two possible approaches to proving Conjecture 1: backward error analysis and shadowing.

4.1 Backward Error Analysis

Typically a pth order numerical method applied to a system of ODEs computes a trajectory that is $\mathcal{O}(\Delta t^p)$ close to the exact trajectory on a finite interval. Backward error analysis is a way of showing that the numerical trajectory is an $\mathcal{O}(\exp(-1/\Delta t))$ approximation to the exact trajectory of a perturbed system. This result can be used in turn to prove results about the stability of the numerical trajectory. See [2] for an early reference and [7, Chap. IX.] for a recent comprehensive treatment of the subject.

If we apply a symplectic integrator to a Hamiltonian system it turns out that the modified system is also Hamiltonian. The Hamiltonian function $\widetilde{H}$ for the new system can be written as $\widetilde{H} = H + \mathcal{O}(\Delta t^2)$. There are two consequences for us. Firstly, the numerical method agrees very closely with the exact solutions of the modified Hamiltonian on short time intervals. If we denote the solution to the modified system with the same initial conditions by $(\tilde{q}, \tilde{p})$ then

$$|\tilde{q}(n\Delta t) - q^n| \leq Ce^{-D/\Delta t} \tag{1}$$

for $T \in [0, B/\Delta t]$, for some appropriate constants [5]. (This alone is not useful for analysing molecular dynamics since T and Δt are both large.) Secondly, the modified Hamiltonian $\widetilde{H}$ is conserved extremely well by the numerical method for long time intervals:

$$\left|\widetilde{H}(q^0, p^0) - \widetilde{H}(q^n, p^n)\right| \leq Ce^{-D/\Delta t},$$

for $n\Delta t \in [0, Ae^{B/\Delta t}]$. Putting this together with $\widetilde{H} = H + \mathcal{O}(\Delta t^2)$ gives

$$\left|H(q^0, p^0) - H(q^n, p^n)\right| \leq E\Delta t^2,$$

for $n\Delta t \in [0, Ae^{B/\Delta t}]$. We chose the bound in Conjecture 1 in analogy with this last result.

Suppose we wanted to bound the error between $\langle R^2_{\Delta t}(t)\rangle$ and $\langle R^2(t)\rangle$ using these estimates. The fact that the initial conditions are random adds an extra level of complication to the problem. We have been using $\langle\cdot\rangle$ to denote the average with respect to the canonical distribution for the Hamiltonian H. The perturbed Hamiltonian $\widetilde{H}$ has a different canonical distribution. We denote averages with respect to it by $\langle\cdot\rangle'$. We let $\widetilde{R}$ denote the net displacement of the tracer particle under the new flow given by $\widetilde{H}$.

We might try bounding the error in the following way:

$$\begin{aligned}|\langle R^2_{\Delta t}(T)\rangle - \langle R^2(T)\rangle| \le\ & |\langle R^2_{\Delta t}(T)\rangle - \langle \widetilde{R}^2(T)\rangle| \\ & + |\langle \widetilde{R}^2(T)\rangle - \langle \widetilde{R}^2(T)\rangle'| \\ & + |\langle \widetilde{R}^2(T)\rangle' - \langle R^2(T)\rangle|\end{aligned}$$

We discuss each of the three terms in turn.

The first term is due to the numerical trajectory not agreeing with the exact trajectory of the modified system with Hamiltonian $\widetilde{H}$. According to (1) we can bound this term by $C\exp(-D/\Delta t)$ for a duration of $B/\Delta t$. The studies in [5] suggest that this is a tight estimate for typical molecular dynamics simulations.

The second term is the difference in the expectation of $\widetilde{R}^2(t)$ due to a perturbation in the measure. Since the two measures are proportional to $\exp(-H/k\mathcal{T})$ and $\exp(-\widetilde{H}/k\mathcal{T})$ respectively, and $H - \widetilde{H} = \mathcal{O}(\Delta t^2)$, we expect this term to be on the order of $\mathcal{O}(\Delta t^2)$ for all T. This probably can be rigourously controlled without much difficulty.

The third term is just the difference in $\langle R^2(t)\rangle$ between the original system and the perturbed system. This is likely to be extremely difficult to bound. However, showing that it is small is not a question about computation but about statistical physics. For now let us assume that it is $\mathcal{O}(\Delta t^2)$ for all T for now.

Already we can see that this approach will not get us the result that we want, even assuming we can bound the third term. The best estimate we have so far is that the error is bounded by $\mathcal{O}(\Delta t^2)$ for $T \in [0, B/\Delta t]$. The bound would hold on an interval much shorter than what is needed. It appears that backward error analysis alone cannot explain the observed convergence.

4.2 Shadowing

The idea of shadowing is complementary to that of backward error analysis. Whereas backward error analysis shows that the numerical trajectory is close to the exact trajectory of a different Hamiltonian system with the same initial condition, shadowing attempts to show that the numerical trajectory is close to an exact trajectory of the same Hamiltonian system with a different initial condition. See [8] for a nice review of shadowing for Hamiltonian systems.

In our situation, if shadowing were possible, something like the following would hold. Suppose we compute a numerical trajectory starting from (q^0, p^0)

with time step Δt, which we denote by $(q^n, p^n), n \geq 0$. If shadowing is possible then there is an exact trajectory $(\tilde{q}(t), \tilde{p}(t))$ of the same Hamiltonian system starting at some other initial condition $(\tilde{q}(0), \tilde{p}(0))$ such that

$$(q^n, p^n) \approx (\tilde{q}(n\Delta t), \tilde{p}(n\Delta t))$$

for $n\Delta t$ in some large range of times. Assuming that it is possible to shadow every numerical trajectory in this way, let us denote the map on the phase space that takes the numerical initial condition to the initial condition of the shadow trajectory by

$$S_{\Delta t}(q^0, p^0) = (\tilde{q}(0), \tilde{p}(0)).$$

The idea of shadowing is used very effectively by Reich in [11]. For a Hamiltonian system for which shadowing holds he demonstrates that long-time averages will be computed accurately by almost all numerical trajectories. That is,

$$\lim_{T\to\infty} \frac{1}{T} \int_0^T g(q(t), p(t))dt \approx \lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N} g(q^n, p^n), \tag{2}$$

for almost all initial conditions $(q^0, p^0) = (q(0), p(0))$, for reasonable functions g. Since the quantity on the left does not depend on $(q(0), p(0))$ in the systems considered in [11] (except for sets of measure zero), it is sufficient that such a map $S_{\Delta t}$ exists to get the result.

In our case we are interested in more general statistical features of trajectories than long-time averages. For example, the variance of the displacement of a single particle in a finite time interval cannot be put into the form of a long-time average such as in (2). This puts more stringent requirements on $S_{\Delta t}$. To show that statistics are captured correctly we cannot consider just single trajectories; we have make sure the entire ensemble's statistics are reproduced correctly. If the shadowing map $S_{\Delta t}$ systematically picked initial conditions for which the tracer particle tended to move to the left, for example, then the computed statistics could be quite inaccurate. See [8] for a discussion of this issue in the context of astrophysics. What is necessary for this shadowing to work is for $S_{\Delta t}$ to leave the canonical ensemble invariant:

$$\langle G(q, p)\rangle = \langle G(S_{\Delta t}(q, p))\rangle \tag{3}$$

for some suitably broad class of functions G on phase space. This is an even more stringent requirement than just that shadowing is possible at all, and it may be quite unlikely to hold for our system.

Fortunately we can weaken some other requirements demanded of shadowing considerably for our problem. We do not need the trajectory of the whole system to be close; we only need the trajectory of a single particle to be close. Suppose that our tracer particle's numerical trajectory is denoted

by (q_1^n, p_1^n) for $n \geq 0$. We say that *weak shadowing* holds if we can select $\tilde{q}(0)$, $\tilde{p}(0)$ such that

$$(q_1^n, p_1^n) \approx (\tilde{q}_1(n\Delta t), \tilde{p}_1(n\Delta t))$$

for $n\Delta t$ in some long range of times.

To see how this fits in with the conjecture suppose that we have both (3) and

$$\|(q_1^n, p_1^n) - (\tilde{q}_1(T)), \tilde{p}_1(T))\| \leq C\Delta t^2, \tag{4}$$

for $T = n\Delta t \in [0, Ae^{B/\Delta t}]$. This means that (assuming we can obtain reasonable bounds on $R^2_{\Delta t}(T)$ and $R^2(T)$) that

$$\begin{aligned}
|\langle R^2_{\Delta t}(T)\rangle - \langle R^2(T)\rangle| &\leq K|\langle \|q_1^n\|\rangle - \langle \|q_1(t)\|\rangle| \\
&\leq K|\langle \|q_1^n\|\rangle - \langle \|\tilde{q}_1(T)\|\rangle| + K|\langle \|\tilde{q}_1(T)\|\rangle - \langle \|q_1(T)\|\rangle| \\
&\leq K\langle \|(q_1^n, p_1^n) - (\tilde{q}_1(T)), \tilde{p}_1(T))\|\rangle \\
&\quad + K|\langle G(S_{\Delta t}(q^0, p^0))\rangle - \langle G(q^0, p^0)\rangle|,
\end{aligned}$$

for $T \in [0, Ae^{B/\Delta t}]$. Here we have let G be the composition of the time T flow map of the Hamiltonian system with the 2-norm. Now the first term above is bounded by $CTe^{-D/\Delta t}$ by (4) and the second term is 0 by (3), thus establishing the conjecture. Simultaneously proving (3) and (4) for some shadowing map $S_{\Delta t}$ may not be easy, but it may be much easier than proving the usual stronger shadowing result.

5 Discussion

Despite the ideas presented in the previous section, the conjecture we have presented is probably not open to attack by existing techniques. The problem is that there is no rigourous mathematical theory of how statistical regularities emerge from the dynamics of generic high-dimensional Hamiltonian systems. Consequently, there is no theory of how perturbations in the Hamiltonian dynamics leads to perturbation in the statistics. A numerical analyst has three choices when faced with this situation:

1. *Take Up Mathematical Physics.* If we are to make progress on the conjecture these entirely non-numerical problems need to be tackled first. Mathematical physicists are interested in proving things like ergodicity and decay of correlations for Hamiltonian systems such as presented here, and it is conceivable that eventually there will a robust body of theory that we can apply to our problem. So one possibility is to work on developing such a theory. This likely will not have much to do with computation.
2. *Relax Standards of Rigour.* Theoretical physicists, as opposed to mathematical physicists, have accepted that much reliable information can be obtained through calculations that cannot be rigourously justified. Typically theoretical physicists study systems about which nothing interesting

can be proved; to do otherwise would be far too restrictive. There is no reason why this informal yet highly fruitful style of reasoning should be restricted to systems themselves and not numerical discretizations of systems. A combination of non-rigourous arguments and careful numerical experiments could do a lot to clarify how the Störmer–Verlet method is able to compute statistics so accurately for our system.

3. *Abandon the Whole Pursuit.* For many, the purpose of numerical analysis is to provide reliable, efficient algorithms. If one is pursuing a theoretical question, it is hoped that it will lead to better algorithms eventually. Sadly, even a complete resolution of the conjecture we have presented in unlikely to have much effect on computational practice. Many people have tried for years to devise an integrator that is more efficient than the Störmer–Verlet method for computing statistically accurate trajectories in molecular dynamics. They have only been successful for Hamiltonian systems with special structure. (The prime example of this is the multiple time stepping methods, see [7, Chap. VIII.4].) In fact, we state another conjecture which is not formulated rigourously.

   Conjecture 2 *No integration scheme can improve the efficiency by more than a factor of two with which Störmer–Verlet computes statistically accurate trajectories for systems like that in Sect. 2.*

   Here even a clear mathematical formulation would be a challenge. Obviously if we already know a lot about a system we can contrive an algorithm which will give correct statistics for a tracer particle, but this does not count. The conjecture is intended to capture the idea that Störmer–Verlet is a very general purpose method; we do not need to know anything about a system to apply it.

At the Abel Symposium participants seemed to prefer the first of the three options: try to prove what one can about the system and its discretization.

Acknowledgements

The author was supported by an NSERC Discovery Grant. He would like to thank Nilima Nigam, Bob Skeel, and Wayne Hayes for helpful comments.

References

1. M. P. Allen and D. J. Tildesley, *Computer Simulation of Liquids*, Oxford University Press, Oxford, 1989.
2. G. Bennetin and A. Giorgilli, *On the Hamiltonian interpolation of near to the identity symplectic mappings with application to symplectic integration algorithms*, J. Stat. Phys. **74**, (1994) 1117–1143.

3. B. Cano, A. M. Stuart, E. Süli, J. O. Warren, *Stiff oscillatory systems, delta jumps and white noise*, Found. Comput. Math. 1 (2001), no. 1, 69–99.
4. E. Cancès, F. Legoll, and G. Stoltz. *Theoretical and numerical comparison of some sampling methods for molecular dynamics.* To appear, Math. Mod. Num. Anal.
5. R. D. Engle, R. D. Skeel, M. Drees, *Monitoring energy drift with shadow Hamiltonians.* J. Comput. Phys. 206 (2005), no. 2, 432–452.
6. D. Frenkel and B. Smit. *Understanding Molecular Simulation: From Algorithms to Applications, 2nd edition.* Academic Press, London, 2002.
7. E. Hairer, C. Lubich, and G. Wanner. *Geometric Numerical Integration: Structure-Preserving Algorithms for Ordinary Differential Equations.* Springer Series in Computational Mathematics, Berlin, 2002.
8. W. Hayes and K. Jackson. *A Survey of Shadowing Methods for Numerical Solutions of Ordinary Differential Equations.* Applied Numerical Mathematics 53:1-2, pp. 299–321 (2005).
9. T. J. Hunt and R. S. MacKay, *Anosov parameter values for the triple linkage and a physical system with a uniformly chaotic attractor.* Nonlinearity 16 (2003), no. 4, 1499–1510.
10. C. Liverani, Interacting Particles, Hard Ball Systems and the Lorentz Gas, in: D. Szász (Ed.), *Hard Ball Systems and the Lorentz Gas*, Springer, Berlin, 2000.
11. S. Reich. *Backward error analysis for numerical integrators.* SIAM J. Numer. Anal. 36 (1999), no. 5, 1549–1570.
12. H. Sigurgeirsson and A. M. Stuart, *Statistics from computations.* Foundations of computational mathematics (Oxford, 1999), 323–344, London Math. Soc. Lecture Note Ser., 284, Cambridge Univ. Press, Cambridge, 2001.
13. R. D. Skeel and P. F. Tupper, editors. *Mathematical Issues in Molecular Dynamics.* Banff International Research Station Reports. 2005.
14. B. Leimkuhler and S. Reich, *Simulating Hamiltonian dynamics.* Cambridge Monographs on Applied and Computational Mathematics, 14. Cambridge University Press, Cambridge, 2004.
15. C. LeBris, *Computational chemistry from the perspective of numerical analysis.* Acta Numer. 14 (2005), 363–444.
16. A. M. Stuart and A. R. Humphries, *Dynamical Systems and Numerical Analysis*, Cambridge University Press, Cambridge, 1996.
17. A. M. Stuart and J. O. Warren, *Analysis and Experiments for a Computational Model of a Heat Bath*, J. Stat. Phys. **97** (1999), 687–723.
18. P. F. Tupper, *Ergodicity and the numerical simulation of Hamiltonian systems.* SIAM J. Appl. Dyn. Syst. 4 (2005), no. 3, 563–587.

The Dynamics of Transition to Turbulence in Plane Couette Flow

D. Viswanath

Department of Mathematics, University of Michigan, 530 Church Street, Ann Arbor, MI 48109, USA, divakar@umich.edu

Summary. In plane Couette flow, the incompressible fluid between two plane parallel walls is driven by the motion of those walls. The laminar solution, in which the streamwise velocity varies linearly in the wall-normal direction, is known to be linearly stable at all Reynolds numbers (Re). Yet, in both experiments and computations, turbulence is observed for $Re \gtrsim 360$.

In this article, we show that for certain *threshold* perturbations of the laminar flow, the flow approaches either steady or traveling wave solutions. These solutions exhibit some aspects of turbulence but are not fully turbulent even at $Re = 4{,}000$. However, these solutions are linearly unstable and flows that evolve along their unstable directions become fully turbulent. The solution approached by a threshold perturbation could depend upon the nature of the perturbation. Surprisingly, the positive eigenvalue that corresponds to one family of solutions decreases in magnitude with increasing Re, with the rate of decrease given by Re^{α} with $\alpha \approx -0.46$.

1 Introduction

1.1 Transition to Turbulence

The classical problem of transition to turbulence in fluids has not been fully solved in spite of attempts spread over more than a century. Transition to turbulence manifests itself in a simple and compelling way in experiments. For instance, in the pipe flow experiment of Reynolds (see [1]), a dye injected at the mouth of the pipe extended in "a beautiful straight line through the tube" at low velocities or low Reynolds numbers (Re). The line would shift about at higher velocities, and at yet higher velocities the color band would mix up with the surrounding fluid all at once at some point down the tube.

A wealth of evidence shows that the incompressible Navier–Stokes equation gives a good description of fluid turbulence. Therefore one ought to be able to understand the transition to turbulence using solutions of the Navier–Stokes equation. However, the nature of the solutions of the Navier–Stokes equation is poorly understood. Thus the problem of transition to turbulence is fascinating both physically and mathematically.

The focus of this paper is on plane Couette flow. In plane Couette flow, the fluid is driven by two plane parallel walls. If the fluid is driven hard enough, the flow becomes turbulent. Such wall driven turbulence occurs in many practical situations such as near the surface of moving vehicles and is technologically important.

The two parallel walls are assumed to be at $y = \pm 1$. The walls move in the x or streamwise direction with velocities equal to ± 1. The z direction is called the spanwise direction. The Reynolds number is a dimensionless constant obtained as $Re = UL/\nu$, where U is half the difference of the wall velocities, L is half the separation between the walls, and ν is the viscosity of the fluid. The velocity of the fluid is denoted by $\mathbf{u} = (u, v, w)$, where u, v, w are the streamwise, wall-normal, and spanwise components.

For the laminar solution, $v = w = 0$ and $u = y$. The laminar solution is linearly stable for all Re. As shown by Kreiss et al. [7], perturbations to the laminar solution that are bounded in amplitude by $O(Re^{-21/4})$ decay back to the laminar solution. However, in experiments and in computations, turbulent spots are observed around $Re = 360$ [2]. The transition to turbulence in such experiments must surely be because of the finite amplitude of the disturbances. By a threshold disturbance, we refer to a disturbance that would lead to transition if it were slightly amplified but which would relaminarize if slightly attenuated. The concept of the threshold for transition to turbulence was highlighted by Trefethen and others [16]. The amplitude of the threshold disturbance depends upon the type of the disturbance. It is believed to scale with Re at a rate given by Re^{α} for some $\alpha <= -1$.

Our main purpose is to explain how certain finite amplitude disturbances of the laminar solution lead to turbulence. The dynamical picture that will be developed in this paper is illustrated in Fig. 1. Historically, the laminar solution itself has been the focus of attempts to understand mechanisms for transition. Our focus however will be on a different solution that is represented as an empty oval in Fig. 1.

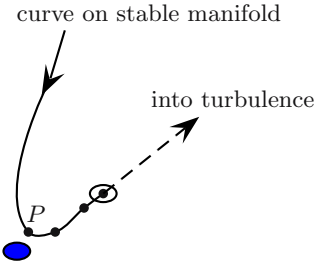


Fig. 1. Schematic sketch of the dynamical picture of transition to turbulence that is developed in this paper. The *solid oval* stands for the laminar solution, and the *empty oval* stands for a steady or traveling wave solution

Solutions that could correspond to the empty oval in Fig. 1 will be called lower-branch solutions [11, 19]. A solution at a certain value of Re can be continued by increasing a carefully chosen parameter. When this parameter is increased, Re first decreases and begins to increase after a bifurcation point and we end up with an "upper branch solution" at the original value of Re. The fact that a continuation procedure can lead to an upper-branch solution appears to have no significance for the dynamics at a fixed value of Re, however.

Depending upon the type of disturbance, the lower-branch solution could either be a steady solution or a traveling wave. Those solutions are not laminar in nature. Neither are they fully turbulent even at high Re. Unlike the laminar solution, these solutions are linearly unstable. The lower-branch solutions remain at an $O(1)$ distance from the laminar solution, while the threshold amplitudes decrease with Re as indicated already. Therefore the threshold disturbances are too tiny to perturb the laminar solution directly onto a lower-branch solution. We will show, however, that some threshold disturbances perturb the laminar solution to a point on the stable manifold of a lower-branch solution (point P in Fig. 1). A slightly larger disturbance brings the flow close to the lower-branch solution, after which the flow follows a branch of its unstable manifold and becomes fully turbulent.

For certain types of disturbances, the perturbed laminar solution does not approach a lower branch solution. Thus the dynamical picture of Fig. 1 is not valid for those disturbances. Instead it flows towards an *edge state* [15]. We give a brief discussion of the nature of the edge states in Sect. 4.

1.2 Connections to Earlier Research

The dynamical picture presented in Fig. 1 is related directly and indirectly to much earlier research. Basic results from hydrodynamic stability show that some eigenmodes that correspond to the least stable eigenvalue of the linearization around the laminar solution do not depend upon the spanwise or z direction. This may lead one to expect that disturbances that trigger transition to turbulence are 2-dimensional. That expectation is not correct, however. As shown by Orszag and Kells [13], spanwise variation is an essential feature of disturbances that trigger transition to turbulence. Accordingly, all the disturbances considered in this paper are 3-dimensional.

Kreiss et al. [7] and Lundbladh et al. [9] investigated disturbances that are non-normal pseudomodes of the linearization of the laminar solution. Since the laminar solution is linearly stable, a slight perturbation along an eigenmode will simply decay back to the laminar solution at a predictable rate. The pseudomodes are chosen to maximize transient growth of the solution of the linearized equation, which is a consequence of the non-normality of the linearization. Such disturbances lead to transition with quite small amplitudes and will be considered again in this paper. It must be noted, however, that any consideration based on the linearization alone can only be valid in a small

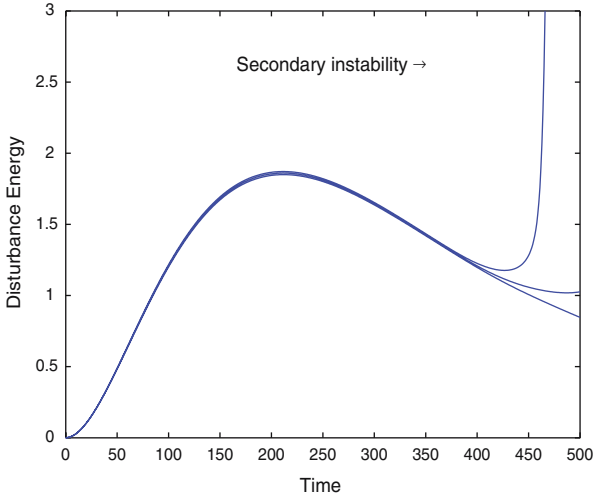


Fig. 2. The plot above shows the secondary instability in a transition computation at $Re = 2{,}000$

region around the laminar solution. The dynamics of transition to turbulence, as sketched in Fig. 1, involves an approach towards a lower-branch solution that lies at an $O(1)$ distance from the laminar solution. It is therefore necessary to work with the fully nonlinear Navier–Stokes equation to explicate the dynamics of transition to turbulence.

Figure 2 shows the variation of the disturbance energy with time for a disturbance that leads to transition. We observe that the disturbance energy increases smoothly initially and is then followed by a spike. The spike is in turn followed by turbulence. The spike corresponds to a secondary instability, as noted by Kreiss et al. [7]. In fact, the so-called secondary instability is just the linear instability of a lower-branch solution as will become clear.

Partly motivated by the secondary instability, there was a search for nonlinear steady solutions related to transition as reviewed in [3]. Early success in this effort was due to Nagata [11, 12] who computed steady solutions of plane Couette flow in the interval $125 \leq Re \leq 300$. Waleffe [18, 19, 20] introduced a more flexible method for computing such solutions, and like Nagata, argued that such solutions could be related to transition to turbulence. The numerical method we use was introduced in [17]. It uses a combination of Krylov space methods and the locally optimally constrained hook step to achieve far better resolution as shown by [4, 17] and this paper.

The computations in [7, 9] imply that threshold amplitudes scale as Re^{α} for $\alpha < -1$. The value of α appears to depend upon the type of perturbation. Our focus is not on determining the scaling of the threshold amplitudes. Nevertheless, we will discuss numerical difficulties that beset determination of threshold amplitudes.

Measuring threshold amplitudes poses experimental challenges as well and it is not always clear from experiments if the thresholds have a simple power scaling with Re. One difficulty is that the turbulent states can be short lived. Schmiegel and Eckhardt [14] have connected the lifetime of turbulence to the possibility that turbulent dynamics in the transition regime is characterized by a chaotic repeller and not a chaotic attractor.

1.3 Connections to Recent Research

Wang et al. [21] have taken steps towards an asymptotic theory of the lower branch solutions and carried their computation beyond $Re = 50,000$. They connect the asymptotics to scalings of the threshold for transition to turbulence. The lower branch states occur as solutions to equations that use periodic boundary conditions. Because such boundary conditions cannot be realized in laboratory setups, the solutions are best thought of as waves. Thus it is pertinent to consider their stability with respect to subharmonic disturbances as in [21]. That paper also suggests that lower branch solutions might be of use for control. A somewhat different suggestion related to control can be found in [5].

Not all disturbances follow the dynamical picture of Fig. 1 as already noted. For the third type of disturbance considered in Sect. 4, the laminar solution perturbed by the threshold disturbance evolves towards a state that looks almost like an invariant object of the underlying differential equation. Those objects have been termed edge states by Schnieder et al. [15]. Lagha et al. [8] make the important point that the dynamical picture of Fig. 1 can be valid for typical disturbances only if the lower-branch solution has a single unstable eigenvalue.

Near the threshold for the third type of disturbance, it appears as if the disturbed state evolves and approaches a traveling wave. Indeed, a crude or under-resolved computation could easily mistake that appearance for a true solution. When we attempted to refine that near-solution using the numerical method reviewed in Sect. 3, the numerical method converged to a traveling wave solution. However, that traveling wave has two unstable eigenvalues and the flow near the threshold does not come as close to that traveling wave as the dynamical picture of Fig. 1 would require.

Visualizing the dynamics in state space is fundamental to the approach to transition to turbulence sketched in this paper and in the articles discussed above. Yet there has so far been no way to obtain revealing visualizations of state space dynamics. Gibson et al. [4] have recently produced revealing visualizations of the state space of turbulent flows. For instance, one of their figures shows a messy-looking turbulent trajectory cleanly trapped by the unstable manifolds of certain equilibrium solutions.

Section 2 reviews some basic aspects of plane Couette flow. The numerical method used to flesh out the dynamical picture of Fig. 1 is given in Sect. 3. In

Sect. 4, we consider three different types of disturbances. The lower-branch solutions (empty oval of Fig. 1) that correspond to the first two types are steady solutions. For a given Re, the solutions that correspond to these two types are identical modulo certain symmetries of plane Couette flow. In Sect. 5, we consider some qualitative aspects of the solutions reported in Sect. 4. A surprising finding is that these these solutions are less unstable for larger Re. The top eigenvalue of these solutions is real and positive. For one family of solutions, the top eigenvalue appears to decrease at the rate Re^{α} for $\alpha \approx -0.46$.

In the concluding Sect. 6, we give additional context for this paper from two points of view. The first point of view is mainly computational and has to do with reduced dimension methods. In this paper, we have taken care to use adequate spatial resolution to ensure that the computed solutions are true solutions of the Navier–Stokes equation. We recognize, however, that resolving all scales may prove computationally infeasible in some practical situations. We argue that transition to turbulence computations can be useful in gaging the possibilities and limitations of methods that do not resolve all scales. Secondly, we briefly discuss the connection of transition computations with transition experiments.

2 Some Aspects of Plane Couette Flow

The Navier–Stokes equation $\partial\mathbf{u}/\partial t + (\mathbf{u}.\nabla)\mathbf{u} = -(1/\rho)\nabla p + (1/Re)\triangle\mathbf{u}$ describes the motion of incompressible fluids. The velocity field $\mathbf{u}$ satisfies the incompressible constraint $\nabla.\mathbf{u} = 0$. For plane Couette flow the boundary conditions are $\mathbf{u} = (\pm 1, 0, 0)$ at the walls, which are at $y = \pm 1$. To render the computational domain finite, we impose periodic boundary conditions in the x and z directions, with periods $2\pi\Lambda_x$ and $2\pi\Lambda_z$, respectively. To enable comparison with [9], we use $\Lambda_x = 1.0$ and $\Lambda_z = 0.5$ throughout this paper.

Certain basic quantities are useful for forming a general idea of the nature of a velocity field of plane Couette flow. The first of these is the rate of energy dissipation per unit volume for plane Couette flow, which is given by

$$D = \frac{1}{8\pi^2\Lambda_x\Lambda_z}\int_0^{2\pi\Lambda_z}\int_{-1}^{+1}\int_0^{2\pi\Lambda_x} |\nabla u|^2 + |\nabla v|^2 + |\nabla w|^2 \; dx\, dy\, dz. \qquad (1)$$

The rate of energy input per unit volume is given by

$$I = \frac{1}{8\pi^2\Lambda_x\Lambda_z}\int_0^{2\pi\Lambda_x}\int_0^{2\pi\Lambda_z} \frac{\partial u}{\partial y}\Big|_{y=1} + \frac{\partial u}{\partial y}\Big|_{y=-1} dx\, dz. \qquad (2)$$

For the laminar solution $(u, v, w) = (y, 0, 0)$, both D and I are normalized to evaluate to 1. Expressions such as (1) and (2) are derived using formal manipulations. The derivations would be mathematically valid if the velocity field $\mathbf{u}$ were assumed to be sufficiently smooth. Although such smoothness

properties of solutions of the Navier–Stokes are yet to be proved, numerical solutions possess the requisite smoothness. Even solutions in the turbulent regime appear to be real analytic in the time and space variables, which is why spectral methods have been so successful in turbulence computations.

In the long run, on physical grounds, we expect the time averages of D and I to be equal because the energy dissipated through viscosity must be input at the walls. For steady solutions and traveling waves, the values of D and I must be equal.

Another useful quantity is the disturbance energy. The disturbance energy of (u, v, w) is obtained by integrating $(u-y)^2+v^2+w^2$ over the computational box. This quantity has already been used in Fig. 2. The disturbance energy is a measure of the distance from the laminar solution.

Two discrete symmetries of the Navier–Stokes equation for plane Couette flow will enter the discussion later. The shift-reflection transformation of the velocity field is given by

$$S_1\mathbf{u} = \begin{pmatrix} u \\ v \\ -w \end{pmatrix} \left(x + \pi\Lambda_x, y, -z \right), \tag{3}$$

and the shift-rotation transformation of the velocity field is given by

$$S_2\mathbf{u} = \begin{pmatrix} -u \\ -v \\ w \end{pmatrix} \left(-x + \pi\Lambda_x, -y, z + \pi\Lambda_z \right). \tag{4}$$

Plane Couette flow is unchanged under both these transformations. Thus if a single velocity field along a trajectory of plane Couette flow satisfies either symmetry, all points along the trajectory must have the same symmetry. However, velocity fields that lie on the stable and unstable manifolds of symmetric periodic or relative periodic solutions need not be symmetric.

3 Numerical Method

The Navier–Stokes equation in the standard form given in Sect. 2 cannot be viewed as a dynamical system because the velocity field $\mathbf{u}$ must satisfy the incompressibility condition and because there is no equation for evolving the pressure p. It can be recast as a dynamical system, however, by using the y components of $\mathbf{u}$ and $\nabla \times \mathbf{u}$, which is the vorticity field. If the resulting system is discretized in space using $M + 1$ Chebyshev points in the y direction, and $2L$ and $2N$ Fourier points in the x and z directions, respectively, the number of degrees of freedom of the spatially discretized system is given by

$$2(M-1) + (2M-4)((2N-1)(2L-1)-1) \tag{1}$$

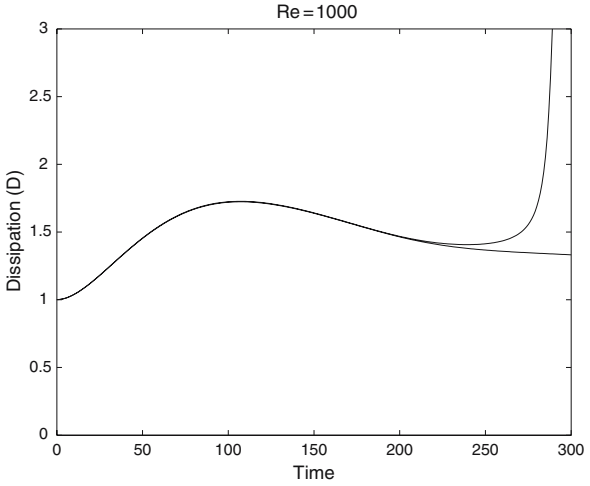


Fig. 3. The plot above shows the variation of D defined by (1) for a disturbance slightly above the threshold and for a disturbance slightly below the threshold

as shown in [17]. We do not use a truncation strategy to discard modes and we employ dealiasing in the directions parallel to the wall.

Given a form of the disturbance P, the threshold for transition is obtained by integrating the disturbed velocity $(y, 0, 0) + \epsilon P$ in time for different ϵ [7]. If ϵ is greater than the threshold value, the flow will spike and become turbulent as evident from Figs. 2 and 3. If ϵ is below the threshold value, the flow will relaminarize. As indicated by Figs. 2 and 3, we may graph either disturbance energy or D to examine a value of ϵ. We may also graph I, which is defined by (2), against time.

The accurate determination of thresholds is beset by numerical difficulties. To begin with, suppose that we are able to integrate the Navier–Stokes equation for plane Couette flow exactly. Then as implied by the dynamical picture in Fig. 1, a disturbance of the laminar solution that is on the threshold will fall into a lower-branch solution, and it will take infinite time to do so. However, computations for determining the threshold, such as that shown in Fig. 2, can only be over a finite interval of time. Thus the finiteness of the time of integration is a source of error in determining thresholds. Two other sources of error are spatial discretization and time discretization.

An accurate determination of the threshold will need to estimate and balance these three sources of error carefully. In our computations, we determine the thresholds with only about two digits of accuracy. That modest level of accuracy is sufficient for our purposes. In Tables 1 and 3, the thresholds are reported using disturbance energy per unit volume.

Once the threshold has been determined, we need to compute a steady solution or a traveling wave to complete the dynamical picture of Fig. 1. The

Table 1. Data for disturbances of the form (1) with unsymmetric noise and for steady solutions that correspond to the empty oval in Fig. 1

Label	Re	D/I	λ_{max}	Re_τ	T	Threshold
$B1$	500	1.3920	0.04326	53	150	$2.46e-4$
$B2$	1,000	1.3486	0.03294	73	300	$5.73e-5$
$B3$	2,000	1.3285	0.02413	103	500	$1.36e-5$
$B4$	4,000	1.3210	0.01732	145	1,000	$3.30e-6$

The steady solutions are labeled $B1$ through $B4$. D and I, which are defined by (1) and (2), correspond to those steady solutions. The next two columns give the eigenvalue with the maximum real part and the frictional Reynolds number for those solutions. T is the time interval used to determine the threshold disturbance and the threshold is reported using disturbance energy per unit volume

initial guess for that lower-branch solution is produced by perturbing the laminar solution by adding the numerically determined threshold disturbance and integrating the perturbed point over the time interval used for determining the threshold (this time interval is 500 in Fig. 2 and 300 in Fig. 3).

That initial guess is fed into the method described in [17] to find a lower-branch solution with good numerical accuracy. That method finds solutions by solving Newton's equations, but the equations are set up and solved in a non-standard way. Suppose that the spatially discretized equation for plane Couette flow is written as $\dot{x} = f(x)$, where the dimension of x is given by (1). To find a steady solution, for instance, it is natural to solve $f(x) = 0$ after supplementing that equation by some conditions that correspond to the symmetries (3) and (4). However that is not the way we proceed. We solve for a fixed point of the time t map $x(t; x_0)$, for a fixed value of t, after accounting for the symmetries. The Newton equations are solved using GMRES. The method does not always compute the full Newton step, however. Instead, the method finds the ideal trust region step within a Krylov subspace as described in [17].

This method can easily handle more than 10^5 degrees of freedom, and thus makes it possible to carry out calculations with good spatial resolution. The reason for setting up the Newton equations in the peculiar way described in the previous paragraph has to do with the convergence properties of GMRES. The matrix that arises in solving the Newton equations approximately has the form $I - \partial x(t; x_0)/\partial x_0$, where I is the identity. Because of viscous damping of high wavenumbers, many of the eigenvalues of that matrix will be close to 1, thus facilitating convergence of GMRES. We may expect the convergence to deteriorate as Re increases, because viscous damping of high wavenumbers is no longer so pronounced, and that is indeed the case. Nevertheless, we were able to go up to $Re = 4{,}000$, and we believe that even higher values of Re can be reached.

4 Disturbances of the Laminar Solution and Transition to Turbulence

In this section, we consider three types of disturbances and determine the threshold amplitudes for various values of Re. To complete the dynamical picture of Fig. 1, we determine for the first two types the steady solution or traveling wave that corresponds to the empty oval of that figure using the numerical method of the previous section.

4.1 Rolls with Unsymmetric Noise

We follow [7] and consider the disturbance,

$$(u, v, w) = \epsilon(0, \psi_z, -\psi_y), \tag{1}$$

where $\psi = (1 - y^2)^2 \sin(z/\Lambda_z)$. This disturbance is unchanged by both S_1, which was defined by (3), and by S_2, which was defined by (4). A disturbance of the laminar solution $\mathbf{u} = (y, 0, 0)$ of the form (1) never leads to transition to turbulence. It is necessary to add some more terms to the disturbance to make the velocity field depend upon the x direction.

To introduce dependence on x, we add modes of the Stokes problem. One can get an eigenvalue problem for $\hat{v}(y)$, where $v = \hat{v}(y) \exp(\imath lx/\Lambda_x + \imath nz/\Lambda_z) \exp(\sigma t)$, or for $\hat{\eta}(y)$, where $\eta = \hat{\eta}(y) \exp(\imath lx/\Lambda_x + \imath nz/\Lambda_z) \exp(\sigma t)$. Here η is the wall-normal component of the vorticity field. For a v mode, $\eta = 0$, and vice versa. For a given mode, the velocity field is recovered using the divergence free condition. The velocity fields of modes with different (l, n) are obviously orthogonal. A calculation shows that the velocity fields for the v and η modes with the same (l, n) are also orthogonal. For a given (l, n), we pick the v and η modes with the least stable σ.

To the disturbance (1), we added both v and η modes for (l, n) with $-3 \le l \le 3$ and $-7 \le n \le 7$. Together the added modes can be called noise. The energy of the noise was equal to 1% of the energy of (1). This energy was equally distributed over the various orthogonal modes. Following [7], we chose random phases for the modes. The threshold can depend upon the choice of phase. Therefore, for accurate determination of thresholds it is better to use non-random phases.

After adding modes of this form to (1), the resulting disturbance in unchanged by neither S_1 nor S_2. Therefore the disturbance is unsymmetric. Table 1 reports data from computations carried out using such an unsymmetric disturbance. The thresholds in that table give the energy of (1) and do not include the energy within the noise terms. The lower-branch solutions $B1$ through $B4$ correspond to the empty oval in Fig. 1. Each of these solutions appears to have a single unstable eigenvalue. We determined the most unstable eigenvalues using simultaneous iteration and the time t map of the Navier–Stokes equation, as in Sect. 3, with $t = 8$. All the solutions seem to

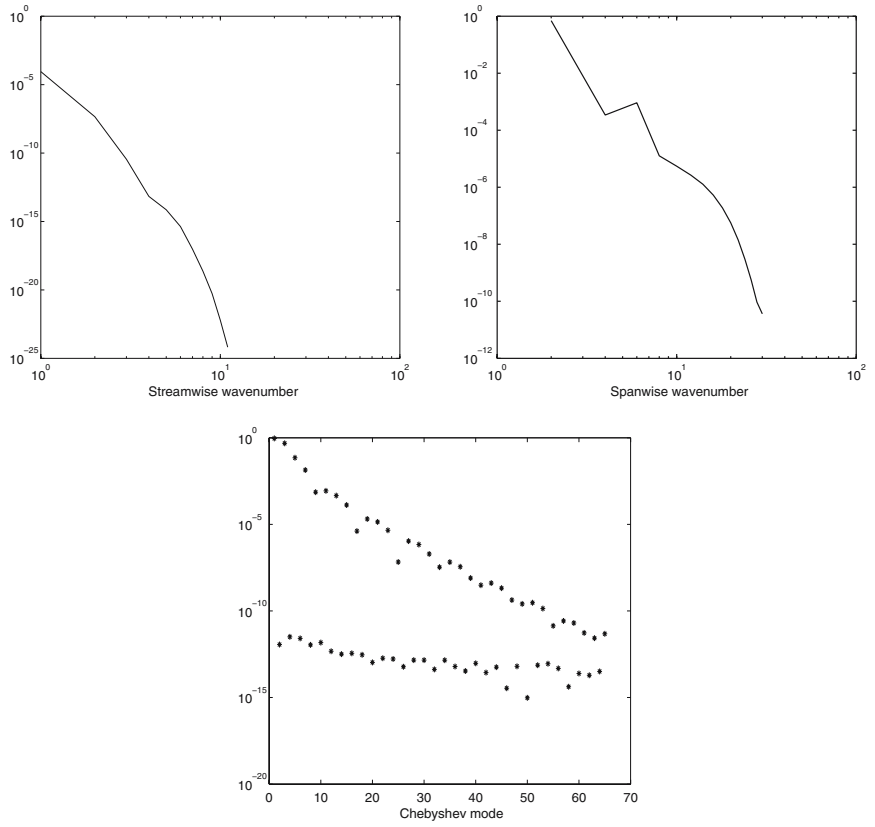


Fig. 4. The plots above graph the energy in the solution $B4$ of Table 1 against streamwise wavenumber, spanwise wavenumber, and Chebyshev mode

have just one unstable eigenvalue. That eigenvalue is real. Surprisingly, it decreases with Re at the rate Re^{α}, where $\alpha \approx -0.46$. Thus the lower-branch solutions become less and less unstable with increasing Re.

All our computations used $(2L, M+1, 2N) = (24, 65, 32)$. By (1), the number of degrees of freedom in the computation for finding the lower-branch solutions is $88,414$. As shown by Fig. 4, that much resolution was entirely adequate. The solutions $B1$ through $B4$ were computed with at least five digits of accuracy.

4.2 Rolls with Symmetric Noise

It has been suggested that one purpose of adding the noise to (1) is to break symmetries and that a symmetric disturbance would lead to drastically increased thresholds [7]. To investigate that matter, we symmetrized the disturbances used to generate Table 1. More specifically, if $\mathbf{u}$ is a disturbed velocity field, we replaced it by $(\mathbf{u} + S_1\mathbf{u} + S_2\mathbf{u} + S_1S_2\mathbf{u})/4$ which is unchanged by

Table 2. Data for disturbances of the form (1) with symmetric noise and steady solutions that correspond to the empty oval in Fig. 1

Label	Re	s_x	s_z	Threshold
$C1$	500	1.5600	0.0016	$2.97e-4$
$C2$	1,000	6.1093	0.0012	$5.72e-5$
$C3$	2,000	0.5075	0.0018	$1.40e-5$
$C4$	4,000	2.8719	0.0013	$3.28e-6$

The solutions Cn are connected to the solutions Bn of Table 1 as follows: $Cn(x + s_x, y, z + s_z) = Bn$

Table 3. Data for disturbances obtained by superposing Orr–Sommerfeld modes and for the corresponding traveling waves labeled $D1$ and $D2$

Label	Re	D/I	λ_{max}	Re_τ	c_x	c_z	T	Threshold
$D1$	500	1.2863	0.0464	51	0.3051	0	100	$8.4e-3$
$D2$	1,000	1.2522	0.0379	72	0.2666	0	200	$1.6e-3$

c_x and c_z give the wave speeds in the x and z directions. The other columns are as in Table 1

both S_1 and S_2. A comparison of Tables 1 and 2 shows that the thresholds are in fact not elevated. Thus we conclude that the purpose of adding the noise is not to break the symmetry but to introduce dependence on the x direction. The lower-branch solutions that correspond to such symmetric disturbances are labeled $C1$ through $C4$ in Table 2.

The solutions $C1$ through $C4$ are just translations of the solutions $B1$ through $B4$ as indicated in Table 2. If the thresholds were determined exactly, the disturbances of Tables 1 and 2 would come arbitrarily close to the corresponding solution in the infinite time limit. Each threshold in those tables was determined inexactly using a finite time interval, and we verified that the disturbed states evolve and come within 2% of the corresponding lower-branch solution. Thus there can be little doubt about the role of these lower-branch solutions in the transition to turbulence. The C family of solutions is the same as the lower-branch family of [20].

Given that the solutions $C1$ through $C4$ are just translations of the solutions $B1$ through $B4$, it is tempting to think that all threshold disturbances, say at $Re = 4{,}000$, might evolve and approach a translate of a single solution such as $C4$. That is not correct, however, as we will now show.

4.3 Superposed Orr–Sommerfeld Modes

The disturbances for Table 3 were obtained by superposing Orr–Sommerfeld modes as in [13]. An Orr–Sommerfeld mode is of the form $(u, v, w) = (\hat{u}(y), \hat{v}(y), \hat{w}(y)) \exp(\imath lx/\Lambda_x + \imath nz/\Lambda_z) \exp(\sigma t)$. We use Orr–Sommerfeld modes with $(l, n) = \pm(1, 0)$ and $(l, n) = \pm(1, 1)$. The phases of the Orr–Sommerfeld

modes were chosen to make $\hat{v}(0)$ real. The disturbance energy was equally distributed across the modes. For given (l, n), we chose the least stable mode and symmetrized it as in (3.2) of [13]. Note that the disturbance depends on both the x and z directions.

The solutions obtained by following the numerical method of Sect. 3 were traveling waves in this case. The wave speeds for both $D1$ and $D2$ in Table 3 are nonzero in the x direction. These traveling waves are unsymmetric and they do not become symmetric even after translations in the x and z directions.

The thresholds for this third type of disturbance are reported in Table 3. Close to the threshold, the flow appears to approach a traveling wave. After a diligent computation, we feel sure that there is no true traveling wave solution or relative periodic solution to complete the dynamical picture of Fig. 1. The flow near the threshold evolves and comes within 10% of $D1$ or $D2$ but no closer. It appears to approach an edge state.

Figure 5 shows plots of the rates of energy input and energy dissipation near an edge state. In that figure, the disturbance is very close to the threshold and the time axis is chosen to correspond to an edge state. Note that the dissipation sags below energy input and then rises above it. Therefore, we do not expect a traveling wave or an equilibrium solution near the edge state. The second crossing of the two curves is below the first. In addition, both the curves spike and transition to turbulence soon after they cross. Therefore, a periodic or relative periodic solution is unlikely to be found near the edge state.

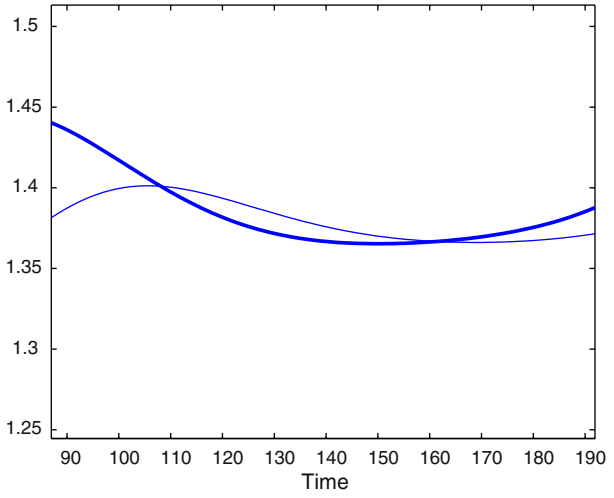


Fig. 5. The *thick line* is a plot of D defined by (1) and the *thin line* is a plot of I defined by (2). The disturbance used to get the plots was a superposition of Orr–Sommerfeld modes at $Re = 500$

In pipe flow transition computations, we have observed that the rate of dissipation and the rate of energy input become almost horizontal lines near the edge states. The rate of dissipation is slightly above the rate of energy input, suggesting that there may be no invariant objects near the edge states in these instances.

As stated earlier, the laminar solution of plane Couette flow is linearly stable. The computations of this section shed some light on the laminar-turbulent separatrix. A part of this separatrix is formed by the stable manifolds of the B and C family of solutions. We have shown that these stable manifolds come closer and closer to the laminar solution as Re increases. The traveling waves $D1$ and $D2$ are also on the separatrix. However, we have not found tiny disturbances to the laminar solution for which the thresholds diminish in magnitude with increasing Re and which approach these solutions as the flow evolves as in Fig. 1. In the next section, we show that the D solutions are qualitatively similar to the B and C solutions.

5 Lower-Branch Solutions of Plane Couette Flow

A notable feature of the solutions of Tables 1–3 is that the solutions are streaky. This feature is illustrated in Fig. 6. The contour lines for the streamwise velocity are approximately parallel to the x axis, but the streamwise velocity varies in a pronounced way in the z direction. We observe that $D1$ is less streaky than $C1$. The contour lines become much straighter when we go from $C1$ to $C4$. This increase in streakiness with Re is in accord with the asymptotic theory sketched in [21].

To show that these solutions are not fully turbulent, we begin by describing the use of frictional or wall units [10]. The mean shear at the wall, which is denoted by $\langle \frac{\partial u}{\partial y}\big|_{y=1} \rangle$, is the basis for frictional units. The frictional units for velocity and length are given by

$$u_f = \sqrt{\nu \Big\langle \frac{\partial u}{\partial y}\Big|_{y=1} \Big\rangle} \qquad \text{and} \qquad l_f = \nu / u_f,$$

respectively. If the width of the channel is L, the frictional Reynolds number is given by $Re_\tau = L u_f/\nu = L/l_f$. The width of the channel in frictional units equals the frictional Reynolds number. The use of frictional units is signaled by using $+$ as a superscript.

The use of frictional units is necessary to state some remarkable properties of turbulent boundary layers. If y^+ measures the distance from the wall and $<u>^+$ is the mean streamwise velocity in frictional units, after making $<u>^+ = 0$ at $y^+ = 0$ by shifting the mean velocities if necessary, then $<u>^+ \approx y^+$ in the viscous sublayer. The viscous sublayer is about five frictional units thick. The buffer layer extends from 5 to about 30 units. It is followed by the logarithmic layer where $<u>^+ \approx A \log y^+ + B$, for constants A and B. These relationships

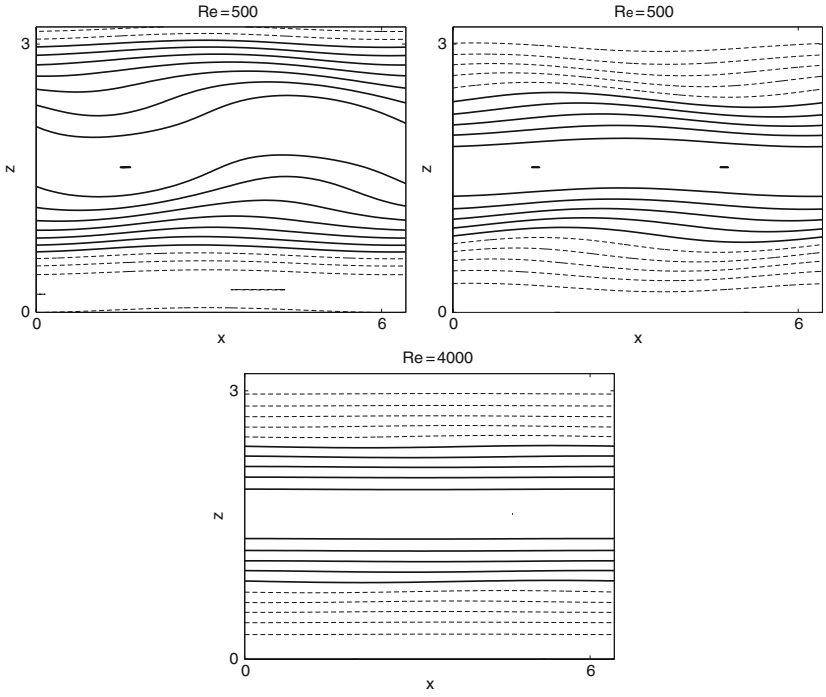


Fig. 6. Contour plots of the streamwise velocity at $y = 0$. The plots correspond to $D1$, $C1$, and $C4$. Contour lines are drawn at 12 equispaced values between the maximum and minimum streamwise velocity in the slice. The lines are *solid* for positive values and *dashed* for negative values. The minimums are $-0.1922, -0.3969, -0.3833$ and the maximums are $0.4146, 0.3969, 0.3833$. In each plot the maximum occurs in the widest gap between the *solid lines*

between $<u>^+$ and y^+ have been confirmed in numerous experiments and in some computations. The experiments are of a very diverse nature as discussed in [10], and it is remarkable that such a simple relationship holds across all those experiments.

There are other relationships that govern the dependence of quantities such as turbulence intensities or turbulent energy production on the distance from the wall. These relationships also characterize turbulent boundary layers. To show that the C and D solutions are not fully turbulent, we will use plots of turbulent energy production. Turbulent energy production equals

$$- <u^* v^*> \frac{\partial <u>}{\partial y},$$

where $u^* = u - <u>$ and $v^* = v - <v>$ are the fluctuating components of the streamwise and wall-normal velocities and $<u>$ is the mean streamwise velocity. Turbulent energy production is easy to measure experimentally and shows

a very sharp peak in the buffer region of turbulent boundary layers [6]. This sharp peak has intrigued experimentalists for a long time. In experiments, the means are calculated by averaging pointwise measurements over long intervals of time. The means involved in the definition of turbulent energy production will be computed by averaging in the x and z directions.

Figure 7 shows plots of turbulent energy production against y^+, the distance from the upper wall in frictional units. In each plot, y^+ varies from 0 to the channel width. The first plot is for a turbulent steady solution of plane Couette flow at $Re = 400$. The data for the velocity field of that solution is from [20]. The second and third plots are for $C1$ and $C4$, respectively. The first plot is strikingly different from the other two. In the first plot, we notice that turbulent energy production peaks inside the buffer layer and then falls off sharply, in a way that is typical of turbulent boundary layers. The second and third plots correspond to higher Re, yet the peak occurs farther away from the wall in frictional units and there is no sharp fall-off. The plots for

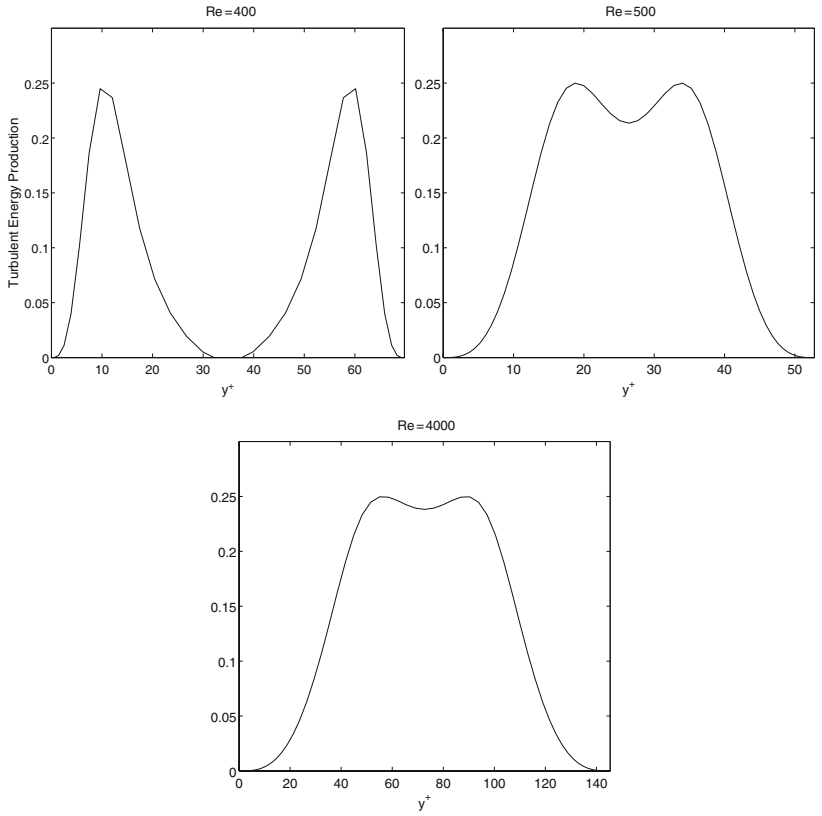


Fig. 7. The plots show the dependence of turbulent energy production in frictional units on y^+ for a turbulent steady solution, $C1$, and $C4$

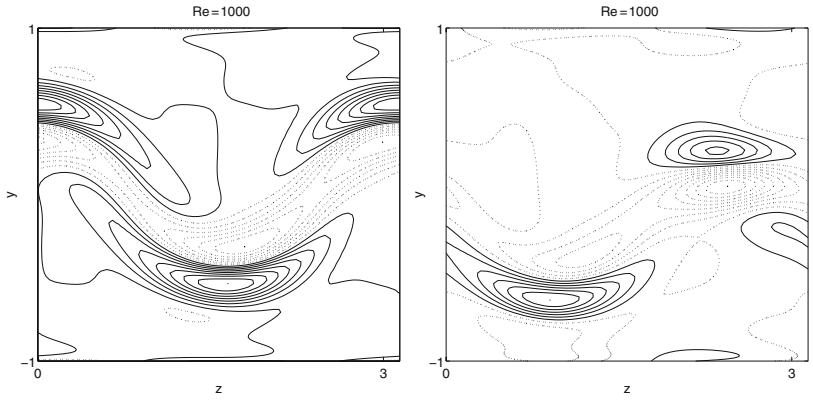


Fig. 8. Contour plots of the streamwise vorticity at $x = \pi$. The contour lines are equispaced between -0.11 and 0.13 for the first plot, which corresponds to $C2$, and between -0.19 and 0.17 for the second plot, which corresponds to $D2$. The lines are *dotted* for negative values of streamwise vorticity

$D1$ and $D2$ are not shown. Those plots are similar to the ones for $C1$ and $C4$ in that they do not match what we expect for turbulent boundary layers. A notable difference is that the plots for $D1$ and $D2$ are not symmetric about the center of the channel. Thus the C and D solutions exhibit some aspects of near-wall turbulence such as the formation of streaks, but do not exhibit many other aspects.

Figure 8 is another illustration of the qualitative similarity between the C and D solutions. In both plots of Fig. 8, one may observe a region near the center of the channel where the streamwise vorticity varies rapidly. Those regions correspond to the critical layer discussed in [21].

6 Conclusion

We verified the dynamical picture for transition to turbulence given in Fig. 1 for certain disturbances. The third type of disturbance considered in Sect. 4.3 shows that that picture does not hold for all disturbances. A more exhaustive study of different types of disturbances of the laminar solution would be desirable.

We found (along with Wang et al. [21]) that the B or C solutions become less unstable as Re increases. This was an unexpected finding. Even a good heuristic explanation of this trend would be interesting.

Transition to turbulence computations would be good targets for reduced dimension methods. Reduced dimension methods are diverse in nature. Although this is not the place to review them, we believe the intricate dynamics of transition of turbulence featuring steady solutions, traveling waves, thresholds and various types of disturbances makes it non-trivial to reduce

dimension. A valid way to reduce dimension must capture the dynamics correctly and not introduce spurious artifacts. It has been known since the work of Orszag and Kells [13] that under-resolved spatial discretizations lead to spurious transitions.

It is important to connect transition computations to experiments. However, connecting transition computations to experiments is impeded by two problems. Firstly, the experiments are performed in much larger domains to eliminate boundary effects. The numerical methods reviewed and discussed in Sect. 2 ought to be able to handle at least 10 million degrees of freedom with a good parallel implementation. Therefore it seems that computations can be performed in much larger domains (i.e., domains with larger Λ_x and Λ_z) and that this problem can be overcome. Secondly, it is very difficult to imagine a way to reproduce the sort of disturbances that have been considered in the computational literature in experiments. The disturbances used in experiments are of a different sort. For instance, one type of disturbance is to inject fluid from the walls. The best way to reconcile this disparity between computation and experiment might be to carry out computations using good models of laboratory disturbances.

Acknowledgments

The author thanks B. Eckhardt, J.F. Gibson, N. Lebovitz, L.N. Trefethen, and F. Waleffe for helpful discussions. This work was supported by the NSF grant DMS-0407110 and by a research fellowship from the Sloan Foundation.

References

1. D. Acheson. *Elementary Fluid Dynamics*. Oxford University Press, Oxford, 1990.
2. K.H. Bech, N. Tillmark, P.H. Alfredsson, and H.I. Andersson. An investigation of turbulent plane Couette flow at low Reynolds number. *Journal of Fluid Mechanics*, 286:291–325, 1995.
3. A. Cherhabili and U. Eherenstein. Finite-amplitude equilibrium states in plane Couette flow. *Journal of Fluid Mechanics*, 342:159–177, 1997.
4. J.F. Gibson, J. Halcrow, and P. Cvitanović. Visualizing the geometry of state space in plane Couette flow. *Journal of Fluid Mechanics*, 2008. To appear. Available at `www.arxiv.org:0705.3957`.
5. G. Kawahara. Laminarization of minimal plane Couette flow: going beyond the basin of attraction of turbulence. *Physics of Fluids*, 17:art. 041702, 2005.
6. S.J. Kline, W.C. Reynolds, F.A. Schraub, and P.W. Rundstadler. The structure of turbulent boundary layers. *Journal of Fluid Mechanics*, 30:741–773, 1967.
7. G. Kreiss, A. Lundbladh, and D.S. Henningson. Bounds for threshold amplitudes in subcritical shear flows. *Journal of Fluid Mechanics*, 270:175–198, 1994.
8. M. Lagha, T.M. Schneider, F. De Lillo, and B. Eckhardt. Laminar-turbulent boundary in plane Couette flow. 2007. preprint.

9. A. Lundbladh, D.S. Henningson, and S.C. Reddy. Threshold amplitudes for transition in channel flows. In M.Y. Hussaini, T.B. Gatski, and T.L. Jackson, editors, *Turbulence and Combustion*. Kluwer, Holland, 1994.
10. A.S. Monin and A.M. Yaglom. *Statistical Fluid Mechanics*. The MIT Press, Cambridge, 1971.
11. M. Nagata. Three dimensional finite amplitude solutions in plane Couette flow: bifurcation from infinity. *Journal of Fluid Mechanics*, 217:519–527, 1990.
12. M. Nagata. Three-dimensional traveling-wave solutions in plane Couette flow. *Physical Review E*, 55:2023–2025, 1997.
13. S.A. Orszag and L.C. Kells. Transition to turbulence in plane Poiseuille and plane Couette flow. *Journal of Fluid Mechanics*, 96:159–205, 1980.
14. T.M. Schmiegel and B. Eckhardt. Fractal stability border in plane Couette flow. *Physical Review letters*, 79:5250, 1997.
15. T.M. Schneider, B. Eckhardt, and J.A. Yorke. Turbulence transition and edge of chaos in pipe flow. *Physical Review Letters*, 99:034502, 2007.
16. L.N. Trefethen, A.E. Trefethen, S.C. Reddy, and T.A. Driscoll. Hydrodynamic stability without eigenvalues. *Science*, 261:578–584, 1993.
17. D. Viswanath. Recurrent motions within plane Couette turbulence. *Journal of Fluid Mechanics*, 580:339–358, 2007.
18. F. Waleffe. Three-dimensional coherent states in plane shear flows. *Physical Review Letters*, 81:4140–4143, 1998.
19. F. Waleffe. Exact coherent structures in channel flow. *Journal of Fluid Mechanics*, 435:93–102, 2001.
20. F. Waleffe. Homotopy of exact coherent structures in plane shear flows. *Physics of Fluids*, 15:1517–1534, 2003.
21. J. Wang, J.F. Gibson, and F Waleffe. Lower branch coherent states in shear flows: transition and control. *Physical Review Letters*, 98:204501, 2007.