Computational Methods in Applied Sciences

B. N. Chetverushkin · W. Fitzgibbon
Y. A. Kuznetsov · P. Neittaanmäki
J. Periaux · O. Pironneau *Editors*

Contributions to Partial Differential Equations and Applications

ECCOMAS
European Community
on Computational Methods
in Applied Sciences

Springer

Computational Methods in Applied Sciences

Volume 47

More information about this series at http://www.springer.com/series/6899

B. N. Chetverushkin · W. Fitzgibbon
Y. A. Kuznetsov · P. Neittaanmäki
J. Periaux · O. Pironneau
Editors

Contributions to Partial Differential Equations and Applications

Springer

Editors

B. N. Chetverushkin
Keldysh Institute of Applied
Mathematics (IPM)
Moscow
Russia

W. Fitzgibbon
College of Technology
University of Houston
Houston, TX
USA

Y. A. Kuznetsov
Department of Mathematics
University of Houston
Houston, TX
USA

P. Neittaanmäki
Department of Mathematical Information
University of Jyväskylä
Jyväskylä
Finland

J. Periaux
University of Jyväskylä
Jyväskylä
Finland

and

Edificio C-1, Campus Norte UPC
CIMNE
Barcelona
Spain

O. Pironneau
LJLL-UPMC
Paris
France

ISSN 1871-3033
Computational Methods in Applied Sciences
ISBN 978-3-319-78324-6 ISBN 978-3-319-78325-3 (eBook)
https://doi.org/10.1007/978-3-319-78325-3

Library of Congress Control Number: 2018935846

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Falaises in Normandy, by Chantal Periaux

Preface

Great things are done by a series of small things brought together.

Vincent Van Gogh

The present volume celebrates the seventieth birthday of three close friends Profs. William Fitzgibbon, Yuri Kuznetsov, and Olivier Pironneau. It is a compendium of papers presented at two conferences on the applied and computational mathematics. The first conference was a two-day event "Contributions to Partial Differential Equations" honoring the Profs. William Fitzgibbon and Yuri Kuznetsov, which was held in September 2015 at the Laboratoire Jacques-Louis Lions of the Universite Pierre et Marie Curie (former Paris VI). The second conference "Applied and Computational Mathematics," a two-day event honoring Prof. Olivier Pironneau, was held at the Department of Mathematics of University of Houston in February 2016.

The volume is comprised of 20 scientific contributions from the invited speakers of the two events and of the three career papers of the honorees. The contributors are internationally recognized experts in the areas of partial differential equations, applied mathematics, and computational mathematics. Profs. Fitzgibbon, Kuznetsov, and Pironneau have a longstanding cooperation in the domains of applied mathematics and computation. The fact that most speakers are close colleagues and/or have collaborated with honorees as well as the fact that many individuals attended both events provides unity to the volume. Indeed, the two events can be viewed as a single sequentially distributed event in two countries.

The papers are listed in alphabetical order by the name of the first listed author (in not all cases the invited presenter). A variety of topics and areas are addressed in this volume. These include modeling of computational fluid dynamics problems, mathematical models for the spatiotemporal spread of infectious disease, numerical tools for partial differential equations and scientific computing, reaction–diffusion systems, nonlinear elasticity, risk analysis, mathematical physics, optimization methods, and algorithms.

The volume incorporates innovative and advanced applied mathematics, sophisticated analysis, mathematical modeling with a high level of complexity and efficient computer implementation. As such, it is a testimony to the scientific endeavor and achievement of the honorees.

It has been the intention of the editors that the volume is of interest to researchers and practitioners as well as advanced students or engineers in applied and computational mathematics. All contributions have been written at an advanced scientific level with no effort made by the editors to make this volume self-contained. It is assumed that the reader is a specialist already who knows the basis of his field of research , has the capability of understanding and is appreciative of the latest developments of this volume.

Many individuals contributed to the success of the Conference in Paris and the International Workshop in Houston. Local arrangements for both events were professionally undertaken by the local committees in each locale. In particular, the following individuals should be recognized F. Hecht, Y. Maday, B. Perthame, O. Pironneau in Paris, W. Fitzgibbon, J. He, Y. Kuznetsov, M. Nicole, M. Olshanskii in Houston.

We acknowledge the thoroughly professional and diligent work done by Ms. Marja-Leena Rantalainen of Jyväskylä University regarding the collection of contributions and manuscript preparation. This volume could not have been produced without her dedicated efforts.

We finally express our gratitude to Ms. Nathalie Jacobs, Senior Springer Publishing Editor, for including the beautiful watercolor painting of the Normandy cliffs by the late Chantal Periaux. The locale holds a special place in the memory of all three individuals honored. In addition to Ms. Jacobs, we also recognize Prof. E. Oñate, CIMNE Director and Editor of the Series Lectures Notes, in Numerical Methods in Engineering and Sciences and thank both individuals for their support of this volume.

Jyväskylä, Finland
July 2017

P. Neittaanmäki
J. Periaux

Contents

Three Faces of Fitz: Science, Communication and Leadership

Jeff Morgan and Jacques Periaux

Abstract In what follows we provide a brief overview of the life and work of Professor William Fitzgibbon (University. of Houston)

People who know William Fitzgibbon bifurcate, with some calling him Fitz and others calling him Bill. Because both authors know him as Fitz, we will refer to him as Fitz. We suspect that this is his preferred appellation, since both his wife and daughter in law call him Fitz, and indeed Fitz reflects his Irish lineage, an importance aspect of the individual that we know.

Fitz is a complex person, and there are segments of his past that are best left to the imagination. He is a serious scholar, an elegant cowboy, a university statesman, a consummate gentleman, an athlete, somewhat of a rogue, and an unreconstructed rebel. Fitz has been known to enjoy quiet scholarship, concert halls, ballet performances, theaters, fine restaurants, sporting events, athletic activity, physical work, a few drinks, duck blinds, deer stands, and Harry's Bar in Paris. He is well-read, an expert with a chain saw, a strong swimmer, and a frustrated baseball player.

Fitz grew up in an intellectual home in Birmingham, Alabama. His father was a research chemist and Air Force officer, and his mother was both a high school teacher and classics scholar. Both parents were accomplished musicians, and interestingly, Fitz struggles to hum a tune. Fitz graduated from high school in 1963, and left Alabama to study at Vanderbilt University. His original intent was to study philosophy, not mathematics. However, he rapidly progressed from classical philosophy to at Vanderbilt to study Mathematical Logic. Shortly thereafter, Fitz became bored with the formalism of logic, and became attracted and engaged by the innovative teaching of Glenn F. Webb. He successfully defended his dissertation under the direction of Professor Webb in 1972. The topic of nonlinear evolution equations and monotone

J. Morgan (✉)
University of Houston, Houston, USA
e-mail: jmorgan@math.uh.edu

J. Periaux
University of Jyväskylä, Jyväskylä, Finland
e-mail: jperiaux@gmail.com

B. N. Chetverushkin et al. (eds.), *Contributions to Partial Differential Equations and Applications*, Computational Methods in Applied Sciences 47,
https://doi.org/10.1007/978-3-319-78325-3_1

and accretive operator theory was a considerable distance from his starting point. Interestingly, Fitz's mathematical career has followed an intellectual arc similar to his studies at Vanderbilt; namely from the abstract to the applied.

The America frontiersman Davy Crockett is reputed to have said,

> You all may go to Hell, and I will go to Texas.

In 1972, Fitz accepted a position at the University of Houston and moved to Texas. Many who have come to know view him as quintessentially Texan. Fitz has worked very hard to teach Yuri Kuznetsov "Texas English". Although he has enjoyed visiting appointments at the University of California at San Diego, Argonne National Laboratory, and the Universities of Bordeaux I and II, he has been a faculty member at the University of Houston over the course of his entire career.

With no pun intended, the corpus of Fitz's work shows continuous evolution. Fitz soon began to apply his abstract work in the area of nonlinear semigroups and evolution equations to study functional differential equations, integral and integro-differential equations, approximations, and singular perturbations. His work on the qualitative theory of semi-linear dissipative systems of partial differential equations led in a natural way to the study of systems of reaction diffusion equations and models of the spread of infectious disease, and creation and spread of atmospheric pollution. Over the course of his career, Fitz has had many collaborators. His major collaborator has been J. J Morgan. Their collaboration began in the 1980s and continues to this day. Their work on reaction diffusion systems produced a long term productive partnership on the spatio-temporal spread of infectious diseases with Michel Langlais in Bordeaux. By virtue of his interest in infectious disease, Fitz established a collaboration with his former advisor, Glenn Webb, with whom, curiously enough, he had not heretofore written a paper. Fitz also wrote several papers with Mary Parrott on such topics as singular perturbations of Hodgkin Huxley models, linearization techniques for partial functional differential equations and age dependent models for the spatial spread of infectious disease. Michel Langlais is not Fitz's only French connection. Roland Glowinski introduced Fitz to Jacques Periaux in 1984, when Roland was applying for a position at UH, and Jacques subsequently involved Fitz in an international project of creating a computational database for hypersonic and supersonic flows created by the atmospheric re-entry of space vehicles. This work enabled Fitz to develop friendships and alliances with Olivier Pironneau, Pierre Perrier, and Antoine Desideri at INRIA Sophia Antipolis, as well as a deep appreciation of the Auberge du Jarrier in the village Biot. These efforts also provided Fitz with an introduction to his good friend and current colleague Yuri Kuznetsov. Other major collaborators of Fitz include Sheila Waggoner, Youcheng You, and Selwyn Hollis.

Fitz has authored well over 120 research papers, plus numerous articles, reviews, and reports. He has edited or co-edited eight volumes, and served on the editorial boards of numerous scientific journals. In additional to his mathematical work, Fitz has received accolades for publications in Technology. In this area, his co-authors include Enrique Barbieri, Heidar Malki and Rupa Iyer. Fitz has lectured extensively in North America, South America, Africa, Europe, and Asia. He has a history of funding–over the course of his career has received funding from a variety of agencies, including the National Science Foundation, the Office of Naval Research, the National Institute of Standards and Technology, the US Department of Education, the Environmental Protection Agency, the US Department of Labor, the Texas Advanced Research Program, NASA, the Texas Workforce Commission, and the Texas Education Agency.

Fitz played a pivotal role in transforming what was a myopic and mediocre department focused on pure mathematics, to an internationally recognized center of applied and computational mathematics. He organized or co-organized a national conference on nonlinear diffusion (with Homer Walker), a special year on partial differential equations and dynamical systems, two international conferences on the mathematics of hydrocarbon discovery and exploitation (with Mary Wheeler), and seminars in both the mathematics of the oil and gas industry (with Peter Purcell) and mathematical biology (with Marek Kimmel), and both a workshop and an international conference in high speed flow fields (with an international network of researchers created by Jacques Periaux), and computational databases as well as numerous regional conferences and workshops in partial differential equations. In 1997 Fitz and Jacques Periaux organized and the signature conference *Computational Science of the 21st Century* honoring the 60th birthday of Roland Glowinski, in Tours, France. Although there is a tradition in mathematics of commemorating the birthdays of prominent senior mathematicians, an event of this scale is rarely seen. It is also notable that in his role as emcee of the banquet, Fitz spoke in French, despite the fact that Fitz is well known by his French friends for butchering the language. Fitz and Professor Periaux continue to collaborate on organizing events and research projects in Transport applications in Europe, and also with Prof. Cliff Dacso in Molecular and Cell Biology at Baylor College of Medicine in Houston.

In discussing his career it is important to emphasize the role of Fitz as an academic leader. Fitz served as President of the University of Houston Faculty Senate from 1999 to 2000. The turn of the century was time of tumult at the University of Houston, and Fitz assumed leadership of group of senior faculty members who laid the groundwork for the transformation the University of Houston from a regional doctoral institution to a nationally recognized comprehensive research university. This effort required political and consensus building skills that are not frequently demonstrated by mathematicians. On a more local level, Fitz served as Chair of the Department of Mathematics from 1999 to 2003, and as Interim co-Head of the Department of Computer Science from 2000 to 2001. Both the Mathematics Department and the Department of Computer Science made major strides under Fitz's leadership. In 2003 Fitz assumed the role of Dean of the College of Technology. This was indeed an unusual move, because Fitz became dean of a college that did not include his own discipline. In the stove piped structure of American universities, such moves are virtually unknown. His work had major impact. When Fitz took the helm of Technology, it was a struggling entity with about 1400 students–when he stepped down as Dean in 2016, the college enrolled over 6000 students, the graduate program had flourished, faculty productivity had dramatically improved, and there was a multifold increase in both external funding and private donations. It is fair to say that Fitz's leadership saved the College of Technology. In his role as Dean, Fitz became very involved with University of Houston Alumni and Athletics. He has also became involved in the Houston arts community and engaged in a wide variety of civic activities. Among other activities, he served on the Board of Directors of Houston's East End Chamber of Commerce, the Board of Trustees of the Houston Maritime Museum and the Advisory Board of Directors of the Houston Technology Center.

Fitz is an intrepid and inveterate traveler who has visited over 50 different countries. He always welcomed the opportunity to meet new people and immerse himself in different cultures. In particular, four countries stand out. To begin we should recall the line from the movie "Casablanca"

> There will always be Paris.

In Fitz's case we should probably adjoin Bordeaux, and perhaps the Cote d'Azur and Normandy. Fitz has spent considerable time in France and has essentially traversed the entire country. Despite his obvious linguistic challenges he has immersed in and developed a deep appreciation for French culture. He relishes the time that he spent in Bordeaux both from an academic and a cultural standpoint. Indeed he has often said that he learned to appreciate good whiskey from his father and honed his appreciation of fine wine and Armagnac with Michel Langlais. India also has been one of his frequent destinations. Fitz recounts how he and his wife were both surprised by how immediately comfortable they felt upon landing in India. Fitz worked with Dinesh Singh to put in place an innovative and often copied program to bring students from India to Houston for graduate study in mathematics, and later computer science and physics. Fitz has continued this work in India and is credited with opening the doors of the University of Houston to India. In recent years Fitz has made numerous trips to Africa. He is justifiably proud of his work in establishing the International University of Grand Bassam in the Cote d'Ivoire. He serves on the Board of Directors of the Institution and chairs the Academic Committee of the Board. He also chairs the board of the US based foundation that supports the institution. Finally, there is Finland. By virtue of his association with Pekka Neittaanmäki, Fitz has frequently visited Jyväskylä and has been able to enjoy the quiet beauty of Finnish winters and opportunity for Nordic skiing. However, he has yet to go caribou hunting.

Those who know Fitz well are aware of his wicked, and often politically incorrect sense of humor. As a Dean, Fitz often had to suppress this. However, now that he has relinquished his diaconal duties, we are pleased to see it reemerging. For many years Fitz was a well-known Houston *bon vivant* and regularly held court at Houston's now defunct River Café. Indeed prior to the ubiquity of the cell phone, scientists across the world learned to call for Fitz at the River Café if they could not reach him in his office. Fitz was rarely home.

Family man might not be the first phrase that would come to mind in describing Fitz, and yet, he is very proud and dedicated to his family. In 1999 Fitz married Jan Brooks, settled down and turned the page of another chapter in his life-au revoir River Café! Jan is a former ballet dancer and instructor who is known for her beauty, grace and Southern charm, as well her infectious laugh. Their son and daughter in law and three grandsons live in Nashville, and we hope that the boys will not cause as much trouble as Fitz did in his younger days.

Unlike *Family Man*, the word *Friend* is one that readily comes to mind, and Fitz sets high standards for others in this arena. He is caring, honest and loyal. Fitz listens well, he is sincere in his compassion, and the light of his spirit shines through in the darkest hours. With all of his success in life, his human character is perhaps his greatest strength. Mark Twain said

> No man is a failure who has friends.

We believe he was speaking of Fitz, well before his time.

Career of Prof. Yuri Kuznetsov

Boris Chetverushkin, William Fitzgibbon and Jacques Periaux

Abstract In what follows we provide a brief overview of the life and work of Professor Yuri Kuznetsov (University of Houston)

As we sit down to trace the course of the career of our friend Prof. Yuri Kuznetsov we are reminded of the words of one of his favorite authors, the American writer Mark Twain:

> Biographies are the clothes and buttons of the man. The biography of the man cannot be written.

This being said we must admit that Yuri Kuznetsov had indeed an impressive tailor. Yuri was born in small village Shuksha of the region Penza on August 7, 1945. He graduated from high school in 1962 and in the same year commenced his studies with the Faculty of Physics at the newly established Novosibirsk State University. He graduated from the university in 1967 and successfully defended his Ph.D. thesis in 1969 under the direction of the outstanding Russian applied mathematician, G. Marchuk. In his thesis Yuri developed the theory of iterative methods in subspaces. In particular this innovative work proposed and investigated the multi-step method based on the minimization of quadratic functional which become commonly known as the generalized minimum residual method. In 1971 Prof. J. -L. Lions invited Yuri to visit France and give a lecture at IRIA (now INRIA). In 1983, Yuri reported his results as invited speaker in Warsaw at the International Congress of Mathematicians, Section of Numerical Analysis. Prof. Kuznetsov has held the MD Anderson

B. Chetverushkin
Keldysh Institute of Applied Mathematics, Russian Academy of Sciences, Moscow, Russia
e-mail: chetver@imamod.ru

W. Fitzgibbon (✉)
University of Houston, Houston, USA
e-mail: wfitzgib@central.uh.edu

J. Periaux
University of Jyväskylä, Jyväskylä, Finland
e-mail: jperiaux@gmail.com

B. N. Chetverushkin et al. (eds.), *Contributions to Partial Differential Equations and Applications*, Computational Methods in Applied Sciences 47,
https://doi.org/10.1007/978-3-319-78325-3_2

Professorship in Mathematics at the University of Houston since Fall 2000—a Chair that has been held by Professor Roland Glowinski and Prof. Mary F. Wheeler among others. Prof. Kuznetsov has played and continues to play a pivotal role in the development of applied and computational mathematics at the University of Houston. His seminal contributions to numerical analysis particularly numerical linear algebra and domain decomposition are recognized worldwide. Of particular note is work in iterative methods and solvers, preconditioned conjugate gradient methods, domain decomposition, fictitious domain methods, and adaptive meshes. Over the course of his career, Prof. Kuznetsov has authored or co-authored four books and well over 100 research papers. Yuri Kuznetsov is a great man—great for his achievements not only in applied and computational mathematics but in life as well.

Yuri's role as one of the founding editors of the East West Journal of Numerical Mathematics, currently Journal of Numerical Mathematics, is symbolic of as well as presages Yuri long effort of facilitating of East West understanding and scientific cooperation. While some may articulate and perhaps fulminate, implementation requires trailblazers possessing the vision, energy, and courage of Yuri Kuznetsov. In the early 1970s, Yuri was one of the key members of a delegation led by Prof. Marchuk and Prof. Yanenko visiting a French team at INRIA Rocquencourt. The leader of the host team on the French side was Prof. J. -L. Lions with R. Glowinski and later O. Pironneau being counterparts of Yuri in the French Russian scientific collaboration. This visit was subsequently followed by a series of French-Russian workshops organized and successfully run in Moscow, Paris, Sophia Antipolis, Tashkent, Marseilles, and Jyväskylä. Repetitive scientific exchanges between Russia and France INRIA (then directed by Prof. A. Bensoussan) and the Lomonosov University of Moscow resulted in the establishment of the Liapunov Center, a center for bilateral collaborative projects in numerical mathematics, automation and computer sciences. Yuri frequently visited Paris meeting with P. Bohn, P. Perrier and J. Periaux of Dassault Aviation, and O. Pironneau of the University of Pierre et Marie Curie. Yuri could apply his expertise in Maxwell's Equations, linear algebra and preconditioning methods to develop computational electromagnetics software. During the 1990s teams of Russian scientists lead by Yuri and Prof. Boris Chetverushkin undertook extensive interactions with their French counterparts. This effort resulted in major activity on the modeling and simulation of the high speed flows for reentry problems around space vehicles.

In 1980, Yuri moved from Novosibirsk. In Moscow he closely collaborated with outstanding Russian numerical mathematicians N. Bakhvalov, V. Lebedev, and V. Voevodin. During the 1980s Professor Kuznetsov became recognized as a leader in effort to compute solutions to partial differential equations. Because the sophistication computational infrastructure available in former Soviet Union lagged behind that available in Western Europe and the United States, his methods needed to be more clever and efficient than those available in the West. Professor Kuznetsov assembled an impressive team of young scientists who were capable of capitalizing on his innovations and implementing his algorithms computationally on the machines available at that time. Yuri and his team systematically visited centers of learning in variety of European countries including Finland, Czechoslovakia, Germany, Italy, Austria and France. The importance of his research on fictitious domains and domain decomposition was rapidly recognized by his European colleagues. During this time period Prof. Kuznetsov met and be became friends with R. Glowinski, O. Pironneau, J.A. Désidéri, M. Feistauer, P. Neittaanmäki, R. Rannacher, R. Hoppe, O. Axelsson, F. Brezzi, and A. Quarteroni among many others scientists.

By the 1990s Yuri's reputation was established and the importance of his scientific contributions were internationally recognized. Yuri worked with his European network to host and initiate numerous scientific meetings. The Domain Decomposition Methods Conferences were successfully launched in Paris in 1987. The 4th DDM Conference was held in 1990 in Moscow, a pivot event linking Soviet scientists to their counterparts organizing. His work in organizing the 4th DDM. Conference in Moscow in 1990 led Yuri Kuznetsov to the conclusion that European computational sciences conferences did not have a sufficiently strong mathematical component. In response he organized together with a group of European colleagues the highly successful ENUMATH series which also continues to this day.

In 1997, Yuri Kuznetsov made a major and courageous decision moving to the United States and joining the faculty of the University of Houston as a professor of mathematics. The path of Yuri Kuznetsov to Houston was lined with suitors runs through Rome, Pavia, Paris, Prague, Tokyo, Nijmegen, Lyon Milano, Jyväskylä as well as New York, Palo Alto, Augsburg, Los Angeles, Heidelberg, Sophia, Denver, Austin, College Station, Laramie, and West Lafayette, Indeed our friend Yuri has been a belle with many beaux. Although Yuri only became conversant in English in mid 1980s, facility with English as opposed to French may have convinced Yuri to choose Houston over Paris. Yuri (with W. Fitzgibbon as a mentor) has rapidly become fluent in Texas English as well.

Prof. Kuznetsov's activities in Houston mirror his work in Moscow. He has assembled a team of young researchers and has lead the effort build strength in computational and applied mathematics in Houston. He expanded his network of scientific collaborators and associates by connecting researchers in the United States with to network he built in Russia and Europe. His expanded network included: D. Young, M. Wheeler, R. Ewing, G. Golub, R. Varga, J. Douglas and R. Glowinski. As a consequence, the University of Houston is now recognized as an international center of scientific computation. He continues to build his legacy with superb Ph.D. graduates. He continues to produce highly innovative work and has been successful in attracting support from both federal agencies and the oil and gas industry, in particular Exxon/Mobil. Most recently he has focused on new discretization method for diffusion equations in heterogeneous media with general polyhedral meshes, nonconforming mixed finite element discretization on polyhedral meshes, monotone discretizations for diffusion and convection diffusion equations and computational basin modeling.

In a sense Yuri never left Russia but brought Russia to Houston. In 1998 Yuri led a delegation of Houston scientists to Russia. The University of Houston now has a sizeable Russian contingent and one frequently hears Russian along its corridors. As a result of this visit a pioneering agreement of understanding and cooperation was put between Lomonosov University and the University Houston was signed.

The Department of Mathematics has profited from a steady stream of visitors from Russia, in particular from a group of young researchers in the Institute of Numerical Mathematics in Moscow led by a former Yuri's student Yuri Vassilevski The Russian language is commonly heard in the corridors of the mathematics building in Houston and more than one Houston scientist has referred to the University of Houston Mathematics Department as "Moscow on the Bayou".

In addition to being a great scientist, Yuri Kuznetsov is a great human being. His friends and colleagues appreciate not his scientific achievement and professionalism but also but also enjoy his human skills, his generosity, kindness and spontaneous hospitality in Moscow, Houston and elsewhere. Professor Jeffrey Morgan, the previous Chair of the Mathematics Department of the University of Houston stated it well when he said:

> Kuznetsov is an amazing person, extremely hardworking, and totally passionate about his work. He is also one of the most unselfish people I have met; you rarely meet people you are so impressed with.

According to the recent book, *Living the Good Long Life: A Practical Guide to Caring for Yourself and Others* the age 70 has become the new 50. By this new standard Yuri has recently become middle aged and has just approached his prime.

Yuri your friends and colleagues still have high expectations and are looking forward to following your itinerary around the world wherever it leads—Moscow, Paris, Tokyo, Prague, Houston or any other destination. We would like to wish you, your lovely wife Ludmila, your children and your grandchildren good health, and prosperity as you continue life's journey. In the words of Charles Darwin, a man's friends are the best measures of his worth. This makes you a truly wealthy man.

Olivier Pironneau Career Paper

William Fitzgibbon and Jacques Periaux

Abstract In what follows we provide a brief overview of the life and work of Professor Olivier Pironneau Fitzgibbon (LJLL Sorbonne Université)

The American Poet and Essayist Ralph Wald Emerson wrote

> Do not go where the path may lead. Go instead where there is no path and leave a trail.

After completing his diploma at the prestigious Ecole Polytechnique in 1968 Olivier followed a route not typically followed by France's intellectual elite and sailed west to the University of California, Berkeley. Berkeley in the 60s was the epicenter of the counter culture revolution. Olivier chose. Although I am sure that Olivier was not oblivious to all was happening, he did not succumb entirely the siren song of Sex, Drugs and Rock and Roll. Instead he maintained his focus on what drew him to Berkeley. Olivier was originally drawn to Berkeley by an interest in control systems and electronics. These were halcyon days control systems and electronics. Much of what is now called applied mathematics took place in the area of control systems. Elijah Polak with his expertise in computer based optimization with applications to electronic circuit design, control system design, and structural design, optimization algorithms, and systems theory, was the Ph.D. mentor for Olivier.

Many students who complete Ph.D.'s at bellwether American universities under the tutelage of well known advisors are content to follow the trail blazed by the advisors to regarding careers. Not so with Olivier Pironneau. Olivier learned that the legendary James Lighthill had an interest in possible applications of control theory to fluid dynamics and aero acoustics. So Olivier with absolutely no prior knowledge

W. Fitzgibbon
University of Houston, Houston, USA
e-mail: wfitzgib@central.uh.edu

J. Periaux (✉)
University of Jyväskylä, Jyväskylä, Finland
e-mail: jperiaux@gmail.com

J. Periaux
CIMNE, Barcelona, Spain

B. N. Chetverushkin et al. (eds.), *Contributions to Partial Differential Equations and Applications*, Computational Methods in Applied Sciences 47,
https://doi.org/10.1007/978-3-319-78325-3_3

of fluid mechanics went east this time to Cambridge University in the UK to work in the Department of Applied Mathematics and Theoretical Physics as a Post Doctoral Fellow working on optimal shape in Fluids. Olivier did indeed become well versed in fluid mechanics through his work and talking to people and perhaps with no pun intended osmosis. Olivier was now well positioned for a career in fluid mechanics and aviation. However, another uncharted path awaited.

We now see electrical engineering, control theory and fluid mechanics in Olivier's back ground but where does mathematics come in? Back in France, Oliver had a serendipitous chance meeting with the preeminent French Mathematician Jacques-Louis Lions on a train. The mathematical work of J. Louis Lions is accurately described by the title—which he chose—of his chair at the Collège de France: "Analyse Mathématique des Systèmes et de leur Contrôle." The systems he had in mind are those described by linear and nonlinear partial differential equations; by analysis he meant everything from the most abstract existence theorems along with underpinning functional analysis and Sobolev Space theory, approximation and numerical issues and to computer implementations. In retrospect Lions must have been deeply impressed by what Olivier had. He asked Olivier for some papers and subsequently offered him a research position at the Institut de Recherche en Informatique et en Automatique. Lions continued to be impressed with Olivier and Olivier soon defended his Thèse d'Etat at University Paris 6 [renamed Université Pierre et Marie Curie (UPMC) in 1974].

There is still more to be told about the path of Olivier. J. -L. Lions urged Pironneau to pursue an academic career and Oliver soon accepted a position at Paris Nord. A precondition for this position was to teach computer science. So Olivier strode forward once again into unchartered territory in this case the theory of computer science and compilation. One telling data point about Olivier is his statement that best way to learn a subject is to teach it. His effort of learning computer science gave birth to the idea of having a user-friendly language for people who work with Partial Differential Equations; Free fem++ was born. Olivier transferred to the University Paris 6. In order to teach a course on computer science tools for applied mathematics. Olivier together with a colleague having a strong background in software development rewrote together the MacFem, FreeFem in C++ and then gave it away as open source. This was the first of this type of software which is popular now. You can download it from www.freefem.org.Leadership on this project has been taken over by Frederic Hecht with whom Olivier shares an office and have a marvelous outcome.

So now we have a map of the areas through which Olivier has marked his trail. He is one of the very few individuals if not unique who integrate knowledge of control systems, fluid mechanics, modern partial differential equations, numerics, computer science and software development. It is interesting that Oliver is now working in the mathematics of finance. An interest in the role control theory could play in mathematical economics lead him to Berkeley. I guess uncharted paths can lead us home or as Andre Gide said

> It's only in adventure that some people succeed in knowing themselves—in finding themselves.

One who knows anything about Olivier knows that India and Indian thought is an important component of his story. Olivier first became interested in India as a result of yoga classes at Cambridge. This led to Olivier's ongoing spiritual quest in India. Olivier was introduced to the thought of Sri Aurobindo by his Indian mentor (more properly "guru") M. L. Parahsar. Sri Aurobindo was a celebrated Indian nationalist, poet, philosopher, and spiritualist known for his philosophy on human evolution and his development of Integral Yoga. In 1910 Sri Aurobindo retired from public life and settled in Pondicherry where he dedicated himself to philosophical and spiritual pursuits. Olivier maintains a residence in Pondicherry which he continues to visit regularly as he follows his inner guide.

Olivier's passions are not only of the spirit and intellect. He maintains his affinity for rock music. He is an avid skier (both downhill and cross country). He is an alpine hiker and mountain climber (he experienced life threatening hiking adventures with his friend Claude Bardos!). Finally in the words of Shri Aurobindo,

> Life is Yoga

He has a lovely companion Annette who shares his Indian spiritual life in Pondicherry. He takes great pleasure and pride in the growth and education of his son Gabriel.

Over the course of his career Olivier has authored or co-authored over 300 papers and 8 books. He is a true scholar with deep scientific knowledge in a variety of scientific areas including, partial differential equations, mechanics, physics, computational fluid dynamics, aeronautics, financial mathematics, computer science and computation. His mastery of computer tools as a hobby both hardware and software coupled with the right mathematics has given and continues to give him a remarkable ability to solve outstanding problems in science and engineering. Indeed it is fair to say that not only is Olivier the quintessential computation scientist but that he also defined the term two decades before it became fashionable. Among the numerous accolades bestowed on Olivier are: the Prix Blaise Pascal from the French Academy of Science in 1983; the Ordre National du Merite also in 1983; and the Prix Marcel Dassault awared to him in 2000 by the French National Academy of Sciences. In 2002 he was inducted into the French National Academy of Sciences and he became an associate member of the Russian Academy of Sciences in 2004. The "Ordre National de la Legion d'Honneur Grade de Chevalier" was bestowed upon him in 2006.

Countless students have been inspired by Olivier—from his lectures, classes and from his research supervision. Ph.D. students to Olivier—he has had 30. Here as in other matters Olivier follows a somewhat different path. Inspiration is two way street for Olivier. Not only does he strive to inspire students, he seeks inspiration from his students and gathers immense joy in the discovery of talent and genius in students. The success of these students continues to create his legacy.

Professor Pironneau has served the scientific community in a variety of ways. In France he served as a Deputy for external affairs of INRIA and as a Project leader for Institut de Recherche en Informatique et en Automatique in the famous Bâtiment

16 at Rocquencourt; from 1996 to 2006 he served as Member of the Commission on Nuclear Safety: currently he serves as President the French National Strategic Committee for Super Computing and he chaired the national association GAMNI (Groupe pour l'Avancement des Methodes Numeriques de l'Ingenieur) in the late 1990s. Olivier played a crucial role in the creation and officially launched in 1993 of the European Association ECCOMAS (European Community for Computational Methods in Applied Sciences) serving from 1992 to 2004, as a member of the Board. helped to establish academic programs of study in Computing in India in the late 1990s. He organized in 1982 a Winter School at Bangalore on Applied Mathematics opened by the French President, F. Mitterand and has been a regular visitor, as a member of the Scientific Advisory board of IMS, 2008–2012 to the Institute for Mathematical Sciences (IMS), National University of Singapore. In addition, he is a member the editorial board of several journals including the Comptes Rendus de l'Académie des Sciences.

The above achievements in Olivier's career are not reported exhaustively: many of them are missing and should be mentioned in the bio of an outstanding applied mathematician.

It has been the good fortune of the second author, to meet Olivier in the 70' at the Golden Age of the CFD, so important during this period for the design of airplanes at Dassault Aviation. Despite younger than him, Olivier provided him mentorship, encouragement and also inspiration control on one side, computation of functional numerical gradient in optimization on the other side. He was honored to chair with him in the late 90' the Pole Scientifique UPMC–Dassault Aviation, launched by Prof. J. -L. Lions and is grateful for his warm friendship when he lost his life in 2008…

Olivier, throughout your scientific career as Professor and Researcher you have built an international network of friends and colleagues all over the world (Houston, Jyväskylä and Pondicherry, three frequently visited scientific hubs of the network among many others…). There is no end to your unchartered paths in this network by aging because to your high energy and "good spirit".

Samuel Ullmann, in his poem "Youth", says:

> Youth is not a time of life; it is a state of mind; it is a matter of the will, a quality of the imagination, a vigor of the emotions; youth means a temperamental predominance of courage over timidity of the appetite, for adventure over the love of ease.

Olivier, this will undoubtedly follow you for many years to come along the multi itineraries of your network with a continuous successful inspired research and exciting computational activities!

We wish you, Olivier, “good health”, “good luck” and “happiness” with Annette, Gabriel and your friends.

Lord Krisna will guide you by his music…

Mean Field Games for Modeling Crowd Motion

Yves Achdou and Jean-Michel Lasry

Some people go to priests; others to poetry; I to my friends.
Virginia Woolf (The waves)

Abstract We present a model for crowd motion based on the recent theory of mean field games. The model takes congestion effects into account. A robust and efficient numerical method is discussed. Numerical simulations are presented for two examples. The second example, in which all the agents share a common source of risk and have incomplete information, is of particular interest, because it cannot be dealt with without modeling rational anticipation.

1 Introduction

It is more and more important to forecast crowd motions, particularly in situations of panic, since this aspect is now taken into account for the certification of buildings and infrastructures, see for example [10]. There is a huge literature on models of human crowd motions: some of them, inspired by classical Newtonian mechanics, see the agents as particles and interactions as shocks between the particles, see for example [11] and references therein. In such models, the global tendency of the agents consists of reaching some goal as fast as possible, but their dynamics at fine scale depends on their interactions with their closest neighbors. Macroscopic descriptions can then be derived by upscaling the previously mentioned microscopic models, see, e.g., [13]. Arguably,

Y. Achdou (✉)
Université Paris Diderot, Sorbonne Paris Cité, Laboratoire Jacques-Louis Lions, UMR 7598,
UPMC, CNRS, 75205 Paris, France
e-mail: achdou@ljll.univ-paris-diderot.fr

J.-M. Lasry
CEREMADE, Université de Paris-Dauphine, Paris, France
e-mail: 2007lasry@gmail.com

B. N. Chetverushkin et al. (eds.), *Contributions to Partial Differential Equations and Applications*, Computational Methods in Applied Sciences 47,
https://doi.org/10.1007/978-3-319-78325-3_4

all these models do not contain any ingredient from game theory and therefore do not really take rational anticipation into account.

In this paper, we propose to apply the recent theory of mean field games to crowd motion; mean field type models describing the asymptotic behavior of stochastic differential games (Nash equilibria) as the number of players tends to $+\infty$ have been introduced by J.-M. Lasry and P.-L. Lions in 2006, see [16–18]. In some cases, they lead to systems of evolutive partial differential equations involving two unknown scalar functions: the density of the agents in a given state x, namely $m = m(t, x)$ and the value function $u = u(t, x)$. Since the present work is devoted to crowd motion, we will assume that the dimension of the state space is $d = 2$.

The present work was finished and presented in several scientific meetings in 2013, but we did not write the report since then, for lack of time. Other papers have dealt with mean field games for pedestrian flows, see, e.g., [9, 15], but some aspects of our work are completely original. Indeed, we are going to address two topics which are not treated in the existing literature:

- show the influence of the structure of information upon the motion of a crowd, see point 3 below for more details;
- present efficient and reliable numerical methods which can be applied to mean field games in which the noise affect macroscopic quantities and therefore all the agents in the same way.

More precisely, we shall discuss

1. a special model for taking the effect of congestion into account, which was introduced by P.-L. Lions, see [19];
2. a numerical scheme which keeps the structure of the system of PDEs and for which convergence can be proved, see [1, 3, 7] for models without congestion;
3. an example of a situation which cannot be modeled without taking rational anticipation into account.

In this example, the behavior of the agents depends on the incomplete information that they have on their future: all the agents are affected by the same random events (here the opening of a door at a given time), and they anticipate future having incomplete information on the game (here everybody knows that one among several doors will be opened but not which one). The structure of information has therefore a crucial influence on the crowd behavior, and makes it impossible for mechanistic models to predict the latter. The example, which will be described in Sect. 5.2, was inspired by some models arising in the theory of economical growth, namely the theory of Krussel-Smith in macro-economics, see [5, 14], in which random shocks on macroscopic quantities affect the whole economy. Several examples of PDEs and mean field games models in economics, including the Krussel-Smith model, are described in [2].

2 The Model

Let Ω be a bounded connected open subset of $\mathbb{R}^2$ with a polygonal boundary. The boundary of Ω, i.e. $\partial\Omega$ is partitioned into $\partial\Omega = \overline{\Gamma_N} \cup \overline{\Gamma_D}$, where Γ_N and Γ_D are two disjoint open subsets of $\partial\Omega$.

The typical system of partial differential equations that will be considered is

$$\frac{\partial u}{\partial t}(t,x) + \nu\Delta u(t,x) - H(x, m(t,x), \nabla u(t,x)) = -F(m(t,x)), \tag{1}$$

$$\frac{\partial m}{\partial t}(t,x) - \nu\Delta m(t,x) - \operatorname{div}\left(m(t,\cdot)\frac{\partial H}{\partial p}(\cdot, m(t,\cdot), \nabla u(t,\cdot))\right)(x) = 0, \tag{2}$$

in $(0,T) \times \Omega$, with the initial and terminal conditions

$$u(T,x) = u_T(x), \quad m(0,x) = m_0(x) \quad \text{in } \Omega \tag{3}$$

given a terminal cost function u_T and an initial probability density m_0.

Here, we denote by ν a nonnegative constant and by Δ, ∇ and div, respectively, the Laplace, the gradient and the divergence operator acting on the state variable x. In the cost term $F(m(t,x))$, F is a $\mathscr{C}^1$ regular function defined on $\mathbb{R}_+$.

The system also involves the scalar Hamiltonian $H(x, m, p)$, which is assumed to be convex with respect to p and $\mathscr{C}^1$ regular w.r.t. x, m and p. The possible dependence of the Hamiltonian on the density variable m allows for modeling congestion effects, i.e. the fact that the cost of motion at $x \in \Omega$ is an increasing function of $m(x)$. In particular, we will focus on Hamiltonians of the form

$$H(x, m, p) = \mathscr{H}(x) + \frac{|p|^\beta}{(c_0 + c_1 m)^\alpha} \tag{4}$$

with $c_0 > 0$, $c_1 \geq 0$, $\beta > 1$ and $0 \leq \alpha < 4(\beta - 1)/\beta$. The potential $\mathscr{H}(x)$ is a smooth function of the state variable. The notation $\frac{\partial H}{\partial p}(x, m, q)$ is used for the gradient of $p \mapsto H(x, m, p)$ at $p = q$.

We have chosen to focus on the case when the cost $u_{|t=T}$ depends directly on x. In some realistic situations, the final cost may depend on the distribution of the players, i.e. $u_{|t=T} = \Phi_T[m_{|t=T}](x)$, where Φ_T is an operator acting on probability densities, which may be local or not. We will not discuss this aspect in the present work.

The system is complemented with Neumann boundary conditions

$$\frac{\partial u}{\partial n}(t,x) = 0, \quad \frac{\partial m}{\partial n}(t,x) = 0 \tag{5}$$

in $(0,T) \times \Gamma_N$, and Dirichlet boundary conditions

$$u(t,x) = u_D(x) \quad m(t,x) = 0 \tag{6}$$

in $(0,T) \times \Gamma_D$, where u_D is a given exit cost.

The Hamiltonian in (4) is of the form

$$H(x,m,p) - F(m) = \sup_{\gamma} \left[\gamma \cdot p - L(x,m,\gamma)\right],$$

with

$$L(x,m,\gamma) = (\beta - 1)(c_0 + c_1 m)^{\frac{\alpha}{\beta-1}} \left(\frac{|\gamma|}{\beta}\right)^{\frac{\beta}{\beta-1}} + F(m) - \mathscr{H}(x). \tag{7}$$

Hence, Dynamic Programming arguments, see Bardi-Capuzzo Dolcetta [8], Fleming-Soner [12], show that u is the value function of an optimal control problem for the controlled dynamics defined on Ω by

$$dX_s = -\gamma_s\, ds + \sqrt{2\nu}\, dW_s$$

((W_s) is a Brownian motion reflected on the Neumann boundaries), running cost density

$$L(X_s, m(s, X_s), \gamma_s)$$

and exit cost u_D. The term $-\mathscr{H}(x)$ is the instantaneous cost for an agent to stand at x: a positive and large value of $-\mathscr{H}(x)$ means that x is not a comfortable location. If F is an increasing function, then the term $F(m)$ models crowd aversion (or agoraphobia). Simple models of panic may be constructed by choosing $\mathscr{H}$ with negative and large values in Ω and F increasing and blowing up at $+\infty$: the agents pay a high cost for staying in Ω and this cost is made higher in crowded places.

The first term in (7) stands for the cost of motion: we see that the denser the population is, the more expensive (or difficult) motion becomes. This is precisely what we mean when we speak of *congestion effects*.

Existence and uniqueness have been studied by P.-L. Lions in his lectures (in French) at Collège de France, see [19] for the videos: in particular, it was proved that uniqueness of a classical solution holds if F is an increasing function and if $\alpha \leq 4\frac{\beta-1}{\beta}$, at least in the simpler situation when $\Gamma_D = \emptyset$.

3 Finite Difference Method

For the numerical simulations, we use a finite difference scheme proposed and tested in [4] for a model with no congestion and periodic boundary conditions. We briefly sketch the method.

3.1 Description of the Scheme

Let N_T be a positive integer and $\Delta t = T/N_T,\ t_n = n\Delta t,\ n = 0, \ldots, N_T$. Let $Q = (0,d)^2$ be a square domain in $\mathbb{R}^2$ containing Ω. Let Q_h be a uniform grid on the square $\overline{Q}$ with mesh step $h = \frac{d}{N_h+1}$ and x_{ij} denote a generic point in Q_h, for $0 \leq i, j \leq N_h + 1$. It is implicitly assumed that the index i stands for the x-axis and the index j for the y-axis.

We assume that the boundary of Ω is made of straight line segments which are parallel to the axes and coincide with some lines of nodes in Q_h. The grid Ω_h is obtained as the restriction of Q_h to Ω. For those indices such that $x_{i,j} \in \Omega_h$, the values of u and m at $(x_{i,j}, t_n)$ are respectively approximated by $u^n_{i,j}$ and $m^n_{i,j}$.

Let u^n (resp. m^n) be the vector containing the values $u^n_{i,j}$ (resp. $m^n_{i,j}$), for i, j such that $x_{i,j} \in \Omega_h$ indexed in the lexicographic order. We may refer to such vectors as *grid functions*.

3.1.1 Elementary Finite Difference Operators

First of all, we must make some conventions about indexing the nodes and the unknowns in order to deal with the boundary conditions. To cope with Neumann type conditions, we use a first order finite difference formula: for example, at a boundary node $x_{i,j}$ for which $x_{i+1,j} \in \Gamma_N$, we impose that $u_{i+1,j} = u_{i,j}$. It would be possible to require that $u_{i+1,j} = u_{i-1,j}$, which would lead to a higher order scheme for the diffusion part of the operator, but this would not be very useful since the discrete version of the Hamiltonian is first order only, and we would lose the monotone character of the scheme.

In the simple case when $\Omega = Q$ and $\Gamma_N = \partial\Omega$, this technique can be summarized by the following relations:

$$u_{0,j} \equiv u_{1,j}, \quad u_{N_h+1,j} \equiv u_{N_h,j}, \quad u_{i,0} \equiv u_{i,1}, \quad u_{i,N_h+1} \equiv u_{i,N_h}.$$

The treatment of Dirichlet boundary condition is straightforward: if $x_{i,j}$ is a node on Γ_D, we set $u_{i,j} = u_D(x_{i,j})$.

We apply the finite difference scheme only at the nodes belonging to Ω_h and use the conventions above when the finite difference stencil involves nodes outside Ω_h.

Using these conventions, the difference operators

$$(D_1^+ u)_{i,j} = \frac{u_{i+1,j} - u_{i,j}}{h} \quad \text{and} \quad (D_2^+ u)_{i,j} = \frac{u_{i,j+1} - u_{i,j}}{h}$$

can be defined at any point in Ω_h. We define $[D_h u]_{i,j}$ as the collection of the four possible one sided finite differences at $x_{i,j}$:

$$[D_h u]_{i,j} = \Big((D_1^+ u)_{i,j}, (D_1^+ u)_{i-1,j}, (D_2^+ u)_{i,j}, (D_2^+ u)_{i,j-1} \Big) \in \mathbb{R}^4.$$

We will also need the standard five point discrete Laplace operator

$$(\Delta_h u)_{i,j} = -\frac{1}{h^2}(4u_{i,j} - u_{i+1,j} - u_{i-1,j} - u_{i,j+1} - u_{i,j-1}).$$

3.1.2 Discrete Bellman Equation

Numerical Hamiltonian

In order to approximate the term $H(x, m, \nabla u)$ in (1), we consider a numerical Hamiltonian $g : \overline{\Omega} \times \mathbb{R}_+ \times \mathbb{R}^4 \to \mathbb{R}$, $(x, m, q_1, q_2, q_3, q_4) \mapsto g(x, m, q_1, q_2, q_3, q_4)$ satisfying the following conditions:

Monotonicity g is nonincreasing with respect to q_1 and q_3 and nondecreasing with respect to q_2 and q_4.

Consistency $g\,(x, m, q_1, q_1, q_2, q_2) = H(x, m, q)$, for all $x \in \overline{\Omega}$, for all $q = (q_1, q_2) \in \mathbb{R}^2$.

Differentiability g is of class $\mathscr{C}^1$.

Convexity $(q_1, q_2, q_3, q_4) \mapsto g\,(x, m, q_1, q_2, q_3, q_4)$ is convex.

We will approximate $H(\cdot, m, \nabla u)(x_{i,j})$ by $g(x_{i,j}, m_{i,j}, [D_h u]_{i,j})$. Standard examples of numerical Hamiltonians fulfilling these requirements are provided by Lax-Friedrichs or Godunov type schemes, see [4].

If the Hamiltonian H is of the form (4), the conditions above are all fulfilled by the discrete Hamiltonian given by

$$g(x, m, q) = \mathscr{H}(x) + \frac{G(q_1^-, q_2^+, q_3^-, q_4^+)}{(c_1 + c_2 m)^\alpha}, \tag{8}$$

where, for a real number r, $r^+ = \max(r, 0)$ and $r^- = \max(-r, 0)$ and where $G : (\mathbb{R}_+)^4 \to \mathbb{R}_+$ is given by

$$G(p) = |p|^\beta = (p_1^2 + p_2^2 + p_3^2 + p_4^2)^{\frac{\beta}{2}}.$$

Note that $g(x, m, q)$ is of class $\mathscr{C}^1$ with respect to $q \in (\mathbb{R}_+)^4$, because $\lambda \mapsto \lambda^{\frac{\beta}{2}}$ is $\mathscr{C}^1$ on $\mathbb{R}_+\backslash\{0\}$, $q \mapsto (q_1^-)^2,\ (q_2^+)^2,\ (q_3^-)^2,\ (q_4^+)^2$ are $\mathscr{C}^1$ functions on $\mathbb{R}^4$, and because the differential of $g(x, m, q)$ with respect to q tend to 0 as $q \to (0, 0, 0, 0)$ in $(\mathbb{R}_+)^4$.

Discrete Bellman Equation

The discrete version of the Bellman equation is obtained by applying the following semi-implicit Euler scheme:

$$\frac{u_{i,j}^{n+1} - u_{i,j}^n}{\Delta t} + \nu(\Delta_h u^n)_{i,j} - g(x_{i,j}, m_{i,j}^{n+1}, [D_h u^n]_{i,j}) = -F(m_{i,j}^{n+1}), \tag{9}$$

for all points in Ω_h and all n, $0 \le n < N_T$. Given $(m^n)_{n=1,\dots,N_T}$, (9) and the terminal condition $u_{i,j}^{N_T} = u_T(x_{i,j})$ for all (i, j) completely characterizes $(u^n)_{0\le n\le N_T}$.

3.1.3 Discrete Kolmogorov Equation

Discrete transport operator

In order to approximate Eq. (2), we multiply the nonlinear term in (2) by a test function w vanishing on Γ_D and integrate over Ω, as one would do when writing the weak formulation of (2): this yields the integral $\int_\Omega \operatorname{div}\left(m \frac{\partial H}{\partial p}(\cdot, m, \nabla u)\right)(x)\, w(x)\, dx$, in which m appears twice; we double the variable m in order to define the discrete transport operator, i.e. we consider

$$I = \int_\Omega \operatorname{div}\left(m \frac{\partial H}{\partial p}(\cdot, \tilde{m}, \nabla u)\right)(x)\; w(x)\, dx.$$

By integration by parts, we obtain

$$\begin{aligned} I = &- \int_\Omega m(x) \frac{\partial H}{\partial p}(x, \tilde{m}, \nabla u(x)) \cdot \nabla w(x)\, dx \\ &+ \int_{\Gamma_N} m(x) \frac{\partial H}{\partial p}(x, \tilde{m}, \nabla u(x)) \cdot n(x) w(x)\, ds. \end{aligned} \tag{10}$$

From the Neumann conditions, the last term in (10) vanishes if the Hamiltonian is of the form (4). Indeed

$$\frac{\partial H}{\partial p}(x, \tilde{m}, \nabla u(x)) \cdot n(x) = \frac{\beta}{(c_0 + c_1 \tilde{m})^\alpha} |\nabla u(x)|^{\beta - 2} \nabla u(x) \cdot n(x) = 0,$$

even if $\nabla u = 0$ since $\beta > 1$. Hence,

$$I = - \int_\Omega m(x) \frac{\partial H}{\partial p}(x, \tilde{m}, \nabla u(x)) \cdot \nabla w(x),$$

which will be approximated by

$$-h^2 \sum_{i,j} m_{i,j} \nabla_q g(x_{i,j}, \tilde{m}_{i,j}, [D_h u]_{i,j}) \cdot [D_h w]_{i,j}.$$

In consequence, we define the transport operator $\mathscr{T}$ by

$$h^2 \sum_{i,j} \mathscr{T}_{i,j}(u, m, \tilde{m}) w_{i,j} = -h^2 \sum_{i,j} m_{i,j} \nabla_q g(x_{i,j}, \tilde{m}_{i,j}, [D_h u]_{i,j}) \cdot [D_h w]_{i,j}.$$

This identity completely characterizes $\mathscr{T}_{i,j}(u, m, \tilde{m})$: for example,

- if $x_{i,j}$ is a strongly interior point, i.e. if the neighbors of $x_{i,j}$ all belong to Ω_h, then

$$\mathscr{T}_{i,j}(u,m,\tilde{m}) = \\ \frac{1}{h}\left(\begin{array}{l} \begin{pmatrix} m_{i,j}\frac{\partial g}{\partial q_1}(x_{i,j},\tilde{m}_{i,j},[D_h u]_{i,j}) - m_{i-1,j}\frac{\partial g}{\partial q_1}(x_{i-1,j},\tilde{m}_{i-1,j},[D_h u]_{i-1,j}) \\ +m_{i+1,j}\frac{\partial g}{\partial q_2}(x_{i+1,j},\tilde{m}_{i+1,j},[D_h u]_{i+1,j}) - m_{i,j}\frac{\partial g}{\partial q_2}(x_{i,j},\tilde{m}_{i,j},[D_h u]_{i,j}) \end{pmatrix} \\ + \\ \begin{pmatrix} m_{i,j}\frac{\partial g}{\partial q_3}(x_{i,j},\tilde{m}_{i,j},[D_h u]_{i,j}) - m_{i,j-1}\frac{\partial g}{\partial q_3}(x_{i,j-1},\tilde{m}_{i,j-1},[D_h u]_{i,j-1}) \\ +m_{i,j+1}\frac{\partial g}{\partial q_4}(x_{i,j+1},\tilde{m}_{i,j+1},[D_h u]_{i,j+1}) - m_{i,j}\frac{\partial g}{\partial q_4}(x_{i,j},\tilde{m}_{i,j},[D_h u]_{i,j}) \end{pmatrix} \end{array}\right).$$

- if $x_{i+1,j} \in \Gamma_N$ and $x_{i,j\pm 1} \in \Omega_h$, then

$$\mathscr{T}_{i,j}(u,m,\tilde{m}) = \\ \frac{1}{h}\left(\begin{array}{l} \begin{pmatrix} -m_{i-1,j}\frac{\partial g}{\partial q_1}(x_{i-1,j},\tilde{m}_{i-1,j},[D_h u]_{i-1,j}) \\ \quad -m_{i,j}\frac{\partial g}{\partial q_2}(x_{i,j},\tilde{m}_{i,j},[D_h u]_{i,j}) \end{pmatrix} \\ + \\ \begin{pmatrix} m_{i,j}\frac{\partial g}{\partial q_3}(x_{i,j},\tilde{m}_{i,j},[D_h u]_{i,j}) - m_{i,j-1}\frac{\partial g}{\partial q_3}(x_{i,j-1},\tilde{m}_{i,j-1},[D_h u]_{i,j-1}) \\ +m_{i,j+1}\frac{\partial g}{\partial q_4}(x_{i,j+1},\tilde{m}_{i,j+1},[D_h u]_{i,j+1}) - m_{i,j}\frac{\partial g}{\partial q_4}(x_{i,j},\tilde{m}_{i,j},[D_h u]_{i,j}) \end{pmatrix} \end{array}\right).$$

The Dirichlet conditions are imposed in a straightforward manner.

Discrete Kolmogorov Equation

With the conventions discussed above for the Neumann and Dirichlet conditions on m, we obtain the discrete Kolmogorov equation:

$$\frac{m_{i,j}^{n+1} - m_{i,j}^{n}}{\Delta t} - \nu(\Delta_h m^{n+1})_{i,j} - \mathscr{T}_{i,j}(u^n, m^{n+1}, m^{n+1}) = 0. \tag{11}$$

3.1.4 Summary

The fully discrete scheme for system (1), (2), (3) is therefore the following: for all i, j such that $x_{i,j} \in \Omega_h$ and $0 \le n < N_T$

$$\begin{cases} \dfrac{u_{i,j}^{n+1} - u_{i,j}^{n}}{\Delta t} + \nu(\Delta_h u^n)_{i,j} - g(x_{i,j}, m_{i,j}^{n+1}, [D_h u^n]_{i,j}) = -F(m_{i,j}^{n+1}), \\ \dfrac{m_{i,j}^{n+1} - m_{i,j}^{n}}{\Delta t} - \nu(\Delta_h m^{n+1})_{i,j} - \mathscr{T}_{i,j}(u^n, m^{n+1}, m^{n+1}) = 0, \end{cases} \tag{12}$$

with the the virtual points convention accounting for Neumann conditions and the initial and terminal conditions

$$m^0_{i,j} = \frac{1}{h^2}\int_{|x-x_{i,j}|_\infty \le h/2} m_0(x)dx, \quad u^{N_T}_{i,j} = u_T(x_{i,j}). \tag{13}$$

Mass Conservation or Decay

It can be proved that if $\Gamma_D = \emptyset$, then the scheme (11) is conservative, i.e. it preserves total mass over time.

If $\Gamma_D \neq \emptyset$, then it can be proved that

$$\sum_{x_{i,j}\in\Omega_h} \left(-\nu\Delta_h m^{n+1}_{i,j} - \mathscr{T}_{i,j}(u^n, m^{n+1}, m^{n+1})\right) \ge 0,$$

which shows that the total mass is a nonincreasing function of time. Indeed,

$$-h\sum_{x_{i,j}\in\Omega_h} \mathscr{T}_{i,j}(u^n, m^{n+1}, m^{n+1}) = \left\{\begin{array}{l} -\displaystyle\sum_{x_{i+1,j}\in\Gamma_D} m^{n+1}_{i,j}\frac{\partial g}{\partial q_1}(x_{i,j}, m^{n+1}_{i,j}, [D_h u^n]_{i,j}) \\ +\displaystyle\sum_{x_{i-1,j}\in\Gamma_D} m^{n+1}_{i,j}\frac{\partial g}{\partial q_2}(x_{i,j}, m^{n+1}_{i,j}, [D_h u^n]_{i,j}) \\ -\displaystyle\sum_{x_{i,j+1}\in\Gamma_D} m^{n+1}_{i,j}\frac{\partial g}{\partial q_3}(x_{i,j}, m^{n+1}_{i,j}, [D_h u^n]_{i,j}) \\ +\displaystyle\sum_{x_{i,j-1}\in\Gamma_D} m^{n+1}_{i,j}\frac{\partial g}{\partial q_4}(x_{i,j}, m^{n+1}_{i,j}, [D_h u^n]_{i,j}) \end{array}\right.$$

is nonnegative.

From these observations on the conservation/decay of the total mass, existence for the discrete problem (12)–(13) is obtained by applying Brouwer fixed point theorem, see, e.g., [4].

3.2 Fundamental Identity Leading to Uniqueness

3.2.1 Basic Lemma

Hereafter, when they are meaningful, the notations $g_q(x,m,q)$, $g_{mq}(x,m,q)$, $g_{qq}(x,m,q)$, respectively, stand for the gradient of g with respect to q, its partial derivative with respect to m, and the Hessian matrix of $q \mapsto g(x,m,q)$.

Lemma 1 *Let us introduce the diagonal matrix* $\mathscr{D} = \mathrm{diag}(-1,1,-1,1) \in \mathbb{R}^{4\times 4}$. *For all* $r \in \mathbb{R}^4$, *we have*

$$g_q(x,m,q)\cdot r = \frac{\beta|p|^{\beta-2}}{(c_1+c_2 m)^\alpha} p \cdot \mathscr{D}r, \tag{14}$$

where $p \in (\mathbb{R}_+)^4$ is given by

$$p = (q_1^-, q_2^+, q_3^-, q_4^+). \tag{15}$$

For all $q \in \mathbb{R}^4$, $r \in \mathbb{R}^4$, and p given by (15)

$$g_{m,q}(x, m, q) \cdot r = -\frac{c_2 \alpha \beta |p|^{\beta-2}}{(c_1 + c_2 m)^{\alpha+1}} p \cdot \mathscr{D} r, \tag{16}$$

and

$$r \cdot g_{qq}(x, m, q) r = \frac{1}{(c_1 + c_2 m)^\alpha} \Big(\beta(\beta - 2)|p|^{\beta-4} (p \cdot \mathscr{D} r)^2 + \beta |p|^{\beta-2} |\mathscr{D} r|^2 \Big). \tag{17}$$

Proof The identities (14), (16) and (17) follow from direct calculations, see [3]. For example, (17) comes from the observation that for all $p \in (\mathbb{R}_+)^4$, $G_p(p) = \beta|p|^{\beta-2} p$ and

$$G_{pp}(p) = \beta|p|^{\beta-2} I_4 + \beta(\beta - 2)|p|^{\beta-4} p \otimes p.$$

3.2.2 Nonlinear Functional $\mathscr{E}(m, u, \tilde{m}, \tilde{u})$

Let us define the nonlinear functional $\mathscr{E}$ acting on grid functions by

$$\mathscr{E}(m, u, \tilde{m}, \tilde{u}) = \sum_{n=1}^{N_T} \sum_{x_{i,j} \in \Omega_h} E \left(x_{i,j}, m_{i,j}^n, [Du^{n-1}]_{i,j}, \tilde{m}_{i,j}^n, [D\tilde{u}^{n-1}]_{i,j} \right)$$

where $m_{i,j}^n$ and $\tilde{m}_{i,j}^n$ are nonnegative and

$$\begin{aligned} E(x, m, q, \tilde{m}, \tilde{q}) = {} & (m - \tilde{m}) \left(g(x, \tilde{m}, \tilde{q}) - g(x, m, q) \right) \\ & - \big(m g_q(x, m, q) - \tilde{m} g_q(x, \tilde{m}, \tilde{q}) \big) \cdot (\tilde{q} - q). \end{aligned}$$

3.2.3 Fundamental Identity

In this paragraph, we discuss a key identity which leads to the stability of the finite difference scheme under additional assumptions. Consider a perturbed system:

$$\begin{cases} \dfrac{\tilde{u}_{i,j}^{n+1} - \tilde{u}_{i,j}^n}{\Delta t} + \nu(\Delta_h \tilde{u}^n)_{i,j} - g(x_{i,j}, \tilde{m}_{i,j}^{n+1}, [D_h \tilde{u}^n]_{i,j}) = -F(\tilde{m}_{i,j}^{n+1}) + a_{i,j}^{n+1}, \\ \dfrac{\tilde{m}_{i,j}^{n+1} - \tilde{m}_{i,j}^n}{\Delta t} - \nu(\Delta_h \tilde{m}^n)_{i,j} - \mathscr{T}_{i,j}(\tilde{u}^{n+1}, \tilde{m}^n, \tilde{m}^n) = b_{i,j}^n, \end{cases} \tag{18}$$

with the same boundary conditions as above.

Multiplying the first equations in (18) and (12) by $m^{n+1}_{i,j} - \tilde{m}^{n+1}_{i,j}$ and subtracting, then summing the results for all $n = 0, \dots, N_T - 1$ and all (i, j), we obtain

$$\begin{aligned}\sum_{n=1}^{N_T} \frac{1}{\Delta t}((u^n - \tilde{u}^n) - (u^{n-1} - \tilde{u}^{n-1}), (m^n - \tilde{m}^n))_2 + \nu(\Delta_h(u^{n-1} - \tilde{u}^{n-1}), m^n - \tilde{m}^n)_2 \\ - \sum_{n=1}^{N_T}\sum_{i,j}(g(x_{i,j}, m^n_{i,j}, [D_h u^{n-1}]_{i,j}) - g(x_{i,j}, \tilde{m}^n_{i,j}, [D_h \tilde{u}^{n-1}]_{i,j}))(m^n_{i,j} - \tilde{m}^n_{i,j}) \\ = -\sum_{n=1}^{N_T}\left(F(m^n) - F(\tilde{m}^n), m^n - \tilde{m}^n\right)_2 - \sum_{n=1}^{N_T}(a^n, m^n - \tilde{m}^n)_2, \quad (19)\end{aligned}$$

where $(X, Y)_2 = \sum_{i,j} X_{i,j} Y_{i,j}$. Similarly, subtracting the second equation in (18) from the second equation in (12), multiplying the result by $u^n_{i,j} - \tilde{u}^n_{i,j}$ and summing for all $n = 0, \dots, N_T - 1$ and all (i, j) leads to

$$\begin{aligned}\sum_{n=0}^{N_T-1} \frac{1}{\Delta}t((m^{n+1} - m^n) - (\tilde{m}^{n+1} - \tilde{m}^n), (u^n - \tilde{u}^n))_2 - \nu((m^{n+1} - \tilde{m}^{n+1}), \Delta_h(u^n - \tilde{u}^n))_2 \\ + \sum_{n=0}^{N_T-1}\sum_{i,j} m^{n+1}_{i,j}[D_h(u^n - \tilde{u}^n)]_{i,j} \cdot g_q\left(x_{i,j}, m^{n+1}_{i,j}, [D_h u^n]_{i,j}\right) \\ - \sum_{n=0}^{N_T-1}\sum_{i,j} \tilde{m}^{n+1}_{i,j}[D_h(u^n - \tilde{u}^n)]_{i,j} \cdot g_g\left(x_{i,j}, \tilde{m}^{n+1}_{i,j}[D_h \tilde{u}^n]_{i,j}\right) = -\sum_{n=0}^{N_T-1}(b^n, u^n - \tilde{u}^n)_2. \quad (20)\end{aligned}$$

Adding (19) and (20) leads to the fundamental identity

$$\begin{aligned}\frac{1}{\Delta}t(m^{N_T} - \tilde{m}^{N_T}, u^{N_T} - \tilde{u}^{N_T})_2 - \frac{1}{\Delta}t(m^0 - \tilde{m}^0, u^0 - \tilde{u}^0)_2 \\ + \mathscr{E}(m, u, \tilde{m}, \tilde{u}) + \sum_{n=1}^{N_T}(F(m^n) - F(\tilde{m}^n), m^n - \tilde{m}^n)_2 \\ = -\sum_{n=1}^{N_T}(a^n, m^n - \tilde{m}^n)_2 - \sum_{n=0}^{N_T-1}(b^n, u^n - \tilde{u}^n)_2. \\ (21)\end{aligned}$$

3.2.4 Uniqueness for the Discrete Problem

Let $(u^n_{i,j}, m^n_{i,j})$ and $(\tilde{u}^n_{i,j}, \tilde{m}^n_{i,j})$ are two solutions of (12)–(13) with the same boundary conditions on Γ_D and Γ_N. The fundamental identity (21) boils down to

$$\mathscr{E}(m,u,\tilde{m},\tilde{u})+\sum_{n=1}^{N_T}(F(m^n)-F(\tilde{m}^n),m^n-\tilde{m}^n)_2=0. \tag{22}$$

Our goal is to show that this implies that $u^n=\tilde{u}^n$ and that $m^n=\tilde{m}^n$.

Proposition 1 *We have that*

$$E(m,q,\tilde{m},\tilde{q})\ge 0,\quad \forall q\in\mathbb{R}^4,\ \forall\tilde{q}\in\mathbb{R}^4,\ \forall m\ge 0,\ \forall\tilde{m}\ge 0, \tag{23}$$

if and only if

$$r\cdot g_{qq}(x,m,q)r+zg_{qm}(x,m,q)\cdot r-z^2g_m(x,m,q)\ge 0,\quad \forall z\in\mathbb{R},\ r\in\mathbb{R}^4, \tag{24}$$

for all $x\in\Omega_h$, $q\in\mathbb{R}^4$ and $m\in\mathbb{R}^+$ such that $(q_1^-,q_2^+,q_3^-,q_4^+)\neq 0$.

Proof (23) implies (24). Take $q\in\mathbb{R}^4$ such that $p=(q_1^-,q_2^+,q_3^-,q_4^+)\neq 0$. Take also $\tilde{q}=q+\varepsilon r$ and $\tilde{m}=m+\varepsilon z$ in (23): dividing by ε^2 and passing to the limit, we get (24).

(24) implies (23). Consider $\tilde{q}=q+r$, $q_t=q+tr$, $\tilde{m}=m+z$, $m_t=m+tz$. Assume first that $p=(q_1^-,q_2^+,q_3^-,q_4^+)\neq 0$ and $\tilde{p}=(\tilde{q}_1^-,\tilde{q}_2^+,\tilde{q}_3^-,\tilde{q}_4^+)\neq 0$.

Consider an open subinterval J in [0, 1] such that $p_t=(q_{t,1}^-,q_{t,2}^+,q_{t,3}^-,q_{t,4}^+)\neq 0$ for all $t\in J$. The function $h:t\mapsto\frac{E(m,q,m_t,q_t)}{t}$ has a derivative on J:

$$\frac{dh}{dt}(t)=(m+tz)r\cdot g_{qq}(x,m_t,q_t)r+(m+tz)zg_{qm}(x,m_t,q_t)\cdot r \\ -z^2g_m(x,m_t,q_t)\ge 0.$$

On the other hand, h is constant in the intervals in which $p_t=(q_{t,1}^-,q_{t,2}^+,q_{t,3}^-,q_{t,4}^+)=0$.

The observation above imply that h is nondecreasing on [0, 1]. Moreover, $\lim_{t\to 0+}h(t)=0$. Therefore $E(m,q,m_t,q_t)\ge 0$ for all $t\in[0,1]$, which implies in particular that $E(m,q,\tilde{m},\tilde{q})\ge 0$.

Proposition 2 *Take $m>0$ and $q\in\mathbb{R}^4$ such that $p=(q_1^-,q_2^+,q_3^-,q_4^+)\neq 0$. A sufficient condition for (24) is that*

$$\alpha\le\frac{4(\beta-1)}{\beta}. \tag{25}$$

Proof Consider $m>0$, and $q\in\mathbb{R}^4$ such that $p=(q_1^-,q_2^+,q_3^-,q_4^+)\neq 0$, (24) can be written as

$$0\le\frac{m}{(c_1+c_2m)^\alpha}\Big(\beta(\beta-2)|p|^{\beta-4}(p\cdot\mathscr{D}r)^2+\beta|p|^{\beta-2}|\mathscr{D}r|^2\Big) \\ -z\frac{c_2\alpha\beta m|p|^{\beta-2}}{(c_1+c_2m)^{\alpha+1}}p\cdot\mathscr{D}r+\frac{c_2\alpha}{(c_1+c_2m)^{\alpha+1}}|p|^\beta z^2.$$

Taking $\tilde{r} = \mathscr{D}r$, the inequality above is equivalent to

$$m\Big(\beta(\beta-2)|p|^{\beta-4}(p\cdot\tilde{r})^2 + \beta|p|^{\beta-2}|\tilde{r}|^2\Big) - z\frac{c_2\alpha\beta m|p|^{\beta-2}}{c_1+c_2m}p\cdot\tilde{r} + z^2\frac{c_2\alpha}{c_1+c_2m}|p|^\beta \geq 0.$$

A sufficient condition is that

$$\left(\frac{c_2\alpha\beta m|p|^{\beta-2}}{c_1+c_2m}p\cdot\tilde{r}\right)^2 - 4\frac{c_2\alpha m}{c_1+c_2m}|p|^\beta\Big(\beta(\beta-2)|p|^{\beta-4}(p\cdot\tilde{r})^2 + \beta|p|^{\beta-2}|\tilde{r}|^2\Big) \leq 0,$$

which is equivalent to

$$(p\cdot\tilde{r})^2\left(\frac{c_2\alpha\beta m}{c_1+c_2m} - 4(\beta-2)\right) - 4|\tilde{r}|^2|p|^2 \leq 0.$$

The latter is a consequence of (25).

Corollary 1 *If $\nu > 0$, F is an increasing function and (25) holds, then the discrete problem has at most a solution.*

Proof From (22) and (23), we infer that $m^n_{i,j} = \tilde{m}^n_{i,j}$ for all i, j, n. Then, uniqueness for the discrete Bellman equation leads to $u^n_{i,j} = \tilde{u}^n_{i,j}$ for all i, j, n.

4 Strategy for Solving (12)–(13)

System (12)–(13) can be seen as a backward discrete Bellman equation for u with a Cauchy condition at $t = T$ coupled with a forward discrete Kolmogorov equation for m with a Cauchy condition at $t = 0$. This structure prohibits the use of a straightforward time-marching solution procedure.

In this paragraph, we assume that H and g are respectively given by (4) and (8) with $\beta \geq 2$. We also assume that F is a $\mathscr{C}^1$ and strictly increasing function. We introduce two auxiliary unknowns λ and μ in (1)–(6):

$$\lambda(t,x) = F(m(t,x)) \quad \text{and} \quad \mu(t,x) = (c_1 + c_2 m(t,x))^{-\alpha}.$$

We consider the nonlinear map $\Xi : (\lambda, \mu) \mapsto (\tilde{\lambda}, \tilde{\mu})$, where

$$\tilde{\lambda}(t,x) = F(m(t,x)) \quad \text{and} \quad \tilde{\mu}(t,x) = (c_1 + c_2 m(t,x))^{-\alpha}$$

and (u, m) is the solution of the system of nonlinear equations

$$\frac{\partial u}{\partial t}(t,x) + \nu\Delta u(t,x) - \mu(t,x)|\nabla u(t,x)|^\beta = -\lambda(t,x), \tag{26}$$

$$\frac{\partial m}{\partial t}(t,x) - \nu\Delta m(t,x) - \operatorname{div}\left(\beta m(t,\cdot)\mu(t,\cdot)|\nabla u(t,\cdot)|^{\beta-2}\nabla u(t,\cdot))\right)(x) = 0, \tag{27}$$

in $(0,T)\times\Omega$, supplemented with (3), (5) and (6). Given λ and μ, the system (26)–(27) with (3), (5) and (6) can be solved in a decoupled manner as follows:

1. The value function u is first obtained by solving the Bellman equation (26), with the terminal condition $u(T,\cdot) = u_T(\cdot)$ and the boundary conditions coming from (5) and (6).
2. Once u is available, then m is obtained by solving the Kolmogorov equation (27) with the initial condition $m(0,\cdot) = m_0(\cdot)$ and the boundary conditions coming from (5) and (6).

The equilibrium is equivalent to the fixed point problem

$$(\lambda,\mu) = \Xi(\lambda,\mu).$$

Here we carry out the same program at the discrete level: we introduce the auxiliary unknowns $(\lambda^n_{i,j},\mu^n_{i,j})_{n,i,j}$ which are bound to coincide with $(F(m^n_{i,j}),(c_1 + c_2 m^n_{i,j})^{-\alpha})_{n,i,j}$ for the solution $(u^n_{i,j}, m^n_{n,i,j})_{n,i,j}$ of (12)–(13). It will be useful to define

$$\tilde{g}(q) = ((q_1^-)^2 + (q_2^+)^2 + (q_3^-)^2 + (q_4^+)^2)^{\frac{\beta}{2}},$$

and

$$\tilde{\mathscr{T}}_{i,j}(u,m,\mu) = \frac{1}{h}\left(\begin{array}{c} \left(\begin{array}{l} m_{i,j}\mu_{i,j}\frac{\partial\tilde{g}}{\partial q_1}([D_h u]_{i,j}) - m_{i-1,j}\mu_{i-1,j}\frac{\partial\tilde{g}}{\partial q_1}([D_h u]_{i-1,j}) \\ +m_{i+1,j}\mu_{i+1,j}\frac{\partial\tilde{g}}{\partial q_2}([D_h u]_{i+1,j}) - m_{i,j}\mu_{i,j}\frac{\partial\tilde{g}}{\partial q_2}([D_h u]_{i,j}) \end{array}\right) \\ + \\ \left(\begin{array}{l} m_{i,j}\mu_{i,j}\frac{\partial\tilde{g}}{\partial q_3}([D_h u]_{i,j}) - m_{i,j-1}\mu_{i,j-1}\frac{\partial\tilde{g}}{\partial q_3}([D_h u]_{i,j-1}) \\ +m_{i,j+1}\mu_{i,j+1}\frac{\partial\tilde{g}}{\partial q_4}([D_h u]_{i,j+1}) - m_{i,j}\mu_{i,j}\frac{\partial\tilde{g}}{\partial q_4}([D_h u]_{i,j}) \end{array}\right) \end{array}\right).$$

We then introduce the discrete version Ξ_h of Ξ, namely $\Xi_h : (\lambda^n_{i,j},\mu^n_{i,j})_{n,i,j} \mapsto (\tilde{\lambda}^n_{i,j},\tilde{\mu}^n_{i,j})_{n,i,j}$, where

$$\tilde{\lambda}^n_{i,j} = F(m^n_{i,j}) \quad \text{and} \quad \tilde{\mu}^n_{i,j} = (c_1 + c_2 m^n_{i,j})^{-\alpha}$$

and $(u^n_{i,j}, m^n_{i,j})$ is the solution of the system of nonlinear equations

$$\frac{u^{n+1}_{i,j} - u^n_{i,j}}{\Delta t} + \nu(\Delta_h u^n)_{i,j} - \mu^{n+1}_{i,j}\tilde{g}([D_h u^n]_{i,j}) = -\lambda^{n+1}_{i,j} - \mathscr{H}(x_{i,j}), \tag{28}$$

$$\frac{m^{n+1}_{i,j} - m^n_{i,j}}{\Delta t} - \nu(\Delta_h m^{n+1})_{i,j} - \tilde{\mathscr{T}}_{i,j}(u^n, m^{n+1}, \mu^{n+1}) = 0, \tag{29}$$

with (13). The discrete equilibrium is equivalent to the fixed point problem

$$(\lambda^n_{i,j}, \mu^n_{i,j})_{n,i,j} = \Xi_h(\lambda^n_{i,j}, \mu^n_{i,j})_{n,i,j}. \tag{30}$$

Given $(\lambda^n_{i,j}, \mu^n_{i,j})_{n,i,j}$, the system (28), (29) and (13) can be solved in a decoupled manner as follows:

1. The value function u is first obtained by solving the discrete Bellman equation (28) with the terminal condition coming from (13), by marching backward in time. At each time step, Newton iterations are used for solving the system of nonlinear equation.
2. Once $(u^n_{i,j})_{n,i,j}$ is available, then $(m^n_{i,j})_{n,i,j}$ is obtained by solving the discrete Kolmogorov equation (29) with the initial condition coming from (13), by marching forward in time. At each time step, one has to solve a system of linear equations. This is done by using the library UMFPACK [20] which contains an Unsymmetric MultiFrontal method for solving linear systems.

The main advantage of working with the auxiliary unknowns λ and μ is that it allows for using the power of Newton iterations for (30) (fast convergence if the initial guess is not too far from the solution) and at the same time for preserving the positivity and the total mass of the discrete function $(m^n_{i,j})_{n,i,j}$. Newton iterations applied directly to (12)–(13) with the unknowns $(u^n_{i,j}, m^n_{i,j})_{n,i,j}$, have been used in previous articles, see, e.g., [6], but the positivity of $(m^n_{i,j})_{n,i,j}$ was not guaranteed.

Newton iterations require differentiating the map Ξ_h. For that, the main step consists of computing the differential of $(u^n, m^n)_{n,i,j}$ with respect to $(\lambda^n_{i,j})_{n,i,j}$ and $(\mu^n_{i,j})_{n,i,j}$. This requires differentiating the discrete Bellman and Kolmogorov equation in (28)–(29): let us give some details:

Let N be the number of grid points in Ω_h. Call $\mathscr{U}$ and $\mathscr{M}$ the vectors of $\mathbb{R}^{N_T N}$ such that $(\mathscr{U}_{nN}, \dots, \mathscr{U}_{(n+1)N-1})$ coincide with the unknowns $(u^{n-1}_{i,j})_{i,j}$, and $(\mathscr{M}_{nN}, \dots, \mathscr{M}_{(n+1)N-1})$ coincide with the unknowns $(m^n_{i,j})_{i,j}$ ordered lexicographically, (recall that $u^{N_T}_{i,j}$ and $m^0_{i,j}$ are given). With the slight abuse of notation consisting of writing λ and μ for the vectors containing $(\lambda^n_{i,j})_{n,i,j}$ and $(\mu^n_{i,j})_{n,i,j}$, the system (28)–(29) can be written

$$\mathscr{F}_U(\mathscr{U}, \mu) = -\lambda, \quad \text{and} \quad \mathscr{F}_M(\mathscr{U}, \mathscr{M}, \mu) = 0,$$

with

- $\mathscr{F}_U(\mathscr{U}, \mu) = -\lambda \Leftrightarrow$ (28) for all n, $0 \le n < N_T$ and i, j such that $x_{i,j} \in \Omega_h$,
- $\mathscr{F}_M(\mathscr{U}, \mathscr{M}, \mu) = 0 \Leftrightarrow$ (29) for all n, $0 \le n < N_T$ and i, j such that $x_{i,j} \in \Omega_h$.

We also use the following notation:

$$\begin{aligned} A_{U,U}(\mathscr{U}, \mu) &= D_{\mathscr{U}}\, \mathscr{F}_{\mathscr{U}}\,(\mathscr{U}, \mu), & & \\ A_{M,U}(\mathscr{U}, \mathscr{M}, \mu) &= D_{\mathscr{U}}\, \mathscr{F}_{\mathscr{M}}\,(\mathscr{U}, \mathscr{M}, \mu), & A_{M,M}(\mathscr{U}, \mathscr{M}, \mu) &= D_{\mathscr{M}}\, \mathscr{F}_{\mathscr{M}}\,(\mathscr{U}, \mathscr{M}, \mu), \\ B_{U,\mu}(\mathscr{U}, \mu) &= D_{\mu} \mathscr{F}_{\mathscr{U}}\,(\mathscr{U}, \mu), & B_{M,\mu}(\mathscr{U}, \mathscr{M}, \mu) &= D_{\mu} \mathscr{F}_{\mathscr{M}}\,(\mathscr{U}, \mathscr{M}, \mu). \end{aligned}$$

The matrix $A_{UU}(\mathscr{U}, \mu)$ has the form

$$A_{UU} = \begin{pmatrix} D_1 & \frac{1}{\Delta} tI & 0 & \dots & 0 \\ 0 & D_2 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & \frac{1}{\Delta} tI \\ 0 & \dots & \dots & 0 & D_{N_T} \end{pmatrix}.$$

The blocks of $A_{UU}(\mathscr{U}, \mu)$ are sparse matrices. The block D_n corresponds to the discrete operator $(z_{i,j}) \mapsto \left(- \frac{1}{\Delta t} z_{i,j} + \nu(\Delta_h z)_{i,j} - \mu^n_{i,j} [D_h z]_{i,j} \cdot \tilde{g}_q([D_h u^{n-1}]_{i,j}) \right)$ coming from the linearization of the discrete Bellman equation. From the monotonicity of the scheme, $-D_n$ is a M-matrix, thus A_{UU} is invertible. The matrices $A_{MM}(\mathscr{U}, \mathscr{M}, \mu)$ and $A_{MU}(\mathscr{U}, \mathscr{M}, \mu)$ have the form

$$A_{MM} = A_{UU}^T, \qquad A_{MU} = \begin{pmatrix} E_1 & 0 & \dots & \dots & 0 \\ 0 & E_2 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & E_{N_T-1} & 0 \\ 0 & \dots\dots & & 0 & E_{N_T} \end{pmatrix}.$$

The block A_{MM} corresponds to a discrete linear transport equation. Note that

$$\mathscr{V}^T E_n \mathscr{W} = \sum_{i,j} m^n_{i,j} \mu^n_{i,j} [D_h v]_{i,j} \cdot \tilde{g}_{q,q}(x_{i,j}, [D_h u^{n-1}]_{i,j}) [D_h w]_{i,j}.$$

From the convexity of g, we see that the block E_n is symmetric and positive semi-definite if m^n and μ^n are nonnegative grid functions. From this, we can differentiate $\mathscr{U}$ and $\mathscr{M}$ with respect to λ and μ: we obtain, in particular, that

$$D_{\lambda,\mu} \mathscr{M} \cdot (\tilde{\lambda}, \tilde{\mu}) = A^{-1}_{MM} A_{MU} A^{-1}_{UU} \tilde{\lambda} + A^{-1}_{MM} \left(A_{MU} A^{-1}_{UU} B_{U,\mu} - B_{M\mu} \right) \tilde{\mu}.$$

The latter allows for computing the differential of Ξ_h. Newton iterations require solving (at least approximatively) systems of linear equations involving the differential of Ξ_h. For that, we use iterative methods, namely BiCGstab, see [21].

In Fig. 1, we plot the convergence of the Newton iterations for the example discussed in Sect. 5.1. The graph seems to indicate a quadratic convergence, i.e. that for a positive constant C, $C\|r^{(n+1)}\| \leq \left(C\|r^{(n)}\| \right)^2$, where $\|r^{(n)}\|$ is the quadratic norm of the residual after n steps of the Newton method. In Fig. 2, we show the convergence of the BiCGstab iterations for one of the linear systems arising in the Newton loop.

Remark 1 If there is no congestion, the same ideas can be used with only one additional unknown λ. Then the linear problems in the Newton method involves matrices of the type

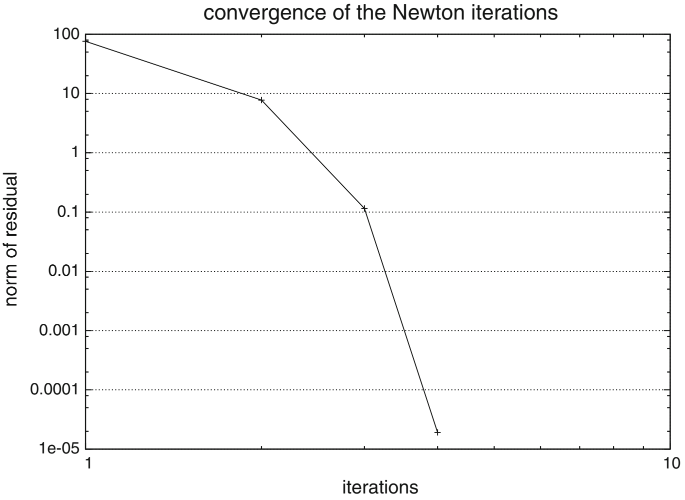
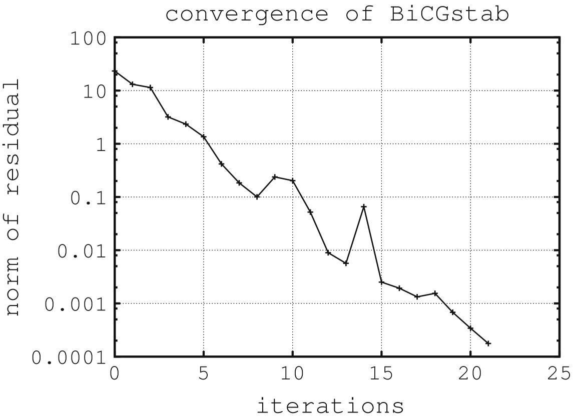


Fig. 1 The convergence of the Newton iterations for the example discussed in Sect. 5.1

Fig. 2 A typical graph of convergence of the BiCGstab iterations for a linear system arising in the Newton method for the example discussed in Sect. 5.1



$$D + A_{MM}^{-1} A_{MU} A_{UU}^{-1}, \tag{31}$$

where D is a diagonal matrix with diagonal entries $\frac{1}{f'(m_{i,j}^n)}$. The matrix in (31) is symmetric and positive definite if f is a strictly increasing function.

5 Examples

5.1 Exit from a Hall

5.1.1 Geometry and Parameters

The domain Ω is obtained by removing several rectangular closed sets from a square with a 50 m side, see Fig. 3. The rectangular subsets stand for obstacles into which the crowd cannot penetrate. These regions can be rows of seats for example, or forbidden zones where the pedestrians cannot seat or even go. Hence, the walls of these obstacles are part of Γ_N. There are two possible exits (which are closed at $t = 0$) which are located at the two ends of the wider part of the domain: the possible exits coincide with the straight line segments $\{0\} \times [0, 2.5]$ and $\{50\} \times [0, 2.5]$. There is no terminal cost, i.e. $u_T(x) = 0$ for all $x \in \Omega$ and no exit cost, i.e. $u_D(x) = 0$ for all $x \in \Gamma_D$.

The initial distribution of pedestrians m_0 is piecewise constant and takes two values 0 and 4: the pedestrians are gathered in some regions, in which the density is 4 people/m^2. Moreover, $\int_\Omega m_0(x)dx \sim 2900$, which means that there are ~ 2900 individuals in the room at the initial time (Fig. 4).

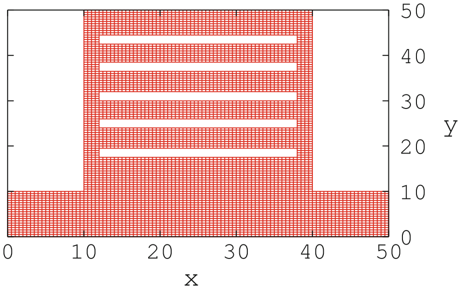
The parameters are

- $\nu = 1/3 \sim 0.33$
- $H(x, m, p) = |p|^2 \left(\frac{25}{1+6m}\right)^{\frac{3}{2}}$
- $F(m) = 5.\,10^{-4}m$
- $\mathscr{H}(x) = -10^{-3}$

which leads to the following HJB equation

$$\frac{\partial u}{\partial t} + \frac{1}{3}\Delta u - \left(\frac{25}{1 + 6m}\right)^{\frac{3}{2}} |\nabla u|^2 = -10^{-4}\ (5m + 10).$$

Fig. 3 The geometry of the problem and the grid



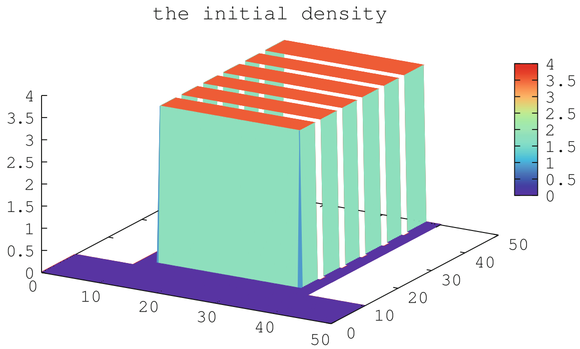


Fig. 4 The initial distribution of pedestrians

5.1.2 Results

The horizon is $T = 20$ min. The two doors stay open from $t = 0$ to $t = T$. The number of pedestrians in the room is plotted in Fig. 5: almost everybody has left the room at $t = T$.

Some snapshots of the distribution of pedestrians at different times are plotted in Fig. 6. In Fig. 7, three snapshots of the optimal feedback $-\gamma = -H_p(x, m, \nabla u)$ at different times are displayed. Since the initial distribution of pedestrians and the geometry of the model are symmetric with respect to the axis $x = 25$, the distribution of pedestrians stays symmetric. The maximal velocity is of the order of 2 m/s, the maximum being reached at the doors. The flow of the crowd has a complex structure in the zones where two streams meet.

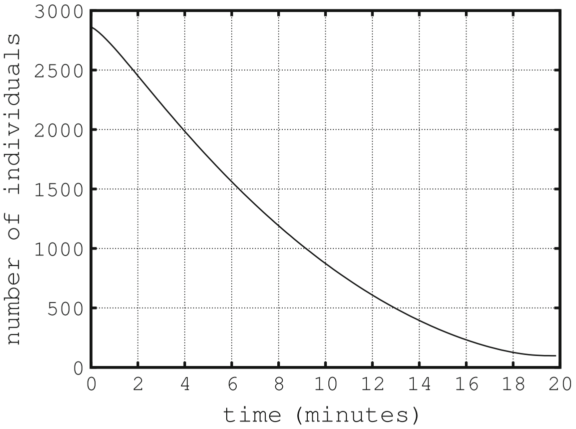


Fig. 5 The number of pedestrians in the room versus time

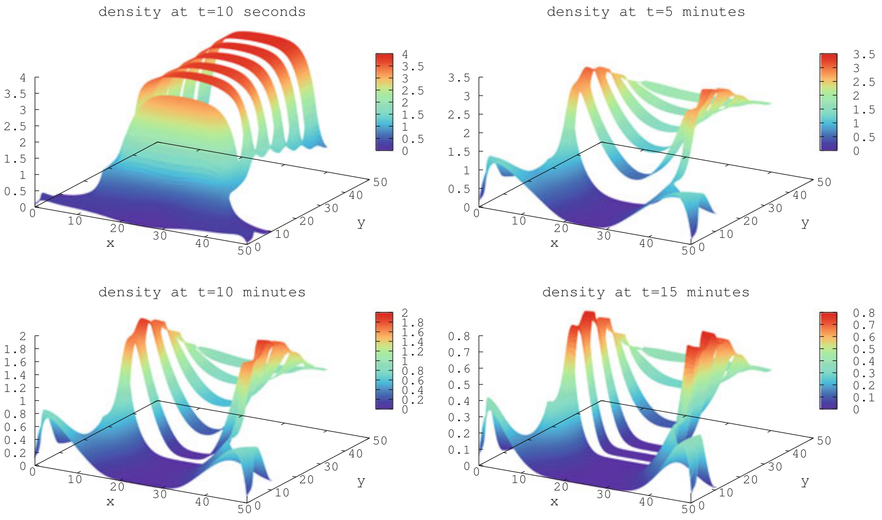


Fig. 6 The distribution of pedestrians at $t = 10\,\text{s}, 5,\ 10,\ 15\,\text{min}$. Note that the density scale varies w.r.t. t

5.2 *Exit from a Hall with Incomplete Information*

5.2.1 Model

The horizon is $T = 40$ min. Before $t = T/2$, the doors are closed. The agents know that one of the two doors will be opened at $t = T/2$ and will stay open until $t = T$, but they do not know which one. At $T/2$, the probability that a given door be opened is $1/2$.

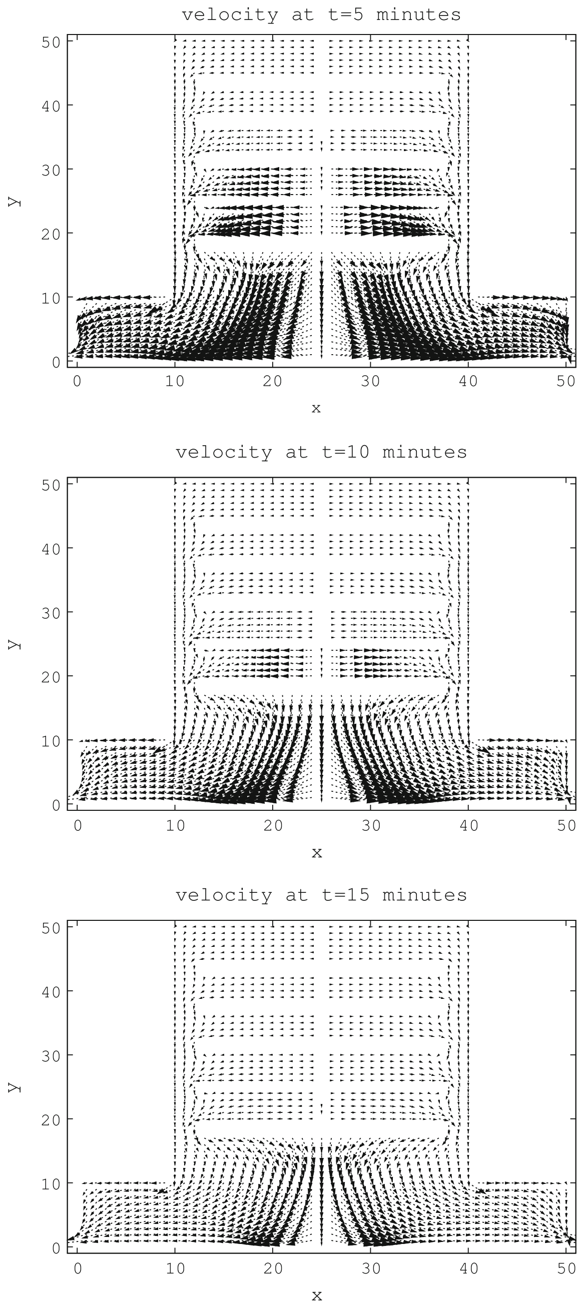
Hence the model involves three pairs of unknown functions

- (u^C, m^C) is defined on $(0, T/2) \times \Omega$ and corresponds to the situation when the room is closed.
- (u^L, m^L) and (u^R, m^R) are defined on $(T/2, T) \times \Omega$ and correspond respectively to the case when the left (resp. right) door is open.

The boundary value problem to be solved is

$$\frac{\partial u^C}{\partial t}(t,x) + \nu \Delta u^C(t,x) - H(x, m^C(t,x), \nabla u^C(t,x)) = -F(m^C(t,x)),$$
$$\frac{\partial m^C}{\partial t}(t,x) - \nu \Delta m^C(t,x) - \operatorname{div}\left(m^C(t,\cdot) \frac{\partial H}{\partial p}(\cdot, m^C(t,\cdot), \nabla u^C(t,\cdot)) \right)(x) = 0,$$

in $(0, T/2) \times \Omega$, with the boundary conditions


velocity at t=5 minutes

velocity at t=10 minutes

velocity at t=15 minutes

Fig. 7 The optimal feedback $-\gamma$ at $t = 5, 10, 15$ min

$$\frac{\partial u^C}{\partial n} = \frac{\partial m^C}{\partial n} = 0 \quad \text{on } \left(0, \frac{T}{2}\right) \times \partial\Omega$$

and for $j = L, R$,

$$\begin{aligned}
\frac{\partial u^j}{\partial t}(t,x) + \nu\Delta u^j(t,x) - H(x, m^j(t,x), \nabla u^j(t,x)) &= -F(m^j(t,x)),\\
\frac{\partial m^j}{\partial t}(t,x) - \nu\Delta m^j(t,x) - \operatorname{div}\left(m^j(t,\cdot)\frac{\partial H}{\partial p}(\cdot, m^j(t,\cdot), \nabla u^j(t,\cdot))\right)(x) &= 0,
\end{aligned}$$

in $(T/2, T) \times \Omega$, with the boundary conditions

$$\begin{aligned}
\frac{\partial u^j}{\partial n} = \frac{\partial m^j}{\partial n} &= 0 \qquad \text{on } \left(\frac{T}{2}, T\right) \times \Gamma_N^j,\\
u^j = m^j &= 0 \qquad \text{on } \left(\frac{T}{2}, T\right) \times \Gamma_D^j
\end{aligned}$$

where $\Gamma_D^L = \{0\} \times (0, 2.5)$, $\Gamma_D^R = \{50\} \times (0, 2.5)$, and $\Gamma_N^j = \partial\Omega \backslash \overline{\Gamma}_D^j$, $j = L, R$.

These equations have to be supplemented with the initial and terminal conditions

$$m^C(0,x) = m_0(x), \quad u^L(T,x) = u^R(T,x) = u_T(x) \quad \text{in } \Omega,$$

and the transmission conditions at $t = T/2$:

$$\begin{aligned}
m^L\left(\frac{T}{2}, x\right) = m^R\left(\frac{T}{2}, x\right) &= m^C\left(\frac{T}{2}, x\right) \qquad \text{in } \Omega,\\
u^C\left(\frac{T}{2}, x\right) &= \frac{u^L(\frac{T}{2}, x) + u^R(\frac{T}{2}, x)}{2} \qquad \text{in } \Omega.
\end{aligned}$$

5.2.2 Results

Since the geometry, the initial distribution of pedestrians and the final cost are symmetric with respect to the axis $x = 25$, and since the left and right door have the same probability to be opened at $t = T/2$, the distribution of pedestrians m^L and m^R must be symmetric to each other, and m^C must be symmetric with respect to the axis $x = 25$. We have not used this remark in our numerical simulations, although it would have been helpful in the present case.

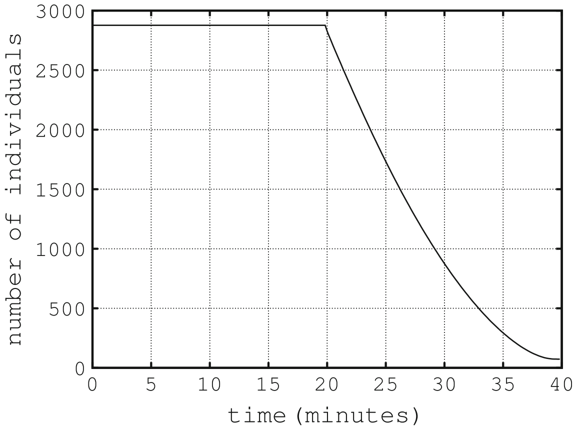


Fig. 8 The number of pedestrians in the room versus time

The number of pedestrians in the room is plotted in Fig. 8: we see that this number remains constant for $t < T/2$ and decays for $t > T/2$. Almost everybody has left the room at $t = T$.

We plot the fields corresponding to the case when the left door is opened at $t = T/2$, i.e. $m = m^C$ for $t \leq T/2$ and $m = M^L$ for $t > T/2$. Similarly, the optimal feedback $-\gamma$ (also named *velocity* below) is $-\gamma = -H_p(x, m^C, \nabla u^C)$ for $t \leq T/2$ and $-\gamma = -H_p(x, m^L, \nabla u^L)$ for $t > T/2$.

Snapshots of the distribution of pedestrians at different times are plotted on Fig. 9. In Fig. 10, three snapshots of the optimal feedback $-\gamma$ at different times are plotted. We see that for $t < T/2$, the crowd moves toward the doors. As expected, the distribution stays symmetric with respect to the axis $x = 25$ for $t \leq T/2$. For $t < T/2$, the maximal velocity is of the order of 0.5 m/s. The regions near the ends of the inner obstacles are interesting, because it is where two flows of pedestrians meet, the first one walking in a direction parallel to the outer walls, the second one walking in a direction parallel to the inner walls.

One sees from Fig. 9 that just before the door opening, the density increases in front of the two doors, even though only one door is going to be opened. Just after the left door opening, the symmetry is broken. The parameters have been chosen in such a way the maximal velocity is of the order of 3 m/s for $t > T/2$.

Acknowledgements The first author would like to affectionately dedicate this work to Yuri Kuznetsov and Olivier Pironneau for their seventieth birthdays. The first author was partially funded by the ANR projects ANR-12-MONU-0013 and ANR-12-BS01-0008-01. The two authors acknowledge the support of the Chaire “Finance et développement durable”.

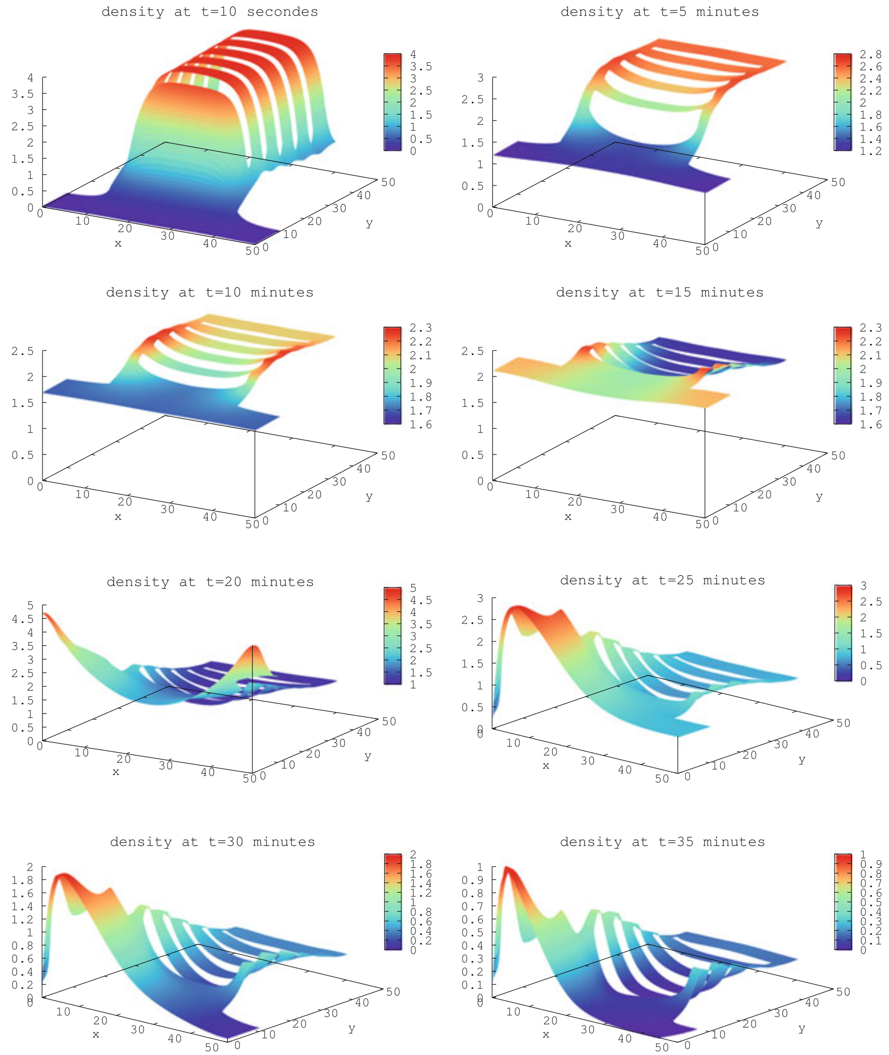


Fig. 9 The distribution of pedestrians at $t = 10\,\mathrm{s}, 5,\ 10,\ 15,\ 20,\ 25,\ 30,\ 35\,\mathrm{min}$. Note that the density scale varies w.r.t. t

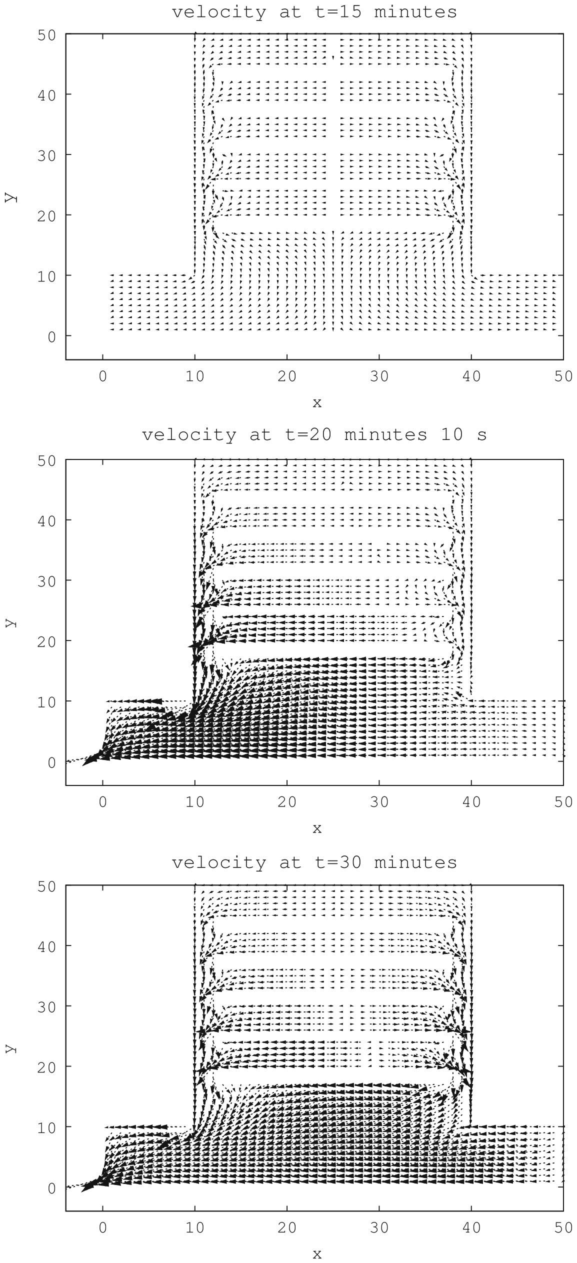


Fig. 10 The optimal feedback $-\gamma$ at $t = 15\,\mathrm{min}$, $20\,\mathrm{min}\ 10\,\mathrm{s.}$, $30\,\mathrm{min}$. We see that the velocity is larger when t is large, i.e. when the room is less crowded. Note the dissymmetry of the velocity between the inner obstacles just after the left door opening, i.e. for $t > 20\,\mathrm{min}$

References

1. Achdou Y (2013) Finite difference methods for mean field games. In: Hamilton-Jacobi Equations: approximations, numerical analysis and applications, volume 2074 of Lecture Notes in Mathematics. Springer, Heidelberg, pp 1–47
2. Achdou Y, Buera FJ, Lasry J-M, Lions P-L, Moll B (2014) Partial differential equation models in macroeconomics. Philos Trans R Soc Lond Ser A Math Phys Eng Sci 372(2028):20130397 (19 pp)
3. Achdou Y, Camilli F, Capuzzo-Dolcetta I (2013) Mean field games: convergence of a finite difference method. SIAM J Numer Anal 51(5):2585–2612
4. Achdou Y, Capuzzo-Dolcetta I (2010) Mean field games: numerical methods. SIAM J Numer Anal 48(3):1136–1162
5. Achdou Y, Han J, Lasry J-M, Lions P-L, Moll B (2015) Heterogeneous agent models in continuous time. Working papers, Princeton University
6. Achdou Y, Perez V (2012) Iterative strategies for solving linearized discrete mean field games systems. Netw Heterog Media 7(2):197–217
7. Achdou Y, Porretta A (2016) Convergence of a finite difference scheme to weak solutions of the system of partial differential equations arising in mean field games. SIAM J Numer Anal 54(1):161–186
8. Bardi M, Capuzzo-Dolcetta I (1997) Optimal control and viscosity solutions of Hamilton-Jacobi-Bellman equations. Birkhäuser, Boston, MA (With appendices by M. Falcone and P. Soravia)
9. Burger M, Di Francesco M, Markowich PA, Wolfram M-T (2014) Mean field games with nonlinear mobilities in pedestrian dynamics. Discrete Contin Dyn Syst Ser B 19(5):1311–1333
10. Djehiche B, Tcheukam A, Tembine H A mean-field game of evacuation in multi-level building. IEEE Trans Automat Control. https://doi.org/10.1109/TAC.2017.2679487 (to appear)
11. Faure S, Maury B (2015) Crowd motion from the granular standpoint. Math Models Methods Appl Sci 25(3):463–493
12. Fleming WH, Soner HM (2006) Controlled Markov processes and viscosity solutions, volume 25 of Stochastic modelling and applied probability, 2nd edn. Springer, New York
13. Hughes RL (2003) The flow of human crowds. In: Annual review of fluid mechanics, volume 35 of Annual Review of Fluid Mechanics. Annual Reviews, Palo Alto, CA, pp 169–182
14. Krusell P, Smith AA Jr (1998) Income and wealth heterogeneity in the macroeconomy. J Polit Econ 106(5):867–896
15. Lachapelle A, Wolfram M-T (2011) On a mean field game approach modeling congestion and aversion in pedestrian crowds. Transp Res Part B: Methodol 45(10):1572–1589
16. Lasry J-M, Lions P-L (2006) Jeux à champ moyen. I. Le cas stationnaire. C R Math Acad Sci Paris 349:619–625
17. Lasry J-M, Lions P-L (2006) Jeux à champ moyen. II. Horizon fini et contrôle optimal. C R Math Acad Sci Paris 343(10):679–684
18. Lasry J-M, Lions P-L (2007) Mean field games. Jpn J Math 2(1):229–260
19. Lions P-L (2007–2011) Cours du Collège de France. http://www.college-de-france.fr/site/pierre-louis-lions/_course.htm
20. UMFPACK. https://en.wikipedia.org/wiki/UMFPACK
21. van der Vorst HA (1992) Bi-CGSTAB: a fast and smoothly converging variant of Bi-CG for the solution of nonsymmetric linear systems. SIAM J Sci Statist Comput 13(2):631–644

Remarks About Spatially Structured SI Model Systems with Cross Diffusion

Verónica Anaya, Mostafa Bendahmane, Michel Langlais and Mauricio Sepúlveda

Abstract One of the simplest deterministic mathematical model for the spread of an epidemic disease is the so-called SI system made of two Ordinary Differential Equations. It exhibits simple dynamics: a bifurcation parameter $\mathscr{T}_0$ yielding persistence of the disease when $\mathscr{T}_0 > 1$, else extinction occurs. A natural question is whether this gentle dynamic can be disturbed by spatial diffusion. It is straightforward to check it is not feasible for linear/nonlinear diffusions. When cross diffusion is introduced for suitable choices of the parameter data set this persistent state of the ODE model system becomes linearly unstable for the resulting initial and no-flux boundary value problem. On the other hand "natural" weak solutions can be defined for this initial and no-flux boundary value problem and proved to exist provided nonlinear and cross diffusivities satisfy some constraints. These constraints are not fully met for the parameter data set yielding instability. A remaining open question is: to which solutions does this apply? Periodic behaviors are observed for a suitable range of cross diffusivities.

The original version of this chapter was revised: Belated correction has been incorporated. The correction to this chapter is available at https://doi.org/10.1007/978-3-319-78325-3_24

V. Anaya
Departamento de Matemática and GIMNAP, Universidad del Bío-Bío,
Concepción, Chile
e-mail: vanaya@ubiobio.cl

M. Bendahmane
Institut de Mathématiques de Bordeaux, UMR CNRS 5251,
Université de Bordeaux, 33405 TALENCE Cedex, Bordeaux, France
e-mail: mostafa.bendahmane@u-bordeaux.fr

M. Langlais (✉)
26 rue Cornac, Bordeaux, France
e-mail: michel.raymond.langlais@gmail.com

M. Sepúlveda
Universidad de Concepción, DIM and CI2 MA, Esteban Iturra s/n,
Barrio Universitario, Concepción, Chile
e-mail: mauricio@ing-mat.udec.cl

B. N. Chetverushkin et al. (eds.), *Contributions to Partial Differential Equations and Applications*, Computational Methods in Applied Sciences 47,
https://doi.org/10.1007/978-3-319-78325-3_5

1 Introduction

This study is related to a well-known generic question: can diffusion destabilize the otherwise globally asymptotically stable (GAS) stationary state of a system of ordinary differential equations (ODE), a question usually referred to as Turing bifurcation cf. [26]. We are more specifically interested in basic planar systems modeling the spread of an epidemic disease within a host population distributed over a spatial habitat and dispersing there. A natural question is whether a spatial structure can destabilize the spatially homogeneous endemic state yielding pattern formation.

There is a worldwide concern in Public Health Agencies about vector-borne diseases related to dispersal of mosquitoes infected by viruses such as zika or dengue to name a few or malaria parasites transmissible to Humans. Vector-borne diseases may also affect wild ruminants and domestic cattle with high economic consequences. The introduction in 2006 of the serotype 8 of the blue tongue virus—transmitted by midges to cattle—in Europe resulted in massive vaccination to control the disease, cf. [8] and references therein. The occurrence of both local or regional epidemics driven by spatiotemporal heterogeneities in distribution and abundance of host and vector populations was not considered, nor livestock movements managed by farmers and subsequent passive and active vector dispersal. A modeling approach was devised in [8] to assess through numerical experiments of a Reaction-Diffusion system how heterogeneities and dispersal effects can impact local and regional BT epidemics using various hypothetical scenarii illustrated in a real geographic area. It is observed there that spatial structures and dispersal do have a strong impact on observed dynamics, a mathematical analytical approach being beyond the scope of the study and likely out of reach!

It is our goal in this work to shed some light on this coupling between a spatial structure and population dispersal on the spread of an epidemic disease and the occurrence of epidemic patterns, using a simple planar Reaction-Diffusion model system and various dispersal fluxes, from linear to cross diffusion. A simple deterministic mathematical model system for the spread of an epidemic disease within an unstructured population is the SI system made of two ODEs, cf. [6, 7, 10] and references therein. It exhibits simple dynamics: there is a bifurcation parameter $\mathscr{T}_0$ yielding extinction of the disease when $\mathscr{T}_0 < 1$ and persistence of the disease when $\mathscr{T}_0 > 1$ in which case the SI-ODE system has a unique and GAS endemic or persistent stationary state (with positive entries), cf. [11] for details.

Our question is whether this is modified for a spatially structured host population. To this end one may devise Reaction-Diffusion model systems (RD), [15, 18, 22, 23, 25] and references therein, featuring linear, nonlinear or cross diffusions. For SI-RD systems linear diffusion does not provide any instability, cf. [11, 12]. This carries over in a straightforward fashion to nonlinear diffusion. It remains to consider cross diffusion. Supplying (partial) answers to this point is the main goal of this study.

When nonlinear and cross diffusions are considered two questions arise at once for the resulting initial and no-flux boundary value problem: (a) the motivating one that is finding "linear" stability/instability conditions for the spatially constant stationary state of the ODE system, and, (b), defining and proving the existence of suitable "weak solutions" to which this applies.

Looking at the linearized system about the persistent state of the ODE system and following routine algebraic calculations, cf. [22], one finds that for some choices of the parameter data set this persistent state may become unstable. This is illustrated in Figs. 1 and 2 for Turing bifurcation and Figs. 3, 4 and 5 for pattern formations. An endemic state is found in "algebraic" closed form only in two specific cases and given the large number (15) of parameters we are not yet able to handle this stability / instability analysis through a formal calculation for the full system.

Following [2, 3] "naive" weak solutions are defined in the spirit of [19]. Existence is proved when diffusivities satisfy a set of constraints, see Proposition 2. These constraints are not fully satisfied for the parameter data set yielding instability so that a further remaining open question is: to which type of weak solutions does this bifurcation analysis apply? Classical solutions are found under quite restrictive conditions, see Sect. 4.3, in which case instability is not possible, see Sect. 5.2. Turing bifurcation in Lotka-Volterra population dynamic models with cross diffusion is the scope of a large literature, cf. [4, 9, 15, 18, 25] and references therein.

In Sects. 2 and 3 we review known results for the SI-ODE system and the SI-RD system with linear/nonlinear diffusion. In Sect. 4 we look at the existence of weak solutions for the SI-RD initial and no-flux boundary conditions system with linear, nonlinear and cross diffusions. Section 5 is dedicated to stability/instability analysis of the endemic state of the SI-ODE system for the SI-RD system. A Conclusion and Perspective section completes this manuscript.

2 Underlying ODE Model System

Let us introduce a two component Ordinary Differential Equation system

$$\begin{aligned} S' &= -\sigma SI + bS + \theta b_I I - (m + kP)S, \\ I' &= \sigma SI - \alpha I + (1-\theta)b_I I - (m + kP)I \end{aligned} \tag{1}$$

with $P = S + I$. System (1) is commonly referred to as a spatially unstructured SI epidemic model system, cf. [6, 7, 10] and their references. The total host population density, P, is split into susceptible, S, and infectious, I, individuals. Concerning vital dynamics b is the fertility rate of susceptibles and b_I that of infectious, a proportion θ of offsprings born from infectious parents being susceptible at birth, m is the natural mortality rate while k represents a density dependent pressure on mortality. Concerning disease transmission σ is an aggregated transmission rate of the disease from S to I and α is an additional mortality rate due to the disease.

Throughout this work one shall assume

(H1) $\sigma > 0, b > m, k > 0$ as well as $\alpha \geq 0, 0 \leq b_I \leq b$ and $0 < \theta \leq 1$.

The limiting case $\theta = 0$ is left to the reader, cf. [11].

Given initial conditions satisfying $S(0) \geq 0$, $I(0) \geq 0$ with $S(0) + I(0) > 0$ the system (1) has a unique componentwise nonnegative and bounded solution. Set

$$\mathscr{T}_0 = \frac{\sigma K}{b + \alpha - (1-\theta)b_I}, \;\; K = \frac{b-m}{k}. \tag{2}$$

Then, cf. [11], $\mathscr{T}_0$ is a threshold parameter: provided $S(0) + b_0 > 0$ and $I(0) > 0$ when

$$\begin{cases} \mathscr{T}_0 < 1 & \text{the semi-trivial stationary state } S = K, I = 0 \text{ is GAS;} \\ \mathscr{T}_0 > 1 & \text{the semi-trivial stationary state } S = K, I = 0 \text{ is unstable and there exists} \\ & \text{a unique endemic or persistent state, } S^* > 0,\ I^* > 0, \text{ that is GAS.} \end{cases}$$

3 RD Model System with Linear/Nonlinear Diffusion

Let Ω be a bounded domain in R^N with smooth boundary $\partial\Omega$, Ω locally lying on one side of $\partial\Omega$. Assume host individuals disperse through their habitat by means of the nonlinear Fickian diffusion, $-d_u(U)\nabla U$ being the population flux for $U = S, I$. Set

$$D_u(U) = \int_0^U d_u(\upsilon)\, d\upsilon.$$

A Reaction-Diffusion system can be devised: for $x \in \Omega$ and $t > 0$

$$\begin{aligned} \partial_t S &= \triangle[D_1(S)] - \sigma SI + bS + \theta b_I I - (m + kP)S, \\ \partial_t I &= \triangle[D_2(I)] + \sigma SI - \alpha I + (1-\theta)b_I I - (m + kP)I \end{aligned} \tag{3}$$

equipped with no flux boundary conditions

$$d_1(S)(x,t)\nabla S(x,t) \cdot \eta(x) = d_2(I)(x,t)\nabla I(x,t) \cdot \eta(x) = 0, \quad x \in \partial\Omega,\ t > 0 \tag{4}$$

η being a unit normal vector to Ω along $\partial\Omega$, cf. [7, 13, 22, 23] and their references. One still set $P = S + I$ and assume the condition (H1) holds. One prescribes

$$0 < d_{\min} \leq d_1(S), d_2(I); \quad d_i \in C^2([0,\infty)),\ i = 1, 2. \tag{5}$$

Given bounded initial conditions satisfying $S(0,x) = S_0(x) \geq 0$, $I(0,x) = I_0(x) \geq 0$ with $S_0(x) + I_0(x) \not\equiv 0$ the system (3)–(4) has a unique componentwise nonnegative and bounded classical solution, cf. [11, 12, 14].

Proposition 1 *Assume $\mathscr{T}_0 > 1$. The unique endemic state, $S^* > 0$, $I^* > 0$, of the system (1) remains GAS for (3)–(4) for those nonnegative and bounded initial conditions such that $S_0 \not\equiv 0$ and $I_0 \not\equiv 0$.*

Proof Given positive u and v in $C^0(\bar{\Omega})$ and positive numbers ν_S and ν_I set

$$\mathscr{L}(u, v) = \nu_S \int_\Omega \Big(u(x) - S^* - S^* \ln \frac{u(x)}{S^*}\Big) dx + \nu_I \int_\Omega \Big(v(x) - I^* - I^* \ln \frac{v(x)}{I^*}\Big) dx,$$

cf. [16]. Rather straightforward calculations, cf. [11, 12], lead to

$$\frac{d}{dt}\mathscr{L}(S(\cdot, t), I(\cdot, t)) \le 0, \quad t > 0,$$

yielding $\mathscr{L}$ is a Lyapunov functional. From the LaSalle invariance principle, cf. [17], $\mathscr{L}$ is constant on the largest invariant subset of the ω-limit set of (3)–(4) in $[C^0(\bar{\Omega})]^2$ and this ω-limit set reduces to (S^*, I^*). □

4 A Prototypical RD Model System with Cross Diffusion 1

Let us now consider a prototypical model system involving nonlinear and cross diffusion posed on the same $\Omega \subset \mathbb{R}^N$. For $x \in \Omega$ and $t > 0$ it reads

$$\begin{aligned} \partial_t S &= \triangle[(d_1 + d_{11}S + d_{12}I)S] - \sigma SI + bS + \theta b_I I - (m + kP)S; \\ \partial_t I &= \triangle[(d_2 + d_{21}S + d_{22}I)I] + \sigma SI - \alpha I + (1 - \theta)b_I I - (m + kP)I \end{aligned} \tag{6}$$

where one has set $P = S + I$, cf. [15, 18, 22, 23, 25] and their references. This also reads for suitably smooth functions S and I

$$\begin{aligned} \partial_t S &= \operatorname{div}[(d_1 + 2d_{11}S + d_{12}I)\nabla S + d_{12}S\nabla I] - \sigma SI + bS + \theta b_I I - (m + kP)S; \\ \partial_t I &= \operatorname{div}[(d_2 + d_{21}S + 2d_{22}I)\nabla I + d_{21}I\nabla S] + \sigma SI - \alpha I + (1 - \theta)b_I I - (m + kP)I \end{aligned} \tag{7}$$

equipped with initial conditions

$$S(x, 0) = S_0(x) \ge 0, \quad I(x, 0) = I_0(x) \ge 0, \quad x \in \Omega \tag{8}$$

and no flux boundary conditions for $x \in \partial\Omega$ and $t > 0$

$$\begin{aligned} &\Big[(d_1 + 2d_{11}S + d_{12}I)\nabla S + d_{12}S\nabla I\Big](x, t) \cdot \eta(x) \\ &\qquad = [(d_2 + d_{21}S + 2d_{22}I)\nabla I + d_{21}I\nabla S]\,(x, t) \cdot \eta(x) = 0 \end{aligned} \tag{9}$$

η being a normal unit vector to Ω along $\partial\Omega$. Conditions (9) are equivalent to a linear system in the $(\nabla S(x, t) \cdot \eta(x), \nabla I(x, t) \cdot \eta(x))$ variables. For nonnegative

$(S(x,t), I(x,t))$ and $(d_{ij})_{i,j=1,2}$ and positive $(d_i)_{i=1,2}$ this linear system is invertible and conditions (9) are equivalent to the homogeneous Neumann boundary conditions

$$\nabla S(x,t)\cdot\eta(x) = \nabla I(x,t)\cdot\eta(x) = 0, \quad x\in\partial\Omega,\ t>0. \tag{10}$$

4.1 Towards Some Ellipticity Condition(s)

At several steps of our analysis given $f(S)\geq 0$, $f(I)\geq 0$ we are to find the signum of a quadratic form in the $(\nabla S, \nabla I)\in(L^2(\Omega))^{2N}$ variable

$$\int_\Omega \left[2d_{11}|\nabla S|^2 + d_{12}\nabla I\cdot\nabla S + d_{21}|\nabla I|^2\right] f(S)\,dx \\ + \int_\Omega \left[d_{12}|\nabla S|^2 + d_{21}\nabla S\cdot\nabla I + 2d_{22}|\nabla I|^2\right] f(I)\,dx. \tag{11}$$

A pointwise underlying $2N\times 2N$ symmetrical matrix is a 2×2 block matrix

$$M(S,I) = \begin{pmatrix} (2d_{11}f(S)+d_{12}f(I))Id_N & \frac{1}{2}(d_{12}f(S)+d_{21}f(I))Id_N \\ \frac{1}{2}(d_{12}f(S)+d_{21}f(I))Id_N & (2d_{22}f(I)+d_{21}f(S))Id_N \end{pmatrix}$$

wherein Id_N is the identity matrix in $\mathscr{R}^N$.

Lemma 1 *Assume $f(S)\geq 0$, $f(I)\geq 0$. When $8d_{11}d_{21}\geq d_{12}^2$ and $8d_{22}d_{12}\geq d_{21}^2$ then pointwise matrix $M(S,I)$ is nonnegative. It follows the quantity in (11) is nonnegative. Else pointwise matrix $M(S,I)$ is nonnegative provided $(f(S), f(I))$ satisfy some set of pointwise additional constraints.*

Proof Matrix $M(S,I)$ characteristic polynomial factors out

$$P(\lambda) = \frac{1}{4^N}\left[4\lambda^2 - 4\left(d_{21}f(S)+2d_{11}f(S)+d_{12}f(I)+2d_{22}f(I)\right)\lambda + R\right]^N.$$

Setting $f(I)=\rho f(S)$, $\rho\geq 0$, one gets $R = 4f(S)^2Q(\rho)$

$$Q(\rho) = (8d_{11}d_{21}-d_{12}^2) + (16d_{11}d_{22}+2d_{12}d_{21})\rho + (8d_{22}d_{12}-d_{21}^2)\rho^2.$$

Then $Q(\rho)\geq 0$ for $\rho\geq 0$ provided $8d_{11}d_{21}\geq d_{12}^2$ and $8d_{22}d_{12}\geq d_{21}^2$. □

4.2 Main Existence Result (Weak Solutions)

Definition 1 Let $T > 0$ be fixed. A componentwise nonnegative weak solution to the system (7), (9), (8) is a duet $(S, I) \in [L^3((0,T) \times \Omega)^+ \cap L^\infty(0,T; L^2(\Omega))]^2$ such that

$$(\nabla S, \nabla I) \in \left(L^2((0,T) \times \Omega)\right)^{2N};$$
$$(\partial_t S, \partial_t I) \in L^{\frac{6}{5}}\left(0,T; [W^{1,6}(\Omega)]'\right) \times L^{\frac{6}{5}}\left(0,T; [W^{1,6}(\Omega)]'\right)$$

satisfying (8) and such that for any $(\varphi, \psi) \in [L^6(0,T; W^{1,6}(\Omega))]^2$ one has

$$\begin{aligned}\int_\Omega \langle \partial_t S, \varphi\rangle\, dx &= -\int_\Omega [(d_1 + 2d_{11}S + d_{12}I)\nabla S + d_{12}S\nabla I] \cdot \nabla\varphi\, dx \\ &\quad + \int_\Omega [\theta b_I I + (b - m - k(S+I))S - \sigma(S,I)]\,\varphi\, dx; \\ \int_\Omega \langle \partial_t I, \psi\rangle\, dx &= -\int_\Omega [(d_2 + d_{21}S + 2d_{22}I)\nabla I + d_{21}I\nabla S] \cdot \nabla\psi\, dx \\ &\quad + \int_\Omega [\sigma(S,I) + ((1-\theta)b_I - m - k(S+I) - \alpha)I]\,\psi\, dx.\end{aligned} \tag{12}$$

Herein $\langle \cdot, \cdot \rangle$ is the duality pairing between $W^{1,6}(\Omega)$ and its dual space $[W^{1,6}(\Omega)]'$.

Proposition 2 *Let $d_i > 0$, $i = 1, 2$ and $d_{ij} > 0$, $1 \le i, j \le 2$. Assume either uncoupled requirements for reactive and diffusive terms*

(H2) Coefficients σ and k satisfy

$$0 < \sigma < 2(1+\sqrt{2})k \iff \frac{\sigma^2}{4k(k+\sigma)} < 1;$$

(H3) Diffusivities satisfy $8d_{11}d_{21} \ge d_{12}^2$ and $8d_{22}d_{12} \ge d_{21}^2$;

or coupled requirements for diffusive and reactive terms

(H23) There exists a $\gamma > \frac{\sigma^2}{4k(k+\sigma)}$ such that $8d_{11}d_{21} \ge \gamma d_{12}^2$; $8\gamma d_{22}d_{12} \ge d_{21}^2$.

Given bounded and nonnegative initial conditions (S_0, I_0) satisfying $S_0(x) + I_0(x) \not\equiv 0$, for any $T > 0$ the system (7)–(9) has a componentwise nonnegative weak solution according to Definition 1.

Remark 1 When either conditions (H3) or (H23) is strengthened into strict inequalities one gets a componentwise nonnegative slightly modified weak solution, that is a duet $(S, I) \in [L^3((0,T) \times \Omega)^+ \cap L^\infty(0,T; L^2(\Omega))]^2$ such that

$$(1+\sqrt{X})\|\nabla Y\| \in L^2\left((0,T)\times\Omega\right), \quad (X,Y)\in\{S,I\};$$
$$(\partial_t S, \partial_t I) \in L^{\frac{3}{2}}\left(0,T;[W^{1,3}(\Omega)]^{'}\right) \times L^{\frac{3}{2}}\left(0,T;[W^{1,3}(\Omega)]^{'}\right)$$

a solution to (12) for $(\varphi, \psi) \in [L^3(0,T;W^{1,3}(\Omega))]^2$.

4.3 Existence of Classical Solutions

Assume

$$d_1 = d_2 = d^* > 0, \; d_{ij} = d^{**} > 0, \quad i,j = 1,2 \tag{13}$$

so that condition (H3) is obviously satisfied. In that case upon adding the equations for S and I one gets a logistic-like inequality for the total population $P = S + I$

$$\partial_t P \le \operatorname{div}\left[(d^* + 2d^{**}P)\nabla P\right] + bP - (m + kP)P$$

equipped with initial and no-flux boundary conditions. From this a uniform a-priori spatio-temporal L^∞ estimate follows at once yielding global existence of a classical solution, starting from a local classical solution whose existence is found in [1, 27].

4.4 Sketch of the Proof of Proposition 2

Cf. [2, 3].

4.4.1 Approximating Scheme for Fixed $\varepsilon > 0$

Let X^+ be the positive part of X, say $X^+ = \frac{X+|X|}{2}$. Set

$$f_\varepsilon(x) = \frac{x}{1+\varepsilon x}, \;\; x \ge 0; \quad \sigma_\varepsilon(x,y) = \sigma\frac{xy}{1+\varepsilon(x+y)}, \quad x \ge 0, \; y \ge 0.$$

One shall consider the following approximating scheme:

$$\begin{aligned} \partial_t S &= \operatorname{div}\left[(d_1 + 2d_{11}f_\varepsilon(S^+) + d_{12}f_\varepsilon(I^+))\nabla S + d_{12}f_\varepsilon(S^+)\nabla I\right] \\ &\quad + \theta b_I|I| + (b - m - k(|S|+|I|))S - \sigma_\varepsilon(S^+, I^+); \\ \partial_t I &= \operatorname{div}\left[(d_2 + d_{21}f_\varepsilon(S^+) + 2d_{22}f_\varepsilon(I^+))\nabla I + d_{21}f_\varepsilon(I^+)\nabla S\right] \\ &\quad + \sigma_\varepsilon(S^+, I^+) + ((1-\theta)b_I - m - k(|S|+|I|) - \alpha)I \end{aligned} \tag{14}$$

equipped with initial conditions (8) and no flux boundary conditions (10).

Definition 2 A weak solution to (14), (8), (10) over $(0, T) \times \Omega$ is a duet (S, I) lying in $[L^2((0, T) \times \Omega) \cap L^2(0, T; H^1(\Omega))]^2$ with $(\partial_t S, \partial_t I) \in [L^2(0, T; [H^1(\Omega)]')]^2$ satisfying (8) and such that for any $(\varphi, \psi) \in [L^2((0, T) \times \Omega) \cap L^2(0, T; H^1(\Omega))]^2$

$$
\begin{aligned}
&\int_\Omega \langle \partial_t S, \varphi\rangle \, dx = \\
&- \int_\Omega \left[\left(d_1 + 2d_{11} f_\varepsilon(S^+) + d_{12} f_\varepsilon(I^+)\right) \nabla S + d_{12} f_\varepsilon(S^+) \nabla I\right] \cdot \nabla \varphi \, dx \\
&+ \int_\Omega \left[\theta b_I |I| + (b - m - k(|S| + |I|))\, S - \sigma_\varepsilon(S^+, I^+)\right] \varphi \, dx; \\
&\int_\Omega \langle \partial_t I, \psi\rangle \, dx = \\
&- \int_\Omega \left[\left(d_2 + d_{21} f_\varepsilon(S^+) + 2d_{22} f_\varepsilon(I^+)\right) \nabla I + d_{21} f_\varepsilon(I^+) \nabla S\right] \cdot \nabla \psi \, dx \\
&+ \int_\Omega \left[\sigma_\varepsilon(S^+, I^+) + ((1 - \theta) b_I - m - k(|S| + |I|) - \alpha)\, I\right] \psi \, dx.
\end{aligned}
$$

Herein $\langle \cdot, \cdot \rangle$ is the duality pairing between $H^1(\Omega)$ and its dual space $[H^1(\Omega)]'$.

Note that for fixed $\varepsilon > 0$ and nonnegative entries $f_\varepsilon(x)$ is bounded while $\sigma_\varepsilon(x, y)$ is sub-linear. A proof of the existence of a such a weak solution is found in [2, 3].

4.4.2 Nonnegativity of Components of $(S_\varepsilon, I_\varepsilon)_{\varepsilon>0}$

For sake of simplicity let us drop subscript ε. Setting $\varphi = -S^-$ in Definition 2 yields

$$
\begin{aligned}
&- \int_\Omega \partial_t S S^-(x, t) \, dx = \\
&\int_\Omega \left[\left(d_1 + 2d_{11} f_\varepsilon(S^+) + d_{12} f_\varepsilon(I^+)\right) \nabla S \cdot \nabla S^- + d_{12} f_\varepsilon(S^+) \nabla I \cdot \nabla S^-\right] dx \\
&- \theta b_I \int_\Omega |I| S^- \, dx - \int_\Omega (b - m - k(|S| + |I|)) S S^- dx + \int_\Omega \sigma_\varepsilon(S^+, I^+) S^- \, dx
\end{aligned}
$$

that implies

$$
\frac{1}{2} \frac{d}{dt} \int_\Omega (S^-)^2(x, t) \, dx \leq (b - m) \int_\Omega (S^-)^2(x, t) \, dx.
$$

As a consequence of $S(x, 0) \geq 0$ one finds $S(x, t) \geq 0$ for $x \in \Omega$ and $t > 0$. Proceeding in a similar fashion, one finds $I(x, t) \geq 0$ for $x \in \Omega$ and $t > 0$.

4.4.3 Uniform Bounds for $\varepsilon > 0$

Let us again drop the index ε. One has a set of componentwise nonnegative weak solutions to (14), (8), (10) according to Definition 2.

Setting $\varphi = \psi = 1$ in the equations for S and I from Definition 2 and adding the resulting equations, non-negativity arguments yield for $P = S + I$

$$\frac{d}{dt}\int_{\Omega} P(x,t)\,dx \le \int_{\Omega} (b - m - kP)P\,dx.$$

From Cauchy-Schwarz's inequality it follows P is a solution to

$$\frac{d}{dt}\int_{\Omega} P(x,t)\,dx \le \int_{\Omega} (b-m)P\,dx - \frac{k}{|\Omega|}\left(\int_{\Omega} P(x,t)\,dx\right)^2$$

and there exists a continuous function $M_2 : [0,+\infty) \to [0,+\infty)$ such that

$$\begin{cases} \|P(\cdot,t)\|_{1,\Omega} \le \max\left(\|P(\cdot,0)\|_{1,\Omega}, \frac{b-m}{k}|\Omega|\right), & t > 0; \\ \|P\|_{2,\Omega\times(0,T)} \le M_2(T), & T > 0. \end{cases}$$

Setting $\varphi = S$ in the equation for S from Definition 2, $\psi = I$ in the equation for I from Definition 2 and adding the resulting equations one gets

$$\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\int_{\Omega} S^2(x,t)\,dx + \frac{1}{2}\frac{d}{dt}\int_{\Omega} I^2(x,t)\,dx + d_1\int_{\Omega} |\nabla S|^2\,dx + d_2\int_{\Omega} |\nabla I|^2\,dx \\
&+ \int_{\Omega} \left[2d_{11}|\nabla S|^2 + d_{12}\nabla I\cdot\nabla S + d_{21}|\nabla I|^2\right] f_\varepsilon(S)\,dx \\
&+ \int_{\Omega} \left[d_{12}|\nabla S|^2 + d_{21}\nabla S\cdot\nabla I + 2d_{22}|\nabla I|^2\right] f_\varepsilon(I)\,dx \\
&\le \theta b_I \int_{\Omega} ISdx + \int_{\Omega} (b - m - k(S+I))S^2\,dx - \int_{\Omega} \sigma_\varepsilon(S,I)S\,dx \\
&+ \int_{\Omega} \sigma_\varepsilon(S,I)I\,dx + \int_{\Omega} ((1-\theta)b_I - m - k(S+I) - \alpha)\,I^2\,dx.
\end{aligned} \tag{15}$$

From Cauchy-Schwarz's inequality one gets for $\rho > 0$

$$\int_{\Omega} \sigma_\varepsilon(S,I)I\,dx \le \frac{\sigma^2}{4k}(1+\rho)\int_{\Omega} \frac{1}{1+\varepsilon P}S^2 I\,dx + \frac{k}{1+\rho}\int_{\Omega} \frac{1}{1+\varepsilon P} I^3\,dx.$$

It follows from (15) and the condition (H3)

$$\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\int_{\Omega} S^2(x,t)\,dx + \frac{1}{2}\frac{d}{dt}\int_{\Omega} I^2(x,t)\,dx + d_1\int_{\Omega} |\nabla S|^2\,dx + d_2\int_{\Omega} |\nabla I|^2\,dx\\
&\qquad \le \theta b_I \int_{\Omega} IS\,dx + \int_{\Omega} (b-m)S^2\,dx + \int_{\Omega} ((1-\theta)b_I - m - \alpha)I^2\,dx\\
&\quad -k\int_{\Omega} S^3\,dx - \frac{\rho}{1+\rho}k\int_{\Omega} I^3\,dx - \int_{\Omega} \left[k(1+\varepsilon P)+\sigma - \frac{\sigma^2}{4k}(1+\rho)\right]\frac{S^2 I}{1+\varepsilon P}\,dx.
\end{aligned}$$

Using the condition (H2) or equivalently $4k(k+\sigma)-\sigma^2 > 0$ for small enough $\rho > 0$ there exists a continuous function $M_3 : [0,+\infty) \to [0,+\infty)$ such that

$$\max_{0<t<T} \|P(\cdot,t)\|_{2,\Omega} + \|\nabla S\|_{2,\Omega\times(0,T)} + \|\nabla I\|_{2,\Omega\times(0,T)} + \|P\|_{3,\Omega\times(0,T)} \le M_3(T). \tag{16}$$

From Hölder's inequality one gets

$$\begin{aligned}
\|f_\varepsilon(X_\varepsilon)\nabla Y_\varepsilon\|_{\frac{6}{5},\Omega\times(0,T)} &\le \|X_\varepsilon\|_{3,\Omega\times(0,T)}\|\nabla Y_\varepsilon\|_{2,\Omega\times(0,T)}, \quad (X,Y)\in\{S,I\}\\
\|\sigma_\varepsilon(S_\varepsilon,I_\varepsilon)\|_{\frac{3}{2},\Omega\times(0,T)} &\le \|\sigma I_\varepsilon S_\varepsilon\|_{\frac{3}{2},\Omega\times(0,T)} \le \sigma\|S_\varepsilon\|_{3,\Omega\times(0,T)}\|I_\varepsilon\|_{3,\Omega\times(0,T)}.
\end{aligned}$$

The a priori estimate in (16) implies the existence of a continuous function $M^{'}$: $[0,+\infty) \to [0,+\infty)$ such that for each $T > 0$

$$\|\partial_t S_\varepsilon\|_{L^{\frac{6}{5}}(0,T;[W^{1,6}(\Omega)]')} + \|\partial_t I_\varepsilon\|_{L^{\frac{6}{5}}(0,T;[W^{1,6}(\Omega)]')} \le M^{'}(T). \tag{17}$$

4.4.4 Convergence as $\varepsilon \to 0$

Above estimates show the sequence $(S_\varepsilon, I_\varepsilon)_{\varepsilon>0}$ is bounded in $\big(L^2((0,T);\, H^1(\Omega))\big)^2$ while the sequence $(\partial_t S_\varepsilon, \partial_t I_\varepsilon)_{\varepsilon>0}$ is bounded in $\Big(L^{\frac{6}{5}}(0,T;\,[W^{1,6}(\Omega)]')\Big)^2$ as $\varepsilon \to 0$. Aubin's Lemma asserts that for each $T > 0$ the sequence $(S_\varepsilon, I_\varepsilon)_{\varepsilon>0}$ lies in a compact set of $\big(L^2(\Omega)\times(0,T)\big)^2 \simeq \big(L^2((0,T);\,L^2(\Omega))\big)^2$.

One may extract a subsequence still denoted $(S_\varepsilon, I_\varepsilon)_{\varepsilon>0}$ converging to some nonnegative (S,I) as $\varepsilon \to 0$, strongly in $\big(L^2((0,T);\,L^2(\Omega))\big)^2$, a.e. $(x,t)\in\Omega\times(0,T)$ and weakly in $\big(L^2((0,T);\,H^1(\Omega))\big)^2$. One may also assume $(\partial_t S_\varepsilon, \partial_t I_\varepsilon)_{\varepsilon>0}$ is weakly converging to $(\partial_t S, \partial_t I)$ in $\Big(L^{\frac{6}{5}}(0,T;\,[W^{1,6}(\Omega)]')\Big)^2$ as $\varepsilon \to 0$.

Nonlinear terms such as $(f_\varepsilon(I_\varepsilon)\nabla S_\varepsilon)_{\varepsilon>0}$ are bounded in $L^{\frac{6}{5}}((0,T);\,L^{\frac{6}{5}}(\Omega))$. Set

$$f_\varepsilon(I_\varepsilon)\nabla S_\varepsilon = \big[f_\varepsilon(I_\varepsilon) - I\big]\nabla S_\varepsilon + I\nabla S_\varepsilon = \Lambda_\varepsilon^1 + \Lambda_\varepsilon^2.$$

One has $\Lambda_\varepsilon^1 \to 0$ strongly in $L^1((0,T);\,L^1(\Omega))$ as $\varepsilon \to 0$ because

$$|f_\varepsilon(I_\varepsilon) - I| \times \|\nabla S_\varepsilon\| \le \big[|I_\varepsilon - I| + \varepsilon|I|\big] \times \|\nabla S_\varepsilon\|$$

while $\Lambda_\varepsilon^2 \to I\nabla S$ weakly in $L^{\frac{6}{5}}((0,T); L^{\frac{6}{5}}(\Omega))$ as $\varepsilon \to 0$. Hence one may assume $(f_\varepsilon(I_\varepsilon)\nabla S_\varepsilon)_{\varepsilon>0}$ is weakly converging to $I\nabla S$ in $L^{\frac{6}{5}}((0,T); L^{\frac{6}{5}}(\Omega))$ as $\varepsilon \to 0$.

Next $(\sigma_\varepsilon(S_\varepsilon, I_\varepsilon))_{\varepsilon>0}$ is bounded in $L^{\frac{6}{5}}((0,T); L^{\frac{6}{5}}(\Omega))$. A mere splitting yields

$$|\sigma_\varepsilon(S_\varepsilon, I_\varepsilon) - \sigma_0(S, I)| \leq \sigma|S_\varepsilon - S|\,|I| + \sigma|I - I_\varepsilon|\,|S_\varepsilon| + \varepsilon\sigma|S_\varepsilon + I|\,|S|S_\varepsilon|$$

and $(\sigma_\varepsilon(S_\varepsilon, I_\varepsilon))_{\varepsilon>0}$ converges to $\sigma_0(S, I)$ strongly in $L^1((0,T); L^1(\Omega))$ as $\varepsilon \to 0$, and weakly in $L^{\frac{6}{5}}((0,T); L^{\frac{6}{5}}(\Omega))$.

Classical arguments show one may extract a subsequence converging to a componentwise nonnegative weak solution to system (7), (9), (8).

4.4.5 Alternate Take wrt the Condition (H23)

Setting $\varphi = \gamma S$, $\psi = I$ in the equations from Definition 2 and adding the resulting equations similar computations yield a weak solution.

Remark 2 When condition (H3) is satisfied with strict inequalities, see Remark 1, one may derive from (15) an additional a-priori estimate

$$\int_0^T \int_\Omega \left(|\nabla S_\varepsilon|^2 + |\nabla I_\varepsilon|^2\right)\left(f_\varepsilon(S_\varepsilon^+) + f_\varepsilon(I_\varepsilon^+)\right)(x,t)\,dx\,dt \leq M_2^\sharp(T).$$

Our nonnegative weak solutions will now satisfy

$$\int_0^T \int_\Omega (|\nabla S|^2 + |\nabla I|^2)(S + I)(x,t)\,dx\,dt \leq M_2^\sharp(T).$$

5 Prototypical RD Model System with Cross Diffusion 2

When $\mathscr{T}_0 > 1$ a natural question to address is whether the unique endemic state of the ODE system (1) remains stable for those nonnegative and bounded initial conditions such that $S_0 \not\equiv 0$ and $I_0 \not\equiv 0$ for the RD system (6).

The linearized diffusion matrix from system (6) evaluated at (S^*, I^*) is

$$D^* = \begin{pmatrix} d_1 + 2d_{11}S^* + d_{12}I^* & d_{12}S^* \\ d_{21}I^* & d_2 + d_{21}S^* + 2d_{22}I^* \end{pmatrix}$$

from which one easily gets $\mathrm{trace}(D^*) > 0$, $\det(D^*) > 0$.

On the other hand let J^* be the Jacobian matrix for the ODE system (1) evaluated at the endemic state (S^*, I^*). When $\mathscr{T}_0 > 1$ one has $\mathrm{trace}(J^*) < 0$, $\det(J^*) > 0$.

5.1 Linearized System Assuming $\mathscr{T}_0 > 1$

Linearizing system (6)–(9) about the persistent stationary state (S^*, I^*) yields

$$\partial_t \begin{pmatrix} u \\ v \end{pmatrix} = D^* \triangle \begin{pmatrix} u \\ v \end{pmatrix} + J^* \begin{pmatrix} u \\ v \end{pmatrix} \tag{18}$$

equipped with no flux boundary conditions (10).

Let $(\mu_j \geq 0, \varphi_j)_{j \geq 0}$ be the eigenvalues and eigenfunctions to the eigenvalue problem

$$-\triangle\varphi(x) = \mu\varphi(x), \quad x \in \Omega; \qquad \nabla\varphi(x) \cdot \eta(x) = 0, \quad x \in \partial\Omega. \tag{19}$$

Looking for a solution to (18), (10) of the parametric form $\exp(\lambda t)\varphi_j(x)\binom{u}{v}$ one gets a familiar eigenvalue problem in $\mathbb{R}^2$, say (Id_2 being the identity matrix in $\mathbb{R}^2$)

$$\left(\lambda Id_2 - [-\mu_j D^* + J^*]\right) \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

The question now is whether λ can be positive that is whether instability can be driven by cross diffusion. Setting

$$M_0 = J^* - \mu_j D^*,$$

it follows trace$(M_0) < 0$ so that instability is feasible if and only if $\det(M_0) < 0$.

A first calculation yields $\det(M_0)$ is a quadratic function of μ_j: for some linear function Θ of diffusivities (d_i, d_{ij}) one has

$$\det(M_0) = \mu_j^2 \det(D^*) + \mu_j \Theta(d_1, d_2, d_{11}, d_{12}, d_{21}, d_{22}) + \det(J^*). \tag{20}$$

A necessary condition for instability, that is $\det(M_0) < 0$, is to find a set of nonnegative diffusivities implying $\Theta(d_1, d_2, d_{11}, d_{12}, d_{21}, d_{22}) < 0$:

1. According to Sect. 3 this cannot come from linear or non-linear diffusion. Actually for $d_1 > 0$, $d_2 > 0$, $d_{11} > 0$ and $d_{22} > 0$ one finds

$$\begin{cases} \Theta(d_1, 0, 0, 0, 0, 0) = d_1 k I^* > 0; \\ \Theta(0, d_2, 0, 0, 0, 0) = d_2 \dfrac{kS^{*2} + \theta b_I I^*}{S^*} > 0; \\ \Theta(0, 0, d_{11}, 0, 0, 0) = d_{11} 2kS^* I^* > 0; \\ \Theta(0, 0, 0, 0, 0, d_{22}) = d_{22} 2\left[kS^* I^* + \theta b_I \dfrac{I^*}{S^*} I^*\right] > 0. \end{cases}$$

2. Looking at cross diffusivities leaves some hope to get a bifurcation. One gets

$$\begin{cases} \Theta(0,0,0,d_{12},0,0) = -d_{12}(kS^* - kI^* - \sigma S^*)I^*; \\ \Theta(0,0,0,0,d_{21},0) = d_{21}(kS^{*2} - kS^*I^* - \sigma S^*I^* + 2\theta b_I I^*). \end{cases}$$

5.2 Some Negative Results

Proposition 3 *Assume $\mathscr{T}_0 > 1$ and diffusivities are positive. Then the following hold:*

(i) $\det(M_0) > 0$ *when diffusivities satisfy*

$$k(2d_{11} + 2d_{22} - d_{12} - d_{21}) + \sigma(d_{12} - d_{21}) \geq 0. \tag{21}$$

When the condition (13) holds no Turing bifurcation can be expected.

(ii) When $\sigma > k$ then $\Theta(0,0,0,d_{12},0,0) > 0$ for $d_{12} > 0$.

(iii) When either $b_I = 0$ or $\theta = 1$ then $\Theta(0,0,0,d_{12},0,0) > 0$ for $d_{12} > 0$.

Proof (i) It is enough to check that

$$\Theta(0,0,d_{11},d_{12},d_{21},d_{22}) \geq [k(2d_{11} + 2d_{22} - d_{12} - d_{21}) + \sigma(d_{12} - d_{21})]\, S^*I^*.$$

When condition (13) holds true then condition (21) is hopefully satisfied.

(ii) When $\sigma > k$ then $kS^* - kI^* - \sigma S^* < 0$ and the conclusion follows at once.

(iii) When either $b_I = 0$ or $\theta = 1$ then

$$\mathscr{T}_0 = \frac{\sigma K}{b + \alpha} > 1 \iff \frac{\sigma}{k}\frac{b - m}{b + \alpha} > 1 \Rightarrow \sigma > k.$$

□

This states that when $\mathscr{T}_0 > 1$ and condition (13) hold then the unique endemic state of the ODE system (1) remains linearly stable for classical solutions of the PDE system (6) emanating from nonnegative and bounded initial conditions such that $S_0 \not\equiv 0$ and $I_0 \not\equiv 0$.

5.3 Algebraically Closed Form Endemic States

This occurs in two limiting cases. We shall repeatedly use

$$\Theta(0,0,0,d_{12},0,0) = d_{12}\Theta(0,0,0,1,0,0);$$
$$\Theta(0,0,0,0,d_{21},0) = d_{21}\Theta(0,0,0,0,1,0).$$

Case 1. $b_I = b$ and $\alpha = 0$.

One finds

$$\mathscr{T}_0 = \frac{\sigma}{\theta b}K = \frac{\sigma}{\theta b}\frac{b-m}{k}, \quad S^* + I^* = K, \quad S^* = \frac{\theta b}{\sigma}, \quad \text{and } I^* = K - S^*.$$

First

$$\Theta(0,0,0,1,0,0) = \frac{1}{\sigma}\left[-2\theta bk + (b - m + \theta b)\sigma\right] I^* = \frac{1}{\sigma}L(\sigma)I^*.$$

Keeping in mind $\mathscr{T}_0 > 1$ after some algebra one gets $\Theta(0,0,0,1,0,0) < 0$ if and only if σ lies in the range $\left(\frac{\theta b}{b-m}k, \frac{2\theta b}{b-m+\theta b}k\right)$, this range being non-empty provided $b - m - \theta b > 0$.

Second, one has

$$\Theta(0,0,0,0,1,0) = \frac{\theta b}{\sigma^2}\left[2\theta bk^2 - k(b - m + \theta b)\sigma + (b-m)\sigma^2\right] = \frac{1}{\sigma}Q_1(\sigma)S^*$$

Q_1 being a quadratic convex function. Keeping in mind $\mathscr{T}_0 > 1$ after some algebra this implies $\Theta(0,0,0,0,1,0) > 0$ in the range $\frac{\theta b}{b-m}k < \sigma < \frac{2\theta b}{b-m+\theta b}k$.

As a conclusion when $\mathscr{T}_0 > 1$ $\Theta(0,0,0,d_{12},0,0)$ and $\Theta(0,0,0,0,d_{21},0)$ cannot be both negative within the same range of the parameter data set.

The case of classical solutions driven by (13)
Let us still assume $b_I = b$ and $\alpha = 0$ as well as $d_1 = d_2 = d^* > 0$, $d_{ij} = d^{**} > 0$, $i, j = 1, 2$. One gets a logistic-like equation for P

$$\partial_t P = \text{div}\left[(d^* + 2d^{**}P)\nabla P\right] + bP - (m + kP)P$$

equipped with initial and no-flux boundary conditions. This implies $P(t,\cdot) \to K = \frac{b-m}{k}$ as $t \to +\infty$ in any reasonably strong norm wrt building some ω-limit-set.

Substituting this back in the equation for I one finds a further logistic PDE for I

$$\partial_t I = \text{div}\left[(d^* + d^{**}K)\nabla I\right] + \left[(\sigma K - \theta b) - \sigma I\right] I$$

equipped with initial and no-flux boundary conditions. This implies $I(t,\cdot) \to K - \frac{\theta b}{\sigma}$ as $t \to +\infty$. Thus when $b_I = b$, $\alpha = 0$ and (13) holds the unique endemic state of the ODE system (1) is GAS for those nonnegative and bounded initial conditions such that $S_0 \not\equiv 0$ and $I_0 \not\equiv 0$ for classical solutions of the PDE system (6).
Case 2. $b_I = 0$ and $\alpha = 0$.

One finds

$$\mathscr{T}_0 = \frac{\sigma}{b}K = \frac{\sigma}{b}\frac{b-m}{k}, \quad S^* + I^* = \frac{b}{\sigma}, \quad \text{and} \quad S^* = \frac{m\sigma + kb}{\sigma^2}.$$

First $\mathscr{T}_0 > 1$ implies $\Theta(0,0,0,d_{12},0,0) > 0$ for $d_{12} > 0$ by Proposition 3.

Second, one has

$$\Theta(0,0,0,0,1,0) = \frac{m\sigma + bk}{\sigma^4}\left[2km\sigma + 2bk^2 - b\sigma^2 + m\sigma^2\right] = \frac{1}{\sigma^2}Q_2(\sigma)S^*$$

Q_2 being a quadratic concave function achieving its maximum value at $\sigma = \frac{m}{b-m}k$.

Now $\mathscr{T}_0 > 1 \Rightarrow \sigma > \frac{b}{b-m}k$ while

$$Q_2(\sigma) = 0 \quad \text{and} \quad \sigma > 0 \iff \sigma = \frac{m + \sqrt{2b^2 - 2mb + m^2}}{b-m}k > \frac{b}{b-m}k.$$

Thus when $\mathscr{T}_0 > 1$ and $\sigma < 2(1+\sqrt{2})k$ are both satisfied $\Theta(0,0,0,0,d_{12},0) < 0$ in the range $\frac{m+\sqrt{2b^2-2mb+m^2}}{b-m}k < \sigma < 2(1+\sqrt{2})k$ provided $b > 2\frac{4+3\sqrt{2}}{5+4\sqrt{2}}m$.

Example 1 The data set is

$$b = 2, \quad b_I = 2, \quad \theta = 0.3, \quad m = 1, \quad \alpha = 0.2, \quad k = 0.2, \quad \sigma = 1.6,$$
$$d_1 = d_2 = 0.05, \quad d_{11} = d_{22} = d_{21} = 0.$$

We selected d_{12} and μ as bifurcation parameters for analyzing the signum of $\det(M_0)$ in Fig. 1: it is negative within the red curve and positive outside.

Choosing $\Omega = (0,4) \times (0,4)$ the first two positive eigenvalues of the spectral problem (19) are $\frac{2}{16}\pi^2$ and $\frac{5}{16}\pi^2$. The first mode is destabilized upon increasing d_{12}.

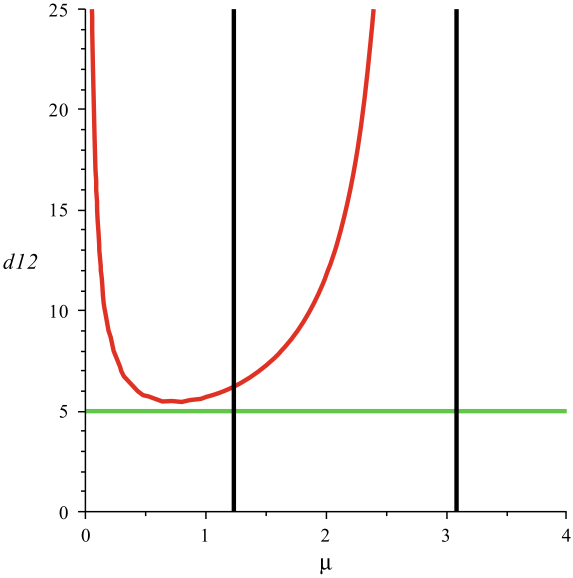


Fig. 1 Numerical illustrations for Example 1. The red curve depicts the level set $\det(M_0) = 0$ as a function of d_{12} and μ, $\det(M_0)$ being negative above this red curve. Black vertical lines represents $\mu = \frac{2}{16}\pi^2$ and $\mu = \frac{5}{16}\pi^2$ the first two positive eigenvalues of the spectral problem (19) in $\Omega = (0,4) \times (0,4)$

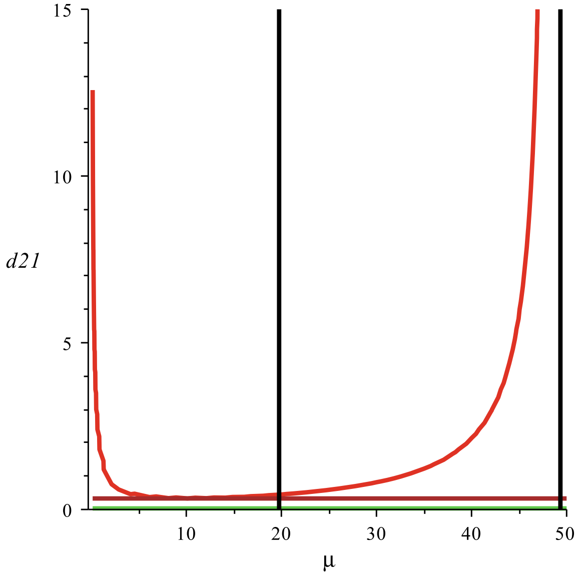


Fig. 2 Numerical illustrations for Example 2. The red curve depicts the level set $\det(M_0) = 0$ as a function of d_{21} and μ, $\det(M_0)$ being negative above this red curve. Black vertical lines represents $\mu = 2\pi^2$ and $\mu = 5\pi^2$ the first two positive eigenvalues of the spectral problem (19) in $\Omega = (0, 1) \times (0, 1)$. The brown horizontal line is $d_{21} = 0.33$

Example 2 The data set is:

$$b = 2, \quad b_I = 0, \quad \theta = 0.3, \quad m = 0.01, \quad \alpha = 0, \quad k = 0.4, \quad \sigma = 1.6,$$
$$d_1 = 0.02, \quad d_2 = 0.01, \quad d_{11} = 0.025, \quad d_{22} = 0.04, \quad d_{12} = 0.01.$$

We selected d_{21} and μ as bifurcation parameters for analyzing the signum of $\det(M_0)$ in Fig. 2: it is negative within the red curve and positive outside.

Choosing $\Omega = (0, 1) \times (0, 1)$ the first two positive eigenvalues of the spectral problem (19) are $2\pi^2$ and $5\pi^2$. The numerical procedure used in the computations is found in [2, 3], based on finite element volumes. The "numerical" initial conditions are $S_0(x) = \frac{b-m}{k}$ and $I_0(x) = 1$ at the bottom left volume of Ω and 0 elsewhere.

The endemic state (S^*, I^*) of the underlying ODE system becomes unstable along the vertical line $\mu = 2\pi^2$ while crossing the level set $\det(M_0) = 0$. As d_{21} increases spatial densities of S and I first reach spatially heterogeneous stable (in time) profiles ($d_{21} = 0.5$ in Fig. 3 and $d_{21} = 5$ in Fig. 4) to become periodic in time ($d_{21} = 100$ in Fig. 5).

6 Conclusion and Perspectives

We introduced a planar ODE system that arises in modeling the transmission of an epidemic disease within a host population or SI model, see (1). From a dynamical point of view the main feature is a threshold parameter, $\mathscr{T}_0$, such that when $\mathscr{T}_0 > 1$

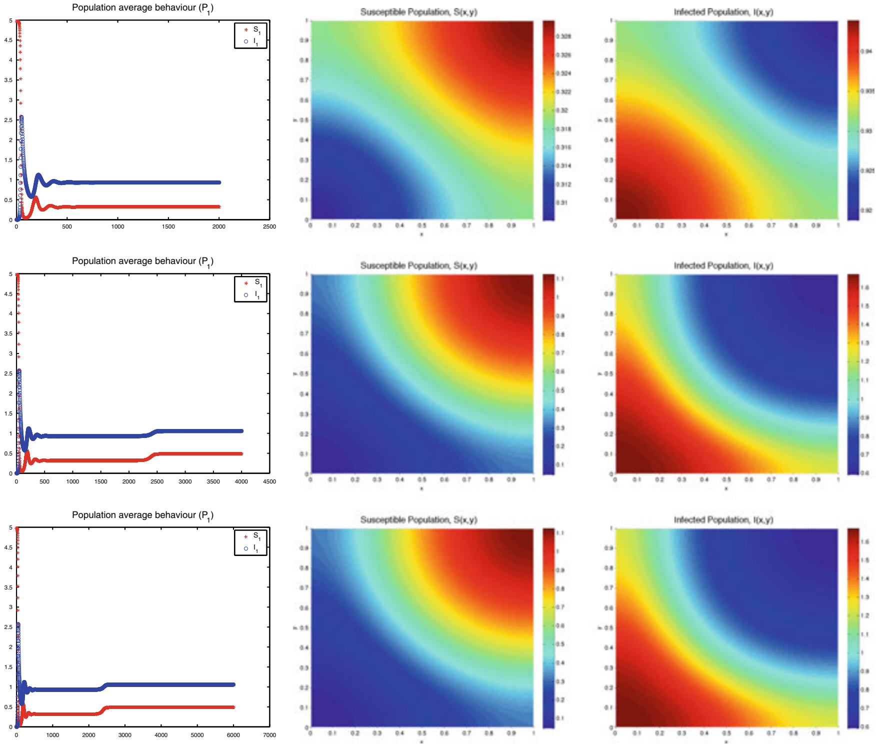


Fig. 3 Numerical illustrations for Example 2 with $d_{21} = 0.5$. The endemic state of the ODE system becomes unstable: spatial average densities (left column), spatial densities of S and I (center and right columns) at times $t = 100$, 200 and 300 or steps 2000, 4000 and 6000

there exists a unique endemic or persistent stationary state that is GAS for solutions emanating from positive initial data and no endemic state when $\mathscr{T}_0 < 1$.

Then we introduced a spatial structure allowing individuals to disperse through their preferred habitat and built several planar Reaction-Diffusion systems featuring linear, nonlinear and cross diffusions, cf. [18, 22, 23]. The goal was to analyze whether the endemic state of the underlying ODE system remains stable for these RD systems. Results in [11, 12] tell us that linear diffusion cannot destabilize this endemic state. This remains true for nonlinear diffusion, see Proposition 1. Then we looked at a prototypical planar system with quadratic nonlinear and cross diffusions. Previous numerical suggested that destabilization of the endemic state and pattern formations occur resulting in a Turing bifurcation.

To support our study, we defined weak solutions according to [2, 3], cf. [19]. This is consistent with our numerical procedures. Some constraints on diffusivities are required to get existence of a weak solution, see (H2)–(H3) or (H23) in Proposition 2—as well as convergence of the approximating scheme, cf. [2, 3].

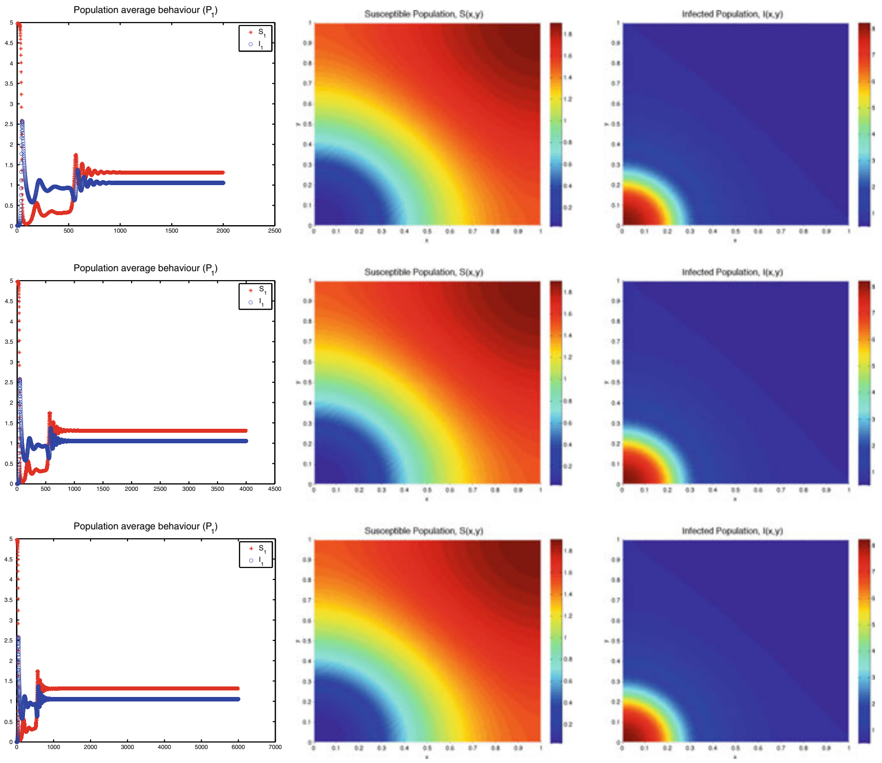


Fig. 4 Numerical illustrations for Example 2 with $d_{21} = 5$. The endemic state of the ODE system becomes unstable: spatial average densities (left column), spatial densities of S and I (center and right columns) at times $t = 100$, 200 and 300 or steps 2000, 4000 and 6000

It turns out from the stability analysis this may not be an optimal approach to defining a weak solution: either weaker solutions without $L^2(\Omega)$ gradient control must be considered, cf. [4], or $L^\infty(\Omega)$ estimates must be a priori established, see Morgan [20, 21] and Pierre [24] or for special cases of the parameter data set. This is beyond the scope of this note. Classical solutions may exist, see Sect. 4.3 and the condition (13), in which case the endemic state remains at least linearly stable, see Proposition 3 and Case 1 in Sect. 5.3 for nonlinear stability.

Then we focused on linear stability / instability of the endemic state when $\mathscr{T}_0 > 1$, following routine procedures, see [22]. Algebraic calculations lead to checking the sign a determinant, $\det(M_0)$, given explicitly in (20) as a function of the endemic state, the parameter data set and the eigenvalue to be destabilized. To contrast with similar problems for Lotka-Volterra like population dynamics model systems involving cross diffusion with algebraically closed form endemic states, see [15, 18] or [25], this sign remains much more complex to find.

In the two cases wherein an algebraically closed form endemic state exists, see Sect. 5.3, one may check that cross diffusivities d_{12} and d_{21} are "pushing" the sign of

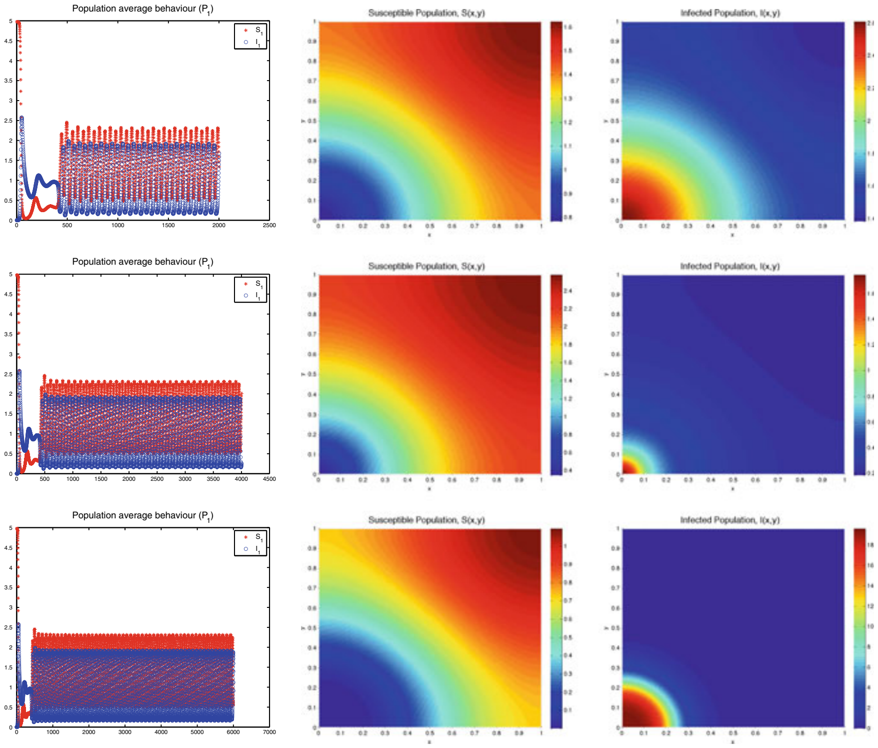


Fig. 5 Numerical illustrations for Example 2 with $d_{21} = 100$. The endemic state of the ODE system becomes unstable: spatial average densities (left column), spatial densities of S and I (center and right columns) at times $t = 100$, 200 and 300 or steps 2000, 4000 and 6000

$\det(M_0)$ either toward the $+$ sign yielding linear stability or in opposite directions. Thus numerical experiments become handy.

In Example 1 we numerically check that upon increasing d_{12} $\det(M_0)$ can pass from positive to negative, see Fig. 1. As a result the first mode of (19) becomes unstable. Other nonlinear and cross diffusivities are set to 0 so that unfortunately (or not) constraints to get a weak solutions are not met!

Example 2 is more comprehensive. In Fig. 2 we numerically check that upon increasing d_{21} $\det(M_0)$ can pass from positive to negative destabilizing the first mode of (19). For d_{21} small enough yielding $\det(M_0) > 0$ the set of constraints to get a weak solutions are met while this is not true anymore upon increasing d_{21} to get $\det(M_0) < 0$!

We supply three data sets numerically exhibiting various patterns observed upon increasing d_{21}. In the first two data sets "stable" heterogeneous spatial profiles are found, see Figs. 3 and 4. In the last one a time periodic heterogeneous spatial profile is achieved, see Fig. 5, suggesting a further bifurcation takes place. Instability occurs after a more or less long transient stage: see the component L^1 norms.

Periodic outbreaks of epidemic diseases is a known phenomenon, cf. [5] for measles epidemics. Spatial heterogeneities in host abundance and patterns of contact between individuals, similar to those used in [8] for BTV, is one out of many explanations for periodicity. Figure 5 suggests that cross diffusion yields an analogous epidemic behavior when increasing d_{21}, that is increasing the weight of susceptible density in the flux of infectious.

One may conclude that existence of weak solutions and bifurcation of the spatially homogeneous persistent state require more analytical and numerical efforts.

References

1. Amann H (1995) Linear and quasilinear parabolic problems, vol I: abstract linear theory. Birkhäuser, Boston, MA
2. Anaya V, Bendahmane M, Langais M, Sepúlveda M (2015) A convergent finite volume method for a model of indirectly transmitted diseases with nonlocal cross-diffusion. Comput Math Appl 70(2):132–157
3. Anaya V, Bendahmane M, Langais M, Sepúlveda M (2015) Pattern formation for a reaction diffusion system with constant and cross diffusion. In: Numerical mathematics and advanced applications—ENUMATH 2013, volume 103 of lecture notes in computational science and engineering. pp 153–161
4. Bendahmane M, Lepoutre T, Marrocco A, Perthame B (2009) Conservative cross diffusions and pattern formation through relaxation. J Math Pures Appl (9) 92(6):651–667
5. Bier M, Brak B (2015) A simple model to quantitatively account for periodic outbreaks of the measles in the Dutch Bible Belt. Eur Phys J B 88(4):107 (11 p)
6. Busenberg S, Cooke K (1993) Vertically transmitted diseases, vol 23. Biomathematics. Springer, Berlin
7. Capasso V (1993) Mathematical structures of epidemic systems, vol 23. Lecture Notes in Biomathematics. Springer, Berlin
8. Charron M, Kluiters G, Langlais M, Seegers H, Baylis M, Ezanno P (2013) Seasonal and spatial heterogeneities in host and vector abundances impact the spatiotemporal spread of bluetongue. Vet Res 44:44
9. Desvillettes L, Lepoutre T, Moussa A (2014) Entropy, duality and cross diffusion. SIAM J Math Anal 46(1):820–853
10. Diekmann O, Heesterbeek H, Britton T (2012) Mathematical tools for understanding infectious disease dynamics, Princeton series in theoretical and computational biology. Princeton University Press, Princeton
11. Ducrot A, Langlais M, Magal P (2012) Qualitative analysis and travelling wave solutions for the SI model with vertical transmission. Commun Pure Appl Anal 11(1):97–113
12. Fitzgibbon WE, Langlais M (2003) A diffusive S.I.S. model describing the propagation of F.I.V. Commun Appl Anal 7(2–3):387–403
13. Fitzgibbon WE, Langlais M (2008) Simple models for the transmission of microparasites between host populations living on noncoincident spatial domains, vol 1936 of lecture notes in mathematics. In: Magal P, Ruan S (eds) Structured Population Models in Biology and Epidemiology, vol 1936. Springer, Berlin, pp 115–164
14. Fitzgibbon WE, Langlais M, Morgan JJ (2004) A reaction-diffusion system on noncoincident spatial domains modeling the circulation of a disease between two host populations. Differ Int Equ 17(7–8):781–802
15. Gambino G, Lombardo MC, Sammartino M (2012) Turing instability and traveling fronts for a nonlinear reaction-diffusion system with cross-diffusion. Math Comput Simul 82(6):1112–1132

16. Goh BS (1978) Global stability in a class of prey-predator models. Bull Math Biol 40(4):525–533
17. Hale JK (1988) Asymptotic behavior of dissipative systems. American Mathematical Society, Providence, RI
18. Iida M, Mimura M, Ninomiya H (2006) Diffusion, cross-diffusion and competitive interaction. J Math Biol 53(4):617–641
19. Ladyzhenskaya OA, Solonnikov VA, Ural'ceva NN (1968) Linear and quasi-linear equations of parabolic type. Translations of mathematical monographs. American Mathematical Society, Providence, RI
20. Morgan J (1989) Global existence for semilinear parabolic systems. SIAM J Math Anal 20(5):1128–1144
21. Morgan J (1990) Boundedness and decay results for reaction-diffusion systems. SIAM J Math Anal 21(5):1172–1189
22. Murray JD (2007/2003) Mathematical biology, vol I: an introduction, vol II: spatial models and biomedical applications. Springer, New York, 3rd edn
23. Okubo A, Levin SA (2002) Diffusion and ecological problems: modern perspectives, 2nd edn. Springer, New York
24. Pierre M (2003) Weak solutions and supersolutions in L^1 for reaction-diffusion systems. J Evol Equ 3(1):153–168 (dedicated to Philippe Bénilan)
25. Tian C, Lin Z, Pedersen M (2010) Instability induced by cross-diffusion in reaction-diffusion systems. Nonlinear Anal Real World Appl 11(2):1036–1045
26. Turing AM (1952) The chemical basis of morphogenesis. Philos Trans Roy Soc Lond Ser B 237(641):37–72
27. Yagi A (2010) Abstract parabolic evolution equations and their applications. Springer monographs in mathematics. Springer, Berlin

Automatic Clustering in Large Sets of Time Series

Robert Azencott, Viktoria Muravina, Rasoul Hekmati, Wei Zhang and Michael Paldino

Abstract To study large sets of interacting time series, we combine spectral analysis of graph Laplacians with simulated annealing to automatically generate optimized clustering of time series, by minimization of cost functions characterizing clustering quality. We apply these techniques to evaluation of connectivity between cortex regions, via analysis of cortex activity recordings by sequences of 3-dimensional fMRI images.

Keywords Time series clustering · Graph Laplacians · Spectral clustering
Mutual information · Simulated annealing · kernel k-means

1 Large Sets of Interacting Time Series

The fast increasing availability of massive data sets has boosted up the use of sophisticated machine learning algorithms and automated clustering techniques to analyze large sets of interactive time series, such as time indexed recordings of brain activity, atmospheric and oceanic evolutions, intraday stock prices, etc. In such contexts, one is confronted to a dynamic system $\mathrm{Sys}(N, T)$ concretely described by N observed time series $X_1(t) \dots X_N(t)$ indexed by discretized times $t = 1, \dots, T$.

For each j the observations $X_j(t)$ can either all be real valued or take values in some fixed finite set. For concrete applications N can range from 100 to 5000, and T can range from a few hundreds to tens of thousands. Interactivity between time series is generally unknown, and dynamic models of such systems are often non existent, or involve very large numbers of unknown parameters to be fitted to the data.

To roughly characterize interactivity without explicit dynamics modelling, one interesting strategy is to define and compute a symmetric matrix S of "affinity" or

R. Azencott (✉) · V. Muravina · R. Hekmati
Department of Mathematics,University of Houston, Houston, TX, USA
e-mail: razencot@math.uh.edu

W. Zhang · M. Paldino
Neuroradiology Section,Texas Children's Hospital, Houston, TX, USA

B. N. Chetverushkin et al. (eds.), *Contributions to Partial Differential Equations and Applications*, Computational Methods in Applied Sciences 47,
https://doi.org/10.1007/978-3-319-78325-3_6

"similarity" coefficient, where $S(i, j) \geq 0$ roughly quantifies the "interaction level" between time series X_i and X_j.

One can then consider the set of "nodes" $G = \{1, \ldots N\}$ as a weighted graph where the edge (i, j) has "weight" $S(i, j) = S(j, i)$. A first natural goal is then to partition the set G into k disjoint clusters $C_1, \ldots, C_k$, where affinities are "high" within each cluster, and "weak" between distinct clusters.

"Spectral Clustering" techniques have been developed to generate "good" clusterings within weighted graphs, and are based on the spectral analysis of the graph Laplacian, as recalled below. These so-called relaxation methods, however, do not fully solve the clustering problem and in this paper we propose to further optimize the results of spectral clustering by stochastic gradient descent, via implementable simulated annealing optimization schemes.

We illustrate our methodology by an ongoing study of fMRI recordings of cortex activity for epileptic children, in collaboration with Michael Paldino (MD), Neuro-Radiology, Texas Children Hospital (TCH).

2 Quantifying Affinities Between Time Series

Given two real valued time series X and Y, one can explicitly eliminate low frequencies by many classical linear filters, such as substraction of a suitable moving average. This type of "detrending" generally improves rough second order stationarity, and one may then characterize "affinity" between X and Y by the absolute value $\text{abscor}(X, Y) \leq 1$ of their correlation after detrending. High values of abscor then indicate approximate affine relations between the detrended X and Y.

Detecting non linear relations between X and Y is better achieved by entropy based mutual information $\text{mut}(X, Y)$. When X and Y take values in two finite sets, $\text{mut}(X, Y)$ is given by

$$\text{mut}(X, Y) = H(X) + H(Y) - H(X, Y)$$

where entropies such as $H(X)$ are computed by

$$H(X) = -\sum p(x) \log p(x)$$

with $p(x) = \text{Pty}(X = x)$.

The relative mutual information

$$\text{rmut}(X, Y) = \text{mut}(X, Y)/min(H(X), H(Y))$$

lies in [0, 1] and $\text{rmut}(X, Y)$ equals 1 iff there is an invertible deterministic between X and Y.

For the more generic case where X, Y are real valued time series, automated analysis of empirical histograms based on T observations $X(t)$ and $Y(t)$ provides finitely discretized approximations $\hat{X}, \hat{Y}$ of X and Y, and one can define an "Entropy Based Affinity" coefficient between X and Y by $\mathrm{rmut}(\hat{X}, \hat{Y})$.

For N real valued time series $X_j(t)$ with $j = 1 \ldots N$ and $t = 1, \ldots, T$ the preceding approaches generate symmetric $N \times N$ affinity matrices $A(i, j) = \mathrm{rmut}(\hat{X}_i, \hat{X}_j)$. Note that A depends on the discretizations implemented on the time series X_j. Natural variants of these techniques can produce other matrices $S(i, j)$ of affinities.

3 Laplacian of a Weighted Graph

Given an observed dynamic system $\mathrm{Sys}(N, T)$ of N time series $X_j(t)$, select and fix an $N \times N$ symmetric matrix S of non negative affinities, as indicated above. Call G the set of N "nodes" $\{1, \ldots, N\}$, and define G as a *weighted graph* by assigning to each edge (i, j) the weight $S(i, j)$. Two nodes i, j are called "connected" if they can be linked by a finite sequence of edges of with positive weights. Define the $N \times N$ diagonal matrix D by $D(j, j) = \sum_{m=1}^{N} S(j, m)$, and call the degree of node j. Assume that all nodes j have positive "degree" $D(j, j)$.

Then all the rows of $Q = D^{-1}S$ have sum 1, and Q is the transition matrix of a Markov Chain $n \to Z_n$ with finite state space G. Let I be the $N \times N$ identity matrix. For any initial line vector $w_0 \in R^N$ of nodes "activity levels" or "energies", stochastic transmission of "activity" via the Markov chain Z_n yields the average evolution $w_n = w_{n-1}Q$, which is classically controlled by the second highest eigenvalue of Q.

The discrete "infinitesimal generator" $\varDelta = I - Q$ of the Markov semi-group $n \to Q^n$ is called the Laplacian operator of the graph G, and is the discrete analogue of the Laplace-Beltrami operator on a Riemannian manifold.

The eigenvalues of $\varDelta$ are the same as those of the so called "normalized Laplacian"

$$L = I - D^{-1/2}SD^{-1/2} = D^{-1/2}\varDelta D^{-1/2}$$

L is always symmetric and semi-positive definite, with eigenvalues $0 \leq \lambda_1 \leq \ldots \leq \lambda_N$. The multiplicity of eigenvalue 0 is the number of connected components in the weighted graph G. The lowest positive eigenvalue of $\varDelta$ (and hence of L) is $(1 - \lambda)$ where λ is the 2nd eigenvalue of the highest eigenvalue of Q, and hence controls the speed of stationarization for the random walk Z_n. We now indicate how the clustering of nodes in G can be linked to the graph Laplacian $\varDelta$.

4 Spectral Clustering and Graph Laplacian

Given the set of nodes G with edges (i, j) endowed with weights or affinities $S(i, j) \geq 0$, one main goal is to find a partition $\mathrm{PAR} = C(1), \ldots, C(k)$ of G into k disjoint clusters. Define the $k \times k$ matrix U of *cluster affinities* by

$$U(m,r) = \sum_{i \in C(m),\ j \in C(r)} S(i,j).$$

Define the "volume" $\operatorname{vol}(m)$ of $C(m)$ and the cost $\operatorname{cut}(m)$ of the "cut" between $C(m)$ and its complement by

$$\operatorname{cut}(m) = \sum_{r \neq m} U(m,r) \quad \text{and} \quad \operatorname{vol}(m) = U(m,m) + \operatorname{cut}(m).$$

Good clusterings should decrease the cost $\operatorname{cut}(m)$ while increasing the self-affinities $U(m,m)$. These requirements can be implemented by minimization of the following cost function

$$\text{COST(PAR)} = \sum_{m=1}^{k} \frac{\operatorname{cut}(m)}{\operatorname{vol}(m)} = \sum_{m=1}^{k} g\left(\frac{\operatorname{cut}(m)}{U(m,m)}\right), \tag{1}$$

where g is the increasing function $g(x) = \frac{x}{1+x}$.

Each cluster C_r defines a column vector f_r in R^N by $f_r(n) = 1/\operatorname{vol}(r)^{1/2}$ for n in C_r and $f_r(n) = 0$ otherwise. The matrix $F = [f_1, \ldots, f_k]$ and the normalized graph Laplacian L then verifies

$$F * DF = I, \quad f_r * Lf_r = \operatorname{cut}(r)/\operatorname{vol}(r) \quad \text{for each } r$$

and hence

$$\text{COST(PAR)} = \operatorname{trace}(F * LF).$$

Minimizing COST(PAR) is then equivalent to finding an $N \times k$ matrix F verifying $F * DF = I$ and minimizing $trace(F * LF)$, under the "combinatorial constraint" that within each column of F, the coefficients take only two values.

If one relaxes the constraints on F by eliminating the combinatorial constraint, then one simply has to minimize $trace(F * LF)$ under the constraint $F * DF = I$. As is easily proved (see [6, 7]), the relaxed solution $\hat{F}$ is given by the k eigenvectors g_r associated to the k lowest eigenvalues of the graph Laplacian $\Delta = I - D^{-1}S$.

Since the coordinates of each g_r take much more than two values, one still needs to associate an actual clustering to $g_1, \ldots, g_k$. A classical approach is to identify each node n with a standard basis vector $e(n)$ in R^N and to compute the projection $v(n)$ of $e(n)$ onto the k-dimensional subspace of R^N generated by the g_r.

One then partitions the cloud of N vectors $v(n) \in R^k$ by the well-known K-means algorithm (see [2]) which generates a partition of R^k into k polyhedral cells. The replacement of K-means by their Hilbert space analog "kernel K-means" often improves the geometric quality of final clusterings, since kernel K-means based on polynomial or Gaussian kernels do partition R^k into k cells separated by hypersurfaces.

However, K-means or kernel K-means algorithms are not specifically designed to minimize the cost function (1). So a terminal partition *par* of G generated by kernel K-means is not necessarily a minimizer of COST(PAR).

We now outline how to improve further the minimization of COST(PAR).

5 Clustering Optimization by Stochastic Descent

Minimizing COST(PAR) over all partitions PAR of G into k clusters is an NP-hard combinatorial problem which can, however, be reasonably attacked by stochastic gradient descent algorithms such as *simulated annealing* (see [1]). Indeed given a current partition PAR $= \{C(1), \ldots, C(k)\}$, one can modify it by one of the following basic moves $\mathrm{MOV}(n, m)$, indexed by $n = 1, \ldots N$ and $m = 1, \ldots, k$ which removes the node n from the cluster $C(j(n))$ to which it currently belongs, and then inserts n into the cluster $C(m)$. Of course if $m = j(n)$ this move does not modify PAR. But for $m \neq j(n)$ the current partition PAR becomes a new partition newPAR where only the two clusters $C(j(n))$ and $C(m)$ have been modified. The change in cost

$$\mathrm{dCOST}(n, m) = \mathrm{COST}(\mathrm{newPAR}) - \mathrm{COST}(\mathrm{PAR})$$

is then easily computed by an elementary formula.

We now indicate how to implement a stochastic descent by simulated annealing in this context.

Select a decreasing sequence of “virtual temperatures” $\mathrm{temp}(s) > 0$ converging slowly to 0 as $s \to \infty$. Ideally one should require $\sum_s \mathrm{temp}(s) = +\infty$. But practical implementations usually select $\mathrm{temp}(s) = a^s$ for some $0 < a < 1$ with a very close to 1, such as $a = 0.99$.

Fix an infinite periodic sequence of nodes $s \to n_s$ visiting successively all nodes in G with periodicity N. Start from any initial partition PAR_0. At each step s, we modify the current partition $\mathrm{PAR}_s = C_s(1), \ldots, C_s(k)$ as follows to generate PAR_{s+1}.

Let $j = j_s$ be the index of the cluster $C_s(j)$ currently containing n_s.

For each m in $1, \ldots, k$, apply the basic move $\mathrm{MOV}(n_s, m)$ to PAR_s to generate a new partition $\mathrm{newPAR}(s, m)$ and compute the associated change in cost $\mathrm{dCOST}(n_s, m)$. Select a cluster index m_s minimizing over m the cost changes $m \to \mathrm{dCOST}(n_s, m)$, and let

$$\mathrm{dc}(s) = \mathrm{dCOST}(n_s, m_s) = \min_{m=1\ldots k} \mathrm{dCOST}(n_s, m).$$

If $\mathrm{dc}(s) < 0$ define $\mathrm{PAR}_{s+1} = \mathrm{newPAR}(s, m_s)$.

If $\mathrm{dc}(s) \geq 0$, pick a random number B equal to 1 or 0 with

$$\mathrm{Pty}(B = 1) = e^{-\,\mathrm{dc}(s)/\,\mathrm{temp}(s)}.$$

Then define $\mathrm{PAR}_{s+1} = \mathrm{newPAR}(m_s)$ if $B = 1$ and $\mathrm{PAR}_{s+1} = \mathrm{PAR}_s$ if $B = 0$.

As has been proved in more generic contexts (see [1]), when the series of temperatures temp(s) diverges, the stochastic sequence of costs $\mathrm{COST}(\mathrm{PAR}_s)$ converge with probability 1 to an absolute minimum of the cost function $\mathrm{COST}(\mathrm{PAR})$ over all partitions PAR of G into k clusters.

6 A Study of Cortex Activity for Epileptic Children

6.1 fMRI Recordings

At TCH NeuroRadiology, an ongoing study led by Dr Michael Paldino (MD) gathers sequences of functional Magnetic Resonance Images in 3D to record cortex activity for selected young epileptic children (see [3–5]). Each full recording is a sequence of fMRI images J_t, with $t = 1, \ldots, T = 295$, acquired at intervals ≈ 2.5 s between J_t and J_{t+1}.

Each 3D image ($\approx$55,000 voxels located on the patient's cortex) is then algorithmically registered at TCH onto a standard pre-segmented cortex atlas, and thus partitioned into 148 anatomically well-identified "cortex regions" $Reg(m)$ with $m = 1, \ldots, 148$, see Figs. 1 and 2.

Each $Reg(m)$ is pre-segmented into smaller "parcels" with surfaces $\approx 150\ mm^2$, yielding a total of roughly $N = 1700$ disjoint cortex parcels CP_i. The average voxel intensity within CP_i at time t is denoted $Y_i(t)$.

The fMRI recording of brain activity for each epileptic patient can thus be viewed as an array $\mathrm{Sys}(N, T)$ of N time series $Y_i(t)$ with $i = 1 \ldots N$ and discrete time $t = 1, \ldots, T$. To evaluate and compare "crude scale" cortex connectivity in a group of 32 young patients, we have begun applying the spectral clustering techniques outlined above to each one of these fMRI recordings.

6.2 Pairwise Mutual Information Between fMRI Time Series

Each recorded time series Y_i oscillates around a mean value M_i which is essentially linked to mean levels of blood irrigation in cortex parcel CP_i. Within CP_i we expect neural activity at time t to be roughly reflected by the normalized oscillations $Z_i(t) = (Y_i(t) - M_i)/M_i$.

So we have first quantified affinity between Y_i and Y_j by the relative mutual information $A(i, j) = \mathrm{rmut}(Z_i, Z_j)$.

We then characterize for each patient P_r with $r = 1, \ldots, 22$ the probability distribution μ_r of the N^2 numbers $A(i, j)$ by its vector of quantiles $q_r =$

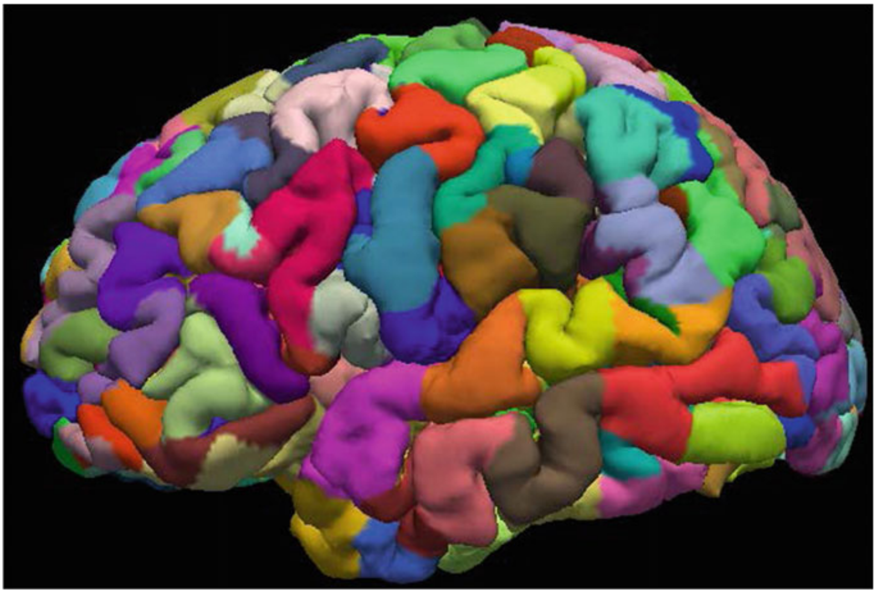


Fig. 1 TCH neuro-radiology, M. Paldino (MD) and collaborators: 3D-image segmentation of cortex into 148 anatomically identified cortex regions

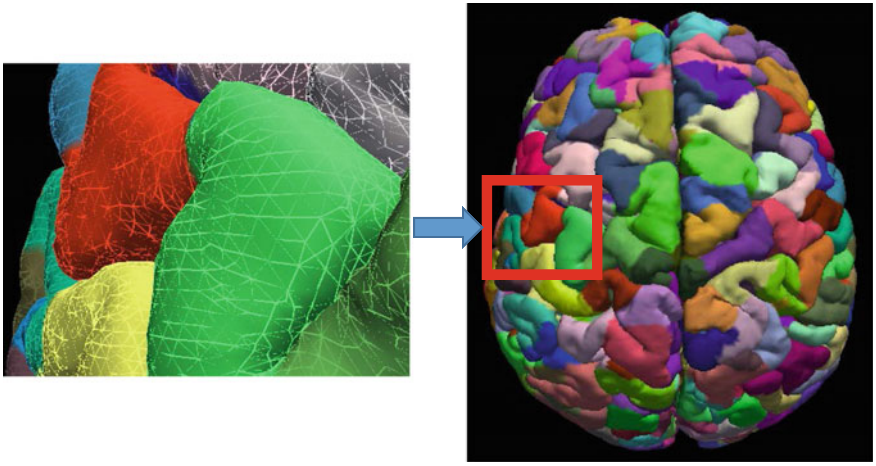


Fig. 2 Voxel level details of 3D-image cortex segmentation

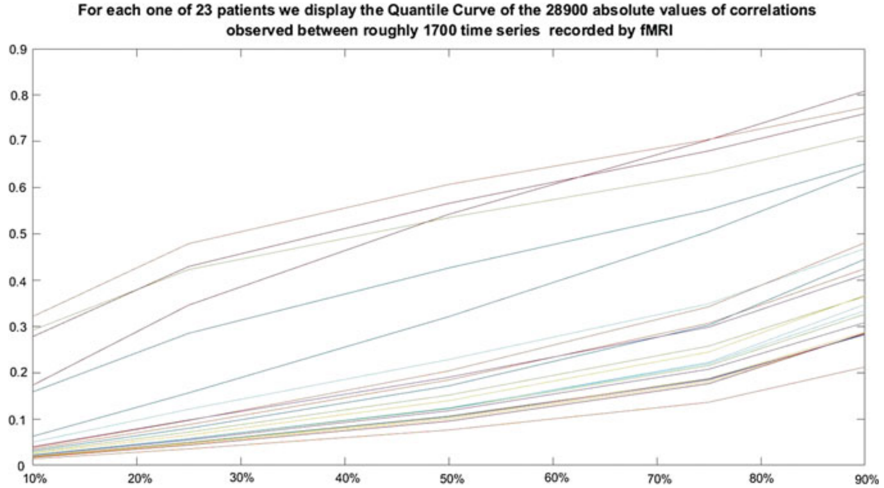


Fig. 3 Quantiles of absolute values of 289,000 pairwise mutual informations between 1700 cortex parcels

$[q(10\%), \ldots, q(90\%)]$, where extreme quantiles are not retained because they are of course less accurately estimated.

We have displayed these quantile curves for several patients in Fig. 3. This essentially ranks patients in terms of overall connectivity, since whenever the quantile curve q_r is above q_s, the $\approx$ 28,900 pairwise affinities of patient P_r are *stochastically larger* than those of patient P_s.

6.3 Interactivities Between Cortex Regions

Since mutual information coefficients with low absolute values have poor statistical significance, we select for each patient P_r a truncation level $q_r(75\%)$ for its affinity matrix A, and we replace A by a truncated version $\hat{A}$ with $\hat{A}(i, j) = A(i, j)$ when $A(i, j) > q_r(75\%)$ and $\hat{A}(i, j) = 0$ otherwise.

The interactivity $S(m, n) \geq 0$ between two cortex regions $Reg(m)$ and $Reg(n)$ is then defined by the sum $S(m, n)$ of all the $\hat{A}(i, j)$ such that $\text{CP}_i \subset Reg(m)$ and $\text{CP}_j \subset Reg(n)$. For patient P_r this defines a symmetric 148×148 matrix S_r of affinities between cortex regions.

6.4 Laplacian Spectrum for Cortex Regions Affinities

For each patient P_r the affinity matrix S_r defines as in Sect. 3 the degree $D_r(m, m) = \sum_{n=1}^{148} S_r(m, n)$ of region m and the transition matrix $Q_r(m, n) = \frac{S_r(m,n)}{D_r(m,m)}$ of a random walk on the weighted graph $G_{\text{reg}} = \{1, \ldots, 148\}$ of cortex regions.

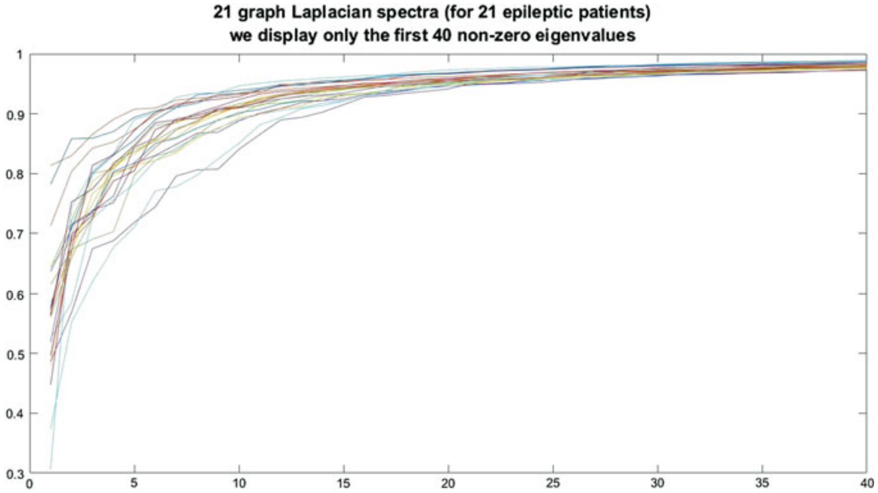


Fig. 4 Spectra of graph Laplacians for cortex regions connectivity graphs

We then compute as indicated above the Laplace operator $\Delta_r = I - Q_r$ of the random walk, and its spectrum $SP_r = (\lambda_j)$ with $j = 1 \ldots 148$, listing the eigenvalues of Δ_r in increasing order, with $\lambda_1 = 0$. Recall that as explained above the important eigenvalues of the Laplacian Δ_r are the lowest ones.

For the 32 patients already studied, the eigenvalues λ_j of Δ_r become very close to 1 as soon as $j >= 13$, as displayed in Fig. 4. The first five positive eigenvalues of the spectra (see Fig. 5) provide a five-dimensional vector of patient features which we will use (among other features) at a later stage to implement quantitative comparisons between patients.

6.5 *Automatic Clustering of Cortex Regions*

We have begun applying the automatic clustering techniques presented above to our group of fMRI recordings, at the level of cortex regions.

The first computational clustering level is currently based on spectral clustering into five clusters after projection on the first 5 eigenvectors of the graph Laplacian, and is actually implementable at rather fast computational speed. Further optimization of this first clustering via simulated annealing is a heavier but still reasonable computational task for each patient.

Here we only illustrate the simpler spectral clustering approach in Fig. 6 where the 148 cortex regions of one patient have been projected on the first three eigenvectors of the graph Laplacian, and then automatically segmented into three clusters of cortex regions.

The neuro-radiology experts in our team will still have to anatomically interpret the optimized clustering results, a time consuming task dedicated to qualitatively link

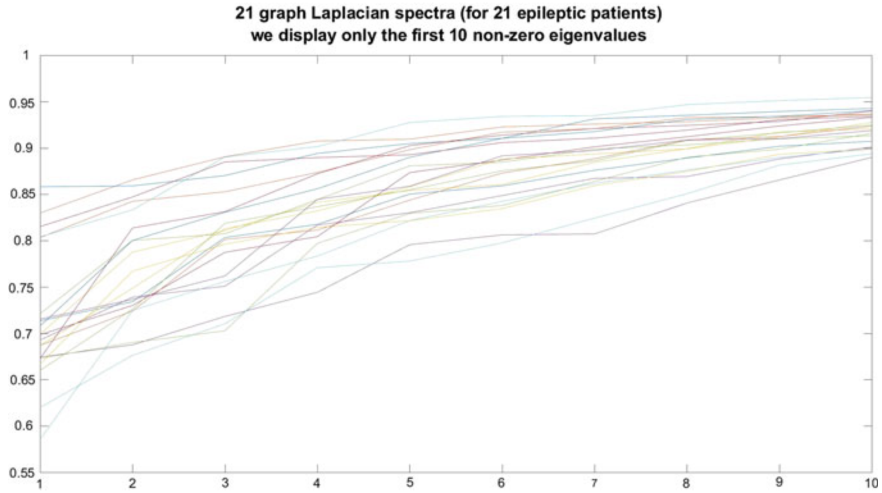


Fig. 5 First 10 lowest eigenvalues of Laplacians for cortex regions connectivity graphs

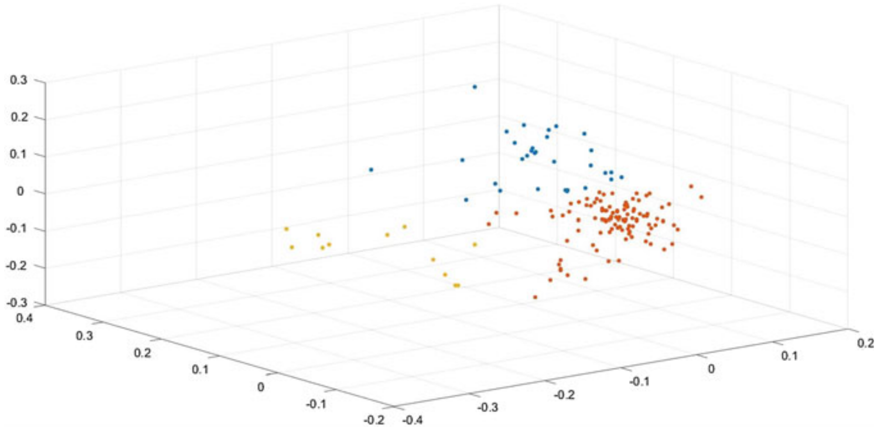


Fig. 6 Partition of 148 cortex regions into three disjoint subsets by spectral clustering. This example is computed by spectral clustering in 3D for cortex regions activities recorded on one patient

cortex regions interactivity to the epileptic focus diagnosed for each patient. Interactive software tools are currently developed to facilitate this medical interpretation task.

7 Meta-similarities Between Weighted Graphs

A natural further goal in the contexts described above is to efficiently define and quantify "meta-similarities" metsim(r, s) between pairs of observed dynamic systems $\text{Sys}_r(N, T)$ and $\text{Sys}_s(N, T)$.

In brain activity recordings by EEG helmets of N sensors, or by fMRI sequences of 3D images with N voxels, the recorded N time series change from subject to subject, or from experiment to experiment. Automated comparisons of brain activity recordings across medium sized groups of subjects or patients P_r can be facilitated by the algorithmic use of adequate meta-similarities.

We propose to formally compare dynamic systems $\text{Sys}_r(N, T)$ with $r = 1...p$ by first associating to each dynamic system a weighted graph G with an affinity matrix S_r of non-negative weights, and then computing the associated normalized graph Laplacian L_r.

Since the L_r belong to the manifold $POS(M)$ of symmetric semi-positive definite matrices of fixed size $M \times M$, which is the closure of a well known Riemannian symmetric space, one can specify mathematically various *computable* positive definite kernels $K(W1, W2)$ defines for all pairs $W1, W2$ in $POS(M)$. The meta-similarity between patients r and s will then be quantified by the scalar product $K(L_r, L_s) = \langle L_r, L_s \rangle_H$, where H is the self-reproducing Hilbert space defined by the kernel K.

The advantage of positive definite kernels to define meta-similarities is that they provide efficient computational tools for automatic non-linear clustering of patients into disjoint patients groups. One can then also use the very efficient Support Vector Machines associated to the kernel K to implement machine learning focused on generating automatic classification of subjects into pre-assigned target groups.

In future work we intend to apply this meta-similarities approach to the data base of fMRI image sequences recorded on epileptic patients at TCH NeuroRadiology.

References

1. Azencott R (ed) (1992) Simulated annealing: parallelization techniques. Wiley, New York
2. Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning: data mining, inference, and prediction, 2nd edn. Springer, New York
3. Paldino MJ, Golriz F, Chapieski ML, Zhang W, Chu ZD (2017) Brain network architecture and global intelligence in children with focal epilepsy. Am J Neuroradiol 38(2):349–356
4. Paldino MJ, Hedges K, Rodrigues KM, Barboriak DP (2014) Repeatability of quantitative metrics derived from MR diffusion tractography in paediatric patients with epilepsy. Br J Radiol 87(1037):20140095
5. Paldino MJ, Hedges K, Zhang W (2014) Independent contribution of individual white matter pathways to language function in pediatric epilepsy patients. NeuroImage Clin 6:327–332
6. Shi J, Malik J (2000) Normalized cuts and image segmentation. IEEE Trans Pattern Anal Mach Intell 22(8):888–905
7. von Luxburg U (2007) A tutorial on spectral clustering. Stat Comput 17(4):395–416

Zero Viscosity Boundary Effect Limit and Turbulence

Claude Bardos

Abstract This contribution is based on a theorem of Kato which relates for time dependent problems the appearance of turbulence with the anomalous energy dissipation, giving for the Cauchy problem an avatar of a basic idea of the statistical theory of turbulence. Some variant of this theorem are given and then it is shown how this point of view is in full agreement with several issues of fluid mechanic ranging from Prandtl's problem to Kutta-Jukowsky's equations.

1 Introduction

After a convenient scaling i.e. put in a-dimensional form the incompressible Navier-Stokes equations

$$\partial_t u_\nu + u_\nu \cdot \nabla u_\nu - \frac{1}{\mathscr{R}} \Delta u_\nu + \nabla p_\nu = 0; \quad \nabla \cdot u_\nu = 0 \ \text{ in } \Omega \times \mathbb{R}_t^+ \tag{1}$$

involve only the parameter $\mathscr{R}$ which is the Reynolds number supposed to be large. Then $\nu = \mathscr{R}^{-1}$ is the rescaled viscosity and the incompressible Euler-equations correspond to the limit case $\mathscr{R} = \infty$ or $\nu = 0$.

Moreover when the problem is considered in a domain $\Omega \neq \mathbb{R}^d$ boundary conditions are prescribed. For $\nu = 0$ the equation (1) become the Euler equations and the corresponding solution (for instance, with the same initial data) is denoted by u. The most natural choice for the boundary condition is the impermeability condition: $u \cdot \mathbf{n} = 0$ where $\mathbf{n}$ denotes the outward normal boundary condition. For the Navier-Stokes equations a larger class of boundary conditions can be given but in the present contribution emphasizes is put on the Dirichlet (or no slip boundary) boundary condition $u_\nu = 0$ on $\partial\Omega$. This is not because it seems to be the simplest but rather because it corresponds to the situation where the obstacle exercises the strongest effect on the

C. Bardos (✉)
Laboratoire Jacques-Louis Lions, Université Pierre et Marie Curie (Paris VI),
France and Wolfgang Pauli Institute Vienna, Paris, Austria
e-mail: claude.bardos@gmail.com

B. N. Chetverushkin et al. (eds.), *Contributions to Partial Differential Equations and Applications*, Computational Methods in Applied Sciences 47,
https://doi.org/10.1007/978-3-319-78325-3_7

behavior of the fluid, in particular for $\nu \to 0$. With such boundary conditions one has (at least for smooth solutions) for u_ν solution of the Navier-Stokes and u solution of the Euler equation, the energy balance equation:

$$\frac{d}{dt}\frac{|u_\nu|^2_{L^2(\Omega)}}{2} + \nu \int_\Omega |\nabla u_\nu|^2 dx = 0 \quad \text{and} \quad \frac{d}{dt}\frac{|u|^2_{L^2(\Omega)}}{2} = 0, \tag{2}$$

With given initial data $u(x,0) = u_0(x) \in L^2(\Omega)$ one deduces that the solutions of the Navier-Stokes equations and of the Euler equation are uniformly bounded in $L^\infty(\mathbb{R}^+_t; L^2(\Omega))$ and that the expression $u \cdot \nabla u$ which appears in both equations is well defined in the sense of distributions according to the relation:

$$\langle u \nabla u, \phi \rangle = -\langle u \otimes u : \nabla \phi \rangle.$$

This seems to indicate that, when ν goes to 0, the solution of the Navier-Stokes converges to the solutions of the Euler (with the same initial data) while the energy dissipation $\nu \int_\Omega |\nabla u_\nu|^2 dx$ goes to zero because at the limit the Euler dynamic is an hamiltonian system which conserves the energy. This is what happens if one considers a problem in the whole space or in a periodic box, assuming that the Euler equation supports a smooth solution (With smooth initial data this fact is well established in dimension 2 and at least for a finite time in dimension 3 (cf. [3] and classical references therein).

On the other hand, in the presence of boundary effects, even with a smooth solution of the Euler equation having the same initial data, the behavior of $\lim_{\nu \to 0} u_\nu$ is a widely open problem. Only the normal component of the velocity on the boundary may remain equal to 0 and since there is no control on the tangential component of the velocity on this boundary a singularity in the behavior of the tangent component of the velocity may appear. Moreover since the Navier-Stokes equations are non linear such singularity may (and will in most physical cases) propagate inside de domain and may generate a turbulent wake.

From the relation (2) one deduces the existence of a weak (modulo extraction of subsequence) limit $\overline{u}$ of solutions of the Navier-Stokes equation equations. However, for such weak limit one may have

$$\epsilon = \lim_{\nu \to 0} \nu \int_0^t \int_\Omega |\nabla u_\nu|^2 dx ds > 0. \tag{3}$$

This implies that if $\overline{u}$ is a weak solution it satisfies the relation $|\overline{u}(t)|^2_{L^2(\Omega)} < |\overline{u}(0)|^2_{L^2(\Omega)}$. In other words, the formula

$$\int_\Omega u \cdot \nabla u \cdot u dx = \int_\Omega \nabla \cdot (u \otimes u) \cdot u dx = 0,$$

which for smooth solutions, with Green formula, follows from the relation $\nabla \cdot u = 0$ and from the boundary condition $u \cdot \mathbf{n} = 0$, may fail to be valid for $\overline{u}$.

As it is well known these issues have been approached by Kolmogorov in the spirit of statistical theory of turbulence where average of flows are considered. According to Kolmogorov (cf. [12, 15], and references therein) turbulent flows are characterized by the relation:

$$\epsilon = \lim_{\nu \to 0} \nu \langle |\nabla u_\nu|^2 \rangle > 0$$

which is the statistical counterpart of the assertion (3). Then under this hypothesis, by a scaling argument, follows the relation:

$$\lim_{\nu \to 0} \langle |u_\nu(x+l,t) - u_\nu(x,t)| \rangle \simeq \epsilon^{\frac{1}{3}} \langle |l|^{\frac{1}{3}} \rangle$$

which indicates for "turbulent" flow an Hölder type regularity of order $\frac{1}{3}$. In particular, for flow more regular one should have:

$$\int_\Omega \nabla \cdot (u \otimes u) \cdot u dx = 0$$

which would hint at the conservation of energy.

The deterministic counter part of this observation was made by Onsager [19]; in 1949 he proved that any weak solution of the Euler equation $u(x,t)$ which satisfies the relation

$$|u(x+l,t) - u(x,t)| \le C|l|^\alpha \quad \text{with } \alpha > \frac{1}{3}$$

conserves the energy. Onsager did not provide a full proof in the sense of mathematical rigor. However, such full proof was later obtained by Constantin and Titi [8] giving rise to a series of refined versions (cf., for instance, [7]). As observed by Shvydkoy [25] all the proofs rely on the same intuitive Onsager argument which goes as follows: In the expression

$$\int_\Omega \nabla(u \otimes u) \cdot u dx = 0 \tag{4}$$

one should distribute the derivative on the three argument leading to a formal expression

$$\int_\Omega \left(\nabla^{\frac{1}{3}} u \otimes \nabla^{\frac{1}{3}} u \otimes \nabla^{\frac{1}{3}} u \right) dx.$$

Then if $\nabla^{\frac{1}{3}} u$ is bounded say in L^p with $3 \le p \le \infty$ it could be approximated by regular functions in such a way that the relation (4) should be obtained by a limit process.

On the other hand, in (1993) and (1997) Scheffer and Shnirleman (cf. [23, 24]) gave examples of functions $u \in L^\infty(\mathbb{R}_t; L^2(\mathbb{R}^2))$ weak solutions (in the sense of distributions):

$$\partial_t u + \nabla(u \otimes u) + \nabla p = 0, \quad \nabla \cdot u = 0$$

of the Euler equations with space and time compact support. Of course such solutions are not physical otherwise they would represent a fluid starting from rest and later returning to rest with no external force. As coined by Villani "The existence of these solutions would solve the energy crisis". More recently starting in 2009 with a series of breathtaking contributions De Lellis and Székelyhidi introduced in the subject the tools of Functional Integration (cf. [11]). Eventually after several other contributions Buckmaster, De Lellis and Székelyhidi (cf. [6] and references therein) proved the existence of weak solutions

$$u \in L^1_t\left(C_x^{\frac{1}{3}-\epsilon}\right)$$

which support energy decay hence almost completing the Onsager conjecture.

These contributions underline the validity of a deterministic approach for problems related to turbulence. With a theorem of T. Kato (1984) [14], recalled and extended in the next section, one shows that it is in presence of boundary effect that in this approach the relation between anomalous energy dissipation and turbulence is the most natural.

It was shown in [4] for some specific examples and conjectured in more general cases that the small viscosity limit be a selection principle to discard the unphysical solutions of the Euler equations. Moreover the importance of these boundary effects was already foreboded by d'Alembert. Working on a 1749 Prize Problem of the Berlin Academy on flow drag, he concluded: "It seems to me that the theory (potential flow), developed in all possible rigor, gives, at least in several cases, a strictly vanishing resistance, a singular paradox which I leave to future Geometers to elucidate" and this became the famous d'Alembert paradox, most probably the origin of the theory of Navier-Stokes equations. Eventually (cf. Sect. 3) one can in a natural way relate the anomalous energy dissipation due to viscosity effects on the boundary with the force exerted on the body (for instance, the lift on a wing).

2 Energy Estimate, Weak Convergence, and the Kato Criteria

In a domain $\Omega \times [0,T] \subset \mathbb{R}^d_x \times \mathbb{R}^+_t$, with $d = 2$ or 3 we assume the existence of smooth solutions both for the Navier-Stokes equations and of the Euler equations with the same initial data:

$$(u_\nu, u) \in C^1(\Omega \times [0,T]); \quad u_\nu(x,0) = u(x,0) = u_0(x); \tag{5a}$$

$$\begin{gathered}\partial_t u_\nu + u_\nu \cdot \nabla u_\nu - \nu \Delta u_\nu + \nabla p_\nu = 0, \quad \nabla \cdot u = 0 \quad \text{in } \Omega \times [0,T] \\ u_\nu = 0 \quad \text{on } \partial\Omega \times [0,T]; \end{gathered} \tag{5b}$$

$$\begin{gathered}\partial_t u + u \cdot \nabla u + \nabla p = 0, \quad \nabla \cdot u = 0 \quad \text{in } \Omega \times [0,T], \\ \mathbf{n} \cdot u = 0 \quad \text{on } \partial\Omega \times [0,T]. \end{gathered} \tag{5c}$$

Hence for the Navier-Stokes equation one has the energy balance relation

$$\begin{gathered}\frac{1}{2}\frac{d}{dt}\int_\Omega |u_\nu(x,\tau)|^2 dx\, d\tau + \nu \int_\Omega |\nabla u_\nu|^2 dx = 0 \quad \text{or} \\ \int_\Omega \frac{|u_\nu(x,t)|^2}{2} dx + \nu \int_0^t \int_\Omega |\nabla u_\nu(x,s)|^2 dx\, ds = \int_\Omega \frac{|u(x,0)|^2}{2} dx \end{gathered} \tag{6}$$

and for the Euler equation the conservation of energy:

$$\forall t \in [0,T] \quad \int_\Omega \frac{|u_\nu(x,t)|^2}{2} dx = \int_\Omega \frac{|u(x,0)|^2}{2} dx.$$

From (6) one deduces that the family u_ν is uniformly bounded in $L^\infty([0,T]; L^2(\Omega))$ and, therefore, modulo extraction of a subsequence it converges in $L^\infty([0,T]; L^2(\Omega))$ weak star to a function $\overline{u}$ satisfying.

$$\partial_t \overline{u} + \nabla_x (\overline{u} \otimes \overline{u}) + \nabla \overline{p} + \nabla_x (\overline{u \otimes u} - (\overline{u} \otimes \overline{u})) = 0$$

with $\overline{u \otimes u}$ denoting the weak limit of the sequence $\overline{u_\nu \otimes u_\nu}$. Here the notion of weak limit plays the role of the ensemble average and

$$\mathscr{R} = \overline{u \otimes u} - \overline{u} \otimes \overline{u}$$

is a non negative tensor which has no reason to be equal to 0 and which is the avatar of the Reynolds stress tensor in the statistical theory of turbulence.

Remark 1 Moreover it was observed in [5] that the vanishing of $\mathscr{R}$ may not be a sufficient condition for the absence of anomalous dissipation.

On the other hand, since u is smooth one can introduce the stress tensor:

$$S(u) = \frac{\nabla u + \nabla^t u}{2}$$

and, from the equations (5b) and (5c), obtain (by substraction, multiplication by $(u - u_\nu)$ and use of the Green formula) the relation:

$$\frac{d}{dt}\frac{1}{2}|u_\nu - u|^2_{L^2(\Omega)} + \nu\int_\Omega |\nabla u_\nu|^2 dx \le |(u_\nu - u, S(u)(u_\nu - u))|$$
$$-\nu\int_\Omega (\nabla u_\nu \cdot \nabla u)dx + \nu\int_{\partial\Omega} \partial_{\mathbf{n}} u_\nu \cdot u d\sigma. \tag{7}$$

The term

$$\nu\int_{\partial\Omega} \partial_{\mathbf{n}} u_\nu \cdot u d\sigma$$

plays a crucial role in the analysis below. In the absence of boundary (in a periodic domain or in the whole space) one deduces that (recalling that we have assumed the existence of a smooth solution of the Euler equation), for $\nu \to 0$ the solution of (5b) converges in $L^\infty([0,T]; L^2(\Omega))$ to the solution of (5c). However, in the presence of no slip boundary effects and in such a general setting the only result is the Kato Theorem which reads as follow:

Theorem 1 *In the presence of a smooth solution of the Euler equation with the same initial data, the following facts are equivalent:*

$$u_\nu(t) \to u(t) \textit{ in } L^2(\Omega) \textit{ uniformly in } t \in [0,T], \tag{8}$$
$$u_\nu(t) \to u(t) \textit{ weakly in } L^2(\Omega) \textit{ for each } t \in [0,T], \tag{9}$$
$$\lim_{\nu\to 0} \nu\int_0^T\int_\Omega |\nabla u_\nu(x,t)|^2 dx\, dt = 0, \tag{10}$$
$$\lim_{\nu\to 0} \nu\int_0^T\int_{\Omega\cap\{d(x,\partial\Omega)<\nu\}} |\nabla u_\nu(x,t)|^2 dx\, dt = 0. \tag{11}$$

and eventually the fact that for all tangent to the boundary vector field $w(x,t) \in \mathscr{D}((0,T)\times\partial\Omega)$ one has:

$$\lim_{\nu\to 0} \nu\int_0^T\int_{\partial\Omega} \frac{\partial u_\nu}{\partial\mathbf{n}}(\sigma,t)w(\sigma,t)d\sigma\, dt = 0. \tag{12}$$

The implications (8) ⇒ (9) ⇒ (10) are direct consequences of the energy dissipation for the Navier-Stokes equation and energy conservation for smooth solutions of the Euler equation. The fact that (11) implies (8) is the essential contribution of Kato [14]. It is done with a well adapted construction of a boundary layer corrector.

The fact that (11) ⇒ (12) was done in [3]. There, following the construction of Kato one introduces a convenient family w_ν of extensions in Ω of any given smooth vector field $w(\sigma,t)$ tangent to $\partial\Omega$. These $w_\nu(x,t)$ are supported in the region

$$\{(x,t)\} \subset \{x \in \Omega \mid d(x,\partial\Omega) < \nu\} \times]0,T[\,.$$

Then one multiplies the Navier-Stokes equations by w_ν and from the relation

$$\begin{aligned}&\nu\int_{\partial\Omega}\frac{\partial u_\nu}{\partial\mathbf{n}}(\sigma,t)w(\sigma,t)d\sigma\\ =&\nu(\nabla u_\nu,\nabla w_\nu)_{L^2(\Omega)}-(u_\nu\otimes u_\nu,\nabla w_\nu)_{L^2(\Omega)}+(\partial_t u_\nu,w_\nu)_{L^2(\Omega)}\end{aligned}$$

one deduces (12). Eventually (12) ⇒ (8) follows from (7).

2.1 Equivalent Form of the Kato Criteria

Cases where the Kato criteria do not apply seem to be, as discussed below, the general situation rather than the exception. They correspond to real or numerical observations. They are the most common way of generating turbulence even homogenous isotropic turbulence, for instance when a grid is used to generate such turbulence (cf. [12] Fig 1.11 page 9). On the other hand, the appearance of such situations depends on many unrelated parameters. Therefore it is interesting, (recalling that we assume for the same initial data the existence in $\Omega\times[0,T]$ of a smooth solution $u(x,t)$ of the Euler Equation), to quote several other fully equivalent criteria which where recently derived and which relate to the fact that at the limit $\nu\to 0$ the energy dissipation does not go to 0 with the non convergence u_ν to this smooth solution. The relation

$$\liminf_{\nu\to 0}\nu\int_0^T\int_\Omega|\nabla u_\nu(x,t)|^2dx\,dt=\epsilon>0 \tag{13}$$

implies among others the following facts.

In [5] it is proven that (13) implies the existence of at least one point

$$p_{\text{turb}}=(x_{\text{turb}},t_{\text{turb}})\in\partial\Omega\times[0,T]$$

such that for any neighborhood U of this point and any n one has:

$$\sup_{\nu\to 0}\|u_\nu\|_{C^{0,\frac{1}{n}}(U)}=\infty.$$

On the other hand, this condition can be even refined following [9]. There, in $2d$, one shows that the existence for some $\nu_0>0$ of a constant C such that (with u^τ_ν and u^n_ν denoting the tangent and the normal component of the flow near $\partial\Omega$)

$$\int_0^T\int_{\Omega\cap d(x,\partial\Omega)<\nu_0}|u^\tau_\nu(x,t)u^n_\nu(x,t)|dx\,dt\le C\nu$$

rules out the anomalous dissipation of energy (i.e. in (13) $\epsilon=0$).

Also in $2d$ cf. [10] one has no anomalous energy dissipation if:

$$\lim_{\nu\to 0} \nu \int_0^T \int_{\partial\Omega} |\inf(0, \nabla \wedge u_\nu)(x,t)| d\sigma \, dt = 0$$

Eventually in [9] other criteria and corresponding references are given.

Remark 2 As it is well known most of the solutions of time dependent compressible Euler equation, even starting from smooth initial data, become singular after a finite time (formation of shocks). However, it was observed in [2] that most of the above results can be adapted to the compressible case as long as the solution of the compressible Euler equations, with the same initial data remains smooth.

2.2 Comparison with the Prandtl Equations and the Triple Deck Ansatz

In (1904) Prandtl [21] proposed to represent the solution near the boundary by a parabolic boundary layer

$$u_\nu \simeq U_\tau\left(\frac{d(x,\partial\Omega)}{\sqrt{\nu}}, x_\tau, t\right) + \sqrt{\nu} U_{\mathbf{n}}\left(\frac{d(x,\partial\Omega)}{\sqrt{\nu}}, x_\tau, t\right),$$

where $d(x, \partial\Omega)$ denotes the distance to the boundary while the indices τ and $\mathbf{n}$ refer to tangent and normal components of the space variable x and of the fluid velocity $U(x, t)$.

Inserting this ansatz in the Navier-Stokes equation and discarding terms of order ν Prandtl obtained the equations that carry his name. Such system has been used with success over the years to compute the air around an airfoil but only in region with no recirculation. Since, in general, the flow does not remain parallel to the boundary one expect the appearance of a series of pathologies which would be first related to phenomena of detachment and recirculation. This justify the reason for the Prandtl equations to be an ill posed problem and this has already been observed long time ago. In the most recent contribution [13] one finds a convenient list of references and a series of sufficient condition (which prevent the appearance of detachment) and insure the well posed-ness of the system. In particular, as observed by Asano [1] and Caflisch and Sammartino [22], with analytic initial data and during a short time (a very non physical and stringent condition) the solution of the Prandtl equations exists and provides a good approximation of the viscous solution.

Remark 3 The following remark seems important: If for some time interval the Prandtl equation have a smooth solution and provide an approximation for $\nu \to 0$ of the solution of the Navier-Stokes equations then the Kato criteria is satisfied. This is not surprising. Otherwise that would lead to a contradiction because the Kato criteria involves a boundary layer of size $O(\nu)$ while the Prandtl ansazt contains oscillations of order $\sqrt{\nu}$. Eventually this follows from the estimate:

$$\begin{aligned}&\nu\int_0^T\int_{\Omega\cap d(x,\partial\Omega)<\nu}|\nabla u_\nu(x,t)|^2dx\,dt\\ &\simeq\nu\int_0^T\int_{\Omega\cap d(x,\partial\Omega)<\nu}\left|\nabla\left(U_\tau\left(\frac{d(x,\partial\Omega)}{\sqrt{\nu}},x_\tau,t\right)+\sqrt{\nu}U_{\mathbf{n}}\left(\frac{d(x,\partial\Omega)}{\sqrt{\nu}},x_\tau,t\right)\right)\right|^2dx\,dt\\ &\qquad\le C\int_0^T\int_{\Omega\cap d(x,\partial\Omega)<\nu}dx\,dt\to 0.\end{aligned}$$

However, the fact that the existence of a smooth solution (whenever it exists) provides a uniform (for $\nu\to 0$) approximation of the Navier-Stokes equation, is except in the case considered in [1, 22], and in full generality, an open problem.

On the other hand, Kato criteria may hold in cases where the validity of Prandtl ansatz fails because such ansatz requires the absence of recirculation in a layer of size $O(\sqrt{\nu})$ while the Kato criteria requires only some regularity in a much smaller region of the order of ν.

In fact, some ansatz that would allow recirculation in a region of order ν have been proposed. The most classical being the triple deck ansatz proposed by Stewartson in 1974, cf. [26], which introduces three layers according to the formula:

1. In the Upper Deck $\{x\backslash\sqrt{\nu}<d(x,\partial\Omega)\}\cap\Omega$ the solution is described by the Euler flow.
2. In the Lower Deck $\{x\backslash 0<d(x,\partial\Omega)<\nu^{\frac{5}{8}}\}\cap\Omega$ the solution is described by the above Prandtl boundary layer ansatz.
3. In the Middle Deck $\{\nu^{\frac{5}{8}}<d(x,\partial\Omega)<\sqrt{\nu}\}\cap\Omega$ which connects the two above regions the following scaling is proposed.

$$u_\nu(x,t)\simeq(\nu^{\frac{1}{8}}U_\tau(\frac{d(x,\partial\Omega)}{\nu^{\frac{5}{8}}},\frac{x_\tau}{\nu^{\frac{3}{8}}},t),\nu^{\frac{3}{8}}U_{\mathbf{n}}(\frac{d(x,\partial\Omega)}{\nu^{\frac{5}{8}}},\frac{x_\tau}{\nu^{\frac{3}{8}}},t)).\tag{14}$$

Once again one observes that if the above formulas provide in the region $0\le d(x,\partial\Omega)\le\sqrt{\nu}$ an uniform approximation of the solution of the Navier-Stokes equation with no slip boundary condition then, the Kato criteria is satisfied.

2.3 Kato Criteria and Turbulent Layer

As already said above cases where the energy dissipation does not vanish with $\nu\to 0$ and hence where the Kato criteria is not satisfied seem to be rather the general cases than the exception. And in such general case turbulence generation in a region of size ν is the basic cause of the phenomena. In spite of the absence of any type of proof I think that it may be useful to compare this issue with the rule for turbulence at the boundary. Since convergence in a strong norm is not expected, a turbulent boundary layer for $\overline{u_\nu}$ should be present, in general, around some part of the boundary.

To the best of my knowledge, the only practical thing available is a description based on experiment, numerical analysis and dimension analysis, the Von Karman-Prandtl turbulent layer (1932). It provides an ansatz for the tangential component of the velocity $u_\tau(x_{\mathbf{n}}, x_\tau, t)$ in the layer

$$B_{\text{turbulent}} = \{x, d(x, \Omega) < \nu\} \cap \mathscr{W}$$

with $\mathscr{W}$ denoting a neighborhood of a part of the boundary.

On $\partial\Omega \cap \overline{\mathscr{W}}$ the quantity

$$u^* = \sqrt{\nu \partial_{\mathbf{n}} u_\tau} \tag{15}$$

which has the dimension of a velocity, is assumed to be of the order of unity.

Then in $B_{\text{turbulent}}$ one has:

$$u_\tau(x_{\mathbf{n}}, x_\tau) = u^* U_\tau(s), \quad s = u^* \frac{x_{\mathbf{n}}}{\nu} \tag{16}$$

with $U_\tau(s)$ an intrinsic function of the “number” s. With phenomenological argument this function is almost linear for $0 < s < 20$ and given by a Prandtl-Von Karman wall law

$$U_\tau(s) = \kappa \log s + \beta \quad \text{for} \quad 20 < s < 100. \tag{17}$$

However, either with (15) which implies that

$$\nu \partial_{\mathbf{n}} (u_\tau)_{|\partial\Omega} \geq \alpha > 0$$

or with (16) which implies

$$\nu \int_{\{x \in \Omega \backslash d(x, \partial\Omega) < \nu^{\frac{1}{2}}\}} |\nabla u_\nu(x, t)|^2 dx \geq \epsilon > 0 \tag{18}$$

one observes that the existence of such boundary layer is consistent with the fact that $\overline{u_\nu}$ does not converge to the smooth solution u and (for instance) is not in $C^{0,\alpha}$ (for any α) uniformly with respect to ν in some neighborhood of a part of the boundary which is necessary for the appearance of a turbulent wake.

3 The d’Alembert Paradox, the Wake Behind an Obstacle, and the Kutta-Jukowski Condition

In this section the case where $\Omega \subset \mathbb{R}^d$ with $d = 2, 3$ is the complement of a bounded obstacle (or body) is considered and the fluid $u \in L^\infty(\Omega \times \mathbb{R}_t)$ is assumed to satisfy:

$$\nabla\cdot u=0 \quad\text{and}\quad \nabla\wedge u=0 \quad\text{in }\Omega;\qquad u\cdot\mathbf{n}=0 \quad\text{on }\partial\Omega,$$
$$\lim_{|x|\to\infty} u(x)=\mathbf{u}_\infty=(0,u_\infty)\quad\text{for } d=2 \quad\text{or}\quad \mathbf{u}_\infty=(0,0,u_\infty)\quad\text{for } d=3$$

then u is a potential flow. Moreover, with the relation:

$$u\cdot\nabla u=\sum_{1\le j\le d} u_j\partial_{x_j}u_i=u\cdot\nabla u=\sum_{1\le j\le d} u_j\partial_{x_j}u_j=\frac{1}{2}\nabla_x|u|^2$$

u is a stationary solution and for any $R>0$ with the Green formula one has

$$\begin{aligned}F&=\int_{\partial\Omega} p\mathbf{n}d\sigma=\int_{\partial\Omega}(p\mathbf{n}+(\mathbf{n}\cdot u)u)d\sigma\\&=\int_{\Omega\cap\{|x|<R\}}(\nabla p+u\cdot\nabla u)dx-\int_{|x|=R}\left(\frac{\mathbf{x}}{|x|}p+\left(\frac{\mathbf{x}}{|x|}\cdot u\right)u\right)d\sigma\\&=-\int_{|x|=R}\left(\frac{\mathbf{x}}{|x|}p+\left(\frac{\mathbf{x}}{|x|}\cdot u\right)u\right)d\sigma.\end{aligned}\tag{19}$$

Letting $R\to\infty$ and using the properties of bounded harmonic functions one deduces from (19) the two following theorems:

Theorem 2 (D'Alembert paradox) *In* 3 *space variables, a potential flow with bounded velocity produces no force on the obstacle.*

Theorem 3 (Kutta-Joukowski theorem) *In two space variables the lift F_+ produced by a potential flow with constant horizontal velocity at ∞*

$$\lim_{|x|\to\infty} u(x)=(u_\infty>0,0)$$

is given, in term of the circulation of the fluid around the obstacle by the Kutta-Joukowsky formula:

$$F_+=-u_\infty\int_{\partial\Omega} u\wedge\mathbf{n}d\sigma.$$

The proofs of these classical theorems can be found in [15] and for a more modern presentation in [17]. As an illustration of this discussion one observes that the d'Alembert paradox can be lifted if instead of considering the above potential flow one considers the limit for $\nu\to 0$ of a solution of Navier-Stokes equation with no slip boundary condition. In general, this solution may not be (and will not be in most cases) a regular solution of the Euler equation. Therefore the Kato criteria shows how the appearance of a force on the body is related to the anomalous energy dissipation. This may contribute to understand how this energy dissipation through viscosity effect creates forces and in particular lift on the body.

Behind an obstacle the physical solution exhibits a wake. When the cross section of the obstacle is very elongated, with a width (or span) large in comparison with the other dimensions and in particular with a sharp trailing edge the thickness of the wake is very small. In such cases one shows (cf. [15][Chapter III, § 37]) that the Kutta-Jukowski relation remains valid (both in dimension 2 and with a natural extension in dimension 3). The ultimate approximation consists in identifying this wake with a line in 2 space variables and a plane in 3 spaces variables, called the Trefftz line or Trefftz plane Γ.

In the complementary of this line (resp. plane) the fluid is irrotational hence potential $u = \nabla\phi$. On Γ with the relation $\nabla \cdot u = 0$ the normal component of the velocity is continuous while the tangential component has a jump leading to formulas:

$$[\nabla\phi] \cdot \mathbf{n} = 0, \quad \nabla \wedge u = [u_\tau] \otimes \delta_{\Gamma(t)} = [\nabla\phi] \otimes \delta_{\Gamma(t)}, \tag{20a}$$

$$\partial_t u + \nabla u \otimes u + \nabla p = 0, \quad \nabla \cdot u = 0. \tag{20b}$$

In the article of Périaux [20] there are several examples showing that such ansazt is well adapted to finite elements discretization. This confirms also the pertinence of the model.

However, one should keep in mind that even if the Trefftz plane is a good model for computations, it does not provides a description of what happens in the wake. The surface $(\Gamma(t), t) \subset \mathbb{R}^d_x \times \mathbb{R}^+_t$ do not exist, in general. The reason is that solutions of (20) with a C^2 surface leads to an ill posed problems. This has been already observed by Kelvin and Helmholtz and therefore such instability carry the name of Kelvin Helmholtz. In the last century this was also the object of many contributions. Basically it was shown that if a solution of (20) exists with a vorticity supported by a C^2 surface then in fact the surface is analytic (cf. [16, 27]). A contrario that means that what happens in the wake cannot, in general, be described by a "smooth surface" But once again that does not prevent the models like the Trefftz plane to be efficient for computations.

4 Conclusion

This contribution contains both some reviews of already published results and some new considerations. It is based on a theorem of T. Kato which concerns the zero viscosity limit of solutions of Navier-Stokes equations with no slip boundary condition. The first observation was that this is the simplest and may the only (to the best of my knowledge) case where there is a natural explicit relation between the anomalous dissipation of energy and the appearance of turbulence.

One drawback of this presentation is the fact that it is based on a finite time evolution (this theory of partial differential equation carries the name of Cauchy problem) while most practical problems concern stationary (or stationary in the average) regimes. However, the similarity between these two regimes (even in the absence of complete proofs) is striking enough to be underlined and eventually rules derived from the stationary regimes are applied in time dependent contributions. A good example being the Prandtl-Von Karman law. As explained above it is compatible with the appearance of turbulence and is used in simulations as explained in Mohammadi and Pironneau [18, p. 15].

Acknowledgements Warm thanks to the organizers of the International Workshop on Applied and Computational Mathematics, Houston, February 2016 for their invitation to give a talk and to propose an article in the present volume, giving me an opportunity to congratulate Olivier Pironneau for his 70th birthday and to express my gratitude for his everlasting friendship and support.

References

1. Asano A (1991) Zero-viscosity limit of the incompressible Navier-Stokes equations. In: Fourth workshop on mathematical aspects of fluid and plasma dynamics. Kyoto
2. Bardos C, Nguyen TN (2016) Remarks on the inviscid limit for the compressible flows. In: Radulescu D, Sequeira A, Solonnikov VA (eds) Recent advances in partial differential equations and applications, volume 666 of Contemp. Math. Providence, RI. AMS, pp 55–67
3. Bardos C, Titi ES (2013) Mathematics and turbulence: where do we stand? J Turbul 14(3):42–76
4. Bardos C, Titi ES, Wiedemann E (2012) The vanishing viscosity as a selection principle for the Euler equations: the case of 3D shear flow. CR Math Acad Sci Paris 350(15–16):757–760
5. Bardos K Jr, Sekelikhidi L, Videmann E (2014) On the absence of uniqueness for the Euler equations: the effect of the boundary. Uspekhi Mat Nauk 69(2(416)):3–22. Translation in Russian Math Surv 69(2):189–207
6. Buckmaster T, De Lellis C, Székelyhidi L Jr (2016) Dissipative Euler flows with Onsager-critical spatial regularity
7. Cheskidov A, Constantin P, Friedlander S, Shvydkoy R (2008) Energy conservation and Onsager's conjecture for the Euler equations. Nonlinearity 21(6):1233–1252
8. Constantin P, Weinan E, Titi ES (1994) Onsager's conjecture on the energy conservation for solutions of Euler's equation. Comm Math Phys 165(1):207–209
9. Constantin P, Elgindi T, Ignatova M, Vicol V (2017) Remarks on the inviscid limit for the Navier-Stokes equations for uniformly bounded velocity fields. SIAM J Math Anal 49(3):1932–1946
10. Constantin P, Kukavica I, Vicol V (2015) On the inviscid limit of the Navier-Stokes equations. Proc Amer Math Soc 143(7):3075–3090
11. De Lellis C, Székelyhidi L Jr (2009) The Euler equations as a differential inclusion. Ann Math (2) 170(3):1417–1436
12. Frisch U (1995) Turbulence. Cambridge University Press, Cambridge
13. Gerard-Varet D, Masmoudi N (2015) Well-posedness for the Prandtl system without analyticity or monotonicity. Ann Sci Éc Norm Supér (4) 48(6):1273–1325, 2015
14. Kato T (1984) Remarks on zero viscosity limit for nonstationary Navier-Stokes flows with boundary. In: Seminar on nonlinear partial differential equations (Berkeley, CA, 1983), volume 2 of Math Sci Res Inst Publ. Springer, New York
15. Landau LD, Lifshitz EM (1959) Fluid mechanics. Pergamon Press, London
16. Lebeau G (2002) Régularité du problème de Kelvin-Helmholtz pour l'équation d'Euler 2d. ESAIM Control Optim Calc Var 8:801–825
17. Marchioro C, Pulvirenti M (1994) Mathematical theory of incompressible nonviscous fluids, vol 96 of Applied Mathematical Sciences. Springer, New York
18. Mohammadi B, Pironneau O (1994) Analysis of the k-epsilon turbulence model. Wiley, Chichester
19. Onsager L (1949) Statistical hydrodynamics. Nuovo Cimento (9) 6(2):279–287
20. Periaux J (1975) Three dimensional analysis of compressible potential flows with the finite element method. Int J Numer Methods Eng 9(4):775–831

21. Prandtl L (1904) Uber flüssigkeits-bewegung bei sehr kleiner reibung. Actes du 3ème Congrès international des Mathématiciens (Heidelberg). Teubner, Leipzig, pp 484–491
22. Sammartino M, Caflisch RE (1998) Zero viscosity limit for analytic solutions, of the Navier-Stokes equation on a half-space. I. Existence for Euler and Prandtl equations. Comm Math Phys 192(2):433–461
23. Scheffer V (1993) An inviscid flow with compact support in space-time. J Geom Anal 3(4):343–401
24. Shnirelman A (1997) On the nonuniqueness of weak solution of the Euler equation. Comm Pure Appl Math 50(12):1261–1286
25. Shvydkoy R Private communication
26. Stewartson K (1974) Multistructured boundary layers on flat plates and related bodies. Adv Appl Mech 14:145–239
27. Wu S (2002) Recent progress in mathematical analysis of vortex sheets. In: Proceedings of the international congress of mathematicians, vol III. Beijing, Higher Ed. Press, pp 233–242

Parabolic Equations with Quadratic Growth in $\mathbb{R}^n$

Alain Bensoussan, Jens Frehse, Shige Peng and Sheung Chi Phillip Yam

Abstract We study here quasi-linear parabolic equations with quadratic growth in $\mathbb{R}^n$. These parabolic equations are at the core of the theory of PDE; see Friedman (Partial differential equations of parabolic type. Prentice-Hall, Englewood Cliffs, 1964) [6], Ladyzhenskaya et al. (Translations of Mathematical Monographs. AMS, 1968) [4] for details. However, for the applications to physics and mechanics, one deals mostly with boundary value problems. The boundary is often taken to be bounded and the solution is bounded. This brings an important simplification. On the other hand, stochastic control theory leads mostly to problems in $\mathbb{R}^n$. Moreover, the functions are unbounded and the Hamiltonian may have quadratic growth. There may be conflicts which prevent solutions to exist. In stochastic control theory, a very important development deals with BSDE (Backward Stochastic Differential Equations). There is a huge interaction with parabolic PDE in $\mathbb{R}^n$. This is why, although we do not deal with BSDE in this paper, we use many ideas from Briand and Hu (Probab Theory Relat Fields 141(3–4):543–567, 2008) [1], Da Lio and Ley (SIAM J Control Optim 45(1):74–106, 2006) [2], Karoui et al. (Backward stochastic differential equations and applications, Princeton BSDE Lecture Notes, 2009) [3], Kobylanski (Ann Probab 28(2):558–602, 2000) [5]. Our presentation provided here is slightly innovative.

A. Bensoussan (✉)
International Center for Decision and Risk Analysis, Jindal School of Management, University of Texas at Dallas, Richardson, USA
e-mail: axb046100@utdallas.edu

A. Bensoussan
College of Science and Engineering, Systems Engineering and Engineering Management, City University Hong Kong, Kowloon Tong, Hong Kong

J. Frehse
Insitute for Applied Mathematics, University of Bonn, Bonn, Germany

S. Peng
Shandong University, Jinan, People's Republic of China

S. C. P. Yam
Department of Statistics, The Chinese University of Hong Kong, Sha Tin, Hong Kong

B. N. Chetverushkin et al. (eds.), *Contributions to Partial Differential Equations and Applications*, Computational Methods in Applied Sciences 47,
https://doi.org/10.1007/978-3-319-78325-3_8

1 Presentation of the Problem

1.1 General Framework

We are considering the following problem:

$$\begin{cases} -\dfrac{\partial u}{\partial t} - \dfrac{1}{2}\triangle u = H(x,u,Du), \\ \qquad u(x,T) = h(x). \end{cases} \tag{1}$$

We have taken $-\frac{1}{2}\triangle u$ instead of a general coercive operator A, to simplify calculations. The argument, space variable, is $x \in \mathbb{R}^n$. The Hamiltonian $H(x,y,z)$ is a measurable function on $\mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^n$ such that

$$y, z \to H(x,y,z) \text{ is continuous, for each } x \in \mathbb{R}^n, \tag{2}$$

$$|H(x,y,z)| \leq \lambda(x) + k|y| + \frac{\gamma}{2}|z|^2. \tag{3}$$

The functions $\lambda(x)$ and $|h(x)|$ are not bounded, but cannot have a growth more than quadratic, with limited growth rates, depending on the size of T. If T is sufficiently small, general quadratic growth is possible. We shall make this assumption more precise in the next section. The key difficulty is that there is a conflict between the growth of the functions $\lambda(x)$, $|h(x)|$ and the quadratic growth of the Hamiltonian in z. When the Hamiltonian has a growth lower than quadratic, things become much easier. We shall prove existence of a solution of (1), in a convenient functional space to be introduced. More conditions are required to obtain uniqueness.

1.2 Discussion on Growth

To understand the issue of growth, we apply the inequality (3) to Eq. (1) to obtain:

$$-\frac{\partial u}{\partial t} - \frac{1}{2}\triangle u \leq \lambda(x) + k|u| + \frac{\gamma}{2}|Du|^2.$$

This leads to introducing the PDE:

$$\begin{cases} -\dfrac{\partial \bar{u}}{\partial t} - \dfrac{1}{2}\triangle \bar{u} - k\bar{u} - \dfrac{\gamma}{2}|D\bar{u}|^2 = \lambda(x), \\ \qquad\qquad \bar{u}(x,T) = |h(x)|, \end{cases} \tag{4}$$

and we expect that

$$|u(x,t)| \leq \bar{u}(x,t).$$

The issue is that the PDE (4) may fail to have a solution. The growth conditions guarantee the solvability of this equation. In fact, by simple calculations the function $\bar{v}(x,t) := \exp(\gamma \bar{u}(x,t) \exp(kt))$ satisfies the inequality

$$\begin{cases} -\dfrac{\partial \bar{v}}{\partial t} - \dfrac{1}{2}\triangle \bar{v} \leq \gamma \lambda(x) \exp(kt)\, \bar{v}, \\ \qquad \bar{v}(x,T) = \exp(\gamma |h(x)| \exp(kT)). \end{cases} \tag{5}$$

This introduces naturally the linear equation

$$\begin{cases} -\dfrac{\partial v^*}{\partial t} - \dfrac{1}{2}\triangle v^* = \gamma \lambda(x) \exp(kt)\, v^*, \\ \qquad v^*(x,T) = \exp(\gamma |h(x)| \exp(kT)) \end{cases} \tag{6}$$

and $\bar{u}(x,t) \leq \dfrac{1}{\gamma} \exp(-kt) \log v^*(x,t)$. The problem reduces to the solvability of (6), which is not warranted, in spite of the linearity. One good way to capture the solvability issue is to use the probabilistic interpretation of $v^*(x,t)$. We have, by a formal application of Feymann-Kac's,

$$v^*(x,t) = \mathbb{E}\left(\exp\left(\gamma \left(|h|(w_{xt}(T)) \exp(kT) + \int_t^T \lambda(w_{xt}(s)) \exp(ks)\, ds \right)\right)\right), \tag{7}$$

in which $w_{xt}(s) = x + w(s) - w(t)$, where $w(s)$ is a standard Wiener process in $\mathbb{R}^n$. It is clear that the expectation of the right-hand side of (7) can be $+\infty$, unless the growth of λ and $|h|$ is restricted.

1.3 Sufficient Condition

We introduce a function $\zeta(x)$ such that

$$\begin{cases} (\mu - k)\zeta - \dfrac{1}{2}\triangle\zeta - \dfrac{\gamma}{2}|D\zeta|^2 \geq \lambda(x) \exp(\mu T), \\ \qquad\qquad \zeta(x) \geq |h(x)| \exp(\mu T) \end{cases} \tag{8}$$

for a suitable choice of $\mu > k$. We have the first bounding result:

Lemma 1 *It holds*

$$\bar{u}(x,t) \leq \zeta(x). \tag{9}$$

Proof Let $\bar{\zeta}(x,t) = \exp\left(\gamma\zeta(x)\exp(-(\mu-k)t)\right)$. We get

$$
\begin{aligned}
-\frac{\partial\bar{\zeta}}{\partial t} - \frac{1}{2}\triangle\bar{\zeta} &= \gamma\bar{\zeta}\exp(-(\mu-k)t)\left[(\mu-k)\zeta - \frac{1}{2}\triangle\zeta - \frac{\gamma}{2}\exp(-(\mu-k)t)\,|D\zeta|^2\right]\\
&\geq \gamma\bar{\zeta}\lambda(x)\exp(kt)\ \exp(\mu(T-t))\\
&\geq \gamma\lambda(x)\exp(kt)\,\bar{\zeta},
\end{aligned}
$$

and $\bar{\zeta}(x,T) \geq \exp(\gamma|h(x)|)\exp(kT)$. Comparing with the inequalities (5), by the maximum principle, we see that $\bar{\zeta}(x,t) \geq \bar{v}(x,t)$ hence $\exp\left[\gamma\zeta(x)\exp kt\ \exp(-\mu t)\right] \geq \exp((\gamma\bar{u}(x,t))\exp(kt)$, which implies immediately the result (9).

Therefore, a sufficient condition for the solvability of (4) is to find a function $\zeta(x)$ satisfying (8). We can then state that:

Lemma 2 *We assume that $\lambda(x)$, $|h(x)|$ satisfy the following respective growth conditions:*

$$
\begin{aligned}
\lambda(x) &\leq |x|^2\frac{\exp(-(2+kT))}{2\gamma T^2} + C, \qquad (10)\\
|h(x)| &\leq |x|^2\frac{\exp(-(2+kT))}{2\gamma T^2} + C.
\end{aligned}
$$

Then the function

$$
\zeta(x) = \frac{|x|^2}{2\gamma T} + B,
$$

and $\mu - k = \dfrac{2}{T}$ satisfy (8), for a sufficiently large, in comparison with C, constant $B > 0$.

Proof The result is obtained by direct calculation. We must take, since $\mu T = 2 + kT$,

$$
B \geq (C\exp(2+kT)) \vee \left(\frac{n}{4\gamma} + \frac{CT}{2}\exp(2+kT)\right).
$$

1.4 Main Result

We then state the main theorem in this article:

Theorem 1 *We assume (2), (3) and (10). Then there exists a solution of (1).*

The proof will be given in the next section. We can give a uniqueness result under some additional assumptions; indeed, if we assume further that

$$|H(x, y, z) - H(x, y', z)| \leq k|y - y'|, \tag{11}$$

$$z \mapsto H(x, y, z) \text{ is convex, for each } (x, y), \tag{12}$$

we then have

Theorem 2 *With the assumptions of Theorem 1, (11), (12), and the growth condition (23) stated below, the solution of (1) is unique.*

2 Methodology

In this section, we aim to prove Theorems 1 and 2. We proceed by first making more assumptions, which will permit to construct an approximation. We begin by the standard case: the Lipschitz case.

Proposition 1 *We assume that, for any (x, y, z) and (x', y', z'), for some $C > 0$,*

$$|H(x, y, z) - H(x, y', z')| \leq C(|y - y'| + |z - z'|), \tag{13}$$

$$|H(x, y, z)| \leq \lambda(x) + k(|y| + |z|), \tag{14}$$

with $\lambda(x)$, $|h(x)|$ being of polynomial growth. There exists one and only one solution of (1) in a Sobolev space with a weight – a positive function $\Phi_m(x) := \dfrac{1}{(1 + |x|^2)^{\frac{m}{2}}}$ for $x \in \mathbb{R}^n$, for some suitable choice of $m \in \mathbb{Z}^+$.

Proof To simplify the notation, if there is no cause of ambiguity, we may skip the subscript m in Φ_m in the rest of this proof. We see that $|D\Phi(x)| \leq c_0\Phi(x)$, for some C_0 depending on m. The growth assumption on $\lambda(x)$ and $|h|(x)$, together with a large enough m, can warrant that

$$\int_{\mathbb{R}^n} \lambda^2(x)\Phi^2(x)dx < +\infty, \quad \int_{\mathbb{R}^n} h^2(x)\Phi^2(x)dx < +\infty, \tag{15}$$

and we say that $\lambda, h \in L^2_\Phi(\mathbb{R}^n)$. We next define the Sobolev space:

$$H^1_\Phi(\mathbb{R}^n) := \left\{ v \mid \int_{\mathbb{R}^n} v^2(x)\Phi^2(x)dx < +\infty \text{ and } \int_{\mathbb{R}^n} |Dv|^2(x)\Phi^2(x)dx < +\infty \right\},$$

and we also consider the space $\mathscr{H} := L^2(0, T; H^1_\Phi(\mathbb{R}^n))$ with the norm, for any $v \in \mathscr{H}$,

$$\|v\|^2 := \int_0^T \int_{\mathbb{R}^n} (|v|^2 + |Dv|^2)\Phi^2(x) \exp(\mu t)\, dx\, dt,$$

in which μ is a positive constant to be determined later. We shall obtain the solution as the fixed point of a contraction map Θ. For $v \in \mathscr{H}$, we define the map $\Theta(v) = u$

as the solution of

$$\begin{cases} -\dfrac{\partial u}{\partial t} - \dfrac{1}{2}\triangle u = H(x, v, Dv), \\ \qquad u(x, T) = h(x), \end{cases}$$

which is feasible as $|D\Phi(x)| \leq c_0\Phi(x)$. Note that $\dfrac{\partial u}{\partial t} \in L^2(0, T; (H^1_\Phi(\mathbb{R}^n))')$, where $(H^1_\Phi(\mathbb{R}^n))'$ is the dual space of $H^1_\Phi(\mathbb{R}^n)$. Besides, $u(\cdot, T)$ is defined as an element of $L^2_\Phi(\mathbb{R}^n)$. Thanks to the assumption (14) and condition (15) we see easily that the map Θ is well-defined, and maps $\mathscr{H}$ into itself. To prove that for some choice of μ that the map Θ is a contraction, consider two functions v_1, v_2 and the images $\Theta(v_1) = u_1$ and $\Theta(v_2) = u_2$. We also set $\tilde{v} = v_1 - v_2,\ \tilde{u} = u_1 - u_2$. We have

$$-\frac{\partial \tilde{u}}{\partial t} - \frac{1}{2}\triangle \tilde{u} = H(x, v_1, Dv_1) - H(x, v_2, Dv_2),$$
$$\tilde{u}(x, T) = 0.$$

Testing this equation with $\tilde{u}\Phi^2$ and integrating we obtain

$$\frac{1}{2}\int_{\mathbb{R}^n} \tilde{u}^2(x, 0)\Phi^2(x)dx + \frac{\mu}{2}\int_0^T \int_{\mathbb{R}^n} \tilde{u}^2\Phi^2 \exp \mu t\, dx\, dt$$
$$+\frac{1}{2}\int_0^T \int_{\mathbb{R}^n} |D\tilde{u}|^2\Phi^2 \exp \mu t\, dx\, dt + \int_0^T \int_{\mathbb{R}^n} D\tilde{u} \cdot D\Phi\tilde{u}\Phi \exp \mu t\, dx\, dt$$
$$\leq C\int_0^T \int_{\mathbb{R}^n} (|\tilde{v}| + |D\tilde{v}|)|\tilde{u}|\Phi^2 \exp \mu t\, dx\, dt.$$

Using the condition on $|D\Phi|$, it implies

$$\frac{\mu}{2}\int_0^T \int_{\mathbb{R}^n} \tilde{u}^2\Phi^2 \exp \mu t\, dx\, dt + \frac{1}{2}\int_0^T \int_{\mathbb{R}^n} |D\tilde{u}|^2\Phi^2 \exp \mu t\, dx\, dt$$
$$\leq c_0\int_0^T \int_{\mathbb{R}^n} |D\tilde{u}||\tilde{u}|\Phi^2 \exp \mu t\, dx\, dt + C\int_0^T \int_{\mathbb{R}^n} (|\tilde{v}| + |D\tilde{v}|)|\tilde{u}|\Phi^2 \exp \mu t\, dx\, dt.$$

By easy majorations we obtain

$$\frac{1}{4}\int_0^T \int_{\mathbb{R}^n} |D\tilde{u}|^2\Phi^2 \exp \mu t\, dx\, dt + \Big(\frac{\mu}{2} - 4C^2 - c_0^2\Big)\int_0^T \int_{\mathbb{R}^n} \tilde{u}^2\Phi^2 \exp \mu t\, dx\, dt \leq \frac{1}{8}\|\tilde{v}\|^2.$$

Choosing μ such that $\dfrac{\mu}{2} - 4C^2 - c_0^2 \geq \dfrac{1}{4}$, it follows $\dfrac{1}{4}\|\tilde{u}\|^2 \leq \dfrac{1}{8}\|\tilde{v}\|^2$, which proves that Θ is a contraction. This completes the proof. ∎

We next state an important comparison result. Consider another Hamiltonian $H'(x, y, z)$ and another initial condition h'. We assume that

$$H'(x, y, z) \geq H(x, y, z),\ h'(x) \geq h(x).$$

We do not make on $H'(x, y, z)$ the same assumptions as for $H(x, y, z)$. However, we assume that there exists a solution $u'(x, t)$ of the equation

$$-\frac{\partial u'}{\partial t} - \frac{1}{2}\Delta u' = H(x, u', Du'),$$
$$u'(x, T) = h'(x)$$

in the space $\mathscr{H}$. We state

Proposition 2 *We have the comparison property* $u'(x, t) \geq u(x, t)$.

Proof Setting $\tilde{u} = u' - u$ then

$$\begin{aligned}-\frac{\partial \tilde{u}}{\partial t} - \frac{1}{2}\Delta \tilde{u} &= H'(x, u', Du') - H(x, u, Du)\\ &\geq H(x, u', Du') - H(x, u, Du)\end{aligned}$$

and $\tilde{u}(x, T) \geq 0$, and $\tilde{u}$ belongs to $\mathscr{H}$.We test the inequality with $(\tilde{u})^- \Phi^2 \exp \mu t$. We note that

$$(H(x, u', Du') - H(x, u, Du))(\tilde{u})^- \geq -C(|(\tilde{u})^-|^2 + (\tilde{u})^-|D(\tilde{u})^-|).$$

After integration we get easily

$$\begin{aligned}\Big(\frac{\mu}{2} - C\Big)\int_0^T \int_{\mathbb{R}^n} |(\tilde{u})^-|^2 \Phi^2 \exp \mu t \, dx\, dt + \frac{1}{2}\int_0^T \int_{\mathbb{R}^n} |D(\tilde{u})^-|^2 \Phi^2 \exp \mu t \, dx\, dt\\ \leq (C + c_0)\int_0^T \int_{\mathbb{R}^n} |D(\tilde{u})^-||(\tilde{u})^-|\Phi^2 \exp \mu t \, dx\, dt\end{aligned}$$

and from the choice of μ, we necessarily have $(\tilde{u})^- = 0$, hence the result. ■

2.1 Continuous Hamiltonian

The objective is to relax the assumption that the Hamiltonian is Lipschitz. We still assume (14) but we replace (13) by (2), with the same growth conditions on $\lambda(x)$, $|h(x)|$. We then have

Proposition 3 *We assume (2), (14) and* $\lambda(x)$, $|h(x)|$ *have polynomial growth. Then the set of solutions of (1) in* $\mathscr{H}$ *is not empty and has a minimum and a maximum solution.*

Proof We first check that the solutions, if they exist, lie in an interval. We define indeed the unique solution $\bar{u}(x, t)$ of

$$-\frac{\partial \bar{u}}{\partial t} - \frac{1}{2}\triangle \bar{u} = \lambda(x) + k(|\bar{u}| + |D\bar{u}|),$$
$$\bar{u}(x,T) = |h(x)|.$$

The existence and uniqueness of $\bar{u}$ in the Hilbert space $\mathscr{H}$ is a simple application of Proposition 1. The next step is to prove that if u is a solution of (1) in the Hilbert $\mathscr{H}$ then we have

$$|u(x,t)| \leq \bar{u}(x,t). \tag{16}$$

Let us consider $\tilde{u}(x,t) = u(x,t) + \bar{u}(x,t)$ and let us prove that it is positive. Indeed, $\tilde{u}$ satisfies

$$\begin{aligned} -\frac{\partial \tilde{u}}{\partial t} - \frac{1}{2}\triangle \tilde{u} &= \lambda(x) + k(|\bar{u}| + |D\bar{u}|) + H(x,u,Du) \\ &\geq k(|\bar{u}| + |D\bar{u}|) - k(|u| + |Du|) \\ &\geq -k|\tilde{u}| - k|D\tilde{u}|, \\ \tilde{u}(x,T) &\geq 0. \end{aligned}$$

We test with $(\tilde{u})^{-}\Phi^2 \exp \mu t$. After integration and already seen majorations we can write

$$\begin{aligned} 0 \geq (\frac{\mu}{2} - k)\int_0^T \int_{\mathbb{R}^n} |(\tilde{u})^{-}|^2\, \Phi^2 \exp \mu t\, dx\, dt + \frac{1}{2}\int_0^T \int_{\mathbb{R}^n} |D(\tilde{u})^{-}|^2 \Phi^2 \exp \mu t\, dx\, dt \\ - (c_0 + k)\int_0^T \int_{\mathbb{R}^n} |D(\tilde{u})^{-}||(\tilde{u})^{-}|\Phi^2 \exp \mu t\, dx\, dt \end{aligned}$$

and for a convenient choice of μ, we obtain $(\tilde{u})^{-} = 0$. Therefore $u(x,t) \geq -\bar{u}(x,t)$. Similarly, we prove $u(x,t) \leq \bar{u}(x,t)$. Hence the result (16) is proven. We now construct a minimum solution. We define the sequence of Hamiltonians

$$H_N(x,y,z) = \inf_{\xi,\eta}(H(x,\xi,\eta) + N(|\xi - x| + |\eta - y|))$$

and N will tend to $+\infty$.We have $H_N(x,y,z) \leq H(x,y,z)$. Also

$$\begin{aligned} H(x,\xi,\eta) + N(|\xi - x| + |\eta - y|) &\geq -\lambda(x) - k|\xi| - k|\eta| + N(|\xi - x| + |\eta - y|) \\ &\geq -\lambda(x) - k|x| - k|y| + (N - k)(|\xi - x| + |\eta - y|). \end{aligned}$$

So for $N > k$, it goes to $+\infty$, as $|\xi|, |\eta| \to +\infty$. Therefore, since $H(x,\xi,\eta)$ is continuous in ξ, η the infimum is attained at a point ξ_N, η_N. Also

$$H(x,y,z) \geq -\lambda(x) - k|x| - k|y| + (N - k)(|\xi_N - x| + |\eta_N - y|).$$

This implies $\xi_N \to x$, $\eta_N \to y$, as $N \to +\infty$. Now

$$H_N(x,y,z) \geq H(x,\xi_N,\eta_N)$$

and $\liminf\ H_N(x, y, z) \geq H(x, y, z)$. Therefore

$$H_N(x, y, z) \uparrow H(x, y, z). \tag{17}$$

Since also from above, $H_N(x, y, z) \geq -\lambda(x) - k|x| - k|y|$, we have the estimate

$$|H_N(x, y, z)| \leq \lambda(x) + k(|y| + |z|).$$

On the other hand,

$$|H_N(x, y, z) - H_N(x, y', z')| \leq N(|y - y'| + |z - z'|)$$

so H_N is Lipschitz. Therefore, from Proposition 1, there exists a unique solution $u_N(x, t)$ of the equation

$$-\frac{\partial u_N}{\partial t} - \frac{1}{2}\Delta u_N = H_N(x, u_N, Du_N),$$
$$u_N(x, T) = h(x)$$

in the space $\mathscr{H}$. Although we should take a norm in $\mathscr{H}$ depending on N, we can see easily that we can take a norm not depending on N. Indeed, testing with $u_N \Phi^2 \exp \mu t$ we obtain the inequality

$$\begin{aligned}(\frac{\mu}{2} - k)\int_0^T \int_{\mathbb{R}^n} (u_N)^2 \Phi^2 \exp \mu t\, dx\, dt + \frac{1}{2}\int_0^T \int_{\mathbb{R}^n} |Du_N|^2 \Phi^2 \exp \mu t\, dx\, dt \\ \leq (c_0 + k)\int_0^T \int_{\mathbb{R}^n} |u_N||Du_N|\Phi^2 \exp \mu t\, dx\, dt + \int_0^T \int_{\mathbb{R}^n} \lambda(x)|u_N|\Phi^2 \exp \mu t\, dx\, dt \\ + \exp \mu T \int_{\mathbb{R}^n} h^2(x)\Phi^2(x) dx\end{aligned}$$

and we see that we can take the norm in $\mathscr{H}$ independent of N. We have also

$$|u_N(x, t)| \leq \bar{u}(x, t),\ \ \|u_N\|_{\mathscr{H}} \leq C$$

and since $H_N(x, y, z)$ is monotone increasing in N, we can assert from the comparison property, Proposition 2, $u_N(x, t) \leq u_{N+1}(x, t)$. Therefore the sequence $u_N(x, t)$ converges pointwise to $u(x, t) \leq \bar{u}(x, t)$. Also from the estimate in $\mathscr{H}$ we can assert for a subsequence

$$u_N \rightharpoonup u,\ \text{weakly in } \mathscr{H}.$$

We want to prove that u_N converges strongly in $\mathscr{H}$. Let $M > N$. We are going to let M tend to $+\infty$,for fixed N. Set $\tilde{u}_M = u_M - u_N$. We have

$$-\frac{\partial \tilde{u}_M}{\partial t} - \frac{1}{2}\Delta \tilde{u}_M = H_M(x, u_M, Du_M) - H_N(x, u_N, Du_N),$$
$$\tilde{u}_M(x, T) = 0.$$

We shall use the inequality

$$|H_M(x, u_M, Du_M) - H_N(x, u_N, Du_N)| \leq k|\tilde{u}_M| + k|D\tilde{u}_M| + 2\lambda(x) + k(|u_N| + |Du_N|).$$

We test the equation with $\tilde{u}_M \Phi^2 \exp \mu t$, and after integration and already seen majorations we obtain

$$\begin{aligned}\left(\frac{\mu}{2} - k\right) \int_0^T \int_{\mathbb{R}^n} (\tilde{u}_M)^2 \Phi^2 \exp \mu t \, dx \, dt + \frac{1}{2} \int_0^T \int_{\mathbb{R}^n} |D\tilde{u}_M|^2 \Phi^2 \exp \mu t \, dx \, dt \\ \leq (c_0 + k) \int_0^T \int_{\mathbb{R}^n} |\tilde{u}_M||D\tilde{u}_M| \Phi^2 \exp \mu t \, dx \, dt \\ + 2 \int_0^T \int_{\mathbb{R}^n} (\lambda(x) + k(|u_N| + |Du_N|)|\tilde{u}_M| \Phi^2 \exp \mu t \, dx \, dt\end{aligned}$$

and for μ fixed, sufficiently large, it follows

$$\frac{1}{4} \int_0^T \int_{\mathbb{R}^n} |D\tilde{u}_M|^2 \Phi^2 \exp \mu t \, dx \, dt \leq 2 \int_0^T \int_{\mathbb{R}^n} (\lambda(x) + k(|u_N| + |Du_N|)|\tilde{u}_M| \Phi^2 \exp \mu t \, dx \, dt.$$

We now let $M \to +\infty$. From the weak convergence in $\mathscr{H}$, the weak lower semi-continuity of the norm and the strong pointwise convergence, we can assert that

$$\begin{aligned}\frac{1}{4} \int_0^T \int_{\mathbb{R}^n} |D(u - u_N)|^2 \Phi^2 \exp \mu t \, dx \, dt \\ \leq 2 \int_0^T \int_{\mathbb{R}^n} (\lambda(x) + k(|u_N| + |Du_N|)|u - u_N| \Phi^2 \exp \mu t \, dx \, dt.\end{aligned}$$

As $N \to +\infty$, the right-hand side goes to 0. This proves that $u_N \to u$ in $\mathscr{H}$ strongly.

We can then extract a subsequence, also noted u_N such that $Du_N(x, t) \to Du(x, t)$ a.e. From this subsequence we can also extract a new one such that

$$\int_0^T \int_{\mathbb{R}^n} \sup_N |Du_N|^2 \Phi^2 dx \, dt < +\infty \tag{18}$$

This is a classical result. It comes from the fact that Du_N is also a Cauchy sequence in $L^2(0, T; L^2_\Phi(\mathbb{R}^n))$. Therefore, we can find a subsequence Du_{N_j} such that

$$\left(\int_0^T \int_{\mathbb{R}^n} |Du_{N_{j+1}} - Du_{N_j}|^2 \Phi^2 dx \, dt\right)^{\frac{1}{2}} \leq \frac{1}{2^j}$$

and, consequently, the function

$$g = |Du_{N_0}| + \sum_{j=0}^{+\infty} |Du_{N_{j+1}} - Du_{N_j}|$$

is in $L^2(0, T; L^2_\Phi(\mathbb{R}^n))$, and $|Du_{N_j}| \leq g$, for all j. Thus $\sup_j |Du_{N_j}| \leq g$ and (18) follows, for this subsequence. The next claim is, for this subsequence, still denoted u_N, that

$$H_N(x, u_N, Du_N) \to H(x, u, Du) \text{ a.e.} \tag{19}$$

This is done in a way very similar to (17), together with the pointwise convergence of u_N, Du_N. Also

$$\sup_N |H_N(x, u_N, Du_N)| \leq \lambda(x) + k\bar{u}(x, t) + k \sup_N |Du_N(x, t)|$$

and the right-hand side belongs to $L^2(0, T; L^2_\Phi(\mathbb{R}^n))$. Therefore, we can apply Lebesgue's theorem, to conclude, from (19)

$$\int_0^T \int_{\mathbb{R}^n} |H_N(x, u_N, Du_N)|^2 \Phi^2 dx\, dt \to \int_0^T \int_{\mathbb{R}^n} |H(x, u, Du)|^2 \Phi^2 dx\, dt.$$

This implies that $H_N(x, u_N, Du_N) \to H(x, u, Du)$ in $L^2(0, T; L^2_\Phi(\mathbb{R}^n))$. This proves that u is a solution of (1). This solution is the minimum one. Indeed, if we have another one in $\mathscr{H}$, called u^*,then

$$-\frac{\partial u^*}{\partial t} - \frac{1}{2}\triangle u* = H(x, u*, Du*),$$
$$u * (x, T) = h(x).$$

But $H(x, u*, Du*) \geq H_N(x, u*, Du*)$ and from the comparison result, Proposition 2, we conclude that $u* \geq u_N$. Therefore, also $u* \geq u$. This proves the assertion.

The maximum solution is constructed, by considering the approximation

$$H^N(x, y, z) = \sup_{\xi, \eta}(H(x, \xi, \eta) - N(|\xi - x| + |\eta - y|))$$

and the proof has been completed. ■

We can also give a comparison result. Consider two Hamiltonians $H(x, y, z)$, $H'(x, y, z)$ satisfying

$$H(x, y, z),\ \ H'(x, y, z) \text{ are continuous in } y, z$$
$$|H(x, y, z)|, |H'(x, y, z)| \leq \lambda(x) + k(|y| + |z|)$$

and

$$H(x, y, z) \geq H'(x, y, z), \quad h(x) \geq h'(x).$$

We state

Proposition 4 *Let u, u' be the minimum (or the maximum) solution of*

$$-\frac{\partial u}{\partial t} - \frac{1}{2}\triangle u = H(x, u, Du),$$
$$u(x, T) = h(x)$$

and

$$-\frac{\partial u'}{\partial t} - \frac{1}{2}\triangle u' = H(x, u', Du'),$$
$$u'(x, T) = h(x).$$

Then we have $u \geq u'$.

Proof Considering the approximations $H_N(x, y, z)$ and $H'_N(x, y, z)$, we have $H_N(x, y, z) \geq H'_N(x, y, z)$. Therefore, the corresponding solutions $u_N(x, t)$ and $u'_N(x, t)$ satisfy $u_N \geq u'_N$. Hence for the limit, $u \geq u'$. This implies the result. ■

2.2 Proof of Theorem 1

We construct an approximation of H as follows:

$$H_{\varepsilon\eta}(x, y, z) = \frac{H^+(x, y, z)}{1 + \varepsilon|z|^2} - \frac{H^-(x, y, z)}{1 + \eta|z|^2}.$$

Clearly

$$|H_{\varepsilon\eta}(x, y, z)| \leq \lambda(x) + k|y| + \frac{\gamma}{2}|z|^2,$$
$$H_{\varepsilon\eta}(x, y, z) \text{ is decreasing in } \varepsilon \text{ and increasing in } \eta,$$
$$-\lambda(x) - k|y| - \frac{\gamma}{2\eta} \leq H_{\varepsilon\eta}(x, y, z) \leq \lambda(x) + k|y| + \frac{\gamma}{2\varepsilon}.$$

To simplify notation we set $\rho = (\varepsilon, \eta)$. The Hamiltonian $H_\rho(x, y, z)$ satisfies the assumptions of Proposition 3. So considering the problem

$$-\frac{\partial u_\rho}{\partial t} - \frac{1}{2}\Delta u_\rho = H_\rho(x, u_\rho, Du_\rho),$$
$$u_\rho(x, T) = h(x),$$

there exists a minimum solution in the space $\mathscr{H}$. Consider now Eq. (4). From the growth conditions, we can assert that

$$|u_\rho(x,t)| \le \bar{u}(x,t) \le \zeta(x).$$

Indeed,

$$H_\rho(x,y,z) \le \lambda(x) + k|y| + \frac{\gamma}{2}\frac{|z|^2}{1+\varepsilon|z|^2}$$

and by the comparison property we get $u_\rho(x,t) \le \bar{u}_\varepsilon(x,t)$ solution of

$$-\frac{\partial \bar{u}_\varepsilon}{\partial t} - \frac{1}{2}\Delta \bar{u}_\varepsilon = \lambda(x) + k|\bar{u}_\varepsilon| + \frac{\gamma}{2}\frac{|D\bar{u}_\varepsilon|^2}{1+\varepsilon|D\bar{u}_\varepsilon|^2},$$
$$\bar{u}_\varepsilon(x,T) = |h(x)|.$$

The solution is unique and positive, so we do not need the absolute value. We then see that $\bar{u}_\varepsilon(x,t) \le \bar{u}(x,t)$. Similarly, we check that $u_\rho(x,t) \ge -\bar{u}(x,t)$.

Because of the quadratic growth of the Hamiltonian, we cannot use the same Hilbert space $\mathscr{H}$ as before. We shall need to use a weight function which goes fast to 0 at infinity. Let $\delta > 0$, to be chosen later. We take the weight

$$\Phi(x) = \exp -\frac{\delta}{2}\zeta^2(x).$$

The property $|D\Phi| \le c_0\Phi$ does not hold, but we have $|D\zeta| \le c_1\zeta$, since ζ is quadratic. We can assert that

$$\int_{\mathbb{R}^n} \lambda^2(x)\Phi^2(x)\exp\frac{\delta}{2}\zeta^2(x)dx < +\infty, \int_{\mathbb{R}^n} \zeta^2(x)\Phi^2(x)\exp\frac{\delta}{2}\zeta^2(x)dx \quad < +\infty,$$
$$\int_{\mathbb{R}^n} |D\Phi|^2\exp\frac{\delta}{2}\zeta^2(x)dx < +\infty. \tag{20}$$

To save notation, we still denote $L^2_\Phi(\mathbb{R}^n) = \{v \mid \int_{\mathbb{R}^n} v^2(x)\Phi^2(x)dx\}$ and the Sobolev space $H^1_\Phi(\mathbb{R}^n) = \{v \mid v \in L^2_\Phi(\mathbb{R}^n),\;\; Dv \in (L^2_\Phi(\mathbb{R}^n))^n\}$. We now test (2.2) with $u_\rho\Phi^2\exp\frac{\delta}{2}(u_\rho)^2$. We obtain

$$\frac{1}{2}\int_0^T\int_{\mathbb{R}^n} |Du_\rho|^2(1+\delta u_\rho^2)\Phi^2\exp\frac{\delta}{2}(u_\rho)^2dx\,dt + \int_0^T\int_{\mathbb{R}^n} Du_\rho\cdot D\Phi u_\rho\Phi\exp\frac{\delta}{2}(u_\rho)^2dx\,dt$$
$$\le \int_0^T\int_{\mathbb{R}^n}\left(\lambda(x) + k|u_\rho| + \frac{\gamma}{2}|Du_\rho|^2\right)|u_\rho|\Phi^2\exp\frac{\delta}{2}(u_\rho)^2dx\,dt + \frac{1}{\delta}\int_{\mathbb{R}^n}\Phi^2\exp\frac{\delta}{2}h^2dx.$$

For any $\beta > 0$, we can then write

$$\frac{1}{2}\int_0^T\int_{\mathbb{R}^n}|Du_\rho|^2(1+(\delta-\beta)u_\rho^2-\gamma|u_\rho|)\Phi^2\exp\frac{\delta}{2}(u_\rho)^2dx\,dt$$
$$\leq\frac{1}{2\beta}\int_0^T\int_{\mathbb{R}^n}|D\Phi|^2\exp\frac{\delta}{2}(u_\rho)^2dx\,dt+\int_0^T\int_{\mathbb{R}^n}(\lambda(x)+k|u_\rho|)|u_\rho|\Phi^2\exp\frac{\delta}{2}(u_\rho)^2dx\,dt$$
$$+\frac{1}{\delta}\int_{\mathbb{R}^n}\Phi^2\exp\frac{\delta}{2}h^2dx.$$

Since $|u_\rho|\leq\zeta$, we majorize the right-hand side, by replacing u_ρ by ζ. We next choose β and δ such that $\delta>2\beta+\frac{\gamma^2}{4}$. Defining $a_0=\min(\beta,1-\frac{\gamma^2}{4(\delta-2\beta)})$, we finally obtain the inequality

$$\frac{a_0}{2}\int_0^T\int_{\mathbb{R}^n}|Du_\rho|^2(1+u_\rho^2)\Phi^2\exp\frac{\delta}{2}(u_\rho)^2dx\,dt\leq\frac{T}{2}\int_{\mathbb{R}^n}\lambda^2\Phi^2\exp\frac{\delta}{2}\zeta^2(x)dx$$
$$+T\left(k+\frac{1}{2}\right)\int_{\mathbb{R}^n}\Phi^2\zeta^2\exp\frac{\delta}{2}\zeta^2dx+\frac{T}{2\beta}\int_{\mathbb{R}^n}|D\Phi|^2\exp\frac{\delta}{2}\zeta^2dx$$
$$+\frac{1}{\delta}\int_{\mathbb{R}^n}\Phi^2\exp\frac{\delta}{2}h^2dx. \quad (21)$$

Thanks to (20) the right-hand side is bounded, and therefore $u_\rho$remains in a bounded set of $L^2(0,T;H^1_\Phi(\mathbb{R}^n))$.

Next, we proceed in two steps. In $\rho=(\varepsilon,\eta)$, we fix η and we let $\varepsilon\to 0$. To simplify notation, we simply write $\rho=\varepsilon$. We have the following properties:

$$H_\varepsilon(x,y,z)\uparrow H^+(x,y,z)-\frac{H^-(x,y,z)}{1+\eta|z|^2}.$$

From the comparison property, we can assert that

$$u_\varepsilon\uparrow u\leq\zeta$$

and from the bound (21) we get also, for a subsequence

$$Du_\varepsilon\rightharpoonup Du,\ \text{in } L^2(0,T;L^2_\Phi(\mathbb{R}^n))\text{ weakly.}$$

We are going to show the strong convergence. However, we shall need to use a weight Ψ more stringent than Φ. Let $\varepsilon'<\varepsilon$, we test (2.2) with $(u_\varepsilon-u_{\varepsilon'})\Psi^2\exp\frac{\delta}{2}(u_\varepsilon-u_{\varepsilon'})^2$. Remember that ρ has been replaced with ε (η is fixed). After integration and rearrangements, we can state the inequality

$$\frac{1}{2}\int_0^T\int_{\mathbb{R}^n}|D(u_\varepsilon-u_{\varepsilon'})|^2[1-\beta+\delta(u_\varepsilon-u_{\varepsilon'})^2-\gamma|u_\varepsilon-u_{\varepsilon'}|]\Psi^2\exp\frac{\delta}{2}(u_\varepsilon-u_{\varepsilon'})^2dx\,dt$$
$$\leq\frac{1}{2\beta}\int_0^T\int_{\mathbb{R}^n}|D\Psi|^2(u_\varepsilon-u_{\varepsilon'})^2\exp\frac{\delta}{2}(u_\varepsilon-u_{\varepsilon'})^2dx\,dt$$
$$+\int_0^T\int_{\mathbb{R}^n}[2\lambda{+}2k\zeta+3\gamma|Du_\varepsilon-Du|^2+3\gamma|Du|^2]|u_\varepsilon-u_{\varepsilon'}|\Psi^2\exp\frac{\delta}{2}(u_\varepsilon-u_{\varepsilon'})^2dx\,dt.$$

We choose $\beta<\frac{1}{2}$, $\delta>\frac{\gamma^2}{2(\frac{1}{2}-\beta)}$, then the preceding inequality implies

$$\frac{1}{4}\int_0^T\int_{\mathbb{R}^n}|D(u_\varepsilon-u_{\varepsilon'})|^2[1+\delta(u_\varepsilon-u_{\varepsilon'})^2]\Psi^2\exp\frac{\delta}{2}(u_\varepsilon-u_{\varepsilon'})^2dx\,dt$$
$$\leq\frac{1}{2\beta}\int_0^T\int_{\mathbb{R}^n}|D\Psi|^2(u_\varepsilon-u_{\varepsilon'})^2\exp\frac{\delta}{2}(u_\varepsilon-u_{\varepsilon'})^2dx\,dt$$
$$+\int_0^T\int_{\mathbb{R}^n}[2\lambda+2k\zeta+3\gamma|Du_\varepsilon-Du|^2+3\gamma|Du|^2]|u_\varepsilon-u_{\varepsilon'}|\Psi^2\exp\frac{\delta}{2}(u_\varepsilon-u_{\varepsilon'})^2dx\,dt. \tag{22}$$

We want to let ε' tend to 0, while ε remains fixed. Recalling that $|u_\varepsilon|,|u_{\varepsilon'}|\leq\zeta$, and calling

$$F_{\varepsilon'}=\left[\frac{|D\Psi|^2}{2\beta}(u_\varepsilon-u_{\varepsilon'})^2+[2\lambda+2k\zeta\right.$$
$$\left.+3\gamma|Du_\varepsilon-Du|^2+3\gamma|Du|^2]|u_\varepsilon-u_{\varepsilon'}|\Psi^2\right]\exp\frac{\delta}{2}(u_\varepsilon-u_{\varepsilon'})^2,$$

we see that

$$F_{\varepsilon'}\to\left[\frac{|D\Psi|^2}{2\beta}(u_\varepsilon-u)^2+[2\lambda+2k\zeta\right.$$
$$\left.+3\gamma|Du_\varepsilon-Du|^2+3\gamma|Du|^2]|u_\varepsilon-u|\Psi^2\right]\exp\frac{\delta}{2}(u_\varepsilon-u)^2$$

pointwise. Also

$$F_{\varepsilon'}\leq 2\frac{|D\Psi|^2}{\beta}\zeta^2\exp 2\delta\zeta^2+2[2\lambda+2k\zeta$$
$$+3\gamma|Du_\varepsilon-Du|^2+3\gamma|Du|^2]\Psi^2\zeta\exp 2\delta\zeta^2.$$

So if we choose $\Psi(x)=\exp-2\delta\zeta^2(x)$, then we get

$$F_{\varepsilon'}\leq\frac{32\delta^2}{\beta}\zeta^2|D\zeta|^2\exp-2\delta\zeta^2(x)+2[2\lambda+2k\zeta$$
$$+3\gamma|Du_\varepsilon-Du|^2+3\gamma|Du|^2]\zeta\exp-2\delta\zeta^2(x)$$

which is a function independent of ε', integrable since $\int_0^T\int_{\mathbb{R}^n}|Du_\varepsilon|^2\exp-\delta\zeta^2dx\,dt<+\infty$, $\int_0^T\int_{\mathbb{R}^n}|Du|^2\exp-\delta\zeta^2dx\,dt<+\infty$. So we can go to the limit in the right-hand side of (22). For the left-hand side we first consider

$$X_{\varepsilon'\alpha}=D(u_\varepsilon-u_{\varepsilon'})\frac{[1+\delta(u_\varepsilon-u_{\varepsilon'})^2]^{\frac12}\exp\dfrac{\delta}{4}(u_\varepsilon-u_{\varepsilon'})^2}{1+\alpha(1+2\sqrt{\delta}\zeta)\exp\delta\zeta^2}.$$

Since $|u_\varepsilon|,|u_{\varepsilon'}|\le\zeta$, the quantity $\dfrac{[1+\delta(u_\varepsilon-u_{\varepsilon'})^2]^{\frac12}\exp\frac{\delta}{4}(u_\varepsilon-u_{\varepsilon'})^2}{1+\alpha(1+2\sqrt{\delta}\zeta)\exp\delta\zeta^2}$ is bounded by $\dfrac{1}{\alpha}$ and converges pointwise as $\varepsilon'\to0$ to $\dfrac{[1+\delta(u_\varepsilon-u)^2]^{\frac12}\exp\frac{\delta}{4}(u_\varepsilon-u)^2}{1+\alpha(1+2\sqrt{\delta}\zeta)\exp\delta\zeta^2}$. Since $D(u_\varepsilon-u_{\varepsilon'})$ converges weakly to $D(u_\varepsilon-u)$ in $L^2(0,T;L^2_\Phi(\mathbb{R}^n))$, the same is true for $X_{\varepsilon'\alpha}$. Therefore

$$\begin{aligned}&\int_0^T\int_{\mathbb{R}^n}|D(u_\varepsilon-u)|^2\frac{[1+\delta(u_\varepsilon-u)^2]\exp\dfrac{\delta}{2}(u_\varepsilon-u)^2}{[1+\alpha(1+2\sqrt{\delta}\zeta)\exp\delta\zeta^2]^2}\Psi^2dx\,dt\\&\le\liminf_{\varepsilon'\to0}\int_0^T\int_{\mathbb{R}^n}|D(u_\varepsilon-u_{\varepsilon'})|^2\frac{[1+\delta(u_\varepsilon-u_{\varepsilon'})^2]\exp\dfrac{\delta}{2}(u_\varepsilon-u_{\varepsilon'})^2}{[1+\alpha(1+2\sqrt{\delta}\zeta)\exp\delta\zeta^2]^2}\Psi^2dx\,dt\\&\le\liminf_{\varepsilon'\to0}\int_0^T\int_{\mathbb{R}^n}|D(u_\varepsilon-u_{\varepsilon'})|^2[1+\delta(u_\varepsilon-u_{\varepsilon'})^2]\exp\frac{\delta}{2}(u_\varepsilon-u_{\varepsilon'})^2\Psi^2dx\,dt\end{aligned}$$

and from the above considerations regarding $F_{\varepsilon'}$, we can state

$$\begin{aligned}&\le\frac{2}{\beta}\int_0^T\int_{\mathbb{R}^n}|D\Psi|^2(u_\varepsilon-u)^2\exp\frac{\delta}{2}(u_\varepsilon-u)^2dx\,dt+\\&+4\int_0^T\int_{\mathbb{R}^n}[2\lambda+2k\zeta+3\gamma|Du_\varepsilon-Du|^2+3\gamma|Du|^2]|u_\varepsilon-u|\Psi^2\exp\frac{\delta}{2}(u_\varepsilon-u)^2dx\,dt.\end{aligned}$$

Letting $\alpha\to0$, we conclude

$$\begin{aligned}&\int_0^T\int_{\mathbb{R}^n}|D(u_\varepsilon-u)|^2[1+\delta(u_\varepsilon-u)^2]\exp\frac{\delta}{2}(u_\varepsilon-u)^2\Psi^2dx\,dt\\&\qquad\le\frac{2}{\beta}\int_0^T\int_{\mathbb{R}^n}|D\Psi|^2(u_\varepsilon-u)^2\exp\frac{\delta}{2}(u_\varepsilon-u)^2dx\,dt\\&+4\int_0^T\int_{\mathbb{R}^n}[2\lambda+2k\zeta+3\gamma|Du_\varepsilon-Du|^2+3\gamma|Du|^2]|u_\varepsilon-u|\Psi^2\exp\frac{\delta}{2}(u_\varepsilon-u)^2dx\,dt.\end{aligned}$$

So, also

$$\begin{aligned}\int_0^T\int_{\mathbb{R}^n} &|D(u_\varepsilon-u)|^2[1+\delta(u_\varepsilon-u)^2-12\gamma|u_\varepsilon-u|]\exp\frac{\delta}{2}(u_\varepsilon-u)^2\Psi^2dx\,dt\\ &\le\frac{2}{\beta}\int_0^T\int_{\mathbb{R}^n}|D\Psi|^2(u_\varepsilon-u)^2\exp\frac{\delta}{2}(u_\varepsilon-u)^2dx\,dt\\ &+4\int_0^T\int_{\mathbb{R}^n}[2\lambda+2k\zeta+3\gamma|Du|^2]|u_\varepsilon-u|\Psi^2\exp\frac{\delta}{2}(u_\varepsilon-u)^2dx\,dt.\end{aligned}$$

We remember that β and δ were chosen so that

$$\beta<\frac{1}{2},\quad \delta>\frac{\gamma^2}{2(\frac{1}{2}-\beta)}.$$

By picking $\beta>\frac{1}{2}-\frac{1}{288}$, we can insure that $\delta>144\gamma^2$, so that $\frac{1}{2}+\frac{\delta}{2}(u_\varepsilon-u)^2-12\gamma|u_\varepsilon-u|\ge 0$. Therefore

$$\begin{aligned}\frac{1}{2}\int_0^T\int_{\mathbb{R}^n}&|D(u_\varepsilon-u)|^2[1+\delta(u_\varepsilon-u)^2]\exp\frac{\delta}{2}(u_\varepsilon-u)^2\Psi^2dx\,dt\\ &\le\frac{2}{\beta}\int_0^T\int_{\mathbb{R}^n}|D\Psi|^2(u_\varepsilon-u)^2\exp\frac{\delta}{2}(u_\varepsilon-u)^2dx\,dt\\ &\quad+8\int_0^T\int_{\mathbb{R}^n}(\lambda+k\zeta)|u_\varepsilon-u|\Psi^2\exp\frac{\delta}{2}(u_\varepsilon-u)^2dx\,dt.\end{aligned}$$

The right-hand side tends to 0, as $\varepsilon\to 0$. Therefore $u_\varepsilon\to u$, in $L^2(0,T;H^1_\Psi(\mathbb{R}^n))$ strongly. We can choose a subsequence such that

$$u_\varepsilon\uparrow u,\ Du_\varepsilon\to Du,\ \text{a.e.}\ ,\int_0^T\int_{\mathbb{R}^n}\sup_\varepsilon|Du_\varepsilon|^2\Psi^2dx\,dt<+\infty.$$

One easily checks that

$$H_\varepsilon(x,u_\varepsilon,Du_\varepsilon)\to H^+(x,u,Du)-\frac{H^-(x,u,Du)}{1+\eta|Du|^2},\quad\text{a.e.}$$

and in $L^1(0,T;L^1_\Psi(\mathbb{R}^n))$. Therefore u is a solution of

$$\begin{aligned}-\frac{\partial u}{\partial t}-\frac{1}{2}\triangle u&=H^+(x,u,Du)-\frac{H^-(x,u,Du)}{1+\eta|Du|^2},\\ u(x,T)&=h(x).\end{aligned}$$

Of course, this solution depends on η, called u_η (not to be mistaken with u_ε, with was a short for u_ρ) and we have the property u_η is monotone increasing in η. This is because $u_{\varepsilon,\eta_1}(x,t) \leq u_{\varepsilon,\eta_2}(x,t)$ if $\eta_1 < \eta_2$, with fixed ε. We can then let η tend to 0,monotonically, and operate in a way similar to that for ε. This leads to a solution of (1). This concludes the proof of Theorem 1.

2.3 Formal Proof of Theorem 2

We turn now to the uniqueness property. We make the assumptions (11), (12). We use a trick introduced by Da Lio-Ley [2], which takes advantage of the convexity assumption. Suppose we have 2 solutions u_1, u_2, in $L^2(0,T;H^1_\Psi(\mathbb{R}^n))$. In fact, we are going to assume sufficient smoothness of these functions. This is the formal aspect. For $0<\theta<1$, we consider $\tilde{u} = u_1 - \theta u_2$. Eventually θ will tend to 1. We have

$$-\frac{\partial \tilde{u}}{\partial t} - \frac{1}{2}\triangle \tilde{u} = H(x,u_1,Du_1) - \theta H(x,u_2,Du_2),$$
$$\tilde{u}(x,T) = (1-\theta)h(x).$$

From the convexity assumption we have

$$\begin{aligned} H(x,u_2,Du_1) &= H\left(x,u_2,\theta Du_2 + (1-\theta)\frac{Du_1-\theta Du_2}{1-\theta}\right) \\ &\leq \theta H(x,u_2,Du_2) + (1-\theta)H\left(x,u_2,\frac{D\tilde{u}}{1-\theta}\right). \end{aligned}$$

Hence

$$\begin{aligned} &H(x,u_1,Du_1) - \theta H(x,u_2,Du_2) \\ &\quad \leq H(x,u_1,Du_1) - H(x,u_2,Du_1) + (1-\theta)H\left(x,u_2,\frac{D\tilde{u}}{1-\theta}\right). \end{aligned}$$

Also

$$H(x,u_1,Du_1) - H(x,u_2,Du_1) \leq k|u_1-u_2| \leq k|\tilde{u}| + k(1-\theta)|u_2|.$$

Collecting estimates, one obtains

$$\begin{aligned} -\frac{\partial \tilde{u}}{\partial t} - \frac{1}{2}\triangle \tilde{u} - k|\tilde{u}| - \frac{\gamma}{2(1-\theta)}|D\tilde{u}|^2 &\leq (1-\theta)(\lambda(x) + 2k|u_2|) \\ \tilde{u}(x,T) &\leq (1-\theta)|h(x)|. \end{aligned}$$

But then $\tilde{v} = \dfrac{\tilde{u}}{1-\theta}$ satisfies

$$-\frac{\partial \tilde{v}}{\partial t} - \frac{1}{2}\Delta \tilde{v} - k|\tilde{v}| - \frac{\gamma}{2}|D\tilde{v}|^2 \leq \lambda(x) + 2k|u_2|,$$
$$\tilde{v}(x,T) \leq |h(x)|$$

and $(\tilde{v})^+(x,t) \leq z(x,t)$, solution of

$$-\frac{\partial z}{\partial t} - \frac{1}{2}\Delta z - kz - \frac{\gamma}{2}|Dz|^2 \exp kt = \lambda(x) + 2k\zeta(x),$$
$$z(x,T) = |h(x)|.$$

Here we make the growth condition (announced in the statement of the theorem):

$$\text{The function } z(x,t) \text{ is well defined.} \tag{23}$$

The inequality can be seen as follows. Define $\chi(x,t) = \exp(\gamma(\tilde{v})^+(x,t)\exp kt)$ and $\zeta(x,t) = \exp(\gamma z(x,t)\exp kt)$. It suffices to show that $\chi(x,t) \leq \zeta(x,t)$. Note first that ζ is the solution of the linear equation

$$-\frac{\partial \zeta}{\partial t} - \frac{1}{2}\Delta \zeta - \zeta\gamma \exp kt(\lambda(x) + 2k\zeta(x)) = 0,$$
$$\zeta(x,T) = \exp(\gamma|h(x)|\exp kT).$$

When $(\tilde{v})^+(x,t) = 0$, we have $\chi(x,t) = 1 \leq \zeta(x,t)$, since $z(x,t) \geq 0$. Also $\chi(x,T) \leq \exp(\gamma|h(x)|\exp kT) = \zeta(x,T)$. Here comes the formal argument: On the domain $(\tilde{v})^+(x,t) > 0$, $\tilde{v}(x,t)$ satisfies the inequality

$$-\frac{\partial \tilde{v}}{\partial t} - \frac{1}{2}\Delta \tilde{v} - k\tilde{v} - \frac{\gamma}{2}|D\tilde{v}|^2 \leq \lambda(x) + 2k\zeta$$

and $\chi(x,t) = \exp(\gamma\tilde{v}(x,t)\exp kt)$ satisfies

$$-\frac{\partial \chi}{\partial t} - \frac{1}{2}\Delta \chi - \chi\gamma \exp kt(\lambda(x) + 2k\zeta) \leq 0.$$

The formal aspect lies in the definition of the domain $(\tilde{v})^+(x,t) > 0$, which requires some smoothness of the solutions u_1, u_2. To conclude, we note that we have proven that $(\tilde{u})^+(x,t) \leq (1-\theta)z(x,t)$, or $(u_1 - \theta u_2)^+(x,t) \leq (1-\theta)z(x,t)$. But $z(x,t)$ is a fixed function, not depending on θ. So, we can let $\theta \to 1$, which leads to $u_1 - u_2 \leq 0$. Reversing the roles of u_1, u_2 we obtain the opposite inequality, hence $u_1 = u_2$, and the uniqueness of the solution. This concludes the proof. ■

Acknowledgements Alain Bensoussan's research is supported by the National Science Foundation under grants DMS-1303775, DMS-1612880 and the Research Grants Council of the Hong Kong Special Administrative Region (CityU 500113, HKGRF CityU 11303316: Mean Field Control with Partial Information).
Phillip Yam acknowledges the financial supports from HKRGC GRF14301015 with the project title: Advance in Mean Field Theory, and Direct Grant for Research 2015/16 with project code: 4053141 offered by CUHK. He also acknowledges the financial support from Department of Statistics of Columbia University in the City of New York during the period he was a visiting faculty member.

References

1. Briand P, Hu Y (2008) Quadratic BSDEs with convex generators and unbounded terminal conditions. Probab Theory Relat Fields 141(3–4):543–567
2. Da Lio F, Ley O (2006) Uniqueness results for second-order Bellman-Isaacs equations under quadratic growth assumptions and applications. SIAM J Control Optim 45(1):74–106
3. El Karoui N, Hamadène S, A. Matoussi (2009) Backward stochastic differential equations and applications, Princeton BSDE Lecture Notes
4. Friedman A (1964) Partial differential equations of parabolic type. Prentice-Hall, Englewood Cliffs
5. Kobylanski M (2000) Backward stochastic differential equations and partial differential equations with quadratic growth. Ann Probab 28(2):558–602
6. Ladyzhenskaya OA, Solonnikov VA, Ural'ceva NN (1968) Linear and quasi-linear equations of parabolic type. Transl Math Monogr. AMS

On the Sensitivity to the Filtering Radius in Leray Models of Incompressible Flow

Luca Bertagna, Annalisa Quaini, Leo G. Rebholz and Alessandro Veneziani

Abstract One critical aspect of Leray models for the Large Eddy Simulation (LES) of incompressible flows at moderately large Reynolds number (in the range of few thousands) is the selection of the filter radius. This drives the effective regularization of the filtering procedure, and its selection is a trade-off between stability (the larger, the better) and accuracy (the smaller, the better). In this paper, we consider the classical Leray-α and a recently introduced (by one of the authors) Leray model with a deconvolution-based indicator function, called Leray-α-NL. We investigate the sensitivity of the solutions to the filter radius by introducing the sensitivity systems, analyzing them at the continuous and discrete levels, and numerically testing them on two benchmark problems.

1 Introduction

The Direct Numerical Simulation (DNS) of the Navier-Stokes equations (NSE) computes the evolution of all the significant flow structures by resolving them with a properly refined mesh. Unfortunately, when convection dominates the dynamics – which happens in many practical applications – this requires very fine meshes, making DNS computationally unaffordable for practical purposes. A possible way to

L. Bertagna
Department of Scientific Computing, Florida State University, Tallahassee, FL, USA
e-mail: lbertagna@fsu.edu

A. Quaini (✉)
Department of Mathematics, University of Houston, Houston, TX, USA
e-mail: quaini@math.uh.edu

L. G. Rebholz
Department of Mathematical Sciences, Clemson University, Clemson, SC, USA
e-mail: rebholz@clemson.edu

A. Veneziani
Department of Mathematics and Computer Science, Emory University,
Atlanta, GA, USA
e-mail: ale@mathcs.emory.edu

B. N. Chetverushkin et al. (eds.), *Contributions to Partial Differential Equations and Applications*, Computational Methods in Applied Sciences 47,
https://doi.org/10.1007/978-3-319-78325-3_9

limit the computational costs associated with DNS without sacrificing accuracy is to solve for the flow average, and model properly the effects of the (not directly solved for) small scales on the (resolved) larger scales.

The Leray-α model (1)–(4) has emerged as a useful model for turbulent flow predictions, thanks to the seminal work of Guerts, Holm, Titi and co-workers [11, 13–15] in the early-mid 2000s. The name of the model was given by Titi to honor Leray, who used a similar model in 1934 as a theoretical tool to help in understanding the well-posedness problem of the NSE [31]. The Leray-α model in [31] describes the small scale effects by a set of equations to be added to the discrete NSE formulated on the under-refined mesh. It was shown in [11, 13–15] that the Leray-α model is well-posed, it can accurately predict turbulent flow on the large scales, where it preserves Kolmogorov's-5/3 law. Moreover, the model can accurately predict the boundary layer. Over the last decade, much more theoretical and computational work has been done to the Leray-α model and several variations of it [2, 7, 10, 16, 18, 19, 23, 29, 32, 33, 39], most of which gives further evidence of the usefulness of the model as an effective tool for coarse-mesh predictions of higher Reynolds number flow.

The filtering radius α plays a central role in the Leray-α model, and Leray type models in general, since it determines the amount of regularization to apply. In particular, larger values lead to more regularized solutions, while for $\alpha = 0$, the models reduce to the NSE; see (1)–(4). Our interest herein is to understand how solutions of the classical Leray-α model and one possible generalization, called Leray-α-NL, depend on α. Parameter sensitivity investigations in fluid flow problems are critical in understanding the reliability of computed solutions [1, 3–5, 17, 21, 34, 36–38]. However, it is often prohibitively costly to identify the appropriate value by running many computations with different choices, especially when the flow problems require fine meshes. An attractive alternative is the sensitivity equation method that computes explicitly the derivative of the solution with respect to the parameter. This system can then be solved simultaneously with the model at each time step of the simulation. Depending on the specific model, the solution of the sensitivity system may be challenging, and its analysis and efficient discretization design require specific investigation. This is exactly the purpose of this paper for the two models of choice.

The outline of the paper is as follows. In Sect. 2 we introduce the continuous Leray-α and Leray-α-NL models and derive the corresponding sensitivity equations. In Sect. 3 we propose efficient and stable numerical schemes for the approximation of both models and their sensitivity systems. Finally, in Sect. 4 we test the proposed numerical schemes against two benchmark problems. Conclusions are drawn in Sect. 5.

2 Problem Definition

We consider a spacial domain $\Omega \subset \mathbb{R}^d$ ($d = 2$ or 3) and time interval $(0, T)$, with $T > 0$. The classical Leray-α model takes the form:

$$\boldsymbol{u}_t + \overline{\boldsymbol{u}} \cdot \nabla \boldsymbol{u} + \nabla p - \nu \Delta \boldsymbol{u} = \boldsymbol{f} \quad \text{in } \Omega \times (0, T), \tag{1}$$
$$\nabla \cdot \boldsymbol{u} = 0 \quad \text{in } \Omega \times (0, T), \tag{2}$$
$$\nabla \lambda - \alpha^2 \Delta \overline{\boldsymbol{u}} + \overline{\boldsymbol{u}} = \boldsymbol{u} \quad \text{in } \Omega \times (0, T), \tag{3}$$
$$\nabla \cdot \overline{\boldsymbol{u}} = 0 \quad \text{in } \Omega \times (0, T), \tag{4}$$

endowed with suitable boundary conditions, e.g.,

$$\boldsymbol{u} = \overline{\boldsymbol{u}} = \boldsymbol{u}_{\text{in}} \quad \text{on } \Gamma_{\text{in}} \times (0, T), \tag{5}$$
$$\boldsymbol{u} = \overline{\boldsymbol{u}} = \boldsymbol{0} \quad \text{on } \Gamma_{\text{wall}} \times (0, T), \tag{6}$$
$$(\nu \nabla \boldsymbol{u} - pI) \cdot \boldsymbol{n} = (\alpha^2 \nabla \overline{\boldsymbol{u}} - \lambda I) \cdot \boldsymbol{n} = \boldsymbol{0} \quad \text{on } \Gamma_{\text{out}} \times (0, T), \tag{7}$$

and initial condition $\boldsymbol{u} = \boldsymbol{u}_0$ in $\Omega \times \{0\}$. In (1)–(7), $\boldsymbol{u}$ represents the fluid velocity (which is considered “averaged” in some sense), p the fluid pressure, $\nu > 0$ the kinematic viscosity, $\boldsymbol{f}$ a body force, and $\boldsymbol{u}_{\text{in}}$ and $\boldsymbol{u}_0$ are given. The equations (3), (4) represent the α-filter applied to $\boldsymbol{u}$, where $\overline{\boldsymbol{u}}$ is the resulting filtered variable and $\alpha > 0$ is the filtering radius. This is the radius of the neighborhood where the filter extracts information from the unresolved scales. The Lagrange multiplier λ is necessary to enforce a solenoidal $\overline{\boldsymbol{u}}$ in non-periodic flows. The inlet and outlet sections are denoted by Γ_{in} and Γ_{out}, while Γ_{wall} is the rest of the boundary. We note that the correct boundary conditions for $\overline{\boldsymbol{u}}$ on solid walls is unsettled in the LES community, although the computational experience of the authors is that a no-slip condition generally produces good results.

We also consider also the following generalized version of the Leray-α model, proposed in [8]:

$$\boldsymbol{u}_t + \widetilde{\boldsymbol{u}} \cdot \nabla \boldsymbol{u} + \nabla p - \nu \Delta \boldsymbol{u} = \boldsymbol{f} \quad \text{in } \Omega \times (0, T), \tag{8}$$
$$\nabla \cdot \boldsymbol{u} = 0 \quad \text{in } \Omega \times (0, T), \tag{9}$$
$$\nabla \lambda - \alpha^2 \nabla \cdot (a(\boldsymbol{u}) \nabla \widetilde{\boldsymbol{u}}) + \widetilde{\boldsymbol{u}} = \boldsymbol{u} \quad \text{in } \Omega \times (0, T), \tag{10}$$
$$\nabla \cdot \widetilde{\boldsymbol{u}} = 0 \quad \text{in } \Omega \times (0, T), \tag{11}$$

endowed with boundary conditions

$$\boldsymbol{u} = \widetilde{\boldsymbol{u}} = \boldsymbol{u}_{\text{in}} \quad \text{on } \Gamma_{\text{in}} \times (0, T), \tag{12}$$
$$\boldsymbol{u} = \widetilde{\boldsymbol{u}} = \boldsymbol{0} \quad \text{on } \Gamma_{\text{wall}} \times (0, T), \tag{13}$$
$$(\nu \nabla \boldsymbol{u} - pI) \cdot \boldsymbol{n} = (\alpha^2 a(\boldsymbol{u}) \nabla \widetilde{\boldsymbol{u}} - \lambda I) \cdot \boldsymbol{n} = \boldsymbol{0} \quad \text{on } \Gamma_{\text{out}} \times (0, T). \tag{14}$$

The scalar function $a(\boldsymbol{u})$, called the *indicator function*, is crucial for the success of model (8)–(11), and satisfies:

$$a(\boldsymbol{u}) \simeq 0 \quad \text{where the velocity } \boldsymbol{u} \text{ does not need regularization,}$$
$$a(\boldsymbol{u}) \simeq 1 \quad \text{where the velocity } \boldsymbol{u} \text{ does need regularization,}$$

so to detect the regions of the domain where regularization is needed. Notice that the choice $a(\boldsymbol{u}) = 1$ in (10), (11) corresponds to system (3), (4). In fact, in this way the operator in the filter equations is linear and constant in time. However, its effectivity is rather limited, since it introduces the same amount of regularization in every region of the domain, hence causing overdiffusion in those region where the flow is already smooth.

Different choices of $a(\cdot)$ have been proposed and compared in [7, 8, 28, 30]. Here, we focus on a class of deconvolution-based indicator functions:

$$a(\boldsymbol{u}) = a_D(\boldsymbol{u}) = |\boldsymbol{u} - D(F(\boldsymbol{u}))|^2 , \tag{15}$$

where F is a linear filter (an invertible, self-adjoint, compact operator from a Hilbert space to itself) and D is a bounded regularized approximation of F^{-1}. A popular choice for D is the Van Cittert deconvolution operator D_N, defined as

$$D_N = \sum_{n=0}^{N} (I - F)^n .$$

The evaluation of a_D with $D = D_N$ (deconvolution of order N) requires then to apply the filter F a total of $N + 1$ times. Since F^{-1} is not bounded, in practice N is chosen to be small, as the result of a trade-off between accuracy (for a regular solution) and filtering (for a non-regular one). In this paper, we consider $N = 0$, corresponding to $D_0 = I$. Numerical tests for $N = 1$ are considered for instance in [2].

We select F to be the linear Helmholtz filter operator F_H defined by

$$F = F_H = \left(I - \alpha^2 \Delta\right)^{-1} .$$

It is possible to prove [12] that

$$\boldsymbol{u} - D_N(F_H(\boldsymbol{u})) = (-1)^{N+1} \delta^{2N+2} \Delta^{N+1} F_H^{N+1} \boldsymbol{u}.$$

Therefore, $a_{D_N}(\boldsymbol{u})$ is close to zero in the regions of the domain where $\boldsymbol{u}$ is smooth. Let us set $\hat{\boldsymbol{u}} = F_H(\boldsymbol{u})$. With $D = D_0$ and $F = F_H$, the indicator function (15) reads

$$a_{D_0}(\boldsymbol{u}) = \left|\boldsymbol{u} - \hat{\boldsymbol{u}}\right|^2 . \tag{16}$$

System (8)–(11) with indicator function given by (16) is what we call Leray-α-NL.

2.1 Sensitivity Equation for Leray-α

Let us define

$$\boldsymbol{s} := \frac{\partial \boldsymbol{u}}{\partial \alpha},\ \boldsymbol{r} := \frac{\partial \overline{\boldsymbol{u}}}{\partial \alpha},\ \phi := \frac{\partial p}{\partial \alpha},\ \psi := \frac{\partial \lambda}{\partial \alpha}.$$

We develop the Leray-α sensitivity equation by differentiating model (1)–(4) with respect to α:

$$\boldsymbol{s}_t + \boldsymbol{r}\cdot\nabla\boldsymbol{u} + \overline{\boldsymbol{u}}\cdot\nabla\boldsymbol{s} + \nabla\phi - \nu\Delta\boldsymbol{s} = \boldsymbol{0} \qquad \text{in } \Omega\times(0,T), \tag{17}$$

$$\nabla\cdot\boldsymbol{s} = 0 \qquad \text{in } \Omega\times(0,T), \tag{18}$$

$$\nabla\psi - \alpha^2\Delta\boldsymbol{r} + \boldsymbol{r} - \boldsymbol{s} = 2\alpha\Delta\overline{\boldsymbol{u}} \quad \text{in } \Omega\times(0,T), \tag{19}$$

$$\nabla\cdot\boldsymbol{r} = 0 \qquad \text{in } \Omega\times(0,T). \tag{20}$$

System (17)–(20) is supplemented with boundary conditions:

$$\boldsymbol{s} = \boldsymbol{r} = \boldsymbol{0} \qquad \text{on } \Gamma_{\text{in}}\cup\Gamma_{\text{wall}}\times(0,T), \tag{21}$$

$$(\nu\nabla\boldsymbol{s} - \phi I)\cdot\boldsymbol{n} = \boldsymbol{0} \qquad \text{on } \Gamma_{\text{out}}\times(0,T), \tag{22}$$

$$(\alpha^2\nabla\boldsymbol{r} - \psi I)\cdot\boldsymbol{n} = -2\alpha\nabla\overline{\boldsymbol{u}} \quad \text{on } \Gamma_{\text{out}}\times(0,T), \tag{23}$$

and initial condition $\boldsymbol{s} = \boldsymbol{0}$ in $\Omega\times\{0\}$. It is important to note that $\overline{\boldsymbol{s}} \neq \boldsymbol{r}$, i.e. filtering does not commute with differentiation in α. In addition, for both $\boldsymbol{s}$ and $\boldsymbol{r}$ we have homogeneous Dirichlet conditions at the inlet section and on the walls.

Sensitivity system (17)–(20) is a new system of partial differential equations, and thus it is important to consider its well-posedness. Its similarity to NSE and Leray models limits our well-posedness study to the case of periodic boundary conditions. Although this setting is typically not physically meaningful, we argue that a lack of well posedness for (17)–(20) with periodic boundary conditions would prevent a successful analysis for physical conditions such as (21)–(23).

The following result is promptly deduced from [11].

Lemma 1 *Suppose $\alpha > 0$, $\boldsymbol{f} \in L^2(0,T;L^2(\Omega)^d)$ and $\boldsymbol{u}_0 \in H^1(\Omega)^d$. Then the Leray-α model (1)–(4) equipped with periodic boundary conditions has a unique weak solution with $\boldsymbol{u} \in L^\infty(0,T;H^1(\Omega)^d)\cap L^2(0,T;H^2(\Omega)^d)$.*

Using this lemma, we can prove that system (17)–(20) with periodic boundary conditions is well-posed.

Theorem 1 *Under the assumptions of Lemma 1, the system (17)–(20) has a unique weak solution satisfying **s**, $\boldsymbol{r} \in L^\infty(0,T;H^1(\Omega)^d)\cap L^2(0,T;H^2(\Omega)^d)$.*

Proof The proof of this theorem follows standard arguments, since the sensitivity system is linear, and the smoothness assumptions of the data yield a sufficiently smooth velocity $\boldsymbol{u}$ and filtered velocity $\overline{\boldsymbol{u}}$. □

2.2 Sensitivity Equation for Leray-α-NL

We define

$$\boldsymbol{s} := \frac{\partial \boldsymbol{u}}{\partial \alpha}, \quad \boldsymbol{r} := \frac{\partial \widetilde{\boldsymbol{u}}}{\partial \alpha}, \quad \boldsymbol{w} := \frac{\partial \hat{\boldsymbol{u}}}{\partial \alpha}, \quad \phi := \frac{\partial p}{\partial \alpha}, \quad \psi := \frac{\partial \lambda}{\partial \alpha}.$$

By differentiating the model (8)–(11) [with indicator function given by (16)] with respect to α, we obtain:

$$\boldsymbol{s}_t + \boldsymbol{r} \cdot \nabla \boldsymbol{u} + \widetilde{\boldsymbol{u}} \cdot \nabla \boldsymbol{s} + \nabla \phi - \nu \Delta \boldsymbol{s} = \boldsymbol{0} \qquad \text{in } \Omega \times (0, T), \tag{24}$$

$$\nabla \cdot \boldsymbol{s} = 0 \qquad \text{in } \Omega \times (0, T), \tag{25}$$

$$\begin{aligned}\nabla \psi - \alpha^2 \nabla \cdot [(2(\boldsymbol{u} - \hat{\boldsymbol{u}}) \cdot (\boldsymbol{s} - \boldsymbol{w})) \nabla \widetilde{\boldsymbol{u}} + |\boldsymbol{u} - \hat{\boldsymbol{u}}|^2 \nabla \boldsymbol{r}] + \boldsymbol{r} - \boldsymbol{s} \\ = 2\alpha \nabla \cdot |\boldsymbol{u} - \hat{\boldsymbol{u}}|^2 \nabla \widetilde{\boldsymbol{u}} \quad \text{in } \Omega \times (0, T),\end{aligned} \tag{26}$$

$$\nabla \cdot \boldsymbol{r} = 0 \qquad \text{in } \Omega \times (0, T), \tag{27}$$

$$-\alpha^2 \Delta \boldsymbol{w} + \boldsymbol{w} - \boldsymbol{s} = 2\alpha \Delta \hat{\boldsymbol{u}} \qquad \text{in } \Omega \times (0, T). \tag{28}$$

The latter equation follows from the fact that $\hat{\boldsymbol{u}} = F_H(\boldsymbol{u}) \Rightarrow \hat{\boldsymbol{u}} - \alpha^2 \Delta \hat{\boldsymbol{u}} = \boldsymbol{u}$. System (24)–(28) is supplemented with boundary conditions

$$\begin{aligned}
\boldsymbol{s} = \boldsymbol{r} = \boldsymbol{w} = \boldsymbol{0} \qquad & \text{on } \Gamma_{\text{in}} \cup \Gamma_{\text{wall}} \times (0, T), \\
(\nu \nabla \boldsymbol{s} - \phi I) \cdot \boldsymbol{n} = (\alpha^2 \nabla \boldsymbol{w}) \cdot \boldsymbol{n} = \boldsymbol{0} \qquad & \text{on } \Gamma_{\text{out}} \times (0, T), \\
(\alpha^2 [(2(\boldsymbol{u} - \hat{\boldsymbol{u}}) \cdot (\boldsymbol{s} - \boldsymbol{w})) \nabla \widetilde{\boldsymbol{u}} + |\boldsymbol{u} - \hat{\boldsymbol{u}}|^2 \nabla \boldsymbol{r}] - \psi I) \cdot \boldsymbol{n} & \\
= -2\alpha \nabla \cdot |\boldsymbol{u} - \hat{\boldsymbol{u}}|^2 \nabla \widetilde{\boldsymbol{u}} \qquad & \text{on } \Gamma_{\text{out}} \times (0, T),
\end{aligned}$$

and initial condition $\boldsymbol{s} = \boldsymbol{0}$ in $\Omega \times \{0\}$.

For the Leray-α-NL sensitivity system (24)–(28), we are not able to establish a well-posedness result. This is due to the fact that the well-posedness of Leray-α-NL has not been proven yet. The major difficulty is the nonlinear filter, which would not provide the extra regularity of $\widetilde{\boldsymbol{u}}$ from the regularity of $\boldsymbol{u}$, since $\boldsymbol{u} - \hat{\boldsymbol{u}}$ could be zero. Hence we would need to apply different techniques from the ones used for the classical Leray-α model. We leave this study for a separate work. For now, we conjecture that Leray-α-NL, and its associated sensitivity system, is well-posed for sufficiently smooth data.

3 Discrete Schemes for the Leray-α and Leray-α-NL Models and Associated Sensitivity Systems

Let $\Delta t > 0, t^n = n\Delta t$, with $n = 0, ..., M$ and $T = M\Delta t$. Moreover, we denote by y^n the approximation of a generic quantity y at the time t^n. For the time discretization, we adopt Backward Differentiation Formula of order 2 (BDF2, see, e.g., [35]).

We assume $\mathscr{T}_h$ to be a regular, conforming triangulation (tetrahedralization), with maximum element diameter h. The velocity and pressure finite element spaces $(X^h, Q^h) \subset (H^1(\Omega)^d, L^2(\Omega))$ are assumed to be LBB stable, i.e. it holds that

$$\inf_{q_h\in Q_h} \sup_{\boldsymbol{v}_h\in X_h} \frac{(\nabla \cdot \boldsymbol{v}_h, q_h)}{\|\nabla \boldsymbol{v}_h\| \|q_h\|} \geq \beta,$$

with β independent of h. Taylor-Hood elements (P_k, P_{k-1}) with $k \geq 2$ on triangles and tetrahedra are popular examples of LBB stable pairs [9, 26]. The usual modifications of these spaces can be made when non-homogeneous Dirichlet boundary conditions are imposed on the velocity.

Finally, we introduce the skew-symmetric form of the nonlinear term in the NSE is given by

$$b^*(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}) := \frac{1}{2}(\boldsymbol{u} \cdot \nabla \boldsymbol{v}, \boldsymbol{w}) - \frac{1}{2}(\boldsymbol{u} \cdot \nabla \boldsymbol{w}, \boldsymbol{v}), \quad \text{with } \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w} \in H^1(\Omega)^d.$$

If $\nabla \cdot \boldsymbol{u} = 0$, then $b^*(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}) = (\boldsymbol{u} \cdot \nabla \boldsymbol{v}, \boldsymbol{w})$. An important property of this operator is that $b^*(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{v}) = 0$ even if $\nabla \cdot \boldsymbol{u} \neq 0$, which can occur in discretizations.

For simplicity, when analyzing the discrete schemes we will consider wall-bounded flows, i.e. homogeneous Dirichlet conditions on all the boundary. The analyses that follow can be promptly adapted to fit the case of other boundary conditions.

Remark 1 The use of the skew-symmetric form of the nonlinearity is for analysis purposes only, and in our computations we use the usual convective formulation. In general, on sufficiently fine discretizations, very little difference between solutions from these formulations is observed. In practice, particularly in the case of zero traction outflow boundary conditions, the usual convective form is much more commonly used (since the skew-symmetric form requires a nonlinear boundary integral be incorporated into the formulation).

3.1 Discrete Scheme for Leray-α

Given $T, \Delta t, \alpha > 0$, $\boldsymbol{f} \in L^\infty(0, T; H^{-1}(\Omega)^d)$, and $\boldsymbol{u}_h^0, \boldsymbol{u}_h^1 \in X_h$, we propose the following decoupled finite element discretization for the Leray-α model (1)–(4) with an implicit-explicit (also called semi-implicit) treatment of the nonlinear term:

Algorithm 1

For $n = 1, \ldots, M-1$, given $\boldsymbol{u}_h^n, \boldsymbol{u}_h^{n-1}, \overline{\boldsymbol{u}}_h^n, \overline{\boldsymbol{u}}_h^{n-1} \in X_h$ find $\boldsymbol{u}_h^{n+1}, \overline{\boldsymbol{u}}_h^{n+1} \in X_h$ and $p_h^{n+1}, \lambda_h^{n+1} \in Q_h$ satisfying:

$$\begin{aligned} \frac{1}{2\Delta t}\left(3\boldsymbol{u}_h^{n+1} - 4\boldsymbol{u}_h^n + \boldsymbol{u}_h^{n-1}, \boldsymbol{v}_h\right) + b^*(2\overline{\boldsymbol{u}}_h^n - \overline{\boldsymbol{u}}_h^{n-1}, \boldsymbol{u}_h^{n+1}, \boldsymbol{v}_h) \\ - \left(p_h^{n+1}, \nabla\cdot\boldsymbol{v}_h\right) + \nu\left(\nabla\boldsymbol{u}_h^{n+1}, \nabla\boldsymbol{v}_h\right) = (\boldsymbol{f}(t^{n+1}), \boldsymbol{v}_h), \end{aligned} \tag{29}$$

$$\left(\nabla\cdot\boldsymbol{u}_h^{n+1}, q_h\right) = 0, \tag{30}$$

$$-(\lambda_h, \nabla\cdot\boldsymbol{z}_h) + \alpha^2(\nabla\overline{\boldsymbol{u}}_h^{n+1}, \nabla\boldsymbol{z}_h) + (\overline{\boldsymbol{u}}_h^{n+1}, \boldsymbol{z}_h) = (\boldsymbol{u}_h^{n+1}, \boldsymbol{z}_h), \tag{31}$$

$$(\nabla\cdot\overline{\boldsymbol{u}}_h^{n+1}, \eta_h) = 0, \tag{32}$$

for every $\boldsymbol{v}_h, \boldsymbol{z}_h \in X_h$ and $q_h, \eta_h \in \times Q_h$.

Algorithm 1 decouples the filtering from the mass/momentum system. It is a straightforward extension of the analysis in [6] (for a linearized Crank-Nicolson temporal discretization with inf-sup stable finite elements) to prove that Algorithm 1 is unconditionally stable with respect to the time step size:

$$\|\boldsymbol{u}_h^M\|^2 + \nu\Delta t\sum_{n=2}^{M}\|\nabla\boldsymbol{u}_h^n\|^2 \le C(\boldsymbol{u}_h^0, \boldsymbol{u}_h^1, \nu^{-1}, \boldsymbol{f}, \Omega).$$

Moreover, it converges optimally (under the usual smoothness assumptions) to the Leray-α solution in the following sense: if Taylor-Hood elements are used, then

$$\|\boldsymbol{u}(T) - \boldsymbol{u}_h^M\|^2 + \nu\Delta t\sum_{n=2}^{M}\|\nabla(\boldsymbol{u}(t^n) - \boldsymbol{u}_h^n)\|^2 \le C(\Delta t^4 + h^{2k}). \tag{33}$$

We propose an analogous algorithm for the sensitivity system. At each time step, after solving the Leray-α discrete system we approximate the solution of sensitivity equation (17)–(20) as follows. We take $\boldsymbol{s}_h^0 = \boldsymbol{s}_h^1 = \boldsymbol{0}$. For $n = 1, \ldots, M-1$, given $\boldsymbol{s}_h^n, \boldsymbol{s}_h^{n-1}, \boldsymbol{r}_h^n, \overline{\boldsymbol{r}}_h^{n-1} \in X_h$ we find $\boldsymbol{s}_h^{n+1}, \boldsymbol{r}_h^{n+1} \in X_h$ and $\phi_h^{n+1}, \psi_h^{n+1} \in Q_h$ satisfying

$$\begin{aligned} \frac{1}{2\Delta t}(3\boldsymbol{s}_h^{n+1} - 4\boldsymbol{s}_h^n + \boldsymbol{s}_h^{n-1}, \boldsymbol{v}_h) + b^*(2\overline{\boldsymbol{u}}_h^n - \overline{\boldsymbol{u}}_h^{n-1}, \boldsymbol{s}_h^{n+1}, \boldsymbol{v}_h) \\ - \left(\phi_h^{n+1}, \nabla\cdot\boldsymbol{v}_h\right) + \nu\left(\nabla\boldsymbol{s}_h^{n+1}, \nabla\boldsymbol{v}_h\right) = -b^*(2\boldsymbol{r}_h^n - \boldsymbol{r}_h^{n-1}, \boldsymbol{u}_h^{n+1}, \boldsymbol{v}_h), \end{aligned} \tag{34}$$

$$\left(\nabla\cdot\boldsymbol{s}_h^{n+1}, q_h\right) = 0, \tag{35}$$

$$-(\psi_h^{n+1}, \nabla\cdot\boldsymbol{z}_h) + \alpha^2\left(\nabla\boldsymbol{r}_h^{n+1}, \nabla\boldsymbol{z}_h\right) + \left(\boldsymbol{r}_h^{n+1}, \boldsymbol{z}_h\right) = (\boldsymbol{s}_h^{n+1}, \boldsymbol{z}_h) - 2\alpha(\nabla\overline{\boldsymbol{u}}_h^{n+1}, \nabla\boldsymbol{z}_h), \tag{36}$$

$$\left(\nabla\cdot\boldsymbol{r}_h^{n+1}, \eta_h\right) = 0, \tag{37}$$

for all $\boldsymbol{v}_h, \boldsymbol{z}_h \in X_h$ and $q_h, \eta_h \in \times Q_h$.

Remark 2 The discrete sensitivity system (34)–(37) can be solved efficiently. In fact, system (34), (35) is decoupled from (36), (37). Furthermore, at each time step the linear system arising from (34)–(37) has exactly the same matrix as system (29)–(32) allowing for the reusing of the preconditioner.

The following lemma proves that the discrete sensitivity system for the Leray-α model is stable with respect to the time step size under a mild restriction on the mesh size relative to the time step.

Lemma 2 *The discrete sensitivity system (34)–(37) with* $(X_h, Q_h) = (P_k, P_{k-1})$ *is stable provided the mesh size h and time step* Δt *are chosen to satisfy* $\Delta t^3 \leq ch \leq \Delta t^{\frac{1}{2k-2}}$. *Then we have*

$$\|\boldsymbol{s}_h^M\|^2 + \nu \Delta t \sum_{n=2}^{M} \|\nabla \boldsymbol{s}_h^{n+1}\|^2 \leq C,$$

where C depends only on the problem data.

Proof We take $\boldsymbol{v}_h = \boldsymbol{s}_h^{n+1}$ and $q_h = \phi_h^{n+1}$ in (34)–(35) and get that

$$\tfrac{1}{2\Delta t}\left(\|\boldsymbol{s}_h^{n+1}\|^2 - \|\boldsymbol{s}_h^n\|^2 + \|2\boldsymbol{s}_h^{n+1} - \boldsymbol{s}_h^n\|^2 - \|2\boldsymbol{s}_h^n - \boldsymbol{s}_h^{n-1}\|^2 + \|\boldsymbol{s}_h^{n+1} - 2\boldsymbol{s}_h^n + \boldsymbol{s}_h^{n-1}\|^2\right)$$
$$+\nu\|\nabla \boldsymbol{s}_h^{n+1}\|^2 = -b^*(2\boldsymbol{r}_h^n - \boldsymbol{r}_h^{n-1}, \boldsymbol{u}_h^{n+1}, \boldsymbol{s}_h^{n+1}).$$

Let $\boldsymbol{e}_{\boldsymbol{u}}^{n+1} := \boldsymbol{u}_h^{n+1} - \boldsymbol{u}(t^{n+1})$. The right-hand side term is handled by first adding and subtracting the true solution $\boldsymbol{u}(t^{n+1})$ to $\boldsymbol{u}_h^{n+1}$, then using Holder's inequality, and Sobolev embeddings to obtain

$$\begin{aligned}
&|b^*(2\boldsymbol{r}_h^n - \boldsymbol{r}_h^{n-1}, \boldsymbol{u}_h^{n+1}, \boldsymbol{s}_h^{n+1})| \\
&\leq |b^*(2\boldsymbol{r}_h^n - \boldsymbol{r}_h^{n-1}, \boldsymbol{u}(t^{n+1}), \boldsymbol{s}_h^{n+1})| + |b^*(2\boldsymbol{r}_h^n - \boldsymbol{r}_h^{n-1}, \boldsymbol{e}_{\boldsymbol{u}}^{n+1}, \boldsymbol{s}_h^{n+1})| \\
&\leq C\|2\boldsymbol{r}_h^n - \boldsymbol{r}_h^{n-1}\| \left(\|\boldsymbol{u}(t^{n+1})\|_{L^\infty} + \|\nabla \boldsymbol{u}(t^{n+1})\|_{L^3}\right) \|\nabla \boldsymbol{s}_h^{n+1}\| \\
&\quad + C\|2\boldsymbol{r}_h^n - \boldsymbol{r}_h^{n-1}\| \left(\|\boldsymbol{e}_{\boldsymbol{u}}^{n+1}\|_{L^\infty} + \|\nabla \boldsymbol{e}_{\boldsymbol{u}}^{n+1}\|_{L^3}\right) \|\nabla \boldsymbol{s}_h^{n+1}\|.
\end{aligned}$$

By the assumed smoothness of the true solution, the first term is bounded by

$$\begin{aligned}
&C\|2\boldsymbol{r}_h^n - \boldsymbol{r}_h^{n-1}\| \left(\|\boldsymbol{u}(t^{n+1})\|_{L^\infty} + \|\nabla \boldsymbol{u}(t^{n+1})\|_{L^3}\right) \|\nabla \boldsymbol{s}_h^{n+1}\| \\
&\qquad \leq \frac{\nu}{2}\|\nabla \boldsymbol{s}_h^{n+1}\|^2 + C\nu^{-1}\|2\boldsymbol{r}_h^n - \boldsymbol{r}_h^{n-1}\|^2.
\end{aligned}$$

Thanks to the generalized inverse inequality (see, e.g. [9]), and well-known interpolation theory, we bound the second term as

$$\begin{aligned}C\|2\boldsymbol{r}_h^n - \boldsymbol{r}_h^{n-1}\|\Big(\|\boldsymbol{e}_{\boldsymbol{u}}^{n+1}\|_{L^\infty} + \|\nabla \boldsymbol{e}_{\boldsymbol{u}}^{n+1}\|_{L^3}\Big)\|\nabla \boldsymbol{s}_h^{n+1}\| \\ \le \frac{\nu}{2}\|\nabla \boldsymbol{s}_h^{n+1}\|^2 + C\nu^{-1}h^{-1}\|\nabla \boldsymbol{e}_{\boldsymbol{u}}^{n+1}\|^2\|2\boldsymbol{r}_h^n - \boldsymbol{r}_h^{n-1}\|^2.\end{aligned}$$

Combining the bounds and summing over n yields

$$\begin{aligned}\|\boldsymbol{s}_h^M\|^2 + \|2\boldsymbol{s}_h^M - \boldsymbol{s}_h^{M-1}\|^2 + \nu\Delta t\sum_{n=2}^{M}\|\nabla \boldsymbol{s}_h^{n+1}\|^2 \\ \le C\nu^{-1}\Delta t\sum_{n=2}^{M-1}\|2\boldsymbol{r}_h^n - \boldsymbol{r}_h^{n-1}\|^2\left(1 + h^{-1}\|\nabla \boldsymbol{e}_{\boldsymbol{u}}^{n+1}\|^2\right).\end{aligned} \tag{38}$$

Next, we use Eqs. (36) and (37) along with Cauchy-Schwarz and Young's inequalities to reveal

$$\alpha^2\|\nabla(2\boldsymbol{r}_h^n - \boldsymbol{r}_h^{n-1})\|^2 + \|2\boldsymbol{r}_h^n - \boldsymbol{r}_h^{n-1}\|^2 \le \|2\boldsymbol{s}_h^n - \boldsymbol{s}_h^{n-1}\|^2 + 4\|\nabla\left(2\overline{\boldsymbol{u}}_h^n - \overline{\boldsymbol{u}}_h^{n-1}\right)\|^2.$$

Combining this with (38) gives

$$\begin{aligned}&\|\boldsymbol{s}_h^M\|^2 + \|2\boldsymbol{s}_h^M - \boldsymbol{s}_h^{M-1}\|^2 + \nu\Delta t\sum_{n=2}^{M}\|\nabla \boldsymbol{s}_h^{n+1}\|^2 \\ &\le C\nu^{-1}\Delta t\sum_{n=2}^{M-1}\|2\boldsymbol{s}_h^n - \boldsymbol{s}_h^{n-1}\|^2\left(1 + h^{-1}\|\nabla \boldsymbol{e}_u^{n+1}\|^2\right) \\ &\quad + C\nu^{-1}\Delta t\sum_{n=2}^{M-1}\|\nabla\left(2\overline{\boldsymbol{u}}_h^n - \overline{\boldsymbol{u}}_h^{n-1}\right)\|^2\left(1 + h^{-1}\|\nabla \boldsymbol{e}_u^{n+1}\|^2\right).\end{aligned} \tag{39}$$

Using the convergence result (33), we have that

$$h^{-1}\|\nabla \boldsymbol{e}_{\boldsymbol{u}}^{n+1}\|^2 \le C\Delta t^{-1}h^{-1}\left(\Delta t^4 + h^{2k}\right) = C\left(\frac{\Delta t^3}{h} + \frac{h^{2k-1}}{\Delta t}\right).$$

Inserting this bound into (39) and applying the discrete Gronwall inequality (see e.g. [24]) gives the stated result. We note that there is no $\boldsymbol{s}_h^M$ on the right-hand side. Thus there is no time step restriction associated with the discrete Gronwall inequality. □

3.2 Discrete Scheme for Leray-α-NL

Given $T, \Delta t, \alpha > 0$, $\boldsymbol{f} \in L^\infty(0, T; H^{-1}(\Omega)^d)$, and $\boldsymbol{u}_h^0, \boldsymbol{u}_h^1 \in X_h$, we propose the following decoupled finite element discretization for the Leray-α-NL model (8)–(11) with indicator function given by (16) and an implicit-explicit treatment of the nonlinear term:

Algorithm 2

For $n = 1, \ldots, M-1$ find $\boldsymbol{u}_h^n, \widetilde{\boldsymbol{u}}_h^n, \hat{\boldsymbol{u}}_h^n \in X_h$ and $p_h^n, \lambda_h^n \in Q_h$ satisfying:

$$\frac{1}{2\Delta t}\left(3\boldsymbol{u}_h^{n+1} - 4\boldsymbol{u}_h^n + \boldsymbol{u}_h^{n-1}, \boldsymbol{v}_h\right) + b^*(2\widetilde{\boldsymbol{u}}_h^n - \widetilde{\boldsymbol{u}}_h^{n-1}, \boldsymbol{u}_h^{n+1}, \boldsymbol{v}_h) - \left(p_h^{n+1}, \nabla \cdot \boldsymbol{v}_h\right) + \nu\left(\nabla \boldsymbol{u}_h^{n+1}, \nabla \boldsymbol{v}_h\right) = (\boldsymbol{f}(t^{n+1}), \boldsymbol{v}_h), \tag{40}$$

$$\left(\nabla \cdot \boldsymbol{u}_h^{n+1}, q_h\right) = 0, \tag{41}$$

$$-(\lambda_h, \nabla \cdot \boldsymbol{z}_h) + \alpha^2(|\boldsymbol{u}_h^{n+1} - \hat{\boldsymbol{u}}_h^{n+1}|^2 \nabla \widetilde{\boldsymbol{u}}_h^{n+1}, \nabla \boldsymbol{z}_h) + (\widetilde{\boldsymbol{u}}_h^{n+1}, \boldsymbol{z}_h) = (\boldsymbol{u}_h^{n+1}, \boldsymbol{z}_h), \tag{42}$$

$$(\nabla \cdot \widetilde{\boldsymbol{u}}_h^{n+1}, \eta_h) = 0, \tag{43}$$

$$\alpha^2(\nabla \hat{\boldsymbol{u}}_h^{n+1}, \nabla \boldsymbol{y}_h) + (\hat{\boldsymbol{u}}_h^{n+1}, \boldsymbol{y}_h) = (\boldsymbol{u}_h^{n+1}, \boldsymbol{y}_h), \tag{44}$$

for every $\boldsymbol{v}_h, \boldsymbol{z}_h, \boldsymbol{y}_h \in X_h$ and $q_h, \eta_h \in Q_h$.

Note that the $\hat{\boldsymbol{u}}_h^n$, $\widetilde{\boldsymbol{u}}_h^n$ velocities for $n = 0, 1$ can be determined from Eqs. (42)–(44), since $\boldsymbol{u}_h^0$ and $\boldsymbol{u}_h^1$ are given.

Algorithm 2 efficiently decouples the mass/momentum system from the two filters. First, system (40), (41) is solved for $\boldsymbol{u}_h^{n+1}$, p_h^{n+1}, then Eq. (44) is solved for $\hat{\boldsymbol{u}}_h^{n+1}$, and finally Eqs. (42)–(43) are solved for $\widetilde{\boldsymbol{u}}_h^{n+1}$, λ_h^{n+1}.

Algorithm 2 was studied in [7] with the only difference that the $|\boldsymbol{u}_h^{n+1} - \hat{\boldsymbol{u}}_h^{n+1}|$ in (42) term was not squared. This change does not affect the stability result proven in [7], which states

$$\|\boldsymbol{u}_h^M\|^2 + \nu\Delta t \sum_{n=2}^{M} \|\nabla \boldsymbol{u}_h^n\|^2 \le C(\boldsymbol{u}_h^0, \boldsymbol{u}_h^1, \nu^{-1}, f, \Omega),$$

$$\|\hat{\boldsymbol{u}}_h^n\| \le \|\boldsymbol{u}_h^n\|, \quad \|\nabla \hat{\boldsymbol{u}}_h^n\| \le \|\nabla \boldsymbol{u}_h^n\|, \quad \|\widetilde{\boldsymbol{u}}_h^n\| \le \|\boldsymbol{u}_h^n\| \quad \text{for } 0 \le n \le M.$$

It is known from [6, 8] that the scheme (40)–(44) converges to a smooth Navier-Stokes solution $\boldsymbol{u}_{NSE}$ as h, Δt, and α tends to 0. If Taylor-Hood elements are used, we have

$$\|\boldsymbol{u}_{NSE}(T) - \boldsymbol{u}_h^M\|^2 + \nu\Delta t \sum_{n=2}^{M} \|\nabla(\boldsymbol{u}_{NSE}(t^n) - \boldsymbol{u}_h^n)\|^2 \le C(\Delta t^4 + h^{2k} + \alpha^4).$$

At each time step, after solving the Leray-α-NL discrete system we approximate the solution of sensitivity equation (24)–(27) as follows. We take $\boldsymbol{s}_h^0 = \boldsymbol{s}_h^1 = \boldsymbol{0}$. For $n = 1, \dots, M-1$, given $\boldsymbol{s}_h^n, \boldsymbol{s}_h^{n-1}, \boldsymbol{r}_h^n, \overline{\boldsymbol{r}}_h^{n-1} \in X_h$ we find $\boldsymbol{s}_h^{n+1}, \boldsymbol{r}_h^{n+1}, \boldsymbol{w}_h^{n+1} \in X_h$ and $\phi_h^{n+1}, \psi_h^{n+1} \in Q_h$ satisfying:

$$\begin{aligned}
\frac{1}{2\Delta t}(3\boldsymbol{s}_h^{n+1} - 4\boldsymbol{s}_h^n + \boldsymbol{s}_h^{n-1}, \boldsymbol{v}_h) + b^*(2\widetilde{\boldsymbol{u}}_h^n - \widetilde{\boldsymbol{u}}_h^{n-1}, \boldsymbol{s}_h^{n+1}, \boldsymbol{v}_h) - \left(\phi_h^{n+1}, \nabla\cdot\boldsymbol{v}_h\right)& \\
+ \nu\left(\nabla\boldsymbol{s}_h^{n+1}, \nabla\boldsymbol{v}_h\right) = -b^*(2\boldsymbol{r}_h^n - \boldsymbol{r}_h^{n-1}, \boldsymbol{u}_h^{n+1}, \boldsymbol{v}_h),& \qquad (45)\\
\left(\nabla\cdot\boldsymbol{s}_h^{n+1}, q_h\right) = 0,& \qquad (46)\\
\alpha^2\left(|\boldsymbol{u}_h^{n+1} - \hat{\boldsymbol{u}}_h^{n+1}|^2\nabla\boldsymbol{r}_h^{n+1}, \nabla z_h\right) - (\psi_h^{n+1}, \nabla\cdot z_h) + \left(\boldsymbol{r}_h^{n+1}, z_h\right)& \\
= -2\alpha^2\left(\left((\boldsymbol{u}_h^{n+1} - \hat{\boldsymbol{u}}_h^{n+1})\cdot(\boldsymbol{s}_h^{n+1} - \boldsymbol{w}_h^{n+1})\right)\nabla\widetilde{\boldsymbol{u}}_h^{n+1}, \nabla z_h\right)& \\
+ (\boldsymbol{s}_h^{n+1}, z_h) - 2\alpha(|u_h^{n+1} - \hat{\boldsymbol{u}}_h^{n+1}|^2\nabla\widetilde{\boldsymbol{u}}_h^{n+1}, \nabla z_h),& \qquad (47)\\
\left(\nabla\cdot\boldsymbol{r}_h^{n+1}, \eta_h\right) = 0,& \qquad (48)\\
\alpha^2(\nabla\boldsymbol{w}_h^{n+1}, \nabla\boldsymbol{y}_h) + (\boldsymbol{w}_h^{n+1}, \boldsymbol{y}_h) = (\boldsymbol{s}_h^{n+1}, \boldsymbol{y}_h),& \qquad (49)
\end{aligned}$$

for all $\boldsymbol{v}_h, z_h, \boldsymbol{y}_h \in X_h$ and $q_h, \eta_h \in Q_h$.

This scheme can also be efficiently computed. In fact, system (45), (46) is computed first, followed by system (49) and (47), (48). Moreover, the matrices for the linear systems are exactly the same as for (40)–(44).

Theorem 2 *The discrete sensitivity scheme is stable (45)–(49): for all $\Delta t > 0$ we have*

$$\|\boldsymbol{s}_h^M\|^2 + \nu\Delta t\sum_{n=1}^{M}\|\nabla\boldsymbol{s}_h^n\|^2 \le C(\boldsymbol{u}, \nu^{-1}, T),$$

and for any n

$$2\alpha^2\|\nabla\boldsymbol{w}_h^n\|^2 + \|\boldsymbol{w}_h^n\|^2 \le \|\boldsymbol{s}_h^n\|^2, \quad \|\boldsymbol{r}_h^n\| \le \|\boldsymbol{s}_h^n\|.$$

Proof By taking $\boldsymbol{v}_h = \boldsymbol{s}_h^{n+1}$ in (45) and $q_h = \phi_h^{n+1}$ in (46) along with Holder's inequality and Sobolev embedding theorems, we get

$$\begin{aligned}
\frac{1}{4\Delta t}\left(\|\boldsymbol{s}_h^{n+1}\|^2 - \|\boldsymbol{s}_h^n\|^2 + \|2\boldsymbol{s}_h^{n+1} - \boldsymbol{s}_h^n\|^2 - \|2\boldsymbol{s}_h^n - \boldsymbol{s}_h^{n-1}\|^2 + \|\boldsymbol{s}_h^{n+1} - 2\boldsymbol{s}_h^n + \boldsymbol{s}_h^{n-1}\|^2\right)& \\
\le +\nu\|\nabla\boldsymbol{s}_h^{n+1}\|^2 C\|2\boldsymbol{r}_h^n - \boldsymbol{r}_h^{n-1}\|(\|\nabla\boldsymbol{u}_h^{n+1}\|_{L^3} + \|\boldsymbol{u}_h^{n+1}\|_{L^\infty})\|\nabla\boldsymbol{s}_h^{n+1}\|.&
\end{aligned}$$

Young's inequality and the assumption of $\boldsymbol{u}_h$ converging sufficiently fast yield:

$$\begin{aligned}
\frac{1}{2\Delta t}(\|\boldsymbol{s}_h^{n+1}\|^2 - \|\boldsymbol{s}_h^n\|^2 + \|2\boldsymbol{s}_h^{n+1} - \boldsymbol{s}_h^n\|^2 - \|2\boldsymbol{s}_h^n - \boldsymbol{s}_h^{n-1}\|^2 + \|\boldsymbol{s}_h^{n+1} - 2\boldsymbol{s}_h^n + \boldsymbol{s}_h^{n-1}\|^2)& \\
\le +\nu\|\nabla\boldsymbol{s}_h^{n+1}\|^2 C\nu^{-1}\|2\boldsymbol{r}_h^n - \boldsymbol{r}_h^{n-1}\|^2.& \qquad (50)
\end{aligned}$$

Next, taking $\boldsymbol{y}_h = \boldsymbol{w}_h^{n+1}$ in (49) and $\boldsymbol{z}_h = \boldsymbol{r}_h^{n+1}$ in (47) provides

$$2\alpha^2 \|\nabla \boldsymbol{w}_h^{n+1}\|^2 + \|\boldsymbol{w}_h^{n+1}\|^2 \leq \|\boldsymbol{s}_h^{n+1}\|^2, \tag{51}$$

and

$$\begin{aligned}
&\alpha^2 \Big\| |\boldsymbol{u}_h^{n+1} - \hat{\boldsymbol{u}}_h^{n+1}| \nabla \boldsymbol{r}_h^{n+1} \Big\|^2 + \|\boldsymbol{r}_h^{n+1}\|^2 \\
&\quad = (\boldsymbol{s}_h^{n+1}, \boldsymbol{r}_h^{n+1}) - 2\alpha^2 \Big(\big((\boldsymbol{u}_h^{n+1} - \hat{\boldsymbol{u}}_h^{n+1}) \cdot (\boldsymbol{s}_h^{n+1} - \boldsymbol{w}_h^{n+1}) \big) \nabla \widetilde{\boldsymbol{u}}_h^{n+1}, \nabla \boldsymbol{r}_h^{n+1} \Big) \\
&\qquad\qquad - 2\alpha (|\boldsymbol{u}_h^{n+1} - \hat{\boldsymbol{u}}_h^{n+1}|^2 \nabla \widetilde{\boldsymbol{u}}_h^{n+1}, \nabla \boldsymbol{r}_h^{n+1}).
\end{aligned}$$

Cauchy-Schwarz and Young's inequalities applied to each term on the right-hand side give the estimate

$$\begin{aligned}
&\alpha^2 \| |\boldsymbol{u}_h^{n+1} - \hat{\boldsymbol{u}}_h^{n+1}| \nabla \boldsymbol{r}_h^{n+1} \|^2 + \|\boldsymbol{r}_h^{n+1}\|^2 \\
&\quad \leq \|\boldsymbol{s}_h^{n+1}\|^2 + 8\alpha^2 \| |\boldsymbol{s}_h^{n+1} - \boldsymbol{w}_h^{n+1}| \nabla \widetilde{\boldsymbol{u}}_h^{n+1} \|^2 + 8 \| |\boldsymbol{u}_h^{n+1} - \hat{\boldsymbol{u}}_h^{n+1}| \nabla \widetilde{\boldsymbol{u}}_h^{n+1} \|^2.
\end{aligned}$$

Assuming $\boldsymbol{u}_h$ converges and using (51), we have that

$$\|\boldsymbol{r}_h^{n+1}\|^2 \leq \|\boldsymbol{s}_h^{n+1}\|^2 + C + C\alpha^2 \| |\boldsymbol{s}_h^{n+1} - \boldsymbol{w}_h^{n+1} \|^2 \leq C(1 + \|\boldsymbol{s}_h^{n+1}\|^2). \tag{52}$$

Inequalities (52) in (50) yield:

$$\begin{aligned}
&\tfrac{1}{2\Delta t} (\|\boldsymbol{s}_h^{n+1}\|^2 - \|\boldsymbol{s}_h^n\|^2 + \|2\boldsymbol{s}_h^{n+1} - \boldsymbol{s}_h^n\|^2 - \|2\boldsymbol{s}_h^n - \boldsymbol{s}_h^{n-1}\|^2 + \|\boldsymbol{s}_h^{n+1} - 2\boldsymbol{s}_h^n + \boldsymbol{s}_h^{n-1}\|^2) \\
&\qquad + \nu \|\nabla \boldsymbol{s}_h^{n+1}\|^2 \leq C\nu^{-1} \left(\|\boldsymbol{s}_h^n\|^2 + \|\boldsymbol{s}_h^{n-1}\|^2 \right).
\end{aligned}$$

To complete the proof we sum over n and apply Gronwall's inequality. There is no time step restriction since the power of $\boldsymbol{s}_h$ on the right-hand side is less than $n + 1$ [24]. □

4 Numerical Testing

In this section we compute solutions to the Leray-α and Leray-α-NL models, and associated sensitivities for two test problems. For both tests, we use Taylor-Hood elements, i.e. P_2 elements for velocities and relative sensitivities, and P_1 elements for pressures and Lagrange multipliers. The computations were performed using Freefem software [22].

4.1 Channel Flow Past a Forward-Backward Step

We consider the two dimensional channel flow past a forward-backward step. The domain is a 40×10 rectangle, with a 1×1 step placed five units in, see Fig. 1. We impose boundary conditions (5)–(7) for the Leray-α model and (12)–(14) for the Leray-α-NL model with $\boldsymbol{u}_{\text{in}} = (y(10-y)/25, 0)^T$. The boundary conditions for the sensitivity systems are as reported in Sect. 2. We set $\boldsymbol{f} = \boldsymbol{0}$ and $\nu = 1/600$. The correct physical behavior for a NSE solution is a smooth velocity profile, with eddies forming and detaching behind the step; see, e.g., [20, 27].

We consider a Delaunay triangulated mesh (shown in Fig. 1), with a total of 2,575 total degrees of freedom. The time step is set to $\Delta t = 0.1$. We let the simulations run until $T = 40$. We show in Fig. 2 the streamlines over velocity magnitude contours given by the Leray-α model with $\alpha = 0.25$ and $\alpha = 0.1$ at time $T = 40$. We observe the solutions are similar away from the step, but behind the step they exhibit very different behavior: for $\alpha = 0.25$ there is no eddy separation, while for $\alpha = 0.1$ the correct transient behavior of eddy shedding is predicted. This sensitivity to α near the step and lack of sensitivity away from the step are predicted in the plot of the velocity sensitivity magnitude $|\boldsymbol{s}_h|$ for $\alpha = 0.25$ reported in Fig. 3.

The same test was run with the Leray-α-NL model. Figure 4 displays the streamlines over velocity magnitude contours given by the Leray-α-NL model with $\alpha = 0.25$ and $\alpha = 0.1$ at time $T = 40$. Here we observe that both solutions correctly predict eddy shedding behind the step. Moreover, we observe that the velocity sensitivity magnitude for $\alpha = 0.25$ shown in Fig. 5 is quite small. In fact even though $|\boldsymbol{s}_h|$ is largest behind the step, just as in the Leray-α case, for the nonlinear model the magnitude of sensitivity is almost 2 orders of magnitude smaller: at $T = 40$

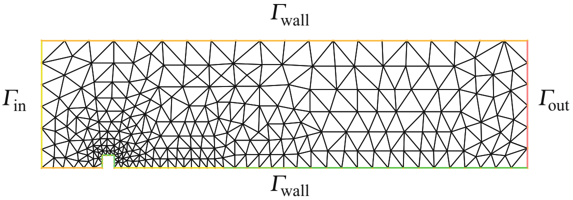


Fig. 1 Mesh used for the computations of the channel flow past a forward-backward step

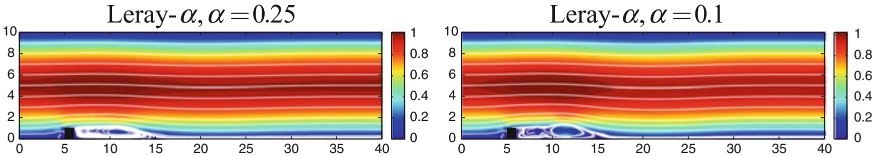


Fig. 2 Streamlines over velocity magnitude contours given by the Leray-α model with $\alpha = 0.25$ (left) and $\alpha = 0.1$ (right) at time $T = 40$

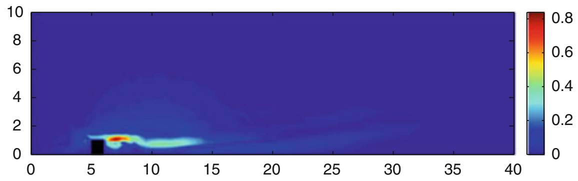


Fig. 3 Velocity sensitivity magnitude $|s_h|$ for the Leray-α model with $\alpha = 0.25$ at time $T = 40$

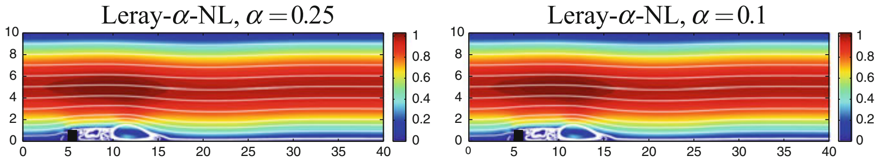


Fig. 4 Streamlines over velocity magnitude contours given by the Leray-α-NL model with $\alpha = 0.25$ (left) and $\alpha = 0.1$ (right) at time $T = 40$

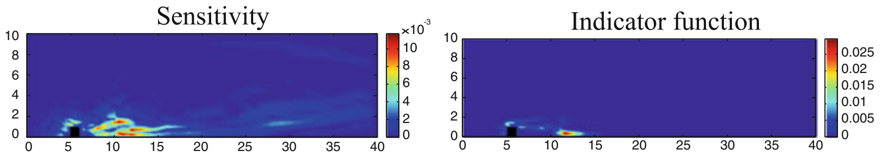


Fig. 5 Velocity sensitivity magnitude $|s_h|$ (left) and indicator function $a(\boldsymbol{u}_h) = |\boldsymbol{u}_h - \hat{\boldsymbol{u}}_h|^2$ (right) for the Leray-α-NL model with $\alpha = 0.25$ at time $T = 40$

$\|\boldsymbol{s}_h\|_{L^\infty} \approx 0.01$ for the Leray-α-NL model, while $\|\boldsymbol{s}_h\|_{L^\infty} \approx 0.80$ for the Leray-α model. Hence the Leray-α-NL correctly predicts the physical behavior with both choices of α, and is much less sensitive to the parameter choice than the classical Leray-α model. Figure 5 reports also the indicator function $a(\boldsymbol{u}_h) = |\boldsymbol{u}_h - \hat{\boldsymbol{u}}_h|^2$ for the Leray-α-NL model with $\alpha = 0.25$ at time $T = 40$. We see that the indicator function takes larger values in the region behind the step, as expected.

4.2 Channel Flow with a Contraction and Two Outlets

The second numerical test is taken from Heywood et al. [25]: channel flow with a contraction, one inlet on the left side, and outlets at the top and right. We impose boundary conditions (5)–(7) for the Leray-α model and (12)–(14) for the Leray-α-NL model with $\boldsymbol{u}_{\text{in}} = (4y(1-y), 0)^T$. The boundary conditions for the sensitivity systems are as reported in Sect. 2. We set $\boldsymbol{f} = \boldsymbol{0}$, $\nu = 0.001$, and $\boldsymbol{u}_0 = \boldsymbol{0}$. We let the simulations run until $T = 4$. The Navier-Stokes velocity magnitude on a fully resolved mesh is shown in Fig. 7 for $T = 4$. This solution was obtained using a fully implicit Crank-Nicolson temporal discretization with time step of $\Delta t = 0.005$ and (P_3, P_2) grad-div stabilized Taylor-Hood elements on the triangular mesh with 260,378 total degrees of freedom.

We consider a coarse Delaunay generated triangulation (shown in Fig. 6), with a total of 24,553 total degrees of freedom, that is one order of magnitude less than a fully resolved mesh. Figure 8 shows the velocity magnitude contours given by the Leray-α model with $\alpha = 0.16$ and $\alpha = 0.14$ at time $T = 4$. First of all, we note these solutions do not match well the solution given by DNS shown in Fig. 7. Comparing to each other, the solutions for $\alpha = 0.16$ and $\alpha = 0.14$ in Fig. 8 appear similar on the left half of the channel, but on the right-hand side the 'jet' for $\alpha = 0.14$ extends slightly farther. Also there are discrepancies near the top outlet; see zoomed-in views in Fig. 8. These differences are predicted by the velocity sensitivity solution for $\alpha = 0.16$ reported in Fig. 9.

The same test was run with Leray-α-NL. Figure 10 displays the velocity magnitude contours given by the Leray-α-NL model with $\alpha = 0.16$ and $\alpha = 0.14$ at time $T = 4$. These solutions match each other well and match the general pattern of the solution given by DNS shown in Fig. 7. Examining the sensitivity solution for $\alpha = 0.16$ in Fig. 11 we see greater sensitivity near the top outlet. However, the velocity sensitivity magnitude $|\boldsymbol{s}_h|$ is smaller than for the classical Leray-α model; compare Fig. 11 with Fig. 9. Also for this second test, the Leray-α-NL correctly predicts the physical behavior with both choices of α, and is less sensitive to the parameter choice than the classical Leray-α model. Finally, Fig. 11 reports also the indicator function $a(\boldsymbol{u}_h) = |\boldsymbol{u}_h - \hat{\boldsymbol{u}}_h|^2$ for the Leray-α-NL model with $\alpha = 0.16$ at time $T = 4$. Figure 11 shows that it is a suitable indicator function since it correctly selects the regions of the domain where the velocity does need regularization.

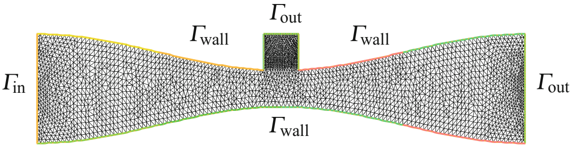


Fig. 6 Mesh used for the computations of the channel flow with a contraction and two outlets

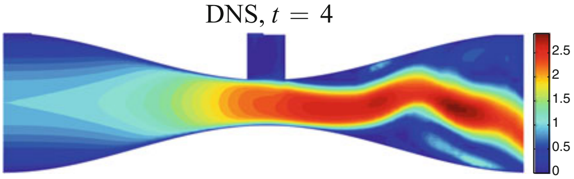


Fig. 7 Velocity magnitude contours given by DNS (NSE on a fully resolved mesh) at time $T = 4$

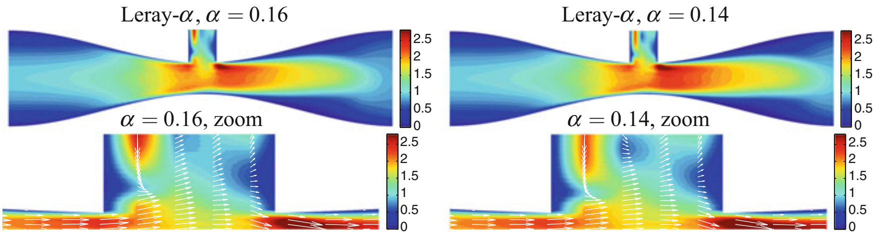


Fig. 8 Velocity magnitude contours given by the Leray-α model with $\alpha = 0.16$ (top left) and $\alpha = 0.14$ (top right) at time $T = 4$ and respective zoomed in views (bottom)

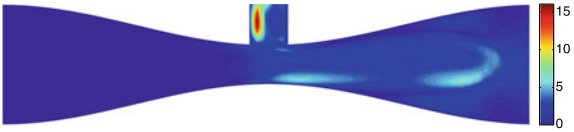


Fig. 9 Velocity sensitivity magnitude $|\boldsymbol{s}_h|$ for the Leray-α model with $\alpha = 0.16$ at time $T = 4\,\text{s}$

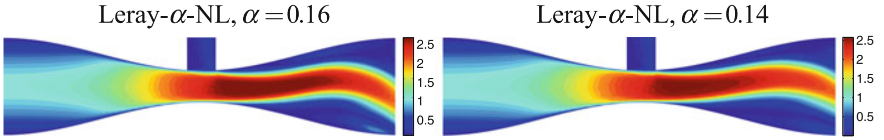


Fig. 10 Velocity magnitude contours given by the Leray-α-NL model with $\alpha = 0.16$ (left) and $\alpha = 0.14$ (right) at time $T = 4$

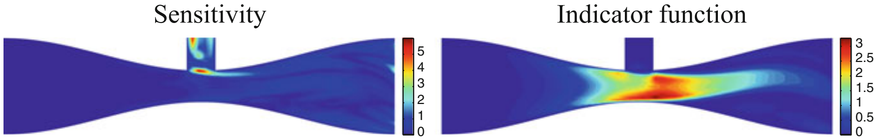


Fig. 11 Velocity sensitivity magnitude $|\boldsymbol{s}_h|$ (left) and indicator function $a(\boldsymbol{u}_h) = |\boldsymbol{u}_h - \hat{\boldsymbol{u}}_h|^2$ (right) for the Leray-α-NL model with $\alpha = 0.16$ at time $T = 4$

5 Conclusions

In this paper, we applied the sensitivity equation method to study the sensitivity to the filtering radius α of the classical Leray-α and a Leray model with a deconvolution-based indicator function, called Leray-α-NL. We proposed efficient and stable numerical schemes for the approximation of both models and their respective sensitivity systems, and we tested them on two benchmark problems. We showed that the velocity sensitivity magnitude correctly identifies the region of the domain where the velocity is sensitive to variations of α. Moreover, we showed that the Leray-α-NL model correctly predicts the physical solution for different values of α, and is much less sensitive to the parameter choice than the classical Leray-α model.

This is a preliminary work aiming at assessing numerical schemes for the sensitivity equations. Clearly, we expect to use the sensitivity results to perform specific strategies for the selection of the filter radius. This will be based on the following steps: (1) Compute the LES solution and the sensitivity with a conservative choice of the radius ($\alpha = \alpha_0$ "large"); (2) Rapidly recompute the solution for smaller values of α according to the expansion

$$\boldsymbol{u}(\alpha) \approx \boldsymbol{u}(\alpha_0) + \boldsymbol{s}(\alpha_0)(\alpha - \alpha_0).$$

The definition of the appropriate criteria for the identification of the most appropriate radius is expected to be largely problem-dependent and will be subject of forthcoming works.

Acknowledgements The research presented in this work was carried out during LR's visit at the Department of Mathematics and Computer Science at Emory University in the fall semester 2014. This support is gratefully acknowledged. This research has been supported in part by the NSF under grants DMS-1620384/DMS-1620406 (Quaini and Veneziani), DMS-1262385 (Quaini), Emory URC Grant 2015 Numerical Methods for Flows at Moderate Reynolds Numbers in LVAD (Veneziani), and DMS-1522191 (Rebholz).

References

1. Anitescu M, Layton WJ (2007) Sensitivities in large eddy simulation and improved estimates of turbulent flow functionals. SIAM J Sci Comput 29(4):1650–1667
2. Bertagna L, Quaini A, Veneziani A (2016) Deconvolution-based nonlinear filtering for incompressible flows at moderately large Reynolds numbers. Int J Numer Methods Fluids 81(8):463–488
3. Borggaard J, Burns J (1995) A sensitivity equation approach to shape optimization in fluid flows. In: Flow control (Minneapolis, MN, 1992), vol 68 of IMA Vol Math Appl, Springer, New York, pp 49–78
4. Borggaard J, Burns J (1997) A PDE sensitivity equation method for optimal aerodynamic design. J Comput Phys 136(2):366–384
5. Borggaard J, Verma A (2000) On efficient solutions to the continuous sensitivity equation using automatic differentiation. SIAM J Sci Comput 22(1):39–62
6. Bowers AL, Rebholz LG (2012) Increasing accuracy and efficiency in FE computations of the Leray-deconvolution model. Numer Methods Partial Differ Equ 28(2):720–736
7. Bowers AL, Rebholz LG (2013) Numerical study of a regularization model for incompressible flow with deconvolution-based adaptive nonlinear filtering. Comput Methods Appl Mech Engrg 258:1–12
8. Bowers AL, Rebholz LG, Takhirov A, Trenchea C (2012) Improved accuracy in regularization models of incompressible flow via adaptive nonlinear filtering. Internat J Numer Methods Fluids 70(7):805–828
9. Brenner SC, Scott LR (2008) The mathematical theory of finite element methods. Springer, New York
10. Cao Y, Titi ES (2009) On the rate of convergence of the two-dimensional α-models of turbulence to the Navier-Stokes equations. Numer Funct Anal Optim 30(11–12):1231–1271

11. Cheskidov A, Holm DD, Olson E, Titi ES (2005) On a Leray-α model of turbulence. Proc R Soc Lond Ser A Math Phys Eng Sci 461(2055):629–649
12. Dunca A, Epshteyn Y (2006) On the Stolz-Adams deconvolution model for the large-eddy simulation of turbulent flows. SIAM J Math Anal 37(6):1890–1902
13. Geurts BJ, Holm DD (2002) Leray simulation of turbulent shear layers. In: Castro IP, Hancock PE, Thomas TG (eds), Advances in Turbulence IX: Proceedings of the Ninth European Turbulence Conference (Southampton, 2002), CIMNE, pp 337–340
14. Geurts BJ, Holm DD (2003) Regularization modeling for large-eddy simulation. Phys Fluids 15(1):L13–L16
15. Geurts BJ, Holm DD (2006) Leray and LANS-α modelling of turbulent mixing. J Turbul 7(10):33
16. Geurts BJ, Kuczaj AK, Titi ES (2008) Regularization modeling for large-eddy simulation of homogeneous isotropic decaying turbulence. J Phys A 41(34):344008 (29p)
17. Godfrey AG, Cliff EM (1998) Direct calculation of aerodynamic force derivatives: a sensitivity equation approach. In: 36th AIAA Aerospace Sciences Meeting and Exhibit. AIAA, Paper 98-0393, 12p
18. Graham JP, Holm DD, Mininni P, Pouquet A (2011) The effect of subfilter-scale physics on regularization models. J Sci Comput 49(1):21–34
19. Graham JP, Holm DD, Mininni PD, Pouquet A (2008) Three regularization models of the Navier-Stokes equations. Phys Fluids 20:35107 (15p)
20. Gunzburger MD (1989) Finite element methods for viscous incompressible flows: a guide to theory, practice, and algorithms. Academic Press, Boston
21. Gunzburger MD (1999) Sensitivities, adjoints and flow optimization. Int J Numer Methods Fluids 31(1):53–78
22. Hecht F (2012) New development in FreeFem++. J Numer Math 20(3–4):251–265
23. Hecht MW, Holm DD, Petersen MR, Wingate BA (2008) The LANS-α and Leray turbulence parameterizations in primitive equation ocean modeling. J Phys A 41(34):344009 (23p)
24. Heywood JG, Rannacher R (1990) Finite-element approximation of the nonstationary Navier-Stokes problem. IV: Error analysis for second-order time discretization. SIAM J Numer Anal 27(2):353–384
25. Heywood JG, Rannacher R, Turek S (1996) Artificial boundaries and flux and pressure conditions for the incompressible Navier-Stokes equations. Int J Numer Methods Fluids 22(5):325–352
26. Layton W (2008) Introduction to the numerical analysis of incompressible viscous flows. SIAM
27. Layton W, Manica CC, Neda M, Rebholz LG (2008) Helicity and energy conservation and dissipation in approximate deconvolution LES models of turbulence. Adv Appl Fluid Mech 4(1):1–46
28. Layton W, Mays N, Neda M, Trenchea C (2014) Numerical analysis of modular regularization methods for the BDF2 time discretization of the Navier-Stokes equations. ESAIM Math Model Numer Anal 48(3):765–793
29. Layton W, Rebholz L (2012) Approximate deconvolution models of turbulence: analysis, phenomenology and numerical analysis. Springer, Heidelberg
30. Layton W, Rebholz L, Trenchea C (2012) Modular nonlinear filter stabilization of methods for higher Reynolds numbers flow. J Math Fluid Mech 14(2):325–354
31. Leray J (1934) Sur le mouvement d'un liquide visqueux emplissant l'espace. Acta Math 63(1):193–248
32. Liu Y, Tucker P, Kerr R (2008) Linear and nonlinear model large-eddy simulations of a plane jet. Comput Fluids 37(4):439–449
33. Lunasin E, Kurien S, Titi ES (2008) Spectral scaling of the Leray-α model for two-dimensional turbulence. J Phys A 41(34):344014 (10p)
34. Pahlevani F (2006) Sensitivity computations of eddy viscosity models with an application in drag computation. Int J Numer Methods Fluids 52(4):381–392
35. Quarteroni A, Sacco R, Saleri F (2007) Numerical mathematics, 2nd edn. Springer, Berlin

36. Sagaut P, Lê T (1997) Some investigations of the sensitivity of large eddy simulation. Technical report, ONERA
37. Sagaut P, Lê TH (1997) Some investigations on the sensitivity of large eddy simulation. In: Chollet J-P, Voke PR, Kleiser L (eds) Direct and Large-Eddy Simulation II: Proceedings of the ERCOFTAC Workshop (Grenoble, 1996). Springer, Dordrecht, pp 81–92
38. Stanley L, Stewart D (2002) Design sensitivity analysis: computational issues of sensitivity equation methods. Number 25 in Frontiers in Applied Mathematics. SIAM, Philadelphia, PA
39. Verstappen R (2008) On restraining the production of small scales of motion in a turbulent channel flow. Comput Fluids 37(7):887–897

Model Order Reduction for Problems with Large Convection Effects

Nicolas Cagniart, Yvon Maday and Benjamin Stamm

Abstract The reduced basis method allows to propose accurate approximations for many parameter dependent partial differential equations, almost in real time, at least if the Kolmogorov n-width of the set of all solutions, under variation of the parameters, is small. The idea is that any solutions may be well approximated by the linear combination of some well chosen solutions that are computed offline once and for all (by another, more expensive, discretization) for some well chosen parameter values. In some cases, however, such as problems with large convection effects, the linear representation is not sufficient and, as a consequence, the set of solutions needs to be transformed/twisted so that the combination of the proper twist and the appropriate linear combination recovers an accurate approximation. This paper presents a simple approach towards this direction, preliminary simulations support this approach.

N. Cagniart · Y. Maday (✉)
Laboratoire Jacques-Louis Lions (LJLL), Sorbonne Université,
Université Paris-Diderot SPC, CNRS, 75005 Paris, France
e-mail: maday@ann.jussieu.fr

N. Cagniart
e-mail: cagniartn@ljll.math.upmc.fr

Y. Maday
Institut Universitaire de France, Paris, France

Y. Maday
Division of Applied Mathematics, Brown University, Providence, RI, USA

B. Stamm
Center for Computational Engineering Science,
RWTH Aachen University, Aachen, Germany
e-mail: stamm@mathcces.rwth-aachen.de

B. Stamm
Computational Biomedicine (IAS-5 and INM-9),
Forschungszentrum Jülich, Jülich, Germany

B. N. Chetverushkin et al. (eds.), *Contributions to Partial Differential Equations and Applications*, Computational Methods in Applied Sciences 47,
https://doi.org/10.1007/978-3-319-78325-3_10

1 Introduction

Fast reliable solutions to many queries parametric Partial Differential Equations (PDE) have many applications among which real time systems, optimization problems and optimal control. Many different methods for reducing the complexity of the computations when such many queries are required have blossomed for answering this specific need. One of the approaches that have emerged is reduced order modeling (ROM). Methods in this category have been developed and are now well understood and set on firm grounds, both for steady cases or time dependent problems where time can be considered as another parameter.

The reduced basis method, which is the method that we focus on in this paper, enters in this frame and consists in, (1) defining a sequence of low dimensional spaces for the approximation of the whole set of the solutions to the parametric PDE when the parameters vary (called hereafter the solution manifold associated to our problem); (2) once such a sequence of low dimensional spaces (known as reduced basis spaces) is determined, an approximate solution is sought in such a chosen reduced space to the PDE for the values of the parameter we are interested in. The approximation is often based on a Galerkin formulation. For such reduced basis methods, both the variety of applications and the theory are now quite sound. For instance, reliable algorithms with a priori estimates and certified a posteriori errors have been developed for elliptic and parabolic problems, with or without so-called affine parameter dependence, see e.g. the two recent books on the subject [12, 20] and, of course, the publications therein.

Reduced basis methods, classically, consider the solution manifold associated to the parametrized problem as outlined above and are appropriate if this manifold can be approximated accurately by a sequence of finite dimensional spaces. The mathematical frame for this is inherently linked to the notion of *Kolmogorov width* of solution manifolds, i.e. on how well the solution manifold can be approached by a finite dimensional linear space. More precisely, let $\mathscr{M}$ be a manifold embedded in some normed linear space X. The Kolmogorov n-width of $\mathscr{M}$ is defined as:

$$d_n(\mathscr{M}, X) = \inf_{E_n} \sup_{f \in \mathscr{M}} \inf_{g \in E_n} \|f - g\|_X .$$

The first infimum being taken over all linear subspaces E_n of dimension n embedded in X.

Even if, from the practical point of view, there are various ways for checking that $\mathscr{M}$ can be approximated by a series of reduced spaced with small dimension, the first natural mathematical question is to provide an estimation of the Kolmogorov n-width of $\mathscr{M}$. Second, the question of an applied mathematician is if one can actually build an optimal, or close to optimal sequence of basis sets for these spaces?

Of course, in the vast majority of real cases, there is no analytical expression for this dimension but there are some papers giving bounds for some restricted classes of problems in the literature. For instance, in [17] bounds on d_n are found for solution manifolds corresponding to regular elliptic problems and where the parameter dependence is on the forcing term. More general cases can be handled using the results in [8]. The hypothesis therein is on the regularity of the solution with respect to the parameter dependence, it is proven that, under analyticity assumption on the behavior of the parameters in the PDE, the smallness Kolmogorov n-width of the manifold of parameters $\mathscr{C}$ ($\leq cn^{-t}$, $t > 1$) implies the smallness of the Kolmogorov n-width of the associated solutions manifold $\mathscr{M}_{\mathscr{C}}$ ($\leq cn^{-s}$, $s \leq t - 1$).

In practice, instead of the "optimal" linear subspace of dimension n in the sense described earlier, we build a "good" linear subspace. In the literature, the two most classical algorithms are the greedy method based on a certified (or at least fair enough) a posteriori estimator, and the Proper Orthogonal Decomposition (POD). We proceed assuming that the chosen algorithm has given a "good" basis "close" to the optimal one, that is, we assume that our reduced family of spaces $\{X_n\}_n$ satisfies:

$$d_n(\mathscr{M}, X) \approx \sup_{f \in \mathscr{M}} \inf_{g \in X_n} \|f - g\|_X .$$

A first paper on this subject is [16], where the authors derived error bounds on the error for the Reduced Basis Method (RBM) approximation in case of a single parameter dependent elliptic PDE. More general results have been obtained more recently for the greedy approach of the RBM [3, 10]. The optimality considered in the case of POD is slightly different. The POD focuses on minimizing the average error (parameter wise), in some norm. More precisely, we have the well known relation

$$\int_{\mathscr{C}} \|u(\mu) - \Pi_{POD} u(\mu)\|^2 \, d\mu = \sum_{i > N_{POD}} \lambda_i,$$

where Π_{POD} is the orthogonal projection onto the POD reduced space of dimension N_{POD} and the λ_i are the eigenvalues of the associated correlation operator, in decreasing order. The faster the decay of the eigenvalues, the fewer modes are needed for a good (in average) reconstruction of the solution manifold.

Up to now, most of the literature on the subject, deals with problems where one can expect/check/prove/ or hope, that the solution manifold $\mathscr{M}_{\mathscr{C}}$ has a small Kolmogorov n-width. There are, however, cases where the plain approach does not work and some transformation of $\mathscr{M}_{\mathscr{C}}$ needs to be done. An example is for instance the use of the Piola transform in the processing of the velocity field when the PDE is the Stokes or Navier-Stokes problem and the parameter includes the geometry of the computational problem (see, e.g., [14]). The choice of the Piola transform indeed provides better reduction than a simple change of variables.

The most classical and simple example illustrating limitations of reduced models due to large Kolmogorov n-width is the pure transport equation, with constant speed $c > 0$. Formally, we consider the following parametric PDE over the domain $\Omega = (a, b) \subset \mathbb{R}$

$$\begin{cases} \partial_t u(x,t) + c\partial_x u(x,t) = 0, & \text{in } \Omega \times]0,T[\,, \\ u(x,0) = u_0(x), & \text{in } \Omega, \\ c \in \mathscr{C} := [c_{\min}, c_{\max}]. \end{cases} \tag{1}$$

The analytic solution is given by

$$u(x,t;c) = u_0(x - ct).$$

We can consider two solution manifolds. Either the space time solution manifold

$$\mathscr{M}_{\mathscr{C}}^{x,t} = \{u(\cdot,\cdot;c),\ c \in \mathscr{C}\}\,,$$

or a more natural solution manifold in our context is the snapshot solution manifold

$$\mathscr{M}_{\mathscr{C}}^{x} = \{u(\cdot,t;c),\ t \in [0,T],\ c \in \mathscr{C}\}\,. \tag{2}$$

We will first give an illustrative idea of $d_n(\mathscr{M}_c^x)$, i.e. for a fixed convection parameter. Thus, the only "parameter" left is time and $d_n(\mathscr{M}_c^x)$ is, of course, smaller than $d_n(\mathscr{M}_{\mathscr{C}}^x)$.

Suppose now that our initial solution is compactly supported and let ℓ denote the Lebesgue measure of its support. Let us assume in addition that its support is included in $]a, a+\ell[$. Then, there are at least $(b-a)/\ell$ snapshots $\{u(\cdot,t^k;c)\}_k$ obtained for $t_k = k\ell/c$ that are two by two orthogonal proving that a lower bound of the Kolmogorov n-width is $(b-a)/\ell$. For a given accuracy, reducing ℓ, we can make the size of the reduced basis needed arbitrarily large. Another example of badly behaved manifold space can also be found in [21].

The objective here is to give a proper framework and to introduce notations generalizing the following observation: apart from translation, the solution manifold for the whole time simulation can be represented by a unique basis. However, let us stress that this translation is not a linear process hence the Kolmogorov process cannot capture it. An additional ingredient to existing reduced order methods has thus to be added so as to capture this very simple problem structure.

Most of the works in the reduced order modeling community on convection dominated problem have been done on the stabilization issue, and not on the reduction of the Kolmogorov n-wdith. For instance, the authors in [9] have proven that using, as usual, the residual of the PDE as a surrogate for the true error, is not adapted if convection is dominating as the relative a posteriori estimator is not fair enough. Their method involves other norms than the natural ones, and increases the stability at each iteration by enriching the trial space. Once again, their method improves the stability of the construction of a reduced basis, but does not handle the fact that the solution manifold can have a large Kolmogorov n-width.

In the same direction let us quote the papers related to the so called GNAT approach [6, 7] where the authors propose also an alternative reduction approach for these type of problems.

In [1], the authors address the stability issue in another direction. They give ideas and show numerical examples illustrating the fact that using L^1-minimisation, instead of the — more classical — L^2-minimisation (corresponding to a Galerkin scheme, which is natural in the reduced modeling context), does a better job for handling shocks (as appears in non linear convection problems) and provides more stable results. However, this approach does not cure the problem that we have indicated above related to the large dimension of the solution manifold.

Let us also mention at this level, as an intermediate approach, the paper [5]. As standard reduced order modeling fails, the author chooses, in a preprocessing step, to "chop off" the reduced basis functions resulting in a kind of adaptive coarse enriched finite element method.

Very few papers take the n-width issue directly. In [21], the authors propose a method that is the first attempt to use shock fitting related ideas in the context of reduced order modeling. The idea is to decompose the spatial domain into zones separated by shocks. In each zone, classical reduced order modeling is performed, and the shocks dynamic is handled using another equation. For them, it is given by Rankine-Hugoniot conditions. This method, just as any other shock fitting method, is somehow limited to one dimensional problems.

In [11], the authors develop a method where the POD basis is reconstructed at each time step to follow the propagation of the phenomenon. More precisely, by referring to Lax–pairs, they choose as reduced basis the modes of the Schrödinger operator where the potential is taken as the solution at the previous time step. Even if no theoretical proof of this ansatz is presented, the numerical results presented in that paper illustrate the interest of the approach for selecting the reduced space and adding stability to the process without curing, however, the large increase of the dimension of the reduced space when the accuracy requirement increases.

The method presented in [13] is similar to our work in many aspects, in particular in looking for a change of variable for better representing the solution manifold. Their approach relies on the existence of a main mode u_0 that, by convection, represents most of the solution. The proper change of variable (written as a sum of advection modes) is fitted by evaluating Wasserstein distances between the snapshots in $\mathscr{M}^x_{\mathscr{C}}$, with modes being obtained by solving Monge-Kantorovich optimal transport problems w.r.t. the reference mode u_0. Various numerical results illustrate the approach, however, only in cases where the solution exhibits indeed such a main mode u_0 which is doubtful in nonlinear processes. We will come back on their ideas in the following sections.

The last approach in this direction developed in [2] and in [19] uses the same initial idea. Their formal and general presentation is quite interesting and enlightening, however, the restrictions imposed on the formulation of the transformed equations seems to be somehow too stringent for many reduction processes.

This paper is the first of a series where we develop our approach in different situations. We start here by presenting the general framework and notations. We introduce the notion of "preconditioning" of a solution manifold based on our knowledge of the process (differing here somehow from the optimal transport problem approach in [13]).

We then apply our method to the specific problem of the one dimensional unsteady viscous Burger equation and we then present some numerical simulations confirming the feasibility of the method. The end of the paper states some perspectives.

2 Formal Presentation

Let us consider a general time dependent parametric PDE in some physical space $\Omega \subset \mathbb{R}^d$, $d = 1, 2, 3$,

$$\begin{cases} u_t + \mathscr{L}(u; \mu) = 0 & \text{in } [0, T] \times \Omega, \\ u(\cdot, t = 0; \mu) = u_0(\cdot, \mu) & \text{in } \Omega, \\ B(u; \mu) = 0 & \text{on } \partial\Omega, \end{cases} \tag{3}$$

where μ varies in some compact parameter space $\mathscr{C}$. Our approach considers the corresponding snapshot solution manifold $\mathscr{M}^x_{\mathscr{C}}$ as defined in (2) that is embedded in X, that, for the sake of conveniency, we choose equal to $L^2(\Omega)$.

Let us assume that the solution manifold has a simple structure, not reflected though by the Kolmogorov n-width but hidden by a transformation of the solution manifold. As stated in the introduction, we can think of the transport equation as being the simplest example for which this is occurring. The objective is to find, through a "preconditioning" step, how to recover the simple structure of the solution manifold.

In this "preconditioning" step, we target a family of (smooth) invertible mappings

$$\mathscr{F}_{\mathscr{C}} = \left\{ F : \overline{\Omega} \mapsto \overline{\Omega} \right\}$$

in which there exists well chosen applications

$$\begin{aligned} [0, T] \times \mathscr{C} &\to \mathscr{F}_{\mathscr{C}}, \\ (t, \mu) &\mapsto F_{t;\mu} \end{aligned}$$

such that the corresponding preconditioned solution manifold, defined as:

$$\mathscr{M}^x_{\mathscr{F},\mathscr{C}} := \left\{ u(F^{-1}_{t;\mu}(\cdot), t; \mu),\ \mu \in \mathscr{C},\ t \in [0, T] \right\} \tag{4}$$

has a smaller Kolmogorov n-width than $\mathscr{M}_{\mathscr{C}}$. The definition of the set $\mathscr{F}_{\mathscr{C}}$ is based on a priori expertise on the behavior of the solution. We aim to conceive and design it during a preprocessing step (generally called "offline" in the RBM community).

In what follows, we explain how we use this preconditioned solution manifold and how we pick the correct application $F_{t;\mu}$, in a computationally efficient way.

2.1 Algorithm

For simplicity, let us assume that we are using an explicit Euler scheme for the time discretization. Extensions to implicit, higher order time discretization, or more involved conservative numerical scheme, is straightforward and will be reported in a future paper.[1] Our semi-discretized PDE then becomes

$$\begin{cases} \dfrac{u^{n+1}-u^n}{dt} + \mathscr{L}(u^n;\mu) = 0 & \text{in } \Omega, \\ u(\cdot, t=0;\mu) = u_0(\cdot,\mu) & \text{in } \Omega, \\ B(u^n;\mu) = 0 & \text{on } \partial\Omega. \end{cases} \tag{5}$$

Here, as is classical, dt denotes the time step, and u^n an approximation for the solution to (3) at time ndt.

Assume that we have a basis $\{\phi_i\}$ such that span$\{\phi_i\}$ approaches the preconditioned solution manifold $\mathscr{M}^x_{\mathscr{F},\mathscr{C}}$ defined in (4) to a given accuracy. Since $\mathscr{M}^x_{\mathscr{F},\mathscr{C}}$ is assumed to be of small Kolmogorov n-width, we expect that we can find such a basis of moderate size. At each time step, we look for coordinates $(\alpha_i^{n+1})_i$ on the reduced basis and an application $F_{n+1} \in \mathscr{F}_{\mathscr{C}}$ such that $u(\cdot, t^{n+1};\mu)$ is well approximated by:

$$u^{n+1} := \sum_{i=1}^{M} \alpha_i^{n+1} \phi_i \circ F_{n+1}.$$

In order to expect the search for F_{n+1} be computationally tractable, let us assume that our family $\mathscr{F}_{\mathscr{C}}$ can be parametrized by a few parameters: that is

$$\forall F_{t;\mu} \in \mathscr{F}_{\mathscr{C}}, \quad \exists (\gamma_j)_j, \quad \text{such that } \; F_{t;\mu} = F\left[\gamma_1(t;\mu), \ldots, \gamma_m(t;\mu)\right].$$

In the discrete setting, the search for F_{n+1} then reduces to the search for $(\gamma_j^{n+1})_j$, and we set $F_{n+1} = F\left[\gamma_1^{n+1}, \ldots, \gamma_m^{n+1}\right]$.

We are thus simultaneously looking for an appropriate reduced space (defined as the span of the $(\phi_i \circ F_{n+1})_i$) and for coordinates on this reduced space. We have chosen to derive our solution from some minimization problem of the form

$$(\gamma_j^{n+1}, \alpha_i^{n+1}) = \operatorname*{arg\,min}_{(\gamma_j,\alpha_i)} \left\| \sum_i \alpha_i \, \phi_i \circ F([\gamma_j]_j) - u^n + dt\mathscr{L}(u^n;\mu) \right\| \tag{6}$$

for some appropriate norm $\|\cdot\|$ on X.

[1] Note that, of course, this choice of an explicit scheme involves a limitation on the time step due to a CFL condition that can be severe for an accurate finite element or finite difference scheme but reveals to be moderate in the reduced basis framework.

Remark 1 It is interesting to note that our approach, in this context, may be presented as a shock fitting method, and thus one may fear that it will suffer from the classical drawback of this class of approach, especially the difficulty to generalize to multidimensional framework. One reassuring element is that the position of the fitting F_n is not defined through the Rankine-Hugoniot conditions but through the minimization process (6), and the evolution in time can be chosen to follow any appropriate conservative numerical scheme. [This does not mean, however, that the extension to two dimensional problems does not lead to some difficulties! This is under investigation and will be presented in a future paper (see also [4]).]

Several choices are possible for the sense in which we will minimize this quantity. One example will be given in the next section. We propose the following generic algorithm.

Algorithm 3

Step 1. Initialize α_i and γ_j:

$$(\alpha_i^{n+1,0}, \gamma_j^{n+1,0}) = (\alpha_i^{ini}, \gamma_j^{ini})$$

α_i^{ini} and γ_j^{ini} will depend on the previous timesteps, namely on $(\alpha_i^k)_i$ and $(\gamma_j^k)_j$ for $k \leq n$. Then, assuming that $(\alpha_i^{n+1,q}, \gamma_j^{n+1,q})$ are known for some internal iteration $q \geq 0$, we proceed

Step 2. Fit the α_i given $[\gamma_j^{n+1,q}]_j$: Find $(\alpha_i^{n+1,q+1})_i$ that minimizes the following quantity (in some sense):

$$\sum_i \alpha_i^{n+1,q+1} \phi_i \circ F\left([\gamma_j^{n+1,q}]_j\right) - u^n + dt * \mathscr{L}(u^n; \mu)$$

Step 3. Fit the γ_j given $(\alpha_i^{n+1,q+1})_i$: Find $(\gamma_j^{n+1,q+1})_j$ that minimizes the following quantity (in some sense):

$$\sum_i \alpha_i^{n+1,q+1} \phi_i \circ F\left([\gamma_j^{n+1,q+1}]_j\right) - u^n + dt * \mathscr{L}(u^n; \mu)$$

until convergence (for which, say $q = q^*$). Then, we set

$$(\alpha_i^{n+1}, \gamma_j^{n+1}) = (\alpha_i^{n+1,q^*+1}, \gamma_j^{n+1,q^*+1}).$$

2.2 *Discussion*

The two closest methods to ours are first the one developed in [2, 19], second the one presented in [13].

In the former method, there is also the search for a phase component: $F(t; \mu)$ and a shape component: $v := u(\cdot, t; \mu) \circ F(t; \mu)$ that they name the “calibrated solution”. They present the approach in the frame of Lie group action and thus introduce the notion of equivariance with respect to the group action for the calibrated solution. Hence, instead of best fitting these two objects from the discrete equation (5) by solving the optimal problem (6), the idea is to find an equation satisfied by the calibrated solution, i.e. an operator $\tilde{\mathscr{L}}$ such that v is solution to the following equation

$$v_t + \tilde{\mathscr{L}}(v;\mu) = 0 \quad \text{in } [0,T]\times\Omega.$$

To close the system, they need to add a well chosen equation on $F(t;\mu)$. Well chosen here means that the calibrated equation has to be well posed, and the dynamics of the shape component should be much simpler than those of the original solution.

The second paper [13] differs on two points. There is a unique reference mode u_0 that allows to characterize the mapping, obtained from an optimal transport problem. The interest is that there is no need to have an expertise on what is the set $\mathscr{F}_{\mathscr{C}}$ — the Monge-Kantorovich optimal transport problem doing the job —, the drawback is that the problem should have a unique natural reference mode.

3 Illustration on the Viscous Burger's Equation in One Dimension

The viscous Burger's equation has already received some attention in the reduced modeling context. We mention [22] for the stationary case and when the solution manifolds can be well represented by a small finite dimensional linear space, without any preconditioning.

We consider $\Omega = (-1, 1)$, and solve for the time dependent viscous Burger equation with no forcing term and periodic boundary conditions (we will see later why these are important in our analysis):

$$\begin{cases} u_t + \nu u u_x - \varepsilon u_{xx} = 0 \quad \text{in } [0,T]\times\Omega, \\ u|_{t=0} = u_0, \\ u \text{ periodic.} \end{cases} \tag{7}$$

The extension to non periodic boundary conditions is under control and will be presented in a future paper (see also [4]).

The parameters of this problem are the triplets: $\mu = (u_0, \nu, \varepsilon)$. We want to choose a parameter domain $\mathscr{C}$ in order that the problem is

- convection dominated so that the solution manifold has a large Kolmogorov n-width
- not too stiff so as not to be bothered by stabilization issues as mentioned in the introduction, hence, we shall only consider the cases $\varepsilon \geq \varepsilon_0 > 0$ (see the recent paper [15] that tackles this problem).

We have chosen the following:

$$\mathscr{C} = \begin{cases} \lambda \in [0.5, 1.3], \\ \nu \in [4., 6.], \\ \varepsilon \in [0.04, 0.2]. \end{cases}$$

3.1 Variational Formulation and Truth Approximation

For the truth approximation to the solution of problem (7), let us consider a semi implicit scheme (so as not to be bothered by a two stringent stability constraint) with time step dt_{truth}. Let

$$X = H^1_{\text{per}}(\Omega)$$

and let us denote by $\langle \cdot, \cdot \rangle$ and $\| \cdot \|$ the usual L^2 inner product and norm. For each $\mu = (u_0, \nu, \varepsilon) \in \mathscr{C}$, for the semi-discrete (in time) truth problem, we are looking for $u^{n+1} \in X$ (approximation of $u(\cdot, (n+1)dt_{\text{truth}}, \mu)$) such that: $\forall v \in X$

$$\langle u^{n+1}(\mu), v \rangle + dt_{\text{truth}} \varepsilon a(u^{n+1}(\mu), v) = \langle u^n(\mu), v \rangle - dt_{\text{truth}} \nu c(u^n(\mu), u^n(\mu), v)$$

where

$$c(w, z, v) = \int_\Omega w\, z_x\, v \quad \text{and} \quad a(w, v) = \int_\Omega w_x\, v_x.$$

This semi-discretized problem is trivially well posed. In order to finalize the discretization, let us introduce an appropriate finite element discretization, the truth approximation space, $X^{\mathcal{N}}$. We pick it fine enough so that, with the chosen time step dt_{truth}, it is able to represent well our solution manifold. From now on, we will consider that the exact solution $u(\cdot, t; \mu)$ and the "truth" solution $u^{\mathcal{N}}(\cdot, t, \mu)$ cannot be distinguished.

3.2 Model Order Reduction — Offline Stage

As mentioned earlier, the first question we need to answer is: does our solution manifold $\mathscr{M}^x_{\mathscr{C}}$ (in practice represented by $\mathscr{M}^{x,truth}_{\mathscr{C}}$) have a large Kolmogorov n-width? And if so, can we find better behaved "calibrated" manifold solution? Fig. 1 shows some snapshots $\left\{u(\cdot, t^k; \mu),\ k \in 1 \ldots K\right\}$ taken in $\mathscr{M}^x_{\mathscr{C}}$ for some parameters.

From basic expertise on the Burger's equation, we choose the mapping family $\mathscr{F} = \left\{F_{t;\mu}\right\}$, where $F_{t;\mu}$ are defined as translation operators:

$$F_{t;\mu} : \Omega \mapsto \Omega, \quad x \mapsto x - \gamma(t; \mu)$$

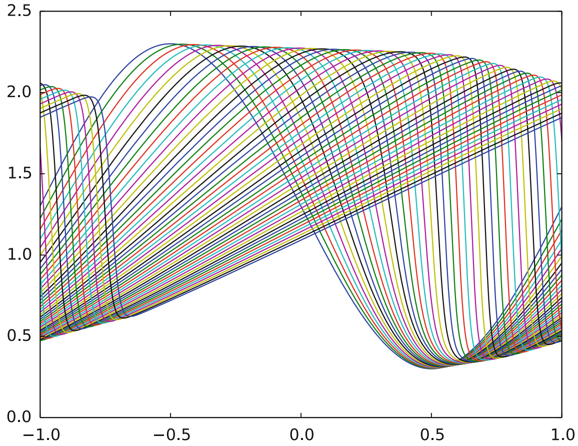


Fig. 1 Snapshots of the solution to the unsteady viscous Burger equation with $u_0 = \lambda + \sin(x)$, $\lambda = 1.3$, $\nu = 4$, $\varepsilon = 0.04$

with $\gamma(t;\mu) \in \mathbb{R}$. With this choice, our family of mappings is a one parameter family, i.e.:

$$\mathscr{F} = \{F(\gamma) \mid \gamma \in \mathbb{R}\}\,.$$

Unlike in the pure translation problem of the introduction (1), our parameter γ is not constant (it is a function of μ and time) and has no analytical expression. Our calibrated solution manifold is then

$$\mathscr{M}^x_{\mathscr{F},\mathscr{C}} = \{u(\cdot - \gamma(t;\mu), t;\mu),\ t \text{ in } [0,T],\ \mu \in \mathscr{C}\},$$

that is represented in Fig. 2 where we understand that the Kolmogorov n-width of $\mathscr{M}^x_{\mathscr{F},\mathscr{C}}$ is smaller than the original one represented in Fig. 1.

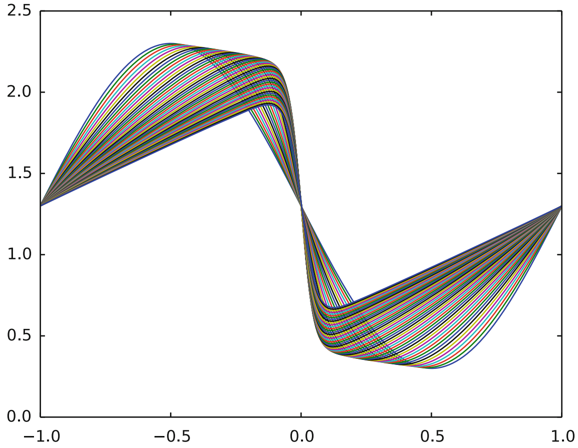


Fig. 2 Calibrated set of the above snapshots for $u_0 = \lambda + \sin(x)$, $\nu = 4$, $\varepsilon = 0.04$

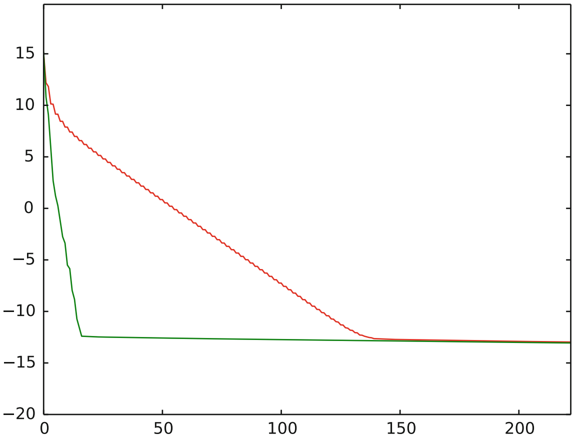


Fig. 3 Eigenvalues of the POD decomposition of the original set of snapshots (in red) and of the calibrated set of snapshots (in green)

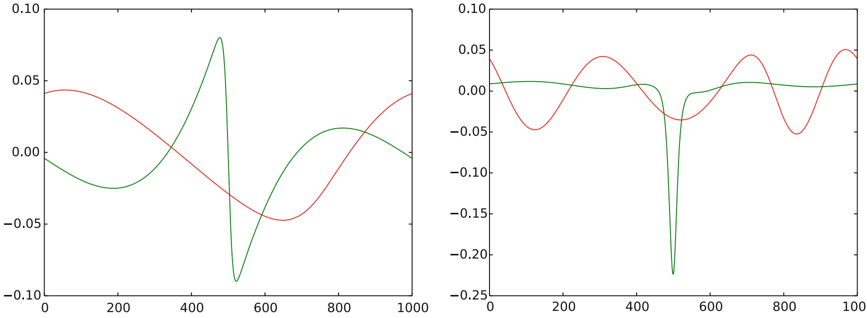


Fig. 4 3rd (left) and 6th (right) POD modes for the calibrated (green) and original (red) simulations

This is confirmed in Fig. 3 which presents the decay of the POD eigenvalues in logarithmic scale for $\mathscr{M}^x_{\mathscr{C}}$ and $\mathscr{M}^x_{\mathscr{F},\mathscr{C}}$. As we could have expected, to achieve a fixed accuracy, the number of POD modes needed to represent the calibrated manifold is much smaller than the number of modes needed for the original solution set.

To confirm this, we present in Fig. 4 the 3rd and 6th POD modes of the calibrated and non calibrated simulations. As we can see, in the calibrated case, with just 3 modes, our L^2-projection focuses on reproducing the shock, whereas in the non-calibrated case, the modes desperately try to represent shocks centered anywhere in Ω. We mention again the fact that, even in the calibrated case, Algorithm 3 could be improved using L^1-minimization.

We present in Fig. 5 the projection of one of the snapshot on the first three POD modes. With 10 POD modes in the uncalibrated case, the projection shown on Fig. 5 exhibit the oscillatory behaviour as described in [1].

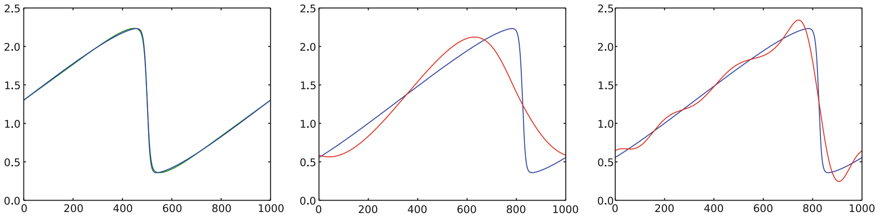


Fig. 5 Projection of a snapshot (blue): on 3 POD modes in the calibrated case (left figure), on 3 POD modes in the non calibrated case (central figure), on 10 POD modes in the non calibrated case (right figure)

At this stage, we suppose that we have found a "calibrated" solution manifold, with nice Kolmogorov n-width decay. That is, we have calibrated an original dataset, and obtained a reduced orthonormal basis:

$$\operatorname{span}\{\phi_i,\ i = 1\ldots M\} \subset X \tag{8}$$

that approximates well the calibrated solution manifold $\mathscr{M}^{x}_{\mathscr{F},\mathscr{C}}$.

We now need to explicit Algorithm 3. The biggest question is: How do we pick the $F \in \mathscr{F}$ at each time step?

3.3 Model Order Reduction — Online Stage

As was introduced in the previous section [see (5)], for the time semi-discretization of the RBM approach, we use a forward Euler discretization with a time step dt that may be different from dt_{truth}. At each time step we are looking for the solution to the elliptic problem[2]

$$u^{n+1} = u^n - dt\nu u^n u^n_x + dt\varepsilon u^n_{xx}$$

with periodic boundary conditions over $(-1, 1)$. This leads to the following variational formulation that will be used to provide the Galerkin formulation of the RBM: Knowing u^n, compute $u^{n+1} \in X$ such that

$$\forall v \in X,\ \langle u^{n+1}(\mu), v\rangle = \langle u^n(\mu), v\rangle - dt\nu c(u^n(\mu), u^n(\mu), v) - dt\varepsilon a(u^n(\mu), v).$$

One could fear that a problem with this discretization is the stringent CFL condition on the time-step. Our reduced basis formulation will allow for very fast computation, which will mitigate this issue on which we shall dwell upon later. As said already, we could also consider an implicit Euler scheme. We refer to [22] (stationary) and [18] (non-stationary), for the development of reduced order model in that case.

[2] Indeed there is no reason why using the same discretization in time for the truth solution and for the reduced basis scheme.

The full RBM discretization starts from the knowledge of the (supposedly accurate) approximation of u^n as an expansion

$$u^n := \sum_{i=1}^{M} \alpha_i^n \phi_i \circ F_n, \tag{9}$$

where the $\{\phi_i\}_i$ are the reduced basis elements of the good approximation of the calibrated solution manifold that have been introduced in (8) as a result of the offline process. F_n is here $F(\gamma^n)$ where γ^n is the current translation value. In order to deduce the next approximation,

$$u^{n+1} := \sum_{i=1}^{M} \alpha_i^{n+1} \phi_i \circ F_{n+1}, \quad \text{where } F_{n+1} = F(\gamma_{n+1})$$

as described in the previous section, we iterate between the search for the reduced coordinates $(\alpha_i^{n+1})_i$ and for the mapping F_{n+1} i.e. for the translation parameter γ^{n+1}. We initialize these quantities as follows:

$$\begin{aligned} \alpha_i^{n+1,0} &= \alpha_i^n \\ \gamma^{n+1,0} &= \gamma^n + \left(\gamma^n - \gamma^{n-1}\right). \end{aligned}$$

In the first part of the iterative step indexed by q, assuming we know $((\alpha_i^{n+1,q})_i, \gamma^{n+1,q})$ we fit the α_i for a fixed translation parameter γ, i.e.we are looking for $(\alpha_i^{n+1,q+1})_i$ that satisfy

$$\{\alpha_i^{n+1,q+1}\} = \underset{(\alpha_i)_i \in \mathbb{R}^{\mathbb{N}}}{\arg\min} \left\| \sum_i \alpha_i \phi_i \circ F(\gamma^{n+1,q}) - u^n - dt\nu u^n u_x^n + dt\varepsilon u_{xx}^n \right\|_2^2 .$$

The nice feature with the chosen norm is that we pick our reduced coordinates such that our residual is orthogonal to the translated reduced space, the space spanned by the $\{\phi_i \circ F(\gamma^{n+1,q})\}_i$. Using u^n's expansion on its reduced basis, the coefficients $\{\alpha_i^{n+1,q+1}\}_i$ are given by the first-order optimality condition:

$$\begin{aligned} \alpha_i^{n+1,q+1} = &\sum_j \alpha_j^n \langle \phi_j \circ F(\gamma^n), \phi_i \circ F(\gamma^{n+1,q}) \rangle \\ &- dt\nu \sum_j \sum_p \alpha_j^n \alpha_p^n \langle \phi_j \circ F(\gamma^n) \left(\phi_p \circ F(\gamma^n)\right)_x, \phi_i \circ F(\gamma^{n+1,q}) \rangle \\ &\qquad - dt\varepsilon \sum_j \alpha_j^n \langle \left(\phi_j \circ F(\gamma^n)\right)_x, \left(\phi_i \circ F(\gamma^{n+1,q})\right)_x \rangle. \end{aligned}$$

In order to evaluate this expression, we need to compute the following integrals:

$$\begin{cases} \forall i,j, & \int_\Omega \phi_j \circ F(\gamma^n)(x)\phi_i \circ F(\gamma^{n+1,q})(x) \\ \forall i,j,p, & \int_\Omega \phi_j \circ F(\gamma^n)(x) \left(\phi_p \circ F(\gamma^n)\right)_x (x)\phi_i \circ F(\gamma^{n+1,q})(x) \\ \forall i,j, & \int_\Omega \phi_j \circ F(\gamma^n)_x(x)\phi_i \circ F(\gamma^{n+1,q})_x(x) \end{cases} \tag{10}$$

We will see in the next subsection how to achieve efficient offline/online decomposition for these quantities.

Once this is done, we fit the γ. Let us define first the residual function $r(\gamma)$:

$$r(\gamma) = \left\| \sum_i \alpha_i^{n+1,q+1} \phi_i \circ F_\gamma - u^n - dt\nu u^n u_x^n + dt\varepsilon u_{xx}^n \right\|_2^2,$$

then we choose $\gamma^{n+1,q+1}$ as the "best", i.e. residual minimizing, translation parameter. It is given by:

$$\gamma^{n+1,q+1} = \arg\min_\gamma r(\gamma)$$

Next we develop $r(\gamma)$:

$$r(\gamma) = \left\| \sum_i \alpha_i^{n+1,q+1} \phi_i \circ F(\gamma) \right\|_2^2 + \left\| u^n - dt\nu u^n u_x^n + dt\varepsilon u_{xx}^n \right\|_2^2 \\ - 2\langle \sum_i \alpha_i^{n+1,q+1} \phi_i \circ F(\gamma), u^n - dt\nu u^n u_x^n + dt\varepsilon u_{xx}^n \rangle.$$

The second term is independent of γ. The first one, using periodicity, happens also to be independent of γ. We can thus replace the minimization of r by the minimisation of the following quantity $\tilde{r}$:

$$\tilde{r}(\gamma) = -\langle \sum_i \alpha_i^{n+1,q+1} \phi_i \circ F(\gamma), u^n - dt\nu u^n u_x^n + dt\varepsilon u_{xx}^n \rangle. \tag{11}$$

Here again, we need to evaluate the quantities

$$\begin{cases} \forall i,j, & \int_\Omega \phi_j \circ F(\gamma^n)(x)\phi_i \circ F(\gamma)(x) \\ \forall i,j,p, & \int_\Omega \phi_j \circ F(\gamma^n)(x) \left(\phi_p \circ F(\gamma^n)\right)_x (x)\phi_i \circ F(\gamma)(x) \\ \forall i,j, & \int_\Omega \phi_j \circ F(\gamma^n)_x(x)\phi_i \circ F(\gamma)_x(x) \end{cases}$$

for various values of γ in order to derive the value of γ that minimizes r (or $\tilde{r}$).

3.4 Offline/Online Decomposition of the Expressions Depending on γ

In both the search for γ [see (11)] and $(\alpha_i)_i$ [see (10)], we need to compute scalar products of the form

$$\langle \psi_i \circ F(\gamma^n), \psi_j \circ F(\gamma) \rangle, \tag{12}$$

where ψ can be one of the POD basis or one of its x-derivatives. γ^n and γ can take any value in Ω. Our key ingredient here is that, due to translation invariance (because we are in a periodic settings), we can replace the previous terms by

$$\langle \psi_i \circ F(\gamma^n - \gamma), \psi_j \rangle = \langle \psi_i \circ F(\Delta\gamma), \psi_j \rangle. \tag{13}$$

We have plotted in Fig. 6 these quantities (after rescaling) as a function of $\Delta\gamma$ for some pairs of chosen ψ's and we notice that, as can be expected because we are essentially using a primitive function of the integrant, these are regular functions of $\Delta\gamma$.

For a sufficiently small time step, we expect $\Delta\gamma$ to be of order $dt * c$ where c is some local characteristic velocity. We have chosen the following method:

- Precompute the scalar products for a predefined set of values of $\Delta\gamma$;
- Using some regularity hypothesis, use spline interpolation to get approximated values for all γ in $[-dt * c_{\max}, dt * c_{\max}]$, where $c_{\max}$ is the maximum expected shock speed during the simulation.

Remark 2 For the optimization of $\tilde{r}$ we have also tested to linearize our problem around γ^n which leads to a doable method but does not work better that the above.

Remark 3 A common comment about this method is about mesh interpolation. Indeed, would such a mesh interpolation be required in the online process, this would

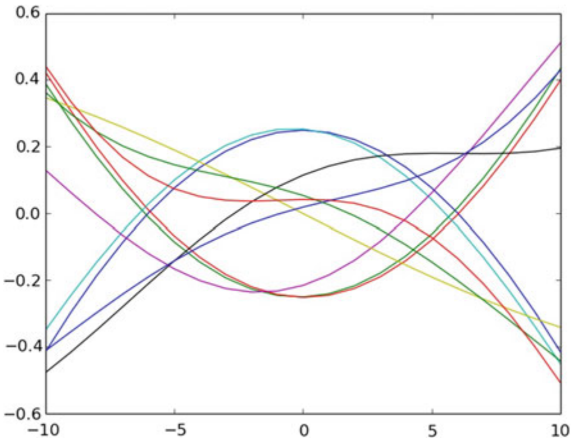


Fig. 6 A few values of the quantities (12) as a function of $\Delta\gamma$. The x axis is scaled to multiples of $c * \Delta t$

preclude any extension to 2D or 3D problems. Fortunately this is not the case. In the offline part, we have indeed to interpolate between meshes. As the computational time is not much an issue, these can be done as precisely as required. For the online now, the only thing that is required (and leads to mastered errors) is the interpolation between the discrete quantities computed in (13). This error can be quantified offline. See Fig. 6 for an idea of the quantities that we are interpolating.

4 Numerical Results

4.1 *About the CFL Condition*

We represent in Fig. 7 the value of the CFL condition of our reduced scheme using the space calibrated $\mathscr{M}^{x}_{\mathscr{F},\mathscr{C}}$ as a function of the dimension M of the discrete space expressed in the equation (9). Of course, the bigger the reduced basis, the smaller the time step required for stability. We remark that there is a plateau for large values of M that is above the CFL-condition for the truth solver. More importantly, for $M = 5$, we can use a discrete time step 3, 000 times bigger than the one of the fine (finite element) scheme (that was $dt_{\text{truth}} \leq 10^{-6}$).

4.2 *Convergence Illustration*

In Fig. 8, we have plotted the L^2-error of the solution of (6) in case of problem (7) as a function of time for different values of the reduced basis for $dt = 2.5\ 10^{-4}$. The different colors represent various values of M used in (9) (Note that on the same

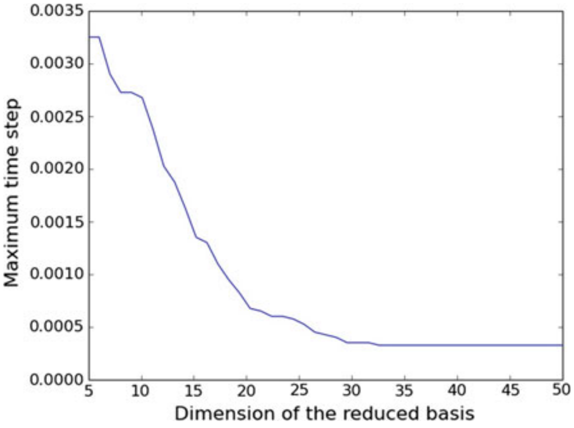


Fig. 7 A few values of the quantities (12) as a function of $\Delta\gamma$. The x axis is scaled to multiples of $c * \Delta t$

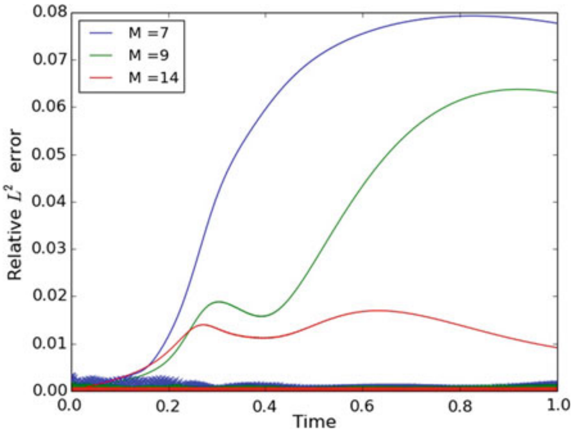


Fig. 8 Relative L^2-error of the solution as a function of time for different values of the reduced basis. The three curves close to the x-axis (almost overlapping at this scale) are the associated best approximation errors

figure, the plots close the x axis represent the projection errors (best approximation) of the solution onto the set $\mathscr{M}^x_{\mathscr{F},\mathscr{C}}$ with the exact value of the translation γ). We see that our numerical scheme is convergent, as a function of M. The final accuracy is somehow difficult to grasp since it is a function of Δt and the number of degrees of freedom used in the spatial direction (here M) as for any discretization of an evolution problem. It is also a function of the way the value of γ is found as each time step as a solution of the full minimization problem (6).

5 Conclusion

This paper is the first of a series that explain how to correct the impossibility of the standard reduced basis method (or actually most model reduction methods) to approximate well convection dominated phenomenon. The additional ingredient is to propose a change of variable, that should also be represented by few coefficients, that are updated thanks to the numerical scheme that is used classically for the discretization of the convection dominated problem. This simple approach can be implemented in an online/offline paradigm that allows online to contribute with a complexity that, at each time step, is a function of the number of reduced basis that are used for the approximation. This paper that allows to set the scene of this new approach deals with a problem with periodic boundary conditions to focus on the main feature of the approach.

The reduced basis here, is composed of snapshots of the solution. Note that we could also use the gradient of these snapshots in order to diminish the effect of the mismatched of the correct value of γ in the iterative process used to solve (6). In the toy problem used here this does not improve the accuracy but in multidimensional situations it may be useful, more tests on this are under investigations.

Acknowledgements This paper is dedicated to Professor Olivier Pironneau on the occasion of his 70th birthday. The French government is greatly acknowledged for its funding of the FUI MECASIF project.

References

1. Abgrall R, Amsallem D, Crisovan R (2016) Robust model reduction by L^1-norm minimization and approximation via dictionaries: application to linear and nonlinear hyperbolic problems. Adv Model Simul Eng Sci 3(1)
2. Beyn W-J, Thümmler V (2004) Freezing solutions of equivariant evolution equations. SIAM J Appl Dyn Syst 3(2):85–116
3. Binev P, Cohen A, Dahmen W, DeVore R, Petrova G, Wojtaszczyk P (2011) Convergence rates for greedy algorithms in reduced basis methods. SIAM J Math Anal 43(3):1457–1472
4. Cagniart N (2017) PhD thesis, University Pierre et Marie Curie, In preparation
5. Carlberg K (2015) Adaptive *h*-refinement for reduced-order models. Int J Numer Methods Engrg 102(5):1192–1210
6. Carlberg K, Bou-Mosleh C, Farhat C (2011) Efficient non-linear model reduction via a least-squares Petrov-Galerkin projection and compressive tensor approximations. Int J Numer Methods Engrg 86(2):155–181
7. Carlberg K, Farhat C, Cortial J, Amsallem D (2013) The GNAT method for nonlinear model reduction: Effective implementation and application to computational fluid dynamics and turbulent flows. J Comput Phys 242:623–647
8. Cohen A, DeVore R (2016) Kolmogorov widths under holomorphic mappings. IMA J Numer Anal 36(1):1–12
9. Dahmen W, Plesken C, Welper G (2014) Double greedy algorithms: reduced basis methods for transport dominated problems. ESAIM Math Model Numer Anal 48(3):623–663
10. DeVore R, Petrova G, Wojtaszczyk P (2013) Greedy algorithms for reduced bases in Banach spaces. Constr Approx 37(3):455–466
11. Gerbeau J-F, Lombardi D (2014) Approximated Lax pairs for the reduced order integration of nonlinear evolution equations. J Comput Phys 265:246–269
12. Hesthaven JS, Rozza G, Stamm B (2016) Certified reduced basis methods for parametrized partial differential equations. Springer, Cham (BCAM Basque Center for Applied Mathematics, Bilbao)
13. Iollo A, Lombardi D (2014) Advection modes by optimal mass transfer. Phys Rev E 89(2):022923
14. Lø vgren AE, Maday Y, Rø nquist EM (2006) A reduced basis element method for the steady Stokes problem. M2AN Math Model Numer Anal 40(3):529–552
15. Maday Y, Manzoni A, Quarteroni A (2016) An online intrinsic stabilization strategy for the reduced basis approximation of parametrized advection-dominated problems. C R Math Acad Sci Paris 354(12):1188–1194
16. Maday Y, Patera AT, Turinici G (2002) A priori convergence theory for reduced-basis approximations of single-parameter elliptic partial differential equations. J Sci Comput 17(1–4):437–446
17. Melenk J-M (2000) On *n*-widths for elliptic problems. J Math Anal Appl 247(1):272–289
18. Nguyen N-C, Rozza G, Patera AT (2009) Reduced basis approximation and a posteriori error estimation for the time-dependent viscous Burgers' equation. Calcolo 46(3):157–185
19. Ohlberger M, Rave S (2013) Nonlinear reduced basis approximation of parameterized evolution equations via the method of freezing. C R Math Acad Sci Paris 351(23–24):901–906
20. Quarteroni A, Manzoni A, Negri F (2016) Reduced basis methods for partial differential equations: an introduction, vol 92. Unitext. Springer, Cham

21. Taddei T, Perotto S, Quarteroni A (2015) Reduced basis techniques for nonlinear conservation laws. ESAIM Math Model Numer Anal 49(3):787–814
22. Veroy K, Prud'homme C, Patera AT (2003) Reduced-basis approximation of the viscous Burgers equation: rigorous a posteriori error bounds. C R Math Acad Sci Paris 337(9):619–624

Parametric Optimization of Pulsating Jets in Unsteady Flow by Multiple-Gradient Descent Algorithm (MGDA)

Jean-Antoine Désidéri and Régis Duvigneau

Abstract Two numerical methodologies are combined to optimize six design characteristics of a system of pulsating jets acting on a laminar boundary layer governed by the compressible Navier-Stokes equations in a time-periodic regime. The flow is simulated by second-order in time and space finite-volumes, and the simulation provides the drag as a function of time. Simultaneously, the sensitivity equations, obtained by differentiating the governing equations w.r.t. the six parameters are also marched in time, and this provides the six-component parametric gradient of drag. When the periodic regime is reached numerically, one thus disposes of an objective-function, drag, to be minimized, and its parametric gradient, at all times of a period. Second, the parametric optimization is conducted as a multi-point problem by the Multiple-Gradient Descent Algorithm (MGDA) which permits to reduce the objective-function at all times simultaneously, and not simply in the sense of a weighted average.

Keywords Active-flow control · Time-dependent Navier-Stokes equations Finite-volume schemes · Sensitivity equations · Multi-objective differentiable optimization · Descent methods · Robust design

1 Introduction: Active Flow Control Issues

This article aims at providing a numerical technique for optimizing parameters in the context of time-dependent problems. We are considering a test-case in which a time-periodic flowgoverned by the compressible Navier-Stokes equations in the

J.-A. Désidéri (✉) · R. Duvigneau
INRIA Acumes Team, 2004, Route des Lucioles, 06902 Sophia-Antipolis, France
e-mail: Jean-Antoine.Desideri@inria.fr

R. Duvigneau
e-mail: Regis.Duvigneau@inria.fr

B. N. Chetverushkin et al. (eds.), *Contributions to Partial Differential Equations and Applications*, Computational Methods in Applied Sciences 47,
https://doi.org/10.1007/978-3-319-78325-3_11

laminar regime includes pulsating jets as a device of active-flow control [10] in the perspective of drag reduction.

In this context, a major difficulty is related to the choice of actuation parameters, such as excitation frequency, amplitude, location, to obtain the expected flow response. In cases implying a single isolated actuator, it is relatively easy to carry out an experimental or numerical study to determine efficient control parameters. However, in the perspective of industrial applications involving hundreds of actuators, this task is far from being straightforward and the use of an automated optimization strategy is thus proposed, in the spirit of previous works [7–9].

The application of an optimization procedure to such problems is faced to the following difficulties: first, the choice of the optimization algorithm is conditioned by the huge computational time of the unsteady-flow simulation, and second, it is necessary to consider several objectives concurrently. Typically, the improvement of the single time-averaged performance is usually not satisfactory for realistic applications. Secondly, sensitivity analysis is tedious in the context of unsteady flows, due to the backward integration of the adjoint equation, which requires the storage, or partial storage / partial re-computation, of the unsteady solution.

The proposed work is based on two methodological ingredients to overcome the difficulties described above: the Sensitivity Equation Method (SEM) for unsteady flows on one side, which allows to compute the gradient of a cost-functional with respect to (w.r.t.) control parameters *at any time* using a *forward time-integration*, and the Multiple Gradient Descent Algorithm (MGDA) on the other side, which is an extension of the classical steepest-descent method to multiobjective problems and permits to compute a descent direction *common to a possibly-large set of cost-functions*. In this way, the optimization acts at all time and not simply by the control of time-averages.

2 Problem Description: Optimization of Pulsating Jets

We consider as a model problem the two dimensional compressible flow over a flat plate equipped with three periodically oscillating jets (see Fig. 1). The Reynolds number based on the length h is $R = 10^3$ and the flow is laminar, while the Mach number is $M = 10^{-1}$. For the three jets, the crosswise velocity is imposed as:

$$v_k(x,t) = A_k \sin(2\pi N_k t + \varphi_k)\zeta(x) \quad k = 1, 2, 3, \tag{1}$$

where $\zeta(x)$ corresponds to a squared sine distribution. The jet frequencies are set to the fixed values $N_1 = N_\infty$, $N_2 = 2N_\infty$ and $N_3 = 1/2N_\infty$, with $N_\infty = u_\infty/h$. The jet amplitudes and phases are considered as control parameters $\mathbf{x} = \{A_1, A_2, A_3, \varphi_1, \varphi_2, \varphi_3\}$. Initial values are chosen somewhat arbitrarily as $\mathbf{x}_0 = \{u_\infty, 3/2\,u_\infty, 2\,u_\infty, 0, \pi/4, 3\pi/4\}$.

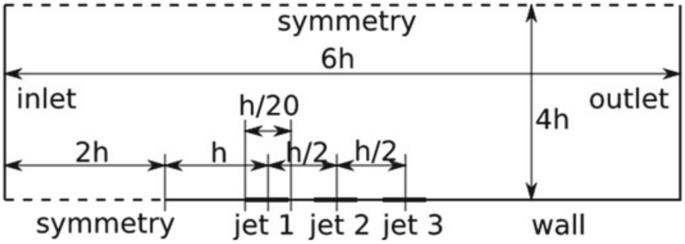


Fig. 1 Problem description

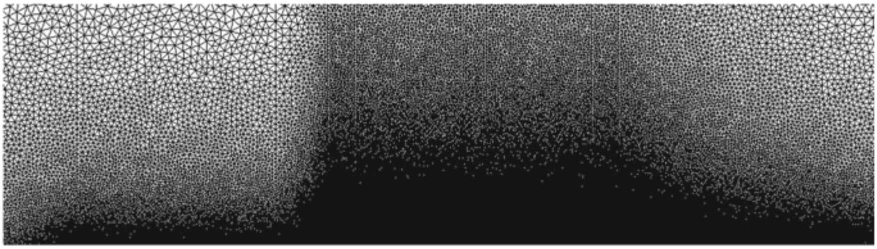


Fig. 2 Computational mesh

The grid employed in this study counts 111 161 nodes (see Fig. 2). The initial solution corresponds to uniform flow based on inlet conditions. The time step is set to $\Delta t = 1/(400N_\infty)$. The unsteady flow tends rapidly to a periodic regime, whose period corresponds to the lowest actuation frequency $1/2N_\infty$ and is thus described by 800 time-steps. As transient effects have vanished, one period is defined as the observation interval (see Fig. 3). Instantaneous velocity fields are shown in Figs. 4 and 5 as illustration.

We consider a set of objective-functions $\{f_j(\mathbf{x})\}_{j=1,\dots,m}$, defined as the values of the drag $\mathscr{J}$, estimated at discrete times $\{t_j\}_{j=1,\dots,m}$, chosen in the observation interval:

$$f_j(\mathbf{x}) = \mathscr{J}(\mathbf{W}_j) \quad \text{with } \mathbf{W}_j = \mathbf{W}(\mathbf{x}, t_j) \quad \forall j \in \{1, \cdots, m\}, \tag{2}$$

where $\mathbf{x} \in \mathbb{R}^n$ represents the vector of the $n = 6$ control parameters and $\mathbf{W}$ the flow variables. The objective of this work is to *reduce simultaneously* these m cost-functions. The following sections describe how the gradient $\nabla_{\mathbf{x}} f_j$ of f_j w.r.t. $\mathbf{x}$ is evaluated, and the optimization algorithm proposed to conduct the simultaneous optimization of these functions.

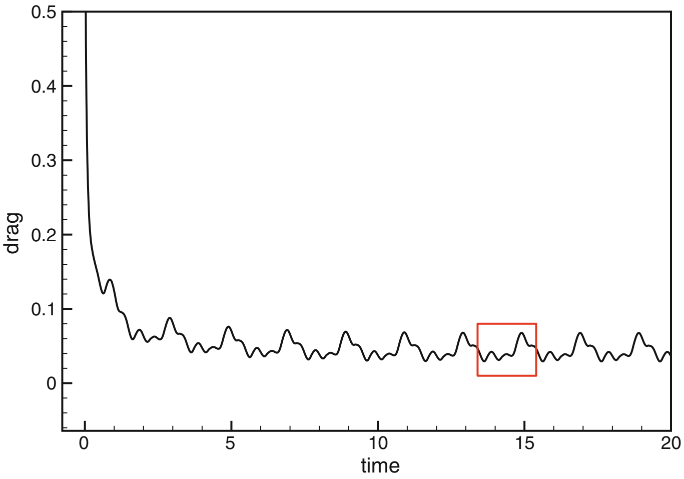


Fig. 3 History of drag for initial parameters and observation area

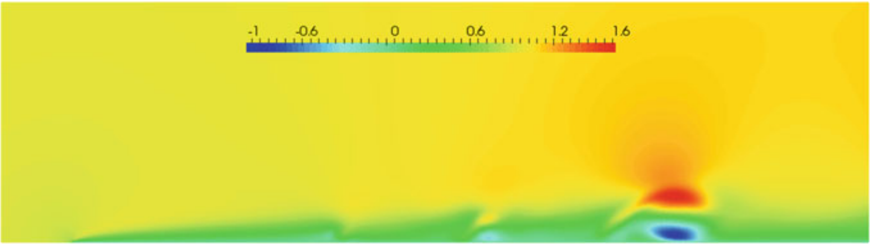


Fig. 4 Snapshot of the streamwise velocity field

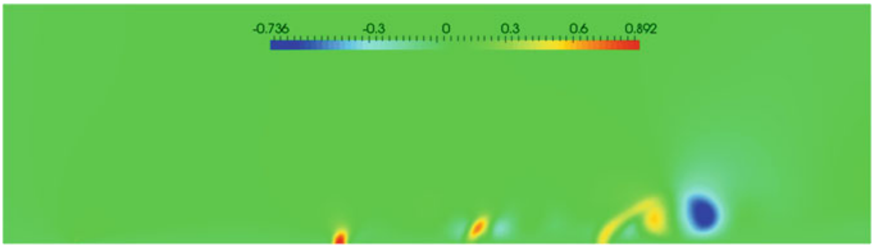


Fig. 5 Snapshot of the crosswise velocity field

3 Sensitivity Analysis for Unsteady Flow

3.1 Method

The governing flow equations are written in conservative form as follows:

$$\frac{\partial \mathbf{W}}{\partial t} + \nabla \cdot \mathscr{F} = \nabla \cdot \mathscr{G}, \tag{3}$$

where $\mathbf{W} = (\rho, \rho u, \rho v, \rho e)$ is the vector of conservative mean-flow variables, ρ is density, u and v are the velocity components, and e the total energy per unit mass; $\mathscr{F} = (\mathbf{F}_x(\mathbf{W}), \mathbf{F}_y(\mathbf{W}))$ and $\mathscr{G} = (\mathbf{G}_x(\mathbf{W}), \mathbf{G}_y(\mathbf{W}))$ are the vectors of convective and diffusive fluxes respectively. Here ∇ stands for the gradient w.r.t. the spatial Cartesian coordinates x and y, and $(\nabla\cdot)$ for the divergence operator. The pressure p is obtained from the perfect-gas state equation:

$$p = \rho(\gamma - 1)(e - \frac{u^2 + v^2}{2}) = \rho(\gamma - 1)e_i \tag{4}$$

where $\gamma = \frac{7}{5}$ is the ratio of the specific heats for diatomic gas, and e_i the internal energy.

The inviscid fluxes are given by:

$$\mathbf{F}_x(\mathbf{W}) = \begin{pmatrix} \rho u \\ \rho u^2 + p \\ \rho u v \\ \rho u (e + \frac{p}{\rho}) \end{pmatrix} \quad \mathbf{F}_y(\mathbf{W}) = \begin{pmatrix} \rho v \\ \rho v u \\ \rho v^2 + p \\ \rho v (e + \frac{p}{\rho}) \end{pmatrix}. \tag{5}$$

The viscous fluxes are written as:

$$\mathbf{G}_x(\mathbf{W}) = \begin{pmatrix} 0 \\ \tau_{xx} \\ \tau_{yx} \\ u\tau_{xx} + v\tau_{yx} - q_x \end{pmatrix} \quad \mathbf{G}_y(\mathbf{W}) = \begin{pmatrix} 0 \\ \tau_{xy} \\ \tau_{yy} \\ u\tau_{xy} + v\tau_{yy} - q_y \end{pmatrix}, \tag{6}$$

where $\bar{\bar{\tau}}$ is the symmetric viscous stress tensor and $\mathbf{q}$ the heat flux.

We can now introduce the sensitivity field $\mathbf{W}'$, which is defined as the derivative of the flow solution $\mathbf{W}$ w.r.t. a given control parameter a, component of $\mathbf{x}$:

$$\mathbf{W}' = \frac{\partial \mathbf{W}}{\partial a}. \tag{7}$$

The equations governing the sensitivity field can be obtained by differentiating (3) w.r.t. a:

$$\frac{\partial}{\partial a}\left(\frac{\partial \mathbf{W}}{\partial t}\right) + \frac{\partial}{\partial a}\left(\nabla \cdot \mathscr{F}\right) = \frac{\partial}{\partial a}\left(\nabla \cdot \mathscr{G}\right). \tag{8}$$

By switching the derivatives w.r.t. a and those w.r.t time or space coordinates, one obtains:

$$\frac{\partial}{\partial t}\left(\frac{\partial \mathbf{W}}{\partial a}\right) + \nabla \cdot \left(\frac{\partial \mathscr{F}}{\partial a}\right) = \nabla \cdot \left(\frac{\partial \mathscr{G}}{\partial a}\right), \tag{9}$$

or:

$$\frac{\partial \mathbf{W}'}{\partial t} + \nabla \cdot \mathscr{F}' = \nabla \cdot \mathscr{G}', \tag{10}$$

which is formally similar to (3), by introducing the sensitivity of the convective flux $\mathscr{F}' = (\mathbf{F}'_x(\mathbf{W}, \mathbf{W}'), \mathbf{F}'_y(\mathbf{W}, \mathbf{W}'))$ and the sensitivity of the diffusive flux $\mathscr{G}' = (\mathbf{G}'_x(\mathbf{W}, \mathbf{W}'), \mathbf{G}'_y(\mathbf{W}, \mathbf{W}'))$. The sensitivity of the convective fluxes can be expressed as:

$$\mathbf{F}'_x(\mathbf{W}, \mathbf{W}') = \begin{pmatrix} (\rho u)' \\ (\rho u)'u + (\rho u)u' + p' \\ (\rho u)'v + (\rho u)v' \\ (\rho u)'(e + \frac{p}{\rho}) + (\rho u)(e' + (\frac{p}{\rho})') \end{pmatrix} \tag{11}$$

$$\mathbf{F}'_y(\mathbf{W}, \mathbf{W}') = \begin{pmatrix} (\rho v)' \\ (\rho v)'u + (\rho v)u' \\ (\rho v)'v + (\rho v)v' + p' \\ (\rho v)'(e + \frac{p}{\rho}) + (\rho v)(e' + (\frac{p}{\rho})') \end{pmatrix}. \tag{12}$$

The sensitivity of the diffusive fluxes reads:

$$\mathbf{G}'_x(\mathbf{W}, \mathbf{W}') = \begin{pmatrix} 0 \\ \tau'_{xx} \\ \tau'_{yx} \\ u'\tau_{xx} + v'\tau_{yx} + u\tau'_{xx} + v\tau'_{yx} - q'_x \end{pmatrix} \tag{13}$$

$$\mathbf{G}'_y(\mathbf{W}, \mathbf{W}') = \begin{pmatrix} 0 \\ \tau'_{xy} \\ \tau'_{yy} \\ u'\tau_{xy} + v'\tau_{yy} + u\tau'_{xy} + v\tau'_{yy} - q'_y \end{pmatrix}, \tag{14}$$

where $\bar{\bar{\tau}}'$ is the sensitivity of the viscous stress tensor and $\mathbf{q}'$ the sensitivity of the heat flux. The boundary conditions for the sensitivity equations are obtained by differentiating the boundary conditions applied to the flow.

Since the flow and sensitivity equations are formally similar, both are solved using the same finite-volume approach [6], based on a second-order vertex-centered discretization scheme. Temporal integration relies on a second-order implicit backward method, with a dual time-stepping technique. Note that the implicit part of the scheme is the same for the flow and sensitivity equations, since both involve the same Jacobian matrix.

The sensitivity equation depends on the parameter a, component of $\mathbf{x}$ of interest. Therefore, six sensitivity equations have to be solved, possibly in parallel, to estimate the components of the gradient for all m cost-functions:

$$\nabla_a f_j = \partial_{\mathbf{W}} \mathscr{J}(\mathbf{W}_j) \cdot \mathbf{W}'(t_j) \quad \forall j \in \{1, \cdots, m\}. \tag{15}$$

We emphasize that if m is large, the solution of six sensitivity equations is far more cost-efficient than solving m adjoint equations backward in time. Additionally, the memory-storage requirement remains moderate.

3.2 Verification

To verify the implementation of the sensitivity equations, we compute a neighboring solution according to a first-order extrapolation $\mathbf{W}(a) + \mathbf{W}'\delta a$ and compare it with the solution $\mathbf{W}(a + \delta a)$. This exercise is conducted in a simplified case including a single jet, a being the jet amplitude A_1. Figures 6 and 7 provide illustrations for the flow fields at selected times and Fig. 8 for the resulting drag history, for a perturbation of the jet amplitude $\delta A_1 = A_1/4$. A similar exercise has been achieved for the phase, to fully verify the gradient estimation.

4 Multiobjective Descent Algorithm MGDA

Equipped with procedures for calculating the objective-functions and their gradients, we now turn to the issue of constructing the multi-objective optimization method.

The Multiple-Gradient Descent Algorithm (MGDA) was originally introduced in [1, 2] to solve general multi-objective optimization problems involving differentiable cost-functions. Variants were proposed in [3], but more recently the algorithm was slightly revised in [4] to apply to cases where the number m of objective-functions exceeds the dimension n of the working design space. We recall here the basic definition of the revised version and provide some details about the application to the present parametric optimization.

4.1 Multi-objective Problem Statement

Let m and n be two arbitrary integers, and consider the multi-objective optimization problem consisting in minimizing m differentiable objective-functions $\{f_j(\mathbf{x})\}$ in some open admissible domain $\Omega_a \subseteq \mathbb{R}^n$ ($j = 1, \ldots, m$; $f_j \in C^1(\Omega_a)$). Given a

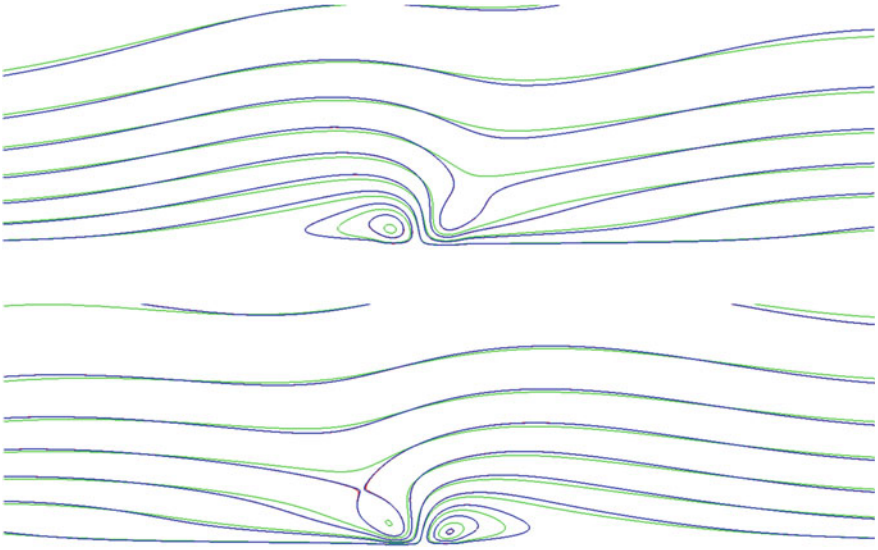


Fig. 6 Linear extrapolation of streamwise velocity field u w.r.t. jet amplitude A_1 for blowing (top) and suction (bottom) phases: reference state $u(A_1)$ in green, extrapolated state $u(A_1) + u'\delta A_1$ in red, non-linear perturbed state $u(A_1 + \delta A_1)$ in blue, for $\delta A_1 = A_1/4$

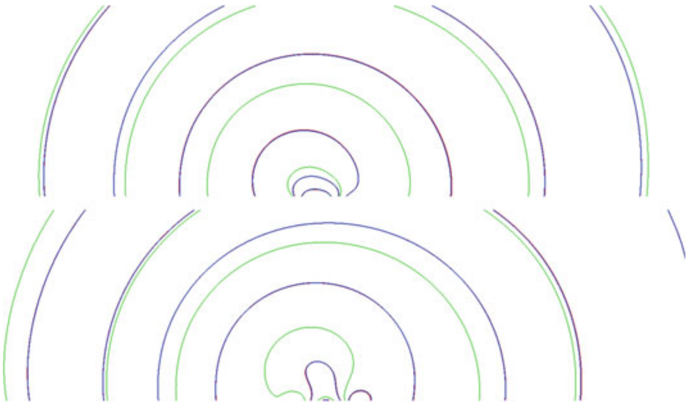


Fig. 7 Linear extrapolation of pressure field p w.r.t. jet amplitude A_1 for blowing (top) and suction (bottom) phases: reference state $p(A_1)$ in green, extrapolated state $p(A_1) + p'\delta A_1$ in red, non-linear perturbed state $p(A_1 + \delta A_1)$ in blue, for $\delta A_1 = A_1/4$

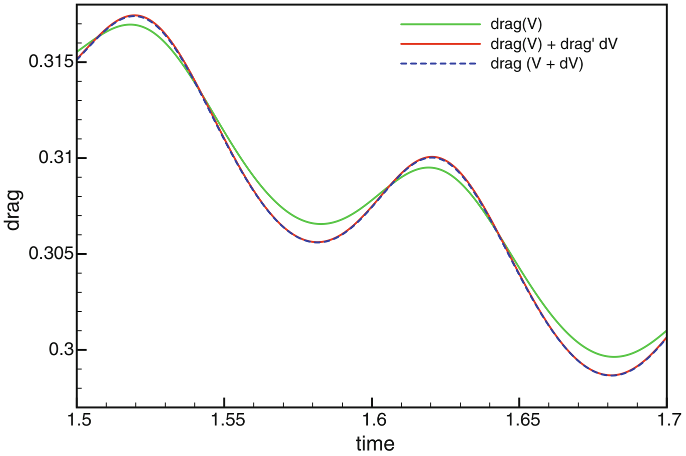


Fig. 8 Linear extrapolation of the drag w.r.t. jet amplitude A_1: reference drag $\mathscr{J}(A_1)$ in green, extrapolated drag $\mathscr{J}(A_1) + \mathscr{J}'\delta A_1$ in red, non-linear perturbed drag $\mathscr{J}(A_1 + \delta A_1)$ in blue, for $\delta A_1 = A_1/4$

starting point $\mathbf{x}_0 \in \Omega_a$ and a vector $\mathbf{d} \in \mathbb{R}^n$, one forms the directional derivatives

$$f_j' = \left[\nabla_{\mathbf{x}} f_j(\mathbf{x}_0)\right]^t \mathbf{d} \tag{16}$$

where $\nabla_{\mathbf{x}}$ is the symbol for the gradient w.r.t. $\mathbf{x}$ and the superscript ^{t} stands for transposition. One seeks for a vector $\mathbf{d}$ such that

$$f_j' > 0 \quad (\forall j). \tag{17}$$

If such a vector $\mathbf{d}$ exists, the direction of vector $(-\mathbf{d})$ is said to be a local descent direction common to all objective-functions. Then evidently, infinitely-many other such directions also exist, and our algorithm permits to identify at least one.

4.2 Convex Hull, Two Lemmas and Basic MGDA

We recall the following:

Definition 1 The convex hull of a family of m vectors $\{\mathbf{u}_j\}$ $(j = 1, \dots, m; \mathbf{u}_j \in \mathbb{R}^n)$, is the set of all their convex combinations:

$$\overline{\mathsf{U}} = \left\{ \mathbf{u} \in \mathbb{R}^n \text{ such that } \mathbf{u} = \sum_{j=1}^{m} \alpha_j \mathbf{u}_j;\ \alpha_j \in \mathbb{R}+;\ \sum_{j=1}^{m} \alpha_j = 1 \right\}. \tag{18}$$

Then, we have:

Lemma 1 *Given an $n \times n$ real-symmetric positive-definite matrix $\mathbf{A}_n$, the associated scalar product*

$$\big(\mathbf{u}, \mathbf{v}\big) = \mathbf{u}^t \mathbf{A}_n \mathbf{v} \quad (\mathbf{u}, \mathbf{v} \in \mathbb{R}^n), \tag{19}$$

and Euclidean norm

$$\|\mathbf{u}\| = \sqrt{\mathbf{u}^t \mathbf{A}_n \mathbf{u}}, \tag{20}$$

the convex hull $\overline{\mathsf{U}}$ admits a unique element ω of minimum norm.

Proof Existence: $\overline{\mathsf{U}}$ is closed and $\|\cdot\|$ is a continuous function.

Uniqueness: Suppose that ω_1 and ω_2 are two realizations of the minimum $\mu = \arg\min_{\mathbf{u}\in\overline{\mathsf{U}}} \|\mathbf{u}\|$ so that $\mu = \|\omega_1\| = \|\omega_2\|$ and let

$$\omega_s = \frac{1}{2}\left(\omega_2 + \omega_1\right), \quad \omega_d = \frac{1}{2}\left(\omega_2 - \omega_1\right),$$

so that:

$$\big(\omega_s, \omega_d\big) = \frac{1}{4}\big(\omega_2 + \omega_1, \omega_2 - \omega_1\big) = \frac{1}{4}\left(\|\omega_2\|^2 - \|\omega_1\|^2\right) = 0.$$

Hence $\omega_s \perp \omega_d$, and since $\omega_s \in \overline{\mathsf{U}}$, $\|\omega_s\| \geq \mu$, and:

$$\mu^2 = \|\omega_2\|^2 = \|\omega_s + \omega_d\|^2 = \|\omega_s\|^2 + \|\omega_d\|^2 \geq \mu^2 + \|\omega_d\|^2 \Longrightarrow \omega_d = 0. \qquad \square$$

Lemma 2 *The minimum-norm element ω defined in Lemma 1 satisfies:*

$$\forall u \in \overline{\mathsf{U}}, \quad \big(\mathbf{u}, \omega\big) \geq \|\omega\|^2. \tag{21}$$

Proof Let $\mathbf{u} \in \overline{\mathsf{U}}$, arbitrary. Let $\delta = \mathbf{u} - \omega$; by convexity of $\overline{\mathsf{U}}$:

$$\forall \varepsilon \in [0, 1], \ (1 - \varepsilon)\omega + \varepsilon \mathbf{u} = \omega + \varepsilon\delta \in \overline{\mathsf{U}},$$

and by definition of ω, $\|\omega + \varepsilon\delta\| \geq \|\omega\|$, that is:

$$\big(\omega + \varepsilon\delta, \omega + \varepsilon\delta\big) - \big(\omega, \omega\big) = 2\varepsilon\big(\omega, \delta\big) + \varepsilon^2 \|\delta\|^2 \geq 0,$$

and this requires that the coefficient of ε be non-negative. $\square$

Then consider the case where

$$\mathbf{u}_j = \nabla_{\mathbf{x}} f_j(\mathbf{x}_0) \quad (\forall j). \tag{22}$$

If the vector ω defined in Lemma 1 is nonzero, the vector

$$\mathbf{d} = \mathbf{A}_n \omega \tag{23}$$

is also nonzero, and is a solution to the problem stated in (16)-(17) since by virtue of Lemma 2:

$$\big(\mathbf{u}_j, \omega\big) = \mathbf{u}_j^t \mathbf{A}_n \omega = \mathbf{u}_j^t \mathbf{d} \geq \|\omega\|^2 > 0. \tag{24}$$

The situation in which $\omega = 0$, or equivalently,

$$\exists \alpha = \{\alpha_j\} \in \mathbb{R}+^m \text{ such that } \sum_{j=1}^{m} \alpha_j \nabla f_j(\mathbf{x}_0) = 0 \text{ and } \sum_{j=1}^{m} \alpha_j = 1, \tag{25}$$

is said to be one of "Pareto-stationarity". The relationship between Pareto-optimality and Pareto-stationarity was made precise by the following [3, 4]:

Theorem 1 *If the objective-functions are differentiable and convex in some open ball $\mathscr{B} \subseteq \Omega_a$ about $\mathbf{x}_0$, and if $\mathbf{x}_0$ is Pareto-optimal, then the Pareto-stationarity condition is satisfied at $\mathbf{x}_0$.*

Hence, the Pareto-stationarity condition generalizes to the multi-objective context, the classical stationarity condition expressing that an unconstrained differentiable function is extremal.

We now return to the non-trivial case of a point $\mathbf{x}_0$ that is *not Pareto-stationary* and we suppose that the vectors ω and $\mathbf{d}$ ($\omega \neq 0$; $\mathbf{d} \neq 0$) have been identified (see next subsection). Then we define MGDA as the iteration which transforms $\mathbf{x}_0$ in

$$\mathbf{x}_1 = \mathbf{x}_0 - \rho \mathbf{d} \tag{26}$$

where $\rho > 0$ is some appropriate step-size. Thus MGDA is an extension to the multi-objective context of the classical steepest-descent method, in which the direction of search is taken to be the vector $\mathbf{d}$ defined above. At convergence, the limiting point is Pareto-stationary.

We now examine how can the vector $\mathbf{d}$ be computed in practice.

4.3 QP Formulation and Hierarchical Gram-Schmidt Orthogonalization

By letting

$$\omega = \sum_{j=1}^{m} \alpha_j \mathbf{u}_j = \mathbf{U}\alpha \tag{27}$$

where $\mathbf{u}_j = \nabla f_j(\mathbf{x}_0)$, $\mathbf{U}$ is the $n \times m$ matrix whose jth column contains the n components of vector $\mathbf{u}_j$, the identification of vector ω can be made by solving the following Quadratic-Programming (QP) problem for the unknown vector of coefficients $\alpha = \{\alpha_j\}$:

$$\omega = \arg\min_{\alpha \in \mathbb{R}^m} \frac{1}{2}\alpha^t \mathbf{H}\alpha \tag{28}$$

subject to:

$$\alpha_j \geq 0 \ (\forall j), \quad \sum_{j=1}^{m} \alpha_j = 1, \tag{29}$$

where $\mathbf{H} = \mathbf{U}^t \mathbf{A}_n \mathbf{U}$. Note that if vector ω is unique, vector α may not be.

If the family of gradients is linearly-independent, which requires in particular that $m \leq n$, it is possible to choose the scalar product, through the definition of matrix $\mathbf{A}_n$, in such a way that these gradients form an orthogonal basis of their span. Then vector ω is explicitly determined by the orthogonal projection of 0 onto the convex hull $\overline{\mathsf{U}}$:

$$\alpha_j = \frac{1}{\left\|\mathbf{u}_j\right\|^2 \sum_{k=1}^{m} \frac{1}{\|\mathbf{u}_k\|^2}} \tag{30}$$

In the inverse case where $m > n$ (and even $m \gg n$), focus of interest presently, let $\mathsf{r} \leq n < m$ be the rank of the family of gradients. Using first the standard Euclidean scalar product ($\mathbf{A}_n = \mathbf{I}_n$), the Gram-Schmidt orthogonalization process stops in r steps and produces an orthogonal basis (in the usual sense) $\{\mathbf{v}_j\}$ ($j = 1, \ldots, \mathsf{r}$), that span a subspace $\mathscr{G}$ of dimension r. The orthogonal vectors, $\{\mathbf{v}_j\}$, are calculated from a subfamily of the original vectors, $\{\mathbf{u}_j\}$, $j \in J$, where J is a subfamily of r indices from 1 to m. In this process, a hierarchical principle was introduced in [4] to select these indices in such a way that the cone bounded by the reduced family $\{\mathbf{u}_j\}$ ($j \in J$) be as large as possible to contain the directions, in the most favorable situation, of all the other gradients, unused in the Gram-Schmidt process by redundancy. When this occurs, the subfamily $\{\mathbf{u}_j\}$ ($j \in J$) not only is a basis of the subspace $\mathscr{G}$, but also, its convex hull contains all the directions of interest. In the more general case where the directions of some gradients among the unused vectors $\{\mathbf{u}_j\}$ ($j \notin J$) are not in the cone, we resort to the QP-formulation, but expressed after changing the basis to become $\{\mathbf{u}_j\}$ ($j \in J$) and by choosing a new scalar product to make this subfamily orthogonal. The technical steps are the following [4]:

- Once the $n \times m$ matrix $\mathbf{U}$ is formed with the components of the given gradients $\{\mathbf{u}_j\}$ ($\mathbf{u}_j \in \mathbb{R}^n$, $j = 1, \ldots, m$), these gradients, or column-vectors are made (physically) dimensionless, by component-wise normalization to form the initial matrix $\mathbf{G}$:

$$\forall i \in 1, \ldots, n: \ s_i = \max_j \left|\mathbf{u}_{i,j}\right|, \quad \mathbf{S} = \mathbf{Diag}(s_i), \quad \mathbf{G} = \mathbf{S}^{-1}\mathbf{U}. \tag{31}$$

This normalization is essential to make the subsequent calculations of scalar products physically meaningful and computationally well-balanced.

- Throughout the Gram-Schmidt process, columns of the $\mathbf{G}$ matrix are permuted by the hierarchical selection of basis vectors. Upon exit, the actually used gradients are placed in the first r columns. The corresponding $n \times \mathsf{r}$ leftmost block of the final matrix $\mathbf{G}$ is then denoted $\underline{\mathbf{G}}$. Note that in the present version of the Gram-Schmidt process, the computed orthogonal vectors $\{\mathbf{v}_j\}$ are not normalized to unity, but in a special way for which r directional derivatives are equal [4]. Let these vectors be stored in matrix $\mathbf{V}$, and define the following diagonal matrix:

  $$\Delta = \mathbf{Diag}\big(\mathbf{v}_j^t \mathbf{v}_j\big). \tag{32}$$

  Then:

  $$\mathbf{A}_n = \mathbf{W}^t\mathbf{W} + (\mathbf{I} - \Pi)^2, \quad \mathbf{W} = \big(\underline{\mathbf{G}}^t\underline{\mathbf{G}}\big)^{-1}\underline{\mathbf{G}}^t\mathbf{V}\Delta^{-1}\mathbf{V}^t, \tag{33}$$

  where $\Pi = \mathbf{V}\Delta^{-1}\mathbf{V}^t$ is the projection matrix onto subspace $\mathscr{G}$ [4].

In this way, the QP-formulation is well-conditioned and easily solved by a library procedure. We have used the procedure qpsolve from the Scilab library which is equivalent to the quadprog procedure from the MATLAB library. As a result of this, exactly r directional derivatives $\{f_j'\}$ are equal, and by experience, the remaining ones differ only slightly.

5 Results

For each control parameter, component of $\mathbf{x} = \{A_1, A_2, A_3, \varphi_1, \varphi_2, \varphi_3\}$, the SEM is applied to obtain sensitivity fields. We provide as illustration (see Figs. 9 and 10) instantaneous sensitivity fields of the velocity w.r.t. the amplitude for the first jet. The derivatives of the drag w.r.t. the control parameters are then computed at all time-steps, yielding 800 values of cost-function and gradient for the whole observation period, as illustrated in Fig. 11.

We aim now at determining the vector of parameters $\mathbf{x} = \{A_1, A_2, A_3, \varphi_1, \varphi_2, \varphi_3\}$ that reduces simultaneously the 800 cost-functions associated with the observation

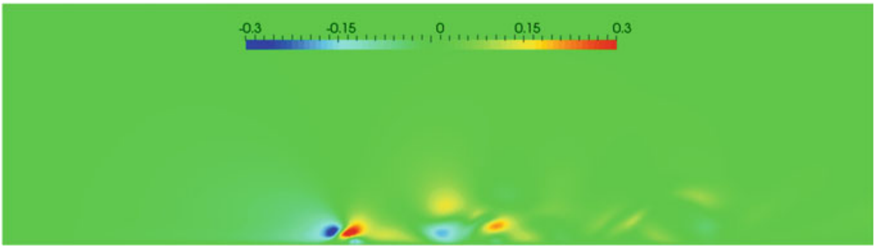


Fig. 9 Sensitivity of streamwise velocity w.r.t. first jet amplitude

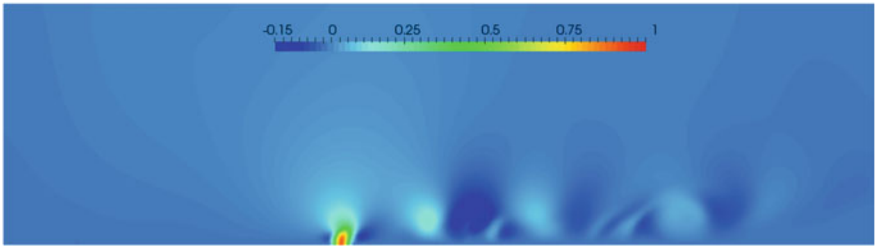


Fig. 10 Sensitivity of crosswise velocity w.r.t. first jet amplitude

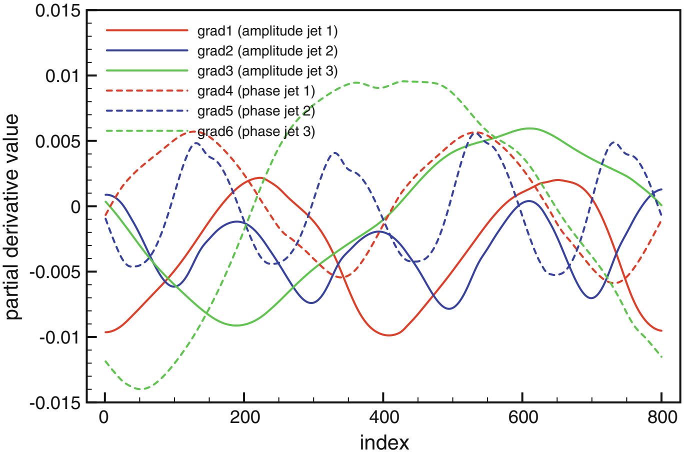


Fig. 11 Gradient components for the 800 cost-functionals

period. To reduce somewhat the computational complexity without altering greatly the transient behavior, the MGDA approach is applied to only $m = 20$ homogenized gradients, obtained by averaging the gradients by time-intervals of 40 time-steps. As a result, one disposes of a common descent direction associated with vector $\mathbf{d}$ satisfying (17).

Once the vector $\mathbf{d}$ is determined, a practical step-size ρ must be estimated. For this, we first note that a natural scale for the variations of a time-dependent objective-function is given by its standard deviation, $\bar{\sigma}$, a more significant value than its average which can be 0. Then if $\delta\mathbf{x} = -\bar{\rho}\mathbf{d}$, the variation of the objective-function average can be estimated as $-\bar{\rho}\bar{g} \cdot \mathbf{d}$ where $\bar{g}$ is the average gradient. Thus, a meaningful reference step-size can be defined by the condition:

$$\bar{\rho}\bar{g} \cdot \mathbf{d} = \bar{\sigma} \tag{34}$$

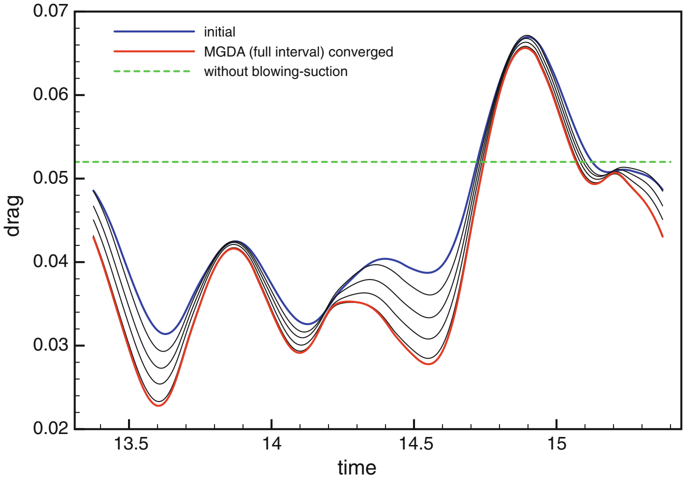


Fig. 12 Evolution of the drag history w.r.t. optimization iterations: case of MGDA approach (blue: initial, red: final, black: intermediate iterations)

where $\bar{\sigma} = \sqrt{\frac{1}{m}\sum_{j=1}^{m}\left(f_j - \bar{f}\right)^2}$, $\bar{f} = \frac{1}{m}\sum_{j=1}^{m} f_j$, and $\bar{g} = \frac{1}{m}\sum_{j=1}^{m} \nabla f_j(\mathbf{x}_0)$. This gives:

$$\bar{\rho} = \frac{\bar{\sigma}}{\bar{g}\cdot\mathbf{d}}. \tag{35}$$

In practice, in the present experiments, we have used the step-size $\rho = \frac{1}{10}\bar{\rho}$ to update the control parameters by the descent method, and this resulted in a successful iteration, stable and effective.

The history of the drag in the observation period is represented in Fig. 12, from the baseline flow to full convergence of the optimization approach. As expected, *each update of the design vector has resulted in a diminished drag over the entire observation period*. As it can be noticed, some points in time are more critical than others. In contrast, when one applies, more classically, the steepest-descent method to the time-averaged cost-function $\overline{\mathscr{J}}$, by setting the search direction to the average gradient, an increase of the drag can be observed at some times, as illustrated in Fig. 13. Finally, also note that actuation permits a significant drag reduction w.r.t. the case without suction/blowing for which the value of drag is indicated on the figure by a dotted horizontal line.

Finally, a second exercise is conducted: the drag values computed over the last 40% of the observation period only are considered as optimization criteria in the MGDA approach (in this interval, the drag is especially high for the baseline flow). The history of the drag in the observation period is represented in Fig. 14, for the first 16 iterations of the optimization algorithm. As expected, a more significant decrease

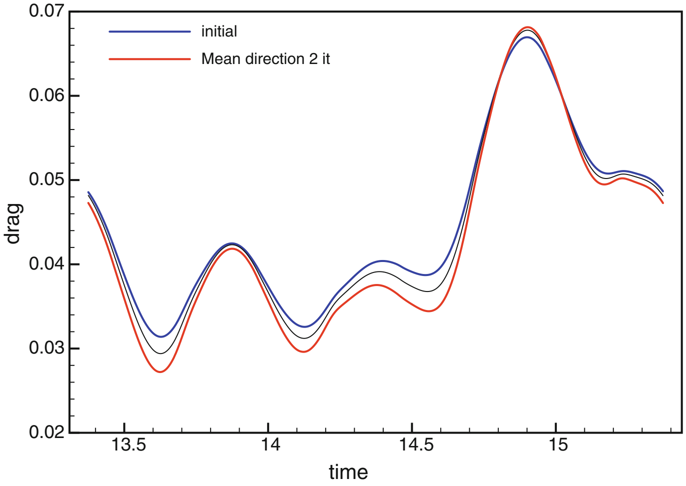


Fig. 13 Evolution of the drag history w.r.t. optimization iterations: case of a mean direction descent (blue: initial, red: after 2 iterations, black: intermediate iteration)

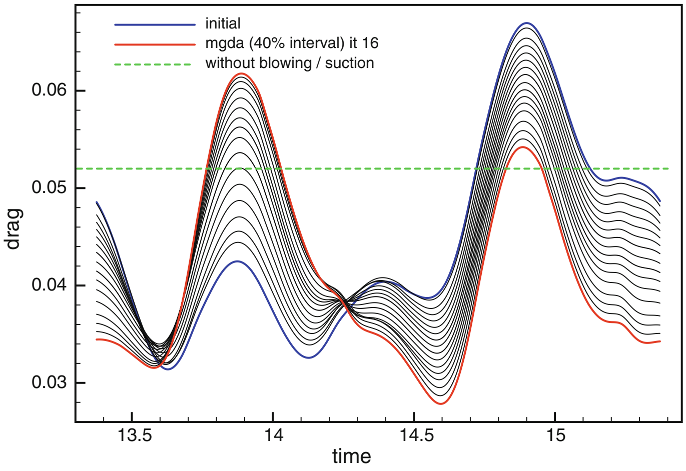


Fig. 14 Evolution of the drag history w.r.t. optimization iterations: case of MGDA approach based on the last 40% of the observation period (blue: initial, red: after 7 iterations, black: intermediate iterations)

is achieved during the last 40% of the observation period, whereas the drag is free to vary in the first 60%, and in fact, increases.

Remark 1 In the present test-case, the flow is periodic in time. However, the periodic boundary conditions are not all known. The simulated flow relaxes towards the time-periodic solution. In practice, due to the small time-step used in this simulation in relation with the pulsating jets frequencies, it takes an integration interval of about 12 periods for the numerical solution to be almost exactly periodic. The solution over a time-interval of 11 periods is simply discarded, and the gradients are calculated only over the following time-interval of one period. Then, the jets parameters are updated, and a new integration is carried out over 12 periods of time, and so on. We proceeded in this way to produce verified results, and clear conclusions. However, in the future, for greater efficiency, the optimization update and the time integration should be performed simultaneously, in a "one-shot" type method.

Remark 2 In the present jet-optimization problem, increasing the number m of accounted gradients is not necessarily a cause of worse-conditioning, if it is combined with an appropriate homogenization of the gradients in which only average gradients are used to compute the descent direction. Here the objective-functions are in large number because they are discrete realizations of a function that evolves with time, in fact rather smoothly. Thus, the critical issue is to account sufficiently accurately for the unsteady features of the time-dependent phenomenon. When this is achieved, increasing m has little or no effect on the convex hull and thus, on the descent direction. For example, in the above experiment, at iteration 6, after computing a new descent direction based on 20 average gradients, the corresponding directional derivatives associated with the 800 initially-available gradients were all observed *a posteriori* to be positive with an acceptable dispersion (standard deviation of about 56 % of the average). This observation confirmed that considering only 20 averages was adequate. However, we have not studied the efficacy and stability of the MATLAB procedure in case of a very large redundant set of input gradients.

Remark 3 When approaching convergence, $\omega \to 0$, and the descent step becomes ineffective. In the above experiment, this was observed after some ten iterations. From the numerical viewpoint, the determination of ω involves a Gram-Schmidt process, a basis change and a call to the MATLAB procedure quadprog. If round-off errors lead to a small but erroneous ω, and a departure from the neighborhood of the Pareto-front, the next MGDA iteration involves a new ω of larger norm, thus more accurately identified, and a step back to the front. Therefore, if the step-size is well controlled, the risk of instability is small.

Remark 4 As mentioned above, when approaching convergence, $\|\omega\| \to 0$. In the limit, the Pareto-stationarity condition is satisfied:

$$\sum_{j=1}^{m} \alpha_j \nabla f_j(\mathbf{x}^\star) = 0 \tag{36}$$

at some limit point $\mathbf{x}^\star$ and for some coefficients $\{\alpha_j\}$. These elements depend on the convergence path, hence to some extent, on the initial point, and are numerically known. If the Pareto front, or a portion of it, is to be determined more completely, the trivial possibility exists to restart the MGDA iteration from a broad set of starting points from scratch, preferably by exploiting parallel computing resources. However, alternately, one can also exploit the fact that a point on the Pareto set has been identified already. At $\mathbf{x} = \mathbf{x}^\star$, the auxiliary objective-function

$$f_A(\mathbf{x}) = \sum_{j=1}^{m} \alpha_j f_j(\mathbf{x}) \tag{37}$$

is stationary in the usual sense. One way to obtain other neighboring points of the Pareto front, is to alternate MGDA with a virtual Nash game. The Nash game involves two players, A and B. Player A's strategy attempts to maintain the Pareto-stationarity condition by iterations directed to minimize $f_A(\mathbf{x})$. Player B's strategy consists of minimization iterations of an alternate and conflicting criterion J_B. In the jets problem, the objective-function $f_B(\mathbf{x})$ could be the average drag over all time-steps, or over a relevant zone of interest. It has been established [5], that by an appropriate split of territory, in which the design space is split into two well-defined supplementary subspaces related to the diagonalization of the Hessian of f_A, possibly under equality constraints, a continuum of points in the neighborhood of $\mathbf{x}^\star$ can be obtained in this way, producing a path in function-space tangent to the Pareto front. Then, by alternating MGDA with the Nash game, a portion of the Pareto-front in the neighborhood of $\mathbf{x}^\star$ can be identified piecewise by zig-zag paths. This technique was applied successfully in problems of optimum-shape design in aerodynamics, in [12] in a case of optimization of the simplified fuselage-wing configuration of a supersonic business jet with respect to wave drag reduction under lift constraint versus sonic-boom reduction, and in [11] in a case of optimization of an helicopter rotor blade with respect to figure of merit in hover conditions versus mechanical power to be developed to permit forward motion.

6 Conclusion

In this work, we solved for demonstration an exercise of active-flow control in which drag over a flat plate has been reduced by three pulsating jets acting on the boundary layer. The flow, governed by the time-dependent compressible Navier-Stokes equations, has been simulated numerically by second-order in time and space finite-volumes, yielding, when the periodic regime is achieved, drag as a function of time. The simultaneous solution of the sensitivity equations has provided additionally the six-component gradient of drag w.r.t. the design characteristics of the jets. The accuracy of these gradients has been verified by comparison with finite-differences via fine-mesh computations.

By the Multiple-Gradient Descent Algorithm (MGDA), a direction of search has been identified permitting to reduce drag at all times of the period, or a selected segment of it. The process was repeated iteratively, and at all intermediate steps of the optimization process as well as at convergence, the drag was effectively reduced over the entire observation time-interval.

Hence, we dispose of a numerical optimization tool whose efficacy is demonstrated uniformly, that is, over a possibly-large range of operational conditions, here different discretization times. This contrasts with more classical approaches in which a single functional, usually defined as a somewhat arbitrary weighted average, is minimized at the risk of a degradation of certain elements composing the average.

This method is currently being extended to solve more general robust design problems.

References

1. Désidéri J-A (2012) Multiple-gradient descent algorithm (MGDA). Research Report 6953, Inria, 2009 (revised version November 5). http://hal.inria.fr/inria-00389811/fr/
2. Désidéri J-A (2012) Multiple-gradient descent algorithm (MGDA) for multiobjective optimization. C R Math Acad Sci Paris 350(5–6):313–318
3. Désidéri J-A (2014) Multiple-gradient descent algorithm for pareto-front identification. In: Fitzgibbon W, Kuznetsov YA, Neittaanmäki P, Pironneau O (eds), Modeling, Simulation and Optimization for Science and Technology. Computational Methods in Applied Sciences, vol 34, Springer, Berlin, pp 41–58
4. Désidéri J-A (2015) Révision de l'algorithme de descente à gradients multiples (MGDA) par orthogonalisation hiérarchique. Research Report 8710, Inria. https://hal.inria.fr/hal-01139994
5. Désidéri J-A, Duvigneau R, Habbal A (2014) Multi-objective design optimization using Nash games. In: Vassile M, Becerra VM (eds) Computational Intelligence in Aerospace Sciences. American Institute of Aeronautics and Astronautics, Reston, pp 583–641
6. Duvigneau R (2015) A sensitivity equation method for unsteady compressible flows: implementation and verification. Research Report 8739, Inria
7. Duvigneau R, Hay A, Visonneau M (2007) Optimal location of a synthetic jet on an airfoil for stall control. J Fluids Eng 129(7):825–833
8. Duvigneau R, Labroquère J, Guilmineau E (2016) Comparison of turbulence closures for optimized active control. Comput Fluids 124:67–77
9. Duvigneau R, Visonneau M (2006) Optimization of a synthetic jet actuator for aerodynamic stall control. Comput Fluids 35(6):624–638
10. Gad El-Hack M, Pollard A, Bonnet J-P (1998) Flow control: fundamentals and practices. Springer, Berlin
11. Roca León E (2014) Simulations aéro-mécaniques pour l'optimisation de rotors d'hélicoptère en vol d'avancement (Aero-mechanical simulations for the optimization of rotor blades of helicopter in forward motion). PhD thesis, University Nice Sophia-Antipolis. http://www.theses.fr/2014NICE4076
12. Minelli A (2013) Aero-acoustic Shape Optimization of a Supersonic Business Jet. PhD thesis, University Nice Sophia-Antipolis. http://www.theses.fr/2013NICE4107

Mixed Formulation of a Linearized Lubrication Fracture Model in a Poro-elastic Medium

Vivette Girault, Mary F. Wheeler, Kundan Kumar and Gurpreet Singh

Abstract We analyse and discretize a mixed formulation for a linearized lubrication fracture model in a poro-elastic medium. The displacement of the medium is expressed in primary variables while the flows in the medium and fracture are written in mixed form, with an additional unknown for the pressure in the fracture. The fracture is treated as a non-planar surface or curve according to the dimension, and the lubrication equation for the flow in the fracture is linearized. The resulting equations are discretized by finite elements adapted to primal variables for the displacement and mixed variables for the flow. Stability and a priori error estimates are derived. A fixed-stress algorithm is proposed for decoupling the computation of the displacement and flow and a numerical experiment is included.

Keywords Poro-elasticity · Biot · Lubrication · Mixed formulation
Finite-elements · Fixed stress split algorithm

V. Girault (✉)
Sorbonne Universités, UPMC Univ. Paris 06, CNRS, UMR 7598,
Laboratoire Jacques-Louis Lions, 4, place Jussieu, 75005 Paris, France
e-mail: girault@ann.jussieu.fr

M. F. Wheeler · G. Singh
Center for Subsurface Modeling, Institute for Computational Engineering
and Sciences, The University of Texas at Austin, Austin, TX 78712, USA
e-mail: mfw@ices.utexas.edu

G. Singh
e-mail: gurpreet@ices.utexas.edu

K. Kumar
Porous Media Group, Mathematics Institute, University of Bergen,
Allegaten 41, 5007 Bergen, Norway
e-mail: kundan.kumar@uib.no

B. N. Chetverushkin et al. (eds.), *Contributions to Partial Differential Equations and Applications*, Computational Methods in Applied Sciences 47,
https://doi.org/10.1007/978-3-319-78325-3_12

1 Introduction

The injection of large volumes of fluids in the subsurface such as during carbon sequestration or during hydraulic fracturing operations can cause geomechanical deformation of the rock mass in the vicinity of the injection well. In addition, recovery predictions from a fractured reservoir are essential for long term production in shale oil and gas fields. Understanding interactions between in-situ stresses, injection fluid pressure and fracture is a difficult and challenging issue because of the complexity of rock properties and physical aspects of rock failure and fracture. In this work, we consider a simplified model for the coupled reservoir-fracture flow which accounts for varying reservoir geometries and complexities including non-planar fractures of small width. Here we utilize different flow models such as Darcy flow and Reynolds' lubrication equation for fractures and reservoir respectively to closely capture the physics.

Furthermore, the geomechanics effects have been included by considering Biot's model. An accurate modeling of solid deformations necessitates a better estimation of fluid pressure inside fractures. We model the fractures and reservoirs explicitly, which allows us to capture the flow details and impact of fractures more accurately. The small width assumption allows modeling fractures geometrically as curves in two dimensions or surfaces in three dimensions, while their variable widths are taken into account by jumps in the displacements along the fractures. Numerically, this model has the advantage of avoiding the mesh of the very narrow regions occupied by fractures. The approach presented here is in contrast with existing averaging approaches such as dual and discrete dual porosity models where the effects of fractures are averaged out.

1.1 Problem Setting

The coupled reservoir-fracture flow problem is discretized by a mixed finite element method, because this method is locally mass conservative and the flux values are continuous, see Ingram et al. [19] and Wheeler et al. [27]. The pressure degrees of freedom are defined at the grid cell centers, similar to the finite difference scheme widely used in petroleum reservoir simulations. Moreover, our motivation in applying a mixed formulation is that in realistic engineering settings, it is necessary to transport proppant; thus local conservation is essential. The coupled flow and geomechanics model developed, for fractured porous medium, has the following advantages:

1. The fracture flow problem is resolved explicitly resulting in an accurate fracture pressure used as a traction boundary condition for reservoir geomechanics.
2. A physically accurate formulation of fractures and reservoir flow problems is achieved by using different constitutive equations and capillary pressure curves for each of the two domains in the case of multiphase flow.
3. Non-planar fractures can be captured using a coarser mesh (lower computational cost) due to non-planar faces of the general hexahedral elements inherent to the discretization.

In this work, we prove existence and uniqueness of the solution of a coupled linearized system with one fracture under fairly weak assumptions on the data. To this date, the analysis of the coupled non-linear system is still an open problem. Our results here represent an extension of a previous article, see [18], in which a continuous Galerkin method for flow was analyzed. In the present situation, switching from a continuous Galerkin scheme to a mixed scheme is not completely straightforward, because the discontinuous approximation of the pressure in the fracture (such as piecewise constants in each element) requires a special analysis in coupling the flow in the fracture with that in the reservoir. This coupling requires the derivation of an inf-sup condition in a norm that is weaker than that used in the mixed form of the exact problem, compare (47) and (27). Such discrepancy in the norms, that arises from the discontinuity of the pressure, complicates the numerical analysis. The resulting system is then solved by a fixed stress splitting algorithm introduced and analyzed by Mikelić and Wheeler in [23] for a Biot system without fracture, and in Girault et al. [14] with a fracture. Of course, the numerical experiment reported in this work is applied to the fully non-linear system, where the permeability in the fracture is related to its width.

The Biot system without fracture has been analyzed by a number of authors who established existence, uniqueness, and regularity, see Showalter [26] and references therein, Phillips and Wheeler [24], Girault et al. [15]. Several articles by Mikelić et al. (see, for instance, [4, 11]) treat homogenization of flows through fractured porous media. Another approach consists in treating a fracture as a thin domain in the framework of domain decomposition. We refer the reader to the extensive work of Jaffré, Roberts and co-authors on Darcy flow, see [1, 22].

After this introduction, the paper is organized as follows. The modeling equations are described in Sect. 2. In Sect. 3, the equations are linearized and set into variational formulations. Existence and uniqueness of solutions of the linearized formulation are established in Sect. 4. In Sect. 5, we propose and analyze a fully discrete scheme: backward Euler in time, continuous Galerkin for elasticity and mixed finite elements for flow, more precisely RT_k on simplices and enhanced BDM on quadrilaterals or hexahedra. The enhanced BDM elements are necessary to guarantee sufficient accuracy in the case of quadrilaterals or hexahedra, which cannot be achieved by RT_k elements. In Sect. 6 we present a fixed-stress method as a decoupling computational algorithm. Numerical results are presented in Sect. 7.

1.2 Notation

Let Ω be a bounded domain (open and connected) of $\mathbb{R}^d$, where the dimension $d = 2$ or 3, with a Lipschitz continuous boundary $\partial\Omega$, and let Γ be an open subset of $\partial\Omega$ with positive measure. When $d = 3$, we assume that the boundary of Γ is also Lipschitz continuous. Let $\mathfrak{D}(\Omega)$ be the space of all functions that are infinitely differentiable and with compact support in Ω and let $\mathfrak{D}'(\Omega)$ be its dual space, i.e., the space of distributions in Ω. As usual, for $1 \le p < \infty$, we define the Banach space $W^{1,p}(\Omega)$ by

$$W^{1,p}(\Omega) = \{v \in L^p(\Omega) \mid \nabla v \in L^p(\Omega)^d\},$$

normed by

$$|v|_{W^{1,p}(\Omega)} = \|\nabla v\|_{L^p(\Omega)}, \quad \|v\|_{W^{1,p}(\Omega)} = \big(\|v\|^p_{L^p(\Omega)} + |v|^p_{W^{1,p}(\Omega)}\big)^{\frac{1}{p}},$$

with the usual modification when $p = \infty$. When $p = 2$, $W^{1,2}(\Omega)$ is the classical Hilbert Sobolev space $H^1(\Omega)$. The space of traces of functions of $H^1(\Omega)$ on Γ (or on any Lipschitz curve when $d = 2$, or surface when $d = 3$, in $\overline{\Omega}$) is $H^{\frac{1}{2}}(\Gamma)$, which is a proper subspace of $L^2(\Gamma)$. Its dual space is denoted by $H^{-\frac{1}{2}}(\Gamma)$. Several equivalent norms can be used on this space. Here, it is convenient to use the semi-norm and norm, see, for example, [21]:

$$|v|_{H^{\frac{1}{2}}(\Gamma)} = \left(\int_\Gamma \int_\Gamma \frac{|v(\boldsymbol{x}) - v(\boldsymbol{y})|^2}{|\boldsymbol{x} - \boldsymbol{y}|^d} d\boldsymbol{x}\, d\boldsymbol{y}\right)^{\frac{1}{2}}, \quad \|v\|_{H^{\frac{1}{2}}(\Gamma)} = \big(\|v\|^2_{L^2(\Gamma)} + |v|^2_{H^{\frac{1}{2}}(\Gamma)}\big)^{\frac{1}{2}}.$$

Then we define

$$H^1_0(\Omega) = \{v \in H^1(\Omega)\,;\; v|_{\partial\Omega} = 0\},$$

and more generally

$$H^1_{0,\Gamma}(\Omega) = \{v \in H^1(\Omega)\,;\; v|_\Gamma = 0\}.$$

For a vector $\boldsymbol{v}$ in $\mathbb{R}^d$, recall the strain (or symmetric gradient) tensor $\boldsymbol{\varepsilon}(\boldsymbol{v})$:

$$\boldsymbol{\varepsilon}(\boldsymbol{v}) = \frac{1}{2}\big(\nabla \boldsymbol{v} + (\nabla \boldsymbol{v})^T\big). \tag{1}$$

In the sequel we shall use Poincaré's, Korn's, and some trace inequalities. Poincaré's inequality in $H^1_{0,\Gamma}(\Omega)$ reads: There exists a constant $\mathscr{P}_\Gamma$ depending only on Ω and Γ such that

$$\forall v \in H^1_{0,\Gamma}(\Omega), \quad \|v\|_{L^2(\Omega)} \le \mathscr{P}_\Gamma |v|_{H^1(\Omega)}. \tag{2}$$

Next, recall Korn's first inequality in $H^1_{0,\Gamma}(\Omega)^d$: There exists a constant C_κ depending only on Ω and Γ such that

$$\forall \boldsymbol{v} \in H^1_{0,\Gamma}(\Omega)^d, \quad |\boldsymbol{v}|_{H^1(\Omega)} \le C_\kappa \|\boldsymbol{\varepsilon}(\boldsymbol{v})\|_{L^2(\Omega)}. \tag{3}$$

We shall use the following trace inequality in $H^1(\Omega)$: There exists a constant C_τ depending only on Ω and Γ such that

$$\forall \varepsilon > 0,\ \forall v \in H^1(\Omega), \quad \|v\|_{L^2(\Gamma)} \le \varepsilon \|\nabla v\|_{L^2(\Omega)} + \left(\frac{C_\tau}{\varepsilon} + \varepsilon\right) \|v\|_{L^2(\Omega)}. \tag{4}$$

This inequality follows, for instance, from the interpolation inequality (see Brenner and Scott [5])

$$\forall v \in H^1(\Omega), \quad \|v\|_{L^2(\Gamma)} \le C\|v\|_{L^2(\Omega)}^{\frac{1}{2}} \|v\|_{H^1(\Omega)}^{\frac{1}{2}},$$

and Young's inequality. Besides (4), by combining (2) and (3), we immediately derive the alternate trace inequality, with a constant C_D depending only on Ω and Γ:

$$\forall \boldsymbol{v} \in H^1_{0,\Gamma}(\Omega)^d, \quad \|\boldsymbol{v}\|_{L^2(\Gamma)} \le C_D \|\boldsymbol{\varepsilon}(\boldsymbol{v})\|_{L^2(\Omega)}.$$

As far as the divergence operator is concerned, we shall use the space

$$H(\mathrm{div}; \Omega) = \{\boldsymbol{v} \in L^2(\Omega)^d \mid \nabla \cdot \boldsymbol{v} \in L^2(\Omega)\},$$

equipped with the norm

$$\|\boldsymbol{v}\|_{H(\mathrm{div};\Omega)} = \left(\|\boldsymbol{v}\|^2_{L^2(\Omega)} + \|\nabla \cdot \boldsymbol{v}\|^2_{L^2(\Omega)}\right)^{\frac{1}{2}}.$$

As usual, for handling time-dependent problems, it is convenient to consider functions defined on a time interval $]a, b[$ with values in a functional space, say X (cf. [21]). More precisely, let $\|\cdot\|_X$ denote the norm of X; then for any number r, $1 \le r \le \infty$, we define

$$L^r(a, b; X) = \left\{ f \text{ measurable in }]a, b[\mid \int_a^b \|f(t)\|_X^r \, dt < \infty \right\},$$

equipped with the norm

$$\|f\|_{L^r(a,b;X)} = \left(\int_a^b \|f(t)\|_X^r dt\right)^{\frac{1}{r}},$$

with the usual modification if $r = \infty$. This space is a Banach space if X is a Banach space, and for $r = 2$, it is a Hilbert space if X is a Hilbert space. To simplify, we sometimes denote derivatives with respect to time with a prime and we define for any r, $1 \le r \le \infty$,

$$W^{1,r}(a, b; X) = \{f \in L^r(a, b; X) \mid f' \in L^r(a, b; X)\}.$$

For any $r \ge 1$, as the functions of $W^{1,r}(a, b; X)$ are continuous with respect to time, we define

$$W^{1,r}_0(a, b; X) = \{f \in W^{1,r}(a, b; X) \mid f(a) = f(b) = 0\},$$

and we denote by $W^{-1,r}(a, b; X)$ the dual space of $W^{1,r'}_0(a, b; X)$, where r' is the dual exponent of r, $\frac{1}{r'} + \frac{1}{r} = 1$.

2 Domain and Model Formulations

Let the reservoir Ω be a bounded domain of $\mathbb{R}^d$, $d = 2$ or 3, with a piecewise smooth Lipschitz boundary $\partial\Omega$ and exterior normal $\boldsymbol{n}$. Let the fracture $\mathscr{C} \Subset \Omega$ be a simple closed piecewise smooth curve with endpoints $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$ when $d = 2$ or a simple closed piecewise smooth surface with piecewise smooth Lipschitz boundary $\partial\mathscr{C}$ when $d = 3$, see Fig. 1. The reservoir contains both the matrix and the fractures; thus the reservoir matrix is $\Omega \setminus \mathscr{C}$.

Sections 2.1, 2.2, and 2.3 present a very succinct mechanical derivation of the model. Then in Sect. 3, the problem is set in a mixed variational formulation.

2.1 Equations in $\Omega \setminus \mathscr{C}$

The displacement of the solid is modeled in $\Omega \setminus \mathscr{C}$ by the quasi-static Biot equations for a linear elastic, homogeneous, isotropic, porous solid saturated with a slightly compressible viscous fluid (see [3]). The constitutive equation for the Cauchy stress tensor $\boldsymbol{\sigma}^{\text{por}}$ is

$$\boldsymbol{\sigma}^{\text{por}}(\boldsymbol{u}, p) = \boldsymbol{\sigma}(\boldsymbol{u}) - \alpha\, p \boldsymbol{I}, \tag{5}$$

where $\boldsymbol{I}$ is the identity tensor, $\boldsymbol{u}$ is the solid's displacement, p is the fluid pressure, $\boldsymbol{\sigma}$ is the effective linear elastic stress tensor:

$$\boldsymbol{\sigma}(\boldsymbol{u}) = \lambda(\nabla \cdot \boldsymbol{u})\boldsymbol{I} + 2\, G\boldsymbol{\varepsilon}(\boldsymbol{u}),$$

see (1) for the definition of $\boldsymbol{\varepsilon}(\boldsymbol{u})$. Here $\lambda > 0$ and $G > 0$ are the Lamé constants and $\alpha > 0$ is the dimensionless Biot coefficient. Then the balance of linear momentum in the solid reads

$$-\operatorname{div} \boldsymbol{\sigma}^{\text{por}}(\boldsymbol{u}, p) = \boldsymbol{f} \quad \text{in } \Omega \setminus \mathscr{C}, \tag{6}$$

where $\boldsymbol{f}$ is a body force, i.e., a gravity loading term. For the fluid, we use a linearized slightly compressible single-phase model. Let p_r be a reference pressure, $\rho_f > 0$ the fluid phase density, $\rho_{f,r} > 0$ a constant reference density relative to p_r, and c_f the fluid compressibility. We consider the simplified case when ρ_f is a linear function of pressure:

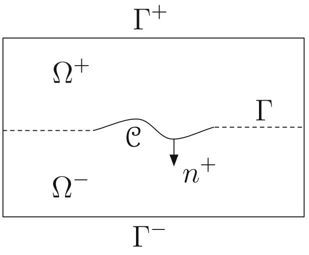


Fig. 1 Diagram of domain, fracture, and boundaries

$$\rho_f = \rho_{f,r}\big(1 + c_f(p - p_r)\big). \tag{7}$$

Next, let φ^* denote the fluid content (or reservoir fluid fraction) of the medium defined by

$$\varphi^* = \varphi(1 + \nabla \cdot \boldsymbol{u}),$$

where φ is the porosity of the medium. For a poroelastic material with small deformation, φ^* can be approximated by

$$\varphi^* = \varphi_0 + \alpha \nabla \cdot \boldsymbol{u} + \frac{1}{M} p, \tag{8}$$

where φ_0 is the initial porosity and M a Biot constant. The velocity of the fluid $\boldsymbol{v}^D$ in $\Omega \setminus \mathscr{C}$ obeys Darcy's Law:

$$\boldsymbol{v}^D = -\frac{1}{\mu_f} \boldsymbol{K}\big(\nabla p - \rho_f g \nabla \eta\big), \tag{9}$$

where $\boldsymbol{K}$ is the absolute permeability tensor, assumed to be symmetric, bounded, uniformly positive definite in space and constant in time, $\mu_f > 0$ is the constant fluid viscosity, g is the gravitation constant, and η is a signed distance in the vertical direction, variable in space, but constant in time. The fluid mass balance in $\Omega \setminus \mathscr{C}$ reads

$$\frac{\partial}{\partial t}\big(\rho_f \varphi^*\big) + \nabla \cdot (\rho_f \boldsymbol{v}^D) = q, \tag{10}$$

where q is a mass source or sink term taking into account injection into or extraction from the reservoir. Let us neglect small quantities by means of the following approximations:

$$\begin{gathered}
\frac{1}{M}(1 + c_f(p - p_r)) \approx \frac{1}{M}, \\
c_f \left(\varphi_0 + \alpha \nabla \cdot \boldsymbol{u} + \frac{1}{M} p \right) \approx c_f \varphi_0, \\
\rho_{f,r}(1 + c_f(p - p_r))\alpha \approx \rho_{f,r}\alpha, \\
\rho_{f,r}(1 + c_f(p - p_r))\boldsymbol{v}^D \approx \rho_{f,r}\boldsymbol{v}^D, \\
\rho_{f,r}(1 + c_f(p - p_r)) g \nabla \eta \approx \rho_{f,r} g \nabla \eta.
\end{gathered}$$

Then by substituting (7), (8), and (9) into (10), and setting $\tilde{q} = \frac{q}{\rho_{f,r}}$, we obtain

$$\frac{\partial}{\partial t}\left(\left(\frac{1}{M} + c_f \varphi_0 \right) p + \alpha \nabla \cdot \boldsymbol{u} \right) - \nabla \cdot \left(\frac{1}{\mu_f} \boldsymbol{K}\big(\nabla p - \rho_{f,r} g \nabla \eta\big) \right) = \tilde{q}. \tag{11}$$

Thus the poro-elastic system we are considering for modeling the displacement $\boldsymbol{u}$ and pressure p in $\Omega \setminus \mathscr{C}$ is governed by (5), (6) and (11).

2.2 Equation in $\mathscr{C}$

Recall that in our model, since the fracture is assumed to be very narrow, it is approximated geometrically by a single curve when $d = 2$ or a single surface when $d = 3$. For the moment, we assume that the fluid pressure p belongs at least to $H^1(\Omega)$; therefore it has a well defined trace on $\mathscr{C}$, say p_c. We denote by $\overline{\nabla}$ the surface gradient operator on $\mathscr{C}$. It is the tangential trace of the gradient, that is well defined for functions in $H^1(\Omega)$, cf., for example, [16]. The physical width of the fracture is represented by a non-negative function w defined on $\mathscr{C}$; it is the jump of the displacement $\boldsymbol{u}$ in the normal direction. Since the medium is elastic and the energy is finite, w must be bounded and must vanish on the boundary of the fracture. Then the volumetric flow rate $\mathscr{Q}$ on $\mathscr{C}$ satisfies

$$\mathscr{Q} = -\frac{w^3}{12\mu_f}\big(\overline{\nabla}\, p_c - \rho_f g \overline{\nabla}\, \eta\big),$$

and the conservation of mass in the fracture reads

$$\frac{\partial}{\partial t}(\rho_f w) = -\overline{\nabla} \cdot (\rho_f \mathscr{Q}) + q_W - q_L,$$

where q_W is a known injection term into the fracture and q_L is an unknown leakage term from the fracture into the reservoir matrix that guarantees the conservation of mass in the system. This cubic model is fairly standard (see, e.g., [28]), and is a form of Reynolds' lubrication equation obtained by averaging Stokes model in the vertical direction of a given fracture. This model, designed for narrow fractures, approximates geometrically the fractures by regions in $d - 1$ dimensions, i.e., with no width. It assumes sufficient permeability in the vertical direction so that, as a result of averaging, the pressure remains continuous across the fracture. This explains why the pressure p is assumed to be globally in $H^1(\Omega)$. Then neglecting again small quantities and setting $\tilde{q}_W = \frac{q_W}{\rho_{f,r}}$, $\tilde{q}_L = \frac{q_L}{\rho_{f,r}}$, we derive the lubrication equation in $\mathscr{C}$:

$$\frac{\partial}{\partial t} w - \overline{\nabla} \cdot \left(\frac{w^3}{12\mu_f}(\overline{\nabla}\, p_c - \rho_{f,r} g \overline{\nabla}\, \eta)\right) = \tilde{q}_W - \tilde{q}_L. \tag{12}$$

In order to specify the relation between the displacement $\boldsymbol{u}$ of the medium and the width w of the fracture, let us distinguish the two sides (or faces) of $\mathscr{C}$ by the superscripts $+$ and $-$; a specific choice must be selected but is arbitrary. To simplify the discussion, we use a superscript $\star$ to denote either $+$ or $-$. Let $\Omega^\star$ denote the part of Ω adjacent to $\mathscr{C}^\star$ and let $\boldsymbol{n}^\star$ denote the unit normal vector to $\mathscr{C}$ exterior to $\Omega^\star$, $\star = +, -$. As the fracture is represented by two geometrically coincident surfaces, the normal vectors are related by $\boldsymbol{n}^- = -\boldsymbol{n}^+$. For any function f defined in $\Omega \setminus \mathscr{C}$ that has a trace, let $f^\star$ denote the trace of f on $\mathscr{C}^\star$, $\star = +, -$. Then we define the jump of f on $\mathscr{C}$ in the direction of $\boldsymbol{n}^+$ by

$$[f]_{\mathscr{C}} = f^+ - f^-.$$

The width w is the jump of $\boldsymbol{u} \cdot \boldsymbol{n}^-$ on $\mathscr{C}$:

$$w = -[\boldsymbol{u}]_{\mathscr{C}} \cdot \boldsymbol{n}^+. \tag{13}$$

Therefore the only unknown in (12) is the leakage term $\tilde{q}_L$.

Summarizing, the equations in $\Omega \setminus \mathscr{C}$ are (6) and (11), and the equation in $\mathscr{C}$ is (12); the corresponding unknowns are $\boldsymbol{u}, p$ and $\tilde{q}_L$. These equations are complemented in the next section by interface, boundary and initial conditions.

2.3 Interface, Boundary, and Initial Conditions

Let $\boldsymbol{\tau}_j^\star$, $1 \le j \le d-1$, be a set of orthonormal tangent vectors on $\mathscr{C}^\star$, $\star = +, -$. The balance of the normal traction vector and the conservation of mass yield the interface conditions on each side (or face) of $\mathscr{C}$:

$$(\boldsymbol{\sigma}^{\mathrm{por}}(\boldsymbol{u}, p))^\star \boldsymbol{n}^\star = -p_c \boldsymbol{n}^\star, \quad \star = +, -. \tag{14}$$

Then the continuity of p through $\mathscr{C}$ yields

$$[\boldsymbol{\sigma}^{\mathrm{por}}(\boldsymbol{u}, p)]_{\mathscr{C}} \boldsymbol{n}^\star = \boldsymbol{0}.$$

Formula (14) also implies

$$\boldsymbol{\sigma}^{\mathrm{por}}(\boldsymbol{u}, p)\boldsymbol{n}^\star \cdot \boldsymbol{n}^\star = -p_c, \quad \boldsymbol{\sigma}^{\mathrm{por}}(\boldsymbol{u}, p)\boldsymbol{n}^\star \cdot \boldsymbol{\tau}^\star = 0.$$

With the above approximations, the conservation of mass at the interface is expressed as

$$\frac{1}{\mu_f}[\boldsymbol{K}(\nabla p - \rho_{f,r} g \nabla \eta)]_{\mathscr{C}} \cdot \boldsymbol{n}^+ = \tilde{q}_L. \tag{15}$$

General conditions on the exterior boundary $\partial\Omega$ of Ω can be prescribed for the poro-elastic system, but to simplify our analysis, we assume that the displacement $\boldsymbol{u}$ vanishes as well as the flux $\boldsymbol{K}(\nabla p - \rho_{f,r} g \nabla \eta) \cdot \boldsymbol{n}$. According to the above hypotheses on the energy and medium, we assume that w is bounded in $\mathscr{C}$ and vanishes on $\partial\mathscr{C}$. Finally, considering that the time derivative in (11) acts on $(\frac{1}{M} + c_f \varphi_0)p + \alpha \nabla \cdot \boldsymbol{u}$, we prescribe at initial time (see [26]):

$$\left(\left(\frac{1}{M} + c_f \varphi_0\right) p + \alpha \nabla \cdot \boldsymbol{u}\right)(0) = \left(\frac{1}{M} + c_f \varphi_0\right) p_0 + \alpha \nabla \cdot \boldsymbol{u}_0, \tag{16}$$

where p_0 is measured and all other initial data are deduced from it: $\boldsymbol{u}_0$ is the displacement associated with p_0 by (6) at initial time, the trace of p_0 on $\mathscr{C}$ is denoted by p_c^0, and the initial value of w is deduced from the normal jump of $\boldsymbol{u}_0$ on $\mathscr{C}$. Strictly speaking, the pressure does not have sufficient regularity in time to define its initial value; therefore p_0 in (16) cannot be related to $p(0)$. However, for practical purposes, we shall assume that p is sufficiently smooth, so that $p(0)$ is indeed p_0.

Therefore the complete problem statement is

Problem 1 Find $\boldsymbol{u}, p$, and $\tilde{q}_L$ satisfying (5), (6), (11) in $\Omega \setminus \mathscr{C}$ and (12) in $\mathscr{C}$, for all time $t \in]0, T[$, with the interface conditions (14) and (15) on $\mathscr{C}$ and initial condition (16):

$$-\operatorname{div} \boldsymbol{\sigma}^{\text{por}}(\boldsymbol{u}, p) = \boldsymbol{f}, \quad \text{in } \Omega \setminus \mathscr{C}, \tag{P1.1}$$

$$\boldsymbol{\sigma}^{\text{por}}(\boldsymbol{u}, p) = \boldsymbol{\sigma}(\boldsymbol{u}) - \alpha p \boldsymbol{I}, \quad \text{in } \Omega \setminus \mathscr{C}, \tag{P1.2}$$

$$\frac{\partial}{\partial t}\left(\left(\frac{1}{M} + c_f \varphi_0\right) p + \alpha \nabla \cdot \boldsymbol{u}\right) - \nabla \cdot \left(\frac{1}{\mu_f} \boldsymbol{K} \nabla (p - \rho_{f,r} g \eta)\right) = \tilde{q}, \quad \text{in } \Omega \setminus \mathscr{C}, \tag{P1.3}$$

$$\frac{\partial}{\partial t} w - \overline{\nabla} \cdot \left(\frac{w^3}{12\mu_f} \overline{\nabla}(p - \rho_{f,r} g \eta)\right) = \tilde{q}_W - \tilde{q}_L, \quad \text{in } \mathscr{C}, \tag{P1.4}$$

$$(\boldsymbol{\sigma}^{\text{por}}(\boldsymbol{u}, p))^\star \boldsymbol{n}^\star = -p|_{\mathscr{C}} \boldsymbol{n}^\star, \quad \star = +, - \text{ on } \mathscr{C}, \tag{P1.5}$$

$$\frac{1}{\mu_f} [\boldsymbol{K} \nabla (p - \rho_{f,r} g \eta)]_{\mathscr{C}} \cdot \boldsymbol{n}^+ = \tilde{q}_L, \quad \text{on } \mathscr{C}, \tag{P1.6}$$

where

$$w = -[\boldsymbol{u}]_{\mathscr{C}} \cdot \boldsymbol{n}^+, \tag{P1.7}$$

with the boundary conditions

$$\boldsymbol{u} = \boldsymbol{0}, \quad \boldsymbol{K} \nabla (p - \rho_{f,r} g \eta) \cdot \boldsymbol{n} = 0 \quad \text{on } \partial \Omega, \tag{P1.8}$$

and the initial condition at time $t = 0$,

$$\left(\left(\frac{1}{M} + c_f \varphi_0\right) p + \alpha \nabla \cdot \boldsymbol{u}\right)(0) = \left(\frac{1}{M} + c_f \varphi_0\right) p_0 + \alpha \nabla \cdot \boldsymbol{u}_0.$$

3 Variational Formulation

Here, we use a mixed formulation for the flow because it leads to locally conservative schemes.

3.1 Spaces

We shall see below that the width function w acts as a weight on the flow velocity in the fracture. For the practical applications we have in mind, w has the following properties when $d = 3$; the statement easily extends to $d = 2$:

Hypothesis 1 The non-negative function w is H^1 in time and is smooth in space away from the fracture's front, i.e., the boundary $\partial\mathscr{C}$. It vanishes on $\partial\mathscr{C}$ and in a neighborhood of any point of $\partial\mathscr{C}$, w is asymptotically of the form:

$$w(x, y) \simeq x^{\frac{1}{2}+\varepsilon} f(y), \quad \text{with small } \varepsilon > 0, \tag{17}$$

where y is locally parallel to the fracture's front, x is the distance to $\partial\mathscr{C}$, and f is smooth.

The Assumption (17) is motivated by the stress intensity factor modelling the stress near the crack-tip and is widely used in fracture mechanics, see, for instance, [10]. The stress near the crack-tip varies as the square root's inverse of the distance from the crack-tip and so the assumption on the width assumes that the displacement is nearly the square-root of the distance from the crack-tip.

The spaces for our unknowns are described below. To simplify the notation, the spaces related to $\mathscr{C}$ are written $L^2(\mathscr{C})$, $H^{\frac{1}{2}}(\mathscr{C})$, etc., although they are defined in the interior of $\mathscr{C}$. Regarding $\boldsymbol{x}$, it is convenient (but not fundamental) to introduce an auxiliary partition of Ω into two non-overlapping subdomains Ω^+ and Ω^- with Lipschitz interface Γ containing $\mathscr{C}$, $\Omega^\star$ being adjacent to $\mathscr{C}^\star$, $\star = +, -$. The precise shape of Γ is not important as long as Ω^+ and Ω^- are both Lipschitz. Let $\Gamma^\star = \partial\Omega^\star \setminus \Gamma$. For any function f defined in Ω, we extend the star notation to $\Omega^\star$ and set $f^\star = f_{|\Omega^\star}$, $\star = +, -$. Let $W = H^1(\Omega^+ \cup \Omega^-)$ with norm

$$\|v\|_W = \left(\|v^+\|^2_{H^1(\Omega^+)} + \|v^-\|^2_{H^1(\Omega^-)}\right)^{\frac{1}{2}}.$$

The space for the displacement is $L^\infty(0, T; \mathbb{V})$, where $\mathbb{V}$ a closed subspace of $H^1(\Omega \setminus \mathscr{C})^d$:

$$\mathbb{V} = \{\boldsymbol{v} \in W^d \mid [\boldsymbol{v}]_{\Gamma\setminus\mathscr{C}} = \boldsymbol{0}, \boldsymbol{v}^\star_{|\Gamma^\star} = \boldsymbol{0},\ \star = +, -\},$$

with the norm of W^d:

$$\|\boldsymbol{v}\|_{\mathbb{V}} = \left(\sum_{i=1}^{d} \|v_i\|_W^2\right)^{\frac{1}{2}}.$$

As stated previously, the pressure p is essentially in $H^1(\Omega)$ (see more precisely (21)), but to set the problem in mixed form, we reduce the regularity of p and take p in $L^\infty(0,T;L^2(\Omega))$. As the functions of $L^2(\Omega)$ have no trace, we introduce an auxiliary variable p_c in the space $L^2(0,T;H^{\frac{1}{2}}(\mathscr{C}))$ that is treated as an unknown variable and is intended to represent the pressure's trace on $\mathscr{C}$.

We associate with the pressure in $\Omega \setminus \mathscr{C}$ an auxiliary velocity $\boldsymbol{z}$ defined by

$$\boldsymbol{z} = -\boldsymbol{K}\nabla(p - \rho_{f,r} g\eta), \tag{18}$$

and we associate with the pressure in $\mathscr{C}$, a surface velocity $\boldsymbol{\zeta}$ defined by

$$\boldsymbol{\zeta} = -w^{\frac{3}{2}}\overline{\nabla}(p_c - \rho_{f,r} g\eta). \tag{19}$$

The space for the reservoir matrix velocity is $L^2(0,T;\boldsymbol{Z})$, where

$$\boldsymbol{Z} = \{\boldsymbol{q} \in H(\mathrm{div};\Omega^+ \cup \Omega^-) \mid [\boldsymbol{q}]\cdot\boldsymbol{n}^+ = 0 \text{ on } \Gamma\setminus\mathscr{C},\ \boldsymbol{q}\cdot\boldsymbol{n} = 0 \text{ on } \partial\Omega\}, \tag{20}$$

normed by

$$\|\boldsymbol{q}\|_Z = \left(\|\boldsymbol{q}\|^2_{H(\mathrm{div};\Omega^+)} + \|\boldsymbol{q}\|^2_{H(\mathrm{div};\Omega^-)}\right)^{\frac{1}{2}}.$$

Strictly speaking, in (20) we should write $[\boldsymbol{q}\cdot\boldsymbol{n}^+] = 0$ on $\Gamma\setminus\mathscr{C}$. However, since $\boldsymbol{n}^+$ does not jump, we abuse the notation and write it as $[\boldsymbol{q}]\cdot\boldsymbol{n}^+ = 0$. Let $\boldsymbol{n}_{\Omega^\star}$ be the unit normal to $\partial\Omega^\star$, exterior to $\Omega^\star$. The trace properties of $H(\mathrm{div};\Omega^\star)$ imply that $\boldsymbol{q}\cdot\boldsymbol{n}_{\Omega^\star}$ belongs to $H^{-\frac{1}{2}}(\partial\Omega^\star)$ which is defined globally (see, for example, [16]). However, as $\boldsymbol{q}\cdot\boldsymbol{n}$ vanishes on $\partial\Omega$, following the work of Galvis and Sarkis in [12], we can prove first that the jump $[\boldsymbol{q}]\cdot\boldsymbol{n}^+$ belongs to $H^{-\frac{1}{2}}(\Gamma)$, and since it vanishes on $\Gamma\setminus\mathscr{C}$ then it is well-defined in $H^{-\frac{1}{2}}(\mathscr{C})$ with continuous dependence on $\|\boldsymbol{q}\|_{\boldsymbol{Z}}$. Therefore $\boldsymbol{Z}$ is a closed subspace of $H(\mathrm{div};\Omega^+\cup\Omega^-)$ and of $H(\mathrm{div};\Omega\setminus\mathscr{C})$.

The space for the velocity in the fracture is $L^2(0,T;\boldsymbol{Z}_{\mathscr{C}})$, where

$$\boldsymbol{Z}_{\mathscr{C}} = \{\boldsymbol{q}_c \in L^2(\mathscr{C})^{d-1} \mid \overline{\nabla}\cdot(w^{\frac{3}{2}}\boldsymbol{q}_c) \in H^{-\frac{1}{2}}(\mathscr{C})\},$$

equipped with the graph norm

$$\|\boldsymbol{q}_c\|_{Z_{\mathscr{C}}} = \left(\|\boldsymbol{q}_c\|^2_{L^2(\mathscr{C})} + \|\overline{\nabla}\cdot(w^{\frac{3}{2}}\boldsymbol{q}_c)\|^2_{H^{-\frac{1}{2}}(\mathscr{C})}\right)^{\frac{1}{2}}.$$

This space is closely related to the pressure's space in the fracture introduced in [18]):

$$H^1_w(\mathscr{C}) = \{z \in H^{\frac{1}{2}}(\mathscr{C}) \mid w^{\frac{3}{2}}\overline{\nabla} z \in L^2(\mathscr{C})^{d-1}\},$$

equipped with the norm

$$\|z\|_{H^1_w(\mathscr{C})} = \left(\|z\|^2_{H^{\frac{1}{2}}(\mathscr{C})} + \|w^{\frac{3}{2}}\overline{\nabla}\, z\|^2_{L^2(\mathscr{C})}\right)^{\frac{1}{2}},$$

so that p belongs to

$$Q = \{q \in H^1(\Omega) \mid q|_{\mathscr{C}} \in H^1_w(\mathscr{C})\}. \tag{21}$$

Remark 1 Strictly speaking, the space $\boldsymbol{Z}_{\mathscr{C}}$ does not correspond to a standard mixed space for the velocity in the fracture since the divergence of its functions is in $H^{-\frac{1}{2}}(\mathscr{C})$ instead of $L^2(\mathscr{C})$. We cannot prescribe this last regularity because the leakage term $\tilde{q}_L$ is not a data: it is the jump in the normal fluxes, which in general cannot be expected to be in $L^2(\mathscr{C})$. Thus the pressure p_c in the fracture must be taken in $H^{\frac{1}{2}}(\mathscr{C})$. This extra regularity will be relaxed in the numerical applications because the discrete jump in the normal fluxes is always in $L^2(\mathscr{C})$. □

To simplify, we denote the scalar products in space by parentheses; if the domain of integration is not indicated, then it is understood that the integrals are taken over $\Omega^+ \cup \Omega^-$. All dualities are denoted by chevrons, eg. the notation $\langle\cdot,\cdot\rangle_{\mathscr{C}}$ stands for a duality pairing on $\mathscr{C}$.

Recall the properties of $H^1_w(\mathscr{C})$ established in [18]:

Theorem 1 *Under Hypothesis 1, $H^1_w(\mathscr{C})$ is a separable Hilbert space, $W^{1,\infty}(\mathscr{C})$ is dense in $H^1_w(\mathscr{C})$, and the following Green formula holds for all θ in $H^1_w(\mathscr{C})$ such that $\overline{\nabla}\cdot(w^3\overline{\nabla}\,\theta)$ belongs to $H^{-\frac{1}{2}}(\mathscr{C})$:*

$$\forall\lambda \in H^1_w(\mathscr{C}), \quad -\langle\overline{\nabla}\cdot(w^3\overline{\nabla}\,\theta), \lambda\rangle_{\mathscr{C}} = (w^{\frac{3}{2}}\overline{\nabla}\,\theta, w^{\frac{3}{2}}\overline{\nabla}\,\lambda)_{\mathscr{C}}.$$

With this result, we can prove that $\boldsymbol{Z}_{\mathscr{C}}$ has the following properties.

Proposition 1 *Let w belong to $L^\infty(\mathscr{C})$. Then $\boldsymbol{Z}_{\mathscr{C}}$ is a separable Hilbert space. Moreover, if w satisfies Hypothesis 1, the following Green formula holds for all θ_c in $H^1_w(\mathscr{C})$:*

$$\forall \boldsymbol{q}_c \in \boldsymbol{Z}_{\mathscr{C}}, \quad -(w^{\frac{3}{2}}\overline{\nabla}\,\theta_c, \boldsymbol{q}_c)_{\mathscr{C}} = \langle\theta_c, \overline{\nabla}\cdot(w^{\frac{3}{2}}\boldsymbol{q}_c)\rangle_{\mathscr{C}}. \tag{22}$$

Proof To show that $\boldsymbol{Z}_{\mathscr{C}}$ is a Hilbert space, it suffices to prove that it is complete. Let $(\boldsymbol{z}_n)_{n\geq 1}$ be a Cauchy sequence of functions in $\boldsymbol{Z}_{\mathscr{C}}$. Then there exists a function $\boldsymbol{z}$ in $L^2(\mathscr{C})^{d-1}$ and a function $v \in H^{-\frac{1}{2}}(\mathscr{C})$ such that

$$\lim_{n\to\infty} \boldsymbol{z}_n = \boldsymbol{z} \quad \text{in } L^2(\mathscr{C})^{d-1}, \qquad \lim_{n\to\infty} \overline{\nabla}\cdot(w^{\frac{3}{2}}\boldsymbol{z}_n) = v \quad \text{in } H^{-\frac{1}{2}}(\mathscr{C}).$$

In order to prove that $v = \overline{\nabla}\cdot(w^{\frac{3}{2}}\boldsymbol{z})$, we take any φ in $H^1_0(\mathscr{C})$. Then

$$\langle\overline{\nabla}\cdot(w^{\frac{3}{2}}\boldsymbol{z}_n), \varphi\rangle_{\mathscr{C}} = -\left(w^{\frac{3}{2}}\boldsymbol{z}_n, \overline{\nabla}\,\varphi\right)_{\mathscr{C}} = -\left(\boldsymbol{z}_n, (w^{\frac{3}{2}}\overline{\nabla}\,\varphi)\right)_{\mathscr{C}}.$$

Passing to the limit in the first and last term, we obtain

$$\langle v, \varphi\rangle_{\mathscr{C}} = -\big(\boldsymbol{z}, (w^{\frac{3}{2}}\overline{\nabla}\varphi)\big)_{\mathscr{C}} = \langle \overline{\nabla}\cdot(w^{\frac{3}{2}}\boldsymbol{z}), \varphi\rangle_{\mathscr{C}},$$

whence $v = \overline{\nabla}\cdot(w^{\frac{3}{2}}\boldsymbol{z})$.

Proving the separability of $\boldsymbol{Z}_{\mathscr{C}}$ is fairly classical. Let $E = L^2(\mathscr{C})^{d-1} \times H^{-\frac{1}{2}}(\mathscr{C})$ normed by

$$\|v\|_E = \big(\|\boldsymbol{v}_1\|^2_{L^2(\mathscr{C})} + \|v_2\|_{H^{-\frac{1}{2}}(\mathscr{C})}\big)^{\frac{1}{2}},$$

where $v = (\boldsymbol{v}_1, v_2)$; exceptionally the parentheses denote pairs, and let Φ be the mapping: $\boldsymbol{Z}_{\mathscr{C}} \mapsto E$ defined by

$$\forall \boldsymbol{z} \in \boldsymbol{Z}_{\mathscr{C}}, \quad \Phi(\boldsymbol{z}) = \big(\boldsymbol{z}, \overline{\nabla}\cdot(w^{\frac{3}{2}}\boldsymbol{z})\big).$$

Then Φ is an isometry and the argument of the completeness proof above yields that the range of Φ, $\mathscr{R}(\Phi)$, is closed in E. Since E is separable, so is $\mathscr{R}(\Phi)$. Then the separability of $\boldsymbol{Z}_{\mathscr{C}}$ follows from the fact that it is isometrically isomorphic to $\mathscr{R}(\Phi)$.

To prove Green's formula (22), we use the density of $W^{1,\infty}(\mathscr{C})$ into $H^1_w(\mathscr{C})$ stated in Theorem 1: Let $(p_n)_{n\geq 1}$ be a sequence of functions of $W^{1,\infty}(\mathscr{C})$ that tend to θ_c in $H^1_w(\mathscr{C})$. Then

$$\forall \boldsymbol{q}_c \in \boldsymbol{Z}_{\mathscr{C}}, \quad -\big(w^{\frac{3}{2}}\overline{\nabla}p_n, \boldsymbol{q}_c\big)_{\mathscr{C}} = -\big(\overline{\nabla}(w^{\frac{3}{2}}p_n), \boldsymbol{q}_c\big)_{\mathscr{C}} + \big(p_n\overline{\nabla}(w^{\frac{3}{2}}), \boldsymbol{q}_c\big)_{\mathscr{C}}.$$

Since $w^{\frac{3}{2}}p_n$ belongs to $H^1_0(\mathscr{C})$, we can write

$$-\big(\overline{\nabla}(w^{\frac{3}{2}}p_n), \boldsymbol{q}_c\big)_{\mathscr{C}} = \langle w^{\frac{3}{2}}p_n, \overline{\nabla}\cdot\boldsymbol{q}_c\rangle_{\mathscr{C}} = \langle p_n, w^{\frac{3}{2}}\overline{\nabla}\cdot\boldsymbol{q}_c\rangle_{\mathscr{C}}.$$

Hence

$$-\big(w^{\frac{3}{2}}\overline{\nabla}p_n, \boldsymbol{q}_c\big)_{\mathscr{C}} = \langle p_n, w^{\frac{3}{2}}\overline{\nabla}\cdot\boldsymbol{q}_c + \overline{\nabla}(w^{\frac{3}{2}})\cdot\boldsymbol{q}_c\rangle_{\mathscr{C}} = \langle p_n, \overline{\nabla}\cdot(w^{\frac{3}{2}}\boldsymbol{q}_c)\rangle_{\mathscr{C}},$$

and (22) follows by letting n tend to infinity in the first and last term above. □

Finally, in view of the jump and boundary conditions of Problem 1, we see that $\boldsymbol{z}$ must satisfy the essential jump condition:

$$\frac{1}{\mu_f}[\boldsymbol{z}]_{\mathscr{C}}\cdot\boldsymbol{n}^+ = -\tilde{q}_L, \quad \text{on } \mathscr{C}. \tag{23}$$

Consequently, the leakage term is in $L^2(0, T; H^{-\frac{1}{2}}(\mathscr{C}))$ and can be eliminated by substituting (23) into the lubrication equation (12).

3.2 Mixed Formulation

The mixed variational formulation of Problem 1 reads:

Problem 2 For given $\boldsymbol{f} \in L^2\left((\Omega \setminus \mathscr{C}) \times]0, T[\right)^d$, $\tilde{q} \in L^2(\Omega \times]0, T[)$, and $\tilde{q}_W \in L^2(0, T; H^{-\frac{1}{2}}(\mathscr{C}))$, find $\boldsymbol{u} \in L^\infty(0, T; \mathbb{V})$, $p \in L^\infty(0, T; L^2(\Omega))$, $p_c \in L^2(0, T; H^{-\frac{1}{2}}(\mathscr{C}))$, $\boldsymbol{z} \in L^2(0, T; \boldsymbol{Z})$, and $\boldsymbol{\zeta} \in L^2(0, T; \boldsymbol{Z}_{\mathscr{C}})$ such that

$$\forall \boldsymbol{v} \in \mathbb{V}, \quad 2G\big(\boldsymbol{\varepsilon}(\boldsymbol{u}), \boldsymbol{\varepsilon}(\boldsymbol{v})\big) + \lambda\big(\nabla \cdot \boldsymbol{u}, \nabla \cdot \boldsymbol{v}\big) - \alpha\big(p, \nabla \cdot \boldsymbol{v}\big) + \big(p_c, [\boldsymbol{v}]_{\mathscr{C}} \cdot \boldsymbol{n}^+\big)_{\mathscr{C}} = \big(\boldsymbol{f}, \boldsymbol{v}\big), \tag{P2.1}$$

$$\forall \theta \in L^2(\Omega), \quad \left(\frac{\partial}{\partial t}\left(\left(\frac{1}{M} + c_f \varphi_0\right)p + \alpha \nabla \cdot \boldsymbol{u}\right), \theta\right) + \frac{1}{\mu_f}\big(\nabla \cdot \boldsymbol{z}, \theta\big) = \big(\tilde{q}, \theta\big), \tag{P2.2}$$

$$\begin{aligned}\forall \theta_c \in H^{\frac{1}{2}}(\mathscr{C}), \quad & -\left\langle \frac{\partial}{\partial t}[\boldsymbol{u}]_{\mathscr{C}} \cdot \boldsymbol{n}^+, \theta_c \right\rangle_{\mathscr{C}} + \frac{1}{12\mu_f}\langle \overline{\nabla} \cdot (w^{\frac{3}{2}} \boldsymbol{\zeta}), \theta_c \rangle_{\mathscr{C}} \\ & - \frac{1}{\mu_f}\langle [\boldsymbol{z}]_{\mathscr{C}} \cdot \boldsymbol{n}^+, \theta_c \rangle_{\mathscr{C}} = \langle \tilde{q}_W, \theta_c \rangle_{\mathscr{C}},\end{aligned} \tag{P2.3}$$

$$\forall \boldsymbol{q} \in \boldsymbol{Z}, \quad \big(\boldsymbol{K}^{-1}\boldsymbol{z}, \boldsymbol{q}\big) = (p, \nabla \cdot \boldsymbol{q}) - \langle p_c, [\boldsymbol{q}]_{\mathscr{C}} \cdot \boldsymbol{n}^+ \rangle_{\mathscr{C}} + \big(\nabla(\rho_{f,r} g \eta), \boldsymbol{q}\big), \tag{P2.4}$$

$$\forall \boldsymbol{q}_c \in \boldsymbol{Z}_{\mathscr{C}}, \quad \big(\boldsymbol{\zeta}, \boldsymbol{q}_c\big)_{\mathscr{C}} = \langle p_c, \overline{\nabla} \cdot (w^{\frac{3}{2}} \boldsymbol{q}_c) \rangle_{\mathscr{C}} + \big(w^{\frac{3}{2}} \overline{\nabla}(\rho_{f,r} g \eta), \boldsymbol{q}_c\big)_{\mathscr{C}}, \tag{P2.5}$$

subject to the initial condition (16):

$$\left(\left(\frac{1}{M} + c_f \varphi_0\right)p + \alpha \nabla \cdot \boldsymbol{u}\right)(0) = \left(\frac{1}{M} + c_f \varphi_0\right)p_0 + \alpha \nabla \cdot \boldsymbol{u}_0.$$

From the assumptions on the data and the choice of spaces for the solution we infer that

$$\begin{aligned}\frac{\partial}{\partial t}\left(\left(\frac{1}{M} + c_f \varphi_0\right)p + \alpha \nabla \cdot \boldsymbol{u}\right) &\in L^2((\Omega \setminus \mathscr{C}) \times]0, T[), \\ \frac{\partial}{\partial t}\big([\boldsymbol{u}]_{\mathscr{C}} \cdot \boldsymbol{n}^+\big) &\in L^2(0, T; H^{-\frac{1}{2}}(\mathscr{C})).\end{aligned} \tag{24}$$

Thus, the first part of (24) implies that the initial condition (16) is meaningful.

We have the following equivalence result.

Theorem 2 *Let* $\boldsymbol{f} \in L^2\left((\Omega \setminus \mathscr{C}) \times]0,T[\right)^d$, $\tilde{q} \in L^2(\Omega \times]0,T[)$, $\tilde{q}_W \in L^2(0,T;H^{-\frac{1}{2}}(\mathscr{C}))$ *and assume Hypothesis 1 holds. Suppose that Problem 1 has a solution with the following regularity:*

$$p \in L^\infty(0,T;Q), \quad \boldsymbol{u} \in L^\infty(0,T;\mathbb{V}), \quad \tilde{q}_L \in L^2(0,T;H^{-\frac{1}{2}}(\mathscr{C})),$$

and such that (24) holds. Then by defining $\boldsymbol{z}$ *and* $\boldsymbol{\zeta}$ *through (18) and (19) respectively, and by setting* p_c *the trace of* p *on* $\mathscr{C}$*, this solution also satisfies (P2.1)–(P2.5) and (16). Conversely, any solution of the mixed formulation (P2.1)–(P2.5) and (16) with* w *and* $\tilde{q}_L$ *defined respectively by (13) and (23), also solves Problem 1.*

Proof Consider first the flow equations in Problem 1. We set $p_c = p|_{\mathscr{C}}$ and we have in particular $p_c \in L^2(0,T;H^1_w(\mathscr{C}))$. From the assumptions on the time derivative of p and $\boldsymbol{u}$ and the regularity of $\tilde{q}$, we infer from the definition (18) of $\boldsymbol{z}$ and (P1.3) that $\boldsymbol{z}$ belongs to $H(\text{div};\Omega \setminus \mathscr{C})$, and thus $\boldsymbol{z} \in L^2(0,T;\boldsymbol{Z})$, owing to the boundary conditions (P1.8). Then (P1.3) gives (P2.2). Similarly, the definition of w and the assumption (24) on its time derivative, the assumption on $\tilde{q}_W$ and $\tilde{q}_L$ and its formula (23), and (P1.4) and (P1.7) imply that $\overline{\nabla} \cdot (w^3 \overline{\nabla} p_c)$ belongs to $L^2(0,T;H^{-\frac{1}{2}}(\mathscr{C}))$. Then we infer from the trace space $H^1_w(\mathscr{C})$ for p on $\mathscr{C}$ and the definition (19) of $\boldsymbol{\zeta}$ that $\boldsymbol{\zeta}$ belongs to $L^2(0,T;\boldsymbol{Z}_{\mathscr{C}})$ and (P1.4), (P1.6), and (P1.7) yield (P2.3).

Next, we turn to the elasticity equation. The assumptions on $\boldsymbol{f}$, $\boldsymbol{u}$ and p imply that each row of $\boldsymbol{\sigma}^{\text{por}}(\boldsymbol{u},p)$ belongs to $L^2(0,T;H(\text{div};\Omega^\star))$, thus implying that the normal trace of $\boldsymbol{\sigma}^{\text{por}}(\boldsymbol{u},p)$ on $\partial\Omega^\star$ is well defined, $\star = +,-$. Thus we can take the scalar product of (P1.1) with $\boldsymbol{v}$ in $\mathbb{V}$, apply Green's formula in $\Omega^\star$, see, for instance, [16], and it remains to recover the boundary term in (P2.1). We know that the normal trace of $\boldsymbol{\sigma}^{\text{por}}(\boldsymbol{u},p)$ is continuous through $\Gamma \setminus \mathscr{C}$. More precisely, each row of $\boldsymbol{\sigma}^{\text{por}}(\boldsymbol{u},p)$ belongs to $L^2(0,T;V_{\text{div}})$, where

$$V_{\text{div}} = \{\boldsymbol{v} \in L^2(\Omega)^d \mid \nabla \cdot \boldsymbol{v} \in L^2(\Omega^\star),\ \star = +,-,\ [\boldsymbol{v}]_{\Gamma\setminus\mathscr{C}} \cdot \boldsymbol{n}^+ = 0\}.$$

Hence (P1.5) is meaningful. Furthermore, since p_c belongs to $L^2(0,T;H^{\frac{1}{2}}(\mathscr{C}))$, the product $p_c\boldsymbol{n}$ is in $L^2(0,T;L^r(\mathscr{C})^d)$ for any real $r \geq 1$ when $d=2$ and $r \in [1,4]$ when $d=3$. Therefore $\boldsymbol{\sigma}^{\text{por}}(\boldsymbol{u},p)\boldsymbol{n}|_{\mathscr{C}}$ belongs to $L^2(0,T;L^r(\mathscr{C})^d)$, and consequently the boundary term in (P2.1) reduces to an integral on $\mathscr{C}$, as $\boldsymbol{v}$ vanishes on $\Gamma^\star$. This justifies the derivation of (P2.1). Finally, we consider the velocity equations. An application of the usual Green formula in (18) yields (P2.4), and (P2.5) follows from (19) and (22).

Conversely, consider a solution of (P2.1)–(P2.5) starting from (16). By choosing $\boldsymbol{q} \in \mathfrak{D}(\Omega^\star)^d$ in (P2.4), $\star = +,-$, we obtain

$$\boldsymbol{z} = -\boldsymbol{K}\nabla(p - \rho_{f,r} g \eta) \quad \text{in } \Omega^\star,\ \star = +,-, \tag{25}$$

which implies that p belongs to $H^1(\Omega^\star)$. Next, by taking the scalar product of (25) with $\boldsymbol{q} \in H^1_0(\Omega)^d$ and applying the usual Green formula, we derive

$$\forall \boldsymbol{q} \in H_0^1(\Omega)^d, \quad (\boldsymbol{K}^{-1}\boldsymbol{z}, \boldsymbol{q}) = (p, \nabla \cdot \boldsymbol{q}) - \big([p]_\Gamma, \boldsymbol{q} \cdot \boldsymbol{n}^+\big)_\Gamma + (\nabla(\rho_{f,r} g \eta), \boldsymbol{q}). \quad (26)$$

Then comparing with (P2.4), we infer that

$$\forall \boldsymbol{q} \in H_0^1(\Omega)^d, \quad \big([p]_\Gamma, \boldsymbol{q} \cdot \boldsymbol{n}^+\big)_\Gamma = 0,$$

whence $[p]_\Gamma = 0$, therefore $p \in H^1(\Omega)$ and (26) reduces to

$$\forall \boldsymbol{q} \in H_0^1(\Omega)^d, \quad (\boldsymbol{K}^{-1}\boldsymbol{z}, \boldsymbol{q}) = -(\nabla(p - \rho_{f,r} g \eta), \boldsymbol{q}),$$

thus implying (18). Then, by substituting (18) into (P2.2), we recover (P1.3). Next, by taking the scalar product of (25) with $\boldsymbol{q} \in H^1(\Omega^\star)^d$, $\star = +, -$, $\boldsymbol{q} = \boldsymbol{0}$ on $\partial\Omega$, applying Green's formula and comparing with (P2.4), we obtain

$$\big(p, [\boldsymbol{q}]_{\mathscr{C}} \cdot \boldsymbol{n}^+\big)_{\mathscr{C}} = \big(p_c, [\boldsymbol{q}]_{\mathscr{C}} \cdot \boldsymbol{n}^+\big)_{\mathscr{C}}.$$

Hence $p_c = p|_{\mathscr{C}}$. Now, from (P2.1), we derive as above (P1.1) and (P1.2) in $\Omega^\star$. The essential homogeneous Dirichlet boundary condition is included in the space $\mathbb{V}$. To recover the interface condition, take the scalar product of (P1.1) with $\boldsymbol{v} \in V$ such that $\boldsymbol{v}^- = \boldsymbol{0}$, $\boldsymbol{v}$ sufficiently smooth in Ω^+, not zero on $\mathscr{C}$, and apply Green's formula. Comparing with (P2.1), this gives

$$\langle \boldsymbol{\sigma}^{\text{por}}(\boldsymbol{u}^+, p)\boldsymbol{n}^+, \boldsymbol{v}^+ \rangle_{\mathscr{C}} = -\big(p_c, \boldsymbol{n}^+ \cdot \boldsymbol{v}^+\big)_{\mathscr{C}}.$$

With the same argument in Ω^-, we recover on one hand (P1.5) and on the other hand (P1.1) and (P1.2) in $\Omega \setminus \mathscr{C}$. Finally, we define $\tilde{q}_L$ by (15) and w by (13). Then (P2.5) yields (19) in the sense of distributions on $\mathscr{C}$ and (P2.3) implies (P1.4). □

In the sequel, we suppose that the assumptions of Theorem 2 hold.

4 Stability and Existence of the Mixed Formulation's Solution

From now on we restrict our study to a linearized version of the mixed problem where the factor $w^{\frac{3}{2}}$ in (P2.3) and (P2.5) is assumed to be known. This would be the case in a time-stepping algorithm, where w is taken at the previous time step.

Proving existence of solutions of the mixed formulation in the above spaces is not trivial because it requires sufficiently smooth solutions. This is the price to pay when switching to a mixed form, and here it is aggravated by the time dependence and the presence of the fracture. Our purpose in this section is to give sufficient conditions on the data and the width w for establishing existence. One of them is restrictive, see Hypothesis 2, but by-passing it is an open question.

We start with a priori estimates; a priori in the sense that they are obtained under the assumption that the mixed problem has a solution. The geometrical setting is the one described in Sect. 2.

4.1 First Set of a Priori Estimates

We derive here a set of a priori estimates under basic regularity assumptions on the solution. Albeit basic, these estimates cannot be derived without an inf-sup condition on the pressure p_c in the fracture.

4.1.1 Inf-sup Condition for p_c

Lemma 1 *There exists a constant* $\beta > 0$ *such that*

$$\forall p_c \in H^{\frac{1}{2}}(\mathscr{C}), \quad \sup_{\boldsymbol{q} \in \boldsymbol{Z}} \frac{\langle p_c, [\boldsymbol{q}]_{\mathscr{C}} \cdot \boldsymbol{n}^+\rangle_{\mathscr{C}}}{\|\boldsymbol{q}\|_{\boldsymbol{Z}}} \geq \beta \|p_c\|_{H^{\frac{1}{2}}(\mathscr{C})}. \tag{27}$$

Proof By duality we write:

$$\|p_c\|_{H^{\frac{1}{2}}(\mathscr{C})} = \sup_{g \in H^{-\frac{1}{2}}(\mathscr{C})} \frac{\langle p_c, g\rangle_{\mathscr{C}}}{\|g\|_{H^{-\frac{1}{2}}(\mathscr{C})}},$$

and the proof relies on relating g to a suitable function $\boldsymbol{q}$ in $\boldsymbol{Z}$. We proceed in two steps:

1. We propose to extend g by zero to $\partial\Omega^+$. For this, let $E(g)$ be defined by

$$\forall \varphi \in H^{\frac{1}{2}}(\partial\Omega^+), \quad \langle E(g), \varphi\rangle_{\partial\Omega^+} = \langle g, \varphi\rangle_{\mathscr{C}}.$$

Then

$$|\langle E(g), \varphi\rangle_{\partial\Omega^+}| \leq \|g\|_{H^{-\frac{1}{2}}(\mathscr{C})}\|\varphi\|_{H^{\frac{1}{2}}(\mathscr{C})} \leq \|g\|_{H^{-\frac{1}{2}}(\mathscr{C})}\|\varphi\|_{H^{\frac{1}{2}}(\partial\Omega^+)}.$$

Thus $E(g) \in H^{-\frac{1}{2}}(\partial\Omega^+)$ and

$$\|E(g)\|_{H^{-\frac{1}{2}}(\partial\Omega^+)} \leq \|g\|_{H^{-\frac{1}{2}}(\mathscr{C})}.$$

Moreover, for all $\varphi \in H^{\frac{1}{2}}(\partial\Omega^+)$ that vanish on $\mathscr{C}$ (i.e. $\varphi \in H^{\frac{1}{2}}_{00}(\partial\Omega^+ \setminus \mathscr{C})$, we have $\langle E(g), \varphi\rangle_{\partial\Omega^+} = 0$. This means that $E(g) = 0$ on $\partial\Omega^+ \setminus \mathscr{C}$. Finally, for all $\varphi \in H^{\frac{1}{2}}_{00}(\mathscr{C})$, we have

$$\langle E(g), \varphi\rangle_{\mathscr{C}} = \langle E(g), \varphi\rangle_{\partial\Omega^+} = \langle g, \varphi\rangle_{\mathscr{C}}.$$

Hence $E(g)$ is the desired extension.

2. As $E(g)$ belongs to $H^{-\frac{1}{2}}(\partial\Omega^+)$, there exists $\boldsymbol{q}^+$ in $H(\mathrm{div};\,\Omega^+)$ such that (see, for instance, [16])

$$\boldsymbol{q}^+ \cdot \boldsymbol{n}^+ = E(g) \quad \text{on } \partial\Omega^+,$$

and, with a constant C that depends only on Ω^+ and $\mathscr{C}$

$$\|\boldsymbol{q}^+\|_{H(\mathrm{div};\Omega^+)} \le C\|E(g)\|_{H^{-\frac{1}{2}}(\partial\Omega^+)} \le C\|g\|_{H^{-\frac{1}{2}}(\mathscr{C})}.$$

Furthermore

$$\boldsymbol{q}^+ \cdot \boldsymbol{n}^+ = 0 \quad \text{on } \partial\Omega^+ \setminus \mathscr{C}.$$

Now, we choose $\boldsymbol{q}^- = \boldsymbol{0}$ in Ω^-. Then $\boldsymbol{q}$ is in $\boldsymbol{Z}$,

$$[\boldsymbol{q}] \cdot \boldsymbol{n}^+ = g \quad \text{on } \mathscr{C},$$

and

$$\|\boldsymbol{q}\|_{\boldsymbol{Z}} \le C\|g\|_{H^{-\frac{1}{2}}(\mathscr{C})}.$$

Thus

$$\|p_c\|_{H^{\frac{1}{2}}(\mathscr{C})} = \sup_{g\in H^{-\frac{1}{2}}(\mathscr{C})} \frac{\langle p_c, E(g)\rangle_{\mathscr{C}}}{\|g\|_{H^{-\frac{1}{2}}(\mathscr{C})}} = \sup_{g\in H^{-\frac{1}{2}}(\mathscr{C})} \frac{\langle p_c, [\boldsymbol{q}]_{\mathscr{C}} \cdot \boldsymbol{n}^+\rangle_{\mathscr{C}}}{\|g\|_{H^{-\frac{1}{2}}(\mathscr{C})}}$$
$$\le C \sup_{\boldsymbol{q}\in\boldsymbol{Z}} \frac{\langle p_c, [\boldsymbol{q}]_{\mathscr{C}} \cdot \boldsymbol{n}^+\rangle_{\mathscr{C}}}{\|\boldsymbol{q}\|_{\boldsymbol{Z}}},$$

and this yields (27), with $\beta = \frac{1}{C}$. □

Note that the bilinear form $\langle p_c, [\boldsymbol{q}]_{\mathscr{C}} \cdot \boldsymbol{n}^+\rangle_{\mathscr{C}}$ is continuous on the product space $H^{\frac{1}{2}}(\mathscr{C}) \times \boldsymbol{Z}$. We associate with this form the operator B and its dual operator B' defined by

$$\forall \boldsymbol{q} \in \boldsymbol{Z},\ \forall p_c \in H^{\frac{1}{2}}(\mathscr{C}), \quad \langle B\boldsymbol{q}, p_c\rangle = \langle p_c, [\boldsymbol{q}]_{\mathscr{C}} \cdot \boldsymbol{n}^+\rangle_{\mathscr{C}} = \langle \boldsymbol{q}, B'p_c\rangle.$$

The kernel of B in $\boldsymbol{Z}$ is the space

$$H_0(\mathrm{div};\,\Omega) = \{\boldsymbol{q} \in H(\mathrm{div};\,\Omega) \mid \boldsymbol{q} \cdot \boldsymbol{n} = 0 \ \text{ on } \partial\Omega\}.$$

Indeed, $B\boldsymbol{q} = 0$ is equivalent to $[\boldsymbol{q}]_{\mathscr{C}} \cdot \boldsymbol{n}^+ = 0$, which means that $\boldsymbol{q} \cdot \boldsymbol{n}^+$ does not jump on $\mathscr{C}$. Then the inf-sup condition (27) has the following consequence.

Corollary 1 *Let $z \in \boldsymbol{Z}$ and $p \in L^2(\Omega)$ be such that*

$$\forall \boldsymbol{q} \in H_0(\mathrm{div};\, \Omega), \quad (p, \nabla \cdot \boldsymbol{q}) - \big(\boldsymbol{K}^{-1}\boldsymbol{z}, \boldsymbol{q}\big) + \big(\nabla(\rho_{f,r} g \eta), \boldsymbol{q}\big) = 0. \tag{28}$$

Then there exists a unique p_c in $H^{\frac{1}{2}}(\mathscr{C})$ such that p_c, p, and $\boldsymbol{z}$ satisfy (P2.4) *and*

$$\|p_c\|_{H^{\frac{1}{2}}(\mathscr{C})} \le \frac{1}{\beta}\Big(\big(\|p\|^2_{L^2(\Omega)} + |\rho_{f,r} g \eta|^2_{H^1(\Omega\setminus\mathscr{C})}\big)^{\frac{1}{2}} + \|\boldsymbol{K}^{-1}\boldsymbol{z}\|_{L^2(\Omega\setminus\mathscr{C})}\Big), \tag{29}$$

where β is the constant of (27).

Proof It stems from (27) and Babuška-Brezzi's theory (see, for instance, [2, 6] or [16]) that the mapping B' is an isomorphism from $H^{\frac{1}{2}}(\mathscr{C})$ onto the subspace of $\boldsymbol{Z}'$:

$$\{\boldsymbol{\ell} \in \boldsymbol{Z}' \mid \forall \boldsymbol{w} \in H_0(\mathrm{div};\, \Omega), \quad \langle \boldsymbol{\ell}, \boldsymbol{w}\rangle = 0\}.$$

Now, for given $\boldsymbol{z} \in \boldsymbol{Z}$ and $p \in L^2(\Omega)$ satisfying (28), let $\boldsymbol{\ell}$ denote the mapping

$$\forall \boldsymbol{q} \in \boldsymbol{Z}, \quad \langle \boldsymbol{\ell}, \boldsymbol{q}\rangle = (p, \nabla \cdot \boldsymbol{q}) - \big(\boldsymbol{K}^{-1}\boldsymbol{z}, \boldsymbol{q}\big) + \big(\nabla(\rho_{f,r} g \eta), \boldsymbol{q}\big).$$

Clearly, $\boldsymbol{\ell}$ belongs to $\boldsymbol{Z}'$ and by assumption $\boldsymbol{\ell}$ vanishes on $H_0(\mathrm{div};\, \Omega)$. Therefore there exists a unique p_c in $H^{\frac{1}{2}}(\mathscr{C})$ such that

$$\forall \boldsymbol{q} \in \boldsymbol{Z}, \quad \langle B' p_c, \boldsymbol{q}\rangle = \langle \boldsymbol{\ell}, \boldsymbol{q}\rangle,$$

i.e.

$$\langle p_c, [\boldsymbol{q}]_{\mathscr{C}} \cdot \boldsymbol{n}^+\rangle_{\mathscr{C}} = (p, \nabla \cdot \boldsymbol{q}) - \big(\boldsymbol{K}^{-1}\boldsymbol{z}, \boldsymbol{q}\big) + \big(\nabla(\rho_{f,r} g \eta), \boldsymbol{q}\big),$$

and the estimate (29) follows easily from this and (27). □

4.1.2 Stability Estimates

As specified at the beginning of this section, we assume that the mixed problem has a sufficiently smooth solution p, $\boldsymbol{u}$, p_c, $\boldsymbol{z}$, and $\boldsymbol{\zeta}$; its precise regularity will be stated further on. A preliminary stability equality is derived by testing (P2.2) with $\theta = p$, (P2.1) with $\boldsymbol{v} = \boldsymbol{u}'$, (P2.3) with $\theta_c = p_c$, (P2.4) with $\boldsymbol{q} = \boldsymbol{z}$ and (P2.5) with $\boldsymbol{q}_c = \boldsymbol{\zeta}$. Recall that scalar products in space are denoted by parentheses and if the domain of integration is not indicated, then it is understood that the integrals are taken over $\Omega^+ \cup \Omega^-$. Thus we obtain the five equalities:

$$
\begin{gathered}
\tfrac{1}{2}\left(\tfrac{1}{M}+c_f\varphi_0\right)\tfrac{d}{dt}\|p(t)\|^2_{L^2(\Omega)}+\alpha(\nabla\cdot\boldsymbol{u}'(t),p(t))+\tfrac{1}{\mu_f}(\nabla\cdot\boldsymbol{z}(t),p(t))=(\tilde{q}(t),p(t)),\\
G\tfrac{d}{dt}\|\boldsymbol{\varepsilon}(\boldsymbol{u}(t))\|^2_{L^2(\Omega\setminus\mathscr{C})}+\tfrac{\lambda}{2}\tfrac{d}{dt}\|\nabla\cdot\boldsymbol{u}(t)\|^2_{L^2(\Omega\setminus\mathscr{C})}-\alpha(p(t),\nabla\cdot\boldsymbol{u}'(t))\\
+\langle p_c(t),[\boldsymbol{u}'(t)]_{\mathscr{C}}\cdot\boldsymbol{n}^+\rangle_{\mathscr{C}}=(\boldsymbol{f}(t),\boldsymbol{u}'(t)),\\
-\langle[\boldsymbol{u}'(t)]_{\mathscr{C}}\cdot\boldsymbol{n}^+,p_c(t)\rangle_{\mathscr{C}}+\tfrac{1}{12\mu_f}\langle\overline{\nabla}\cdot(w^{\frac{3}{2}}(t)\boldsymbol{\zeta}(t)),p_c(t)\rangle_{\mathscr{C}}-\tfrac{1}{\mu_f}\langle[\boldsymbol{z}]_{\mathscr{C}}\cdot\boldsymbol{n}^+,p_c(t)\rangle_{\mathscr{C}}\\
=\langle\tilde{q}_W(t),p_c(t)\rangle_{\mathscr{C}},\\
\tfrac{1}{\mu_f}\|\boldsymbol{K}^{-\frac{1}{2}}\boldsymbol{z}(t)\|^2_{L^2(\Omega\setminus\mathscr{C})}=\tfrac{1}{\mu_f}(p(t),\nabla\cdot\boldsymbol{z}(t))-\tfrac{1}{\mu_f}\langle[\boldsymbol{z}(t)]_{\mathscr{C}}\cdot\boldsymbol{n}^+,p_c(t)\rangle_{\mathscr{C}}\\
+\tfrac{1}{\mu_f}(\nabla(\rho_{f,r}g\eta),\boldsymbol{z}(t)),\\
\tfrac{1}{12\mu_f}\|\boldsymbol{\zeta}(t)\|^2_{L^2(\mathscr{C})}=\tfrac{1}{12\mu_f}\langle\overline{\nabla}\cdot(w^{\frac{3}{2}}(t)\boldsymbol{\zeta}(t)),p_c(t)\rangle_{\mathscr{C}}+\tfrac{1}{12\mu_f}\left(w^{\frac{3}{2}}(t)\overline{\nabla}(\rho_{f,r}g\eta),\boldsymbol{\zeta}(t)\right)_{\mathscr{C}}.
\end{gathered}
\tag{30}
$$

The second equation is problematic because we have no information on the time derivative of $\boldsymbol{u}$ that appears in its right-hand side. Therefore, supposing $\boldsymbol{f}$ is differentiable in time, we write

$$
(\boldsymbol{f}(t),\boldsymbol{u}'(t))=\frac{d}{dt}(\boldsymbol{f}(t),\boldsymbol{u}(t))-(\boldsymbol{f}'(t),\boldsymbol{u}(t))
$$

and use instead

$$
\begin{aligned}
G\frac{d}{dt}\|\boldsymbol{\varepsilon}(\boldsymbol{u}(t))\|^2_{L^2(\Omega\setminus\mathscr{C})}+\frac{\lambda}{2}\frac{d}{dt}\|\nabla\cdot\boldsymbol{u}(t)\|^2_{L^2(\Omega\setminus\mathscr{C})}-\alpha(p(t),\nabla\cdot\boldsymbol{u}'(t))&\\
+\langle p_c(t),[\boldsymbol{u}'(t)]_{\mathscr{C}}\cdot\boldsymbol{n}^+\rangle_{\mathscr{C}}=\frac{d}{dt}(\boldsymbol{f}(t),\boldsymbol{u}(t))-(\boldsymbol{f}'(t),\boldsymbol{u}(t)).&
\end{aligned}
$$

In addition, for the sake of convenience, we rewrite the last two equations as

$$
\begin{aligned}
\frac{1}{\mu_f}\|\boldsymbol{K}^{-\frac{1}{2}}\boldsymbol{z}(t)\|^2_{L^2(\Omega\setminus\mathscr{C})}-\frac{1}{\mu_f}(p(t),\nabla\cdot\boldsymbol{z}(t))+\frac{1}{\mu_f}\langle[\boldsymbol{z}(t)]_{\mathscr{C}}\cdot\boldsymbol{n}^+,p_c(t)\rangle_{\mathscr{C}}&\\
=\frac{1}{\mu_f}(\nabla(\rho_{f,r}g\eta),\boldsymbol{z}(t)),&\qquad(31)\\
\frac{1}{12\mu_f}\|\boldsymbol{\zeta}(t)\|^2_{L^2(\mathscr{C})}-\frac{1}{12\mu_f}\langle\overline{\nabla}\cdot(w^{\frac{3}{2}}(t)\boldsymbol{\zeta}(t)),p_c(t)\rangle_{\mathscr{C}}&\\
=\frac{1}{12\mu_f}\left(w^{\frac{3}{2}}(t)\overline{\nabla}(\rho_{f,r}g\eta),\boldsymbol{\zeta}(t)\right)_{\mathscr{C}}.&\qquad(32)
\end{aligned}
$$

When adding the five equations (30) through (32), we find the following stability equality:

$$\frac{1}{2}\left(\frac{1}{M}+c_f\varphi_0\right)\frac{d}{dt}\|p(t)\|^2_{L^2(\Omega)}+G\frac{d}{dt}\|\boldsymbol{\varepsilon}(\boldsymbol{u}(t))\|^2_{L^2(\Omega\setminus\mathscr{C})}+\frac{\lambda}{2}\frac{d}{dt}\|\nabla\cdot\boldsymbol{u}(t)\|^2_{L^2(\Omega\setminus\mathscr{C})}$$
$$+\frac{1}{\mu_f}\|\boldsymbol{K}^{-\frac{1}{2}}\boldsymbol{z}(t)\|^2_{L^2(\Omega\setminus\mathscr{C})}+\frac{1}{12\mu_f}\|\boldsymbol{\zeta}(t)\|^2_{L^2(\mathscr{C})}$$
$$=(\tilde{q}(t),p(t))+\frac{d}{dt}(\boldsymbol{f}(t),\boldsymbol{u}(t))-(\boldsymbol{f}'(t),\boldsymbol{u}(t))+\langle\tilde{q}_W(t),p_c(t)\rangle_{\mathscr{C}}$$
$$+\frac{1}{\mu_f}(\nabla(\rho_{f,r}g\eta),\boldsymbol{z}(t))+\frac{1}{12\mu_f}\left(w^{\frac{3}{2}}(t)\overline{\nabla}(\rho_{f,r}g\eta),\boldsymbol{\zeta}(t)\right)_{\mathscr{C}}. \quad (33)$$

Corollary 1 will be used to control the pressure p_c in the fracture; it appears in the fourth term of the right-hand side of (33), but cannot be absorbed by any term of the left-hand side.

Theorem 3 *Let the data satisfy* $\boldsymbol{f}\in H^1(0,T;L^2(\Omega\setminus\mathscr{C})^d)$, $\tilde{q}\in L^2(\Omega\times]0,T[)$, $\tilde{q}_W\in L^2(0,T;H^{-\frac{1}{2}}(\mathscr{C}))$*, and suppose that* w *verifies Hypothesis 1 and* $\rho_{f,r}g\eta$ *is independent of time and belongs both to* $H^1(\Omega\setminus\mathscr{C})$ *and* $H^1(\mathscr{C})$*. If* $p\in H^1(0,T;L^2(\Omega))$, $\boldsymbol{u}\in H^1(0,T;H^1(\Omega\setminus\mathscr{C})^d)$, $p_c\in L^2(0,T;H^{\frac{1}{2}}(\mathscr{C}))$, $\boldsymbol{z}\in L^2(0,T;\boldsymbol{Z})$ *and* $\boldsymbol{\zeta}\in L^2(0,T;\boldsymbol{Z}_{\mathscr{C}})$ *is a solution of the mixed problem* (P2.1)–(P2.5) *and* (16)*, then it satisfies the following a priori bound almost everywhere in* $]0,T[$*:*

$$\left(\tfrac{1}{M}+c_f\varphi_0\right)\|p(t)\|^2_{L^2(\Omega)}+G\|\boldsymbol{\varepsilon}(\boldsymbol{u}(t))\|^2_{L^2(\Omega\setminus\mathscr{C})}+\lambda\|\nabla\cdot\boldsymbol{u}(t)\|^2_{L^2(\Omega\setminus\mathscr{C})}$$
$$+\tfrac{1}{\mu_f}\|\boldsymbol{K}^{-\frac{1}{2}}\boldsymbol{z}\|^2_{L^2((\Omega\setminus\mathscr{C})\times]0,t[)}+\tfrac{1}{12\mu_f}\|\boldsymbol{\zeta}\|^2_{L^2(\mathscr{C}\times]0,t[)}$$
$$\leq C\Big[\left(\tfrac{1}{M}+c_f\varphi_0\right)\|p(0)\|^2_{L^2(\Omega)}+G\|\boldsymbol{\varepsilon}(\boldsymbol{u}(0))\|^2_{L^2(\Omega\setminus\mathscr{C})}+\lambda\|\nabla\cdot\boldsymbol{u}(0)\|^2_{L^2(\Omega\setminus\mathscr{C})}$$
$$+\|\boldsymbol{u}(0)\|^2_{L^2(\Omega\setminus\mathscr{C})}+\|\boldsymbol{f}(0)\|^2_{L^2(\Omega\setminus\mathscr{C})}+\|\tilde{q}\|^2_{L^2((\Omega\setminus\mathscr{C})\times]0,t[)}+\|\boldsymbol{f}\|^2_{H^1(0,t;L^2(\Omega\setminus\mathscr{C})^d)}$$
$$+\|\tilde{q}_W\|^2_{L^2(0,t;H^{-\frac{1}{2}}(\mathscr{C}))}+t|\rho_{f,r}g\eta|^2_{H^1(\Omega\setminus\mathscr{C})}+\tfrac{1}{12\mu_f}\|w^{\frac{3}{2}}\overline{\nabla}(\rho_{f,r}g\eta)\|^2_{L^2(\mathscr{C}\times]0,t[)}\Big]\exp(t), \quad (34)$$

$$\|p_c\|_{L^2(0,t;H^{\frac{1}{2}}(\mathscr{C}))}\leq\frac{\sqrt{2}}{\beta}\left(\|p\|^2_{L^2(\Omega\times]0,t[)}+\|\boldsymbol{K}^{-1}\boldsymbol{z}\|^2_{L^2((\Omega\setminus\mathscr{C})\times]0,t[)}+t|\rho_{f,r}g\eta|^2_{H^1(\Omega\setminus\mathscr{C})}\right)^{\frac{1}{2}}, \quad (35)$$

with the constant β *of* (27) *and a constant* C *that depends on* α, $\|\boldsymbol{K}^{-\frac{1}{2}}\|_{L^\infty(\Omega\setminus\mathscr{C})}$, $\|\boldsymbol{K}^{\frac{1}{2}}\|_{L^\infty(\Omega\setminus\mathscr{C})}$, $1/(\frac{1}{M}+c_f\varphi_0)$*, and* $1/\mu_f$*, but is independent of* t*.*

Proof Deriving a bound from the stability equality (33) and the pressure bound (29) is straightforward. Of course (35) is an immediate consequence of (29), and it suffices to derive (34). We integrate (33) over $]0,t[$, $t>0$ and bound the terms in the right-hand side with positive constants δ_i that will be adjusted at the end. The first term is bounded by

$$\left|\int_0^t (\tilde{q}(s), p(s))\, ds\right| \leq \frac{1}{2}\Big(\delta_1 \left(\frac{1}{M} + c_f \varphi_0\right) \|p\|^2_{L^2(\Omega\times]0,t[)} \\ + \frac{1}{\delta_1} \frac{1}{\frac{1}{M} + c_f\varphi_0} \|\tilde{q}\|^2_{L^2((\Omega\setminus\mathscr{C})\times]0,t[)}\Big).$$

For the second term, we use Poincaré's and Korn's inequalities (2) and (3):

$$\left|\int_0^t \frac{d}{ds}(\boldsymbol{f}(s), \boldsymbol{u}(s)) ds\right| \leq \frac{1}{2}\Big(\delta_2 G \|\boldsymbol{\varepsilon}(\boldsymbol{u}(t))\|^2_{L^2(\Omega\setminus\mathscr{C})} + C^2 \frac{1}{\delta_2}\frac{1}{G} \|\boldsymbol{f}(t)\|^2_{L^2(\Omega\setminus\mathscr{C})} \\ + \|\boldsymbol{u}(0)\|^2_{L^2(\Omega\setminus\mathscr{C})} + \|\boldsymbol{f}(0)\|^2_{L^2(\Omega\setminus\mathscr{C})}\Big),$$

where C is the product of the constants in (2) and (3). Similarly, the third term has the bound

$$\Big|\int_0^t (\boldsymbol{f}'(s), \boldsymbol{u}(s)) ds\Big| \leq \frac{1}{2}\Big(\delta_3 G \|\boldsymbol{\varepsilon}(\boldsymbol{u})\|^2_{L^2((\Omega\setminus\mathscr{C})\times]0,t[)} + C^2 \frac{1}{\delta_3}\frac{1}{G} \|\boldsymbol{f}'\|^2_{L^2((\Omega\setminus\mathscr{C})\times]0,t[)}\Big).$$

For the fourth term, applying (29), we write

$$\begin{aligned}\Big|\langle \tilde{q}_W(t), p_c(t)\rangle_{\mathscr{C}}\Big| \leq& \frac{1}{\beta}\|\tilde{q}_W(t)\|_{H^{-\frac{1}{2}}(\mathscr{C})}\Big(\big(\|p(t)\|^2_{L^2(\Omega)} + |\rho_{f,r} g \eta|^2_{H^1(\Omega\setminus\mathscr{C})}\big)^{\frac{1}{2}} \\ &+ \|\boldsymbol{K}^{-\frac{1}{2}}\|_{L^\infty(\Omega\setminus\mathscr{C})} \|\boldsymbol{K}^{-\frac{1}{2}}\boldsymbol{z}(t)\|_{L^2(\Omega\setminus\mathscr{C})}\Big) \\ \leq& \frac{\delta_4}{2}\big(\frac{1}{M} + c_f\varphi_0\big)\big(\|p(t)\|^2_{L^2(\Omega)} + |\rho_{f,r} g\eta|^2_{H^1(\Omega\setminus\mathscr{C})}\big) + \frac{\delta_5}{2\mu_f}\|\boldsymbol{K}^{-\frac{1}{2}}\boldsymbol{z}(t)\|^2_{L^2(\Omega\setminus\mathscr{C})} \\ &+ \frac{1}{2\beta^2}\Big(\frac{1}{\delta_4}\frac{1}{\frac{1}{M} + c_f\varphi_0} + \frac{\mu_f}{\delta_5}\|\boldsymbol{K}^{-\frac{1}{2}}\|^2_{L^\infty(\Omega\setminus\mathscr{C})}\Big)\|\tilde{q}_W(t)\|^2_{H^{-\frac{1}{2}}(\mathscr{C})}.\end{aligned}$$

Hence

$$\begin{aligned}\left|\int_0^t \langle \tilde{q}_W(s), p_c(s)\rangle_{\mathscr{C}}\, ds\right| \leq& \frac{\delta_4}{2}\left(\frac{1}{M} + c_f\varphi_0\right)\left(\|p\|^2_{L^2(\Omega\times]0,t[)} + t|\rho_{f,r} g\eta|^2_{H^1(\Omega\setminus\mathscr{C})}\right) \\ &+ \frac{\delta_5}{2\mu_f}\|\boldsymbol{K}^{-\frac{1}{2}}\boldsymbol{z}\|^2_{L^2((\Omega\setminus\mathscr{C})\times]0,t[)} \\ &+ \frac{1}{2\beta^2}\left(\frac{1}{\delta_4}\frac{1}{\frac{1}{M} + c_f\varphi_0} + \frac{\mu_f}{\delta_5}\|\boldsymbol{K}^{-\frac{1}{2}}\|^2_{L^\infty(\Omega\setminus\mathscr{C})}\right)\|\tilde{q}_W\|^2_{L^2(0,t;H^{-\frac{1}{2}}(\mathscr{C}))}.\end{aligned}$$

The fifth term has the bound

$$\left|\int_0^t \frac{1}{\mu_f}(\nabla(\rho_{f,r}g\eta), \boldsymbol{z}(s))ds\right| \le \frac{1}{2}\frac{\delta_6}{\mu_f}\|\boldsymbol{K}^{-\frac{1}{2}}\boldsymbol{z}\|^2_{L^2((\Omega\setminus\mathscr{C})\times]0,t[)} \\ + \frac{1}{2}\frac{t}{\mu_f}\frac{1}{\delta_6}\|\boldsymbol{K}^{\frac{1}{2}}\|^2_{L^\infty(\Omega\setminus\mathscr{C})}|\rho_{f,r}g\eta|^2_{H^1(\Omega\setminus\mathscr{C})}.$$

Finally, we have for the last term

$$\left|\int_0^t \frac{1}{12\mu_f}\left(w^{\frac{3}{2}}(s)\overline{\nabla}(\rho_{f,r}g\eta), \boldsymbol{\zeta}(s)\right)_{\mathscr{C}} ds\right| \le \frac{\delta_7}{24\mu_f}\|\boldsymbol{\zeta}\|^2_{L^2(\mathscr{C}\times]0,t[)} \\ + \frac{1}{\delta_7}\frac{1}{24\mu_f}\|w^{\frac{3}{2}}\overline{\nabla}(\rho_{f,r}g\eta)\|^2_{L^2(\mathscr{C}\times]0,t[)}.$$

With a suitable choice of positive parameters δ_i, $1 \le i \le 7$, all terms in the above right-hand sides that involve the solution can be absorbed by the corresponding terms in the left-hand side of (33). This yields (34). In view of Corollary 1, (35) is immediate. □

4.2 Additional a Priori Estimates

Theorem 3 gives no information on the divergence of $\boldsymbol{z}$ or on the surface divergence of $w^{\frac{3}{2}}\boldsymbol{\zeta}$ on $\mathscr{C}$. As is usual, a bound for these quantities requires an estimate on the time derivative of p and $\boldsymbol{u}$, and this will also yield a bound for the leakage term $\tilde{q}_L$.

In order to estimate these time derivatives, we test (P2.2) with p' and (P2.3) with p_c', then we differentiate (P2.1), (P2.4), and (P2.5) in time, and test them respectively with $\boldsymbol{u}'$, $\boldsymbol{z}$, and $\boldsymbol{\zeta}$. By summing these five equations we obtain

$$\begin{aligned}\left(\tfrac{1}{M} + c_f\varphi_0\right)\|p'(t)\|^2_{L^2(\Omega)} + 2G\|\boldsymbol{\varepsilon}(\boldsymbol{u}'(t))\|^2_{L^2(\Omega\setminus\mathscr{C})} + \lambda\|\nabla\cdot(\boldsymbol{u}'(t))\|^2_{L^2(\Omega\setminus\mathscr{C})} \\ + \tfrac{1}{2\mu_f}\tfrac{d}{dt}\|\boldsymbol{K}^{-\frac{1}{2}}\boldsymbol{z}(t)\|^2_{L^2(\Omega\setminus\mathscr{C})} + \tfrac{1}{24\mu_f}\tfrac{d}{dt}\|\boldsymbol{\zeta}(t)\|^2_{L^2(\mathscr{C})} \\ -\tfrac{1}{12\mu_f}\langle p_c(t), \overline{\nabla}\cdot((w^{\frac{3}{2}})'(t)\boldsymbol{\zeta}(t))\rangle_{\mathscr{C}} - \tfrac{1}{12\mu_f}((w^{\frac{3}{2}})'(t)\overline{\nabla}(\rho_{f,r}g\eta), \boldsymbol{\zeta}(t))_{\mathscr{C}} \\ = (\tilde{q}(t), p'(t)) + (\boldsymbol{f}'(t), \boldsymbol{u}'(t)) - \langle\tilde{q}_W'(t), p_c(t)\rangle_{\mathscr{C}} + \tfrac{d}{dt}\langle\tilde{q}_W(t), p_c(t)\rangle_{\mathscr{C}},\end{aligned} \tag{36}$$

where we have passed the time derivative to the first factor in $\langle\tilde{q}_W(t), p_c'(t)\rangle_{\mathscr{C}}$. The term involving the time derivative of $w^{\frac{3}{2}}$ is written as follows:

$$\overline{\nabla}\cdot((w^{\frac{3}{2}})'(t)\boldsymbol{\zeta}(t)) = \frac{(w^{\frac{3}{2}})'}{w^{\frac{3}{2}}}(t)\overline{\nabla}\cdot(w^{\frac{3}{2}}(t)\boldsymbol{\zeta}(t)) + w^{\frac{3}{2}}(t)\boldsymbol{\zeta}(t)\cdot\overline{\nabla}\left(\frac{(w^{\frac{3}{2}})'}{w^{\frac{3}{2}}}(t)\right), \tag{37}$$

and the factor $\frac{(w^{\frac{3}{2}})'}{w^{\frac{3}{2}}}$ is controlled via the following assumption, that complements Hypothesis 1.

Hypothesis 2 The width function is the product of two positive functions

$$\forall(\boldsymbol{x},t) \in \mathscr{C} \times]0,T[\,, \quad w(\boldsymbol{x},t) = \varphi(\boldsymbol{x})\psi(t), \tag{38}$$

and there exists a constant C such that

$$\forall t \in [0,T], \quad |\frac{\psi'(t)}{\psi(t)}| \leq C. \tag{39}$$

Under this assumption, we have sharper a priori estimates. To simplify, we do not specify the constant below.

Theorem 4 *We retain the assumptions of Theorem 3 and, in addition, suppose that w satisfies Hypothesis 2, the data satisfy $\tilde{q}_W \in H^1(0,T;H^{-\frac{1}{2}}(\mathscr{C}))$, $\boldsymbol{z}(0) \in L^2(\Omega \setminus \mathscr{C})^d$, $\boldsymbol{\zeta}(0) \in L^2(\mathscr{C})^{d-1}$, and $p_c(0) \in H^{\frac{1}{2}}(\mathscr{C})$. Then this solution satisfies the following a priori bound almost everywhere in* $]0,T[$*:*

$$\begin{aligned}\left(\frac{1}{M} + c_f\varphi_0\right)\|p'\|^2_{L^2(\Omega\times]0,t[)} &+ 2G\|\boldsymbol{\varepsilon}(\boldsymbol{u}')\|^2_{L^2((\Omega\setminus\mathscr{C})\times]0,t[)} + \lambda\|\nabla\cdot(\boldsymbol{u}')\|^2_{L^2((\Omega\setminus\mathscr{C})\times]0,t[)}\\ &+ \frac{1}{2\mu_f}\|\boldsymbol{K}^{-\frac{1}{2}}\boldsymbol{z}(t)\|^2_{L^2(\Omega\setminus\mathscr{C})} + \frac{1}{24\mu_f}\|\boldsymbol{\zeta}(t)\|^2_{L^2(\mathscr{C})}\\ &\leq C(\boldsymbol{K}, \boldsymbol{f}, \tilde{q}, \tilde{q}_W, p(0), p_c(0), \boldsymbol{z}(0), \boldsymbol{\zeta}(0), \rho_{f,r}g\eta, t).\end{aligned} \tag{40}$$

Proof Owing to the decomposition (38), we have

$$\frac{(w^{\frac{3}{2}})'}{w^{\frac{3}{2}}}(\boldsymbol{x},t) = \frac{(\psi^{\frac{3}{2}}(t))'}{\psi^{\frac{3}{2}}(t)},$$

that does not depend on $\boldsymbol{x}$. Consequently, on one hand, the product of this factor with p_c belongs to $H^{\frac{1}{2}}(\mathscr{C})$, and, on the other hand, the second term in (37) vanishes. Hence, after an application of (39), (36) implies

$$\begin{aligned}\left(\frac{1}{M} + c_f\varphi_0\right)\|p'(t)\|^2_{L^2(\Omega)} &+ 2G\|\boldsymbol{\varepsilon}(\boldsymbol{u}')(t)\|^2_{L^2(\Omega\setminus\mathscr{C})} + \lambda\|\nabla\cdot(\boldsymbol{u}')(t)\|^2_{L^2(\Omega\setminus\mathscr{C})}\\ &+ \frac{1}{2\mu_f}\frac{d}{dt}\|\boldsymbol{K}^{-\frac{1}{2}}\boldsymbol{z}(t)\|^2_{L^2(\Omega\setminus\mathscr{C})} + \frac{1}{24\mu_f}\frac{d}{dt}\|\boldsymbol{\zeta}(t)\|^2_{L^2(\mathscr{C})}\\ \leq \frac{C}{12\mu_f}|\langle p_c(t), \overline{\nabla}\cdot(w^{\frac{3}{2}}(t)\boldsymbol{\zeta}(t))\rangle_{\mathscr{C}}| &+ \frac{C}{12\mu_f}\|w^{\frac{3}{2}}(t)\overline{\nabla}(\rho_{f,r}g\eta)\|_{L^2(\mathscr{C})}\|\boldsymbol{\zeta}(t)\|_{L^2(\mathscr{C})}\\ &+\|\tilde{q}(t)\|_{L^2(\Omega\setminus\mathscr{C})}\|p'(t)\|_{L^2(\Omega)} + \|\boldsymbol{f}'(t)\|_{L^2(\Omega\setminus\mathscr{C})}\|\boldsymbol{u}'(t)\|_{L^2(\Omega\setminus\mathscr{C})}\\ &+\|\tilde{q}'_W(t)\|_{H^{-\frac{1}{2}}(\mathscr{C})}\|p_c(t)\|_{H^{\frac{1}{2}}(\mathscr{C})} + \frac{d}{dt}\langle\tilde{q}_W(t), p_c(t)\rangle_{\mathscr{C}} = \sum_{i=1}^{6} T_i.\end{aligned} \tag{41}$$

Indeed,

$$\left|\langle p_c(t), \frac{(w^{\frac{3}{2}})'}{w^{\frac{3}{2}}}(t)\overline{\nabla}\cdot(w^{\frac{3}{2}}(t)\boldsymbol{\zeta}(t))\rangle_{\mathscr{C}}\right| = \left|\left\langle p_c(t), \frac{(\psi^{\frac{3}{2}})'}{\psi^{\frac{3}{2}}}(t)\overline{\nabla}\cdot(w^{\frac{3}{2}}(t)\boldsymbol{\zeta}(t))\right\rangle_{\mathscr{C}}\right|$$
$$= \left|\frac{(\psi^{\frac{3}{2}})'}{\psi^{\frac{3}{2}}}(t)\langle p_c(t), \overline{\nabla}\cdot(w^{\frac{3}{2}}(t)\boldsymbol{\zeta}(t))\rangle_{\mathscr{C}}\right|.$$

Regarding T_1, we immediately derive from (32) that

$$\frac{C}{12\mu_f}\left|\left\langle p_c(t), \overline{\nabla}\cdot\left(w^{\frac{3}{2}}(t)\boldsymbol{\zeta}(t)\right)\right\rangle_{\mathscr{C}}\right|$$
$$\leq \frac{C}{12\mu_f}\Big(\|\boldsymbol{\zeta}(t)\|^2_{L^2(\mathscr{C})} + \|w^{\frac{3}{2}}(t)\overline{\nabla}(\rho_{f,r}g\eta)\|_{L^2(\mathscr{C})}\|\boldsymbol{\zeta}(t)\|_{L^2(\mathscr{C})}\Big)$$
$$\leq \frac{1}{12\mu_f}\left(C\|\boldsymbol{\zeta}(t)\|^2_{L^2(\mathscr{C})} + \frac{\delta_1}{2}\|\boldsymbol{\zeta}(t)\|^2_{L^2(\mathscr{C})} + \frac{C^2}{2\delta_1}\|w^{\frac{3}{2}}(t)\overline{\nabla}(\rho_{f,r}g\eta)\|^2_{L^2\mathscr{C})}\right).$$

The bounds for T_i, $2 \leq i \leq 5$, are straightforward, with positive constants δ_j, $2 \leq j \leq 5$:

$$T_2 \leq \frac{1}{24\mu_f}\left(\delta_2\|\boldsymbol{\zeta}(t)\|^2_{L^2(\mathscr{C})} + \frac{C^2}{\delta_2}\|w^{\frac{3}{2}}(t)\overline{\nabla}(\rho_{f,r}g\eta)\|^2_{L^2\mathscr{C}}\right),$$
$$T_3 \leq \frac{1}{2}\left(\delta_3\big(\frac{1}{M} + c_f\varphi_0\big)\|p'(t)\|^2_{L^2(\Omega)} + \frac{1}{\delta_3}\frac{1}{\frac{1}{M} + c_f\varphi_0}\|\tilde{q}(t)\|^2_{L^2(\Omega\setminus\mathscr{C})}\right),$$
$$T_4 \leq \frac{1}{2}\left(2G\delta_4\|\boldsymbol{\varepsilon}(\boldsymbol{u}')(t)\|^2_{L^2(\Omega\setminus\mathscr{C})} + \frac{C^2}{2G\delta_4}\|\boldsymbol{f}'(t)\|^2_{L^2(\Omega\setminus\mathscr{C})}\right),$$
$$T_5 \leq \frac{1}{2}\left(\delta_5\|p_c(t)\|^2_{H^{\frac{1}{2}}(\mathscr{C})} + \frac{1}{\delta_5}\|\tilde{q}'_W(t)\|^2_{H^{-\frac{1}{2}}(\mathscr{C})}\right).$$

We shall bound T_6 after an integration over $]0, t[$:

$$\left|\int_0^t \frac{d}{ds}\langle\tilde{q}_W(s), p_c(s)\rangle_{\mathscr{C}}\, ds\right| \leq |\langle\tilde{q}_W(t), p_c(t)\rangle_{\mathscr{C}}| + |\langle\tilde{q}_W(0), p_c(0)\rangle_{\mathscr{C}}|$$
$$\leq \|\tilde{q}_W(t)\|_{H^{-\frac{1}{2}}(\mathscr{C})}\|p_c(t)\|_{H^{\frac{1}{2}}(\mathscr{C})} + \|\tilde{q}_W(0)\|_{H^{-\frac{1}{2}}(\mathscr{C})}\|p_c(0)\|_{H^{\frac{1}{2}}(\mathscr{C})}.$$

For the first term, we use formula (29) at time t. Thus

$$\|\tilde{q}_W(t)\|_{H^{-\frac{1}{2}}(\mathscr{C})}\|p_c(t)\|_{H^{\frac{1}{2}}(\mathscr{C})}$$
$$\leq \frac{1}{2}\left(\delta_6\|\boldsymbol{K}^{-\frac{1}{2}}\boldsymbol{z}(t)\|^2_{L^2(\Omega\setminus\mathscr{C})} + \frac{1}{\delta_6\beta^2}\|\boldsymbol{K}^{-\frac{1}{2}}\|^2_{L^\infty(\Omega\setminus\mathscr{C})}\|\tilde{q}_W(t)\|^2_{H^{-\frac{1}{2}}(\mathscr{C})}\right)$$
$$+ \frac{1}{2}\left(\frac{\delta_7}{\beta}\big(\|p(t)\|^2_{L^2(\Omega)} + |\rho_{f,r}g\eta|^2_{H^1(\Omega\setminus\mathscr{C})}\big) + \frac{1}{\delta_7\beta}\|\tilde{q}_W(t)\|^2_{H^{-\frac{1}{2}}(\mathscr{C})}\right).$$

Finally, we substitute the bounds for T_i, $1 \leq i \leq 5$, into (41), we integrate on time over $]0, t[$, and we substitute the bound for T_6. A suitable choice of constants δ_i, $i = 3, 4, 6$ allows to absorb the terms involving $\boldsymbol{u}'$, p', and $\boldsymbol{z}$ into the left-hand side of (41). This yields (40), considering that all other terms are either data or terms that have been bounded by Theorem 3. □

The next corollary complements the bounds on $\boldsymbol{z}$ and $\boldsymbol{\zeta}$ of Theorem 4.

Corollary 2 *Under the assumptions of Theorem 4, we have*

$$\|\nabla \cdot \boldsymbol{z}\|_{L^2((\Omega\setminus\mathscr{C})\times]0,T[)} \leq \mu_f \Bigg[\left(\frac{1}{M} + c_f\varphi_0\right)\|p'\|_{L^2((\Omega\setminus\mathscr{C})\times]0,T[)} + \alpha\|\nabla \cdot \boldsymbol{u}'\|_{L^2((\Omega\setminus\mathscr{C})\times]0,T[)} + \|\tilde{q}\|_{L^2((\Omega\setminus\mathscr{C})\times]0,T[)}\Bigg], \tag{42}$$

$$\|\overline{\nabla} \cdot (w^{\frac{3}{2}}\boldsymbol{\zeta})\|_{L^2(0,T;H^{-\frac{1}{2}}(\mathscr{C}))} \leq 12\mu_f \Bigg[C\|\boldsymbol{u}'\|_{L^2(0,T;\boldsymbol{Z})} + \frac{C}{\mu_f}\|\boldsymbol{z}\|_{L^2(0,T;\boldsymbol{Z})} + \|\tilde{q}_W\|_{L^2(0,T;H^{-\frac{1}{2}}(\mathscr{C}))}\Bigg], \tag{43}$$

where C is the constant of the trace inequality

$$\forall \boldsymbol{q} \in \boldsymbol{Z}, \quad \|[\boldsymbol{q}]_{\mathscr{C}} \cdot \boldsymbol{n}^+\|_{H^{-\frac{1}{2}}(\mathscr{C})} \leq C\|\boldsymbol{q}\|_{\boldsymbol{Z}}. \tag{44}$$

Proof Formula (42) follows directly from (P2.2), and (43) follows from (P2.3) and (44). □

Remark 2 The bounds (44) and (42) lead to an immediate a priori estimate for the leakage term:

$$\begin{aligned}\|\tilde{q}_L\|_{L^2(0,T;H^{-\frac{1}{2}}(\mathscr{C}))} &= \frac{1}{\mu_f}\|[\boldsymbol{z}]_{\mathscr{C}} \cdot \boldsymbol{n}^+\|_{L^2(0,T;H^{-\frac{1}{2}}(\mathscr{C}))} \\ &\leq C\Bigg[\|\boldsymbol{z}\|_{L^2((\Omega\setminus\mathscr{C})\times]0,T[)} + \left(\frac{1}{M} + c_f\varphi_0\right)\|p'\|_{L^2((\Omega\setminus\mathscr{C})\times]0,T[)} \\ &\qquad + \alpha\|\nabla \cdot \boldsymbol{u}'\|_{L^2((\Omega\setminus\mathscr{C})\times]0,T[)} + \|\tilde{q}\|_{L^2((\Omega\setminus\mathscr{C})\times]0,T[)}\Bigg].\end{aligned}$$

□

The Assumption 2 on the fracture width's growth in time is restrictive because it does not allow the fracture to propagate in time, i.e., the fracture must already be present and its width can grow, but it cannot open at points where it is closed. This is only a theoretical sufficient condition for the existence theorem and we do not know if it is necessary. Of course, these two parts on a priori estimates might have been by-passed and replaced by the assumption that Problem 1 has a sufficiently smooth solution, since this implies existence of a solution of the mixed problem, but deriving a priori estimates always gives worthwhile information.

4.3 Existence and Uniqueness of Solutions

The above estimates show that if the problem (P2.1)–(P2.5) and (16) has a solution with the regularity stated in Theorem 4, then this solution is unique.

Regarding existence, rather than directly constructing a solution for the mixed formulation, let us use known existence results for Problem 1 and the equivalence Theorem 2, even though the assumptions may not be optimal. For instance, if $\boldsymbol{f} \in H^2(0,T;L^2(\Omega\setminus\mathscr{C})^d)$, $\tilde{q} \in L^2(\Omega\times]0,T[)$, $\tilde{q}_W \in H^1(0,T;L^2(\mathscr{C}))$, $p_0 \in Q$, and w satisfies Hypotheses 1 and 2, then the solution of Problem 1 satisfies $p \in H^1(0,T;L^2(\Omega\setminus\mathscr{C})) \cap L^\infty(0,T;Q)$, $\boldsymbol{u} \in H^1(0,T;\mathbb{V})$, $\tilde{q}_L \in L^2(0,T;H^1_w(\mathscr{C})')$ and is unique in these spaces, see [17]. Once this is known, additional regularity can be derived from the equations of Problem 1. In particular, with the definition (18) of z, (P1.3) implies that z is in $L^2(0,T;\boldsymbol{Z})$. In view of (P1.6), this means that $\tilde{q}_L$ belongs to $L^2(0,T;H^{-\frac{1}{2}}(\mathscr{C}))$. Then (P1.4), (P1.7), and the definition (19) of $\boldsymbol{\zeta}$ imply that $\boldsymbol{\zeta}$ is in $L^2(0,T;\boldsymbol{Z}_\mathscr{C})$. Hence we are in the setting of Theorem 2, which yields existence of a solution of (P2.1)–(P2.5), (16), (13), and (23) with the above regularity. This is summarized in the next theorem.

Theorem 5 *Let the data $\boldsymbol{f}$, $\tilde{q}$, $\tilde{q}_W$ and p_0 be given in $H^2(0,T;L^2(\Omega\setminus\mathscr{C})^d)$, $L^2(\Omega\times]0,T[)$, $H^1(0,T;L^2(\mathscr{C}))$, and Q, respectively, and let w satisfy Hypotheses 1 and 2. Then the mixed problem* (P2.1)–(P2.5)*,* (16)*,* (13)*, and* (23) *has one and only one solution p, $\boldsymbol{u}$, $\tilde{q}_L$, z, and $\boldsymbol{\zeta}$, respectively, in $H^1(0,T;L^2(\Omega\setminus\mathscr{C}))\cap L^\infty(0,T;Q)$, $H^1(0,T;\mathbb{V})$, $L^2(0,T;H^{-\frac{1}{2}}(\mathscr{C}))$, $L^2(0,T;\boldsymbol{Z})$, and $L^2(0,T;\boldsymbol{Z}_\mathscr{C})$.*

5 Discretization

In this section, we study a space–time discretization of the linearized mixed problem (P2.1)–(P2.5) and (16), with a backward Euler scheme in time and finite elements in space that are conforming for the displacement $\boldsymbol{u}$ and velocity variables z and $\boldsymbol{\zeta}$. In order to avoid handling curved elements or analyzing the approximation of curved surfaces that raises additional technicalities, we assume that both $\partial\Omega$ and the fracture $\mathscr{C}$ are polygonal or polyhedral surfaces.

5.1 General Discrete Spaces

Let $\mathscr{T}_h$ be a regular family of conforming meshes of $\overline{\Omega}$, made of triangles or convex quadrilaterals in 2D and tetrahedra or convex hexahedra in 3D. To simplify, we assume that $\mathscr{T}_h$ meshes Ω^+ and Ω^-, i.e., $\mathscr{C}$ does not cross the elements of $\mathscr{T}_h$. Let $N \geq 1$ be a fixed integer, $\Delta t = T/N$ the time step, and $t_i = i\Delta t$, $0 \leq i \leq N$, the discrete time points.

In each element, if the element is a simplex, the functions are approximated by polynomials P_k of total degree k, and if the element is a quadrilateral or hexahedron, the functions are approximated by images of tensor product polynomials Q_k of degree k in each variable. The displacement, velocity and pressure finite element spaces on any physical element E are defined, respectively, via the vector transformation

$$\boldsymbol{v} \leftrightarrow \hat{\boldsymbol{v}} : \boldsymbol{v} = \hat{\boldsymbol{v}} \circ F_E^{-1},$$

via the Piola transformation

$$\boldsymbol{z} \leftrightarrow \hat{\boldsymbol{z}} : \boldsymbol{z} = \frac{1}{J_E} \mathbb{DF}_E \hat{\boldsymbol{z}} \circ F_E^{-1},$$

and via the scalar transformation

$$w \leftrightarrow \hat{w} : w = \hat{w} \circ F_E^{-1},$$

where F_E denotes a mapping from the reference element $\hat{E}$, unit square or cube according to the dimension, to the physical element E, $\mathbb{DF}_E$ is the Jacobian of F_E, and J_E is its determinant. The advantage of the Piola transformation is that it preserves the divergence and the normal components of the velocity vectors on the sides or faces [16, Ch. III, 4.4] in the following sense:

$$(\nabla \cdot \boldsymbol{v}, w)_E = (\hat{\nabla} \cdot \hat{\boldsymbol{v}}, \hat{w})_{\hat{E}} \quad \text{and} \quad (\boldsymbol{v} \cdot \boldsymbol{n}_e, w)_e = (\hat{\boldsymbol{v}} \cdot \hat{\boldsymbol{n}}_{\hat{e}}, \hat{w})_{\hat{e}}.$$

This is used in constructing the $H(\text{div}; \Omega)$-conforming velocity space $\boldsymbol{Z}_h$ defined below. On $\mathscr{T}_h$, the finite element spaces $\mathbb{V}_h$ for the displacement $\boldsymbol{u}_h$, $\boldsymbol{Z}_h$ for the velocity $\boldsymbol{z}_h$, and Q_h for the pressure p_h are given by

$$\begin{aligned}
\mathbb{V}_h &= \left\{ \boldsymbol{v} \in \mathbb{V} \mid \boldsymbol{v}|_E = \hat{\boldsymbol{v}} \circ F_E^{-1},\ \hat{\boldsymbol{v}} \in \hat{\mathbb{V}}(\hat{E}),\ \forall E \in \mathscr{T}_h \right\}, \\
\boldsymbol{Z}_h &= \left\{ \boldsymbol{z} \in \boldsymbol{Z} \mid \boldsymbol{z}|_E = \frac{1}{J_E} \mathbb{DF}_E \hat{\boldsymbol{z}} \circ F_E^{-1},\ \hat{\boldsymbol{z}} \in \hat{\boldsymbol{Z}}(\hat{E}),\ \forall E \in \mathscr{T}_h \right\}, \\
Q_h &= \left\{ q \in L^2(\Omega) \mid q|_E = \hat{q} \circ F_E^{-1},\ \hat{q} \in \hat{Q}(\hat{E}),\ \forall E \in \mathscr{T}_h \right\},
\end{aligned}$$

where $\hat{\mathbb{V}}(\hat{E})$, $\hat{\boldsymbol{Z}}(\hat{E})$ and $\hat{Q}(\hat{E})$ are suitable finite element spaces on the reference element $\hat{E}$. In particular, we suppose that $\hat{\boldsymbol{Z}}(\hat{E})$ and $\hat{Q}(\hat{E})$ are compatible pairs such as the Raviart-Thomas pairs of elements on simplices or enhanced BDM pairs of elements on quadrilaterals and hexahedra. The enhanced BDM pairs are used on quadrilaterals and hexahedra as described in Sect. 5.3.

By definition, the functions of $\mathbb{V}_h$ and $\boldsymbol{Z}_h$ are conforming in $\mathbb{V}$ and $\boldsymbol{Z}$ respectively. Moreover, we assume that the conformity holds also on the boundary of $\mathscr{C}$, i.e., the functions of $\mathbb{V}_h$ as well as the normal components of functions of $\boldsymbol{Z}_h$ have no jump on $\partial\mathscr{C}$.

5.2 Discretization in the Fracture

Let $\mathscr{C}_h$ denote the trace of $\mathscr{T}_h$ on $\mathscr{C}$. Since $\mathscr{C}$ is assumed to be polygonal or polyhedral, we can map each line segment or plane face of $\mathscr{C}$ onto a segment in the x_1 line (when $d = 2$) or a polygon in the $x_1 - x_2$ plane (when $d = 3$) by a rigid-body motion that preserves both surface gradient and divergence, maps the normal $\boldsymbol{n}^+$ into a unit vector along x_3, for example, $-\boldsymbol{e}_3$, and whose Jacobian is one. After this change in variable, all operations on this line segment or plane face can be treated as the same operations on the x_1 axis or $x_1 - x_2$ plane. To simplify, we do not use a particular notation for this change in variable, and work as if the line segments or plane faces of $\mathscr{C}$ lie on the x_1 line or $x_1 - x_2$ plane. Let $\mathscr{S}_i$, $1 \le i \le I$, denote the line segments or plane faces of $\mathscr{C}$; to simplify, we drop the index i. Again, to simplify the analysis, we take the trace of $\mathscr{T}_h$ on $\mathscr{S}$, say $\mathscr{T}_{\mathscr{S},h}$ as partition of $\mathscr{S}$. Let e denote a generic element of $\mathscr{T}_{\mathscr{S},h}$, with reference element $\hat{e}$, and let the scalar and Piola transforms be defined by the same formula as above, but with respect to e instead of E. Then we define the finite element spaces on $\mathscr{C}$ by:

$$\begin{aligned}\boldsymbol{Z}_{\mathscr{C},h} &= \left\{\boldsymbol{\mu} \in \boldsymbol{Z}_{\mathscr{C}} \mid \boldsymbol{\mu}|_{\mathscr{S}_i} \in \boldsymbol{Z}_{\mathscr{S}_i,h},\ 1 \le i \le I\right\},\\ \Theta_{\mathscr{C},h} &= \left\{q \in L^2(\mathscr{C}) \mid q|_{\mathscr{S}_i} \in \Theta_{\mathscr{S}_i,h},\ 1 \le i \le I\right\},\end{aligned}$$

with

$$\begin{aligned}\boldsymbol{Z}_{\mathscr{S},h} &= \left\{\boldsymbol{\mu} \in \boldsymbol{Z}_{\mathscr{C}} \mid \boldsymbol{\mu}|_e \leftrightarrow \hat{\boldsymbol{\mu}},\ \hat{\boldsymbol{\mu}} \in \hat{\boldsymbol{Z}}_{\mathscr{C}}(\hat{e}),\ \forall e \in \mathscr{T}_{\mathscr{S},h}\right\},\\ \Theta_{\mathscr{S},h} &= \left\{q \in L^2(\mathscr{C}) \mid q|_e \leftrightarrow \hat{q},\ \hat{q} \in \hat{\Theta}_{\mathscr{C}}(\hat{e}),\ \forall e \in \mathscr{T}_{\mathscr{S},h}\right\},\end{aligned}$$

where $\hat{\boldsymbol{Z}}_{\mathscr{C}}(\hat{e})$ and $\hat{\Theta}_{\mathscr{C}}(\hat{e})$ are finite element spaces on the reference element $\hat{e}$. Again, we assume that they are compatible pairs like the Raviart-Thomas pairs on triangles or enhanced BDM pairs on quadrilaterals. This implies that the functions $\boldsymbol{q}_{c,h}$ of $\boldsymbol{Z}_{\mathscr{S},h}$ belong globally to $H(\mathrm{div}; \mathscr{S}) \cap L^\infty(\mathscr{S})^{d-1}$. By expanding $\overline{\nabla} \cdot (w^{\frac{3}{2}}\boldsymbol{q}_{c,h})$ and using the assumption (17) on w, we easily derive that $\overline{\nabla} \cdot (w^{\frac{3}{2}}\boldsymbol{q}_{c,h})$ belongs to $L^2(\mathscr{S})$. This allows to take the discrete pressure $p_{c,h}$ in $L^2(\mathscr{C})$ instead of $H^{\frac{1}{2}}(\mathscr{C})$.

5.3 Elements on Convex Quadrilaterals and Hexahedra

In the case of convex quadrilaterals, $\hat{E}$ is the unit square with vertices $\hat{\mathsf{r}}_1 = (0, 0)^T$, $\hat{\mathsf{r}}_2 = (1, 0)^T$, $\hat{\mathsf{r}}_3 = (1, 1)^T$, and $\hat{\mathsf{r}}_4 = (0, 1)^T$. Denote by r_i, $1 \le i \le 4$, the corresponding vertices of E. In this case, F_E is the bilinear mapping given as

$$F_E(\hat{x}, \hat{y}) = \mathsf{r}_1(1 - \hat{x})(1 - \hat{y}) + \mathsf{r}_2\hat{x}(1 - \hat{y}) + \mathsf{r}_3\hat{x}\hat{y} + \mathsf{r}_4(1 - \hat{x})\hat{y};$$

the space for the displacement is

$$\hat{\mathbb{V}}(\hat{E}) = Q_1(\hat{E})^2,$$

and the space for the flow is the lowest order BDM$_1$ space [8]

$$\hat{\boldsymbol{Z}}(\hat{E}) = P_1(\hat{E})^2 + r\ \mathbf{curl}(\hat{x}^2\hat{y}) + s\ \mathbf{curl}(\hat{x}\hat{y}^2)\ ,\ \hat{Q}(\hat{E}) = P_0(\hat{E}),$$

where r and s are real constants.

In the case of hexahedra, $\hat{E}$ is the unit cube but the element E can have non-planar faces. The vertices of $\hat{E}$ are $\hat{\mathsf{r}}_1 = (0, 0, 0)^T$, $\hat{\mathsf{r}}_2 = (1, 0, 0)^T$, $\hat{\mathsf{r}}_3 = (1, 1, 0)^T$, $\hat{\mathsf{r}}_4 = (0, 1, 0)^T$, $\hat{\mathsf{r}}_5 = (0, 0, 1)^T$, $\hat{\mathsf{r}}_6 = (1, 0, 1)^T$, $\hat{\mathsf{r}}_7 = (1, 1, 1)^T$, and $\hat{\mathsf{r}}_8 = (0, 1, 1)^T$. Denote by $\mathsf{r}_i = (x_i, y_i, z_i)^T$, $1 \le i \le 8$, the eight corresponding vertices of E. In this case F_E is a trilinear mapping given by

$$\begin{aligned} F_E(\hat{x},\hat{y}, \hat{z}) &= \mathsf{r}_1(1 - \hat{x})(1 - \hat{y})(1 - \hat{z}) + \mathsf{r}_2\hat{x}(1 - \hat{y})(1 - \hat{z}) + \mathsf{r}_3\hat{x}\hat{y}(1 - \hat{z}) \\ &\quad + \mathsf{r}_4(1 - \hat{x})\hat{y}(1 - \hat{z}) + \mathsf{r}_5(1 - \hat{x})(1 - \hat{y})\hat{z} + \mathsf{r}_6\hat{x}(1 - \hat{y})\hat{z} + \mathsf{r}_7\hat{x}\hat{y}\hat{z} + \mathsf{r}_8(1 - \hat{x})\hat{y}\hat{z}, \end{aligned}$$

the space for the displacement is defined by

$$\hat{\mathbb{V}}(\hat{E}) = Q_1(\hat{E})^3,$$

the space for the flow is an enhanced BDDF$_1$ space [19]:

$$\begin{aligned} \hat{\boldsymbol{Z}}(\hat{E}) &= \mathrm{BDDF}_1(\hat{E}) + s_2\,\mathbf{curl}(0, 0, \hat{x}^2\hat{z})^T + s_3\,\mathbf{curl}(0, 0, \hat{x}^2\hat{y}\hat{z})^T + t_2\,\mathbf{curl}(\hat{x}\hat{y}^2, 0, 0)^T \\ &\quad + t_3\,\mathbf{curl}(\hat{x}\hat{y}^2\hat{z}, 0, 0)^T + w_2\,\mathbf{curl}(0, \hat{y}\hat{z}^2, 0)^T + w_3\,\mathbf{curl}(0, \hat{x}\hat{y}\hat{z}^2, 0)^T, \\ \hat{Q}(\hat{E}) &= P_0(\hat{E}), \end{aligned}$$

where the BDDF$_1(\hat{E})$ space is defined as [7]:

$$\begin{aligned} \mathrm{BDDF}_1(\hat{E}) = P_1(\hat{E})^3 &+ s_0\,\mathbf{curl}(0, 0, \hat{x}\hat{y}\hat{z})^T + s_1\,\mathbf{curl}(0, 0, \hat{x}\hat{y}^2)^T + t_0\,\mathbf{curl}(\hat{x}\hat{y}\hat{z}, 0, 0)^T \\ &+ t_1\,\mathbf{curl}(\hat{y}\hat{z}^2, 0, 0)^T + w_0\,\mathbf{curl}(0, \hat{x}\hat{y}\hat{z}, 0)^T + w_1\,\mathbf{curl}(0, \hat{x}^2\hat{z}, 0)^T. \end{aligned}$$

In the above equations, s_i, t_i, w_i, $0 \le i \le 3$, are real constants. In all cases the degrees of freedom (DOF) for the displacements are chosen as Lagrangian nodal point values.

The velocity DOF are chosen to be the normal components at the d vertices on each face. The dimension of the space is dn_v, where $d = 2, 3$ is the dimension and n_v is the number of vertices in E. Note that, although the original BDDF$_1$ spaces have only three DOF on square faces, these spaces have been enhanced in [19] to have four DOF on square faces. This special choice is needed in the reduction to a cell-centered pressure stencil in a pure Darcy flow problem as described later in this section.

5.4 Fully Discrete Equations

The assumptions on the data are:

$$\boldsymbol{f} \in H^1(0,T; L^2(\Omega \setminus \mathscr{C})^d), \quad \tilde{q} \in \mathscr{C}^0([0,T]; L^2(\Omega \setminus \mathscr{C})),$$
$$\tilde{q}_W \in H^1(0,T; H^{-\frac{1}{2}}(\mathscr{C})), \quad p(0) = p_0 \in Q$$

with $p_c(0)$ the trace of p_0 on $\mathscr{C}$, and in addition to Hypothesis 1, w is continuous in time.

For each n and for almost every $\boldsymbol{x} \in \Omega^+ \cup \Omega^-$ or Ω, we set

$$\boldsymbol{f}^n(\boldsymbol{x}) = \boldsymbol{f}(\boldsymbol{x}, t_n), \quad \tilde{q}^n(\boldsymbol{x}) = \tilde{q}(\boldsymbol{x}, t_n),$$

and for almost every $s \in \mathscr{C}$

$$w^n(s) = w(s, t_n), \quad \tilde{q}_W^n(s) = \tilde{q}_W(s, t_n).$$

To simplify, we denote the first backward difference in time of any function v (continuous in time) as follows,

$$\delta v^n = v^n - v^{n-1}.$$

We propose the following fully discrete implicit coupled mixed scheme. It is assumed that the finite element functions are sufficiently smooth to give meaning to all integrals below.

Problem 3 At time $t = 0$, let $p_h^0 = r_h(p_0)$, where r_h is the local L^2 projection on each element E of $\mathscr{T}_h$, with values in Q_h. By assumption $p_0 \in Q$ and therefore $p_{c,0}$, its trace on $\mathscr{C}$, belongs to $H_w^1(\mathscr{C}) \subset H^{\frac{1}{2}}(\mathscr{C})$. We take $p_{c,h}^0 = r_{\mathscr{C},h}(p_{c,0})$, where $r_{\mathscr{C},h}$ is the local L^2 projection on each element e of $\mathscr{C}_h$, with values in $\Theta_{\mathscr{C},h}$.

Once p_h^0 and $p_{c,h}^0$ are known, $\boldsymbol{u}(p_h^0)$ is approximated by discretizing the elasticity equation (P2.1) in $\Omega \setminus \mathscr{C}$: Find $\boldsymbol{u}_h^0 \in \mathbb{V}_h$ solution of

$$\forall \boldsymbol{v}_h \in \mathbb{V}_h, \quad 2G\big(\boldsymbol{\varepsilon}(\boldsymbol{u}_h^0), \boldsymbol{\varepsilon}(\boldsymbol{v}_h)\big) + \lambda\big(\nabla \cdot \boldsymbol{u}_h^0, \nabla \cdot \boldsymbol{v}_h\big) \tag{P3.1}$$
$$= \alpha\big(p_h^0, \nabla \cdot \boldsymbol{v}_h\big) - \big(p_{c,h}^0, [\boldsymbol{v}_h]_{\mathscr{C}} \cdot \boldsymbol{n}^+\big)_{\mathscr{C}} + \big(\boldsymbol{f}^0, \boldsymbol{v}_h\big).$$

Similarly, z_h^0 and $\boldsymbol{\zeta}_h^0$ are approximated by discretizing respectively (P2.4) and (P2.5):

$$\forall \boldsymbol{q}_h \in \boldsymbol{Z}_h, \quad \big(\boldsymbol{K}^{-1} z_h^0, \boldsymbol{q}_h\big) = (p_h^0, \nabla \cdot \boldsymbol{q}_h) - \big(p_{c,h}^0, [\boldsymbol{q}_h]_{\mathscr{C}} \cdot \boldsymbol{n}^+\big)_{\mathscr{C}} + \big(\nabla(\rho_{f,r} g \eta), \boldsymbol{q}_h\big), \tag{P3.2}$$

$$\forall \boldsymbol{q}_{c,h} \in \boldsymbol{Z}_{\mathscr{C},h}, \quad \big(\boldsymbol{\zeta}_h^0, \boldsymbol{q}_{c,h}\big)_{\mathscr{C}} = \big(p_{c,h}^0, \overline{\nabla} \cdot \big((w^0)^{\frac{3}{2}} \boldsymbol{q}_{c,h}\big)\big)_{\mathscr{C}} + \big((w^0)^{\frac{3}{2}} \overline{\nabla}(\rho_{f,r} g \eta), \boldsymbol{q}_{c,h}\big)_{\mathscr{C}}. \tag{P3.3}$$

For any n, $1 \leq n \leq N$, $\boldsymbol{u}_h^n$, p_h^n, $p_{c,h}^n$, z_h^n, and $\boldsymbol{\zeta}_h^n$ are approximated by discretizing (P2.1)–(P2.5): Knowing $\boldsymbol{u}_h^{n-1}, p_h^{n-1}$, find $\boldsymbol{u}_h^n \in \mathbb{V}_h, p_h^n \in Q_h, p_{c,h}^n \in \Theta_{\mathscr{C},h}, z_h^n \in \boldsymbol{Z}_h$, and $\boldsymbol{\zeta}_h^n \in \boldsymbol{Z}_{\mathscr{C},h}$ solutions of

$$\forall \boldsymbol{v}_h \in \mathbb{V}_h, \quad 2G\big(\boldsymbol{\varepsilon}(\boldsymbol{u}_h^n), \boldsymbol{\varepsilon}(\boldsymbol{v}_h)\big) + \lambda\big(\nabla \cdot \boldsymbol{u}_h^n, \nabla \cdot \boldsymbol{v}_h\big) - \alpha\big(p_h^n, \nabla \cdot \boldsymbol{v}_h\big) \\ + \big(p_{c,h}^n, [\boldsymbol{v}_h]_{\mathscr{C}} \cdot \boldsymbol{n}^+\big)_{\mathscr{C}} = \big(\boldsymbol{f}^n, \boldsymbol{v}_h\big), \tag{P3.4}$$

$$\forall \theta_h \in Q_h, \quad \Big(\big(\frac{1}{M} + c_f \varphi_0\big) \frac{1}{\Delta t} \delta p_h^n + \frac{\alpha}{\Delta t} \nabla \cdot \delta \boldsymbol{u}_h^n, \theta_h\Big) + \frac{1}{\mu_f}\big(\nabla \cdot z_h^n, \theta_h\big) = \big(\tilde{q}^n, \theta_h\big), \tag{P3.5}$$

$$\forall \theta_{c,h} \in \Theta_{\mathscr{C},h}, \quad -\frac{1}{\Delta t}\big(\delta([\boldsymbol{u}_h^n]_{\mathscr{C}}) \cdot \boldsymbol{n}^+, \theta_{c,h}\big)_{\mathscr{C}} + \frac{1}{12\mu_f}\big(\overline{\nabla} \cdot ((w^n)^{\frac{3}{2}} \boldsymbol{\zeta}_h^n), \theta_{c,h}\big)_{\mathscr{C}} \\ - \frac{1}{\mu_f}\big([z_h^n]_{\mathscr{C}} \cdot \boldsymbol{n}^+, \theta_{c,h}\big)_{\mathscr{C}} = \langle \tilde{q}_W^n, \theta_{c,h} \rangle_{\mathscr{C}}, \tag{P3.6}$$

$$\forall \boldsymbol{q}_h \in \boldsymbol{Z}_h, \quad \big(\boldsymbol{K}^{-1} z_h^n, \boldsymbol{q}_h\big) = (p_h^n, \nabla \cdot \boldsymbol{q}_h) - \big(p_{c,h}^n, [\boldsymbol{q}_h]_{\mathscr{C}} \cdot \boldsymbol{n}^+\big)_{\mathscr{C}} + \big(\nabla(\rho_{f,r} g \eta), \boldsymbol{q}_h\big), \tag{P3.7}$$

$$\forall \boldsymbol{q}_{c,h} \in \boldsymbol{Z}_{\mathscr{C},h}, \quad \big(\boldsymbol{\zeta}_h^n, \boldsymbol{q}_{c,h}\big)_{\mathscr{C}} = \big(p_{c,h}^n, \overline{\nabla} \cdot ((w^n)^{\frac{3}{2}} \boldsymbol{q}_{c,h})\big)_{\mathscr{C}} + \big((w^n)^{\frac{3}{2}} \overline{\nabla}(\rho_{f,r} g \eta), \boldsymbol{q}_{c,h}\big)_{\mathscr{C}}. \tag{P3.8}$$

Problem 3 is a square system of linear equations in finite dimension. Therefore, to show existence of a solution, it suffices to prove that, at each time step, if all data are zero (including the values at the preceding step) then the only solution is the zero solution. Existence and uniqueness of p_h^n, $\boldsymbol{u}_h^n$, z_h^n, and $\boldsymbol{\zeta}_h^n$ follow immediately from the following stability equality, obtained by testing (P3.5) with p_h^n, (P3.6) with $p_{c,h}^n$, (P3.7) with z_h^n, (P3.8) with $\boldsymbol{\zeta}_h^n$, (P3.4) with $\delta \boldsymbol{u}_h^n$, multiplying everything by Δt, and combining the resulting equations:

$$\begin{aligned}\frac{1}{2}\left(\frac{1}{M}+c_f\varphi_0\right)&\Big(\delta\big(\|p_h^n\|^2_{L^2(\Omega)}\big)+\|\delta\, p_h^n\|^2_{L^2(\Omega)}\Big)\\ &+G\Big(\delta\big(\|\boldsymbol{\varepsilon}(\boldsymbol{u}_h^n)\|^2_{L^2(\Omega\setminus\mathscr{C})}\big)+\|\delta\,\boldsymbol{\varepsilon}(\boldsymbol{u}_h^n)\|^2_{L^2(\Omega\setminus\mathscr{C})}\Big)\\ &+\frac{\lambda}{2}\Big(\delta\big(\|\nabla\cdot\boldsymbol{u}_h^n\|^2_{L^2(\Omega\setminus\mathscr{C})}\big)+\|\delta(\nabla\cdot\boldsymbol{u}_h^n)\|^2_{L^2(\Omega\setminus\mathscr{C})}\Big)\\ &+\frac{\Delta t}{\mu_f}\|\boldsymbol{K}^{-\frac{1}{2}}\boldsymbol{z}_h^n\|^2_{L^2(\Omega\setminus\mathscr{C})}+\frac{\Delta t}{12\mu_f}\|\boldsymbol{\zeta}_h^n\|^2_{L^2(\mathscr{C})}\\ =\Delta t(\tilde{q}^n,p_h^n)_\Omega+(\boldsymbol{f}^n,\delta\,\boldsymbol{u}_h^n)&+\Delta t\langle\tilde{q}_W^n,p_{c,h}^n\rangle_{\mathscr{C}}+\frac{\Delta t}{\mu_f}(\nabla(\rho_{f,r}g\eta),\boldsymbol{z}_h^n)\\ &+\frac{\Delta t}{12\mu_f}\big((w^n)^{\frac{3}{2}}\overline{\nabla}(\rho_{f,r}g\eta),\boldsymbol{\zeta}_h^n\big)_{\mathscr{C}}.\end{aligned}$$

From here, existence and uniqueness of $p_{c,h}^n$ will be a consequence of the discrete inf-sup condition established in the next section.

5.5 Discrete Inf-sup Condition for $p_{c,h}$

As in the exact problem, we need an inf-sup condition to control the discrete surface pressure $p_{c,h}$ on $\mathscr{C}$. However, the argument used in deriving Lemma 1 does not carry over to the discrete case because the Raviart-Thomas interpolant R_h, which is the most obvious candidate for discretizing $\boldsymbol{q}$, is not defined in $H(\text{div};\Omega)$. We shall use instead an interior argument that creates a smoother function. In addition, we suppose the following compatibility condition on the finite element spaces.

Hypothesis 3 There exists an approximation operator $R_h\in\mathscr{L}(\boldsymbol{Z}\cap H^s(\Omega^+\cup\Omega^-)^d;\boldsymbol{Z}_h)$ for $s>0$, such that for all $\boldsymbol{q}\in\boldsymbol{Z}\cap H^s(\Omega^+\cup\Omega^-)^d$,

$$\begin{aligned}\forall E\subset\Omega^\star,\star=+,\ \forall\theta_h\in Q_h,\quad &\big(\theta_h,\nabla\cdot(\boldsymbol{q}-R_h(\boldsymbol{q}))\big)_E=0,\\ \forall e\subset\mathscr{C},\ \forall\theta_{c,h}\in\Theta_{\mathscr{C},h},\quad &\langle\theta_{c,h},[\boldsymbol{q}-R_h(\boldsymbol{q}))]_{\mathscr{C}}\cdot\boldsymbol{n}^+\rangle_e=0,\end{aligned}\tag{45}$$

and there exists a constant C independent of h such that for all element E of $\mathscr{T}_h$

$$\begin{aligned}\forall\boldsymbol{q}\in H^s(E)^d,\quad &\|\boldsymbol{q}-R_h(\boldsymbol{q})\|_{L^2(E)}\le C\,h^s|\boldsymbol{q}|_{H^s(E)},\\ \forall\boldsymbol{q}\in H(\text{div};E)\cap H^s(E)^d,\quad &\|\operatorname{div}(\boldsymbol{q}-R_h(\boldsymbol{q}))\|_{L^2(E)}\le\|\operatorname{div}\boldsymbol{q}\|_{L^2(E)}.\end{aligned}\tag{46}$$

These assumptions are satisfied by the Raviart-Thomas RT_k finite elements pairs of degree $k\ge 0$, i.e., $H(\text{div})$ velocity with incomplete degree $k+1$ and discontinuous pressure with degree k. They are also satisfied, for instance, by the enhanced BDM_1 elements pairs described in Sect. 5.3, associated with piecewise constant pressures.

In view of the first part of (45), the compatibility between Q_h and $\boldsymbol{Z}_h$ implies that $\nabla \cdot R_h(\boldsymbol{q})$ is the projection of $\nabla \cdot \boldsymbol{q}$ onto Q_h.

Lemma 2 *Under Hypothesis 3, there exists a constant $\beta_1^* > 0$, independent of h, such that*

$$\forall \theta_{c,h} \in \Theta_{\mathscr{C},h}, \quad \sup_{\boldsymbol{q}_h \in \boldsymbol{Z}_h} \frac{\big(\theta_{c,h}, [\boldsymbol{q}_h]_{\mathscr{C}} \cdot \boldsymbol{n}^+\big)_{\mathscr{C}}}{\|\boldsymbol{q}_h\|_{\mathrm{Z}}} \geq \beta_1^* \|\theta_{c,h}\|_{L^2(\mathscr{C})}. \tag{47}$$

Proof The idea is to construct an adequately smooth function $\boldsymbol{q}$ in $\boldsymbol{Z}$ whose normal jump on $\mathscr{C}$ coincides with $\theta_{c,h}$, and to which R_h can be applied. We proceed in two steps:

1. As we only consider the L^2 norm, we extend $\theta_{c,h}$ by zero to $\partial\Omega^+$ (without changing its notation) and we consider the unique solution $\varphi^+ \in H^1(\Omega^+)$ of the following Laplace equation with Neumann boundary conditions:

$$\begin{aligned} -\Delta \varphi^+ &= -\frac{1}{|\Omega^+|}\big(\theta_{c,h}, 1\big)_{\mathscr{C}} && \text{in } \Omega^+, \\ \frac{\partial}{\partial \boldsymbol{n}}\varphi^+ &= \theta_{c,h} && \text{on } \partial\Omega^+. \end{aligned}$$

Note that the interior and boundary data are compatible; therefore this problem has a unique solution φ^+ such that

$$\|\varphi^+\|_{H^1(\Omega^+)} \leq C\|\theta_{c,h}\|_{L^2(\mathscr{C})}, \tag{48}$$

with a constant C that depends only on Ω^+ and $\mathscr{C}$.
Then we choose $\boldsymbol{q}^+ = \nabla\varphi^+$ in Ω^+ and $\boldsymbol{q}^- = \boldsymbol{0}$ in Ω^-. By construction, $\boldsymbol{q}$ belongs to $H(\mathrm{div}; \Omega^+ \cup \Omega^-)$,

$$\nabla \cdot \boldsymbol{q} = \begin{cases} \dfrac{1}{|\Omega^+|}(\theta_{c,h}, 1)_{\mathscr{C}} & \text{in } \Omega^+, \\ 0 & \text{in } \Omega^-, \end{cases}$$

and

$$[\boldsymbol{q} \cdot \boldsymbol{n}^+]_{\mathscr{C}} = \theta_{c,h}, \quad [\boldsymbol{q} \cdot \boldsymbol{n}^+]_{\Gamma\setminus\mathscr{C}} = 0, \quad \boldsymbol{q} \cdot \boldsymbol{n}|_{\partial\Omega} = 0,$$

so that $\boldsymbol{q}$ belongs to $\boldsymbol{Z}$. Since the extended function $\theta_{c,h}$ belongs to $L^2(\partial\Omega^+)$, the regularity of the Laplace equation with Neumann boundary conditions imply that φ^+ is in $H^{\frac{3}{2}}(\Omega^+)$ (cf. [20]) with continuous dependence on $\theta_{c,h}$. Therefore $\boldsymbol{q} \in H^{\frac{1}{2}}(\Omega^+ \cup \Omega^-)^d$ and there exists a constant C depending only on Ω, Γ and $\mathscr{C}$ such that

$$\|\boldsymbol{q}\|_{H^{\frac{1}{2}}(\Omega^+\cup\Omega^-)} \leq C\|\theta_{c,h}\|_{L^2(\mathscr{C})}. \tag{49}$$

2. The regularity (49) of $\boldsymbol{q}$ and Hypothesis 3 allow to define $\boldsymbol{q}_h = R_h(\boldsymbol{q})$. As $\nabla \cdot \boldsymbol{q}$ is constant in each subdomain, and Q_h contains at least the constant functions, the first part of (45) implies trivially that $\nabla \cdot R_h(\boldsymbol{q}) = \nabla \cdot \boldsymbol{q}$. Thus

$$\|R_h(\boldsymbol{q})\|_Z \leq \|R_h(\boldsymbol{q}) - \boldsymbol{q}\|_Z + \|\boldsymbol{q}\|_Z \leq \|R_h(\boldsymbol{q}) - \boldsymbol{q}\|_{L^2(\Omega^+)} + \|\boldsymbol{q}\|_Z.$$

In view of the first part of (46) with $s = \frac{1}{2}$, and (49),

$$\|R_h(\boldsymbol{q}) - \boldsymbol{q}\|_{L^2(\Omega^+)} \leq C_1 h^{\frac{1}{2}} |\boldsymbol{q}|_{H^{\frac{1}{2}}(\Omega^+)} \leq C_2 C_1 h^{\frac{1}{2}} \|\theta_{c,h}\|_{L^2(\mathscr{C})},$$

where C_1 and C_2 are the constants of (46) and (49) respectively. Similarly,

$$\|\boldsymbol{q}\|_Z \leq \Big(\|\boldsymbol{q}\|^2_{L^2(\Omega)} + \frac{|\mathscr{C}|}{|\Omega^+|}\|\theta_{c,h}\|^2_{L^2(\mathscr{C})}\Big)^{\frac{1}{2}} \leq \Big(C_3^2 + \frac{|\mathscr{C}|}{|\Omega^+|}\Big)^{\frac{1}{2}} \|\theta_{c,h}\|_{L^2(\mathscr{C})},$$

where C_3 is the constant of (48). Combining these two inequalities, we have, on one hand,

$$\|R_h(\boldsymbol{q})\|_Z \leq C_4 \|\theta_{c,h}\|_{L^2(\mathscr{C})}. \tag{50}$$

On the other hand, the second part of (45) yields for all e in $\mathscr{C}$

$$\big(\theta_{c,h}, [R_h(\boldsymbol{q})]_{\mathscr{C}} \cdot \boldsymbol{n}^+\big)_e = \langle \theta_{c,h}, [\boldsymbol{q}]_{\mathscr{C}} \cdot \boldsymbol{n}^+ \rangle_e = \big(\theta_{c,h}, \theta_{c,h}\big)_e = \|\theta_{c,h}\|^2_{L^2(e)}. \tag{51}$$

Then (50) and (51) imply

$$\frac{1}{\|R_h(\boldsymbol{q})\|_Z}\big(\theta_{c,h}, [R_h(\boldsymbol{q})]_{\mathscr{C}} \cdot \boldsymbol{n}^+\big)_{\mathscr{C}} \geq \frac{1}{C_4}\|\theta_{c,h}\|_{L^2(\mathscr{C})},$$

whence (47) with $\beta_1^* = \frac{1}{C_4}$. □

Then we have the analogue of Corollary 1 with the same proof. The operator B is replaced by B_h defined on $\boldsymbol{Z}_h$ as

$$\forall \theta_{c,h} \in \Theta_{\mathscr{C},h}, \quad \langle B_h \boldsymbol{q}_h, \theta_{c,h} \rangle = \big(\theta_{c,h}, [\boldsymbol{q}_h]_{\mathscr{C}} \cdot \boldsymbol{n}^+\big)_{\mathscr{C}}.$$

The operator B_h is linear and since we are in finite dimension where all norms are equivalent, B_h is continuous on $\boldsymbol{Z}_h$ (albeit the continuity constant is not expected to be bounded as h tends to zero). The kernel of B_h in $\boldsymbol{Z}_h$ is:

$$\mathrm{Ker}(B_h) = \{\boldsymbol{q}_h \in \boldsymbol{Z}_h\,;\ \forall \theta_{c,h} \in \Theta_{\mathscr{C},h}, \big(\theta_{c,h}, [\boldsymbol{q}_h]_{\mathscr{C}} \cdot \boldsymbol{n}^+\big)_{\mathscr{C}} = 0\}.$$

Corollary 3 *Let $z_h \in \boldsymbol{Z}_h$ and $p_h \in Q_h$ be such that*

$$\forall \boldsymbol{q}_h \in \mathrm{Ker}(B_h), \quad (p_h, \nabla \cdot \boldsymbol{q}_h) - \big(\boldsymbol{K}^{-1} z_h, \boldsymbol{q}_h\big) + \big(\nabla(\rho_{f,r} g \eta), \boldsymbol{q}_h\big) = 0.$$

Then there exists a unique $p_{c,h} \in \Theta_{\mathscr{C},h}$ such that $p_{c,h}$, p_h, and $\boldsymbol{z}_h$ satisfy (P3.7)*, and*

$$\|p_{c,h}\|_{L^2(\mathscr{C})} \leq \frac{1}{\beta_1^*}\Big(\big(\|p_h\|^2_{L^2(\Omega)} + |\rho_{f,r} g \eta|^2_{H^1(\Omega\setminus\mathscr{C})}\big)^{\frac{1}{2}} + \|\boldsymbol{K}^{-1}\boldsymbol{z}_h\|_{L^2(\Omega\setminus\mathscr{C})}\Big),$$

where β_1^* is the constant of (47).

5.6 Error Estimates

Here we assume that the solution and data are sufficiently smooth in time and space, as needed. It is convenient to split the scheme's error into a time consistency error and a spatial discretization error.

5.6.1 Time Consistency Error

The time consistency error measures the difference between the divided difference in time and the time derivative. More precisely, for $\theta_h \in Q_h$ and $\theta_{c,h} \in \Theta_{\mathscr{C},h}$, we define

$$\begin{aligned} E_n(\theta_h, \theta_{c,h}) = & \left(\frac{1}{M} + c_f \varphi_0\right)\left(\frac{1}{\Delta t}\delta p(t_n) - p'(t_n), \theta_h\right) \\ & + \left(\alpha \nabla \cdot \left(\frac{1}{\Delta t}\delta \boldsymbol{u}(t_n) - \boldsymbol{u}'(t_n)\right), \theta_h\right) - \left(\left[\frac{1}{\Delta t}\delta \boldsymbol{u}(t_n) - \boldsymbol{u}'(t_n)\right]_{\mathscr{C}} \cdot \boldsymbol{n}^+, \theta_{c,h}\right)_{\mathscr{C}}. \end{aligned} \tag{52}$$

In view of the Taylor expansion valid for all functions v in $W^{2,1}(t_{n-1}, t_n)$:

$$\frac{1}{\Delta t}\delta v(t_n) = v'(t_n) - \frac{1}{\Delta t}\int_{t_{n-1}}^{t_n} (s - t_{n-1}) v''(s) ds,$$

the expression for $E_n(\theta_h, \theta_{c,h})$ becomes:

$$\begin{aligned} E_n(\theta_h, \theta_{c,h}) = & -\left(\frac{1}{M} + c_f \varphi_0\right)\frac{1}{\Delta t}\left(\int_{t_{n-1}}^{t_n} (s - t_{n-1}) p''(s) ds, \theta_h\right) \\ & - \frac{\alpha}{\Delta t}\left(\int_{t_{n-1}}^{t_n} (s - t_{n-1}) \nabla \cdot \boldsymbol{u}''(s) ds, \theta_h\right) \\ & + \frac{1}{\Delta t}\left(\int_{t_{n-1}}^{t_n} (s - t_{n-1})([\boldsymbol{u}''(s)]_{\mathscr{C}} \cdot \boldsymbol{n}^+) ds, \theta_{c,h}\right)_{\mathscr{C}}. \end{aligned}$$

5.6.2 Full Discretization Error

Formulas (52), (P3.5), and (P3.6) lead to the following error equality:

$$\begin{aligned}
&\frac{1}{\Delta t}\left(\left(\frac{1}{M}+c_f\varphi_0\right)\delta(p_h^n-p(t_n))+\alpha\nabla\cdot\delta(\boldsymbol{u}_h^n-\boldsymbol{u}(t_n)),\theta_h\right)\\
&\qquad-\frac{1}{\Delta t}\big([\delta(\boldsymbol{u}_h^n-\boldsymbol{u}(t_n))]_{\mathscr{C}}\cdot\boldsymbol{n}^+,\theta_{c,h}\big)_{\mathscr{C}}+\frac{1}{\mu_f}\big(\nabla\cdot(\boldsymbol{z}_h^n-\boldsymbol{z}(t_n)),\theta_h\big)\\
&\quad+\frac{1}{12\mu_f}\langle\overline{\nabla}\cdot((w^n)^{\frac{3}{2}}(\boldsymbol{\zeta}_h^n-\boldsymbol{\zeta}(t_n))),\theta_{c,h}\big\rangle_{\mathscr{C}}-\frac{1}{\mu_f}\langle[\boldsymbol{z}_h^n-\boldsymbol{z}(t^n)]_{\mathscr{C}}\cdot\boldsymbol{n}^+,\theta_{c,h}\rangle_{\mathscr{C}}\\
&\qquad\qquad=-E_n(\theta_h,\theta_{c,h}).
\end{aligned}$$

In addition to r_h, R_h, and $r_{\mathscr{C},h}$, we need an approximation operator from $\mathbb{V}$ into $\mathbb{V}_h$, such as a Scott–Zhang approximation operator [25] or a Lagrange interpolation operator [9], and an approximation operator $R_{\mathscr{C},h}$ from $\boldsymbol{Z}_{\mathscr{C}}\cap H^s(\mathscr{C})^{d-1}$ into $\boldsymbol{Z}_{\mathscr{C},h}$ for some $s>0$. Then by adding and subtracting $I_h(\boldsymbol{u})$, $r_h(p)$, $r_{\mathscr{C},h}(p_c)$, $R_h(\boldsymbol{z})$, and $R_{\mathscr{C},h}(\boldsymbol{\zeta})$, and by denoting the discretization errors and interpolation errors, respectively, by

$$\begin{gathered}
e_p^n=p_h^n-r_h(p(t_n)),\quad e_{c,p}^n=p_{c,h}^n-r_{\mathscr{C},h}(p_c(t_n)),\quad e_{\boldsymbol{u}}^n=\boldsymbol{u}_h^n-I_h(\boldsymbol{u}(t_n)),\\
e_{\boldsymbol{z}}^n=\boldsymbol{z}_h^n-R_h(\boldsymbol{z}(t_n)),\quad e_{\boldsymbol{\zeta}}^n=\boldsymbol{\zeta}_h^n-R_{\mathscr{C},h}(\boldsymbol{\zeta}(t_n)),\\
a_p^n=r_h(p(t_n))-p(t_n),\quad a_{c,p}^n=r_{\mathscr{C},h}(p_c(t_n))-p_c(t_n),\quad a_{\boldsymbol{u}}^n=I_h(\boldsymbol{u}(t_n))-\boldsymbol{u}(t_n),\\
a_{\boldsymbol{z}}^n=R_h(\boldsymbol{z}(t_n))-\boldsymbol{z}(t_n),\quad a_{\boldsymbol{\zeta}}^n=R_{\mathscr{C},h}(\boldsymbol{\zeta}(t_n))-\boldsymbol{\zeta}(t_n),
\end{gathered}$$

we derive the pressure error equation for any $\theta_h\in Q_h$ and $\theta_{c,h}\in\Theta_{\mathscr{C},h}$:

$$\begin{aligned}
&\left(\tfrac{1}{M}+c_f\varphi_0\right)\tfrac{1}{\Delta t}\big(\delta(e_p^n),\theta_h\big)+\tfrac{\alpha}{\Delta t}\big(\nabla\cdot\delta(e_{\boldsymbol{u}}^n),\theta_h\big)\\
&\quad+\tfrac{1}{\mu_f}\big(\nabla\cdot e_{\boldsymbol{z}}^n,\theta_h\big)-\tfrac{1}{\Delta t}\big([\delta(e_{\boldsymbol{u}}^n)]_{\mathscr{C}}\cdot\boldsymbol{n}^+,\theta_{c,h}\big)_{\mathscr{C}}\\
&\quad+\tfrac{1}{12\mu_f}\big(\overline{\nabla}\cdot((w^n)^{\frac{3}{2}}e_{\boldsymbol{\zeta}}^n),\theta_{c,h}\big)_{\mathscr{C}}-\tfrac{1}{\mu_f}\big([e_{\boldsymbol{z}}^n]_{\mathscr{C}}\cdot\boldsymbol{n}^+,\theta_{c,h}\big)_{\mathscr{C}}\\
&\quad=-\left(\tfrac{1}{M}+c_f\varphi_0\right)\tfrac{1}{\Delta t}\big(\delta(a_p^n),\theta_h\big)-\tfrac{\alpha}{\Delta t}\big(\nabla\cdot\delta(a_{\boldsymbol{u}}^n),\theta_h\big)\\
&\quad-\tfrac{1}{\mu_f}\big(\nabla\cdot a_{\boldsymbol{z}}^n,\theta_h\big)+\tfrac{1}{\Delta t}\big([\delta(a_{\boldsymbol{u}}^n)]_{\mathscr{C}}\cdot\boldsymbol{n}^+,\theta_{c,h}\big)_{\mathscr{C}}\\
&-\tfrac{1}{12\mu_f}\langle\overline{\nabla}\cdot((w^n)^{\frac{3}{2}}a_{\boldsymbol{\zeta}}^n),\theta_{c,h}\rangle_{\mathscr{C}}+\tfrac{1}{\mu_f}\langle[a_{\boldsymbol{z}}^n]_{\mathscr{C}}\cdot\boldsymbol{n}^+,\theta_{c,h}\rangle_{\mathscr{C}}-E_n(\theta_h,\theta_{c,h}).
\end{aligned}\tag{53}$$

Note that the above right-hand side simplifies because both $\big(\nabla\cdot a_{\boldsymbol{z}}^n,\theta_h\big)$ and $\langle[a_{\boldsymbol{z}}^n]_{\mathscr{C}}\cdot\boldsymbol{n}^+,\theta_{c,h}\rangle_{\mathscr{C}}$ vanish owing to (45). Likewise, the poro-elastic displacement error equations are, for all $\boldsymbol{v}_h\in\mathbb{V}_h$:

$$\begin{aligned}
&2G\big(\boldsymbol{\varepsilon}(e_{\boldsymbol{u}}^n),\boldsymbol{\varepsilon}(\boldsymbol{v}_h)\big)+\lambda\big(\nabla\cdot e_{\boldsymbol{u}}^n,\nabla\cdot\boldsymbol{v}_h\big)-\alpha\big(e_p^n,\nabla\cdot\boldsymbol{v}_h\big)+\big(e_{c,p}^n,[\boldsymbol{v}_h]_{\mathscr{C}}\cdot\boldsymbol{n}^+\big)_{\mathscr{C}}\\
&\ =-2G\big(\boldsymbol{\varepsilon}(a_{\boldsymbol{u}}^n),\boldsymbol{\varepsilon}(\boldsymbol{v}_h)\big)-\lambda\big(\nabla\cdot a_{\boldsymbol{u}}^n,\nabla\cdot\boldsymbol{v}_h\big)+\alpha\big(a_p^n,\nabla\cdot\boldsymbol{v}_h\big)-\big(a_{c,p}^n,[\boldsymbol{v}_h]_{\mathscr{C}}\cdot\boldsymbol{n}^+\big)_{\mathscr{C}}.
\end{aligned}\tag{54}$$

The fluid velocity error equations in the reservoir reduce to

$$\forall \boldsymbol{q}_h \in \boldsymbol{Z}_h, \quad \big(\boldsymbol{K}^{-1} e_z^n, \boldsymbol{q}_h\big) - \big(e_p^n, \nabla \cdot \boldsymbol{q}_h\big) + \big(e_{c,p}^n, [\boldsymbol{q}_h]_{\mathscr{C}} \cdot \boldsymbol{n}^+\big)_{\mathscr{C}} = -\big(\boldsymbol{K}^{-1} a_z^n, \boldsymbol{q}_h\big), \tag{55}$$

because the choice of r_h and $r_{\mathscr{C},h}$ (local L^2 projections) and the compatibility between the spaces imply that

$$\big(a_p^n, \nabla \cdot \boldsymbol{q}_h\big) = 0\ ,\ \big(a_{c,p}^n, [\boldsymbol{q}_h]_{\mathscr{C}} \cdot \boldsymbol{n}^+\big)_{\mathscr{C}} = 0.$$

In the fracture, the fluid velocity error equations read

$$\begin{aligned}\forall \boldsymbol{q}_{c,h} \in \boldsymbol{Z}_{\mathscr{C},h}, \quad &\big(e_{\boldsymbol{\zeta}}^n, \boldsymbol{q}_{c,h}\big)_{\mathscr{C}} - \big(e_{c,p}^n, \overline{\nabla} \cdot ((w^n)^{\frac{3}{2}} \boldsymbol{q}_{c,h})\big)_{\mathscr{C}} \\ &= -\big(a_{\boldsymbol{\zeta}}^n, \boldsymbol{q}_{c,h}\big)_{\mathscr{C}} + \big(a_{c,p}^n, \overline{\nabla} \cdot ((w^n)^{\frac{3}{2}} \boldsymbol{q}_{c,h})\big)_{\mathscr{C}}.\end{aligned} \tag{56}$$

On one hand, (55) and the inf-sup condition (47) imply the estimate

$$\|e_{c,p}^n\|_{L^2(\mathscr{C})} \leq \frac{1}{\beta_1^\star} E_{p,\boldsymbol{z}}^n, \tag{57}$$

where

$$E_{p,\boldsymbol{z}}^n = \big(\|e_p^n\|_{L^2(\Omega\setminus\mathscr{C})}^2 + \|\boldsymbol{K}^{-1} e_{\boldsymbol{z}}^n\|_{L^2(\Omega\setminus\mathscr{C})}^2\big)^{\frac{1}{2}} + \|\boldsymbol{K}^{-1} a_{\boldsymbol{z}}^n\|_{L^2(\Omega\setminus\mathscr{C})}. \tag{58}$$

On the other hand, by testing (53) with $\theta_h = e_p^n$ and $\theta_{c,h} = e_{c,p}^n$, (54) with $\boldsymbol{v}_h = \delta(e_{\boldsymbol{u}}^n)$, (55) with $\boldsymbol{q}_h = e_z^n$, and (56) with $\boldsymbol{q}_{c,h} = e_{\boldsymbol{\zeta}}^n$, multiplying everything by Δt, and summing the resulting equations, we derive:

$$\begin{aligned}\frac{1}{2}\left(\frac{1}{M} + c_f \varphi_0\right)&\Big(\delta\big(\|e_p^n\|_{L^2(\Omega)}^2\big) + \|\delta e_p^n\|_{L^2(\Omega)}^2\Big) \\ &+ G\Big(\delta\big(\|\boldsymbol{\varepsilon}(e_{\boldsymbol{u}}^n)\|_{L^2(\Omega\setminus\mathscr{C})}^2\big) + \|\boldsymbol{\varepsilon}(\delta e_{\boldsymbol{u}}^n)\|_{L^2(\Omega\setminus\mathscr{C})}^2\Big) \\ &+ \frac{\lambda}{2}\Big(\delta\big(\|\nabla \cdot e_{\boldsymbol{u}}^n\|_{L^2(\Omega\setminus\mathscr{C})}^2\big) + \|\nabla \cdot \delta e_{\boldsymbol{u}}^n\|_{L^2(\Omega\setminus\mathscr{C})}^2\Big) \\ &+ \frac{\Delta t}{\mu_f}\|\boldsymbol{K}^{-\frac{1}{2}} e_{\boldsymbol{z}}^n\|_{L^2(\Omega\setminus\mathscr{C})}^2 + \frac{\Delta t}{12\mu_f}\|e_{\boldsymbol{\zeta}}^n\|_{L^2(\mathscr{C})}^2 \\ = -\Delta t\, E_n(e_p^n, e_{c,p}^n) - A_{\boldsymbol{u},p}^n - A_{\boldsymbol{u}}^n + A_{\delta,\boldsymbol{u}}^n &+ \frac{\Delta t}{\mu_f}\Big(-A_{p,z,\zeta}^n - A_{z,\zeta}^n + \frac{1}{12}A_{\zeta}^n\Big),\end{aligned}$$

where

$$\begin{aligned}
A^n_{\boldsymbol{u},p} &= \left(\frac{1}{M} + c_f\varphi_0\right)\big(\delta(a^n_p), e^n_p\big) + \alpha\big(\nabla\cdot\delta(a^n_{\boldsymbol{u}}), e^n_p\big),\\
A^n_{p,z,\zeta} &= \big(\boldsymbol{K}^{-1}a^n_z, e^n_z\big) + \frac{1}{12}\big(a^n_\zeta, e^n_\zeta\big),\\
A^n_{\boldsymbol{u}} &= 2G\big(\boldsymbol{\varepsilon}(a^n_{\boldsymbol{u}}), \boldsymbol{\varepsilon}(\delta(e^n_{\boldsymbol{u}}))\big) + \lambda\big(\nabla\cdot a^n_{\boldsymbol{u}}, \nabla\cdot(\delta e^n_{\boldsymbol{u}})\big)\\
&\quad - \alpha\big(a^n_p, \nabla\cdot\delta(e^n_{\boldsymbol{u}})\big) + \big(a^n_{c,p}, [\delta(e^n_{\boldsymbol{u}})]_{\mathscr{C}}\cdot\boldsymbol{n}^+\big)_{\mathscr{C}},\\
A^n_{\delta,\boldsymbol{u}} &= \big(e^n_{c,p}, [\delta(a^n_{\boldsymbol{u}})]_{\mathscr{C}}\cdot\boldsymbol{n}^+\big)_{\mathscr{C}},\\
A^n_{z,\zeta} &= \frac{1}{12}\langle\overline{\nabla}\cdot\big((w^n)^{\frac{3}{2}}a^n_\zeta\big), e^n_{c,p}\rangle_{\mathscr{C}},\\
A^n_\zeta &= \big(\overline{\nabla}\cdot\big((w^n)^{\frac{3}{2}}e^n_\zeta\big), a^n_{c,p}\big)_{\mathscr{C}}.
\end{aligned}$$

Therefore, we must derive bounds for these six quantities. First, $A^n_{p,z,\zeta}$ has a straightforward bound:

$$\frac{\Delta t}{\mu_f}|A^n_{p,z,\zeta}| \le \frac{\Delta t}{\mu_f}\Big[\|\boldsymbol{K}^{-\frac{1}{2}}e^n_z\|_{L^2(\Omega\setminus\mathscr{C})}\|\boldsymbol{K}^{-\frac{1}{2}}a^n_z\|_{L^2(\Omega\setminus\mathscr{C})} + \frac{1}{12}\|e^n_\zeta\|_{L^2(\mathscr{C})}\|a^n_\zeta\|_{L^2(\mathscr{C})}\Big].$$

Next, considering that for example

$$\|\delta(r_h(p(t_n)) - p(t_n))\|_{L^2(\Omega)} \le \sqrt{\Delta t}\|r_h(p') - p'\|_{L^2(\Omega\times]t_{n-1},t_n[)},$$

we find a straightforward bound for $A^n_{\boldsymbol{u},p}$:

$$\begin{aligned}
|A^n_{\boldsymbol{u},p}| \le \sqrt{\Delta t}\|e^n_p\|_{L^2(\Omega)}\Big[\left(\frac{1}{M} + c_f\varphi_0\right)\|r_h(p') - p'\|_{L^2(\Omega\times]t_{n-1},t_n[)}\\
+ \alpha\|\nabla\cdot(I_h(\boldsymbol{u}') - \boldsymbol{u}')\|_{L^2((\Omega\setminus\mathscr{C})\times]t_{n-1},t_n[)}\Big].
\end{aligned}$$

Similarly, applying (57), the trace inequality (44), and (58), $A^n_{\delta,\boldsymbol{u}}$ is bounded by

$$\begin{aligned}
|A^n_{\delta,\boldsymbol{u}}| \le \frac{\sqrt{\Delta t}}{\beta_1^\star}C\Big[\big(\|e^n_p\|^2_{L^2(\Omega\setminus\mathscr{C})} + \|\boldsymbol{K}^{-1}e^n_z\|^2_{L^2(\Omega\setminus\mathscr{C})}\big)^{\frac{1}{2}}\\
+ \|\boldsymbol{K}^{-1}a^n_z\|_{L^2(\Omega\setminus\mathscr{C})}\Big]\|I_h(\boldsymbol{u}') - \boldsymbol{u}'\|_{L^2(t_{n-1},t_n;\mathbb{V})},
\end{aligned}$$

with the constant C of (44). Now we proceed with $A^n_{\boldsymbol{u}}$. As it involves factors of the form $\delta(e^n_{\boldsymbol{u}})$ that cannot be absorbed by the left-hand side, and considering that the whole expression needs to be summed over n, we use a summation by parts that switches the difference to the first factor:

$$\sum_{m=1}^{n} a^m(\delta b^m) = -\sum_{m=1}^{n-1}(\delta a^{m+1})b^m + a^n b^n - a^1 b^0.$$

This gives

$$\begin{aligned}\Big|\sum_{m=1}^{n} A_{\boldsymbol{u}}^{m}\Big| \le \sum_{m=1}^{n-1} \Big[&2G\|\boldsymbol{\varepsilon}(\delta(a_{\boldsymbol{u}}^{m+1}))\|_{L^2(\Omega\setminus\mathscr{C})}\|\boldsymbol{\varepsilon}(e_{\boldsymbol{u}}^{m})\|_{L^2(\Omega\setminus\mathscr{C})}\\
&+\lambda\|\nabla\cdot\delta(a_{\boldsymbol{u}}^{m+1})\|_{L^2(\Omega\setminus\mathscr{C})}\|\nabla\cdot e_{\boldsymbol{u}}^{m}\|_{L^2(\Omega\setminus\mathscr{C})}\\
&+\alpha\|\delta(a_{p}^{m+1})\|_{L^2(\Omega\setminus\mathscr{C})}\|\nabla\cdot e_{\boldsymbol{u}}^{m}\|_{L^2(\Omega\setminus\mathscr{C})} + C\|\delta(a_{c,p}^{m+1})\|_{L^2(\mathscr{C})}\|e_{\boldsymbol{u}}^{m}\|_{\mathbb{V}}\Big]\\
&+2G\|\boldsymbol{\varepsilon}(a_{\boldsymbol{u}}^{n})\|_{L^2(\Omega\setminus\mathscr{C})}\|\boldsymbol{\varepsilon}(e_{\boldsymbol{u}}^{n})\|_{L^2(\Omega\setminus\mathscr{C})} + \lambda\|\nabla\cdot a_{\boldsymbol{u}}^{n}\|_{L^2(\Omega\setminus\mathscr{C})}\|\nabla\cdot e_{\boldsymbol{u}}^{n}\|_{L^2(\Omega\setminus\mathscr{C})}\\
&+\alpha\|a_{p}^{n}\|_{L^2(\Omega\setminus\mathscr{C})}\|\nabla\cdot e_{\boldsymbol{u}}^{n}\|_{L^2(\Omega\setminus\mathscr{C})} + C\|a_{c,p}^{n}\|_{L^2(\mathscr{C})}\|e_{\boldsymbol{u}}^{n}\|_{\mathbb{V}}\\
&+2G\|\boldsymbol{\varepsilon}(a_{\boldsymbol{u}}^{1})\|_{L^2(\Omega\setminus\mathscr{C})}\|\boldsymbol{\varepsilon}(e_{\boldsymbol{u}}^{0})\|_{L^2(\Omega\setminus\mathscr{C})} + \lambda\|\nabla\cdot a_{\boldsymbol{u}}^{1}\|_{L^2(\Omega\setminus\mathscr{C})}\|\nabla\cdot e_{\boldsymbol{u}}^{0}\|_{L^2(\Omega\setminus\mathscr{C})}\\
&+\alpha\|a_{p}^{1}\|_{L^2(\Omega\setminus\mathscr{C})}\|\nabla\cdot e_{\boldsymbol{u}}^{0}\|_{L^2(\Omega\setminus\mathscr{C})} + C\|a_{c,p}^{1}\|_{L^2(\mathscr{C})}\|e_{\boldsymbol{u}}^{0}\|_{\mathbb{V}}.\end{aligned}$$

There remains to examine $A_{z,\zeta}^{n}$ and A_{ζ}^{n}. Let us start with A_{ζ}^{n}; it involves a factor that cannot be absorbed by the left-hand side. We cannot use directly the compatibility properties of the spaces on $\mathscr{C}$ and the projection properties because of the variable factor $(w^n)^{\frac{3}{2}}$. By expanding the divergence, we write

$$A_{\zeta}^{n} = \big((w^n)^{\frac{3}{2}}\overline{\nabla}\cdot e_{\zeta}^{n}, a_{c,p}^{n}\big)_{\mathscr{C}} + \big(\overline{\nabla}\,(w^n)^{\frac{3}{2}}\cdot e_{\zeta}^{n}, a_{c,p}^{n}\big)_{\mathscr{C}}. \tag{59}$$

Now, let $\pi_0(w^n)^{\frac{3}{2}}$ denote the average of $(w^n)^{\frac{3}{2}}$ in each e:

$$\pi_0(w^n)^{\frac{3}{2}} = \frac{1}{|e|}\int_e (w^n)^{\frac{3}{2}}.$$

Then the projection property of $r_{\mathscr{C},h}$ and the fact that $\pi_0(w^n)^{\frac{3}{2}}$ is a constant in each e yield

$$\begin{aligned}\big((w^n)^{\frac{3}{2}}\overline{\nabla}\cdot e_{\zeta}^{n}, a_{c,p}^{n}\big)_{\mathscr{C}} &= \big(\overline{\nabla}\cdot e_{\zeta}^{n}, \big((w^n)^{\frac{3}{2}} - \pi_0(w^n)^{\frac{3}{2}}\big)a_{c,p}^{n}\big)_{\mathscr{C}} + \big(\overline{\nabla}\cdot e_{\zeta}^{n}, \pi_0(w^n)^{\frac{3}{2}}a_{c,p}^{n}\big)_{\mathscr{C}}\\
&= \big(\overline{\nabla}\cdot e_{\zeta}^{n}, \big((w^n)^{\frac{3}{2}} - \pi_0(w^n)^{\frac{3}{2}}\big)a_{c,p}^{n}\big)_{\mathscr{C}}.\end{aligned}$$

Moreover,

$$\|(w^n)^{\frac{3}{2}} - \pi_0(w^n)^{\frac{3}{2}}\|_{L^4(e)} \le Ch_e\|\overline{\nabla}((w^n)^{\frac{3}{2}})\|_{L^4(e)}.$$

Therefore, by applying a local inverse inequality in each e, we deduce that

$$\big|\big(\overline{\nabla}\cdot e_{\zeta}^{n}, \big((w^n)^{\frac{3}{2}} - \pi_0(w^n)^{\frac{3}{2}}\big)a_{c,p}^{n}\big)_e\big| \le C\|\overline{\nabla}((w^n)^{\frac{3}{2}})\|_{L^4(e)}\|e_{\zeta}^{n}\|_{L^2(e)}\|a_{c,p}^{n}\|_{L^4(e)}.$$

Hence summing over all e in $\mathscr{C}_h$, we obtain

$$\big|\big(\overline{\nabla}\cdot e^n_{\boldsymbol{\zeta}}, \big((w^n)^{\frac{3}{2}} - \pi_0(w^n)^{\frac{3}{2}}\big)a^n_{c,p}\big)_{\mathscr{C}}\big| \le C\|\overline{\nabla}((w^n)^{\frac{3}{2}})\|_{L^4(\mathscr{C})}\|e^n_{\boldsymbol{\zeta}}\|_{L^2(\mathscr{C})}\|a^n_{c,p}\|_{L^4(\mathscr{C})},$$

where the first factor is bounded in view of (17). We can easily check that the second term in (59) has the same bound. Therefore

$$\frac{\Delta t}{12\mu_f}|A^n_{\boldsymbol{\zeta}}| \le \frac{\Delta t}{12\mu_f}C|(w^n)^{\frac{3}{2}}|_{W^{1,4}(\mathscr{C})}\|e^n_{\boldsymbol{\zeta}}\|_{L^2(\mathscr{C})}\|a^n_{c,p}\|_{L^4(\mathscr{C})}.$$

A similar argument can be applied to $A^n_{z,\zeta}$. Indeed, considering the compatibility of the finite element spaces on $\mathscr{C}$, we have:

$$\big(\overline{\nabla}\cdot a^n_{\boldsymbol{\zeta}}, \pi_0(w^n)^{\frac{3}{2}}e^n_{c,p}\big)_{\mathscr{C}} = 0.$$

Therefore, by writing

$$A^n_{z,\zeta} = \frac{1}{12}\Big[\big((w^n)^{\frac{3}{2}} - \pi_0(w^n)^{\frac{3}{2}}\big)\overline{\nabla}\cdot a^n_{\boldsymbol{\zeta}}, e^n_{c,p}\big)_{\mathscr{C}} + \big(\overline{\nabla}((w^n)^{\frac{3}{2}})\cdot a^n_{\boldsymbol{\zeta}}, e^n_{c,p}\big)_{\mathscr{C}}\Big],$$

we obtain

$$\big|A^n_{z,\zeta}| \le \frac{1}{12}\|e^n_{c,p}\|_{L^2(\mathscr{C})}|(w^n)^{\frac{3}{2}}|_{W^{1,4}(\mathscr{C})}\Big(\|a^n_{\boldsymbol{\zeta}}\|_{L^4(\mathscr{C})} + Ch\|\overline{\nabla}\cdot a^n_{\boldsymbol{\zeta}}\|_{L^4(\mathscr{C})}\Big).$$

Then by substituting (57) and (58) in the above inequality, we derive

$$\frac{\Delta t}{\mu_f}|A^n_{z,\zeta}| \le \frac{\Delta t}{\mu_f}\frac{1}{12\beta^\star_1}\Big[\big(\|e^n_p\|^2_{L^2(\Omega\backslash\mathscr{C})} + \|\boldsymbol{K}^{-1}e^n_z\|^2_{L^2(\Omega\backslash\mathscr{C})}\big)^{\frac{1}{2}} + \|\boldsymbol{K}^{-1}a^n_z\|_{L^2(\Omega\backslash\mathscr{C})}\Big]$$
$$\times|(w^n)^{\frac{3}{2}}|_{W^{1,4}(\mathscr{C})}\Big(\|a^n_{\boldsymbol{\zeta}}\|_{L^4(\mathscr{C})} + Ch\|\overline{\nabla}\cdot a^n_{\boldsymbol{\zeta}}\|_{L^4(\mathscr{C})}\Big).$$

The next theorem collects these results and concludes with a basic error bound. The proof is skipped, as it is a straightforward consequence of repeated applications of Young's inequality with suitable coefficients and a discrete Gronwall's Lemma.

Theorem 6 *Let the data $\boldsymbol{f}$, $\tilde{q}$, $\tilde{q}_W$ and $p(0)$ be sufficiently smooth and let Hypotheses 1 and 3 hold. Suppose that problem* (P2.1)–(P2.5) *and* (16) *has a sufficiently smooth solution. Then the sequence of solutions $(\boldsymbol{u}^n_h, p^n_h, p^n_{c,h}, z^n_h, \boldsymbol{\zeta}^n_h)$ of* (P3.4)–(P3.8) *with starting values $(p^0_h, p^0_{c,h}, \boldsymbol{u}^0_h, z^0_h, \boldsymbol{\zeta}^0_h)$, $\boldsymbol{u}^0_h$, z^0_h, $\boldsymbol{\zeta}^0_h$ being computed respectively by* (P3.1), (P3.2)*,* (P3.3)*, satisfies the following error bounds for any integer n, $1 \le n \le N$:*

$$
\begin{aligned}
&\left(\frac{1}{M} + c_f\varphi_0\right)\left(\|e_p^n\|^2_{L^2(\Omega)} + \sum_{m=1}^{n}\|\delta e_p^m\|^2_{L^2(\Omega)}\right)\\
&\quad + 2G\left(\|\boldsymbol{\varepsilon}(e_{\boldsymbol{u}}^n)\|^2_{L^2(\Omega\setminus\mathscr{C})} + \sum_{m=1}^{n}\|\boldsymbol{\varepsilon}(\delta e_{\boldsymbol{u}}^m)\|^2_{L^2(\Omega\setminus\mathscr{C})}\right)\\
&\quad + \lambda\left(\|\nabla\cdot e_{\boldsymbol{u}}^n\|^2_{L^2(\Omega\setminus\mathscr{C})} + \sum_{m=1}^{n}\|\nabla\cdot\delta e_{\boldsymbol{u}}^m\|^2_{L^2(\Omega\setminus\mathscr{C})}\right)\\
&\quad + \frac{1}{\mu_f}\sum_{m=1}^{n}\Delta t\left(\|\boldsymbol{K}^{-\frac{1}{2}}e_{\boldsymbol{z}}^m\|^2_{L^2(\Omega\setminus\mathscr{C})} + \frac{1}{12}\|e_{\boldsymbol{\zeta}}^m\|^2_{L^2(\mathscr{C})}\right)\\
&\le C\Big[(\Delta t)^2\big(\|p''\|^2_{L^2(\Omega\times]0,t_n[)} + \|\boldsymbol{u}''\|^2_{L^2(0,t_n;\mathbb{V})}\big) + \|e_{\boldsymbol{u}}^0\|^2_{\mathbb{V}}\\
&\quad + \|I_h(\boldsymbol{u}) - \boldsymbol{u}\|^2_{H^1(0,t_n;\mathbb{V})} + \|r_h(p) - p\|^2_{H^1(0,t_n;L^2(\Omega\setminus\mathscr{C}))}\\
&\quad + \|r_{\mathscr{C},h}(p_c) - p_c\|^2_{H^1(0,t_n;L^2(\mathscr{C}))} + \|R_h(\boldsymbol{z}) - \boldsymbol{z}\|^2_{\mathscr{C}^0(0,t_n;L^2(\Omega\setminus\mathscr{C})^d)}\\
&\quad + \|w^{\frac{3}{2}}\|_{\mathscr{C}^0(0,t_n;W^{1,4}(\mathscr{C}))}\big(\|r_{\mathscr{C},h}(p_c) - p_c\|^2_{\mathscr{C}^0(0,t_n;L^4(\mathscr{C}))}\\
&\quad + \|R_{\mathscr{C},h}(\boldsymbol{\zeta}) - \boldsymbol{\zeta}\|^2_{\mathscr{C}^0(0,t_n;L^4(\mathscr{C})^{d-1})} + h\|\overline{\nabla}\cdot(R_{\mathscr{C},h}(\boldsymbol{\zeta}) - \boldsymbol{\zeta})\|^2_{\mathscr{C}^0(0,t_n;L^4(\mathscr{C}))}\big)\Big]\exp(t_n),
\end{aligned}
$$

with a constant C independent of n, h, and Δt, and

$$
\begin{aligned}
\sum_{m=1}^{n}\Delta t\|e_{c,p}^m\|^2_{L^2(\mathscr{C})} \le \frac{2}{(\beta_1^\star)^2}\sum_{m=1}^{n}\Delta t\Big(&\|e_p^m\|^2_{L^2(\Omega)} + \|\boldsymbol{K}^{-1}e_{\boldsymbol{z}}^m\|^2_{L^2(\Omega\setminus\mathscr{C})}\\
&+ \|\boldsymbol{K}^{-1}(R_h(\boldsymbol{z}(t_m)) - \boldsymbol{z}(t_m))\|^2_{L^2(\Omega\setminus\mathscr{C})}\Big),
\end{aligned}
$$

with the constant $\beta_1^\star$ of (47).

Remark 3 Further error bounds, in the spirit of the estimates derived in Sect. 4.2, are more delicate. On one hand, Hypothesis 2 is quite restrictive, and, on the other hand, the choice of the fracture's discrete spaces in (P3.1)–(P3.8) is not consistent with the theoretical setting because the relevant space for the fracture's pressure p_c should be $H^{\frac{1}{2}}(\mathscr{C})$ instead of $L^2(\mathscr{C})$. We use L^2 pressures because they are locally mass conservative and by taking advantage of the finite dimension, they lead to the basic estimates of Sect. 5.6, but it is not clear that they lead to additional satisfactory estimates and in particular to a useful bound for the discrete leakage term. If we want complete estimates, we can modify the scheme so that it matches the setting of (P2.1)–(P2.5), and in particular uses continuous pressures in the fracture. For instance, we can choose

$$
\Theta_{\mathscr{C},h} = \left\{q \in \mathscr{C}^0(\mathscr{C}) \mid q|_{\mathscr{S}_i} \in \Theta_{\mathscr{S}_i,h},\ 1 \le i \le I\right\},
$$

with

$$
\Theta_{\mathscr{S},h} = \left\{q \in \mathscr{C}^0(\mathscr{S}) \mid q|_e \leftrightarrow \hat{q},\ \hat{q} \in \hat{\Theta}_{\mathscr{C}}(\hat{e}),\ \forall e \in \mathscr{T}_{\mathscr{S},h}\right\},
$$

without changing the other spaces. We can prove that $p_{c,h}$ satisfies an inf-sup condition in $H^{\frac{1}{2}}(\mathscr{C})$ by exploiting the fact that if the functions of $\Theta_{\mathscr{C},h}$ are continuous and piecewise polynomials, then they belong to $H^1(\mathscr{C})$. The proof is more complex than that of Lemma 2, but it still requires (46). This hypothesis holds if we raise the degree of the polynomials, which may not be desirable. □

6 Fixed Stress Splitting

We shall use the following fixed stress splitting algorithm for decoupling the computation of the mechanics from that of the flow. To simplify, we describe it at the exact level and we denote the time derivative by ∂_t. It proceeds in two steps. First the flow problem in the reservoir and fracture is solved in a monolithic manner:

Step a. Given $\boldsymbol{u}^n$, we solve for p^{n+1}, $\boldsymbol{z}^{n+1}$, p_c^{n+1}, $\boldsymbol{\zeta}^{n+1}$ such that

$$\Big(\frac{1}{M} + c_f\varphi_0 + \frac{\alpha^2}{\lambda}\Big)\partial_t p^{n+1} + \frac{1}{\mu_f}\nabla\cdot\boldsymbol{z}^{n+1} = \frac{\alpha^2}{\lambda}\partial_t p^n - \alpha\nabla\cdot\partial_t\boldsymbol{u}^n + \tilde{q} \text{ in } \Omega\setminus\mathscr{C}, \tag{60}$$

$$\begin{aligned}\boldsymbol{z}^{n+1} &= -\boldsymbol{K}\nabla\big(p^{n+1} - \rho_{f,r}g\eta\big),\\ \gamma_c\partial_t p_c^{n+1} + \partial_t w^n + \frac{1}{12\mu_f}\overline{\nabla}\cdot\big((w^n)^{\frac{3}{2}}\boldsymbol{\zeta}^{n+1}\big) &= \gamma_c\partial_t p_c^n + \tilde{q}_W + \frac{1}{\mu_f}[\boldsymbol{z}^{n+1}]_{\mathscr{C}}\cdot\boldsymbol{n}^+ \text{ in } \mathscr{C},\end{aligned} \tag{61}$$

$$\begin{aligned}\boldsymbol{\zeta}^{n+1} &= -(w^n)^{\frac{3}{2}}\overline{\nabla}\big(p_c^{n+1} - \rho_{f,r}g\eta\big),\\ w^n &= -[\boldsymbol{u}^n]_{\mathscr{C}}\cdot\boldsymbol{n}^+.\end{aligned}$$

Once the flow is computed, we update the displacement solution.

Step b. Given p^{n+1}, $\boldsymbol{z}^{n+1}$, p_c^{n+1}, $\boldsymbol{\zeta}^{n+1}$, we solve for $\boldsymbol{u}^{n+1}$ satisfying

$$-\operatorname{div}\boldsymbol{\sigma}^{\text{por}}(\boldsymbol{u}^{n+1}, p^{n+1}) = \boldsymbol{f} \quad \text{in } \Omega\setminus\mathscr{C}, \tag{62}$$

$$(\boldsymbol{\sigma}^{\text{por}}(\boldsymbol{u}^{n+1}, p^{n+1}))^{\star}\boldsymbol{n}^{\star} = -p_c^{n+1}\boldsymbol{n}^{\star}, \quad \star = +, - \text{ on } \mathscr{C}, \tag{63}$$

where

$$\boldsymbol{\sigma}^{\text{por}}(\boldsymbol{u}^{n+1}, p^{n+1}) = \boldsymbol{\sigma}(\boldsymbol{u}^{n+1}) - \alpha\, p^{n+1}\boldsymbol{I} \text{ in } \Omega\setminus\mathscr{C}.$$

The stabilizing terms $\frac{\alpha^2}{\lambda}\partial_t p^{n+1}$ and $\gamma_c\partial_t p_c^{n+1}$ are added to the left-hand sides of (60) and (61) respectively, with similar terms on the right-hand sides of the equations for the sake of consistency. The first term is a standard addition in fixed stress splitting, see [23]. Motivated by this, we add a similar term to the fracture equation with an adjustable coefficient γ_c.

The following definition of the volumetric mean stress:

$$\sigma_v = \sigma_{v,0} + \lambda\nabla\cdot\boldsymbol{u} - \alpha(p - p_0),$$

where $\sigma_{v,0}$ denotes the initial volumetric stress, justifies the name of the algorithm. Indeed, as $\sigma_{v,0}$ and p_0 are constant in time, we have

$$-\frac{\alpha}{\lambda}\partial_t\sigma_v^n = \frac{\alpha^2}{\lambda}\partial_t p^n - \alpha\nabla\cdot\partial_t\boldsymbol{u}^n, \tag{64}$$

and we recognize the first two terms in the right-hand side of (60).

The variational form of the algorithm reads as follows:

Step a. Find $p^{n+1} \in L^\infty(0,T;L^2(\Omega))$, $p_c^{n+1} \in L^2(0,T;H^{\frac{1}{2}}(\mathscr{C}))$, $\boldsymbol{z}^{n+1} \in L^2(0,T;\boldsymbol{Z})$, and $\boldsymbol{\zeta}^{n+1} \in L^2(0,T;\boldsymbol{Z}_{\mathscr{C}})$ such that for all $t \in]0,T[$

$$\begin{aligned}
&\forall\theta \in L^2(\Omega), \quad \Big(\Big(\tfrac{1}{M} + c_f\varphi_0 + \tfrac{\alpha^2}{\lambda}\Big)\partial_t p^{n+1}, \theta\Big) + \tfrac{1}{\mu_f}\big(\nabla\cdot\boldsymbol{z}^{n+1}, \theta\big) \\
&\qquad = \Big(-\tfrac{\alpha}{\lambda}\partial_t\sigma_v^n, \theta\Big) + \big(\tilde{q}, \theta\big), \\
&\forall\theta_c \in H^{\frac{1}{2}}(\mathscr{C}), \quad \gamma_c\big(\partial_t p_c^{n+1}, \theta_c\big)_{\mathscr{C}} + \tfrac{1}{12\mu_f}\big(\overline{\nabla}\cdot((w^n)^{\frac{3}{2}}\boldsymbol{\zeta}^{n+1}), \theta_c\big)_{\mathscr{C}} \\
&-\tfrac{1}{\mu_f}\big([\boldsymbol{z}^{n+1}]_{\mathscr{C}}\cdot\boldsymbol{n}^+, \theta_c\big)_{\mathscr{C}} = \Big(\gamma_c\partial_t p_c^n, \theta_c\Big)_{\mathscr{C}} + \big(\partial_t[\boldsymbol{u}^n]_{\mathscr{C}}\cdot\boldsymbol{n}^+, \theta_c\big)_{\mathscr{C}} + \big(\tilde{q}_W, \theta_c\big)_{\mathscr{C}}, \\
&\forall\boldsymbol{q} \in \boldsymbol{Z}, \quad \big(\boldsymbol{K}^{-1}\boldsymbol{z}^{n+1}, \boldsymbol{q}\big) = (p^{n+1}, \nabla\cdot\boldsymbol{q}) - \big(p_c^{n+1}, [\boldsymbol{q}]_{\mathscr{C}}\cdot\boldsymbol{n}^+\big)_{\mathscr{C}} + \big(\nabla(\rho_{f,r}g\eta), \boldsymbol{q}\big), \\
&\forall\boldsymbol{q}_c \in \boldsymbol{Z}_{\mathscr{C}}, \quad \big(\boldsymbol{\zeta}^{n+1}, \boldsymbol{q}_c\big)_{\mathscr{C}} = \big(p_c^{n+1}, \overline{\nabla}\cdot((w^n)^{\frac{3}{2}}\boldsymbol{q}_c)\big)_{\mathscr{C}} + \big((w^n)^{\frac{3}{2}}\overline{\nabla}(\rho_{f,r}g\eta), \boldsymbol{q}_c\big)_{\mathscr{C}},
\end{aligned} \tag{65}$$

with the initial condition, independent of n,

$$p^{n+1}(0) = p_0\,, \;\; p_c^{n+1}(0) = p_0|_{\mathscr{C}}\,.$$

Step b. Given p^{n+1}, $\boldsymbol{z}^{n+1}$, p_c^{n+1}, $\boldsymbol{\zeta}^{n+1}$, find $\boldsymbol{u}^{n+1} \in L^\infty(0,T;\mathbb{V})$ such that for all $t \in]0,T[$,

$$\begin{aligned}
\forall\boldsymbol{v} \in \mathbb{V}, \quad 2G\big(\boldsymbol{\varepsilon}(\boldsymbol{u}^{n+1}), \boldsymbol{\varepsilon}(\boldsymbol{v})\big) + \lambda\big(\nabla\cdot\boldsymbol{u}^{n+1}, \nabla\cdot\boldsymbol{v}\big) - \alpha\big(p^{n+1}, \nabla\cdot\boldsymbol{v}\big) \\
+ \big(p_c^{n+1}, [\boldsymbol{v}]_{\mathscr{C}}\cdot\boldsymbol{n}^+\big)_{\mathscr{C}} = \big(\boldsymbol{f}, \boldsymbol{v}\big).
\end{aligned} \tag{66}$$

We have seen that (66) defines $\boldsymbol{u}^{n+1}(0)$ in terms of p_0 and p_c^0, and in turn $w^{n+1}(0) = -[\boldsymbol{u}^{n+1}(0)]_{\mathscr{C}}\cdot\boldsymbol{n}^+$, all quantities being independent of n. To begin the iteration, for $n = 0$, we assign as initial condition $p^0 = p_0$, $p_c^0 = p_0|_{\mathscr{C}}$, $\boldsymbol{u}^0$ is computed from p^0 and p_c^0 by (P2.1) at time $t = 0$, and $w^0 = -[\boldsymbol{u}^0]_{\mathscr{C}}\cdot\boldsymbol{n}^+$. More specific details can be found in Ganis et al. [13].

Notice that in (65), the right-hand side has been re-written in terms of the volumetric mean total stress as defined in (64). As such, the convergence of this algorithm is an open problem. With a suitable choice of parameter γ_c in terms of the material parameters, constants of the trace and the Korn's inequalities, convergence of a simplified version is established in Girault et al. [14].

A flowchart for a fixed stress splitting scheme is provided in Fig. 2. Here we iterate between the flow solution assuming a fixed stress field and the mechanics solution assuming fixed pressure and saturation fields. For mechanics we apply a Galerkin finite element with continuous piecewise linears and for flow a mixed finite element (MFMFE) is used as described in Sect. 5.3. The simulations were performed using the coupled flow and geomechanics reservoir simulator IPARS (Integrated Parallel Accurate Reservoir Simulator). IPARS is capable of handling complex subsurface flow descriptions such as two-phase, black oil and compositional flow along with chemical equilibrium and kinetic type reactions.

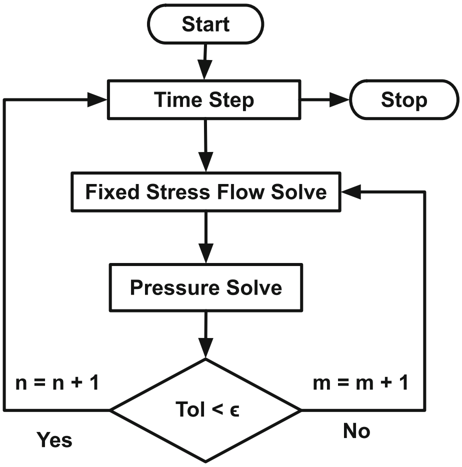


Fig. 2 Flowchart for iteratively coupled flow and poroelasticity in IPARS

7 Numerical Results

In this numerical experiment, we show the stress and displacement fields in a poroelastic domain with two orthogonal fractures. Figure 3 shows a schematic of the problem along with boundary conditions and location of the fractures. A square domain $\Omega = (0, 250\,\text{ft}) \times (0, 250\,\text{ft})$ is considered with two orthogonal fractures along the axes $\{y = 125\}$ ft and $\{z = 150\}$ ft, each 50 ft in length with one end point at (125, 62.5 ft) and (100, 150 ft), respectively. A no flow ($\boldsymbol{z} = \mathbf{0}$) boundary condition is specified on all the edges allowing the pressure in the domain to rise with time. A zero displacement ($\boldsymbol{u} = \mathbf{0}$) boundary condition is specified for the left and bottom edges whereas normal stresses ($\boldsymbol{\sigma}^{por}\boldsymbol{n}$) of $(-6300, 0)$ psi and $(-6400, 0)$ psi are specified at the right and top edges, respectively as shown in Fig. 3. Further, an initial condition of 500 psi for pressure is specified both in the poroelastic domain (Ω) and on the fracture ($\mathscr{C}$). Fluid is injected into the middle of each fracture at 5000 psi.

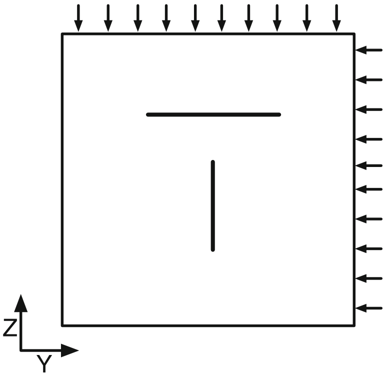


Fig. 3 Problem schematic

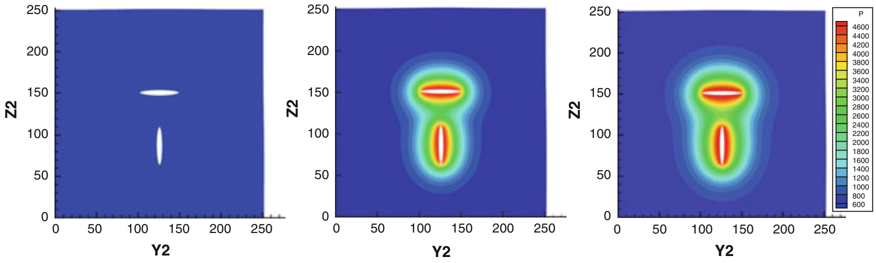


Fig. 4 Pressure profiles at $T = 0.0$, 0.05 and 0.1 day

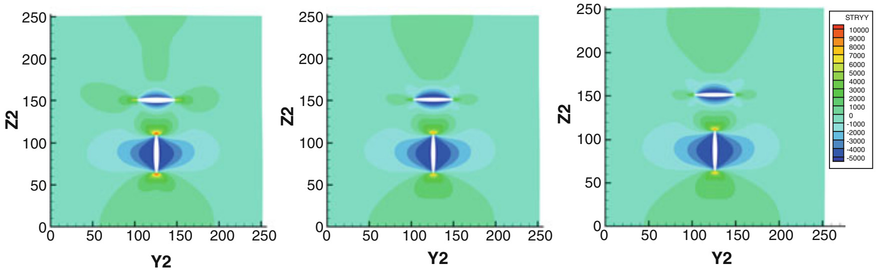


Fig. 5 Stress (σ_{yy}) profiles at $T = 0.0$, 0.05 and 0.1 day

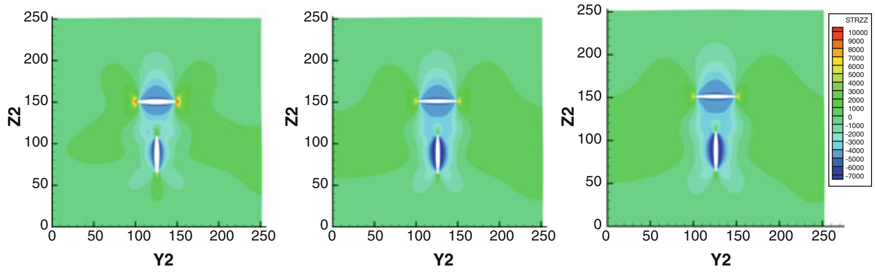


Fig. 6 Stress (σ_{zz}) profiles at $T = 0.0$, 0.05 and 0.1 day

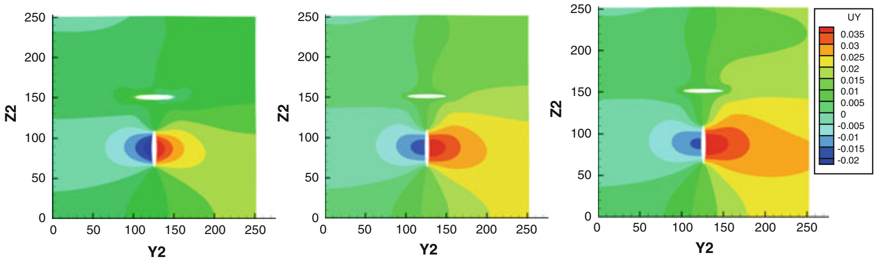


Fig. 7 Y-direction displacement profiles at $T = 0.0$, 0.05 and 0.1 day

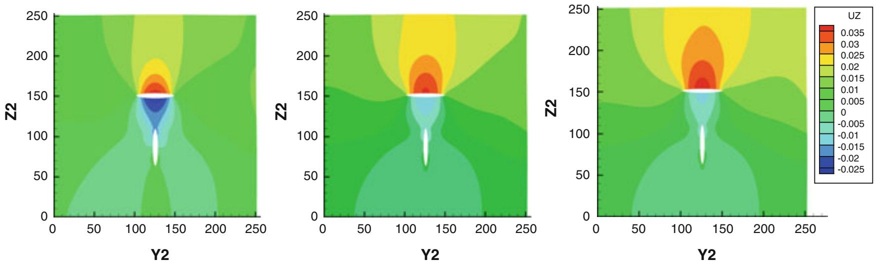


Fig. 8 Z-direction displacement profiles at $T = 0.0$, 0.05 and 0.1 day

A homogeneous porosity value of 0.2 and homogeneous and isotropic permeability tensor of 50 mD is assumed. The fluid is assumed to be slightly compressible with density 62.4 lbm/ft^3 and compressibility 1×10^{-6} psi^{-1}. The Young's modulus and Poisson's ratio of the poroelastic medium are 7.3×10^6 psi and 0.2, respectively.

The domain is discretized into 80×80 structured hexahedral elements with a uniform mesh width of 3.125 ft in both y and z directions. Figures 4, 5, 6, 7 and 8 show the pressure, stress and displacement profiles in y and z directions, respectively, at $T = 0.0$, 0.05 and 0.1 days.

References

1. Alboin C, Jaffré J, Roberts JE, Serres C (2001) Modeling fractures as interfaces for flow and transport in porous media. In: Chen Z, Ewing RE (eds) Fluid flow and transport in porous media: mathematical and numerical treatment. South Hadley, MA. (Vol 295 of Contemporary Mathematics). American Mathematical Society, Providence, RI, pp 13–24 (2002)
2. Babuška I (1972/73) The finite element method with Lagrangian multipliers. Numer Math 20:179–192
3. Biot MA (1941) General theory of three-dimensional consolidation. J Appl Phys 12(2):155–164
4. Bourgeat A, Mikelić A, Piatnitski A (2003) On the double porosity model of a single phase flow in random media. Asymptot Anal 34(3–4):311–332
5. Brenner SC, Scott LR (2008) The mathematical theory of finite element methods, 3rd edn. Springer, New York
6. Brezzi F (1974) On the existence, uniqueness and approximation of saddle-point problems arising from Lagrangian multipliers. Rev Française Automat Inform Rech Opér Anal Numér 8(R–2):129–151
7. Brezzi F, Douglas J Jr, Durán R, Fortin M (1987) Mixed finite elements for second order elliptic problems in three variables. Numer Math 51(2):237–250
8. Brezzi F, Douglas J Jr, Marini LD (1985) Two families of mixed finite elements for second order elliptic problems. Numer Math 47(2):217–235
9. Ciarlet PG (1991) Basic error estimates for elliptic problems. In: Handbook of numerical analysis, Vol II. North-Holland, Amsterdam, pp 17–351
10. Dean RH, Schmidt JH (2009) Hydraulic-fracture predictions with a fully coupled geomechanical reservoir simulator. SPE J 14(04):707–714
11. Fasano A, Mikelić A, Primicerio M (1998) Homogenization of flows through porous media with permeable grains. Adv Math Sci Appl 8(1):1–31

12. Galvis J, Sarkis M (2007) Non-matching mortar discretization analysis for the coupling Stokes-Darcy equations. Electron Trans Numer Anal 26:350–384
13. Ganis B, Girault V, Mear M, Singh G, Wheeler MF (2014) Modeling fractures in a poro-elastic medium. Oil Gas Sci Tech 69(4):515–528
14. Girault V, Kumar K, Wheeler MF (2016) Convergence of iterative coupling of geomechanics with flow in a fractured poroelastic medium. Comput Geosci 20(5):997–1011
15. Girault V, Pencheva G, Wheeler MF, Wildey T (2011) Domain decomposition for poroelasticity and elasticity with DG jumps and mortars. Math Models Methods Appl Sci 21(1):169–213
16. Girault V, Raviart P-A (1986) Finite element methods for Navier-Stokes equations: theory and algorithms, vol 5. Springer Series in Computational Mathematics. Springer, Berlin
17. Girault V, Wheeler MF, Ganis B, Mear ME (2013) A lubrication fracture model in a poro-elastic medium. ICES Report 13-32, Institute for Computational Engineering and Sciences, University of Texas at Austin
18. Girault V, Wheeler MF, Ganis B, Mear ME (2015) A lubrication fracture model in a poro-elastic medium. Math Models Methods Appl Sci 25(4):587–645
19. Ingram R, Wheeler MF, Yotov I (2010) A multipoint flux mixed finite element method on hexahedra. SIAM J Numer Anal 48(4):1281–1312
20. Jerison DS, Kenig CE (1981) The Neumann problem on Lipschitz domains. Bull Am Math Soc (NS) 4(2):203–207
21. Lions J-L, Magenes E (1972) Non-homogeneous boundary value problems and applications, vol I. Springer, New York
22. Martin V, Jaffré J, Roberts JE (2005) Modeling fractures and barriers as interfaces for flow in porous media. SIAM J Sci Comput 26(5):1667–1691
23. Mikelić A, Wheeler MF (2013) Convergence of iterative coupling for coupled flow and geomechanics. Comput Geosci 17(3):455–461
24. Phillips PJ, Wheeler MF (2007) A coupling of mixed and continuous Galerkin finite element methods for poroelasticity. I. The continuous in time case. Comput Geosci 11(2):131–144
25. Scott LR, Zhang S (1990) Finite element interpolation of nonsmooth functions satisfying boundary conditions. Math Comp 54(190):483–493
26. Showalter RE (2000) Diffusion in poro-elastic media. J Math Anal Appl 251(1):310–340
27. Wheeler MF, Xue G, Yotov I (2014) Coupling multipoint flux mixed finite element methods with continuous Galerkin methods for poroelasticity. Comput Geosci 18(1):57–75
28. Witherspoon PA, Wang JSY, Iwai K, Gale JE (1980) Validity of cubic law for fluid flow in a deformable rock fracture. Water Resour Res 16(6):1016–1024

Two Decades of Wave-Like Equation for the Numerical Simulation of Incompressible Viscous Flow: A Review

Roland Glowinski and Tsorng-Whay Pan

Abstract A wave-like equation based method for the numerical solution of the Navier-Stokes equations modeling incompressible viscous flow was introduced nearly twenty years ago. From its inception to nowadays it has been applied successfully to the numerical solution of two and three dimensional flow problems for incompressible Newtonian and non-Newtonian viscous fluids, in flow regions with fixed or moving boundaries. The main goals of this article are: (i) To recall the foundations of the wave-like equation methodology, and (ii) to review some typical viscous flow problems where it has been applied successfully.

Keywords Incompressible viscous flow · Operator splitting time discretization schemes · Wave-like equation method for the numerical treatment of the advection step · Finite element approximations

1 Introduction

Some time ago, the authors of this article were asked to contribute to a volume dedicated to their colleagues and friends *W. Fitzgibbon*, *Y. Kuznetsov* and *O. Pironneau* on the occasion of their 70th anniversary. The authors decided to take advantage of this special volume to celebrate another anniversary: Indeed, nearly twenty years ago, they dropped the nonlinear least-squares methodology they have been using for years for the numerical treatment of the advection operator, encountered in the

Dedicated to W. Fitzgibbon, Y. Kuznetsov and O. Pironneau on the occasion of their 70th anniversary.

R. Glowinski (✉) · T.-W. Pan
Department of Mathematics, University of Houston, Houston, TX 77204, USA
e-mail: roland@math.uh.edu

T.-W. Pan
e-mail: pan@math.uh.edu

R. Glowinski
Department of Mathematics, Hong-Kong Baptist University, Kowloon Tong, Hong Kong

B. N. Chetverushkin et al. (eds.), *Contributions to Partial Differential Equations and Applications*, Computational Methods in Applied Sciences 47,
https://doi.org/10.1007/978-3-319-78325-3_13

Navier-Stokes equations modelling *incompressible viscous flow*, and started employing systematically a novel methodology based on a *wave-like equation* modelling of the advection. From then to now, the wave-like equation method has been successfully applied, by the authors and other people, to the numerical simulation of a rather large variety of incompressible viscous flows, justifying in the authors opinion a relatively detailed dedicated review publication. The content of this article is as follows: In Sect. 2, we will describe the wave-like equation method when applied to the numerical solution of the Navier-Stokes equations modelling incompressible viscous flow, and take advantage of this section to provide related references. In Sects. 3–5 we will describe and comment on several successful applications of the wave-like equation based methodology; they concern Newtonian, visco-elastic and particulate viscous flows.

2 The Wave-Like Equation Method for the Incompressible Navier-Stokes Equations

Our starting point will be the *Navier-Stokes equations* modeling the flow of *incompressible Newtonian viscous fluids*, namely

$$\begin{cases} \rho\left[\dfrac{\partial \mathbf{u}}{\partial t} + (\mathbf{u}\cdot\boldsymbol{\nabla})\mathbf{u}\right] - \mu\boldsymbol{\nabla}^2\mathbf{u} + \boldsymbol{\nabla}p = \mathbf{f} \quad \text{in } \Omega\times(0,T), \\ \boldsymbol{\nabla}\cdot\mathbf{u} = 0 \quad \text{in } \Omega\times(0,T), \\ \mathbf{u}(0) = \mathbf{u}_0 \quad \text{with} \quad \boldsymbol{\nabla}\cdot\mathbf{u}_0 = 0, \\ \mathbf{u} = \mathbf{u}_B \quad \text{on } \Gamma\times(0,T) \quad \text{with} \quad \displaystyle\int_\Gamma \mathbf{u}_B(t)\cdot\mathbf{n}\, d\Gamma = 0 \quad \text{on } (0,T), \end{cases} \tag{1}$$

where:

- Ω (a sub-domain of $\mathbb{R}^d$, $d = 2$ or 3) is the flow region, and $0 < T \leq +\infty$. We denote by Γ the boundary of Ω.
- $\mathbf{u}$ (resp., p) denotes the flow velocity (resp., pressure), and $\mathbf{f}$ a density of external forces.
- ρ and μ are both > 0, and denote the fluid density and viscosity, respectively.
- $\phi(t)$ denotes the function $\mathbf{x} \to \phi(\mathbf{x}, t)$ (with $\mathbf{x} = \{x_i\}_{i=1}^d$).
- $\mathbf{n}$ denotes the unit outward normal vector at Γ.

The numerical solution of problem (1) has generated a most abundant literature (see, in particular, the related references provided by *Google Scholar*). Among the many methods for the numerical solution of (1), we will single out those based on *operator-splitting*. Applying the *Lie scheme* (see, e.g., [20, 21, 25] for a general discussion of that scheme), we obtain (among other possibilities) the following time-discretization of problem (1) (with $\triangle t(>0)$ a time-discretization step and $t^n = n\triangle t$):

$$\mathbf{u}^0 = \mathbf{u}_0. \tag{2}$$

For $n \geq 0, \mathbf{u}^n \to \{\mathbf{u}^{n+1/2}, p^{n+1}\} \to \mathbf{u}^{n+1}$ *via the solution of*

$$\begin{cases} \rho \dfrac{\mathbf{u}^{n+1/2} - \mathbf{u}^n}{\triangle t} - \mu \nabla^2 \mathbf{u}^{n+1/2} + \nabla p^{n+1} = \mathbf{f}^{n+1} & \text{in } \Omega, \\ \nabla \cdot \mathbf{u}^{n+1/2} = 0 & \text{in } \Omega, \\ \mathbf{u}^{n+1/2} = \mathbf{u}_B(t^{n+1}) & \text{on } \Gamma, \end{cases} \tag{3}$$

and

$$\begin{cases} \dfrac{\partial \mathbf{w}}{\partial t} + (\mathbf{u}^{n+1/2} \cdot \nabla)\mathbf{w} = \mathbf{0} & \text{in } \Omega \times (t^n, t^{n+1}), \\ \mathbf{w}(t^n) = \mathbf{u}^{n+1/2}, \\ \mathbf{w}(t) = \mathbf{u}^{n+1/2}(= \mathbf{u}_B(t^{n+1})) & \text{on } \Gamma_-^{n+1} \times (t^n, t^{n+1}), \end{cases} \tag{4.1}$$

$$\mathbf{u}^{n+1} = \mathbf{w}(t^{n+1}), \tag{4.2}$$

with $\Gamma_-^{n+1} = \{\mathbf{x} \mid \mathbf{x} \in \Gamma, \mathbf{u}_B(\mathbf{x}, t^{n+1}) \cdot \mathbf{n}(\mathbf{x}) < 0\}$.

Remark 1 The time discretization of problem (1) by the *Strang symmetrized scheme* (a more sophisticated variant of the Lie scheme) is discussed in [12, 20] (see also the references therein). □

The solution of the (generalized) Stokes problem (3) being a well-documented (and different) issue (see, e.g., [3, 20]), we will focus on the most controversial part of scheme (2)–(4), namely the solution of the initial value problem (4.1). One can easily show that in (4.1), each component of $\mathbf{w}$ is solution of an initial-boundary value problem of the following type:

$$\begin{cases} \dfrac{\partial \phi}{\partial t} + \mathbf{V} \cdot \nabla \phi = 0 & \text{in } \Omega \times (t_0, t_f), \\ \phi(t_0) = \phi_0, \\ \phi = g & \text{on } \Gamma_- \times (t_0, t_f), \end{cases} \tag{5}$$

where $\frac{\partial \mathbf{V}}{\partial t} = \mathbf{0}, \nabla \cdot \mathbf{V} = 0, \frac{\partial g}{\partial t} = 0$, and $\Gamma_- = \{\mathbf{x} \mid \mathbf{x} \in \Gamma, \ \mathbf{V}(\mathbf{x}) \cdot \mathbf{n}(\mathbf{x}) < 0\}$.

The solution of first order problems such as (5) has motivated a very large literature (see, e.g., [28] and the references therein). It seems thus that one has abundance of methods to solve the problem (5); this is definitely true, but things get complicated if one wishes to solve problem (4.1) using the same finite element velocity spaces that one employs for the solution of the problem (3). A conceptually elegant way to achieve that goal is to use the *backward method of characteristics* as done in, e.g., [52, 57, 58] via the so-called *Lagrange-Galerkin* methodology. Albeit conceptually simple the practical implementation of the Lagrange-Galerkin methods requires a lot of 'savoir faire' (see [58] for an evidence of the above statement). Fortunately, there exists a very simple alternative to the method of characteristics, based on a wave-like equation reformulation of the problem (5). We personally encountered this approach when investigating the wavelet solution (see [24]) of

$$\begin{cases} \dfrac{\partial u}{\partial t} + a\dfrac{\partial u}{\partial x} = 0 \quad \text{in } (0, L) \times (0, T), \\ u(0) = u_0, \\ u(0, t) = g(t), \quad t \in (0, T), \end{cases} \tag{6}$$

with: $0 < L < +\infty$, a a positive number, and $0 < T \leq +\infty$. If the functions u_0 and g are smooth enough, one can easily show (by time differentiation of the first equation in (6); see [12, 20] for details) that problem (6) has a unique solution which is also the unique solution of the following (genuine) wave equation problem:

$$\begin{cases} \dfrac{\partial^2 u}{\partial t^2} - a^2\dfrac{\partial^2 u}{\partial x^2} = 0 \quad \text{in } (0, L) \times (0, T), \\ u(0) = u_0, \quad \dfrac{\partial u}{\partial t}(0) = -a\dfrac{\partial u_0}{\partial x}, \\ u(0, t) = g(t), \quad \dfrac{\partial u}{\partial t}(L, t) + a\dfrac{\partial u}{\partial x}(L, t) = 0, \quad t \in (0, T); \end{cases} \tag{7}$$

the boundary condition at $x = L$ can be viewed as a *radiation condition*. Assuming that u is smooth enough, problem (7) has the following variational formulation:

$$\begin{cases} u(t) \in H^1(0, L), \quad u(0, t) = g(t), \quad t \in (0, T), \\ \displaystyle\int_0^L \frac{\partial^2 u}{\partial t^2} v\, dx + a^2 \int_0^L \frac{\partial u}{\partial x}(t)\frac{\partial v}{\partial x}\, dx + a\frac{\partial u}{\partial t}(L, t)v(L) = 0, \\ \forall v \in V_0, \ t \in (0, T), \\ u(0) = u_0, \quad \dfrac{\partial u}{\partial t}(0) = -a\dfrac{\partial u_0}{\partial x}, \end{cases} \tag{8}$$

with $V_0 = \{v \mid v \in H^1(0, L),\ v(0) = 0\}$. Thanks to variational formulation (8), problem (7), and therefore problem (6), can be solved by *finite element methods* of the Lagrange-Galerkin type (including those based on the Courant element, that is piecewise affine, globally continuous approximations) or (as done in [24]) by wavelet-Galerkin methods.

Remark 2 The approach we just advocated for the solution contradicts the popular approach which consists in writing second order in time differential equations as systems of first order ones. □

Actually, the strategy we just described for problem (6) can be easily generalized to problem (5) by observing that the properties $\mathbf{\nabla} \cdot \mathbf{V} = 0$ and $\frac{\partial \mathbf{V}}{\partial t} = \mathbf{0}$ imply that, after time differentiation, any smooth solution of problem (5) is solution of

$$\begin{cases} \dfrac{\partial^2 \phi}{\partial t^2} - \mathbf{\nabla} \cdot ((\mathbf{V} \cdot \mathbf{\nabla}\phi)\mathbf{V}) = 0, \quad \text{in } \Omega \times (t_0, t_f), \\ \phi(t_0) = \phi_0, \quad \dfrac{\partial \phi}{\partial t}(t_0) = -\mathbf{V} \cdot \mathbf{\nabla}\phi_0, \\ \phi = g \quad \text{on } \Gamma_- \times (t_0, t_f), \quad (\mathbf{V} \cdot \mathbf{n})\left(\dfrac{\partial \phi}{\partial t} + \mathbf{V} \cdot \mathbf{\nabla}\phi\right) = 0 \quad \text{on } \Gamma \setminus \Gamma_- \times (t_0, t_f), \end{cases} \tag{9}$$

a wave-like equation problem associated with the *hypo-elliptic* operator

$$\phi \to -\nabla \cdot ((\mathbf{V} \cdot \nabla\phi)\mathbf{V}).$$

Let us define the space V_0 by

$$V_0 = \{\theta \mid \theta \in H^1(\Omega), \quad \theta = 0 \text{ on } \Gamma_-\};$$

assuming that problem (9) has a smooth enough solution, using the *divergence theorem*, one can easily show that the above problem has the following *variational formulation* (with $d\mathbf{x} = dx_1 \dots dx_d$):

$$\begin{cases} \phi(t) \in H^1(\Omega), \quad \phi(t)|_{\Gamma_-} = g, \quad t \in (t_0, t_f), \\ \displaystyle\int_\Omega \frac{\partial^2\phi}{\partial t^2}\theta\, d\mathbf{x} + \int_\Omega (\mathbf{V}\cdot\nabla\phi)(\mathbf{V}\cdot\nabla\theta)\, d\mathbf{x} + \int_{\Gamma\setminus\Gamma_-} \mathbf{V}\cdot\mathbf{n}\frac{\partial\phi}{\partial t}\theta\, d\Gamma = 0, \\ \forall\theta \in V_0, \quad t \in (t_0, t_f), \\ \phi(t_0) = \phi_0, \quad \dfrac{\partial\phi}{\partial t}(t_0) = -\mathbf{V}\cdot\nabla\phi_0. \end{cases} \tag{10}$$

From (10), one can easily show that if problem (9) has a solution, it is unique. Formulation (10) is ideally suited to *Lagrange finite element approximations* as shown in, e.g., [12, 20], where it has been (successfully) applied in combination with the finite element spaces used for the numerical solution of the Stokes-like problem (3). Concerning the *time-discretization* of (10), we have been advocating the following *centered scheme* (with $\tau = \frac{t_f - t_0}{Q}$, the integer Q being > 1):

$$\phi^0 = \phi_0, \quad \phi^1 - \phi^{-1} = 2\tau\phi_1. \tag{11}$$

For $q = 0, 1, \dots, Q-1$, $\{\phi^{q-1}, \phi^q\} \to \phi^{q+1}$ as follows:

$$\begin{cases} \phi^{q+1} \in H^1(\Omega), \quad \phi^{q+1}|_{\Gamma_-} = g, \\ \displaystyle\int_\Omega \frac{\phi^{q+1} + \phi^{q-1} - 2\phi^q}{\tau^2}\theta\, d\mathbf{x} + \int_\Omega (\mathbf{V}\cdot\nabla\phi^q)(\mathbf{V}\cdot\nabla\theta)\, d\mathbf{x} \\ \quad + \displaystyle\int_{\Gamma\setminus\Gamma_-} \mathbf{V}\cdot\mathbf{n}\left(\frac{\phi^{q+1} - \phi^{q-1}}{2\tau}\right)\theta\, d\Gamma = 0, \quad \forall\theta \in V_0, \end{cases} \tag{12}$$

where, in (11), ϕ_1 is solution of the following variational problem:

$$\begin{cases} \phi_1 \in V_0, \\ \displaystyle\int_\Omega \phi_1\theta\, d\mathbf{x} = -\int_\Omega \mathbf{V}\cdot\nabla\phi_0\theta\, d\mathbf{x}, \quad \forall\theta \in V_0. \end{cases} \tag{13}$$

Strictly speaking, the infinite dimensional variational problems (12) and (13) make no sense, in general, unlike, fortunately, their finite dimensional analogues, obtained from (12) and (13) via appropriate finite element approximations (see [12, 20] for details).

Remark 3 As expected, the wave-like equation method described above is not a *stand-alone* one for the numerical solution of advection problems such (5), as shown by the numerical experiments reported in [20]. The reason for that unfortunate situation is easy to understand: the wave-like equation in (9) is a model for propagation in both the $\mathbf{V}$ and $-\mathbf{V}$ directions; with appropriate initial and boundary conditions, there is no 'signal' propagating in the $-\mathbf{V}$ direction. However these ideal circumstances do not hold *exactly* anymore after space-time discretization, explaining the existence of a small (if $\triangle x$ and $\triangle t$ are small) parasitic signal propagating in the $-\mathbf{V}$ direction. The good news are that when using the wave-like equation method to solve the incompressible Navier-Stokes equations, the advection step (4) is combined with the incompressible-viscous step (3), the solver of the problem (3) filtering (at least partially) those unwanted oscillations generated by the solver of the problem (9), (10).

Remark 4 When applying the wave-like equation method to solve the incompressible Navier equations via the Lie-scheme (2)–(4), we advocate taking $\tau = \triangle t/Q$, with $2 \leq Q \leq 5$, in the fully discrete analogue of scheme (11), (12).

Remark 5 To the best of our knowledge, the wave-like equation method for the solution of the incompressible Navier-Stokes equations has been introduced in [10]. Actually, a related method was introduced in 1979 by *Lynch and Gray* for the solution of the *shallow water equations* [44], the convergence of the method being discussed in [7, 8]. See also [60, 61] for the application of a closely related method to the solution of *multi-dimensional transport problems*. □

Since its introduction in 1997 the wave-like equation/operator-splitting method discussed above has been applied by the authors, their students, post-docs and other collaborators and scientists to a large variety of viscous-flow problems, some more complicated than problem (1). Let us mention among others: (i) The numerical simulation of *particulate flow* (see, e.g., [20, 26, 31]). (ii) The numerical solution of the *Boussinesq system* coupling the *Navier-Stokes* and *heat equations*, and modelling *natural convection* [20]. (iii) The simulation of *visco-plastic flow* [11, 20, 27]. (iv) The simulation of *visco-elastic flow* (possibly with particles) [20, 48, 49]. (v) The solution of *free boundary problems* for incompressible viscous flow [22]. (vi) The numerical solution of the system coupling the *Cahn-Hilliard* and *Navier-Stokes equations* and modelling the flow of multiple immiscible incompressible viscous fluids [30]. The references in the above publications are also worth consulting.

Other examples and further references will be given in Sects. 3–5.

To conclude this introductory section we cannot resist mentioning the fact that some of the results from [23], concerning operator-splitting/wave-like equation based simulations of wall-driven incompressible viscous flows in a *semi-circular cavity*, have been used in [59] to validate a *NURBS* (for *Non Uniform Rational Bézier Splines*) based Navier-Stokes solver.

3 On the Simulation of 3-D Incompressible Viscous Flow in a Cube with a Moving Wall

3.1 Generalities

Starting with [18], the wall-driven square cavity flow problem has been for decades the most popular problem used to validate and compare incompressible Navier-Stokes solvers. No surprisingly, it has been used by the two authors and their collaborators to validate the operator-splitting/wave like equation method briefly discussed in Sect. 1, some of the results of the related simulations being reported in [12, 20]. Actually, one has also reported in ([20], Chap. 9) and [47] the results of the simulation of a Newtonian incompressible viscous flow in a cubic cavity when one of the walls is sliding; the maximal Reynolds number (Re) considered in [20, 47] is 10^3. More recently the oscillatory instability of cubic lid-driven cavity flows has been studied in [1, 16, 41]. Numerically, Feldman and Gelfgat [16] obtained that the critical Reynolds number for the transition from a steady flow to an oscillatory one (a Hopf bifurcation) is at $\mathrm{Re}_{cr} = 1914$. Anupindi et al. [1] reported that the critical Re they observed is $\mathrm{Re}_{cr} = 2300$, which was obtained using regularized boundary condition. Experimentally, Liberzon et al. [41] reported that the critical Reynolds number is in the range [1700, 1970], One of our goals in this section is to report on the results we obtained when taking Re beyond 10^3, and to identify as accurately as possible the value of Re at which a Hopf bifurcation does occur.

3.2 Numerical Methods

To speed up the numerical solution of the cubic lid-driven cavity flow problem, we time-discretized the related problem (1), using a three stage operator-splitting scheme, namely: (i) using a L^2-projection Stokes solver à la Uzawa to force the incompressibility condition, (ii) an advection step similar to (4), and (iii) a diffusion step. The resulting scheme reads as follows:

$$\mathbf{u}^0 = \mathbf{u}_0. \tag{14}$$

For $n \geq 0, \mathbf{u}^n \rightarrow \{\mathbf{u}^{n+1/3}, p^{n+1}\} \rightarrow \mathbf{u}^{n+2/3} \rightarrow \mathbf{u}^{n+1}$ via the solution of:

$$\begin{cases} \rho \dfrac{\mathbf{u}^{n+1/3} - \mathbf{u}^n}{\triangle t} + \nabla p^{n+1} = \mathbf{0} & \text{in } \Omega, \\ \nabla \cdot \mathbf{u}^{n+1/3} = 0 & \text{in } \Omega, \\ \mathbf{u}^{n+1/3} \cdot \mathbf{n} = 0 & \text{on } \Gamma, \end{cases} \tag{15}$$

$$\begin{cases} \dfrac{\partial \mathbf{w}}{\partial t} + (\mathbf{u}^{n+1/3} \cdot \nabla)\mathbf{w} = \mathbf{0} & \text{in } \Omega \times (t^n, t^{n+1}), \\ \mathbf{w}(t^n) = \mathbf{u}^{n+1/3}, \\ \mathbf{w}(t) = \mathbf{u}^{n+1/3}(= \mathbf{u}_B(t^{n+1})) & \text{on } \Gamma_-^{n+1} \times (t^n, t^{n+1}), \end{cases} \tag{16.1}$$

$$\mathbf{u}^{n+2/3} = \mathbf{w}(t^{n+1}), \tag{16.2}$$

$$\begin{cases} \rho \dfrac{\mathbf{u}^{n+1} - \mathbf{u}^{n+2/3}}{\triangle t} - \mu \nabla^2 \mathbf{u}^{n+1} = \mathbf{f}^{n+1} & \text{in } \Omega, \\ \mathbf{u}^{n+1} = \mathbf{u}_B(t^{n+1}) & \text{on } \Gamma. \end{cases} \tag{17}$$

Two simplifications take place for the lid-driven cavity flow problem considered here: namely, $\mathbf{f}^{n+1} = \mathbf{0}$ and $\Gamma_-^{n+1} = \emptyset$. For the space discretization, we have used, as in ([20], Chap. 5) and [3], a P_1-*iso*-P_2 (resp., P_1) finite element method for the approximation of the velocity field (resp., pressure), defined from uniform "tetrahedral" meshes $\mathcal{T}_h$ (resp., $\mathcal{T}_{2h}$). Problem (15) is reminiscent of those encountered when applying Chorin's projection method [9].

3.3 Numerical Results

For the lid-driven cavity flow problem in a cube, considered here, we took $\Omega = (0, 1)^3$ as computational domain and defined the Dirichlet data $\mathbf{u}_B$ by

$$\mathbf{u}_B(\mathbf{x}) = \begin{cases} (1, 0, 0)^T & \text{on } \{\mathbf{x} \mid \mathbf{x} = (x_1, x_2, 1)^T,\ 0 < x_1,\ x_2 < 1\}, \\ \mathbf{0} & \text{elsewhere on } \Gamma. \end{cases} \tag{18}$$

We considered that the steady state has been reached when the change between two consecutive time step in the simulation, $\|\mathbf{u}_h^n - \mathbf{u}_h^{n-1}\|_\infty / \triangle t$, is less than 10^{-4}, and then took $\mathbf{u}_h^n$ as the steady state solution.

To validate the numerical methodology, we have considered for the velocity mesh size the values $h = 1/60$ and 1/96 associated with the time step $\triangle t = 0.001$. For $\mathrm{Re} = 400$ and 1000, the results reported in Fig. 1 show a very good agreement with those obtained in [5, 17, 39]. The velocity vectors of the steady flows obtained for $\mathrm{Re} = 400$ and 1000 are shown in Fig. 2. Those velocity field vectors are projected orthogonally to the three planes, $x_2 = 0.5$, $x_1 = 0.5$, and $x_3 = 0.5$, and the length of the vectors has been doubled in the two later planes to observe the flow more clearly.

To study the transition from steady flow to oscillatory flow, we have analyzed the history of the L^2-norm $\|\mathbf{u}_h^n\|$ of the flow field for different values of Re and of the mesh size h. For $h = 1/60$, the flow field evolves to a steady state for $\mathrm{Re} \leq 1860$ and the amplitude of the oscillation of the flow field L^2-norm decreases also in time. At $\mathrm{Re} = 1865$, the steady state criterion is not satisfied and the amplitude of the oscillation increases in time as in Fig. 3. Thus we conclude that the critical Reynolds number Re_{cr}

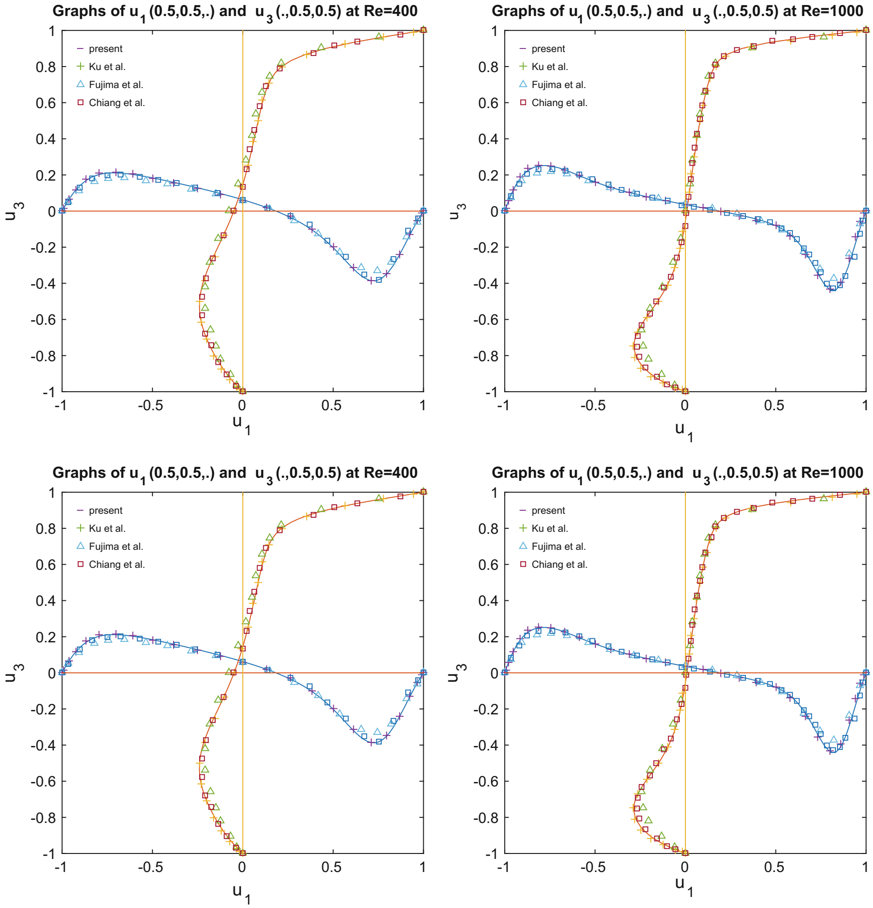


Fig. 1 Comparisons of the numerical results obtained for $h = 1/60$ (top) and $1/96$ (bottom) at $\text{Re} = 400$ (left) and 1000 (right)

for the occurrence of the transition is somewhere between 1860 and 1865. Applying a similar analysis, we obtain that, for $h = 1/96$, Re_{cr} is in (1870, 1875), the histories of the velocity L^2-norm being shown in Fig. 3. The oscillation frequencies of the velocity L^2-norm obtained for $h = 1/60$ and $\triangle t = 1/1000$ are about 0.5937 and 0.5941 for $\text{Re} = 1860$ and 1865, respectively. Those obtained for $h = 1/96$ and $\triangle t = 1/1000$ are about 0.5978 and 0.5973 for $\text{Re} = 1870$ and 1875, respectively.

A documented feature of three-dimensional lid-driven cavity flows, like those considered in this section, is that they may exhibit Taylor-Görtler-like (TGL) vortices if Re is sufficiently large. Indeed, Iwatsu, Hyun and Kuwahara reported (in [34]) such vortices at $\text{Re} = 2000$ for cubic cavity flows similar to those considered in this section.

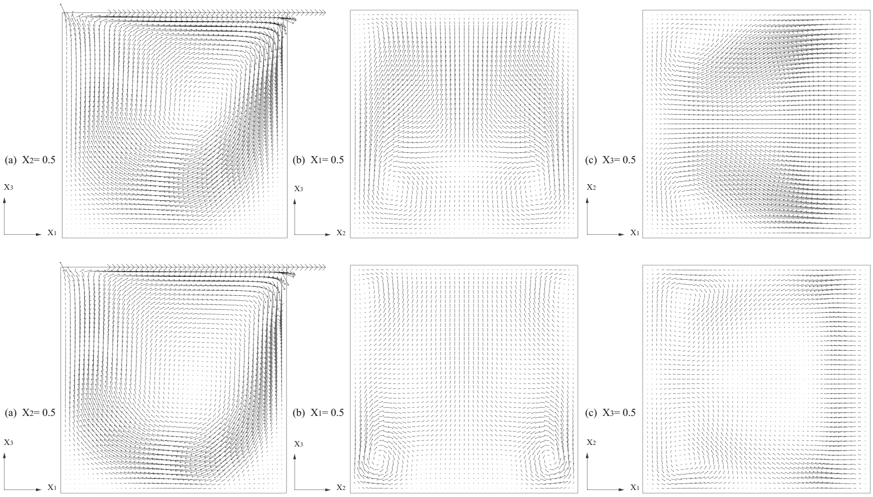


Fig. 2 Steady flow velocity vector of steady flow for Re $= 400$ (top) and 1000 (bottom) projected on the planes $x_2 = 0.5$ (left), $x_1 = 0.5$ (middle), and $x_3 = 0.5$ (right)

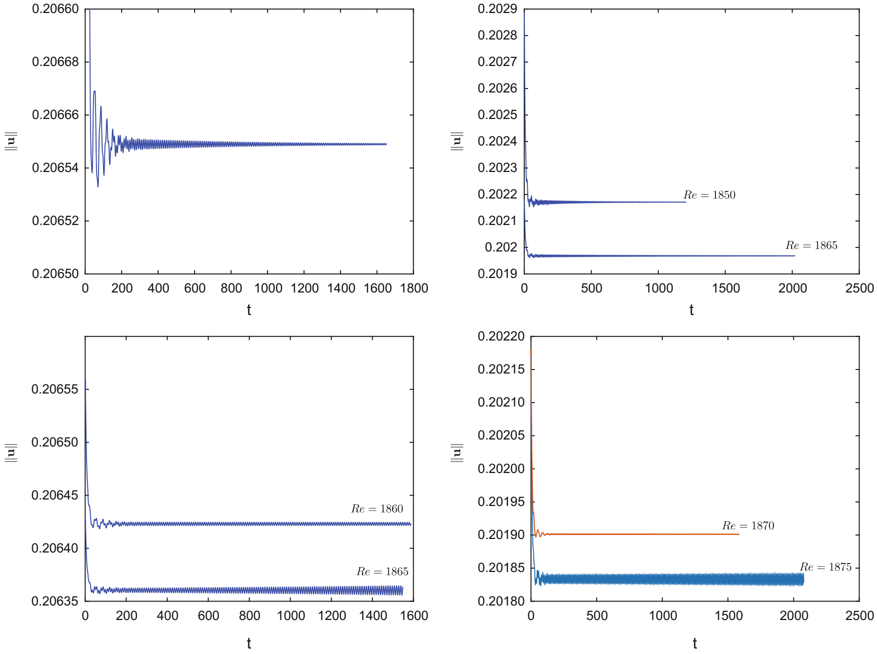


Fig. 3 Histories of the flow field L^2-norm for $h = 1/60$ (left) and 1/96 (right): **a** Re $=$ 1850 (top left), Re $=$ 1860 and 1865 (bottom left); **b** Re $=$ 1850 and 1865 (top right), Re $=$ 1870 and 1875 (bottom right)

Also, as predicted in [16, 41] (and confirmed by our own simulations), a transition from steady flow to oscillatory flow (Hopf bifurcation) occurs at $\mathrm{Re}_{cr} < 2000$. On the other hand, using a global linear stability analysis, Gianetti et al. [19] found that cubic lid-driven cavity flow becomes unstable for Re just above 2000. All these results (ours in particular) lead us to suspect that the Hopf bifurcation is connected to the TGL vortices at Re slightly below 2000.

The bottom left picture of Fig. 3 shows oscillatory regimes at $\mathrm{Re} = 1860$ and 1865, for the flow computed with $h = 1/60$. In order to study the computed flow distortion we have visualized in Figs. 4 (for $\mathrm{Re} = 1860$) and 5 (for $\mathrm{Re} = 1865$) the velocity fields associated with the peak and bottom of the velocity field L^2-norm, and the vector field obtained by difference of the above two velocity fields. The top (resp., bottom) pictures have been obtained by projection of the vector fields on the plane $x_1 = 34/60$ (resp., $x_3 = 1/2$). Figures 4 and 5 show no evidence of TGL vortices for the velocity fields computed with $h = 1/60$ at $\mathrm{Re} = 1860$ and 1865; however, the pictures on the right of Figs. 4 and 5, obtained by the vector field difference detailed above, show a pair of vortices reminiscent of the GTL ones, but with much smaller magnitude since the vector fields have been amplified by a factor of 200 (resp., 50) for $\mathrm{Re} = 1860$

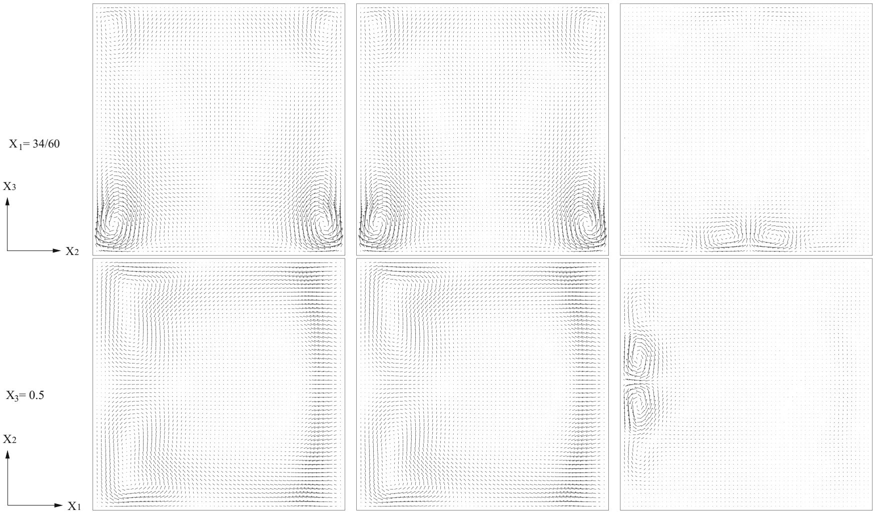


Fig. 4 Left and middle: Projections (at $\mathrm{Re} = 1860$) of the cavity flow velocity vector fields associated with the peak (left) and bottom (middle) of the velocity L^2-norm during an oscillation. Right: Projections (at $\mathrm{Re} = 1860$) of the vector field obtained by difference of the velocity vector fields associated with the peak and bottom of the velocity L^2-norm. All the vector fields are projected on the planes $x_1 = 34/60$ (top) and $x_3 = 0.5$ (bottom). The vector scale for the field obtained by difference (right) is 200 times that of the actual one, while the scale for the two other fields (left and middle) is twice that of the actual one

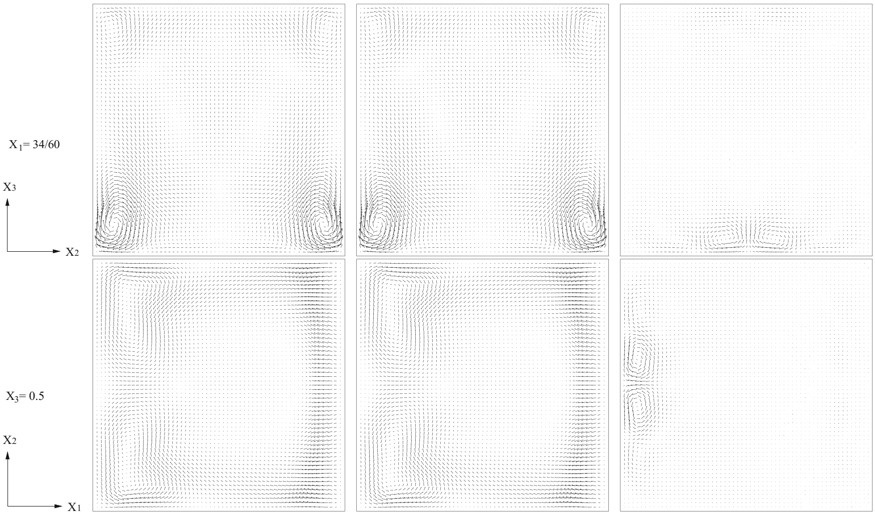


Fig. 5 Left and middle: Projections (at Re = 1865) of the cavity flow velocity vector fields associated with the peak (left) and bottom (middle) of the velocity L^2-norm during an oscillation. Right: Projections (at Re = 1865) of the vector field obtained by difference of the velocity vector fields associated with the peak and bottom of the velocity L^2-norm. All the vector fields are projected on the planes $x_1 = 34/60$ (top) and $x_3 = 0.5$ (bottom). The vector scale for the field obtained by difference (right) is 50 times that of the actual one, while the scale for the two other fields (left and middle) is twice that of the actual one

(resp., 1865) in order to make them visible. On the other hand, at Re = 1875, a pair of TGL vortices becomes visible as shown by Fig. 6 where we have visualized (using a nonlinear scaling to enhance visibility) several snap-shots of the velocity field during an oscillation time period. This pair of TGL vortices is not stationary, however, it remains symmetric with respect to the the mid-plane $x_2 = 1/2$. Figure 6 shows that two tertiary vortices are formed on the left and right parts of the bottom wall, near the large corner vortices at $t = 1526$, 1527 and 1528; next, these tertiary vortices move toward the symmetry plane $x_2 = 0.5$ at $t = 1529$, a pair of TGL vortices being formed in the time interval $[1531, 1533]$; finally, the TGL vortices disappear after $t = 1533$, to reappear during the next time-period. We have reported on Fig. 7 the projection on the plane $x_1 = 33/60$ of the vector field obtained by difference of the velocity flow fields at $t = 1525$ and $t = 1527$. The vortex pair we observe is reminiscent of those visualized on the right of Figs. 4 and 5. This vortex pair keeps hiding there and becomes stronger as Re increases. These results suggest that the TGL vortices observed for Re slightly below 2000 are related to the onset of an oscillatory flow.

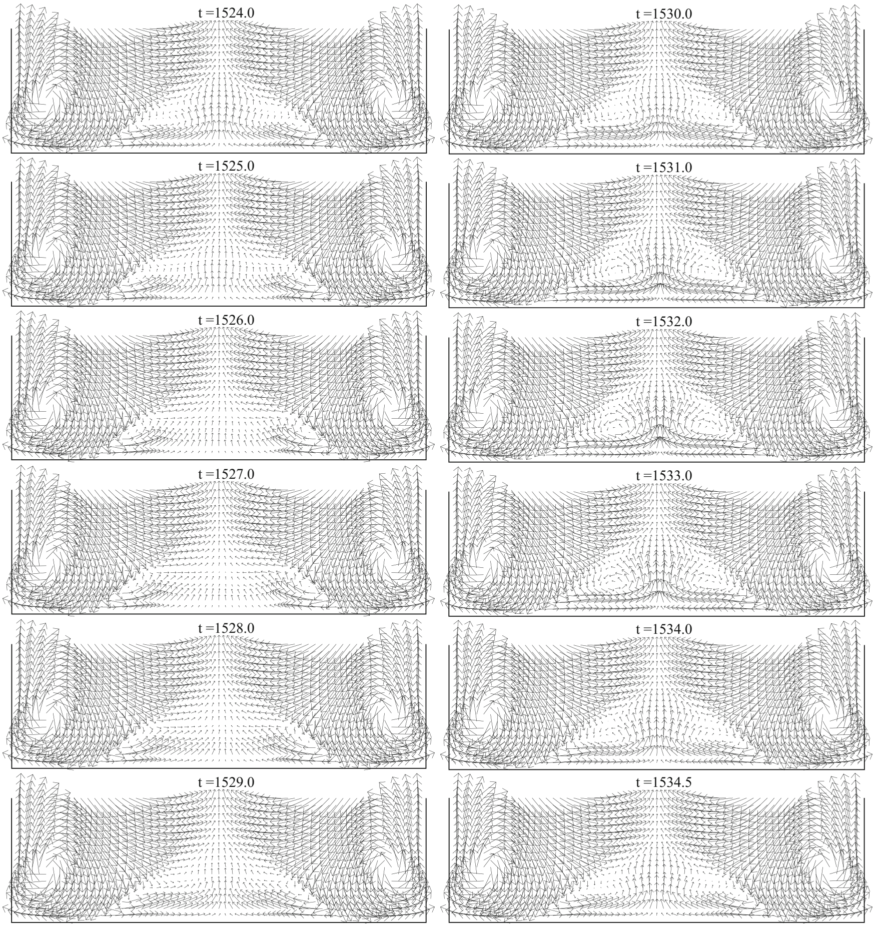


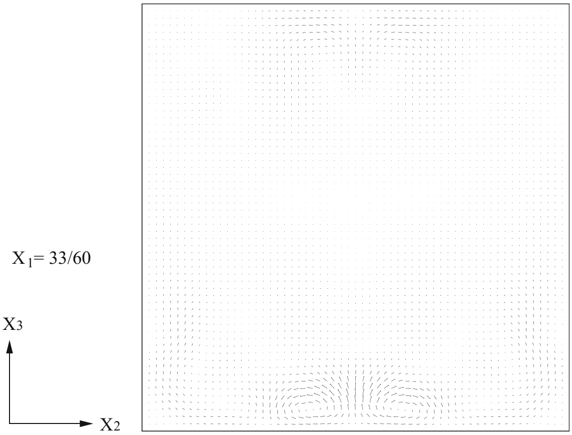
Fig. 6 Projected velocity vector field of the cavity flow at $\mathrm{Re} = 1875$ on the plane $x_1 = 33/60$ at different instants of time during one oscillation of the flow field L^2-norm from $t = 1524$ to 1534.575 [for enhancing the visibility of the TGL vortices we proceeded as follows: (i) for those projected vectors of length ≤ 0.02 the vector scale is 15 times that of the actual one and (ii) for those projected vectors of length >0.02, the length is reduced to 0.02 first and then plotted as in (i)]

4 Particulate Flow: The Orientation of a Neutrally Buoyant Prolate Ellipsoid in a Three-Dimensional Poiseuille Flow

4.1 Generalities

The distributed Lagrange multiplier/fictitious domain (DLM/FD) formulation for particulate flow, and its associated numerical methodologies based on the Lie scheme have been developed in the past 20 years (see, e.g., [20], Chaps. 8 and 9, [22, 25, 26]). It

Fig. 7 Projected velocity vector field of the difference of the velocity fields at $t = 1525$ and $t = 1527$ on the plane $x_1 = 33/60$ for $\mathrm{Re} = 1875$. The vector scale is 20 times that of the actual one



is the (necessarily biased) opinion of the authors of this article that the *direct numerical simulation of particulate flow* has been one of the success stories of the wave-like equation-based methodology.

The motion of particles in a channel is relevant to a variety of applications in many chemical engineering and biological processes, such as suspension process, sedimentation, blood flow, and flow cytometry. Understanding this kind of motion has become even more important with the recent advent of microfluidic devices used for many cell-based assays (see, e.g., [33]). The study of the motion of non-spherical particles in viscous fluids has a long history. Jeffery [35] solved the motion of a free ellipsoid for various types of unbounded shear flow under Stokes flow conditions. He concluded that the final state of a spheroid depends on its initial orientation and corresponds to the minimal energy dissipation. The experiments of Segré and Silberberg [54, 55] have had a large influence on fluid mechanics studies of migration and lift of particles. These autohrs studied the migration of dilute suspensions of neutrally buoyant spheres in a tube flow. The particles migrate away from the wall and centerline and accumulate at about 0.6 of the tube radius from the centerline. Karnis et al. [37] verified the same phenomenon and observed, in contrary to Jeffery's theory, that the inertial effect migrates non-spherical particles to a final equilibrium distance in the tube at which the long axis of a rod-like particle rotates within the plane passing through the central axis of the tube and the mass center of the particle; but a disk-like particle will rotates with its short axis perpendicular to the plane passing through the central axis of the tube and the mass center of the disk. In [46], similar migration and rotational behaviors of a neutrally buoyant ellipsoid were obtained at particle Reynolds numbers up to 52; and it was also found that this ellipsoid rotation exhibits distinctive states depending on the Reynolds number range and on the particle shape. In this section, we have further studied the orientation of a prolate ellipsoid in a three-dimensional Poiseuille flow.

4.2 A Fictitious Domain Formulation of the Model Problem

All the fluid-solid interactions to be considered in this article concern the flow of fluid-solid particle mixtures in a cylindrical tube (denoted by $\mathbf{T}$ in the sequel) with a circular cross-section. In order to take a full advantage of the fictitious domain approach we will embed $\mathbf{T}$ in a cylindrical tube (denoted by Ω) with a square cross-section whose edge length is slightly larger than the diameter of the $\mathbf{T}$ cross-section.

We will start our discussion with a one particle situation. Therefore, let $\Omega \subset \mathbb{R}^3$ be a rectangular parallelepiped. We suppose that Ω is filled with a *Newtonian incompressible viscous fluid* (of *density* ρ_f and *viscosity* μ_f) and that it contains a moving neutrally buoyant rigid particle B centered at $\mathbf{G} = \{G_1, G_2, G_3\}^t$ of *density* ρ_f, as shown in Fig. 8, which shows also the inclusion in Ω of the cylinder $\mathbf{T}$ mentioned above; we suppose that the central axis of both cylinders is parallel to the x_3-axis. The flow is modeled by the *Navier-Stokes equations* while the particle motion is described by the *Euler-Newton equations*. We introduce (with $d\mathbf{x} = dx_1dx_2dx_3$) the following functional spaces:

$$\begin{aligned} W_{0,P} = \{\mathbf{v} \mid \mathbf{v} \in (H^1(\Omega))^3,\ \mathbf{v} = \mathbf{0} \text{ on the top, bottom, front, and back of } \Omega \text{ and}\\ \mathbf{v} \text{ is periodic in the } x_3 \text{ direction}\},\end{aligned}$$

$$L_0^2 = \{q \mid q \in L^2(\Omega), \int_\Omega q\, d\mathbf{x} = 0\},$$

$$\Lambda_0(t) = \{\boldsymbol{\mu} \mid \boldsymbol{\mu} \in (H^1(B(t)))^3, \langle \boldsymbol{\mu}, \mathbf{e}_i \rangle_{B(t)} = 0, \langle \boldsymbol{\mu}, \mathbf{e}_i \times \overrightarrow{\mathbf{Gx}} \rangle_{B(t)} = 0,\ i = 1, 2, 3\},$$

$$\Lambda_T = \{\boldsymbol{\mu} \mid \boldsymbol{\mu} \in (H^1(\Omega \setminus \overline{\mathbf{T}}))^3,\ \boldsymbol{\mu} \text{ is periodic in the } x_3 \text{ direction}\},$$

where $\mathbf{e}_1 = \{1, 0, 0\}^t$, $\mathbf{e}_2 = \{0, 1, 0\}^t$, $\mathbf{e}_3 = \{0, 0, 1\}^t$, and $\langle \cdot, \cdot \rangle_{B(t)}$ (resp., $\langle \cdot, \cdot \rangle_T$) is an inner product on $\Lambda_0(t)$ (resp., Λ_T) (see [26], Sect. 5 and [20], Chap. 8) for further information on the choice of $\langle \cdot, \cdot \rangle_{B(t)}$). Above, and from now on, periodicity in the x_3 direction means periodicity of period L, L being the common length of the truncated cylinders Ω and $\mathbf{T}$. Then, the distributed Lagrange multiplier based fictitious domain

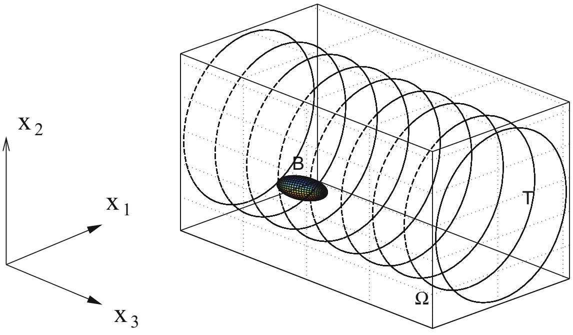


Fig. 8 An example of three-dimensional flow region with one rigid body

formulation for the flow around a freely moving neutrally buoyant particle of general shape inside a cylindrical tube reads as follows (see [20, 50] for a detailed discussion of the non-neutrally buoyant case):

For a.e. $t > 0$, find $\mathbf{u}(t) \in W_{0,P}$, $p(t) \in L_0^2$, $\mathbf{V_G}(t) \in \mathbb{R}^3$, $\mathbf{G}(t) \in \mathbb{R}^3$, $\boldsymbol{\omega}(t) \in \mathbb{R}^3$, $\boldsymbol{\lambda}(t) \in \Lambda_0(t)$, $\boldsymbol{\lambda}_T \in \Lambda_T$ *such that*

$$\begin{cases} \rho_f \displaystyle\int_\Omega \left[\frac{\partial \mathbf{u}}{\partial t} + (\mathbf{u} \cdot \nabla)\mathbf{u}\right] \cdot \mathbf{v}\, d\mathbf{x} + 2\mu_f \int_\Omega \mathbf{D}(\mathbf{u}) : \mathbf{D}(\mathbf{v})\, d\mathbf{x} - \int_\Omega p \nabla \cdot \mathbf{v}\, d\mathbf{x} \\ -\langle \boldsymbol{\lambda}, \mathbf{v} \rangle_{B(t)} - \langle \boldsymbol{\lambda}_T, \mathbf{v} \rangle_T = \rho_f \displaystyle\int_\Omega \mathbf{g} \cdot \mathbf{v}\, d\mathbf{x} + \int_\Omega \mathbf{f} \cdot \mathbf{v}\, d\mathbf{x}, \quad \forall \mathbf{v} \in W_{0,P}, \end{cases} \tag{19}$$

$$\int_\Omega q \nabla \cdot \mathbf{u}(t) d\mathbf{x} = 0, \quad \forall q \in L^2(\Omega), \tag{20}$$

$$\langle \boldsymbol{\mu}, \mathbf{u}(t) \rangle_{B(t)} = 0, \quad \forall \boldsymbol{\mu} \in \Lambda_0(t), \tag{21}$$

$$\langle \boldsymbol{\mu}_T, \mathbf{u}(t) \rangle_T = 0, \quad \forall \boldsymbol{\mu}_T \in \Lambda_T, \tag{22}$$

$$\frac{d\mathbf{G}}{dt} = \mathbf{V_G}, \tag{23}$$

$$\frac{d\mathbf{x}_i}{dt} = \mathbf{V_G} + \boldsymbol{\omega} \times \overrightarrow{\mathbf{G}\mathbf{x}_i}, \quad i = 1, 2, \tag{24}$$

$$\mathbf{V_G}(0) = \mathbf{V}_\mathbf{G}^0,\ \boldsymbol{\omega}(0) = \boldsymbol{\omega}^0,\ \mathbf{G}(0) = \mathbf{G}^0 = \{G_1^0, G_2^0, G_3^0\}^t,\ \mathbf{x}_i(0) = \mathbf{x}_i^0,\ i = 1, 2, \tag{25}$$

$$\mathbf{u}(\mathbf{x}, 0) = \overline{\mathbf{u}}_0(\mathbf{x}) = \begin{cases} \mathbf{u}_0(\mathbf{x}), & \forall \mathbf{x} \in \Omega \setminus \overline{B(0)}, \\ \mathbf{V}_\mathbf{G}^0 + \boldsymbol{\omega}^0 \times \overrightarrow{\mathbf{G}^0\mathbf{x}}, & \forall \mathbf{x} \in \overline{B(0)}. \end{cases} \tag{26}$$

In (19)–(26) $\mathbf{u}$ and p denote *velocity* and *pressure*, respectively, $\boldsymbol{\lambda}$ is a *Lagrange multiplier* associated with relation (21) (from (21) the fluid has a rigid body motion in the region occupied by $B(t)$), $\boldsymbol{\lambda}_T$ is a *Lagrange multiplier* associated with relation (22) (from (22), the fluid velocity is $\mathbf{0}$ in $\bar{\Omega} \setminus \mathbf{T}$), $\mathbf{D}(\mathbf{v}) = \frac{1}{2}(\nabla \mathbf{v} + (\nabla \mathbf{v})^t)$, $\mathbf{g}$ denotes *gravity*, $\mathbf{f}$ is an imposed *pressure gradient* pointing in the x_3-direction inside the cylinder $\mathbf{T}$, $\mathbf{V_G}$ is the *translation velocity* of the particle B, and $\boldsymbol{\omega}$ is the *angular velocity* of B. We suppose that the *no-slip* condition holds on ∂B. We also use, if necessary, the notation $\phi(t)$ for the function $\mathbf{x} \rightarrow \phi(\mathbf{x}, t)$.

Remark 6 The hydrodynamical forces and torque imposed on the rigid body by the fluid are built in (19)–(26) implicitly (see [26] for details), thus we do not need to compute them explicitly in the simulation. Since in (19)–(26) the flow field is defined on the entire domain Ω, it can be computed with a simple structured grid.

Remark 7 In (21), the rigid body motion in the region occupied by the particle is enforced via Lagrange multipliers $\boldsymbol{\lambda}$. To recover the translation velocity $\mathbf{V_G}(t)$ and the angular velocity $\boldsymbol{\omega}(t)$ from $\mathbf{u}(t)$ satisfying (21), we solve the following equations:

$$\begin{cases} \langle \mathbf{e}_i, \mathbf{u}(t) - \mathbf{V_G}(t) - \boldsymbol{\omega}(t) \times \overrightarrow{\mathbf{Gx}} \rangle_{B(t)} = 0, & \text{for } i = 1, 2, 3, \\ \langle \mathbf{e}_i \times \overrightarrow{\mathbf{Gx}}, \mathbf{u}(t) - \mathbf{V_G}(t) - \boldsymbol{\omega}(t) \times \overrightarrow{\mathbf{Gx}} \rangle_{B(t)} = 0, & \text{for } i = 1, 2, 3. \end{cases} \tag{27}$$

Remark 8 In (24), we have to track the motion of two extra points attached to any particle of general shape so that we can determine the region occupied by the particle via its center of mass, the translation velocity of the center of mass and the angular velocity of the particle. In practice we shall track two orthogonal normalized vectors rigidly attached to the body B from the center of mass $\mathbf{G}$.

Remark 9 In (19), $2\int_\Omega \mathbf{D}(\mathbf{u}) : \mathbf{D}(\mathbf{v})\, d\mathbf{x}$ can be replaced by $\int_\Omega \nabla\mathbf{u} : \nabla\mathbf{v}\, d\mathbf{x}$ since $\mathbf{u}$ is divergence free and in $W_{0,P}$. This change can make the computation simpler and faster. Also the gravity $\mathbf{g}$ in (19) can be absorbed into the pressure term.

Remark 10 The details of numerical methodologies for simulating the motion of prolate and oblate spheroids are given in [46]. Applying Lie's scheme to (19)–(26), we have a six stage operator-splitting scheme, namely: (i) using a L^2-projection Stokes solver à la Uzawa to force the incompressibility condition, (ii) an advection step similar to (4), (iii) a diffusion step with the body force $\mathbf{f}$ and the enforcement of zero velocity outside the cylinder $\mathbf{T}$, (iv) a step to predict the particle position and its orientation, (v) a step to enforce the rigid body motion inside the particle and to obtain its updated translation and angular velocity, and (vi) a step to correct the particle position and its orientation. For the space discretization, we have still used a P_1-*iso*-P_2 (resp., P_1) finite element approximation of the velocity field (resp., pressure) defined from uniform "tetrahedral" meshes $\mathcal{T}_h$ (resp., $\mathcal{T}_{2h}$). For the enforcement of the rigid body motion and zero velocity outside the cylinder, we have applied a collocation method (see [46] for details).

4.3 Numerical Results

For the first series of test problems, we have considered the simulation of a neutrally buoyant prolate ellipsoid moving in a fluid filled cylinder (see Fig. 8). We take $\Omega = (0, 1 + 4h) \times (0, 1 + 4h) \times (0, 2)$ as computational domain with h as the space mesh size to construct the flow velocity spaces. The radius R of the cylinder $\mathbf{T}$ is 0.5 and its length is 2. The semi-long axis of the prolate ellipsoid is 0.195 and its two semi-short axes are 0.065, while the common value of the densities of the fluid and particle is 1. The viscosity of the fluid is $\mu_f = 0.5, 0.1$, or 0.05. The force $\mathbf{f}$ in (19) is a constant vector, positively oriented in the Ox_3 direction; $\|\mathbf{f}\|$ has been chosen so that the maximum velocity of the corresponding *Poiseuille flow* (without particle) is 10. We suppose that the prolate ellipsoid is at rest initially and that the initial fluid velocity corresponds to the one of a fully developed Poiseuille flow of maximal velocity 10. Thus the Reynolds numbers based on the diameter of the cylinder are Re $= 10/\mu_f = 20$, 100, and 200, respectively. The initial mass center $\mathbf{G}(0)$ of the ellipsoid is vertically located below the cylinder axis at a distance 0.4 to this axis and the long-axis of the ellipsoid lies on the plane parallel to the x_1x_3-coordinate plane. The initial angle between the long axis and the direction of the x_1-axis has been chosen as 0°, 30°, 60°, or 90°. The one in Fig. 8 corresponds to the case of 90°. We have used uniform tetrahedral meshes to approximate velocity and pressure. The velocity (resp., pressure) mesh size is $h = 1/80$ (resp., $h_p = 2h$), while the time discretization step is $\triangle t = 0.001$.

For all the cases with $\mu_f = 0.5$, the prolate ellipsoid has a tumbling behavior after migrating away from the wall of the cylindrical tube and reaching its equilibrium distance to the central axis of the tube. Its long axis rotates on the plane passing through the cylinder axis and its mass center (e.g., see Fig. 9). The average equilibrium distances of the mass center to the central axis of the tube are 0.5368 R, 0.5396 R, 0.5398 R, and 0.5352 R for the initial angles 0°, 30°, 60°, and 90°, respectively. The particle Reynolds

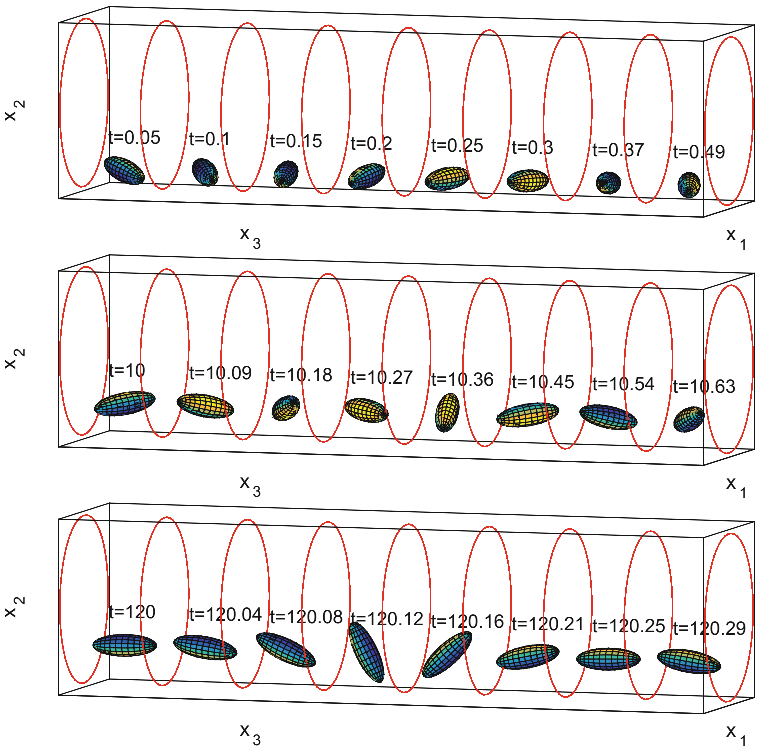


Fig. 9 Visualization of the prolate orientation change from its initial orientation to the rotation with respect to the short axis (tumbling) while reaching its equilibrium distance to the cylinder central axis ($\mu_f = 0.5$, initial angle = 60°)

numbers based on the length of the long axis and the average translation velocity are about 5.4. For all four different initial orientations, the long axis tumbles after it has reached the equilibrium distance to the tube central axis while the center of mass moves along a straight line parallel to the x_3-axis. This behavior is similar to the experimental results of the rod-like particle moving and rotating in the Poiseuille flow reported in [37].

For the cases of $\mu_f = 0.1$, the prolate ellipsoid has two different rotational behaviors after reaching its equilibrium distance to the tube central axis. With the initial angle of 0°, 30° and 60°, the prolate ellipsoid is rotating with respect to its long axis, which is perpendicular to the plane passing through the central axis of the tube and its mass center (see Fig. 10). This motion was not reported in the 1964 paper by Karnis et al. [37], but since this behavior persists after decreasing h and $\triangle t$, the authors strongly believe that it is not a numerical artifact. The average distances of the mass center to the central axis of the tube for both initial angles 0°, 30° and 60° are about $0.519\,R$ for $290 \leq t \leq 300$. Once the center of mass has reached the equilibrium distance to the tube central axis, the ellipsoid does not tumble but rotates with respect to its long axis as

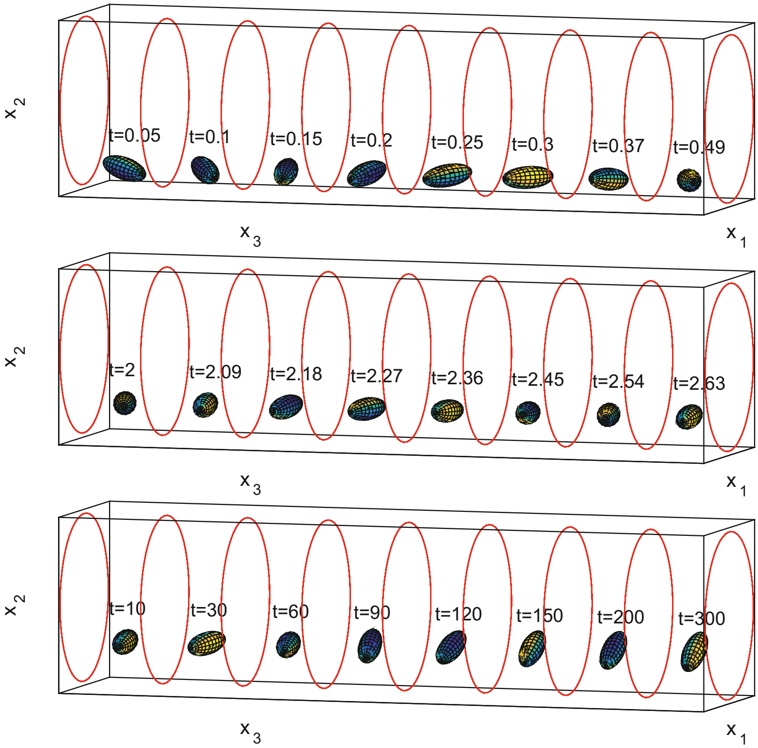


Fig. 10 Visualization of the prolate orientation change from its initial orientation to the rotation with respect to the long axis while reaching its equilibrium distance to the cylinder central axis ($\mu_f = 0.1$, initial angle = 60°)

shown in Fig. 10. The particle Reynolds numbers based on the length of the long-axis and the average translation velocity for $290 \leq t \leq 300$ are about 26.23. For the case of the initial angle equal to 90° (as in Fig. 8), the prolate ellipsoid tumbles just like it does when $\mu_f = 0.5$. For $215 \leq t \leq 225$, the average distance of the mass center to the central axis of the tube is $0.5456\,R$ and the particle Reynolds number is 26.25. The co-existence of two different rotating behaviors at about the same range of Reynolds number is quite unusual. For the two initial angles of 0° and 90°, we have also placed the initial mass center vertically below the cylinder axis at a distance 0.252 to this axis, which is much closer to the cylinder central axis. In both situations, the prolate spheroid migrates away from the cylinder central axis and the rotational motions are eventually the same as those one obtains when the ellipsoid is placed initially closer to the tube boundary.

When decreasing the viscosity to 0.05 and keeping all other parameters the same, we have obtained that, after reaching its equilibrium distance, the prolate spheroid does not tumble but rotate with respect to its long axis for all four different initial angles. But the ellipsoid placed vertically below the cylinder axis at a distance 0.252 to this axis with the initial angles 0° and 90° behaves like it does when $\mu_f = 0.1$.

Thus besides the Reynolds number, the initial distance to the cylinder central axis does matter too. In the near future, we will further study the effect of the initial position and the range of Reynolds number leading to two rotational behaviors.

5 Visco-Elastic Particulate Flow

The motion of particles in non-Newtonian fluids is not only of fundamental theoretical interest, but is also of importance in many applications to industrial processes involving particle-laden materials (see, e.g., [4, 45]). For example, during the hydraulic fracturing operation used in oil and gas wells, suspensions of solid particles in polymeric solutions are pumped into hydraulically-induced fractures. The particles must prop these channels open to enhance the rate of oil recovery [13].

Although numerical methods for simulating particulate flows in Newtonian fluids have been very successful, numerically simulating particulate flows in viscoelastic fluids is much more complicated and challenging. One of the difficulties (e.g., see [2, 38]) for simulating viscoelastic flows is the breakdown of the numerical methods. It has been widely believed that the lack of positive definiteness preserving property of the conformation tensor at the discrete level during the *entire time integration* is one of the reasons for the breakdown. To preserve the positive definiteness property of the conformation tensor, several methodologies have been proposed recently, as in [14, 15, 40, 43]. Lozinski and Owens [43] factored the conformation tensor to get $\boldsymbol{\sigma} = AA^T$ and then they wrote down the equations for A approximately at the discrete level. Hence, the positive definiteness of the conformation tensor is forced with such an

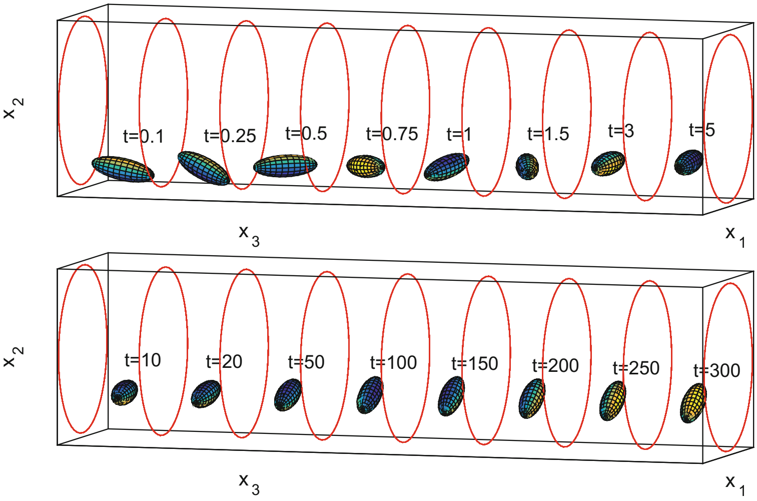


Fig. 11 Visualization of the change of orientation of the prolate ellipsoid: From its initial orientation to its rotation around its long axis, while reaching an equilibrium distance to the axis of the cylinder ($\mu_f = 0.05$, initial angle = 90°)

approach. The methodologies developed in [43] have been applied in [29] together with the FD/DLM method through operator splitting techniques for simulating particulate flows in Oldroyd-B fluid. We have generalized these computational methodologies to viscoelastic fluids of the FENE-CR type, which is a more “realistic” model when compared with the Oldroyd-B model as advocated in [53]. We have compared the particle sedimenting in a vertical two-dimensional channel filled with viscoelastic fluid of either Oldroyd-B or FENE-CR type to find out the effect of the maximum extension of the immersed polymer coils on the chaining (Fig. 11).

5.1 Mathematical Formulations

Following the work developed in [29], we will first address in the following the models and computational methodologies combined with the Lozinski and Owens' factorization approach. Let Ω be a bounded domain in $\mathbb{R}^d$ ($d = 2$ or 3) and let Γ be the boundary of Ω. We suppose that Ω is filled with a viscoelastic fluid of either Oldroyd-B or FENE-CR type of density ρ_f and that it contains N moving rigid particles of density ρ_s (see Fig. 12).

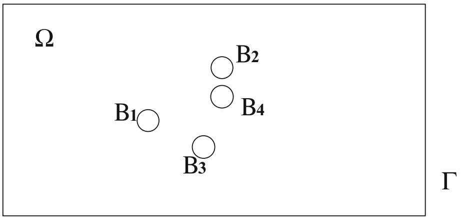


Fig. 12 An example of a two-dimensional flow region with four circular particles

Let $B(t) = \cup_{i=1}^{N} B_i(t)$ where $B_i(t)$ is the ith rigid particle in the fluid for $i = 1, \ldots, N$. We denote by $\partial B_i(t)$ the boundary of $B_i(t)$ for $i = 1, \ldots, N$. For some $T > 0$, the governing equations for the fluid-particle system are

$$\rho_f \left(\frac{\partial \mathbf{u}}{\partial t} + (\mathbf{u} \cdot \boldsymbol{\nabla})\mathbf{u} \right) = \rho_f \mathbf{g} - \boldsymbol{\nabla} p + 2\mu \boldsymbol{\nabla} \cdot \mathbf{D}(\mathbf{u}) + \boldsymbol{\nabla} \cdot \boldsymbol{\sigma}^p \text{ in } \Omega \setminus \overline{B(t)},\ t \in (0, T), \tag{28}$$

$$\nabla \cdot \mathbf{u} = 0 \quad \text{in } \Omega \setminus \overline{B(t)},\ t \in (0, T), \tag{29}$$

$$\mathbf{u}(\mathbf{x}, 0) = \mathbf{u}_0(\mathbf{x}), \quad \forall \mathbf{x} \in \Omega \setminus \overline{B(0)}, \quad \text{with } \nabla \cdot \mathbf{u}_0 = 0, \tag{30}$$

$$\mathbf{u} = \mathbf{g}_0 \quad \text{on } \Gamma \times (0, T), \quad \text{with } \int_\Gamma \mathbf{g}_0 \cdot \mathbf{n}\, d\Gamma = 0, \tag{31}$$

$$\mathbf{u} = \mathbf{V}_{p,i} + \boldsymbol{\omega}_i \times \overrightarrow{\mathbf{G}_i \mathbf{x}}, \quad \forall \mathbf{x} \in \partial B_i(t), \quad i = 1, \ldots, N, \tag{32}$$

$$\frac{\partial \mathbf{C}}{\partial t} + (\mathbf{u} \cdot \boldsymbol{\nabla})\mathbf{C} - (\boldsymbol{\nabla}\mathbf{u})\mathbf{C} - \mathbf{C}(\boldsymbol{\nabla}\mathbf{u})^t = -\frac{f(\mathbf{C})}{\lambda_1}(\mathbf{C} - \mathbf{I}) \text{ in } \Omega \setminus \overline{B(t)},\ t \in (0, T), \tag{33}$$

$$\mathbf{C}(\mathbf{x}, 0) = \mathbf{C}_0(\mathbf{x}), \quad \mathbf{x} \in \Omega \setminus \overline{B(0)}, \tag{34}$$

$$\mathbf{C} = \mathbf{C}_L, \quad \text{on } \Gamma^-, \tag{35}$$

where $\mathbf{u}$ is the flow velocity, p is the pressure, $\mathbf{g}$ is the gravity, $\mu = \eta_1 \lambda_2 / \lambda_1$ is the Newtonian viscosity of the fluid, $\eta = \eta_1 - \mu$ is the elastic viscosity of the fluid, η_1 is the fluid viscosity, λ_1 is the relaxation time of the fluid, λ_2 is the retardation time of the fluid, $\mathbf{n}$ is the outer normal unit vector at Γ, Γ^- is the upstream portion of Γ. The polymeric stress tensor $\boldsymbol{\sigma}^p$ in (28) is given by $\boldsymbol{\sigma}^p = \frac{\eta}{\lambda_1} f(\mathbf{C})(\mathbf{C} - \mathbf{I})$, where the conformation tensor $\mathbf{C}$ is symmetric and positive definite (see [36]) and $\mathbf{I}$ is the identity matrix. Setting f equal to unity corresponds to the Oldroyd-B model while

$$f(\mathbf{C}) = \frac{L^2}{L^2 - tr(\mathbf{C})} \tag{36}$$

corresponds to the FENE-CR model [6], where $tr(\mathbf{C})$ is the trace of the conformation tensor $\mathbf{C}$ and L is the maximum extension of the immersed polymer coils and referred to as the extensibility of the immersed polymer coils. The Oldroyd-B model then is a special case associated with infinite extensibility.

In (32), the no-slip condition holds on the boundary of the ith particle, $\mathbf{V}_{p,i}$ is the translation velocity, $\boldsymbol{\omega}_i$ is the angular velocity and $\mathbf{G}_i$ is the center of mass. The motion of the particles is modeled by Newton's laws:

$$M_{p,i}\,\frac{d\mathbf{V}_{p,i}}{dt} = M_{p,i}\mathbf{g} + \mathbf{F}_i + \mathbf{F}_i^r, \tag{37}$$

$$\frac{d(\mathbf{I}_{p,i}\boldsymbol{\omega}_i)}{dt} = \mathbf{F}_i^t, \tag{38}$$

$$\frac{d\mathbf{G}_i}{dt} = \mathbf{V}_{p,i}, \tag{39}$$

$$\mathbf{G}_i(0) = \mathbf{G}_i^0, \quad \mathbf{V}_{p,i}(0) = \mathbf{V}_{p,i}^0, \quad \boldsymbol{\omega}_i(0) = \boldsymbol{\omega}_i^0, \tag{40}$$

for $i = 1, \ldots, N$, where in (37)–(40), $M_{p,i}$ and $\mathbf{I}_{p,i}$ are the the mass and the inertia tensor of the ith particle, respectively, $\mathbf{F}_i^r$ is a short range repulsion force imposed on the ith particle by other particles and the wall to prevent particle/particle and particle/wall penetration (see [26] for details), and $\mathbf{F}_i$ and $\mathbf{F}_i^t$ denote the hydrodynamic force and the associated torque imposed on the ith particle by the fluid, respectively.

To avoid the frequent remeshing and the difficulty of the mesh generation for a time-varying domain in which the rigid particles can be very close to each other, especially for three dimensional particulate flow, we have extended the governing equations to the entire domain Ω (a fictitious domain). For a fictitious-domain-based variational formulation of the governing equations of the particulate flow, we consider only one rigid particle $B(t)$ (either a disk in 2D or a ball in 3D) in the fluid domain without losing generality. Let us define first the following functional spaces

$$\begin{aligned}
\mathbf{V}_{\mathbf{g}_0(t)} &= \{\mathbf{v} \mid \mathbf{v} \in (H^1(\Omega))^d,\ \mathbf{v} = \mathbf{g}_0(t) \text{ on } \Gamma\},\\
L_0^2(\Omega) &= \{q \mid q \in L^2(\Omega),\ \int_\Omega q\, d\mathbf{x} = 0\},\\
\mathbf{V}_{\mathbf{C}_L(t)} &= \{\mathbf{C} \mid \mathbf{C} \in (H^1(\Omega))^{d\times d},\ \mathbf{C} = \mathbf{C}_L(t) \text{ on } \Gamma^-\},\\
\mathbf{V}_{\mathbf{C}_0} &= \{\mathbf{C} \mid \mathbf{C} \in (H^1(\Omega))^{d\times d},\ \mathbf{C} = 0 \text{ on } \Gamma^-\},\\
\Lambda(t) &= H^1(B(t))^d.
\end{aligned}$$

Following the methodologies developed in [26, 56], a fictitious domain formulation of the governing Eqs. (28)–(40) reads as follows:

For a.e. $t > 0$, *find* $\mathbf{u}(t) \in \mathbf{V}_{\mathbf{g}_0(t)}$, $p(t) \in L_0^2(\Omega)$, $\mathbf{C}(t) \in \mathbf{V}_{\mathbf{C}_L(t)}$, $\mathbf{V}(t) \in \mathbb{R}^d$, $\mathbf{G}(t) \in \mathbb{R}^d$, $\boldsymbol{\omega}(t) \in \mathbb{R}^d$, $\boldsymbol{\lambda}(t) \in \Lambda(t)$ *such that*

$$\begin{cases} \rho_f \displaystyle\int_\Omega \left[\frac{\partial \mathbf{u}}{\partial t} + (\mathbf{u}\cdot\nabla)\mathbf{u}\right]\cdot \mathbf{v}\,d\mathbf{x} + 2\mu \int_\Omega \mathbf{D}(\mathbf{u}) : \mathbf{D}(\mathbf{v})\,d\mathbf{x} - \int_\Omega p\nabla\cdot\mathbf{v}\,d\mathbf{x} \\ -\displaystyle\int_\Omega \mathbf{v}\cdot(\nabla\cdot\boldsymbol{\sigma}^p)\,d\mathbf{x} + (1-\rho_f/\rho_s)\left\{M_p\frac{d\mathbf{V}}{dt}\cdot\mathbf{Y} + \mathbf{I}_p\frac{d\boldsymbol{\omega}}{dt}\cdot\boldsymbol{\theta}\right\} \\ -\langle\boldsymbol{\lambda}, \mathbf{v} - \mathbf{Y} - \boldsymbol{\theta}\times\overrightarrow{\mathbf{G}\mathbf{x}}\rangle_{B(t)} - \mathbf{F}^r\cdot\mathbf{Y} \\ = \rho_f \displaystyle\int_\Omega \mathbf{g}\cdot\mathbf{v}d\mathbf{x} + (1-\rho_f/\rho_s)M_p\mathbf{g}\cdot\mathbf{Y}, \\ \forall\{\mathbf{v},\mathbf{Y},\boldsymbol{\theta}\} \in (H_0^1(\Omega))^d \times \mathbb{R}^d \times \mathbb{R}^d, \end{cases} \tag{41}$$

$$\int_\Omega q\nabla\cdot\mathbf{u}(t)\,d\mathbf{x} = 0, \forall q \in L^2(\Omega), \tag{42}$$

$$\langle\boldsymbol{\mu}, \mathbf{u}(\mathbf{x},t) - \mathbf{V}(t) - \boldsymbol{\omega}(t)\times\overrightarrow{\mathbf{G}(t)\mathbf{x}}\rangle_{B(t)} = 0,\ \ \forall\boldsymbol{\mu}\in\boldsymbol{\Lambda}(t), \tag{43}$$

$$\int_\Omega \left(\frac{\partial\mathbf{C}}{\partial t} + (\mathbf{u}\cdot\nabla)\mathbf{C} - (\nabla\mathbf{u})\mathbf{C} - \mathbf{C}(\nabla\mathbf{u})^t\right) : \mathbf{s}\,d\mathbf{x} \tag{44}$$

$$= -\int_\Omega \frac{f(\mathbf{C})}{\lambda_1}(\mathbf{C}-\mathbf{I}) : \mathbf{s}\,d\mathbf{x}, \quad \forall\mathbf{s}\in\mathbf{V}_{\mathbf{C}_0},\ \text{with}\,\mathbf{C} = \mathbf{I}\ \textit{in}\ B(t),$$

$$\frac{d\mathbf{G}}{dt} = \mathbf{V}, \tag{45}$$

$$\mathbf{C}(\mathbf{x},0) = \mathbf{C}_0(\mathbf{x}), \quad \forall\mathbf{x}\in\Omega,\ \text{with}\,\mathbf{C}_0 = \mathbf{I}\ \textit{in}\ B(0), \tag{46}$$

$$\mathbf{G}(0) = \mathbf{G}_0, \quad \mathbf{V}(0) = \mathbf{V}_0, \quad \boldsymbol{\omega}(0) = \boldsymbol{\omega}_0, \quad B(0) = B_0, \tag{47}$$

$$\mathbf{u}(\mathbf{x},0) = \begin{cases} \mathbf{u}_0(\mathbf{x}), & \forall\mathbf{x}\in\Omega\setminus\overline{B_0}, \\ \mathbf{V}_0 + \boldsymbol{\omega}_0\times\overrightarrow{\mathbf{G}_0\mathbf{x}}, & \forall\mathbf{x}\in\overline{B_0}. \end{cases} \tag{48}$$

In (41) the *Lagrange multiplier* $\boldsymbol{\lambda}$ defined over B can be viewed as an extra body force maintaining the rigid body motion inside B. The conformation tensor $\mathbf{C}$ inside the rigid particle is extended as the identity tensor $\mathbf{I}$ as in (44) since the polymeric stress tensor is zero inside the rigid particle. In Eq. (41), since $\mathbf{u}$ is divergence free and satisfies the Dirichlet boundary conditions on Γ, we have $2\int_\Omega \mathbf{D}(\mathbf{u}) : \mathbf{D}(\mathbf{v})d\mathbf{x} = \int_\Omega \nabla\mathbf{u} : \nabla\mathbf{v}d\mathbf{x}$, $\forall\mathbf{v}\in(H_0^1(\Omega))^d$. This is a substantial simplification from the computational point of view, which is another advantage of the fictitious domain approach. With this simplification, we can use, as shown in the following section, fast solvers for the elliptic problems in order to speed up computations. Also the gravity term $\mathbf{g}$ in (41) can be absorbed in the pressure term.

5.2 Numerical Results

The details of numerical methodologies for simulating the motion of disks sedimenting in Oldroyd-B fluid in a vertical two-dimensional channel are given in [29]. Applying Lie's scheme to (41)–(48), we have used a six-stage operator-splitting scheme, namely:

Stage 1 We use a Neumann preconditioned Uzawa/conjugate gradient algorithm to force (in a L^2 sense) the incompressibility condition of $\mathbf{u}$.
Stage 2 We combine two advection steps similar to (4): one for $\mathbf{u}$ and one for $\mathbf{C}$.
Stage 3 We combine a diffusion step for $\mathbf{u}$ with a step taking into account the remaining operator in the evolution equation verified by $\mathbf{C}$.
Stage 4 We update the position of $\mathbf{G}$.
Stage 5 We force the rigid body motion of the particle, update $\mathbf{V}$ and $\boldsymbol{\omega}$, and impose the condition $\mathbf{C} = \mathbf{I}$ inside the particle.
Stage 6 This is a diffusion step for the velocity, driven by the updated polymeric stress tensor.

We present here the results of two numerical experiments concerning the sedimentation of circular particles in a two-dimensional channel filled with an Oldroyd-B fluid. For the space discretization, we have still used a P_1-*iso*-P_2 (resp., P_1) finite element for the approximation of velocity field (resp., pressure) defined from uniform "triangular" meshes $\mathscr{T}_h$ (resp., $\mathscr{T}_{2h}$). For the finite element approximation of each entry in the conformation tensor, P_1 finite element spaces defined from uniform triangular meshes $\mathscr{T}_h$ have been used. For the enforcement of the rigid body motion, we have applied a collocation method (see [26] for details).

The numerical results concern six circular particles of diameter $D = 0.25$ sedimenting in a channel filled with an Oldroyd-B fluid. The channel is infinitely long and has a width of 1. The computational domain is $\Omega = (0, 1) \times (0, 7)$ initially and then moves down with the mass center of the lowest of the six particles. It is known that when the elasticity number $\mathrm{E} = \mathrm{De}/\mathrm{Re}$ is larger than the critical value ($O(1)$) and the Mach number $\mathrm{M} = \sqrt{\mathrm{DeRe}} < 1$, the particles in this case will form chains that are parallel to the flow [32, 42]. In our simulations, all six particles are lined up along the flow direction, agreeing thus the known observations and experiments. Figure 13 gives the snapshots at various moments of time of the particles lining up phenomenon.

We can see that, after drafting, kissing and chaining, the six particles form approximately a straight line at $t = 20$; at $t = 30$, the trailing particle has been separated from the leading five particles. This observation agrees with experiments showing that, sometimes, the last particle in the chain gets detached as discussed in [51]. It is known that a long chain falls faster than a single particle in the fluid. This long body effect tends to detach the last particle from the chain. The average terminal velocity is 0.1535 for $26 \leq t \leq 30$, the Reynolds number is $\mathrm{Re} = 0.1476$, the Deborah number is $\mathrm{De} = 0.7981$, the elasticity number is $\mathrm{E} = 5.408$ and the Mach number is $\mathrm{M} = 0.3432$. With the same parameters as in the case of Oldroyd-B fluid, we just changed to the FENE-CR model with $L = 5$ for the polymer extension limit. Since the viscoelastic fluid has a shorter polymer extension limit, it cannot hold all six disks together as shown in Fig. 14. For this case, the average terminal velocity is 0.1317 for

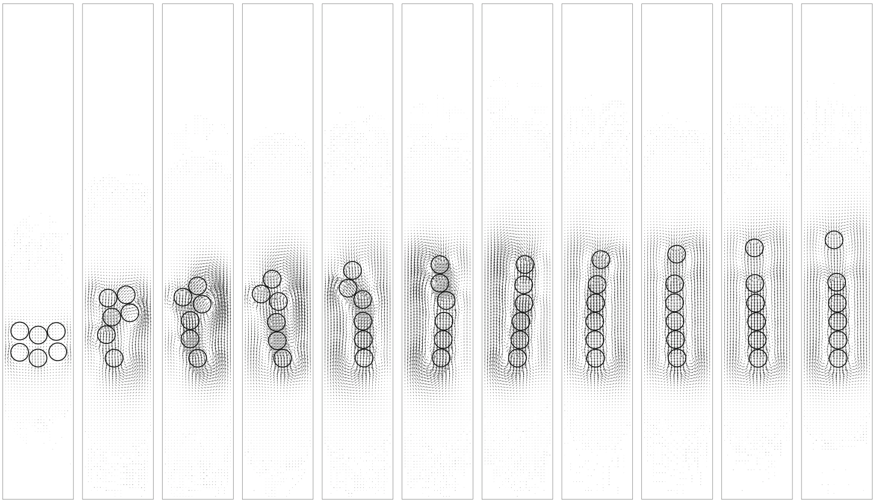


Fig. 13 Snapshots at $t = 2, 10, 12, 14, 16, 18, 20, 24, 26, 28$, and 30 of the positions of six particles lining up in an Oldroyd-B fluid ($h = 1/96$ and $\triangle t = 0.0004$)

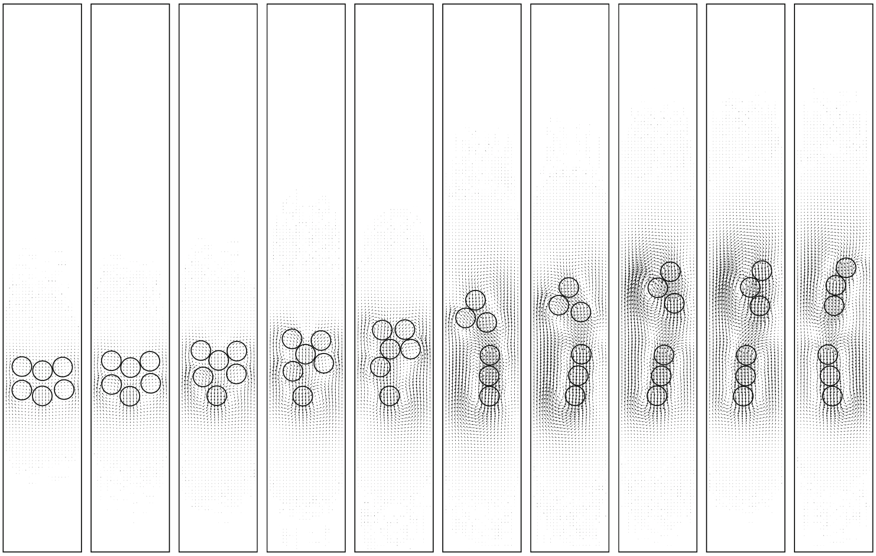


Fig. 14 Snapshots at $t = 2, 4, 6, 8, 10, 18, 20, 24, 26$, and 30 of the positions of six particles lining up in an FENE-CR fluid with $L = 5$ ($h = 1/96$ and $\triangle t = 0.0004$)

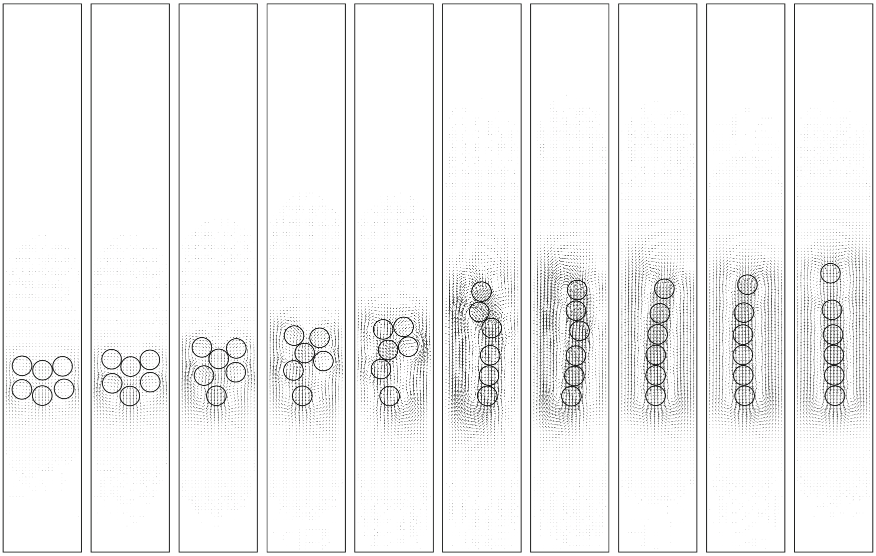


Fig. 15 Snapshots at $t = 2, 4, 6, 8, 10, 18, 20, 24, 26$, and 30 of the positions of six particles lining up in an FENE-CR fluid with $L = 10$ ($h = 1/96$ and $\triangle t = 0.0004$)

$26 \leq t \leq 30$, and the associated numbers are $\mathrm{Re} = 0.1266$, $\mathrm{De} = 0.6847$, $\mathrm{E} = 5.408$ and $\mathrm{M} = 0.2944$. But for the case $L = 10$, the chaining shown in Fig. 15 is much closer to the one obtained for the Oldroyd-B fluid since in (36), $f(\mathbf{C})$ is close to 1 (i.e., the FENE-CR model has almost recovered the Oldroyd-B model). The terminal velocity is 0.1490 for $26 \leq t \leq 30$, and the associated numbers are $\mathrm{Re} = 0.1433$, $\mathrm{De} = 0.7750$, $\mathrm{E} = 5.408$ and $\mathrm{M} = 0.3333$.

6 Conclusion

The wave-like equation based method we introduced twenty years ago, for the numerical simulation of incompressible viscous flow (as an alternative to Lagrange-Galerkin methods) has been further discussed in this article. This method, which allows a purely variational treatment of the advection (well-suited to simple finite element approximations), has been briefly described in Sect. 2 of this article, and applied in Sects. 3–5 to the simulation of Newtonian and non-Newtonian viscous flows in two and three dimensions, possibly involving rigid solid particles. Through the methodology discussed in this article we have been able to reproduce accurately documented phenomena from the physics of fluids, and more importantly to discover new ones, as shown in Sects. 4 and 5. The results reported in this article suggest that, despite being twenty years old, the methodology we discussed in this article is far from being obsolete and should be still helpful in the future.

Acknowledgements The authors of this article are thanking the editors of this volume, a very special thank being due to Marja-Leena Rantalainen, from the University of Jyväskylä in Finland, for her beautiful processing of this contribution. The support of the US National Science Foundation via grant DMS-1418308 is also acknowledged.

References

1. Anupindi K, Lai W, Frankel S (2014) Characterization of oscillatory instability in lid driven cavity flows using lattice Boltzmann method. Comput Fluids 92:7–21
2. Baaijens FPT (1998) Mixed finite element methods for viscoelastic flow analysis: a review. J Non-Newton Fluid Mech 79(2–3):361–385
3. Bristeau MO, Glowinski R, Périaux J (1987) Numerical methods for the Navier-Stokes equations. Application to the simulation of compressible and incompressible viscous flow. Comput Phys Rep 6:73–187
4. Chhabra RP (1993) Bubbles, drops, and particles in non-Newtonian fluids. CRC Press, Boca Raton, FL
5. Chiang TP, Sheu WH, Hwang RR (1998) Effect of Reynolds number on the eddy structure in a lid-driven cavity. Int J Numer Meth Fluids 26(5):557–579
6. Chilcott MD, Rallison JM (1988) Creeping flow of dilute polymer solutions past cylinders and spheres. J Non-Newton Fluid Mech 29:381–432
7. Chippada S, Dawson CN, Martinez ML, Wheeler MF (1998) Finite element approximations to the system of shallow water equations I: continuous-time a priori error estimates. SIAM J Numer Anal 35(2):692–711
8. Chippada S, Dawson CN, Martínez-Canales ML, Wheeler MF (1998) Finite element approximations to the system of shallow water equations, Part II: discrete-time a priori error estimates. SIAM J Numer Anal 36(1):226–250
9. Chorin AJ (1967) A numerical method for solving incompressible viscous flow problems. J Comput Phys 2(1):12–26
10. Dean EJ, Glowinski R (1997) A wave equation approach to the numerical solution of the Navier-Stokes equations for incompressible viscous flow. CR Acad Sci Paris Sér I Math 325(7):783–791
11. Dean EJ, Glowinski R, Guidoboni G (2007) On the numerical simulation of Bingham visco-plastic flow: old and new results. J Non-Newton Fluid Mech 142(1–3):36–62
12. Dean EJ, Glowinski R, Pan T-W (1998) A wave equation approach to the numerical simulation of incompressible viscous fluid flow modelled by the Navier-Stokes equations. In: De Santo JA (ed) Mathematical and numerical aspects of wave propagation (Golden. CO, 1998). Philadelphia, PA, pp 65–74
13. Economides MJ, Nolte KG (1989) Reservoir stimulation. Prentice Hall, Englewood Cliffs, NJ
14. Fattal R, Kupferman R (2004) Constitutive laws for the matrix-logarithm of the conformation tensor. J Non-Newton Fluid Mech 123(2–3):281–285
15. Fattal R, Kupferman R (2005) Time-dependent simulation of viscoelastic flows at high Weissenberg number using the log-conformation representation. J Non-Newton Fluid Mech 126(1):23–37
16. Feldman Y, Gelfgat AY (2010) Oscillatory instability of a three-dimensional lid-driven flow in a cube. Phys Fluids 22:093602
17. Fujima S, Tabata M, Fukasawa Y (1994) Extension to three-dimensional problems of the upwind finite element scheme based on the choice of up- and downwind points. Comput Methods Appl Mech Eng. 112:109–131
18. Ghia UK, Ghia KN, Shin CT (1982) High-Re solutions for incompressible flow using the Navier-Stokes equations and a multigrid method. J Comput Phys 48(3):387–411
19. Giannetti R, Luchini P, Marino L (2009) Linear stability analysis of three-dimensional lid-driven cavity flow. In: Atti del XIX Congresso AIMETA di Meccanica Teorica e Applicata (Ancona, 2009), Aras Edizioni, pp 738.1–738.10

20. Glowinski R (2003) Finite element methods for incompressible viscous flow. In: Ciarlet PG, Lions JL (eds) Handbook of numerical analysis, vol IX. North-Holland, Amsterdam, pp 3–1176
21. Glowinski R (2015) Variational methods for the numerical solution of nonlinear elliptic problems. SIAM, Philadelphia, PA
22. Glowinski R, Dean EJ, Guidoboni G, Juárez LH, Pan T-W (2008) Applications of operator-splitting methods to the direct numerical simulation of particulate and free-surface flows and to the numerical solution of the two-dimensional elliptic Monge-Ampère equation. Jap J Indust Appl Math 25(1):1–63
23. Glowinski R, Guidoboni G, Pan T-W (2006) Wall-driven incompressible viscous flow in a two-dimensional semi-circular cavity. J Comput Phys 216(1):76–91
24. Glowinski R, Lawton W, Ravachol M, Tenenbaum E (1990) Wavelet solution of linear and nonlinear elliptic, parabolic and hyperbolic problems in one space dimension. In: Glowinski R, Lichnewsky A (eds) Computing methods in applied sciences and engineering (Paris, 1990). SIAM, Philadelphia, PA, pp 55–120
25. Glowinski R, Osher S, Yin W (eds) (2016) Splitting methods in communication and imaging, science and engineering. Springer, New York
26. Glowinski R, Pan T-W, Hesla TI, Joseph DD, Périaux J (2001) A fictitious domain approach to the direct numerical simulation of incompressible viscous flow past moving rigid bodies: application to particulate flow. J Comput Phys 169(2):363–426
27. Glowinski R, Wachs A (2011) On the numerical simulation of viscoplastic fluid flow. In: Ciarlet PG, Glowinski R, Xu J (eds) Handbook of numerical analysis, vol XVI. North-Holland, Amsterdam, pp 483–718
28. Gustafsson B, Kreiss H-O, Oliger J (1995) Time dependent problems and difference methods. Wiley, New York
29. Hao J, Pan T-W, Glowinski R, Joseph DD (2009) A fictitious domain/distributed Lagrange multiplier method for the particulate flow of Oldroyd-B fluids: a positive definiteness preserving approach. J Non-Newton Fluid Mech 156(1–2):95–111
30. He Q, Glowinski R, Wang XP (2011) A least-squares/finite element method for the numerical solution of the Navier-Stokes-Cahn-Hilliard system modeling the motion of the contact line. J. Comput. Phys. 230(12):4991–5009
31. Hou S, Pan T-W, Glowinski R (2014) Circular band formation for incompressible viscous fluid-rigid-particle mixtures in a rotating cylinder. Phys Rev E 89(2):023013
32. Huang PY, Hu HH, Joseph DD (1998) Direct simulation of the sedimentation of elliptic particles in Oldroyd-B fluids. J Fluid Mech 362:297–326
33. Hur SC, Choi SE, Kwon S, Di Carlo D (2011) Inertial focusing of non-spherical microparticles. Appl Phys Lett 99(4):044101
34. Iwatsu R, Hyun JM, Kuwahara K (1990) Analyses of three-dimensional flow calculations in a driven cavity. Fluid Dyn Res 6(2):91–102
35. Jeffery GB (1922) The motion of ellipsoidal particles immersed in a viscous fluid. Proc R Soc London A 102:161–179
36. Joseph DD (1990) Fluid dynamics of viscoelastic liquids. Springer, New York
37. Karnis A, Goldsmith HL, Mason SG (1966) The flow of suspensions through tubes: V. Inertial effects. Can J Chem Eng 44(4):181–193
38. Keunings R (2000) A survey of computational rheology. In: Binding DM et al (eds) Proceedings of the 13th international congress on rheology, vol 1. British Society of Rheology, Glasgow, pp 7–14
39. Ku HC, Hirsh RS, Taylor TD (1987) A pseudospectral method for solution of the three-dimensional incompressible Navier-Stokes equations. J Comput Phys 70(2):439–462
40. Lee Y-J, Xu J (2006) New formulations, positivity preserving discretizations and stability analysis for non-Newtonian flow models. Comput Methods Appl Mech Eng 195(9–12):1180–1206
41. Liberzon A, Feldman Y, Gelfgat A (2011) Experimental observation of the steady-oscillatory transition in a cubic lid-driven cavity. Phys Fluids 23:084106

42. Liu YJ, Joseph DD (1993) Sedimentation of particles in polymer solutions. J Fluid Mech 255:565–595
43. Lozinski A, Owens RG (2003) An energy estimate for the Oldroyd-B model: theory and applications. J Non-Newton Fluid Mech 112(2–3):161–176
44. Lynch DR, Gray WG (1979) A wave equation model for finite element tidal computations. Comput Fluids 7(3):207–228
45. McKinley GH (2002) Steady and transient motion of spherical particles in viscoelastic liquids. In: De Kee D, Chhabra RP (eds) Transport processes in bubbles, drops, and particles, 2nd edn. Taylor & Francis, New York, pp 338–375
46. Pan T-W, Chang C-C, Glowinski R (2008) On the motion of a neutrally buoyant ellipsoid in a three-dimensional Poiseuille flow. Comput Methods Appl Mech Eng 197(25–28):2198–2209
47. Pan T-W, Glowinski R (2000) A projection/wave-like equation method for the numerical simulation of incompressible viscous fluid flow modeled by the Navier-Stokes equations. Comput Fluid Dyn J 9(2):28–42
48. Pan T-W, Hao J, Glowinski R (2009) On the simulation of a time-dependent cavity flow of an Oldroyd-B fluid. Int J Numer Meth Fluids 60(7):791–808
49. Pan T-W, Hao J, Glowinski R (2011) Positive definiteness preserving approaches for viscoelastic flow of Oldroyd-B fluids: applications to a lid-driven cavity flow and a particulate flow. In: Ciarlet PG, Glowinski R, Xu J (eds) Handbook of numerical analysis, vol XVI. North-Holland, Amsterdam, pp 433–481
50. Pan T-W, Joseph DD, Glowinski R (2005) Simulating the dynamics of fluid-ellipsoid interactions. Comput Struct. 83(6–7):463–478
51. Patankar NA, Hu HH (2000) A numerical investigation of the detachment of the trailing particle from a chain sedimenting in Newtonian and viscoelastic fluids. J Fluids Eng 122(3):517–521
52. Pironneau O (1989) Finite element methods for fluids. Wiley, Chichester
53. Rallison JM, Hinch EJ (1988) Do we understand the physics in the constitutive equation? J Non-Newton Fluid Mech 29:37–55
54. Segré G, Silberberg A (1961) Radial particle displacements in Poiseuille flow of suspensions. Nature 189:209–210
55. Segré G, Silberberg A (1962) Behaviour of macroscopic rigid spheres in Poiseuille flow Part 1. Determination of local concentration by statistical analysis of particle passages through crossed light beams. J Fluid Mech 14(1):115–135
56. Singh P, Joseph DD, Hesla TI, Glowinski R, Pan T-W (2000) A distributed Lagrange multiplier/fictitious domain method for viscoelastic particulate flows. J Non-Newton Fluid Mech 91(2–3):165–188
57. Süli E (1988) Convergence and nonlinear stability of the Lagrange-Galerkin method for the Navier-Stokes equations. Numer Math 53(4):459–483
58. Süli E (1988) Stability and convergence of the Lagrange-Galerkin method with nonexact integration. In: Whiteman JR (ed) The mathematics of finite elements and applications, VI (Uxbridge, 1987). Academic Press, London, pp 435–442
59. Tagliabue A, Dedè L, Quarteroni A (2014) Isogeometric analysis and error estimates for high order partial differential equations in fluid dynamics. Comput Fluids 102:277–303
60. Wu J (1994) Wave equation models for solving advection-diffusion equation. Int J Numer Methods Eng 37(16):2717–2733
61. Wu J (1997) A wave equation model to solve the multidimensional transport equation. Int J Numer Meth Fluids 24(5):423–439

Arbitrary Lagrangian-Eulerian Finite Element Method Preserving Convex Invariants of Hyperbolic Systems

Jean-Luc Guermond, Bojan Popov, Laura Saavedra and Yong Yang

Abstract We present a conservative Arbitrary Lagrangian Eulerian method for solving nonlinear hyperbolic systems. The key characteristics of the method is that it preserves all the convex invariants of the hyperbolic system in question. The method is explicit in time, uses continuous finite elements and is first-order accurate in space and high-order in time. The stability of the method is obtained by introducing an artificial viscosity that is unambiguously defined irrespective of the mesh geometry/anisotropy and does not depend on any ad hoc parameter.

1 Introduction

This paper is the expanded version of a talk given at the University of Houston in February 2016 at a workshop honoring the 70th birthday of Olivier Pironneau and his long lasting contributions to Numerical Analysis and Scientific Computing [19]. The topic of paper is in the continuation of the groundbreaking work done by Olivier Pironneau on the analysis of the method of characteristics for solving the

J.-L. Guermond (✉) · B. Popov
Department of Mathematics, Texas A&M University 3368 TAMU,
College Station, TX 77843, USA
e-mail: guermond@math.tamu.edu

B. Popov
e-mail: popov@math.tamu.edu

L. Saavedra
Departamento Fundamentos Matemáticos, Departamento de Matematica
Aplicada a la Ingenieria Aeroespacial,
Universidad Politécnica de Madrid,
E.T.S.I. Aeronáuticos, 28040 Madrid, Spain
e-mail: laura.saavedra@upm.es

Y. Yang
Department of Mathematics, Penn State University,
University Park, State College, PA 16802, USA
e-mail: yytamu@gmail.com

B. N. Chetverushkin et al. (eds.), *Contributions to Partial Differential Equations and Applications*, Computational Methods in Applied Sciences 47,
https://doi.org/10.1007/978-3-319-78325-3_14

transport equation [18]. More specifically, our objective is to build a finite element approximation to the entropy solution of the following hyperbolic system written in conservative form:

$$\begin{cases} \partial_t \boldsymbol{u} + \nabla\cdot \boldsymbol{f}(\boldsymbol{u}) = 0, & \text{for } (\boldsymbol{x},t) \in \mathbb{R}^d \times \mathbb{R}_+, \\ \boldsymbol{u}(\boldsymbol{x},0) = \boldsymbol{u}_0(\boldsymbol{x}), & \text{for } \boldsymbol{x} \in \mathbb{R}^d, \end{cases} \tag{1}$$

where the dependent variable $\boldsymbol{u}$ is $\mathbb{R}^m$-valued and the flux $\boldsymbol{f}$ is $\mathbb{R}^{m\times d}$-valued. We investigate in this paper an approximation technique using an Arbitrary Lagrangian Eulerian (ALE) formulation with continuous finite elements and explicit time stepping on non-uniform meshes.

The paper is organized as follows. We introduce some notation and recall important properties about the one-dimensional Riemann problem in Sect. 2. We introduce notation relative to mesh motion and Lagrangian mappings in Sect. 3. The results established in Sects. 2 and 3 are standard and will be invoked in Sects. 4 and 5. It is proved in Sect. 5 that under the appropriate CFL condition the algorithm is conservative, satisfies a local entropy inequality for every admissible entropy pair and preserves invariant domains. The main results of this section are Theorem 1 and Theorem 2. The SSP RK3 extension of the method is tested numerically in Sect. 6 on scalar conservation equations and on the compressible Euler equations. The paper essentially reproduces the arguments developed in [13]. We refer the reader to [13] for details, proofs and extensions of the material presented herein.

2 Riemann Problem and Invariant Domain

We recall in this section elementary properties of Riemann problems that will be used in the paper.

2.1 Notation and Boundary Conditions

The dependent variable $\boldsymbol{u}$ in (1) is considered as a column vector $\boldsymbol{u} = (u_1, \dots, u_m)^\mathsf{T}$. The flux is a matrix with entries $f_{ij}(\boldsymbol{u})$, $1 \le i \le m$, $1 \le j \le d$. We denote by $\boldsymbol{f}_i$ the row vector $(f_{i1}, \dots, f_{id})$, $i \in \{1{:}m\}$. We denote by $\nabla\cdot\boldsymbol{f}$ the column vector with entries

$$(\nabla\cdot\boldsymbol{f})_i = \sum_{1\le j\le d} \partial_{x_j} f_{ij}.$$

For any $\boldsymbol{n} = (n_1 \dots, n_d)^\mathsf{T} \in \mathbb{R}^d$, we denote $\boldsymbol{f}(\boldsymbol{u})\cdot\boldsymbol{n}$ the column vector with entries

$$\boldsymbol{f}_i(\boldsymbol{u}) \cdot \boldsymbol{n} = \sum_{1 \le l \le d} n_l f_{il}(\boldsymbol{u}), \quad i \in \{1{:}m\}.$$

Given two vector fields, say $\boldsymbol{u} \in \mathbb{R}^m$ and $\boldsymbol{v} \in \mathbb{R}^d$, we define $\boldsymbol{u} \otimes \boldsymbol{v}$ to be the $m \times d$ matrix with entries $u_i v_j$, $i \in \{1{:}m\}$, $j \in \{1{:}d\}$. We also define $\nabla \cdot (\boldsymbol{u} \otimes \boldsymbol{v})$ to be the column vector with entries

$$\nabla \cdot (\boldsymbol{u} \otimes \boldsymbol{v})_i = \sum_{j=1}^{d} \partial_j (u_i v_j).$$

The unit sphere in $\mathbb{R}^d$ centered at 0 is denoted by $S^{d-1}(\boldsymbol{0}, 1)$.

To simplify questions regarding boundary conditions, we assume that either the initial data is constant outside a compact set and we solve the Cauchy problem in $\mathbb{R}^d$ or we use periodic boundary conditions.

2.2 One-Dimensional Riemann Problem

We are not going to try to define weak solutions to (1), but instead we assume that there is a clear notion for the solution of the Riemann problem. To stay general we introduce a generic hyperbolic flux $\boldsymbol{h}$ and we say that $(\eta, \boldsymbol{q})$ is an entropy pair associated with the flux $\boldsymbol{h}$ if η is convex and the following identity holds:

$$\partial_{v_k}(\boldsymbol{q}(\boldsymbol{v}) \cdot \boldsymbol{n}) = \sum_{i=1}^{m} \partial_{v_i} \eta(\boldsymbol{v}) \partial_{v_k}(\boldsymbol{h}_i(\boldsymbol{v}) \cdot \boldsymbol{n}), \qquad \forall k \in \{1{:}m\},\ \forall \boldsymbol{n} \in S^{d-1}(\boldsymbol{0}, 1).$$

We refer to [4, Sect. 2] for more details on convex entropies and symmetrization. In the rest of the paper we assume that there exists a nonempty admissible set $\mathscr{A}_{\boldsymbol{h}} \subset \mathbb{R}^m$ such that the following one-dimensional Riemann problem

$$\partial_t \boldsymbol{u} + \partial_x(\boldsymbol{h}(\boldsymbol{u}) \cdot \boldsymbol{n}) = 0, \quad (x,t) \in \mathbb{R} \times \mathbb{R}_+, \qquad \boldsymbol{u}(x,0) = \begin{cases} \boldsymbol{u}_L, & \text{if } x < 0 \\ \boldsymbol{u}_R, & \text{if } x > 0, \end{cases} \tag{2}$$

has a unique entropy satisfying solution for any pair of states $(\boldsymbol{u}_L, \boldsymbol{u}_R) \in \mathscr{A}_{\boldsymbol{h}} \times \mathscr{A}_{\boldsymbol{h}}$ and any unit vector $\boldsymbol{n} \in S^{d-1}(\boldsymbol{0}, 1)$. We henceforth denote the solution to this problem by $\boldsymbol{u}(\boldsymbol{h}, \boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R)$. We also say that $\boldsymbol{u}$ is an entropy satisfying solution of (2) if the following holds in the distribution sense for any entropy pair $(\eta, \boldsymbol{q})$:

$$\partial_t \eta(\boldsymbol{u}) + \partial_x(\boldsymbol{q}(\boldsymbol{u}) \cdot \boldsymbol{n}) \le 0.$$

Since it is unrealistic to expect a general theory of the Riemann problem (2) for arbitrary nonlinear hyperbolic systems with large data, we instead make the following assumption: The unique solution of (2) has a finite speed of propagation for any $\boldsymbol{n}$ and any $(\boldsymbol{u}_L, \boldsymbol{u}_R) \in \mathscr{A}_{\boldsymbol{h}} \times \mathscr{A}_{\boldsymbol{h}}$, i.e., there are $\lambda_L(\boldsymbol{h}, \boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R) \le \lambda_R(\boldsymbol{h}, \boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R)$ s.t.

$$\boldsymbol{u}(x,t) = \begin{cases} \boldsymbol{u}_L, & \text{if } x \le t\lambda_L(\boldsymbol{h}, \boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R) \\ \boldsymbol{u}_R, & \text{if } x \ge t\lambda_R(\boldsymbol{h}, \boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R). \end{cases} \tag{3}$$

This assumption is known to hold for small data when the system is strictly hyperbolic with smooth flux and all the characteristic fields are either genuinely nonlinear or linearly degenerate, see, e.g., [6, Thm. 9.5.1]. The sector $\lambda_L t < x < \lambda_R t$, $0 < t$, is henceforth referred to as the Riemann fan. The maximum wave speed in the Riemann fan is $\lambda_{\max} := \lambda_{\max}(\boldsymbol{h}, \boldsymbol{n}, \boldsymbol{u}_L, \boldsymbol{u}_R) := \max(|\lambda_L|, |\lambda_R|)$.

2.3 Invariant Sets and Domains

The following elementary result is a well-known and important consequence of the Riemann fan assumption (3):

Lemma 1 *Let $\boldsymbol{h}$ be a hyperbolic flux over the admissible set $\mathscr{A}_{\boldsymbol{h}}$ and satisfying the finite wave speed assumption (3). Let $\boldsymbol{v}(\boldsymbol{h}, \boldsymbol{n}, \boldsymbol{v}_L, \boldsymbol{v}_R)$ be the unique solution to the problem $\partial_t \boldsymbol{v} + \partial_x(\boldsymbol{h}(\boldsymbol{v}) \cdot \boldsymbol{n}) = 0$ with initial data $\boldsymbol{v}_L, \boldsymbol{v}_R \in \mathscr{A}_{\boldsymbol{h}}$. Let $(\eta, \boldsymbol{q})$ be an entropy pair associated with the flux $\boldsymbol{h}$. Assume that $t\, \lambda_{\max}(\boldsymbol{h}, \boldsymbol{n}, \boldsymbol{v}_L, \boldsymbol{v}_R) \le \frac{1}{2}$ and let*

$$\overline{\boldsymbol{v}}(t, \boldsymbol{h}, \boldsymbol{n}, \boldsymbol{v}_L, \boldsymbol{v}_R) := \int_{-\frac{1}{2}}^{\frac{1}{2}} \boldsymbol{v}(\boldsymbol{h}, \boldsymbol{n}, \boldsymbol{v}_L, \boldsymbol{v}_R)(x,t) \mathrm{d}\boldsymbol{x}.$$

Then

$$\overline{\boldsymbol{v}}(t, \boldsymbol{h}, \boldsymbol{n}, \boldsymbol{v}_L, \boldsymbol{v}_R) = \tfrac{1}{2}(\boldsymbol{v}_L + \boldsymbol{v}_R) - t\big(\boldsymbol{h}(\boldsymbol{v}_R) \cdot \boldsymbol{n} - \boldsymbol{h}(\boldsymbol{v}_L) \cdot \boldsymbol{n}\big). \tag{4}$$

$$\eta(\overline{\boldsymbol{v}}(t, \boldsymbol{h}, \boldsymbol{n}, \boldsymbol{v}_L, \boldsymbol{v}_R)) \le \tfrac{1}{2}(\eta(\boldsymbol{v}_L) + \eta(\boldsymbol{v}_R)) - t(\boldsymbol{q}(\boldsymbol{v}_R) \cdot \boldsymbol{n} - \boldsymbol{q}(\boldsymbol{v}_L) \cdot \boldsymbol{n}). \tag{5}$$

We now introduce notions of invariant sets that are slightly different from what is usually done in the literature (see, e.g., [5, 9, 15]).

Definition 1 (*Invariant set*) Let $\boldsymbol{h}$ be a hyperbolic flux over the admissible set $\mathscr{A}_{\boldsymbol{h}}$ and satisfying the finite wave speed assumption (3). A convex set $A \subset \mathscr{A}_{\boldsymbol{h}} \subset \mathbb{R}^m$ is said to be invariant for the problem $\partial_t \boldsymbol{v} + \nabla \cdot \boldsymbol{h}(\boldsymbol{v}) = 0$ if for any pair $(\boldsymbol{v}_L, \boldsymbol{v}_R) \in A \times A$, any unit vector $\boldsymbol{n} \in S^{d-1}(\boldsymbol{0}, 1)$, the average of the entropy solution of the Riemann problem

$$\partial_t \boldsymbol{v} + \nabla \cdot (\boldsymbol{h}(\boldsymbol{v}) \cdot \boldsymbol{n}) = 0$$

over the Riemann fan

$$\frac{1}{t(\lambda_R-\lambda_L)}\int_{\lambda_L t}^{\lambda_R t} \boldsymbol{v}(\boldsymbol{h},\boldsymbol{n},\boldsymbol{v}_L,\boldsymbol{v}_R)(x,t)\ \mathrm{d}\boldsymbol{x},$$

remains in A for all $t>0$.

Remark 1 The above definition implies that

$$\frac{1}{I}\int_I \boldsymbol{v}(\boldsymbol{h},\boldsymbol{n},\boldsymbol{v}_L,\boldsymbol{v}_R)(x,t)\,\mathrm{d}\boldsymbol{x}\in A$$

for any $t>0$ and any interval I such that $(\lambda_L t,\lambda_R t)\subset I$.

Lemma 2 (Translation) *Let* $\boldsymbol{W}\in\mathbb{R}^d$ *and let* $\boldsymbol{g}(\boldsymbol{v}):=\boldsymbol{f}(\boldsymbol{v})-\boldsymbol{v}\otimes\boldsymbol{W}$.

(i) The two problems: $\partial_t\boldsymbol{u}+\nabla\cdot\boldsymbol{f}(\boldsymbol{u})=0$ *and* $\partial_t\boldsymbol{v}+\nabla\cdot\boldsymbol{g}(\boldsymbol{v})=0$ *have the same admissible sets and the same invariant sets.*

(ii) $(\eta(\boldsymbol{u}),\boldsymbol{q}(\boldsymbol{u}))$ *is an entropy pair for the flux* $\boldsymbol{f}$ *if and only if* $(\eta(\boldsymbol{v}),\boldsymbol{q}(\boldsymbol{v})-\eta(\boldsymbol{v})\boldsymbol{W})$ *is an entropy pair for the flux* $\boldsymbol{g}$.

3 Geometric Preliminaries

In this section we introduce some notation and recall well known results about Lagrangian mappings. The key results, which will be invoked in Sects. 4 and 5, are Lemmas 3 and 4. The reader who is familiar with these notions is invited to skip this section and to go directly to Sect. 4.

3.1 Jacobian of the Coordinate Transformation

Let $\boldsymbol{\Phi}:\mathbb{R}^d\times\mathbb{R}_+\longrightarrow\mathbb{R}^d$ be a uniformly Lipschitz mapping, and assume that there is $t^*>0$ such that the mapping $\boldsymbol{\Phi}_t:\mathbb{R}^d\ni\boldsymbol{\xi}\longmapsto\boldsymbol{\Phi}_t(\boldsymbol{\xi}):=\boldsymbol{\Phi}(\boldsymbol{\xi},t)\in\mathbb{R}^d$ is invertible for all $t\in[0,t^*]$. Let $\boldsymbol{v}_{\mathrm{A}}:\mathbb{R}^d\times[0,t^*]\longrightarrow\mathbb{R}^d$ be the vector field implicitly defined by

$$\boldsymbol{v}_{\mathrm{A}}(\boldsymbol{\Phi}(\boldsymbol{\xi},t),t):=\partial_t\boldsymbol{\Phi}(\boldsymbol{\xi},t),\quad\forall(\boldsymbol{\xi},t)\in\mathbb{R}\times[0,t^*].\tag{6}$$

This definition makes sense owing to the inversibility assumption on the mapping $\boldsymbol{\Phi}_t$; actually (6) is equivalent to $\boldsymbol{v}_{\mathrm{A}}(\boldsymbol{x},t):=\partial_t\boldsymbol{\Phi}(\boldsymbol{\Phi}_t^{-1}(\boldsymbol{x}),t)$ for any $t\in[0,t^*]$.

Lemma 3 (Liouville's formula) *Let* $\mathbb{J}(\boldsymbol{\xi},t)=\nabla_{\boldsymbol{\xi}}\boldsymbol{\Phi}(\boldsymbol{\xi},t)$ *be the Jacobian matrix of* $\boldsymbol{\Phi}$*, then*

$$\partial_t\det(\mathbb{J}(\boldsymbol{\xi},t))=(\nabla\cdot\boldsymbol{v}_{\mathrm{A}})(\boldsymbol{\Phi}(\boldsymbol{\xi},t),t)\det(\mathbb{J}(\boldsymbol{\xi},t)).\tag{7}$$

Note that the expression $(\nabla\cdot\boldsymbol{v}_{\mathrm{A}})(\boldsymbol{\Phi}(\boldsymbol{\xi},t),t)$ in (7) should not be confused with $\nabla\cdot(\boldsymbol{v}_{\mathrm{A}}(\boldsymbol{\Phi}(\boldsymbol{\xi},t),t))$.

3.2 Arbitrary Lagrangian Eulerian Formulation

The following result is the main motivation for the arbitrary Lagrangian Eulerian formulation that we are going to use in the paper.

Lemma 4 *The following identity holds in the distribution sense (in time) over the interval* $[0,t^*]$ *for every function* $\psi \in C_0^0(\mathbb{R}^d;\mathbb{R})$ *(with the notation* $\varphi(\boldsymbol{x},t) := \psi(\boldsymbol{\Phi}_t^{-1}(\boldsymbol{x}))$*):*

$$\partial_t \int_{\mathbb{R}^d} \boldsymbol{u}(\boldsymbol{x},t)\varphi(\boldsymbol{x},t)\,\mathrm{d}\boldsymbol{x} = \int_{\mathbb{R}^d} \nabla\cdot(\boldsymbol{u}\otimes\boldsymbol{v}_{\mathrm{A}} - \boldsymbol{f}(\boldsymbol{u}))\varphi(\boldsymbol{x},t)\,\mathrm{d}\boldsymbol{x}. \tag{8}$$

Proof Using the chain rule and Lemma 3, we have

$$\begin{aligned}
&\partial_t \int_{\mathbb{R}^d} \boldsymbol{u}(\boldsymbol{x},t)\varphi(\boldsymbol{x},t)\,\mathrm{d}\boldsymbol{x} = \partial_t \int_{\mathbb{R}^d} \boldsymbol{u}(\boldsymbol{\Phi}_t(\boldsymbol{\xi}),t)\det(\mathbb{J}(\boldsymbol{\xi},t))\psi(\boldsymbol{\xi})\,\mathrm{d}\boldsymbol{\xi}\\
&= \int_{\mathbb{R}^d} \big\{\partial_t(\boldsymbol{u}(\boldsymbol{\Phi}_t(\boldsymbol{\xi}),t))\det(\mathbb{J}(\boldsymbol{\xi},t)) + \boldsymbol{u}(\boldsymbol{\Phi}_t(\boldsymbol{\xi}),t)\partial_t(\det(\mathbb{J}(\boldsymbol{\xi},t)))\big\}\psi(\boldsymbol{\xi})\,\mathrm{d}\boldsymbol{\xi}\\
&= \int_{\mathbb{R}^d} \big\{(\partial_t\boldsymbol{u})(\boldsymbol{\Phi}_t(\boldsymbol{\xi}),t) + \partial_t\boldsymbol{\Phi}(\boldsymbol{\xi},t)\cdot(\nabla\boldsymbol{u})(\boldsymbol{\Phi}_t(\boldsymbol{\xi}),t)\big\}\det(\mathbb{J}(\boldsymbol{\xi},t))\psi(\boldsymbol{\xi})\,\mathrm{d}\boldsymbol{\xi}\\
&\qquad + \int_{\mathbb{R}^d} \boldsymbol{u}(\boldsymbol{\Phi}_t(\boldsymbol{\xi}),t)(\nabla\cdot\boldsymbol{v}_{\mathrm{A}})(\boldsymbol{\Phi}_t(\boldsymbol{\xi}),t)\det(\mathbb{J}(\boldsymbol{\xi},t))\psi(\boldsymbol{\xi})\,\mathrm{d}\boldsymbol{\xi}.
\end{aligned}$$

Then using (1) and the definition of the vector field $\boldsymbol{v}_{\mathrm{A}}$ yields

$$\begin{aligned}
\partial_t \int_{\mathbb{R}^d} \boldsymbol{u}(\boldsymbol{x},t)\varphi(\boldsymbol{x},t)\,\mathrm{d}\boldsymbol{x} &= \int_{\mathbb{R}^d} -\nabla\cdot\boldsymbol{f}(\boldsymbol{u})(\boldsymbol{\Phi}_t(\boldsymbol{\xi}),t)\psi(\boldsymbol{\xi})\det(\mathbb{J}(\boldsymbol{\xi},t))\,\mathrm{d}\boldsymbol{\xi}\\
&\quad + \int_{\mathbb{R}^d} \big\{\boldsymbol{v}_{\mathrm{A}}(\boldsymbol{\Phi}_t(\boldsymbol{\xi}),t)\cdot(\nabla\boldsymbol{u})(\boldsymbol{\Phi}_t(\boldsymbol{\xi}),t)\\
&\quad + (\nabla\cdot\boldsymbol{v}_{\mathrm{A}})(\boldsymbol{\Phi}_t(\boldsymbol{\xi}),t)\boldsymbol{u}(\boldsymbol{\Phi}_t(\boldsymbol{\xi}),t)\big\}\psi(\boldsymbol{\xi})\det(\mathbb{J}(\boldsymbol{\xi},t))\,\mathrm{d}\boldsymbol{\xi}\\
&= \int_{\mathbb{R}^d} \big\{-\nabla\cdot\boldsymbol{f}(\boldsymbol{u})(\boldsymbol{\Phi}_t(\boldsymbol{\xi}),t) + \nabla\cdot(\boldsymbol{u}\otimes\boldsymbol{v}_{\mathrm{A}})(\boldsymbol{\Phi}_t(\boldsymbol{\xi}),t)\big\}\psi(\boldsymbol{\xi})\det(\mathbb{J}(\boldsymbol{\xi},t))\,\mathrm{d}\boldsymbol{\xi}.
\end{aligned}$$

We conclude by making the change of variable $\boldsymbol{x} = \boldsymbol{\Phi}(\boldsymbol{\xi},t)$. □

We now state a result regarding the notion of entropy solution in the ALE framework. The proof of this result is similar to that of Lemma 4.

Lemma 5 *Let* $(\eta,\boldsymbol{q})$ *be an entropy pair for* (1)*. The following inequality holds in the distribution sense (in time) over the interval* $[0,t^*]$ *for every non-negative function* $\psi \in C_0^0(\mathbb{R}^d;\mathbb{R}_+)$ *(with the notation* $\varphi(\boldsymbol{x},t) := \psi(\boldsymbol{\Phi}_t^{-1}(\boldsymbol{x}))$*):*

$$\partial_t \int_{\mathbb{R}^d} \eta(\boldsymbol{u}(\boldsymbol{x},t))\varphi(\boldsymbol{x},t)\,\mathrm{d}\boldsymbol{x} \le \int_{\mathbb{R}^d} \nabla\cdot(\eta(\boldsymbol{u})\boldsymbol{v}_{\mathrm{A}} - \boldsymbol{q}(\boldsymbol{u}))\varphi(\boldsymbol{x},t)\,\mathrm{d}\boldsymbol{x}.$$

4 Arbitrary Lagrangian Eulerian Algorithm

We describe in this section the ALE algorithm to approximate the solution of (1). We use continuous finite elements and explicit time stepping. We use two different discrete settings: one for the mesh motion and one for the approximation of (1).

4.1 Geometric Finite Elements and Mesh

Let $(\mathscr{T}_h^0)_{h>0}$ be a shape-regular sequence of matching meshes. The symbol ^{0} in $\mathscr{T}_h^0$ refers to the initial configuration of the meshes. The meshes will deform over time, in a way that has yet to be defined, and we are going to use the symbol ^{n} to say that $\mathscr{T}_h^n$ is the mesh at time t^n for a given $h > 0$. We assume that the elements in the mesh cells are generated from a finite number of reference elements denoted $\widehat{K}_1, \dots, \widehat{K}_\varpi$. For instance, $\mathscr{T}_h^0$ could be composed of a combination of triangles and parallelograms in two space dimensions ($\varpi = 2$ in this case); the mesh $\mathscr{T}_h^0$ could also be composed of a combination of tetrahedra, parallelepipeds, and triangular prisms in three space dimensions ($\varpi = 3$ in this case). The diffeomorphism mapping $\widehat{K}_r$ to an arbitrary element $K \in \mathscr{T}_h^n$ is denoted $T_K^n : \widehat{K}_r \longrightarrow K$ and its Jacobian matrix is denoted $\mathbb{J}_K^n$, $1 \le r \le \varpi$. We now introduce a set of reference Lagrange finite elements $\{(\widehat{K}_r, \widehat{P}_r^{\text{geo}}, \widehat{\Sigma}_r^{\text{geo}})\}_{1\le r\le\varpi}$ (the index $r \in \{1\colon\varpi\}$ will be omitted in the rest of the paper to alleviate the notation). Letting $n_{\text{sh}}^{\text{geo}} := \dim \widehat{P}^{\text{geo}}$, we denote by $\{\widehat{\boldsymbol{a}}_i\}_{i\in\{1:n_{\text{sh}}^{\text{geo}}\}}$ and $\{\widehat{\theta}_i^{\text{geo}}\}_{i\in\{1:n_{\text{sh}}^{\text{geo}}\}}$ the Lagrange nodes of $\widehat{K}$ and the associated Lagrange shape functions.

The unique purpose of the geometric reference element $\{(\widehat{K}, \widehat{P}^{\text{geo}}, \widehat{\Sigma}^{\text{geo}})$ is to construct the geometric transformation T_K^n. Let $\{\boldsymbol{a}_i^n\}_{i\in\{1:I^{\text{geo}}\}}$ be the collection of all the Lagrange nodes in the mesh $\mathscr{T}_h^n$. The Lagrange nodes are organized in cells by means of the geometric connectivity array $\mathsf{j}^{\text{geo}} : \mathscr{T}_h^n \times \{1{:}n_{\text{sh}}^{\text{geo}}\} \longrightarrow \{1{:}I^{\text{geo}}\}$ (assumed to be independent of the time index n). Given a mesh cell $K \in \mathscr{T}_h^n$, the connectivity array is defined such that $\{\boldsymbol{a}_{\mathsf{j}^{\text{geo}}(i,K)}^n\}_{i\in\{1:n_{\text{sh}}^{\text{geo}}\}}$ is the set of the Lagrange nodes describing K^n. More precisely, upon defining the geometric transformation $T_K^n : \widehat{K} \longrightarrow K$ at time t^n by

$$T_K^n(\widehat{\boldsymbol{x}}) = \sum_{i\in\{1:n_{\text{sh}}^{\text{geo}}\}} \boldsymbol{a}_{\mathsf{j}^{\text{geo}}(i,K)}^n \widehat{\theta}_i^{\text{geo}}(\widehat{\boldsymbol{x}}) \tag{9}$$

we have $K := T_K^n(\widehat{K})$. In other words the geometric transformation is fully described by the motion of geometric Lagrange nodes. Recall that constructing the Jacobian matrix $\mathbb{J}_K^n$ from (9) is an elementary operation for any finite element code.

4.2 Approximating Finite Elements

We now introduce a set of reference finite elements $\{(\widehat{K}_r, \widehat{P}_r, \widehat{\Sigma}_r)\}_{1\le r\le\varpi}$ which we are going to use to construct an approximate solution to (1) (the index $r \in \{1\colon\varpi\}$ will be omitted in the rest of the paper to alleviate the notation). The shape functions on the reference element are denoted $\{\widehat{\theta}_i\}_{i\in\{1:n_{\rm sh}\}}$. We assume that the basis $\{\widehat{\theta}_i\}_{i\in\{1:n_{\rm sh}\}}$ has the following key properties:

$$\widehat{\theta}_i(\boldsymbol{x}) \geq 0, \qquad \sum_{i\in\{1:n_{\rm sh}\}} \widehat{\theta}_i(\widehat{\boldsymbol{x}}) = 1, \quad \forall \widehat{\boldsymbol{x}} \in \widehat{K}. \tag{10}$$

These properties hold true for linear Lagrange elements and for Bernstein-Bezier finite elements, see, e.g., [16, Chap. 2], [1].

Given the mesh $\mathscr{T}_h^n$, we denote by D^n the computational domain generated by $\mathscr{T}_h^n$ and we define the scalar-valued space

$$P(\mathscr{T}_h^n) := \{v \in \mathscr{C}^0(D^n;\mathbb{R}) \mid v_{|K}\circ T_K^n \in \widehat{P},\ \forall K \in \mathscr{T}_h^n\}.$$

We also introduce the vector-valued spaces

$$\boldsymbol{P}_d(\mathscr{T}_h^n) := [P(\mathscr{T}_h^n)]^d, \quad \text{and} \quad \boldsymbol{P}_m(\mathscr{T}_h^n) := [P(\mathscr{T}_h^n)]^m.$$

We are going to approximate the ALE velocity in $\boldsymbol{P}_d(\mathscr{T}_h^n)$ and the solution of (1) in $\boldsymbol{P}_m(\mathscr{T}_h^n)$. The global shape functions in $P(\mathscr{T}_h^n)$ are denoted by $\{\psi_i^n\}_{i\in\{1:I\}}$. Recall that these functions form a basis of $P(\mathscr{T}_h^n)$. Let $\mathsf{j} : \mathscr{T}_h^n \times \{1{:}n_{\rm sh}\} \longrightarrow \{1{:}I\}$ be the connectivity array, assumed to be independent of n. This array is defined such that

$$\psi^n_{\mathsf{j}(i,K)}(\boldsymbol{x}) = \widehat{\theta}_i((T_K^n)^{-1}(\boldsymbol{x})), \quad \forall i \in \{1{:}n_{\rm sh}\},\ \forall K \in \mathscr{T}_h^n.$$

This definition together with (10) implies that

$$\psi_i^n(\boldsymbol{x}) \geq 0, \qquad \sum_{i\in\{1:I\}} \psi_i^n(\boldsymbol{x}) = 1, \quad \forall \boldsymbol{x} \in \mathbb{R}^d.$$

We denote by S_i^n the support of ψ_i^n and by $|S_i^n|$ the measure of S_i, $i \in \{1{:}I\}$. We also define $S_{ij}^n := S_i^n \cap S_j^n$ the intersection of the two supports S_i^n and S_j^n. Let E be a union of cells in $\mathscr{T}_h^n$; we define $\mathscr{I}(E) := \{j \in \{1{:}I\} \mid |S_j^n \cap E| \neq 0\}$ the set that contains the indices of all the shape functions whose support on E is of nonzero measure. Note that the index set $\mathscr{I}(E)$ does not depend on the time index n since we have assumed that the connectivity of the degrees of freedom is fixed once for all. We are going to regularly invoke $\mathscr{I}(K)$ and $\mathscr{I}(S_i^n)$ and the partition of unity property: $\sum_{i\in\mathscr{I}(K)} \psi_i^n(\boldsymbol{x}) = 1$ for all $\boldsymbol{x} \in K$.

Lemma 6 *For all $K \in \mathscr{T}_h^n$, all $\boldsymbol{x} \in K$, and all $\boldsymbol{v}_h := \sum_{i\in\{1:I\}} \boldsymbol{V}_i\psi_i^n \in \boldsymbol{P}_m(\mathscr{T}_h^n)$, $\boldsymbol{v}_h(\boldsymbol{x})$ is in the convex hull of $(\boldsymbol{V}_i)_{i\in\mathscr{I}(K)}$ (henceforth denoted* $\mathrm{conv}(\boldsymbol{V}_i)_{i\in\mathscr{I}(K)}$*). Moreover for any convex set A in $\mathbb{R}^m$, we have*

$$\big((\boldsymbol{V}_i)_{i\in\mathscr{I}(K)} \in A\big) \Rightarrow (\boldsymbol{v}_h(\boldsymbol{x}) \in A,\ \forall \boldsymbol{x} \in K)\,. \tag{11}$$

4.3 ALE Algorithm

Let $\mathscr{T}_h^0$ be the mesh at the initial time $t = 0$. Let $(\mathfrak{m}_i^0)_{i\in\{1:I\}}$ be the approximations of the mass of the shape functions at time t^0 defined by $\mathfrak{m}_i^0 = m_i^0 := \int_{\mathbb{R}^d} \psi_i^0(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$. Let $\boldsymbol{u}_{h0} := \sum_{i\in\{1:I\}} \mathsf{U}_i^0 \psi_i^0 \in \boldsymbol{P}_m(\mathscr{T}_h^0)$ be a reasonable approximation of the initial data $\boldsymbol{u}_0$ (we shall make a more precise statement later).

Let $\mathscr{T}_h^n$ be the mesh at time t^n, $(\mathfrak{m}_i^n)_{1\le i\le I}$ be the approximations of the mass of the shape functions at time t^n, and $\boldsymbol{u}_h^n := \sum_{i\in\{1:I\}} \mathsf{U}_i^n \psi_i^n \in \boldsymbol{P}_m(\mathscr{T}_h^n)$ be the approximation of $\boldsymbol{u}$ at time t^n. We denote by $\mathfrak{M}^{L,n}$ the approximate lumped matrix, i.e., $\mathfrak{M}_{ij}^{L,n} = \mathfrak{m}_i^n \delta_{ij}$. We now make the assumption that the given ALE velocity field is a member of $\boldsymbol{P}_d(\mathscr{T}_h^n)$, i.e., $\boldsymbol{W}^n = \sum_{i\in\{1:I\}} \mathsf{W}_i^n \psi_i^n \in \boldsymbol{P}_d(\mathscr{T}_h^n)$. Then the Lagrange nodes of the mesh are moved by using the following rule:

$$\boldsymbol{a}_i^{n+1} = \boldsymbol{a}_i^n + \tau \boldsymbol{W}^n(\boldsymbol{a}_i^n). \tag{12}$$

This fully defines the mesh $\mathscr{T}_h^{n+1}$ as explained at the end of Sect. 4.1. Upon introducing $\psi^{\mathrm{geo}}_{\mathrm{j}^{\mathrm{geo}}(i,K)}(\boldsymbol{\xi}) := \widehat{\theta}_i((T_K^n)^{-1}(\boldsymbol{\xi}))$ and $\boldsymbol{a}_i(t) = \boldsymbol{a}_i^n + (t - t^n)\boldsymbol{W}^n(\boldsymbol{a}_i^n)$ for $t \in [t^n, t^n + \tau]$, this also defines the ALE mapping

$$\boldsymbol{\Phi}_{t|K}(\boldsymbol{\xi}) = \sum_{i\in\{1:n_{\mathrm{sh}}^{\mathrm{geo}}\}} \boldsymbol{a}_{\mathrm{j}^{\mathrm{geo}}(i,K)}(t)\psi^{\mathrm{geo}}_{\mathrm{j}^{\mathrm{geo}}(i,K)}(\boldsymbol{\xi}), \qquad \forall \boldsymbol{\xi} \in K,\ \forall K \in \mathscr{T}_h^n. \tag{13}$$

We now estimate the mass of the shape function $\psi_i^{n+1} := \psi_i^n \circ \boldsymbol{\Phi}_{t^{n+1}}$. Of course we could use $m_i^{n+1} = \int_{\mathbb{R}^d} \psi_i^{n+1}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$. This option leads to many difficulties that are explored in [13]; in particular, extending the method to high-order in time with this definition is problematic. To have a method that is compatible with higher-order strong stability preserving (SSP) time stepping techniques, we define $\mathfrak{m}_i^{n+1}$ by approximating the following identity with a first-order quadrature rule:

$$\int_{\mathbb{R}^d} \psi^{n+1}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} - \int_{\mathbb{R}^d} \psi^n(\boldsymbol{\xi})\,\mathrm{d}\boldsymbol{\xi} = \int_{\mathbb{R}^d} \psi^n(\boldsymbol{\xi}) \left[\int_{t^n}^{t^{n+1}} \partial_\zeta \det(\mathbb{J}(\boldsymbol{\xi},\zeta))\,\mathrm{d}\zeta\right] \mathrm{d}\boldsymbol{\xi}.$$

Note that $\det(\mathbb{J}(\boldsymbol{\xi},\zeta))$ is a polynomial function of ζ of degree d. The first-order approximation of the integral with respect to ζ in the above expression gives:

$$\mathfrak{m}_i^{n+1} = \mathfrak{m}_i^n + \tau \int_{S_i^n} \psi_i^n(\boldsymbol{x}) \nabla\cdot \boldsymbol{W}^n(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}. \tag{14}$$

Taking inspiration from (8), we propose to compute $\boldsymbol{u}_h^{n+1}$ as follows:

$$\frac{\mathrm{m}_i^{n+1}\mathbf{U}_i^{n+1} - \mathrm{m}_i^n\mathbf{U}_i^n}{\tau} - \sum_{j\in\mathscr{I}(S_i^n)} d_{ij}^n\mathbf{U}_j^n \\ + \int_{\mathbb{R}^d} \nabla\cdot\Big(\sum_{j\in\{1:I\}} (\boldsymbol{f}(\mathbf{U}_j^n) - \mathbf{U}_j^n\otimes\mathbf{W}_j^n)\psi_j^n(\boldsymbol{x})\Big)\psi_i^n(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = 0, \quad (15)$$

where $\boldsymbol{u}_h^{n+1} := \sum_{i\in\{1:I\}}\mathbf{U}_i^{n+1}\psi_i^{n+1} \in \boldsymbol{P}_m(\mathscr{T}_h^{n+1})$. Notice that we have replaced the consistent mass matrix by an approximation of the lumped mass matrix to approximate the time derivative. The coefficient d_{ij}^n is an artificial viscosity for the pair of degrees of freedom (i, j) that will be identified by proceeding as in [12]. We henceforth assume that $d_{ij}^n = 0$ if $j \notin \mathscr{I}(S_i^n)$ and

$$d_{ij}^n \geq 0, \text{ if } i \neq j, \qquad d_{ij}^n = d_{ji}^n, \quad \text{and} \quad d_{ii} := \sum_{i\neq j\in\mathscr{I}(S_i^n)} -d_{ji}^n. \quad (16)$$

The entire process is described in Algorithm 4.

Let us reformulate (15) in a form that is more suitable for computations. Let us introduce the vector-valued coefficients

$$\boldsymbol{c}_{ij}^n := \int_{S_i^n} \nabla\psi_j^n(\boldsymbol{x})\psi_i^n(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}. \quad (17)$$

We define the unit vector $\boldsymbol{n}_{ij}^n := \frac{\boldsymbol{c}_{ij}^n}{\|\boldsymbol{c}_{ij}^n\|_{\ell^2}}$. Then we rewrite (15) as follows:

$$\frac{\mathrm{m}_i^{n+1}\mathbf{U}_i^{n+1} - \mathrm{m}_i^n\mathbf{U}_i^n}{\tau} + \sum_{j\in\mathscr{I}(S_i^n)} (\boldsymbol{f}(\mathbf{U}_j^n) - \mathbf{U}_j^n\otimes\mathbf{W}_j^n)\cdot\boldsymbol{c}_{ij}^n - d_{ij}^n\mathbf{U}_j^n = 0. \quad (18)$$

It will be shown in the proof of Theorem 1 that an admissible choice for d_{ij}^n is

$$d_{ij}^n = \max(\lambda_{\max}(\boldsymbol{g}_j^n, \boldsymbol{n}_{ij}^n, \mathbf{U}_i^n, \mathbf{U}_j^n)\|\boldsymbol{c}_{ij}^n\|_{\ell^2}, \lambda_{\max}(\boldsymbol{g}_i^n, \boldsymbol{n}_{ji}^n, \mathbf{U}_j^n, \mathbf{U}_i^n)\|\boldsymbol{c}_{ji}^n\|_{\ell^2}). \quad (19)$$

where $\lambda_{\max}(\boldsymbol{g}_j^n, \boldsymbol{n}_{ij}^n, \mathbf{U}_i^n, \mathbf{U}_j^n)$ is the largest wave speed in the following one-dimensional Riemann problem with the flux $\boldsymbol{g}_j^n(\boldsymbol{v}) := \boldsymbol{f}(\boldsymbol{v}) - \boldsymbol{v}\otimes\mathbf{W}_j^n$:

$$\partial_t\boldsymbol{v} + \partial_x(\boldsymbol{g}_j^n(\boldsymbol{v})\cdot\boldsymbol{n}_{ij}^n) = 0, \quad (x, t)\in\mathbb{R}\times\mathbb{R}_+, \quad \boldsymbol{v}(x, 0) = \begin{cases}\mathbf{U}_i^n & \text{if } x < 0\\ \mathbf{U}_j^n & \text{if } x > 0.\end{cases} \quad (20)$$

Remark 2 (Fastest wave speed) The fastest wave speed in (20) can be obtained by estimating the fastest wave speed in the Riemann problem (2) with the flux $\boldsymbol{f}(\boldsymbol{v})\cdot\boldsymbol{n}_{ij}^n$ and the initial data $(\mathbf{U}_i^n, \mathbf{U}_j^n)$. Let $\lambda_L(\boldsymbol{f}, \boldsymbol{n}_{ij}^n, \mathbf{U}_i^n, \mathbf{U}_j^n)$ and $\lambda_R(\boldsymbol{f}, \boldsymbol{n}_{ij}^n, \mathbf{U}_i^n, \mathbf{U}_j^n)$ be the speed of the leftmost and the rightmost waves in (2), respectively. Then

$$\lambda_{\max}(\boldsymbol{g}_j^n, \boldsymbol{n}_{ij}^n, \mathbf{U}_i^n, \mathbf{U}_j^n) = \max(|\lambda_L(\boldsymbol{f}, \boldsymbol{n}_{ij}^n, \mathbf{U}_i^n, \mathbf{U}_j^n) - \mathbf{W}_j^n \cdot \boldsymbol{n}_{ij}^n|, \\ |\lambda_R(\boldsymbol{f}, \boldsymbol{n}_{ij}^n, \mathbf{U}_i^n, \mathbf{U}_j^n) - \mathbf{W}_j^n \cdot \boldsymbol{n}_{ij}^n|). \tag{21}$$

A fast algorithm to compute $\lambda_L(\boldsymbol{f}, \boldsymbol{n}_{ij}^n, \mathbf{U}_i^n, \mathbf{U}_j^n)$ and $\lambda_R(\boldsymbol{f}, \boldsymbol{n}_{ij}^n, \mathbf{U}_i^n, \mathbf{U}_j^n)$ for the compressible Euler equations is given in [11]; see also [20].

Algorithm 4

```
Require: $\boldsymbol{u}_h^0$ and $\mathfrak{M}^{L,0}$
 1: while $t^n < T$ do
 2:     Use CFL condition to estimate $\tau$.
 3:     if $t^n + \tau > T$ then
 4:         $\tau \leftarrow T - t^n$
 5:     end if
 6:     Estimate/choose $\boldsymbol{W}^n$ and make sure that the transformation $\boldsymbol{\Phi}_t$ defined in (13) is invertible
    over the interval $[t^n, t^{n+1}]$.
 7:     Move mesh from $t^n$ to $t^{n+1}$ using (12).
 8:     Compute $\mathfrak{m}_i^{n+1}$, see (14). Check $\mathfrak{m}_i^{n+1} > 0$; otherwise, go to step 6, reduce $\tau$.
 9:     Compute $\boldsymbol{c}_{ij}^n$ as in (17).
10:     Compute $d_{ij}^n$, see (19) and (16).
11:     Check $1 - \sum_{i \neq j \in \mathscr{I}(S_i^n)} 2d_{ij}^n \frac{\tau}{\mathfrak{m}_i^{n+1}}$ positive. Otherwise, go to step 6 and reduce $\tau$.
12:     Compute $\boldsymbol{u}_h^{n+1}$ by using (18).
13:     $t^n \leftarrow t^n + \tau$
14: end while
```

Since it is important to compare $\mathbf{U}_j^{n+1}$ and $\mathbf{U}_j^n$ to establish the invariant domain property, we rewrite the scheme in a form that is more suitable for this purpose.

Lemma 7 (Non-conservative form) *The scheme* (15) *is equivalent to*

$$\mathfrak{m}_i^{n+1} \frac{\boldsymbol{U}_i^{n+1} - \boldsymbol{U}_i^n}{\tau} = \sum_{j \in \mathscr{I}(S_i^n)} ((\boldsymbol{U}_j^n - \boldsymbol{U}_i^n) \otimes \boldsymbol{W}_j^n - \boldsymbol{f}(\boldsymbol{U}_j^n)) \cdot \boldsymbol{c}_{ij}^n + d_{ij}^n \boldsymbol{U}_j^n, \tag{22}$$

Proof We rewrite (18) as follows:

$$\mathfrak{m}_i^{n+1} \frac{\mathbf{U}_i^{n+1} - \mathbf{U}_i^n}{\tau} + \frac{\mathfrak{m}_i^{n+1} - \mathfrak{m}_i^n}{\tau} \mathbf{U}_i^n = \sum_{j \in \mathscr{I}(S_i^n)} (\mathbf{U}_j^n \otimes \mathbf{W}_j^n - \boldsymbol{f}(\mathbf{U}_j^n)) \cdot \boldsymbol{c}_{ij}^n + d_{ij}^n \mathbf{U}_j^n,$$

Then, recalling the expression $\boldsymbol{W}^n = \sum_{i \in \{1:I\}} \mathbf{W}_i^n \psi_i^n$, and using (14), we infer that $\mathfrak{m}_i^{n+1} = \mathfrak{m}_i^n + \tau \sum_{j \in \mathscr{I}(S_i^n)} \mathbf{W}_j^n \cdot \boldsymbol{c}_{ij}^n$, which in turn implies that

$$(\mathfrak{m}_i^{n+1} - \mathfrak{m}_i^n) \mathbf{U}_i^n = \tau \mathbf{U}_i^n \sum_{j \in \mathscr{I}(S_i^n)} \mathbf{W}_j^n \cdot \boldsymbol{c}_{ij}^n = \tau \sum_{j \in \mathscr{I}(S_i^n)} (\mathbf{U}_i^n \otimes \mathbf{W}_j^n) \cdot \boldsymbol{c}_{ij}^n.$$

□

Remark 3 (*Other discretizations*) The method for computing the artificial diffusion is quite generic, i.e., it is not specific to continuous finite elements. The above method can be applied to any type of discretization that can be put into the form (18).

4.4 SSP Extension

Retaining the invariant domain property (see Sect. 5.1) and increasing the time accuracy can be done by using so-called Strong Stability Preserving (SSP) time discretization methods. The key is to achieve higher-order accuracy in time by making convex combination of solutions of forward Euler sub-steps. More precisely each time step of a SSP method is decomposed into substeps that are all forward Euler solutions, and the end of step solution is a convex combination of the intermediate solutions; see [8, 10, 14] for reviews on SPP techniques. Algorithm 5 illustrates one Euler step of the scheme. SSP techniques are useful when combined with reasonable limitation strategies since the resulting methods are high-order, both in time and space, and invariant domain preserving.

Algorithm 5 Euler step

```
Require: $\mathscr{T}_h^0$, $\boldsymbol{u}_h^0$, ($\mathfrak{m}^0$ or $m^0$), $\boldsymbol{W}^0$, $\tau$
Compute $\widetilde{\boldsymbol{a}}_i^1 = \boldsymbol{a}_i^0 + \tau \boldsymbol{W}^0$, ($\widetilde{\mathfrak{m}}^1$ or $\widetilde{m}^1$), $\widetilde{\boldsymbol{u}}_h^1$, and build new mesh $\widetilde{\mathscr{T}}_h^1$
return $\widetilde{\mathscr{T}}_h^1$, $\widetilde{\boldsymbol{u}}_h^1$, ($\widetilde{\mathfrak{m}}^1$ or $\widetilde{m}^1$)
```

As an illustration we describe the SSP RK3 implementation of the scheme in Algorithm 6. Generalizations to other SSP techniques are left to the reader.

Algorithm 6 SPP RK3

```
Require: $\mathscr{T}_h^0$, $\boldsymbol{u}_h^0$, $\mathfrak{m}^0$, $t^0$
Define the ALE velocity $\boldsymbol{W}^0$ at $t^0$
Call Euler step($\mathscr{T}_h^0$, $\boldsymbol{u}_h^0$, $\mathfrak{m}^0$, $\boldsymbol{W}^0$, $\tau$, $\mathscr{T}_h^1$, $\boldsymbol{u}_h^1$, $\mathfrak{m}^1$)
Define the ALE velocity $\boldsymbol{W}^1$ at $t^0 + \tau$
Call Euler step($\mathscr{T}_h^1$, $\boldsymbol{u}_h^1$, $\mathfrak{m}^1$, $\boldsymbol{W}^1$, $\tau$, $\widetilde{\mathscr{T}}_h^2$, $\widetilde{\boldsymbol{u}}_h^2$, $\widetilde{\mathfrak{m}}^2$)
Set $\boldsymbol{a}^2 = \frac{3}{4}\boldsymbol{a}^0 + \frac{1}{4}\widetilde{\boldsymbol{a}}^2$, $\mathfrak{m}^2 = \frac{3}{4}\mathfrak{m}^0 + \frac{1}{4}\widetilde{\mathfrak{m}}^2$, build mesh $\mathscr{T}_h^2$, $\boldsymbol{u}_h^2 = \frac{3}{4}\frac{\mathfrak{m}^0}{\mathfrak{m}^2}\boldsymbol{u}_h^0 + \frac{1}{4}\frac{\widetilde{\mathfrak{m}}^2}{\mathfrak{m}^2}\widetilde{\boldsymbol{u}}_h^2$
Define the ALE velocity $\boldsymbol{W}^2$ at $t^0 + \frac{1}{2}\tau$
Call Euler step($\mathscr{T}_h^2$, $\boldsymbol{u}_h^2$, $\mathfrak{m}^2$, $\boldsymbol{W}^2$, $\tau$, $\widetilde{\mathscr{T}}_h^3$, $\widetilde{\boldsymbol{u}}_h^3$, $\widetilde{\mathfrak{m}}^3$)
Set $\boldsymbol{a}^3 = \frac{1}{3}\boldsymbol{a}^0 + \frac{2}{3}\widetilde{\boldsymbol{a}}^3$, $\mathfrak{m}^3 = \frac{1}{3}\mathfrak{m}^0 + \frac{2}{3}\widetilde{\mathfrak{m}}^3$, build mesh $\mathscr{T}_h^3$, $\boldsymbol{u}_h^3 = \frac{1}{3}\frac{\mathfrak{m}^0}{\mathfrak{m}^3}\boldsymbol{u}_h^0 + \frac{2}{3}\frac{\widetilde{\mathfrak{m}}^3}{\mathfrak{m}^3}\widetilde{\boldsymbol{u}}_h^3$
return $\mathscr{T}_h^3$, $\boldsymbol{u}_h^3$, $\mathfrak{m}^3$, $t^1 = t^0 + dt$
```

Note that $\boldsymbol{u}_h^2$ is a convex combination of $\boldsymbol{u}_h^0$ and $\widetilde{\boldsymbol{u}}_h^2$ since $1 = \frac{3}{4}\frac{\mathfrak{m}_i^0}{\mathfrak{m}_i^2} + \frac{1}{4}\frac{\widetilde{\mathfrak{m}}_i^2}{\mathfrak{m}_i^2}$. The same observation holds true for $\boldsymbol{u}_h^3$, i.e., $\boldsymbol{u}_h^3$ is a convex combination of $\boldsymbol{u}_h^0$ and $\widetilde{\boldsymbol{u}}_h^3$ since $1 = \frac{1}{3}\frac{\mathfrak{m}_i^0}{\mathfrak{m}_i^3} + \frac{2}{3}\frac{\widetilde{\mathfrak{m}}_i^3}{\mathfrak{m}_i^3}$, for any $i \in \{1{:}I\}$.

5 Stability Analysis

We establish the conservation and the invariant domain property of the scheme (15).

5.1 Invariant Domain Property

We first discuss the conservation properties of the scheme.

Lemma 8 *For the scheme* (15)*, the quantity* $\sum_{i\in\{1:I\}} \mathfrak{m}_i^n \boldsymbol{U}_i^n$ *is independent of* n*, i.e., the total mass is conserved.*

We can now prove a result somewhat similar in spirit to Thm 5.1 from [7], although the present result is more general since it applies to any hyperbolic system. We define the local minimum mesh size $\underline{h}_{ij}^n$ associated with an ordered pair of shape functions (ψ_i^n, ψ_j^n) as follows:

$$\underline{h}_{ij}^n := \frac{1}{\| \|\nabla\varphi_j\|_{\ell^2} \|_{L^\infty(S_{ij}^n)}},$$

where $S_{ij}^n = S_i^n \cap S_j^n$. We then define a local mesh size and a local mesh structure parameter κ_i^n by setting

$$\underline{h}_i^n = \min_{j\in\mathscr{I}(S_i^n)} \underline{h}_{ij}^n, \qquad \kappa_i^n := \frac{\sum_{i\neq j\in\mathscr{I}(S_i^n)} \int_{S_{ij}^n} \psi_i^n(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}}{\int_{S_i^n} \psi_i^n(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}}.$$

Note that the upper estimate $\kappa_i^n \leq \max_{j\in\{1:I\}} \mathrm{card}(\mathscr{I}(S_j(0))) - 1$ implies that κ_i^n is uniformly bounded with respect to n and i.

Theorem 1 (Local invariance) *Let* $n \geq 0$*, and* $i \in \{1{:}I\}$*. Assume the CFL condition*

$$2\tau \frac{\lambda_{i,\max}^n}{\underline{h}_i^n} \kappa_i^n \frac{\mathfrak{m}_i^n}{\mathfrak{m}_i^{n+1}} \leq 1, \tag{23}$$

where $\lambda_{i,\max}^n := \max_{j\in\mathscr{I}(S_i^n)}(\lambda_{\max}(\boldsymbol{g}_j^n, \boldsymbol{n}_{ij}^n, \boldsymbol{U}_i^n, \boldsymbol{U}_j^n), \lambda_{\max}(\boldsymbol{g}_i^n, \boldsymbol{n}_{ji}^n, \boldsymbol{U}_j^n, \boldsymbol{U}_i^n))$*. Let* $B \subset \mathscr{A}_f$ *be a convex invariant set for the flux* $\boldsymbol{f}$*. If* $\{\boldsymbol{U}_j^n \mid j \in \mathscr{I}(S_i^n)\} \subset B$*, then* $\boldsymbol{U}_i^{n+1} \in B$*.*

Proof Let $i \in \{1{:}I\}$ and invoke (22) from Lemma 7 to express $\mathbf{U}_i^{n+1}$ as follows:

$$\mathbf{U}_i^{n+1} = \mathbf{U}_i^n + \frac{\tau}{\mathfrak{m}_i^{n+1}} \sum_{j\in\mathscr{I}(S_i^n)} ((\mathbf{U}_j^n - \mathbf{U}_i^n)\otimes\mathbf{W}_j^n - \boldsymbol{f}(\mathbf{U}_j^n))\cdot\boldsymbol{c}_{ij}^n + d_{ij}^n\mathbf{U}_j^n.$$

Since the partition of unity property implies that $\sum_{j\in\mathscr{I}(S_i^n)} \boldsymbol{c}_{ij}^n = 0$ and we have $\sum_{j\in\mathscr{I}(S_i^n)} d_{ij}^n = 0$ from (16), we can rewrite the above equation as follows:

$$\begin{aligned}
\mathbf{U}_i^{n+1} &= \mathbf{U}_i^n + \sum_{j\in\mathscr{I}(S_i^n)} d_{ij}^n(\mathbf{U}_i^n + \mathbf{U}_j^n)\\
&\qquad + \frac{\tau}{\mathfrak{m}_i^{n+1}} \sum_{j\in\mathscr{I}(S_i^n)} ((\mathbf{U}_j^n - \mathbf{U}_i^n)\otimes\mathbf{W}_j^n + \boldsymbol{f}(\mathbf{U}_i^n) - \boldsymbol{f}(\mathbf{U}_j^n))\cdot\boldsymbol{c}_{ij}^n\\
&= \mathbf{U}_i^n\left(1 + 2d_{ii}^n\frac{\tau}{\mathfrak{m}_i^{n+1}}\right) + \sum_{i\neq j\in\mathscr{I}(S_i^n)} d_{ij}^n(\mathbf{U}_i^n + \mathbf{U}_j^n)\\
&\qquad + \frac{\tau}{\mathfrak{m}_i^{n+1}} \sum_{i\neq j\in\mathscr{I}(S_i^n)} ((\mathbf{U}_j^n - \mathbf{U}_i^n)\otimes\mathbf{W}_j^n + \boldsymbol{f}(\mathbf{U}_i^n) - \boldsymbol{f}(\mathbf{U}_j^n))\cdot\boldsymbol{c}_{ij}^n.
\end{aligned}$$

Let us introduced the auxiliary state $\overline{\mathbf{U}}_{ij}^{n+1}$ defined by

$$\overline{\mathbf{U}}_{ij}^{n+1} = (\boldsymbol{f}(\mathbf{U}_i^n) - \boldsymbol{f}(\mathbf{U}_j^n) - (\mathbf{U}_i^n - \mathbf{U}_j^n)\otimes\mathbf{W}_j^n)\cdot\boldsymbol{n}_{ij}^n\frac{\|\boldsymbol{c}_{ij}^n\|_{\ell^2}}{2d_{ij}^n} + \frac{1}{2}(\mathbf{U}_i^n + \mathbf{U}_j^n),$$

where $\boldsymbol{n}_{ij}^n := \boldsymbol{c}_{ij}^n/\|\boldsymbol{c}_{ij}^n\|_{\ell^2}$. Then, provided we establish that

$$1 - \sum_{i\neq j\in\mathscr{I}(S_i^n)} 2d_{ij}^n\frac{\tau}{\mathfrak{m}_i^{n+1}} \geq 0,$$

we have proved that $\mathbf{U}_i^{n+1}$ is a convex combination of $\mathbf{U}_i^n$ and $(\overline{\mathbf{U}}_{ij}^{n+1})_{i\neq j\in\mathscr{I}(S_i^n)}$:

$$\mathbf{U}_i^{n+1} = \mathbf{U}_i^n\left(1 - \sum_{i\neq j\in\mathscr{I}(S_i^n)} 2d_{ij}^n\frac{\tau}{\mathfrak{m}_i^{n+1}}\right) + \frac{\tau}{\mathfrak{m}_i^{n+1}} \sum_{i\neq j\in\mathscr{I}(S_i^n)} 2d_{ij}^n\overline{\mathbf{U}}_{ij}^{n+1}. \tag{24}$$

Let us now consider the Riemann problem (20). Let $\boldsymbol{v}(\boldsymbol{g}_j^n, \boldsymbol{n}_{ij}^n, \mathbf{U}_i^n, \mathbf{U}_j^n)$ be the solution to (20) with $\boldsymbol{g}_j^n(\boldsymbol{v}) := \boldsymbol{f}(\boldsymbol{v}) - \boldsymbol{v}\otimes\mathbf{W}_j^n$. Let $\lambda_{\max}(\boldsymbol{g}_j^n, \boldsymbol{n}_{ij}^n, \mathbf{U}_i^n, \mathbf{U}_j^n)$ be the fastest wave speed in (20), see (21). Using the notation of Lemma 1, we then observe that $\overline{\mathbf{U}}_{ij}^{n+1} = \overline{\boldsymbol{v}}(t, \boldsymbol{g}_j^n, \boldsymbol{n}_{ij}^n, \mathbf{U}_i^n, \mathbf{U}_j^n)$ with $t = \frac{\|\boldsymbol{c}_{ij}^n\|_{\ell^2}}{2d_{ij}^n}$, provided $t\lambda_{\max}(\boldsymbol{g}_j^n, \boldsymbol{n}_{ij}^n, \mathbf{U}_i^n, \mathbf{U}_j^n) \leq \frac{1}{2}$. Note that the definition of d_{ij}^n, (19), implies that the condition $t\lambda_{\max}(\boldsymbol{g}_j^n, \boldsymbol{n}_{ij}^n, \mathbf{U}_i^n, \mathbf{U}_i^n) \leq \frac{1}{2}$ is satisfied. Since B is an invariant set for the flux $\boldsymbol{f}$, by Lemma 2, B is also an invariant set for the flux $\boldsymbol{g}_j^n$. Since, in addition, B contains the data $(\mathbf{U}_i^n, \mathbf{U}_j^n)$, we conclude that $\overline{\mathbf{U}}_{ij}^{n+1} = \bar{\boldsymbol{v}}(t, \boldsymbol{g}_j^n, \boldsymbol{n}_{ij}^n, \mathbf{U}_i^n, \mathbf{U}_j^n) \in B$; see Remark 1. In conclusion, $\mathbf{U}_i^{n+1} \in B$ since $\mathbf{U}_i^{n+1}$ is a convex combination of objects in B. The rest of the proof consists of verifying that (23) indeed implies

$$1 - \sum_{i \neq j \in \mathscr{I}(S_i^n)} 2d_{ij}^n \frac{\tau}{\mathfrak{m}_i^{n+1}} \geq 0.$$

□

Corollary 1 *Let $n \in \mathbb{N}$. Assume that τ is small enough so that the CFL condition (23) holds for all $i \in \{1{:}I\}$. Let $B \subset \mathcal{A}_f$ be a convex invariant set. Assume that $\{\boldsymbol{U}_i^n \mid i \in \{1{:}I\}\} \subset B$. Then* (i) *$\{\boldsymbol{U}_i^{n+1} \mid i \in \{1{:}I\}\} \subset B$;* (ii) *$\boldsymbol{u}_h^n \in B$ and $\boldsymbol{u}_h^{n+1} \in B$.*

Proof The statement (i) is a direct consequence of Theorem 1. The statement (ii) is a consequence of (11) from Lemma 6. □

Corollary 2 *Let $B \subset \mathcal{A}_f$ be a convex invariant set containing the initial data $\boldsymbol{u}_0$. Assume that $\{\boldsymbol{U}_i^0 \mid i \in \{1{:}I\}\} \subset B$. Let $N \in \mathbb{N}$. Assume that τ is small enough so that the CFL condition (23) holds for all $i \in \{1{:}I\}$ and all $n \in \{0{:}N\}$. Then $\{\boldsymbol{U}_i^n \mid i \in \{1{:}I\}\} \subset B$ and $\boldsymbol{u}_h^n \in B$ for all $n \in \{0{:}N+1\}$.*

Remark 4 (*Construction of $\boldsymbol{u}_h^0$*) Let $B \subset \mathcal{A}_f$ be a convex invariant set containing the initial data $\boldsymbol{u}_0$. If $\boldsymbol{P}_m(\mathscr{T}_h^0)$ is composed of piecewise Lagrange elements, then defining $\boldsymbol{u}_h^0$ to be the Lagrange interpolant of $\boldsymbol{u}_0$, we have $\{\mathbf{U}_i^0 \mid i \in \{1{:}I\}\} \subset B$. Similarly if $\boldsymbol{P}_m(\mathscr{T}_h^0)$ is composed of Bernstein finite elements of degree two and higher, then defining $\boldsymbol{u}_h^0$ to be the Bernstein interpolant of $\boldsymbol{u}_0$ we have $\{\mathbf{U}_i^0 \mid i \in \{1{:}I\}\} \subset B$; see [16, Eq. (2.72)]. In both cases the assumptions of Corollary 2 hold true.

5.2 Discrete Geometric Conservation Law

The ALE scheme (15) preserves constant states. This property is known in the literature as the Discrete Geometric Conservation Law (DGCL).

Corollary 3 (DGCL) *The scheme (15) preserves constant states. In particular if $\boldsymbol{U}_j^n = \boldsymbol{U}_i^n$ for all $j \in \mathscr{I}(S_i^n)$, then $\boldsymbol{U}_i^{n+1} = \boldsymbol{U}_i^n$.*

Proof The partition of unity property implies that $\sum_{j \in \mathscr{I}(S_i^n)} \boldsymbol{c}_{ij}^n = 0$. Moreover, the definition d_{ij}^n implies that $\sum_{j \in \mathscr{I}(S_i^n)} d_{ij}^n = 0$ (see (16)). Since Lemma 7 implies that

$$\mathbf{U}_i^{n+1} = \mathbf{U}_i^n + d_{ij}^n(\mathbf{U}_j^n - \mathbf{U}_i^n) + \frac{\tau}{\mathfrak{m}_i^{n+1}} \sum_{j \in \mathscr{I}(S_i^n)} ((\mathbf{U}_j^n - \mathbf{U}_i^n) \otimes \mathbf{W}_j^n + \boldsymbol{f}(\mathbf{U}_i^n) - \boldsymbol{f}(\mathbf{U}_j^n)) \cdot \boldsymbol{c}_{ij}^n,$$

it is now clear that if $\mathbf{U}_j^n = \mathbf{U}_i^n$ for all $j \in \mathscr{I}(S_i^n)$, then $\mathbf{U}_i^{n+1} = \mathbf{U}_i^n$. □

Remark 5 (*DGCL*) Note that although the DGCL seems to be given some importance in the literature, Corollary 3 has no particular significance. It is a direct consequence of the definition of the mass update (14) which is invoked to rewrite the scheme (15) from the conservative form to the equivalent nonconservative form (22). This equivalence is essential to prove the invariant domain property. In other words, *the DGCL is just a consequence of the equivalence of the discrete conservative and nonconservative formulations.*

5.3 Discrete Entropy Inequality

In this section we prove a discrete entropy inequality which is consistent with the inequality stated in Lemma 5.

Theorem 2 *Let* $(\eta, \boldsymbol{q})$ *be an entropy pair for* (1)*. Let* $n \in \mathbb{N}$ *and* $i \in \{1{:}I\}$*. Assume that all the assumptions of Theorem 1 hold. Then the following discrete entropy inequality holds:*

$$
\begin{aligned}
\frac{1}{\tau}\big(\mathfrak{m}_i^{n+1}\eta(\boldsymbol{U}_i^{n+1}) - \mathfrak{m}_i^{n}\eta(\boldsymbol{U}_i^{n})\big) \le{}& -\sum_{j\in\mathscr{I}(S_i^n)} d_{ij}^n\eta(\boldsymbol{U}_j^n) \\
&- \int_{\mathbb{R}^d} \nabla\cdot\Bigg(\sum_{j\in\mathscr{I}(S_i^n)} (\boldsymbol{q}(\boldsymbol{U}_j^n) - \eta(\boldsymbol{U}_j^n)\boldsymbol{W}_j^n)\psi_j^n(\boldsymbol{x})\Bigg)\psi_i^n(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}. \qquad (25)
\end{aligned}
$$

Proof Let $(\eta, \boldsymbol{q})$ be an entropy pair for the hyperbolic system (1). Let $i \in \{1{:}I\}$ and let $n \in \mathbb{N}$. Then using (24), the CFL condition and the convexity of η, we have

$$
\eta(\mathbf{U}_i^{n+1}) \le \eta(\mathbf{U}_i^n)\Bigg(1 - \sum_{i\neq j\in\mathscr{I}(S_i^n)} 2d_{ij}^n\frac{\tau}{m_i^{n+1}}\Bigg) + \frac{\tau}{m_i^{n+1}}\sum_{i\neq j\in\mathscr{I}(S_i^n)} 2d_{ij}^n\eta(\overline{\mathbf{U}}_{ij}^{n+1}).
$$

This can also be rewritten as follows:

$$
\frac{m_i^{n+1}}{\tau}\big(\eta(\mathbf{U}_i^{n+1}) - \eta(\mathbf{U}_i^n)\big) \le \sum_{i\neq j\in\mathscr{I}(S_i^n)} 2d_{ij}^n(\eta(\overline{\mathbf{U}}_{ij}^{n+1}) - \eta(\mathbf{U}_i^n)).
$$

Owing to (5) from Lemma 1, and recalling that the entropy flux of the Riemann problem (20) is $(\boldsymbol{q}(v) - \eta(v)\mathbf{W}_j^n)\cdot\boldsymbol{n}_{ij}^n$ we infer that

$$
\eta(\overline{\mathbf{U}}_{ij}^{n+1}) \le \tfrac{1}{2}(\eta(\mathbf{U}_i^n) + \eta(\mathbf{U}_j^n)) - t\big(\boldsymbol{q}(\mathbf{U}_j^n) - \eta(\mathbf{U}_j^n)\mathbf{W}_j^n - \boldsymbol{q}(\mathbf{U}_i^n) + \eta(\mathbf{U}_i^n)\mathbf{W}_j^n\big)\cdot\boldsymbol{n}_{ij}^n
$$

with $t = \|\boldsymbol{c}_{ij}^n\|_{\ell^2}/2d_{ij}^n$. Inserting this inequality in the first one, we have

$$\frac{m_i^{n+1}}{\tau}\big(\eta(\mathbf{U}_i^{n+1})-\eta(\mathbf{U}_i^n)\big)\leq\sum_{j\in\mathscr{I}(S_i^n)}d_{ij}^n(\eta(\mathbf{U}_j^n)-\eta(\mathbf{U}_i^n))$$
$$-\sum_{j\in\mathscr{I}(S_i^n)}\|\boldsymbol{c}_{ij}^n\|_{\ell^2}\big(\boldsymbol{q}(\mathbf{U}_j^n)-\boldsymbol{q}(\mathbf{U}_i^n)-(\eta(\mathbf{U}_j^n)-\eta(\mathbf{U}_i^n))\mathbf{W}_j^n\big)\cdot\boldsymbol{n}_{ij}^n.$$

By proceeding as in the proof of Lemma 7, we observe that

$$\frac{\mathfrak{m}_i^{n+1}-\mathfrak{m}_i^n}{\tau}=\sum_{j\in\mathscr{I}(S_i^n)}\mathbf{W}_j^n\cdot\boldsymbol{c}_{ij}^n.$$

Then using that $\|\boldsymbol{c}_{ij}^n\|_{\ell^2}\boldsymbol{n}_{ij}^n=\boldsymbol{c}_{ij}^n$, we obtain (25). This concludes the proof. □

6 Numerical Tests

In this section, we numerically illustrate the performance of the proposed method using SSP RK3. All the tests have been done with two different codes. One code is written in F95 and uses $\mathbb{P}_1$ Lagrange elements on triangles. The other code is based on deal.ii [2], is written in C++ and uses $\mathbb{Q}_1$ Lagrange elements on quadrangles. The mesh composed of triangles is obtained by dividing all the quadrangles into two triangles. The same numbers of degrees of freedom are used for both codes.

6.1 Analytical Scalar-Valued Solution

To test the convergence property of the SSP RK3 version of the method, as described in Algorithm 6, we solve the linear transport equation in the domain $D^0=(0,1)^2$:

$$\partial_t u+\nabla\cdot(\boldsymbol{\beta}u)=0,\quad u_0(\boldsymbol{x})=x_1+x_2, \tag{26}$$

where $\boldsymbol{\beta}=(\sin(\pi x_1)\cos(\pi x_2)\cos(2\pi t),-\cos(\pi x_1)\sin(\pi x_2)\cos(2\pi t))^{\mathsf{T}}$. In both codes the ALE velocity is chosen by setting $\mathbf{W}_i^n=\boldsymbol{\beta}(\boldsymbol{a}_i^n)$, i.e., $\boldsymbol{W}_h^n$ is the Lagrange interpolant of $\boldsymbol{\beta}$ on $\mathscr{T}_h^n$. Notice that there is no issue with boundary condition since $\boldsymbol{\beta}\cdot\boldsymbol{n}_{|\partial D^0}=0$.

We first test the accuracy in time of the algorithm by setting $d_{ij}^n=0$, i.e., the viscosity is removed. The computations are done with $CFL=1$. The error measured in the L^1-norm at time $t=0.5$ is reported in the left part of Table 1. The third-order convergence in time is confirmed. Note that there is no space error due to the particular choice for the ALE velocity and the initial data.

In the second test we put back the viscosity d_{ij}^n. Notice that the particular choice of the ALE velocity implies that $\lambda_{\max}(\boldsymbol{g}_j^n,\boldsymbol{n}_{ij}^n,\mathbf{U}_i^n,\mathbf{U}_j^n)=|(\boldsymbol{\beta}_i^n-\boldsymbol{\beta}_j^n)\cdot\boldsymbol{n}_{ij}^n|$; hence

Table 1 Rotation problem (26) with Lagrangian formulation, CFL = 1.0

#dofs	Without viscosity				With viscosity			
	$\mathbb{Q}_1$, L^1-norm		$\mathbb{P}_1$,L^1-norm		$\mathbb{Q}_1$, L^1-norm		$\mathbb{P}_1$, L^1-norm	
81	6.46E–04	–	1.76E–03	–	1.31E–02	–	1.13E–02	–
289	1.16E–04	2.48	2.46E–04	2.85	4.28E–03	1.61	3.63E–03	1.64
1089	1.41E–05	3.03	3.23E–05	2.93	1.23E–03	1.80	1.04E–03	1.80
4225	1.76E–06	3.01	4.20E–06	2.94	3.29E–04	1.90	2.78E–04	1.90
16641	2.26E–07	2.96	5.76E–07	2.87	8.50E–05	1.95	7.19E–05	1.95
66049	2.82E–08	3.00	9.57E–08	2.59	2.16E–05	1.97	1.83E–05	1.98

the viscosity is second-order in space instead of being first-order. This phenomenon makes the algorithm second-order in space (in addition to being conservative and maximum principle preserving). The error in the L^1-norm at time $t = 0.5$ is shown in the right part of Table 1.

6.2 *Nonlinear Scalar Conservation Equations*

We now test the method on nonlinear scalar conservation equations.

6.2.1 Definition of the ALE Velocity

In nonlinear conservation equations, solutions may develop shocks in finite time. In this case, using the purely Lagrangian velocity leads to a breakdown of the method in finite time which manifests itself by a time step that goes to zero as the current time approaches the time of formation of the shock. One way to avoid this breakdown is to use an ALE velocity that is a modified version of the Lagrangian velocity.

We now propose an algorithm to compute an ALE velocity based on [17]. The only purpose of this algorithm is to be able to run the nonlinear simulations past the time of formation of shocks. We refer the reader to the abundant ALE literature to design other ALE velocities that better suit the reader's goals.

We first deform the mesh by using the Lagrangian motion, i.e., we set

$$\boldsymbol{a}_{i,\mathrm{Lg}}^{n+1} = \boldsymbol{a}_i^n + \tau \nabla_u \boldsymbol{f}(\mathsf{U}_i^n);$$

we recall that $\mathsf{U}_i^n \in \mathbb{R}$ and $\nabla_u \boldsymbol{f}(\mathsf{U}_i^n) \in \mathbb{R}^d$ for scalar equations. Then, given $L \in \mathbb{N} \setminus \{0\}$, we define a smooth version of the Lagrangian mesh by smoothing the position of the geometric Lagrange nodes as follows:

$$\begin{cases} \boldsymbol{a}_i^{n+1,0} := \boldsymbol{a}_{i,\mathrm{Lg}}^{n+1},\ i \in \{1{:}I\} \\ \left(\boldsymbol{a}_i^{n+1,l} := \dfrac{1}{|\mathscr{I}(\mathscr{S}_i)| - 1} \displaystyle\sum_{i \neq j \in \mathscr{I}(\mathscr{S}_i)} \boldsymbol{a}_j^{n+1,l-1},\ i \in \{1{:}I\}\right),\ l \in \{1{:}L\} \\ \boldsymbol{a}_{i,\mathrm{Sm}}^{n+1} := \boldsymbol{a}_i^{n+1,L},\ i \in \{1{:}I\}. \end{cases} \tag{27}$$

Finally, the actual ALE motion is defined by

$$\boldsymbol{a}_i^{n+1} = \omega \boldsymbol{a}_{i,\mathrm{Lg}}^{n+1} + (1-\omega)\boldsymbol{a}_{i,\mathrm{Sm}}^{n+1}, \quad i \in \{1{:}I\},$$

where ω is a user-defined constant. In all our computations, we use $\omega = 0.9$ and $L = 2$. As mentioned in [17], a more advanced method consists of choosing ω pointwise by using the right Cauchy-Green strain tensor. We have not implemented this version of the method since the purpose of the tests in the next sections is just to show that the present method works as advertised for any reasonable ALE velocity.

6.2.2 Burgers Equation

We consider the inviscid Burgers equation in two space dimensions

$$\partial_t u + \nabla \cdot (\tfrac{1}{2}u^2 \boldsymbol{\beta}) = 0, \quad u_0(\boldsymbol{x}) = \mathbb{1}_{\{\|\boldsymbol{x}\|_{\ell^2}\}},$$

where $\boldsymbol{\beta} = (1,1)^{\mathsf{T}}$ and $\mathbb{1}_E$ denotes the characteristic function of the set $E \subset \mathbb{R}^d$. The solution to this problem at time $t > 0$ and at $\boldsymbol{x} = (x_1, x_2)$ is given as follows. Assume first that $x_2 \leq x_1$, then define $\alpha = x_1 - x_2$ and let $\alpha_0 = 1 - \frac{t}{2}$. There are three cases:

1. If $\alpha > 1$, then $u(x_1, x_2, t) = 0$.
2. If $\alpha \leq \alpha_0$, then

$$u(x_1, x_2, t) = \begin{cases} \dfrac{x_2}{t} & \text{if } 0 \leq x_2 < t \\ 1 & \text{if } t \leq x_2 < \frac{t}{2} + 1 - \alpha \\ 0 & \text{otherwise.} \end{cases}$$

3. If $\alpha_0 < \alpha \leq 1$, then

$$u(x_1, x_2, t) = \begin{cases} \dfrac{x_2}{t} & \text{if } 0 \leq x_2 < \sqrt{2t(1-\alpha)} \\ 0 & \text{otherwise.} \end{cases}$$

If $x_2 > x_1$, then $u(x_1, x_2, t) := u(x_2, x_1, t)$. The computation are done up to $T = 1$ in the initial computational domain $D^0 = (-0.25, 1.75)^2$. The boundary of D^n does not move in the time interval $(0, 1)$, i.e., $\partial D^0 = \partial D^n$ for any $n \geq 0$. The results of the

Table 2 Burgers equation, convergence tests, $CFL = 0.1$

# dofs	$\mathbb{Q}_1$				$\mathbb{P}_1$			
	L^2-error		L^1-error		L^2-error		L^1-error	
81	5.79E–01	–	6.00E–01	–	5.80E–01	–	6.17E–01	–
289	4.20E–01	0.46	3.88E–01	0.63	4.43E–01	0.39	4.68E–01	0.40
1089	2.96E–01	0.51	2.32E–01	0.74	3.12E–01	0.51	2.86E–01	0.71
4225	2.14E–01	0.47	1.32E–01	0.82	2.17E–01	0.53	1.55E–01	0.88
16641	1.56E–02	0.45	7.40E–02	0.83	1.23E–01	0.82	7.57E–02	1.04

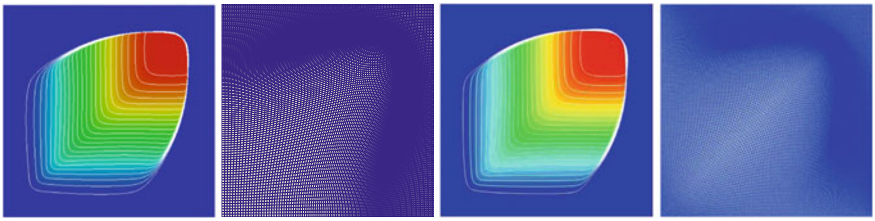


Fig. 1 Burgers equation. Left: $\mathbb{Q}_1$ FEM with 25 contours; Center left: Final $\mathbb{Q}_1$ mesh; Center right: $\mathbb{P}_1$ FEM with 25 contours; Right: Final $\mathbb{P}_1$ mesh

convergence tests are reported in Table 2. The solution is computed on a 128 × 128 mesh. The $\mathbb{Q}_1$ and $\mathbb{P}_1$ meshes at $T = 1$ are shown in Fig. 1.

6.3 Compressible Euler Equations

We finish the series of tests by solving the compressible Euler equations in $\mathbb{R}^2$

$$\begin{cases} \partial_t \rho + \nabla \cdot (\rho \boldsymbol{u}) = 0, \\ \partial_t (\rho \boldsymbol{u}) + \nabla \cdot (\rho \boldsymbol{u} \otimes \boldsymbol{u} + p\mathbb{I}) = 0, \\ \partial_t E + \nabla \cdot (\boldsymbol{u}(E + p)) = 0, \end{cases}$$

with the ideal gas equation of state, $p = (\gamma - 1)(E - \frac{1}{2}\rho\|\boldsymbol{u}\|_{\ell^2}^2)$ where $\gamma > 1$, and appropriate initial and boundary conditions. The motion of the mesh is done as described in (27) with $\boldsymbol{a}_{i,\mathrm{Lg}}^{n+1} = \boldsymbol{a}_i^n + \tau \boldsymbol{u}_h^n(\boldsymbol{a}_i^n)$ where $\boldsymbol{u}_h^n$ is the approximate fluid velocity.

We consider the so-called Noh problem, see, e.g., [3, Sect. 5]. The computational domain at the initial time is $D^0 = (-1, 1)^2$ and the initial data is

$$\rho_0(\boldsymbol{x}) = 1.0, \quad \boldsymbol{u}_0(\boldsymbol{x}) = -\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_{\ell^2}}, \quad p_0(\boldsymbol{x}) = 10^{-15}.$$

Table 3 Noh problem, convergence test, $T = 0.6$, $CFL = 0.2$

# dofs	$\mathbb{Q}_1$				$\mathbb{P}_1$			
	L^2-norm		L^1-norm		L^2-norm		L^1-norm	
961	2.60	–	1.44	–	2.89	–	1.71	–
3721	1.81	0.52	8.45E–01	0.77	2.21	0.39	1.09	0.64
14641	1.16	0.64	4.21E–01	1.01	1.42	0.64	5.15E–01	1.08
58081	7.66E–01	0.60	2.10E–01	0.99	9.39E–01	0.59	2.60E–01	0.99
231361	5.21E–01	0.56	1.06E–01	0.98	6.33E–01	0.57	1.28E–01	1.02

A Dirichlet boundary condition is enforced on all the dependent variables at the boundary of the domain. We use $\gamma = \frac{5}{3}$. The ALE velocity at the boundary of the computational domain is prescribed to be equal to the fluid velocity, i.e., the boundary moves inwards in the radial direction with speed 1. The final time is chosen to be $T = 0.6$ in order to avoid that the shockwave collides with the moving boundary of the computational domain which happens at $t = \frac{3}{4}$ since the shock moves radially outwards with speed $\frac{1}{3}$.

The solution to this problem is known. We show in Table 3 the L^1- and the L^2-norm of the error on the density for various meshes which are uniform at $t = 0$: 30×30, 60×60, etc.

7 Concluding Remarks

In this paper we have developed a framework for constructing ALE algorithms using continuous finite elements. The method is invariant domain preserving on any mesh in arbitrary space dimension. The methodology applies to any hyperbolic system which has such intrinsic property. If the system at hand has an entropy pair, then the method also satisfies a discrete entropy inequality. The time accuracy of the method can be increased by using SSP time discretization techniques. The equivalence between the conservative and non-conservative formulations implies the that DGCL condition holds (preservation of constant states). The new methods have been tested on a series of benchmark problems and the observed convergence orders and numerical performance are compatible with what is reported in the literature.

References

1. Ainsworth M (2014) Pyramid algorithms for Bernstein-Bézier finite elements of high, nonuniform order in any dimension. SIAM J Sci Comput 36(2):A543–A569
2. Bangerth W, Hartmann R, Kanschat G (2007) deal.II – a general purpose object oriented finite element library. ACM Trans Math Softw 33(4):Art. 24, 27 (27 pp)

3. Caramana E, Shashkov M, Whalen P (1998) Formulations of artificial viscosity for multidimensional shock wave computations. J Comput Phys 144(1):70–97
4. Chen G-Q (2005) Euler equations and related hyperbolic conservation laws. In: Evolutionary equations, Vol. II, Handbook of Differential Equations. Elsevier/North-Holland, Amsterdam, pp 1–104
5. Chueh KN, Conley CC, Smoller JA (1977) Positively invariant regions for systems of nonlinear diffusion equations. Indiana Univ Math J 26(2):373–392
6. Dafermos CM (2000) Hyperbolic conservation laws in continuum physics, vol 325. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]. Springer, Berlin
7. Farhat C, Geuzaine P, Grandmont C (2001) The discrete geometric conservation law and the nonlinear stability of ALE schemes for the solution of flow problems on moving grids. J Comput Phys 174(2):669–694
8. Ferracina L, Spijker MN (2005) An extension and analysis of the Shu-Osher representation of Runge-Kutta methods. Math Comp 74(249):201–219
9. Frid H (2001) Maps of convex sets and invariant regions for finite-difference systems of conservation laws. Arch Ration Mech Anal 160(3):245–269
10. Gottlieb S, Ketcheson DI, Shu C-W (2009) High order strong stability preserving time discretizations. J Sci Comput 38(3):251–289
11. Guermond J-L, Popov B (2016) Fast estimation from above of the maximum wave speed in the Riemann problem for the Euler equations. J Comput Phys 321:908–926
12. Guermond J-L, Popov B (2016) Invariant domains and first-order continuous finite element approximation for hyperbolic systems. SIAM J Numer Anal 54(4):2466–2489
13. Guermond J-L, Popov B, Saavedra L, Yang Y (2017) Invariant domains preserving ALE approximation of hyperbolic systems with continuous finite elements. SIAM J Sci Comput 39(2):A385–A414. arXiv:1603.01184 [math.NA].
14. Higueras I (2005) Representations of Runge-Kutta methods and strong stability preserving methods. SIAM J Numer Anal 43(3):924–948
15. Hoff D (1985) Invariant regions for systems of conservation laws. Trans Am Math Soc 289(2):591–610
16. Lai M-J, Schumaker LL (2007) Spline functions on triangulations, vol 110. Encyclopedia of mathematics and its applications. Cambridge University Press, Cambridge
17. Loubère R, Maire P-H, Shashkov M, Breil J, Galera S (2010) ReALE: a reconnection-based arbitrary-Lagrangian-Eulerian method. J Comput Phys 229(12):4724–4761
18. Pironneau O (1981/82) On the transport-diffusion algorithm and its applications to the Navier-Stokes equations. Numer Math 38(3):309–332
19. Pironneau O (1989) Finite element methods for fluids. Wiley, Chichester (translated from the French)
20. Toro EF (2009) Riemann solvers and numerical methods for fluid dynamics: a practical introduction, 3rd edn. Springer, Berlin

Dual-Primal Isogeometric Tearing and Interconnecting Methods

Christoph Hofer and Ulrich Langer

Abstract This paper generalizes the Dual-Primal Finite Element Tearing and Interconnecting (FETI-DP) method, that is well established as parallel solver for large-scale systems of finite element equations, to linear algebraic systems arising from the Isogemetric Analysis of elliptic diffusion problems with heterogeneous diffusion coefficients in two- and three-dimensional multipatch domains with C^0 smoothness across the patch interfaces. We consider different scalings, and derive the expected polylogarithmic bound for the condition number of the preconditioned systems. The numerical results confirm these theoretical bounds, and show incredibly robustness with respect to large jumps in the diffusion coefficient across the interfaces.

1 Introduction

Isogeometric Analysis (IgA) is a relatively new methodology for the numerical solution of partial differential equations (PDEs). IgA was introduced by Hughes, Cottrell and Bazilevs in [20], and has become a very active field of research, see also [2] for the first results on the numerical analysis of IgA, the monograph [10] for a comprehensive presentation of the IgA, and the recent survey article [3] on the mathematical analysis of variational isogeometric methods. In IgA, the basis functions, which are used for the representation of the geometry in computer aided design (CAD) models anyway, are also employed for approximating the solution of the PDE or the system of PDEs describing the physical phenomenon which we are going to simulate. The typical choice for such basis functions are B-Splines or Non-Uniform Rational Basis Spline (NURBS). One advantage of the IgA over the more traditional finite element method (FEM) is certainly the fact that there is no need for decomposing the

C. Hofer (✉) · U. Langer
Johann Radon Institute for Computational and Applied Mathematics (RICAM),
Austrian Academy of Sciences, Altenbergerstr. 69, A-4040 Linz, Austria
e-mail: christoph.hofer@ricam.oeaw.ac.at

U. Langer
e-mail: ulrich.langer@ricam.oeaw.ac.at

B. N. Chetverushkin et al. (eds.), *Contributions to Partial Differential Equations and Applications*, Computational Methods in Applied Sciences 47,
https://doi.org/10.1007/978-3-319-78325-3_15

computational domain into finite elements. Hence, one gets rid of this geometrical error source, at least, in the class of computational domains that are produced by a CAD system. Moreover, it is much easier to build up C^l, $l \geq 1$, conforming basis functions in IgA than in the finite element (FE) case. The major drawback is the fact that the basis functions are not nodal and have a larger support. However, it is still possible to associate basis functions to the interior, the boundary and the vertices of the subdomains (patches), which is crucial for the dual-primal isogeometric tearing and interconnecting (IETI-DP) method, which was introduced in [23]. The IETI-DP method is an extension of the dual-primal finite element tearing and interconnecting method (FETI-DP) to IgA. The FETI-DP was introduced by Farhat, Lesoinne, Le Tellec, Pierson, and Rixen in [15] as a faster alternative to the classical two-level FETI method that was earlier proposed for the parallel solution of large-scale finite element systems by Farhat and Roux in [16]. A comprehensive presentation of different FETI algorithms, including versions nowadays called balanced domain decomposition by constraints (BDDC), and their analysis as well as the corresponding references to the original papers can be found in the monographs [30] and [28]. The analysis of BDDC preconditioners for IgA matrices, which has been done in [6], also applies to the IETI-DP method due to the same spectrum (with the exception of at most two eigenvalues), see [27]. The so-called *deluxe scaling*, that was introduced in [12] for improving the robustness of BDDC preconditioners of finite element stiffness matrices, has recently been also generalized to IgA matrices [8]. We here also mention the recent developments of other IgA domain decomposition (DD) techniques. In particular, we refer to [5], [7] and [9] for isogeometric overlapping Schwarz methods, [18] for isogeometric mortaring discretizations, and [1] for comparison of different domain decomposition methods.

The aim of this paper is to extend the condition number estimates for BDDC preconditioners, presented in [6], to multipatch domains composed of non-overlapping patches which are images of the parameter domain by several different geometrical mappings. Moreover, we present the derivation of an improved bound for the condition number for the so-called *stiffness scaling* in the simplified case of C^0 smoothness across patch interfaces. For simplicity, we consider a diffusion problem with a heterogeneous diffusion coefficient α as model problem, the weak formulation of which reads as follows: find $u \in V_D = \{u \in H^1(\Omega) : u = 0 \text{ on } \Gamma_D\}$ such that

$$a(u, v) = \langle F, v\rangle \quad \forall v \in V_D, \tag{1}$$

where the bilinear form $a(\cdot, \cdot) : V_D \times V_D \to \mathbb{R}$ and the linear form $\langle F, \cdot\rangle : V_D \to \mathbb{R}$ are given by the expressions

$$a(u, v) = \int_\Omega \alpha(x)\, \nabla u(x) \cdot \nabla v(x)\, dx \quad \text{and} \quad \langle F, v\rangle = \int_\Omega f(x) v(x)\, dx + \int_{\Gamma_N} g_N(x) v(x)\, ds,$$

respectively. We assume that the computational (physical) domain Ω is a bounded Lipschitz domain in $\mathbb{R}^d$ with $d \in \{2, 3\}$ that can be represented by a multipatch IgA map from the parameter domain $\hat{\Omega} = (0, 1)^d$. The boundary $\Gamma = \partial\Omega$ of Ω consists of a non-empty Dirichlet part Γ_D and a Neumann part Γ_N. Furthermore, we assume that the Dirichlet boundary Γ_D is always a union of complete domain sides which are uniquely defined in IgA. Without loss of generality, we assume homogeneous Dirichlet conditions. The given data f, g_N and α are assumed to be sufficiently smooth, where the diffusion coefficient α is assumed to be positive and patchwise constant.

The rest of the paper is organized as follows. In Sect. 2, we recall the basic definitions and properties of B-Splines as well as the main principles of IgA. The IETI-DP and the corresponding BDDC methods are explained and analysed in Sect. 3. Section 4 is devoted to the implementation of the IETI-DP method. The numerical examples confirming the theory are presented in Sect. 5. Finally, in Sect. 6, we draw some conclusions and discuss further issues concerning generalizations to multipatch discontinuous Galerkin IgA schemes as constructed and analysed in [24, 25].

2 Some Preliminaries on Multipatch Isogeometric Analysis

B-Splines and NURBS play an important role in computer aided design and computer graphics. Here we will use these splines for building our trial and test spaces for the Galerkin approximations to (1), as proposed in [20]. This section provides the definition of B-Splines in one dimension as well as in higher dimensions via a tensor product structure. We will give an overview of isogeometric discretization and summarize the approximation properties of these B-Splines and NURBS spaces. We refer the reader to [2, 3, 10, 20] for a comprehensive presentation of the IgA basics.

2.1 B-Splines and Multipatch Geometries

Let the vector $\Xi = \{\xi_1 = 0, \xi_2, \dots, \xi_m = 1\}$ with non-decreasing real values ξ_i be a partition of the unit interval [0, 1]. The vector Ξ is called *knot vector*. Given a knot vector Ξ, $p \in \mathbb{N}$ and $M = m - p - 1$, we can define the B-Spline function via the Cox-de Boor formulas:

$$N_{i,0}(\xi) = \begin{cases} 1 & \text{if } \xi_i \leq \xi \leq \xi_{i+1}, \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

$$N_{i,p}(\xi) = \frac{\xi - \xi_i}{\xi_{i+p} - \xi_i} N_{i,p-1}(\xi) + \frac{\xi_{i+p+1} - \xi}{\xi_{i+p+1} - \xi_{i+1}} N_{i+1,p-1}(\xi), \tag{3}$$

where $i = 1, \dots, M$ and p is called *degree*. From this recursion, we can observe that $N_{i,p}$ is a piecewise polynomial of degree p. Furthermore, we only consider open knot vectors, i.e., the first and the last node is repeated p times.

Since we are considering d-dimensional problems, we need to extend the concept of B-Splines to the d-dimensional space, which is done via the tensor product. Let $(p_1, \dots, p_d) \in \mathbb{N}^d$, and let, for all $\iota = 1, \dots, d$, Ξ^ι be a knot vector. Furthermore, we denote the i^ι univariate B-Spline defined on the knot vector Ξ^ι by $N^\iota_{i^\iota,p}(\xi^\iota)$. Then the d-dimensional tensor product B-Spline (TB-Spline) is defined by

$$N_{(i^1,\dots,i^d),(p^1,\dots,p^d)}(\xi) = \prod_{\iota=1}^{d} N^\iota_{i^\iota,p^\iota}(\xi^\iota). \tag{4}$$

In order to avoid cumbersome notations, we will again denote the tensor product B-Spline by $N_{i,p}$ and interpret i and p as multi-indices. Additionally, we define the set of multi-indices $\mathscr{I}$ by $\mathscr{I} := \{(i^1, \dots, i^d) : i^\iota \in \{1, \dots, M_\iota\}\}$, where M_ι are the number of B-Spline basis function for dimension ι.

Now we are in a position to describe our computational domain, called *physical domain*, $\Omega = G((0,1)^d)$ by means of the *geometrical mapping* $G : \hat{\Omega} = (0,1)^d \to \mathbb{R}^d$ defined by $G(\xi) := \sum_{i \in \mathscr{I}} P_i N_{i,p}(\xi)$, where P_i are the given *control points*. Since the knot vector Ξ provides a partition of $\hat{\Omega}$, called *parameter domain* in the following, it introduces a mesh $\hat{\mathscr{Q}}$, and we will denote a mesh element by $\hat{Q}$, called *cell*. By means of the geometrical mapping, we receive a partition of the physical domain in cells Q_i as well, where $Q_i = G(\hat{Q}_i)$, $\hat{Q}_i \in \hat{\mathscr{Q}}$. If we collect all these cells, we get a mesh $\mathscr{Q}_h$ for the physical domain $\mathscr{Q}_h := \{Q = G(\hat{Q}) | \hat{Q} \in \hat{\mathscr{Q}}\}$. For the remainder of the paper, we only consider quasi-uniform meshes $\{\mathscr{Q}_h\}$, defined as follows:

Definition 1 A family of meshes $\{\mathscr{Q}_h\}$, is called *quasi uniform*, if there exists a constant $\theta \geq 1$ for all $\{\mathscr{Q}_h\}$, such that $\theta^{-1} \leq \operatorname{diam}(Q)/\operatorname{diam}(Q') \leq \theta$ for all $Q, Q' \in \mathscr{Q}_h$.

In many practical applications, it is not possible to describe the physical computational domain Ω just with one geometrical mapping G. Therefore, we represent the physical domain Ω by N non-overlapping domains $\Omega^{(k)}$, called *patches*. Each $\Omega^{(k)}$ is the image of an associated geometrical mapping $G^{(k)}$, defined on the parameter domain $\hat{\Omega}$, i.e., $\Omega^{(k)} = G^{(k)}(\hat{\Omega})$ for $k = 1, \dots, N$, and $\overline{\Omega} = \bigcup_{k=1}^{N} \overline{\Omega}^{(k)}$. Clearly, each patch has a mesh $\mathscr{Q}_h^{(k)}$ in the physical domain and a mesh $\hat{\mathscr{Q}}^{(k)}$ in the parameter domain, consisting of cells $Q^{(k)}$ and $\hat{Q}^{(k)}$. We denote the interface between the two patches $\Omega^{(k)}$ and $\Omega^{(l)}$ by $\Gamma^{(k,l)}$, and the collection of all interfaces by Γ, i.e., $\Gamma^{(k,l)} = \overline{\Omega}^{(k)} \cap \overline{\Omega}^{(l)}$ and $\Gamma := \bigcup_{l>k} \Gamma^{(k,l)}$. Furthermore, the boundary of the domain is denoted by $\partial\Omega$. This interface Γ is sometimes called *skeleton*.

2.2 Isogeometric Discretization

The key point in isogeometric analysis is the use of the same functions for representing the geometry as well as for constructing the solution and test spaces in the

Galerkin method. This motives the definition of the basis functions in the physical domain via the push-forward of the basis functions in the parameter domain, i.e., $\check{N}_{i,p} := N_{i,p} \circ G^{-1}$. Thus, we define our finite-dimensional IgA space V_h by $V_h = \operatorname{span}\{\check{N}_{i,p}\}_{i\in\mathscr{I}} \subset H^1(\Omega)$. The function u_h from the IgA space V_h can therefore be represented in the form $u_h(x) = \sum_{i\in\mathscr{I}} u_i\check{N}_{i,p}(x)$. Hence, each function $u_h(x)$ is associated with the vector $\mathbf{u} = (u_i)_{i\in\mathscr{I}}$. This map is known as *Ritz isomorphism*. One usually writes this relation as $u_h \leftrightarrow \mathbf{u}$, and we will use it in the following without further comments. Additionally, we define $\mathrm{S}_h = \operatorname{span}\{N_{i,p}\}_{i\in\mathscr{I}} \subset H^1(\hat{\Omega})$. If we consider a single patch $\Omega^{(k)}$ of a multipatch domain Ω, we will use the notation $V_h^{(k)}, \mathrm{S}_h^{(k)}, \check{N}_{i,p}^{(k)}, N_{i,p}^{(k)}$ and $G^{(k)}$ with the analogous definitions. The discrete function spaces for the whole multipatch domain is then given by $V_h := \{v|\ v|_{\Omega^{(k)}} \in V_h^{(k)}\} \cap H^1(\Omega)$. Based on the work in [6], we can find an important splitting of the space V_h. Since we are using open knot vectors, we can identify basis function on the interface Γ and in the interior of each patch. Let us define the spaces

$$V_{\Gamma,h} := \operatorname{span}\{\check{N}_{i,p}|\ i \in \mathscr{I}_B\} \subset H^1(\Omega) \quad \text{and} \quad V_{I,h}^{(k)} := V_h^{(k)} \cap H_0^1(\Omega^{(k)}), \tag{5}$$

where $\mathscr{I}_B$ denotes all indices of basis functions having support on Γ. This leads to the decomposition $V_h = \prod_{k=1}^N V_{I,h}^{(k)} \oplus \mathscr{H}(V_{\Gamma,h})$, where $\mathscr{H}:\ V_{\Gamma,h} \to V_h$ is the *discrete NURBS harmonic extension* defined by

$$\begin{cases} \text{Find } \mathscr{H}v_B \in V_h : \\ \quad a(\mathscr{H}v_B, v^{(k)}) = 0 \qquad \forall v^{(k)} \in V_{I,h}^{(k)},\ 1 \le k \le N, \\ \quad \mathscr{H}v_B|_{\partial\Omega^{(k)}} = v_B|_{\partial\Omega^{(k)}} \quad 1 \le k \le N. \end{cases} \tag{6}$$

See [6, 29] for a more sophisticated discussion.

2.2.1 Continuous Galerkin IgA Schemes

We look for the Galerkin approximate u_h from the finite dimensional subspace $V_{D,h}$ of V_D, where $V_{D,h}$ is the set of all functions from V_h which vanish on the Dirichlet boundary Γ_D. The Galerkin IgA scheme reads as follows: find $u_h \in V_{D,h}$ such that

$$a(u_h, v_h) = \langle F, v_h\rangle \quad \forall v_h \in V_{D,h}. \tag{7}$$

There exists a unique IgA solution $u_h \in V_{D,h}$ of (7) that converges to the solution $u \in V_D$ of (1) for h tends to 0. Due to Cea's lemma, the usual discretization error estimates in the H^1-norm follow from the corresponding approximation error estimates, see [2] or [3]. A basis for this space is given by the B-Spline functions $\{\check{N}_{i,p}\}_{i\in\mathscr{I}_0}$, where $\mathscr{I}_0$ contains all indices of $\mathscr{I}$ which do not have a support on Γ_D. Hence, the Galerkin IgA scheme (7) is equivalent to the linear system of algebraic equations $\mathbf{K}\mathbf{u} = \mathbf{f}$, where $\mathbf{K} = (\mathbf{K}_{i,j})_{i,j\in\mathscr{I}_0}$ and $\mathbf{f} = (\mathbf{f}_i)_{i\in\mathscr{I}_0}$ denote the stiffness matrix and the load vector, respectively, with $\mathbf{K}_{i,j} = a(\check{N}_{j,p}, \check{N}_{i,p})$ and $\mathbf{f}_i = \langle F, \check{N}_{i,p}\rangle$, and $\mathbf{u}$ is the vector representation of u_h given by the IgA isomorphism. In order to keep the notation simple, we will reuse the symbol $\mathscr{I}$ for the set $\mathscr{I}_0$ in the following.

2.2.2 Schur Complement System

Introducing the bilinear form $s : V_{\Gamma,h} \times V_{\Gamma,h} \to \mathbb{R}$; $s(w_B, v_B) = a(\mathscr{H} w_B, \mathscr{H} v_B)$, one can show that the interface component u_B of the solution to the IgA scheme (7) satisfies the variational identity

$$s(u_B, v_B) = \langle g, v_B \rangle \quad \forall v_B \in V_{\Gamma,h}, \tag{8}$$

where $g \in V^*_{\Gamma,h}$ is a suitable functional. By choosing the B-Spline basis for $V_{\Gamma,h}$, the variational identity (8) is equivalent to the linear system $\mathbf{S}\mathbf{u}_B = \mathbf{g}$. The matrix $\mathbf{S}$ is the Schur complement matrix of $\mathbf{K}$ with respect to the interface dofs. Suppose, we reorder the entries of the stiffness matrix $\mathbf{K}$ and the load vector $\mathbf{f}$, such that the dofs corresponding to the interface come first, i.e.,

$$\mathbf{K} = \begin{bmatrix} \mathbf{K}_{BB} & \mathbf{K}_{BI} \\ \mathbf{K}_{IB} & \mathbf{K}_{II} \end{bmatrix} \text{ and } \quad \mathbf{f} = \begin{bmatrix} \mathbf{f}_B \\ \mathbf{f}_I \end{bmatrix},$$

then it is easy to see that $\mathbf{S}$ and $\mathbf{g}$ are given by $\mathbf{S} = \mathbf{K}_{BB} - \mathbf{K}_{BI}(\mathbf{K}_{II})^{-1}\mathbf{K}_{IB}$ and $\mathbf{g} = \mathbf{f}_B - \mathbf{K}_{BI}(\mathbf{K}_{II})^{-1}\mathbf{f}_I$, respectively. Once $\mathbf{u}_B$ is calculated, we obtain $\mathbf{u}_I$ as the solution of the system $\mathbf{K}_{II}\mathbf{u}_I = \mathbf{f}_I - \mathbf{K}_{BI}\mathbf{u}_B$. Instead of the Schur complement matrix $\mathbf{S}$ we will mostly use its operator representation: $S : V_{\Gamma,h} \to V^*_{\Gamma,h}$; $\langle Sv, w \rangle = (\mathbf{S}\mathbf{v}, \mathbf{w})$.

3 IETI-DP Methods and Their Analysis

The IETI-DP method, that was introduced in [23], is nothing but the adaption of the FETI-DP method (see, e.g., [28, 30]) to isogeometric analysis. According to [27] based on algebraic arguments, the BDDC preconditioner and the FETI-DP method possess the same spectrum up to zeros and ones. Hence a condition number bound for BDDC implies a bound for FETI-DP and vice versa. Since the proof is based on algebraic arguments, it also holds for the IETI-DP method.

3.1 Local Spaces and Jump Operator

Analogously to the splitting introduced in Sect. 2.2, we define the local interface space $W^{(k)} := \text{span}\{\check{N}_{i,p} \mid \text{supp}\{\check{N}_{i,p}\} \cap (\partial\Omega^{(k)} \cap \Gamma) \neq \emptyset,\ i \in \mathscr{I}\}$, which is the restriction of $V_{\Gamma,h}$ to $\Omega^{(k)}$. In the following, in order to avoid cumbersome notation, we define the patch boundary to be just the interface part, i.e. $\partial\Omega^{(k)} := \partial\Omega^{(k)} \cap \Gamma$. Hence, we have $V_h^{(k)} = W^{(k)} \oplus V_{I,h}^{(k)}$, where $V_{I,h}^{(k)}$ is defined as in (5). Furthermore, we define the space of functions, which are locally in $W^{(k)}$, by $W := \prod_{k=1}^{N} W^{(k)}$. Similar to the definition of the global discrete NURBS harmonic extension $\mathscr{H}$ in (6), we define the local patch version $\mathscr{H}^{(k)} : W^{(k)} \to V_h^{(k)}$.

In order to obtain continuous functions, we introduce additional constraints which will enforce the continuity. Let $\mathscr{B}(k,l)$ be the set of all coupled indices between $\Omega^{(k)}$ and $\Omega^{(l)}$, then we enforce the following constraints:

$$\mathbf{w}_i^{(k)} - \mathbf{w}_j^{(l)} = 0 \quad \forall (i,j) \in \mathscr{B}(k,l),\ k > l. \tag{9}$$

The operator $B : W \to U^* := \mathbb{R}^\Lambda$, which realizes constraints (9) in the form $Bw = 0$, is called *jump operator*. The space of all functions in W which belong to the kernel of B is denoted by $\widehat{W}$, and can be identified with $V_{\Gamma,h}$.

3.1.1 Saddle Point Formulation

Due to the multipatch structure of our physical domain, we can decompose the bilinear form and the right-hand side functional as follows:

$$a(u_h, v_h) = \sum_{k=1}^{N} a^{(k)}(u_h^{(k)}, v_h^{(k)}) \quad \text{and} \quad \langle F, v_h\rangle = \sum_{k=1}^{N} \langle F^{(k)}, v_h^{(k)}\rangle,$$

where $u_h, v_h \in V_h$ and $u_h^{(k)}, v_h^{(k)}$ denote its restriction to $\Omega^{(k)}$. The Galerkin IgA scheme (7) can be reformulated as a constrained minimization problem

$$u_{B,h} = \underset{w\in W,\ Bw=0}{\arg\min}\ \frac{1}{2}\langle Sw, w\rangle - \langle g, w\rangle, \tag{10}$$

on the skeleton, where $\langle Sv, w\rangle := \sum_{k=1}^{N} \langle S^{(k)} v^{(k)}, w^{(k)}\rangle$ and $\langle g, w\rangle := \sum_{k=1}^{N} \langle g^{(k)}, w^{(k)}\rangle$ for $v, w \in W$.

In the following, we will only work with the Schur complement system and hence, to simplify the notation, we will use u instead of $u_{B,h}$, when we consider functions in $V_{\Gamma,h}$. If there has to be made a distinction between u_h, $u_{B,h}$ and $u_{I,h}$, we will write the subscripts again.

3.1.2 Intermediate Space and Primal Constraints

In order to guarantee the positive definiteness of S, we are looking for an intermediate space $\widetilde{W}$ in the sense $\widehat{W} \subset \widetilde{W} \subset W$ such that S restricted to $\widetilde{W}$ is SPD. Let $\Psi \subset V_{\Gamma,h}^*$ be a set of linearly independent *primal variables*.

Then we define the spaces $\widetilde{W} := \{w \in W : \forall \psi \in \Psi : \psi(w^{(k)}) = \psi(w^{(l)}), \forall k > l\}$ and $W_\Delta := \prod_{k=1}^{N} W_\Delta^{(k)}$, where $W_\Delta^{(k)} := \{w^{(k)} \in W^{(k)} : \forall \psi \in \Psi : \psi(w^{(k)}) = 0\}$. Moreover, we introduce the space $W_\Pi \subset \widehat{W}$, such that $\widetilde{W} = W_\Pi \oplus W_\Delta$. We call W_Π *primal space* and W_Δ *dual space*. In the literature, there are the following typical choices for ψ:

- Vertex evaluation: $\psi^{\mathscr{V}}(v) = v(\mathscr{V})$,
- Edge averages: $\psi^{\mathscr{E}}(v) = \frac{1}{|\mathscr{E}|}\int_{\mathscr{E}} v\, ds$,
- Face averages: $\psi^{\mathscr{F}}(v) = \frac{1}{|\mathscr{F}|}\int_{\mathscr{F}} v\, ds$.

The typical choices for Ψ are usually called Algorithm A–C:

- Algorithm A: $\Psi = \{\psi^{\mathscr{V}}\}$,
- Algorithm B: $\Psi = \{\psi^{\mathscr{V}}\} \cup \{\psi^{\mathscr{E}}\} \cup \{\psi^{\mathscr{F}}\}$,
- Algorithm C: $\Psi = \{\psi^{\mathscr{V}}\} \cup \{\psi^{\mathscr{E}}\}$.

Moreover, one finds references to two further choices for Ψ, commonly referred to Algorithm D and E, which are aiming for a reduced set of primal variables, see, e.g., Algorithm 6.28 and 6.29 in [30]. These algorithms address the issue of the rapidly increasing number of primal variables.

Remark 1 For domains $\Omega \subset \mathbb{R}^2$, Algorithm A will provide a quasi optimal method for the Poisson problem. By choosing additional primal variables, the coarse problem will grow. Hence, it becomes computationally more demanding. However, it brings benefits in the condition number. For three-dimensional domains, one can show that just choosing vertex evaluation does not lead to a quasi optimal method. In such cases, additional primal variables have to be chosen, see, e.g. Remark 6.39 in [30].

3.2 IETI–DP

Since $\widetilde{W} \subset W$, there is a natural embedding $\widetilde{I} : \widetilde{W} \to W$. Let the jump operator restricted to $\widetilde{W}$ be $\widetilde{B} := B\widetilde{I} : \widetilde{W} \to U^*$. Then we can formulate the saddle point problem in $\widetilde{W}$ as follows: find $(u, \lambda) \in \widetilde{W} \times U$ such that

$$\begin{bmatrix} \widetilde{S} & \widetilde{B}^T \\ \widetilde{B} & 0 \end{bmatrix} \begin{bmatrix} u \\ \lambda \end{bmatrix} = \begin{bmatrix} \widetilde{g} \\ 0 \end{bmatrix}, \tag{11}$$

where $\widetilde{g} := \widetilde{I}^T g$, and $\widetilde{B}^T = \widetilde{I}^T B^T$. Here, $\widetilde{I}^T : W^* \to \widetilde{W}^*$ denotes the adjoint of $\widetilde{I}$.

By construction, $\widetilde{S}$ is SPD on $\widetilde{W}$. Hence, we can define the Schur complement F and the corresponding right-hand side of equation (11) as $F := \widetilde{B}\widetilde{S}^{-1}\widetilde{B}^T$ and $d := \widetilde{B}\widetilde{S}^{-1}\widetilde{g}$. Hence, the saddle point system (11) is equivalent to solving:

$$\text{find}\, \lambda \in U : \quad F\lambda = d. \tag{12}$$

By means of Brezzi's theorem we obtain that (12) has a unique solution up to adding elements from $\ker(\widetilde{B}^T)$ to λ, and $u = \widetilde{S}^{-1}(\widetilde{g} - \widetilde{B}^T\lambda) \in \widehat{W}$ is the unique solution of (10).

We note that F is SPSD on U. If we restrict F to $\widetilde{U} := \mathscr{R}(\widetilde{B})$, then $F|_{\widetilde{U}}$ is SPD, cf., e.g., [28]. Hence, we can solve (12) by means of the PCG.

3.2.1 Preconditioning

In order to receive robustness with respect to the diffusion coefficient α, we use the so called *scaled Dirichlet preconditioner*. The scaling is incorporated in the application of the jump operator. We define the *scaled* jump operator B_D such that the operator enforces the constraints: $\delta_j^{\dagger(l)}\mathbf{w}_i^{(k)} - \delta_i^{\dagger(k)}\mathbf{w}_j^{(l)} = 0$ for $(i,j) \in \mathscr{B}(k,l)$, $k > l$, where $\delta_i^{\dagger(k)} := \rho_i^{(k)} / \sum_l \rho_{j_l}^{(l)}$ and,j_l is the corresponding coefficient index on the neighboring patch $\Omega^{(l)}$. The scaled Dirichlet preconditioner has the following form:

$$M_{sD}^{-1} = B_D S B_D^T. \tag{13}$$

Typical choices for $\rho_i^{(k)}$ are

- Multiplicity Scaling: $\rho_i^{(k)} = 1$,
- Coefficient Scaling: If $\alpha(x)|_{\Omega^{(k)}} = \alpha^{(k)}$, choose $\rho_i^{(k)} = \alpha^{(k)}$,
- Stiffness Scaling: $\rho_i^{(k)} = \mathbf{K}_{i,i}^{(k)}$.

Theorem 1 *Let $H^{(k)}$ be the diameter and $h^{(k)}$ the local mesh size of $\Omega^{(k)}$ and let M_{sD}^{-1} be the scaled Dirichlet preconditioner. Then, under suitable assumptions imposed on the mesh, we have*

$$\kappa(M_{sD}^{-1}F_{\widetilde{U}}) \leq C \max_k \left(1 + \log\left(H^{(k)}/h^{(k)}\right)\right)^2,$$

where the positive constant C is independent of h and H.

In the case of IgA, a more general proof in the sense, that not only C^0 smoothness across patch interfaces is allowed but also C^l, $l \geq 0$ smoothness, can be found in [6]. However, the proof is restricted to the case of a domain decomposition, which is obtained by subdividing a single patch, i.e. performing a decomposition of the parameter domain. Hence, always the same geometrical mapping G is used. Furthermore, due to the C^l, $l \geq 0$, smoothness across interfaces, only a condition number bound of $O\left((1 + \log H/h)H/h\right)$ could be proven for stiffness scaling. In the proceeding section, we will extend the proof given in [6] to multipatch domains, which consists of different geometrical mappings $G^{(k)}$ for each patch. Additionally, for $l = 0$, we again obtain quasi-optimal condition number bounds also for stiffness scaling.

3.3 Analysis of BDDC Preconditioner

In this section we rephrase the results and notations established in [6], and extend them to multipatch domains, consisting of a different geometrical mapping $G^{(k)}$ for each patch. However, we only allow C^0 smoothness across the patch interfaces and restrict the analysis to the 2D case of Algorithm A.

3.3.1 General Results

Let $\check{z}$ be a function from V_h. Its restriction to a patch $\Omega^{(k)}$ belongs to $V_h^{(k)}$, and can be written as $\check{z}^{(k)} := \check{z}|_{\Omega^{(k)}} = \sum_{i \in \mathscr{I}^{(k)}} c_i^{(k)} \check{N}_{i,p}^{(k)}$, where in $\mathscr{I}^{(k)}$ all indices, where the basis functions have a support on the Dirichlet boundary in the physical space $\Omega^{(k)}$, are excluded. The corresponding spline function in the parameter space is denoted by $z^{(k)} \in \mathrm{S}_h^{(k)}$. It is important to note that the geometrical map G and its inverse G^{-1} are independent of h, since it is fixed on a coarse discretization. When the domain becomes refined, G stays the same. Clearly, the same applies for the gradients and it can be assumed, that $G^{(k)} \in W^{1,\infty}((0,1)^d)$ for all $k \in \{1, \ldots, N\}$. The analysis of the method in the physical space is based on the fact that

$$\|\check{z}^{(k)}\|_{L^2(\Omega^{(k)})} \approx \|z^{(k)}\|_{L^2((0,1)^d)} \text{ and } |\check{z}^{(k)}|_{H^1(\Omega^{(k)})} \approx |z^{(k)}|_{H^1((0,1)^d)},$$

where the hidden constants only depend on G, see Lemma 3.2 in [2]. We note that, as in [2], equivalence is only stated for the full H^m-norm, with $m \geq 0$. However, it follows from the proof that the L^2 term is not needed for $m > 0$, and is only incorporated for giving a unified presentation for all $m \geq 0$.

We will now define a local discrete seminorm based on the control points c_i, where we refer to [6] for a motivation and a more sophisticated discussion. Moreover, we denote the coefficient corresponding to basis function $N_{(i^1,\ldots,i^\iota - j,\ldots,i^d),p}$ by $c_{i,i^\iota - j}$, cf. (4).

Definition 2 Let $\check{z} \in V_h^{(k)}$, and z its counterpart in the parameter domain. Then $|z|^2_\nabla := \sum_{\iota=1}^d |z|^2_{\xi^\iota}$ defines a discrete seminorm, where $|z|^2_{\xi^\iota} := \sum_{i \in \mathscr{I}_\iota^{(k)}} |c^{(k)}_{i,i^\iota} - c^{(k)}_{i,i^\iota - 1}|^2$. The set $\mathscr{I}_\iota$ contains the admissible indices such that each summand is well defined.

Proposition 1 *Let $\check{z} \in V_h^{(k)}$ and z its counterpart in the parameter domain. Then $|z|^2_\nabla \approx |z|^2_{H^1((0,1)^2)} \approx \left|\check{z}\right|^2_{H^1(\Omega^{(k)})}$ holds, where the hidden constants are independent of h and H.*

Proof The proof follows from the equivalence of the norms in the parameter and physical domain, see Lemma 3.5 [2], and from Proposition 5.2 in [6].

The second step is to provide properties in the local index spaces. Since we consider only the two dimensional problem, we can interpret the control points

$(c_i)_{i\in\mathscr{I}}$ as entries of a matrix $\mathbf{C} = (c_i)_{i=1}^{M_1,M_2}$. Before doing that we will provide abstract results for an arbitrary matrix. We define the seminorm

$$|||\mathbf{C}|||_\nabla^2 := \sum_{\iota=1}^{2} \sum_{\substack{i=1\\ i^\iota=2}}^{M_\iota} |c_{i,i^\iota} - c_{i,i^\iota-1}|^2$$

for a real valued $M_1 \times M_2$ matrix $\mathbf{C} = (c_i)_{i=1}^{M_1,M_2} \in \mathbb{R}^{M_1\times M_2}$. The entries of the matrix $\mathbf{C}$ can be interpreted as values on a uniform grid $\mathscr{T}$. This motivates the definition of an operator $(\cdot)_I : C([0,1]^2) \to \mathbb{R}^{M_1\times M_2}$, which evaluates a continuous function on the grid points $(x_i) = (x_{i^1 i^2})$, and an operator $\chi : \mathbb{R}^{M_1\times M_2} \to \mathscr{Q}_1(\mathscr{T}) \subset H^1((0,1)^2)$, that provides a piecewise bilinear interpolation of the given grid values, where $\mathscr{Q}_1(\mathscr{T})$ is the space of piecewise bilinear functions on $\mathscr{T}$.

Furthermore, given values on an edge e on $[0,1]^2$, we need to define its linear interpolation and a discrete harmonic extension to the interior. In order to do so, let us denote all indices of grid points x_i associated to e by $\mathscr{I}(e)$. Additionally, let $\mathscr{P}_1(\mathscr{T}|_e)$ be the space of piecewise linear spline functions on $\mathscr{T}|_e$. We define the interpolation of values on $\mathscr{I}(e)$ by the restriction of the operator χ to e, denoted by $\chi_e : \mathbb{R}^{M_\iota} \to H^1(e)$ with an analogous definition. In a similar way, we define the interpolation operator for the whole boundary ∂, denoted by $\chi_\partial : \mathbb{R}^{|\mathscr{I}(\partial)|} \to H^1(\partial[0,1]^2)$, where $\mathscr{I}(\partial) := \{i : x_i \in \partial[0,1]^2\}$.

This leads to a definition of a seminorm for grid points on an edge e via the interpolation to functions from $\mathscr{P}_1(\mathscr{T}|_e)$:

Definition 3 Let e be an edge of $[0,1]^2$ along dimension ι, and let $\mathbf{v}$ be a vector in $\mathbb{R}^{M_\iota}$. Then we define the seminorm $|||\mathbf{v}|||_e := |\chi_e(\mathbf{v})|_{H^{1/2}(e)}$ for all $\mathbf{v} \in \mathbb{R}^{M_\iota}$.

Remark 2 For the interpolation operator χ_e defined on an edge e, it is easy to see that $\chi(\mathbf{C})|_e = \chi_e(\mathbf{C}|_e)$ and $|||\mathbf{C}|_e|||_e = |\chi_e(\mathbf{C}|_e)|_{H^{1/2}(e)} = |\chi(\mathbf{C})|_e)|_{H^{1/2}(e)}$ hold.

Finally, we are able to define the discrete harmonic extension in $\mathbb{R}^{M_1\times M_2}$.

Definition 4 Let $\mathscr{H}_{\mathscr{Q}_1}$ be the standard discrete harmonic extension into the piecewise bilinear space $\mathscr{Q}_1$. This defines the lifting operator $\mathbf{H} : \mathbb{R}^{|\mathscr{I}(\partial)|} \to \mathbb{R}^{M_1\times M_2}$ by

$$\mathbf{b} \mapsto \mathbf{H}(\mathbf{b}) := (\mathscr{H}_{\mathscr{Q}_1}(\chi_\partial(\mathbf{b})))_I.$$

Theorem 2 *Let e be a particular side on the boundary of* $[0,1]^2$ *and the constant $\beta \in \mathbb{R}^+$ such that $\beta^{-1}M_2 \leq M_1 \leq \beta M_2$. Then the following statements hold:*

- *For all $\mathbf{b} \in \mathbb{R}^{2M_1+2M_2-4}$ that vanish on the four components corresponding to the four corners, the estimate $|||\mathbf{H}(\mathbf{b})|||_\nabla^2 \leq c(1+\log^2 M_1)\sum_{e\in\partial[0,1]^2} |||\mathbf{b}|_e|||_e^2$, holds, where the constant c depends only on β.*
- *The estimate $|||\mathbf{C}|||_\nabla \geq c\, |||\mathbf{C}|_e|||_e$ is valid for all $\mathbf{C} \in \mathbb{R}^{M_1\times M_2}$, where the constant c depends only on β.*

Proof See [6]. □

3.3.2 Condition Number Estimate

The goal of this section to establish a condition number bound for $P = M_{\mathrm{BDDC}}^{-1}\hat{S}$. Following [6], we assume that the mesh is quasi-uniform on each subdomain and the diffusion coefficient is globally constant. We focus now on a single patch $\Omega^{(k)}$, $k \in \{1, \dots, N\}$. For notational simplicity, we assume that the considered patch $\Omega^{(k)}$ does not touch the boundary $\partial\Omega$.

We define the four edges of the parameter domain $[0, 1]^2$ by E_r, and their images by $\check{E}_r^{(k)} = G^{(k)}(E_r)$, $r = 1, 2, 3, 4$. Moreover, we denote by $\mathscr{I}(\check{E}_r^{(k)})$ the coefficient indices corresponding to the basis functions on $\check{E}_r^{(k)}$ and by $\mathscr{I}(\Gamma^{(k)})$ the indices corresponding to the whole boundary.

Let $\check{z}^{(k)} \in V_h^{(k)}$, then $\check{z}^{(k)}$ is determined by its coefficients $c_i^z = (c_{i^1,i^2}^z)_{i^1,i^2=1}^{M_1^{(k)},M_2^{(k)}}$, which can be interpreted as a $M_1^{(k)} \times M_2^{(k)}$ matrix $\mathbf{C}^z$. In a similar way, we can identify functions on the trace space $W^{(k)}$.

Finally, let $W_\Delta^{(k)} \subset W^{(k)}$ be the space of spline functions which vanish on the primal variables, i.e., in the corner points. The following theorem provides an abstract estimate of the condition number using the coefficient scaling:

Theorem 3 *Let the counting function $\delta^{\dagger^{(k)}}$ be chosen accordingly to the coefficient scaling strategy. Assume that there exist two positive constants c_*, c^* and a boundary seminorm $|\cdot|_{W^{(k)}}$ on $W^{(k)}$, $k = 1, \dots, N$, such that*

$$|\check{w}^{(k)}|_{W^{(k)}}^2 \leq c^* s^{(k)}(\check{w}^{(k)}, \check{w}^{(k)}) \quad \forall \check{w}^{(k)} \in W^{(k)}, \tag{14}$$

$$|\check{w}^{(k)}|_{W^{(k)}}^2 \geq c_* s^{(k)}(\check{w}^{(k)}, \check{w}^{(k)}) \quad \forall \check{w}^{(k)} \in W_\Delta^{(k)}, \tag{15}$$

$$|\check{w}^{(k)}|_{W^{(k)}}^2 = \sum_{r=1}^{4} |\check{w}^{(k)}|_{\check{E}_r^{(k)}}|_{W_r^{(k)}} \quad \forall \check{w}^{(k)} \in W^{(k)}, \tag{16}$$

where $|\cdot_{W_r^{(k)}}$ is a seminorm associated to the edge spaces $W^{(k)}|_{\check{E}_r^{(k)}}$, with $r = 1, 2, 3, 4$. Then the condition number of the preconditioned BDDC operator P satisfies the bound

$$\kappa(M_{\mathrm{BDDC}}^{-1}\hat{S}) \leq C(1 + c_*^{-1}c^*),$$

where the constant C is independent of h and H.

Proof See [6] or [4]. □

Using this abstract framework, we obtain the following condition number estimate for the BDDC preconditioner.

Theorem 4 *There exists a boundary seminorm such that the constants c_* and c^* of Theorem 3 are bounded by*

$$c^* \leq C_1 \ \textit{and} \ c_*^{-1} \leq C_2 \max_{1 \leq k \leq N} \left(1 + \log^2\left(H^{(k)}/h^{(k)}\right)\right),$$

where the constants C_1 and C_2 are independent of H and h. Therefore, the condition number of the isogeometric preconditioned BDDC operator is bounded by

$$\kappa(M_{\mathrm{BDDC}}^{-1}\hat{S}) \leq C \max_{1\leq k\leq N} \left(1+\log^2\left(H^{(k)}/h^{(k)}\right)\right),$$

where the constant C is independent of H and h.

Proof The proof essentially follows the lines of the proof given in [6] with a minor modification due to the different geometrical mappings $G^{(k)}$. We note that we only consider C^0 continuity across the patch interfaces, which makes the proof less technical.

The first step is to appropriately define the seminorm $|\check{w}^{(k)}|^2_{W^{(k)}}$ in $W^{(k)}$:

$$\begin{aligned}
|\check{w}^{(k)}|^2_{W^{(k)}} &:= \sum_{r=1}^{4} |\check{w}^{(k)}|_{\check{E}_r^{(k)}}|^2_{W_r^{(k)}},\\
|\check{w}^{(k)}|_{\check{E}_1^{(k)}}|^2_{W_1^{(k)}} &:= |||\check{w}^{(k)}|_{\check{E}_1^{(k)}}|||^2_{\check{E}_1^{(k)}} + \sum_{i^2=1}^{M_2{}^{(k)}-1} |c^w_{(1,i^2+1)} - c^w_{(1,i^2)}|^2,\\
|\check{w}^{(k)}|_{\check{E}_2^{(k)}}|^2_{W_2^{(k)}} &:= |||\check{w}^{(k)}|_{\check{E}_2^{(k)}}|||^2_{\check{E}_2^{(k)}} + \sum_{i^2=1}^{M_2{}^{(k)}-1} |c^w_{(M_1,i^2+1)} - c^w_{(M_1,i^2)}|^2,\\
|\check{w}^{(k)}|_{\check{E}_3^{(k)}}|^2_{W_3^{(k)}} &:= |||\check{w}^{(k)}|_{\check{E}_3^{(k)}}|||^2_{\check{E}_3^{(k)}} + \sum_{i^1=1}^{M_1{}^{(k)}-1} |c^w_{(i^1+1,1)} - c^w_{(i^1,1)}|^2,\\
|\check{w}^{(k)}|_{\check{E}_4^{(k)}}|^2_{W_4^{(k)}} &:= |||\check{w}^{(k)}|_{\check{E}_4^{(k)}}|||^2_{\check{E}_4^{(k)}} + \sum_{i^1=1}^{M_1{}^{(k)}-1} |c^w_{(i^1+1,M_2)} - c^w_{(i^1,M_2)}|^2,
\end{aligned} \tag{17}$$

where $M_\iota^{(k)}$ denotes the number of basis functions on patch k in direction ι. Furthermore, we define $|||\check{w}^{(k)}|_{\check{E}_r^{(k)}}|||_{\check{E}_r^{(k)}} := |||\mathbf{v}|||_{E_r}$, where $\mathbf{v}$ are the values $(c_i^w)_{i\in\mathscr{I}(\check{E}_r^{(k)})}$ written as a vector.

Let $\check{z}^{(k)} \in V_h^{(k)}$ be the NURBS harmonic extension of $w^{(k)} = \{c_i^w\} \in W^{(k)}$, and $z^{(k)}$ its representation in the parameter domain. Additionally, let e be any edge of the parameter domain of $\Omega^{(k)}$. Due to the fact that $c_i^w = c_i^z$ for $i \in \mathscr{I}(\Gamma^{(k)})$, and denoting $\mathbf{C}^{(k)} = (c_i^z)_{i\in\mathscr{I}^{(k)}}$, we obtain $|||\check{w}^{(k)}|_e|||^2_e = |||\mathbf{C}^{(k)}|_e|||^2_e \leq c|||\mathbf{C}^{(k)}|||^2_\nabla$ by means of Theorem 2. From the definition of $|||\mathbf{C}^{(k)}|||^2_\nabla$ and the definition of $|\check{w}^{(k)}|_{\check{E}_r^{(k)}}|^2_{W_r^{(k)}}$, we get $|\check{w}^{(k)}|_e|^2_{W_r^{(k)}} \leq c|||\mathbf{C}^{(k)}|||^2_\nabla$. Furthermore, we have

$$|\check{w}^{(k)}|_e|^2_{W_r^{(k)}} \leq c|||\mathbf{C}^{(k)}|||^2_\nabla \leq c|z^{(k)}|^2_\nabla \leq c|z^{(k)}|^2_{H^1((0,1)^d)} \leq c|\check{z}^{(k)}|^2_{H^1(\Omega^{(k)})}.$$

Since $|\check{z}^{(k)}|^2_{H^1(\Omega^{(k)})} = |\mathscr{H}^{(k)}(\check{w}^{(k)})|^2_{H^1(\Omega^{(k)})} = s^{(k)}(\check{w}^{(k)}, \check{w}^{(k)})$, we arrive at the estimate $|\check{w}^{(k)}|_e|^2_{W_r^{(k)}} \leq c\, s^{(k)}(\check{w}^{(k)}, \check{w}^{(k)})$. These estimates hold for all edges of $\Omega^{(k)}$. Hence, it follows that $|\check{w}^{(k)}|^2_{W^{(k)}} \leq c^* s^{(k)}(\check{w}^{(k)}, \check{w}^{(k)})$ for $\check{w} \in W^{(k)}$, where the constant does not depend on h and H. This proves the upper bound, i.e., estimate (14).

Let be $\check{w}^{(k)} \in W_{\Delta}^{(k)}$, $w^{(k)}$ its representation in the parameter domain, and $(c_i^w)_{i\in\mathscr{I}(\Gamma^{(k)})}$ its coefficient representation. We apply the lifting operator $\mathbf{H}^{(k)}$ to $(c_i^w)_{i\in\mathscr{I}(\Gamma^{(k)})}$, and obtain a matrix $\mathbf{H}^{(k)}(w^{(k)})$ with entries $(c_i^{H^{(k)}})_{i\in\mathscr{I}^{(k)}}$. These entries define a spline function $z^{(k)} := \sum_{i\in\mathscr{I}^{(k)}} c_i^{H^{(k)}} N_{i,p}^{(k)}$. It follows the estimate

$$|||\mathbf{H}^{(k)}(w^{(k)})|||^2_{\nabla} = |z^{(k)}|^2_{\nabla} \geq c|z^{(k)}|^2_{H^1((0,1)^d)} \geq c|\check{z}^{(k)}|^2_{H^1(\Omega^{(k)})} \geq c|\mathscr{H}(\check{w}^{(k)})|^2_{H^1(\Omega^{(k)})},$$

where the last inequality holds due to the fact that the discrete NURBS harmonic extension minimizes the energy among functions with given boundary data $\check{w}$. The constant c does not depend on h or H.

Recalling the definition of $|\check{w}^{(k)}|^2_{W^{(k)}}$ and using Theorem 2, we arrive at the estimates

$$|||\mathbf{H}^{(k)}(w^{(k)})|||^2_{\nabla} \leq c(1+\log^2 M^{(k)}) \sum_{e\in\partial[0,1]^2} |||w|_e^{(k)}|||^2_e \leq c(1+\log^2 M^{(k)})|\check{w}^{(k)}|^2_{W^{(k)}}.$$

Due to the mesh regularity, we have $M^{(k)} \approx H^{(k)}/h^{(k)}$, and, hence, we obtain

$$s^{(k)}(\check{w}^{(k)}, \check{w}^{(k)}) = |\mathscr{H}(\check{z}^{(k)})|^2_{H^1(\Omega^{(k)})} \leq c(1+\log^2(H^{(k)}/h^{(k)}))|\check{w}^{(k)}|^2_{W^{(k)}},$$

which provides the desired estimate for c_*^{-1}. □

The next theorem provides the corresponding estimates for the stiffness scaling.

Theorem 5 *Let the counting functions be chosen according to the stiffness scaling strategy. Assume that there exist two positive constants c_*, c^* and a boundary seminorm $|\cdot|_{W^{(k)}}$ on $W^{(k)}$, $k = 1, \ldots, N$, such that the three conditions of Theorem 3 hold. Moreover, we assume that it exits a constant c^*_{STIFF} such that*

$$|\check{w}^{(k)}|_{W^{(k)}} \leq c^*_{\mathrm{STIFF}} s(\delta\check{w}^{(k)}, \delta\check{w}^{(k)}) \quad \forall \check{w}^{(k)} \in W_{\Delta}^{(k)}, \tag{18}$$

where the coefficients of $\delta\check{w}^{(k)}$ are given by $c_i^{(k)}\delta_i^{(k)}$. Then the condition number of the preconditioned BDDC operator $M_{\mathrm{BDDC}}^{-1}\hat{S}$ satisfies the bound

$$\kappa(M_{\mathrm{BDDC}}^{-1}\hat{S}) \leq c(1 + c_*^{-1}c^* + c_*^{-1}c^*_{\mathrm{STIFF}})$$

for some constant c which is independent of h and H.

Proof See [6]. □

According to [6], we apply a modified version of the stiffness scaling where we use one representative of the values $\delta_i^{(k)}$. This is reasonable, since these values are very similar on one patch $\delta_i^{(k)} \approx \delta_j^{(k)}$, which arises from the tensor product structure of B-Splines and the constant material value on a patch.

Lemma 1 *The bound (18) holds with $c^*_{\mathrm{STIFF}} \leq C_1$, where C_1 is the constant appearing in Theorem 4. Hence, the condition number of the BDDC preconditioned system in the case of stiffness scaling is bounded by*

$$\kappa(M_{\mathrm{BDDC}}^{-1}\hat{S}) \leq C \max_{1 \leq k \leq N} \left(1 + \log^2\left(H^{(k)}/h^{(k)}\right)\right),$$

where the constant C is independent of H and h.

Proof The inequality $|\check{w}^{(k)}|^2_{W^{(k)}} \leq c^*_{\mathrm{STIFF}} s(\delta\check{w}^{(k)}, \delta\check{w}^{(k)})$ is equivalent to

$$|\delta^\dagger\check{w}^{(k)}|^2_{W^{(k)}} \leq c^*_{\mathrm{STIFF}} s(\check{w}^{(k)}, \check{w}^{(k)}) \text{ for } \check{w} \in W_\Delta^{(k)}.$$

We have already proven that $|\check{w}^{(k)}|^2_{W^{(k)}} \leq c^* s(\check{w}^{(k)}, \check{w}^{(k)})$ for $\check{w} \in W_\Delta^{(k)} \subset W^{(k)}$. Hence, it is enough to show the inequality $|\delta^\dagger\check{w}^{(k)}|^2_{W^{(k)}} \leq c_{h,H}|\check{w}^{(k)}|^2_{W^{(k)}}$ for $\check{w} \in W_\Delta^{(k)}$, where the constant $c_{h,H}$ may depend on $h^{(k)}$ and $H^{(k)}$. Recalling the definition of $|\delta^\dagger\check{w}^{(k)}|^2_{W^{(k)}}$ as the sum of $|\delta^\dagger\check{w}^{(k)}|^2_{\check{E}_r^{(k)}}|_{W_r^{(k)}}$, $r = 1, 2, 3, 4$, see (17), we have only to estimate the parts, e.g., $|\check{w}^{(k)}|_{\check{E}_1^{(k)}}|_{W_1^{(k)}}$. The other three terms follow analogously. From the fact that $\delta_i^{\dagger(k)} \leq 1$ and $\delta_i^{\dagger(k)} = \delta_{i+1}^{\dagger(k)}$ for all $k \in \{1, \ldots, N\}$ and $i \in \mathscr{I}(\check{E}_1^{(k)})$, c.f. Sect. 6.2. in [6], it follows that

$$|||\delta^\dagger\check{w}^{(k)}|_{\check{E}_1^{(k)}}|||_{\check{E}_1^{(k)}} = \delta^\dagger|||\check{w}^{(k)}|_{\check{E}_1^{(k)}}|||_{\check{E}_1^{(k)}} \leq |||\check{w}^{(k)}|_{\check{E}_1^{(k)}}|||_{\check{E}_1^{(k)}}$$

and

$$\begin{aligned}\sum_{i^2=1}^{{M_2}^{(k)}-1} |\delta^\dagger_{(1,i^2+1)}c^w_{(1,i^2+1)} - \delta^\dagger_{(1,i^2)}c^w_{(1,i^2)}|^2 &= \sum_{i^2=1}^{{M_2}^{(k)}-1} \delta_i^\dagger|c^w_{(1,i^2+1)} - c^w_{(1,i^2)}|^2 \\ &\leq \sum_{i^2=1}^{M_2^{(k)}-1} |c^w_{(1,i^2+1)} - c^w_{(1,i^2)}|^2.\end{aligned}$$

These estimates provide the inequalities $|\delta^\dagger\check{w}^{(k)}|_{\check{E}_1^{(k)}}|_{W_1^{(k)}} \leq |\check{w}^{(k)}|_{\check{E}_1^{(k)}}|_{W_1^{(k)}}$, and, finally, $|\delta^\dagger\check{w}^{(k)}|_{W^{(k)}} \leq |\check{w}^{(k)}|_{W^{(k)}}$. This concludes the proof with $c^*_{\mathrm{STIFF}} \leq c^*$, and the desired condition number bound. □

4 Implementation

Since **F** is symmetric and at least positive semi definite and positive definite on $\widetilde{U}$, we can solve the linear system $\mathbf{F}\boldsymbol{\lambda} = \mathbf{d}$ of the algebraic equations by means of the PCG algorithm, where we use M_{sD}^{-1} as preconditioner. Since it is very expensive to build up the matrices $\mathbf{F}$ and M_{sD}^{-1}, we use a matrix-free version of the PCG, which only needs the functional procedure of the application of a matrix to a vector. The challenging part is the application of $\widetilde{S}^{-1}$, which is part of F. The idea is to split the space $\widetilde{W}$ into $\widetilde{W}_\Pi \oplus \prod \widetilde{W}_\Delta^{(k)}$, such that $\widetilde{W}_\Delta^{(k)} \perp_S \widetilde{W}_\Pi$ for all k.

4.1 Choosing a Basis for $\widetilde{W}_\Pi$

The first step is to provide an appropriate space $\widetilde{W}_\Pi$ and a local basis $\{\widetilde{\phi}_j\}_j^{n_\Pi}$, where n_Π is the number of primal vari ables. We request from the basis that it has to be nodal with respect to the primal variables, i.e., $\psi_i(\widetilde{\phi}_j) = \delta_{i,j}$, for $i,j \in \{1,\ldots,n_\Pi\}$. There are many choices for the subspace $\widetilde{W}_\Pi$. Following the approach presented in [28], we will choose that one which is orthogonal to $\widetilde{W}_\Delta$ with respect to S. Hence, we can define $\widetilde{W}_\Pi := \widetilde{W}_\Delta^{\perp_S}$. This choice, which will simplify the application of $\widetilde{S}^{-1}$ significantly, is known as *energy minimizing primal subspace* in the literature, cf., [11, 28].

In order to construct a nodal basis, we introduce the constraint matrix $C^{(k)} : W^{(k)} \to \mathbb{R}^{n_\Pi^{(k)}}$ for each patch $\Omega^{(k)}$ which realizes the primal variables, i.e., $(C^{(k)}v)_j = \psi_{i(k,j)}(v)$ for $v \in W$ and $j \in \{1,\ldots,n_\Pi^{(k)}\}$, where $n_\Pi^{(k)}$ is the number of primal variables associated with $\Omega^{(k)}$ and $i(k,j)$ the global index of the j-th primal variable on $\Omega^{(k)}$.

$$\begin{bmatrix} K_{BB}^{(k)} & K_{BI}^{(k)} & C^{(k)^T} \\ K_{IB}^{(k)} & K_{II}^{(k)} & 0 \\ C^{(k)} & 0 & 0 \end{bmatrix} \begin{bmatrix} \widetilde{\phi}_j^{(k)} \\ \cdot \\ \widetilde{\mu}_j^{(k)} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \mathbf{e}_j^{(k)} \end{bmatrix}, \tag{19}$$

where $\mathbf{e}_j^{(k)} \in \mathbb{R}^{n_\Pi^{(k)}}$ is the j-th unit vector. Here we use an equivalent formulation with the system matrix instead of $S^{(k)}$. For each patch k, the LU factorization of this matrix is computed and stored.

4.2 Application of $\widetilde{S}^{-1}$

Assume that $f := \{\mathbf{f}_\Pi, \{f_\Delta^{(k)}\}\} \in \widetilde{W}^*$ is already given. We are now looking for $w := \{\mathbf{w}_\Delta, \{w_\Delta^{(k)}\}\} \in \widetilde{W}$ with $w = \widetilde{S}^{-1}f$. Let $S_{\Pi\Pi}, S_{\Delta\Pi}, S_{\Pi\Delta}$ and $S_{\Delta\Delta}$ be the restrictions of $\widetilde{S}$ to the corresponding subspaces with $S_{\Delta\Pi} = S_{\Pi\Delta}^T$. We note that $S_{\Delta\Delta}$ can be seen as a block diagonal operator, i.e., $S_{\Delta\Delta} = \text{diag}(S_{\Delta\Delta}^{(k)})$. Due to our special choice $\widetilde{W}_\Pi := \widetilde{W}_\Delta^{\perp_S}$, we have $S_{\Delta\Pi} = S_{\Pi\Delta} = 0$. Based on this splitting, we have the block forms

$$\widetilde{S} = \begin{bmatrix} S_{\Pi\Pi} & 0 \\ 0 & S_{\Delta\Delta} \end{bmatrix} \text{ and } \widetilde{S}^{-1} = \begin{bmatrix} S_{\Pi\Pi}^{-1} & 0 \\ 0 & S_{\Delta\Delta}^{-1} \end{bmatrix}.$$

Therefore, the application of $\widetilde{S}^{-1}$ reduces to an application of one global coarse problem involving $\mathbf{S}_{\Pi\Pi}^{-1}$ and N local problems involving $S_{\Delta\Delta}^{(k)\ -1}$.

Application of $S_{\Delta\Delta}^{(k)^{-1}}$: The application of $S_{\Delta\Delta}^{(k)^{-1}}$ corresponds to solving a local Neumann problem in the space $\widetilde{W}_{\Delta}$, i.e., $S^{(k)}w^{(k)} = f_{\Delta}^{(k)}$ with the constraint $C^{(k)}w^{(k)} = 0$. This problem can be rewritten as a saddle point problem in the form

$$\begin{bmatrix} K_{BB}^{(k)} & K_{BI}^{(k)} & C^{(k)^T} \\ K_{IB}^{(k)} & K_{II}^{(k)} & 0 \\ C^{(k)} & 0 & 0 \end{bmatrix} \begin{bmatrix} w^{(k)} \\ \cdot \\ \cdot \end{bmatrix} = \begin{bmatrix} f_{\Delta}^{(k)} \\ 0 \\ 0 \end{bmatrix}.$$

From (19), the LU factorization of the matrix is already available.

Application of $\mathbf{S}_{\Pi\Pi}^{(k)^{-1}}$: The matrix $\mathbf{S}_{\Pi\Pi}$ can be assembled from the patch local matrices $\mathbf{S}_{\Pi\Pi}^{(k)}$. Let $\{\widetilde{\phi}_j^{(k)}\}_{j=1}^{n_{\Pi}^{(k)}}$ be the basis of $\widetilde{W}_{\Pi}^{(k)}$. The construction of $\{\widetilde{\phi}_j^{(k)}\}_{j=1}^{n_{\Pi}^{(k)}}$ in (19) provides

$$\begin{aligned} \left(\mathbf{S}_{\Pi\Pi}^{(k)}\right)_{i,j} &= \left\langle S^{(k)}\widetilde{\phi}_i^{(k)}, \widetilde{\phi}_j^{(k)}\right\rangle = -\left\langle C^{(k)^T}\widetilde{\mu}_i^{(k)}, \widetilde{\phi}_j^{(k)}\right\rangle = -\left\langle \widetilde{\mu}_i^{(k)}, C^{(k)}\widetilde{\phi}_j^{(k)}\right\rangle \\ &= -\left\langle \widetilde{\mu}_i^{(k)}, \mathbf{e}_j^{(k)}\right\rangle = -\left(\widetilde{\mu}_i^{(k)}\right)_j, \end{aligned}$$

where $i, j \in \{1, \ldots, n_{\Pi}^{(k)}\}$. Therefore, we can reuse the Lagrange multipliers $\widetilde{\mu}_i^{(k)}$ obtained in (19), and can assemble $\mathbf{S}_{\Pi\Pi}^{(k)}$ from them. Once $\mathbf{S}_{\Pi\Pi}$ is assembled, the LU factorization can be calculated and stored.

4.3 Summary of the Algorithm for $F = \widetilde{B}S^{-1}\widetilde{B}^T$ and M_{sD}^{-1}

The application of F and M_{sD}^{-1} is summarized in Algorithm 7. Let us mention that the implementation of the embedding operator $\widetilde{I}$ and assembling operator $\widetilde{I}^T$ is explained in detail in [28] and is omitted here.

5 Numerical Examples

We test the implemented IETI-DP algorithm for solving large scale systems arising from the IgA discretization of (1) on the so-called YETI-footprint domains illustrated in Fig. 1. The computational domain consists of 21 subdomains in both 2D and 3D. In both cases, one side of a patch boundary has inhomogeneous Dirichlet conditions, whereas all other sides have homogeneous Neumann conditions. Each subdomain has a diameter of H and an associated mesh size of h. The degree of the B-Splines is chosen as $p = 4$. In order to solve the linear system (12), a PCG algorithm with

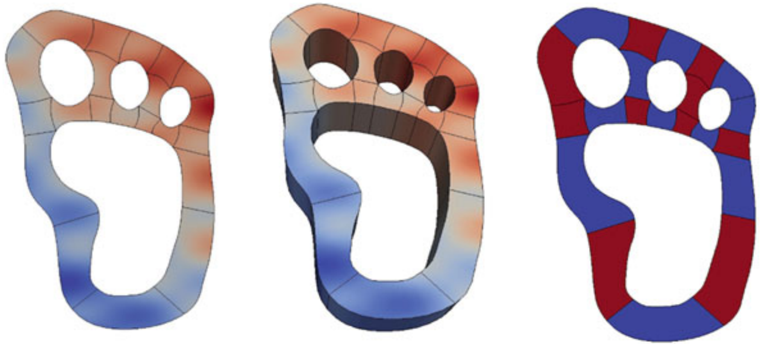


Fig. 1 The domain Ω in 2D (left) and 3D (middle), and the coefficient pattern (right)

Algorithm 7 Algorithm for calculating $\nu = F\lambda$ and $\nu = M_{sD}^{-1}\lambda$ for given $\lambda \in U$

```
procedure $F(\lambda)$
    Application of $B^T$ : $\{f^{(k)}\}_{k=1}^N = B^T\lambda$
    Application of $\widetilde{I}^T$ : $\{\mathbf{f}_\Pi, \{f_\Delta^{(k)}\}_{k=1}^N\} = \widetilde{I}^T\left(\{f^{(k)}\}_{k=1}^N\right)$
    Application of $S^{-1}$ :
    Begin
        $\mathbf{w}_\Delta = \mathbf{S}_{\Pi\Pi}^{-1}\mathbf{f}_\Pi$
        $w_\Delta^{(k)} = S_{\Delta\Delta}^{(k)\ -1} f_\Delta^{(k)} \quad \forall k = 1, \ldots, N$
    End
    Application of $\widetilde{I}$ : $\{w^{(k)}\}_{k=1}^N = \widetilde{I}\left(\{\mathbf{w}_\Delta, \{w_\Delta^{(k)}\}_{k=1}^N\}\right)$
    Application of $B$ : $\nu = B\left(\{w^{(k)}\}_{k=1}^N\right)$
end procedure
procedure $M_{sD}^{-1}(\lambda)$
    Application of $B_D^T$ : $\{w^{(k)}\}_{k=1}^N = B_D^T\lambda$
    Application of $S$ :
    Begin
        Solve $K_{II}^{(k)}x^{(k)} = -K_{IB}^{(k)}w^{(k)} \quad \forall k = 1, \ldots, N$
        $v^{(k)} = K_{BB}^{(k)}w^{(k)} + K_{BI}^{(k)}x^{(k)}. \quad \forall k = 1, \ldots, N$
    End
    Application of $B_D$ : $\nu = B_D\left(\{v^{(k)}\}_{k=1}^N\right)$
end procedure
```

the scaled Dirichlet preconditioner (13) is performed. We use zero initial guess, and a reduction of the initial residual by a factor of 10^{-6} as stopping criterion. The numerical examples illustrate the dependence of the condition number of the IETI-DP preconditioned system on jumps in the diffusion coefficient α, patch size H, mesh size h and the degree p. We use the C++ library G+Smo for describing the geometry and performing the numerical tests, see also [17, 21].

5.1 Homogeneous Diffusion Coefficients

We present numerical tests for problem (1) with a globally constant diffusion coefficient $\alpha = 1$. The 2D results are summarized in Table 1, whereas the 3D results are presented in Table 2. The results confirm that the preconditioned systems with coefficient scaling as well as stiffness scaling provide a quasi optimal condition number bound according to Theorems 4 and 5.

5.2 Jumping Diffusion Coefficients

We investigate numerical examples with patchwise constant diffusion coefficient α, the jumping pattern of which is shown in Fig. 1 (right). The values of α are 10^{-3} (blue) and 10^{3} (red). The 2D results are summarized in Table 3, whereas the 3D results are shown in Table 4. We again observe a quasi optimal condition number bound which is clearly independent of the diffusion coefficient and its jumps across the subdomain interfaces.

Table 1 2D example with homogeneous diffusion coefficient and $p = 4$. Choice of primal variables: vertex evaluation (Alg. A), vertex evaluation and edge averages (Alg. C)

	Alg. A	Unprec. F		Coeff. scal		Stiff. scal.		Alg. C	Unprec. F		Coeff. scal		Stiff. scal.	
#dofs	H/h	κ	It.	κ	It.	κ	It.	H/h	κ	It.	κ	It.	κ	It.
2364	9	45	50	9	21	9	20	9	28	44	1.8	11	1.8	11
4728	13	73	46	11	22	11	22	13	22	39	2	12	2	12
11856	21	133	57	15	24	14	24	21	18	39	2.4	14	2.4	14
352712	37	265	68	18	25	18	25	37	17	38	2.8	15	2.8	15

Table 2 3D example with homogeneous diffusion coefficient and $p = 4$. Choice of primal variables: vertex evaluation (Alg. A), vertex evaluation, edge averages and face averages (Alg. C)

	Alg. A	Unprec. F		Coeff. scal		Stiff. scal.		Alg. B	Unprec. F		Coeff. scal		Stiff. scal.	
#dofs	H/h	κ	It.	κ	It.	κ	It.	H/h	κ	It.	κ	It.	κ	It.
7548	5	3254	393	63	33	63	33	5	2751	341	1.6	10	1.6	10
14368	7	3059	356	86	37	86	37	7	2860	397	1.7	11	1.7	11
38100	10	2170	317	196	45	196	46	10	1697	333	2.0	12	2.3	13
142732	16	7218	397	467	64	468	65	16	1261	333	2.3	13	3.1	16

Table 3 2D example with jumping diffusion coefficient and $p = 4$. Choice of primal variables: vertex evaluation (Alg. A), vertex evaluation and edge averages ($p = 4$)

	Alg. A	Unprec. F		Coeff. scal		Stiff. scal.		Alg. C	Unprec. F		Coeff. scal		Stiff. scal	
#dofs	H/h	κ	It.	κ	It.	κ	It.	H/h	κ	It.	κ	It.	κ	It.
2364	9	1.4e07	317	5.6	13	5.3	13	9	1.5e07	261	1.8	7	1.7	7
4728	13	1.5e07	297	7.0	13	6.4	13	13	1.1e07	267	2.2	8	2	7
11856	21	2.4e07	397	8.7	15	7.8	13	21	9.8e06	291	2.6	8	2.3	8
35712	37	4.0e07	434	10.6	16	9.3	14	37	9.0e06	310	3.0	10	2.7	10

Table 4 3D example with jumping diffusion coefficient and $p = 4$. Choice of primal variables: vertex evaluation (Alg. A), vertex evaluation and edge averages and face averages (Alg. B)

	Alg. A	Unprec. F		Coeff. scal		Stiff. scal		Alg.B	Unprec. F		Coeff. scal		Stiff. scal	
#dofs	H/h	κ	It.	κ	It.	κ	It.	H/h	κ	It.	κ	It.	κ	It.
7548	5	$\geq$1.e16	$\geq$500	47	20	47	18	5	$\geq$1.e16	$\geq$500	1.7	7	1.6	7
14368	7	$\geq$1.e16	$\geq$500	69	20	65	19	7	$\geq$1.e16	$\geq$500	1.8	7	1.7	7
38100	10	$\geq$1.e16	$\geq$500	165	32	152	29	10	$\geq$1.e16	$\geq$500	2.1	8	2.3	8
142732	16	$\geq$1.e16	$\geq$500	405	38	368	34	16	$\geq$1.e16	$\geq$500	4.4	9	3.2	11

5.3 Dependence on the Degree p

We want to examine the dependence of the condition number on the B-Spline degree p, although the theory presented in Sect. 3.3 does not cover the dependence of IETI-DP preconditioned systems on p. We note that, in our implementation, the degree elevation yields an increase in the multiplicity of the knots within each step, resulting in C^1 smoothness on each patch. The computational domain Ω is chosen as the 2D YETI-footprint presented in Fig. 1. The diffusion coefficient α is chosen to be equal to 1. The results are summarized in Table 5, where we observe a possibly logarithmic dependence of the condition number on the polynomial degree p in case of the coefficient scaling as well as of the special version of the stiffness scaling mentioned above. The numerical experiments depict a linear dependence in case of the regular stiffness scaling, see Fig. 2.

5.4 Performance

The algorithm was tested on a Desktop PC with an Intel(R) Xeon(R) CPU E5-1650 v2 @ 3.50 GHz and 16 GB main memory. As already mentioned at the beginning of this section, we used the open source library G+Smo for the materialization of the code with the Sparse-LU factorization of the open source library "Eigen", see [14], for the local solvers. The timings presented in Table 6 are obtained from a sequential implementation of the code. We choose the same setting as presented in Table 3 with Algorithm C. However, we do one more refinement steps and obtain 121824 total degrees of freedom, 1692 Lagrange multipliers, and on each patch approximate 4900

Table 5 2D example with homogeneous diffusion coefficient and fixed initial mesh. Choice of primal variables: vertex evaluation and edge averages (Alg. C)

Alg. C		Unprec. F		Coeff. scal		Stiff. scal.		Stiff. scal. modif.	
#dofs	Degree	κ	It.	κ	It.	κ	It.	κ	It.
800	2	3.24	16	1.65	8	1.64	8	1.63	8
2364	3	8.08	28	1.71	10	1.69	10	1.7	10
4728	4	24.2	51	1.83	11	1.88	12	1.83	11
7892	5	82.8	86	2.03	13	2.16	12	2.01	12
11856	6	296	140	2.23	13	2.42	14	2.18	13
16620	7	1082	230	2.41	14	2.66	14	2.34	14
22184	8	4021	371	2.57	15	2.88	15	2.49	15
28548	9	15034	594	2.72	15	3.09	16	2.63	15
35712	10	56773	968	2.87	16	3.28	16	2.75	15

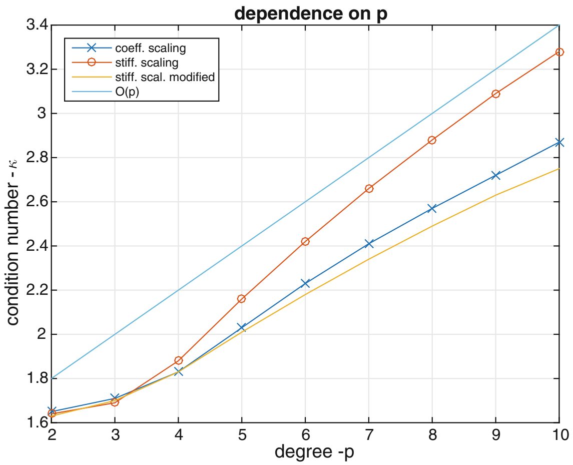


Fig. 2 2D example with homogeneous diffusion coefficient and fixed initial mesh. Condition number κ as a function of polynomial degree p. Choice of primal variables: vertex evaluation and edge averages (Alg. C)

local degrees of freedom. We select a run with coefficient scaling and obtain a condition number of $\kappa = 3.53$ and 11 iterations.

We remark that about 90 % of the total runtime is used for the assembling part of the program including the Schur complement computations, where a majority is spent for calculating the LU-factorizations of the local matrices. This indicates the importance of replacing the direct solver with inexact solvers on each patch, see, e.g., [22, 26] for the finite element case. Furthermore, we note that, especially in 3D, an additional bottleneck is the memory demand of the direct solvers.

Table 6 2D example: Timings of Algorithm C with coefficient scaling

#dofs = 121824	Wall-clock time (s)	Relative time in %
Preparing the bookkeeping	0.011	0.03
Assembling all patch local $K^{(k)}$	6.2	15.42
Partitioning w.r.t. B and I	0.087	0.22
Assembling C	0.016 s	0.04
Calculating LU-fact. of K_{II}	15	37.31
Calculating LU-fact. of (19)	15 s	37.31
Assembling and LU-fact of $S_{\Pi\Pi}$	0.46	1.14
Assemble rhs.	0.094	0.23
Total assembling	37	92.04
One PCG iteration	0.22	–
Solving the system	2.5	6.22
Calculating the solution u	0.5	1.24
Total spent time	40.2	100.00

6 Conclusions

We have derived condition number estimates for the IETI-DP method and extended the existent theory to domains which cannot be represented by a single geometrical mapping. Due to the fact that we only considered open knot vectors, we could identify basis function on the interface and on the interior. This assumption implies that the discrete solution is only C^0 smooth across patch interfaces. However, under this assumption, we were able to find an improved condition number bound of the IETI-DP method using the Dirichlet preconditioner with stiffness scaling. Numerical examples with two and three dimensional domains, different choices of primal variables and different scaling methods confirmed the theoretical results presented in Sect. 3. Moreover, the numerical results indicate the robustness with respect to jumping diffusion coefficients across the interfaces. In [19], we have obtained similar numerical results for solving multipatch discontinuous Galerkin (dG) IgA schemes, proposed and investigated in [25], by means of IETI-DP methods following the approach developed by [13] for composite finite element and dG methods.

Acknowledgements This research work was supported by the National Research Network (NFN) “Geometry + Simulation” funded by the Austrian Science Fund (FWF) under the grand S117-03.

References

1. Apostolatos A, Schmidt R, Wüchner R, Bletzinger K-U (2014) A Nitsche-type formulation and comparison of the most common domain decomposition methods in isogeometric analysis.

Internat J Numer Methods Eng 97(7):473–504

2. Bazilevs Y, Beirão da Veiga L, Cottrell JA, Hughes TJR, Sangalli G (2006) Isogeometric analysis: approximation, stability and error estimates for h-refined meshes. Math Models Methods Appl Sci 16(7):1031–1090
3. Beirão da Veiga L, Buffa A, Sangalli G, Vázquez R (2014) Mathematical analysis of variational isogeometric methods. Acta Numer 23:157–287
4. Beirão da Veiga L, Chinosi C, Lovadina C, Pavarino LF (2010) Robust BDDC preconditioners for Reissner-Mindlin plate bending problems and MITC elements. SIAM J Numer Anal 47(6):4214–4238
5. Beirão da Veiga L, Cho D, Pavarino LF, Scacchi S (2012) Overlapping Schwarz methods for isogeometric analysis. SIAM J Numer Anal 50(3):1394–1416
6. Beirão Da Veiga L, Cho D, Pavarino LF, Scacchi S (2013) BDDC preconditioners for isogeometric analysis. Math Models Methods Appl Sci 23(6):1099–1142
7. Beirão da Veiga L, Cho D, Pavarino LF, Scacchi S (2013) Isogeometric Schwarz preconditioners for linear elasticity systems. Comput Methods Appl Mech Eng 253:439–454
8. Beirão Da Veiga L, Pavarino LF, Scacchi S, Widlund OB, Zampini S (2014) Isogeometric BDDC preconditioners with deluxe scaling. SIAM J Sci Comput 36(3):A1118–A1139
9. Bercovier M, Soloveichik I (2015) Overlapping non-matching meshes domain decomposition method in isogeometric analysis. arxiv:1502.03756 [math.NA]
10. Cottrell JA, Hughes TJR, Bazilevs Y (2009) Isogeometric analysis: toward integration of CAD and FEA. Wiley, USA
11. Dohrmann CR (2003) A preconditioner for substructuring based on constrained energy minimization. SIAM J Sci Comput 25(1):246–258
12. Dohrmann CR, Widlund OB (2013) Some recent tools and a BDDC algorithm for 3D problems in H(curl). In: Bank R, Holst M, Widlund O, Xu J (eds) Domain decomposition methods in science and engineering XX, vol 91 of Lect Notes Comput Sci Eng. Springer, Berlin, pp 15–25
13. Dryja M, Galvis J, Sarkis M (2013) A FETI-DP preconditioner for a composite finite element and discontinuous Galerkin method. SIAM J Numer Anal 51(1):400–422
14. Eigen v3. http://eigen.tuxfamily.org, 2010. Started by B. Jacob (founder) and G. Guennebaud (guru)
15. Farhat C, Lesoinne M, Le Tallec P, Pierson K, Rixen D (2001) FETI-DP: a dual-primal unified FETI method—part I: a faster alternative to the two-level FETI method. Internat J Numer Methods Eng 50(7):1523–1544
16. Farhat C, Roux F-X (1991) A method of finite element tearing and interconnecting and its parallel solution algorithm. Int J Numer Methods Eng 32(6):1205–1227
17. G+Smo (Geometry + Simulation Modules) v0.8.1. http://gs.jku.at/gismo, 2015. Coordinator and maintainer A. Mantzaflaris
18. Hesch C, Betsch P (2012) Isogeometric analysis and domain decomposition methods. Comput Methods Appl Mech Eng 213–216:104–112
19. Hofer C, Langer U (2017) Dual-primal isogeometric tearing and interconnecting solvers for multipatch dG-IgA equations. Comput Methods Appl Mech Engrg 316:2–21
20. Hughes TJR, Cottrell JA, Bazilevs Y (2005) Isogeometric analysis: CAD, finite elements, NURBS, exact geometry and mesh refinement. Comput Methods Appl Mech Eng 194(39–41):4135–4195
21. Jüttler B, Langer U, Mantzaflaris A, Moore SE, Zulehner W (2014) Geometry + Simulation modules: implementing isogeometric analysis. Proc Appl Math Mech 14(1):961–962
22. Klawonn A, Rheinbach O (2007) Inexact FETI-DP methods. Int J Numer Methods Eng 69(2):284–307
23. Kleiss S, Pechstein C, Jüttler B, Tomar S (2012) IETI—isogeometric tearing and interconnecting. Comput Methods Appl Mech Eng 247–248:201–215
24. Langer U, Moore SE (2016) Discontinuous Galerkin isogeometric analysis of elliptic PDEs on surfaces. In: Dickopf T, Gander M, Halpern L, Krause R, Pavarino LF (eds) Domain decomposition methods in science and engineering XXII, vol 104 of Lect Notes Comput Sci Eng. Springer, Berlin, pp 319–326

25. Langer U, Toulopoulos I (2015) Analysis of multipatch discontinuous Galerkin IgA approximations to elliptic boundary value problems. Comput Vis Sci 17(5):217–233
26. Li J, Widlund OB (2007) On the use of inexact subdomain solvers for BDDC algorithms. Comput Methods Appl Mech Eng 196(8):1415–1428
27. Mandel J, Dohrmann CR, Tezaur R (2005) An algebraic theory for primal and dual substructuring methods by constraints. Appl Numer Math 54(2):167–193
28. Pechstein C (2013) Finite and boundary element tearing and interconnecting solvers for multiscale problems. Springer, Berlin
29. Schumaker LL (2007) Spline functions: basic theory, 3rd edn. Cambridge University Press, Cambridge
30. Toselli A, Widlund OB (2005) Domain decomposition methods—algorithms and theory. Springer, Berlin

$\mathbf{C^0}$-Interior Penalty Discontinuous Galerkin Approximation of a Sixth-Order Cahn-Hilliard Equation Modeling Microemulsification Processes

Ronald H. W. Hoppe and Christopher Linsenmann

Abstract Microemulsions can be modeled by an initial-boundary value problem for a sixth order Cahn-Hilliard equation. Introducing the chemical potential as a dual variable, a Ciarlet-Raviart type mixed formulation yields a system consisting of a linear second order evolutionary equation and a nonlinear fourth order equation. The spatial discretization is done by a C^0 Interior Penalty Discontinuous Galerkin (C^0IPDG) approximation with respect to a geometrically conforming simplicial triangulation of the computational domain. The DG trial spaces are constructed by C^0 conforming Lagrangian finite elements of polynomial degree $p \geq 2$. For the semidiscretized problem we derive quasi-optimal a priori error estimates for the global discretization error in a mesh-dependent C^0IPDG norm. The semidiscretized problem represents an index 1 Differential Algebraic Equation (DAE) which is further discretized in time by an s-stage Diagonally Implicit Runge-Kutta (DIRK) method of order $q \geq 2$. Numerical results show the formation of microemulsions in an oil/water system and confirm the theoretically derived convergence rates.

1 Introduction

Microemulsions are thermodynamically stable colloidal dispersions of an oil/water system that typically occur as oil-in-water, water-in-oil, or water/oil droplets with a diameter up to 200 nm. They are thus considerably smaller than ordinary emulsions (macroemulsions). Moreover, in contrast to macroemulsions whose generation

R. H. W. Hoppe (✉)
Department of Mathematics, University of Houston, Houston, TX 77204-3008, USA
e-mail: rohop@math.uh.edu

C. Linsenmann
Institute of Mathematics, University of Augsburg, 86159 Augsburg, Germany
e-mail: christopher.linsenmann@math.uni-augsburg.de

B. N. Chetverushkin et al. (eds.), *Contributions to Partial Differential Equations and Applications*, Computational Methods in Applied Sciences 47,
https://doi.org/10.1007/978-3-319-78325-3_16

requires strong shear forces, microemulsions can be created by simple mixing. Due to their efficient drug solubilization capacity and bioavailability, microemulsions have significant applications in pharmacology as drug carriers for the delivery of hydrophilic as well as lipophilic drugs. Other applications include cleaning and polishing processes, food processing, and cutting oils (cf. [14, 21–23, 26, 27]).

As far as the mathematical modeling is concerned, for ternary oil-water-microemulsions Gompper et al. [15–18] have considered a second order Ginzburg-Landau free energy so that the dynamics of the microemulsification process can be described by an initial-boundary value problem for a sixth order Cahn-Hilliard equation. The existence and uniqueness of strong and weak solutions has been investigated analytically by Pawlow et al. [24, 25, 28].

For the numerical simulation of the microemulsification process, we introduce the chemical potential as a dual variable and consider a Ciarlet-Raviart type mixed formulation as a system consisting of a linear second order evolutionary equation and a nonlinear fourth order elliptic equation. The spatial discretization is taken care of by a C^0-Interior Penalty Discontinuous Galerkin (C^0-IPDG) approximation with respect to a geometrically conforming simplicial triangulation of the computational domain. The DG trial spaces are constructed by C^0 conforming Lagrangian finite elements of polynomial degree $p \geq 2$. We note that IPDG methods for the standard fourth order Cahn-Hilliard equation have been studied in [31] based on IPDG approximations of fourth order problems including the biharmonic equation considered in [5, 10] (cf. also [3, 11–13]). The semidiscretized problem represents an initial value problem for an index 1 Differential Algebraic Equation (DAE) which is discretized in time by an s-stage Diagonally Implicit Runge-Kutta method of order $q \geq 2$ with respect to a partitioning of the time interval (cf., e.g., [1, 7, 19]). The resulting parameter dependent nonlinear algebraic system is numerically solved by a predictor-corrector continuation strategy with the time step size as the continuation parameter featuring constant continuation as a predictor and Newton's method as corrector.

The paper is organized as follows: After some notations and preliminaries in Sect. 2, in Sect. 3 we present the initial-boundary value problem for the sixth order Cahn-Hilliard equation based on a Ginzburg-Landau free energy and introduce a Ciarlet-Raviart type mixed formulation as a system consisting of a linear second order evolutionary equation and a nonlinear fourth order elliptic equation. Then, Sect. 4 is devoted to the semidiscretization in space by the C^0IPDG method. Quasi-optimal a priori error estimates for the global discretization error both in the primal and in the dual variable are derived in Sect. 5. In Sect. 6, very briefly we discuss the discretization in time by an s-stage DIRK method of order q and the numerical solution of the resulting parameter dependent nonlinear algebraic system by a predictor-corrector continuation strategy. In the final Sect. 7, we present numerical results which show the formation of water-in-oil and oil-in-water droplets in a ternary water-oil-microemulsion system and confirm to some extent the theoretically derived convergence rates.

2 Notations and Preliminaries

We use standard notation from Lebesgue and Sobolev space theory (cf., e.g., [29]). In particular, for a bounded domain $\Omega \subset \mathbb{R}^d$, $d \in \mathbb{N}$, we refer to $L^p(\Omega)$, $1 \le p < \infty$, as the Banach space of p-th power Lebesgue integrable functions on Ω with norm $\|\cdot\|_{0,p,\Omega}$ and to $L^\infty(\Omega)$ as the Banach space of essentially bounded functions on Ω with norm $\|\cdot\|_{0,\infty,\Omega}$. For functions $v_i \in L^{p_i}(\Omega)$, $1 \le i \le 3$, where $p_i \in \mathbb{R}_+$, $\sum_{i=1}^3 1/p_i = 1$, the generalized Hölder inequality

$$\int_\Omega \prod_{i=1}^3 |v_i|\,dx \le \prod_{i=1}^3 \|v_i\|_{0,p_i,\Omega} \tag{1}$$

holds true. Further, we denote by $W^{s,p}(\Omega)$, $s \in \mathbb{R}_+$, $1 \le p \le \infty$, the Sobolev spaces with norms $\|\cdot\|_{s,p,\Omega}$. We note that for $p = 2$ the spaces $L^2(\Omega)$ and $W^{s,2}(\Omega) = H^s(\Omega)$ are Hilbert spaces with inner products $(\cdot,\cdot)_{0,2,\Omega}$ and $(\cdot,\cdot)_{s,2,\Omega}$. In the sequel, we will suppress the subindex 2 and write $(\cdot,\cdot)_{0,\Omega}$, $(\cdot,\cdot)_{s,\Omega}$ and $\|\cdot\|_{0,\Omega}$, $\|\cdot\|_{s,\Omega}$ instead of $(\cdot,\cdot)_{0,2,\Omega}$, $(\cdot,\cdot)_{s,2,\Omega}$ and $\|\cdot\|_{0,2,\Omega}$, $\|\cdot\|_{s,2,\Omega}$.

For $T > 0$ and a Banach space V with norm $\|\cdot\|_V$ the space $L^p((0,T),V)$, $1 \le p \le \infty$, refers to the Banach space of all functions v such that $v(t) \in V$ for almost all $t \in (0,T)$ with norm

$$\|v\|_{L^p((0,T),V)} := \begin{cases} \left(\int\limits_0^T \|v(t)\|_V^p\,dt\right)^{1/p}, & 1 \le p < \infty, \\ \operatorname*{ess\,sup}\limits_{t\in(0,T)} \|v(t)\|_V, & p = \infty. \end{cases}$$

The spaces $W^{s,p}((0,T),V)$, $s \in \mathbb{R}_+$, $1 \le p \le \infty$, are defined analogously. Finally, $C([0,T],V)$ denotes the Banach space of functions v such that $v(t) \in V$ for all $t \in [0,T]$ with norm

$$\|v\|_{C([0,T],V)} := \max_{t\in[0,T]} \|v(t)\|_V.$$

3 Sixth Order Cahn-Hilliard Equation

Given a bounded domain $\Omega \subset \mathbb{R}^2$ with boundary $\Gamma = \partial\Omega$ and exterior unit normal vector $\boldsymbol{n}_\Gamma$, denoting by $T > 0$ the final time, and setting $Q := \Omega \times (0,T)$, $\Sigma = \Gamma \times (0,T)$, we consider the following sixth order Cahn-Hilliard equation

$$\sigma \frac{\partial c}{\partial t} - M\Delta\Big(\kappa\Delta^2 c - a(c)\Delta c - \frac{1}{2}a'(c)|\nabla c|^2 + f_0(c)\Big) = 0 \quad \text{in } Q \tag{2a}$$

with the boundary conditions

$$\boldsymbol{n}_\Gamma \cdot \nabla c = \boldsymbol{n}_\Gamma \cdot \nabla \mu(c) = \boldsymbol{n}_\Gamma \cdot \nabla \Delta c = 0 \quad \text{on } \Sigma \tag{2b}$$

and the initial condition

$$c(\cdot, 0) = c_0 \quad \text{in } \Omega. \tag{2c}$$

Here, σ is a surface energy density, M stands for the mobility which in the sequel will be assumed to be a positive constant, κ is a positive constant as well, and the coefficient function $a(c)$ is assumed to be of the form

$$a(c) = a_0 + a_2 c^2, \quad a_0 \in \mathbb{R}, \; a_2 > 0. \tag{3}$$

The function $f_0(c) = \delta F_0(c)/\delta c$ is the variational derivative of the multiwell free energy

$$F_0(c) = \int_\Omega \frac{\beta}{2}\,(c+1)^2(c^2+h_0)(c-1)^2, \quad h_0 \in \mathbb{R},$$

where β is another surface energy density and $h_0 \in \mathbb{R}$ measures the deviation from the oil-water-microemulsion coexistence. Moreover, $\mu(c)$ denotes the chemical potential which is the variational derivative

$$\mu(c) = \frac{\delta F(c)}{\delta c}$$

of the total free energy

$$F(c) = F_0(c) + \int_\Omega \Big(\frac{1}{2}a(c)|\nabla c|^2 + \frac{1}{2}\kappa|\Delta c|^2\Big)\,dx, \tag{4}$$

and c_0 is a given initial condition.

Remark 1 The initial-boundary value problem (2a)–(2c) describes the dynamics of ternary oil-water-microemulsion systems where the solution c is an order parameter representing the local difference between the oil and water concentrations. We note that the Ginzburg-Landau free energy (4) for such systems has been suggested in [15–18].

For bounded convex domains with boundary Γ of class C^6 and initial data c_0 such that $c_0 \in H^5(\Omega)$ with spatial mean

$$c_m := \frac{1}{|\Omega|}\int_\Omega c_0\,dx$$

satisfying the compatibility conditions

$$\boldsymbol{n}_\Gamma \cdot \nabla c_0 = \boldsymbol{n}_\Gamma \cdot \nabla \Delta c_0 = 0 \quad \text{on } \Gamma, \tag{5}$$

it has been shown in [24] that the initial-boundary value problem for the sixth order Cahn-Hilliard equation (2a)–(2c) has a unique solution global in time such that

$$c \in L^2((0,T), H^6(\Omega)) \cap H^1((0,T), H^4(\Omega)),$$
$$c(\cdot,0) = c_0, \quad \frac{1}{|\Omega|} \int_\Omega c(t)\, dx = c_m \quad \text{for all } t \in \mathbb{R}_+.$$

Introducing the chemical potential $\mu(c)$ as an additional unknown $w := \mu(c)$, the sixth order Cahn-Hilliard equation (2a) can be equivalently formulated as a system of a linear second order evolutionary equation and a nonlinear fourth order elliptic equation in (c, w) according to

$$\sigma \frac{\partial c}{\partial t} - M \Delta w = 0 \quad \text{in } Q, \tag{6a}$$

$$\kappa \Delta^2 c - a(c)\Delta c - a_2 c|\nabla c|^2 + f_0(c) - w = 0 \quad \text{in } Q \tag{6b}$$

with the boundary conditions

$$\boldsymbol{n}_\Gamma \cdot \nabla c = \boldsymbol{n}_\Gamma \cdot \nabla w = \boldsymbol{n}_\Gamma \cdot \nabla \Delta c = 0 \quad \text{on } \Sigma \tag{6c}$$

and the initial condition

$$c(\cdot,0) = c_0 \quad \text{in } \Omega. \tag{6d}$$

We set

$$V := H^1(\Omega), \quad Z := \{z \in H^2(\Omega) \mid \boldsymbol{n}_\Gamma \cdot \nabla z = 0 \text{ on } \Gamma\}. \tag{7}$$

Observing

$$\nabla \cdot (a(c)\nabla c) = a(c)\Delta c + 2a_2|\nabla c|^2,$$

we define

$$(g(c), v)_{0,\Omega} := -(a(c)\Delta c, v)_{0,\Omega} - (a_2 c|\nabla c|^2, v)_{0,\Omega} + (f_0(c), v)_{0,\Omega}, \quad v \in Z. \tag{8}$$

A pair (c, w) is said to be a weak solution of (6a)–(6d), if for all $v \in V$ and $z \in Z$ it holds

$$\sigma \left\langle \frac{\partial c}{\partial t}, v \right\rangle_{V^*,V} + M(\nabla w, \nabla v)_{0,\Omega} = 0, \tag{9a}$$

$$\kappa(\Delta c, \Delta z)_{0,\Omega} + (g(c), z)_{0,\Omega} - (w, z)_{0,\Omega} = 0, \tag{9b}$$

and if the initial condition

$$c(\cdot, 0) = c_0. \tag{9c}$$

is satisfied.

Remark 2 The existence and uniqueness of a weak solution satisfying

$$\begin{aligned} &c \in H^1((0,T), V^*) \cap L^\infty((0,T), Z) \cap L^2((0,T), H^3(\Omega)),\\ &w \in L^2((0,T), V) \end{aligned}$$

has been shown in [28].

4 C⁰-Interior Penalty Discontinuous Galerkin Approximation

For semidiscretization in space of the coupled system (6a)–(6d) we will use the C^{0}IPDG method with respect to a simplicial triangulation of the computational domain. Due to the convexity of the computational domain, we can use the Ciarlet-Raviart mixed formulation of (6b) by introducing $z = \Delta c$ as an additional unknown so that (6b) can be written as the following system of two second order equations

$$z = \Delta c, \tag{10a}$$

$$\kappa \Delta z - a(c)\Delta c - a_2 c|\nabla c|^2 + f_0(c) = w. \tag{10b}$$

Multiplying (10a) by a test function $\varphi \in H^1(\Omega)$ and (10b) by a test function $\psi \in H^2(\Omega)$ and integrating over Ω, integration by parts and observing (6c), (8) yields the weak formulation

$$(z, \varphi)_{0,\Omega} = -(\nabla c, \nabla \varphi)_{0,\Omega}, \tag{11a}$$

$$(\kappa z, \Delta \psi)_{0,\Omega} - (\kappa z, \boldsymbol{n} \cdot \nabla \psi)_{0,\Gamma} + (g(c), \psi)_{0,\Omega} = (w, \psi)_{0,\Omega}. \tag{11b}$$

We assume $\mathscr{T}_h(\Omega)$ to be a shape-regular simplicial triangulation of Ω. For $D \subseteq \overline{\Omega}$, we denote by $\mathscr{E}_h(D)$ the sets of nodal points of $\mathscr{T}_h$ in D. For $K \in \mathscr{T}_h(\Omega)$ and $E \in \mathscr{E}_h(\bar{\Omega})$ we further refer to h_K and h_E as the diameter of K and the length of E. We set $h := \max\{h_K \mid K \in \mathscr{T}_h(\Omega)\}$. For two quantities $A, B \in \mathbb{R}_+$ we use the notation $A \lesssim B$, if there exists a constant $C > 0$, independent of h, such that $A \leq CB$.

Denoting by $P_p(K)$, $p \in \mathbb{N}$, the linear space of polynomials of degree $\leq p$ on K, for $p \geq 2$ we set

$$Q_h^{(p)} := \{v_h \in L^2(\Omega) \mid v_h|_K \in P_p(T),\ K \in \mathscr{T}_h\}$$

and refer to

$$V_h^{(p)} := Q_h^{(p)} \cap H^1(\Omega)$$

as the finite element space of Lagrangian finite elements of type p (cf., e.g., [4, 8]). We refer to $\mathscr{N}_h(\Omega)$ as the set of nodal points such that any $v_h \in V_h^{(p)}$ is uniquely determined by its degrees of freedom $v_h(a)$, $a \in \mathscr{N}_h(\Omega)$ and to $I_h : H^s(\Omega) \to V_h^{(p)}$, $s \geq 2$, as the nodal interpolation operator.

In the sequel, we will use the inverse inequalities [30]

$$\|\nabla v_h\|_{0,K} \leq C_{\mathrm{Inv}}^{(1)} p^2\, h^{-1}\, \|v_h\|_{0,K}, \quad v_h \in V_h^{(p)}, \tag{12a}$$

$$\|\Delta v_h\|_{0,K} \leq C_{\mathrm{Inv}}^{(2)} (p-1)^2\, h^{-1}\, \|\nabla v_h\|_{0,K}, \quad v_h \in V_h^{(p)}, \tag{12b}$$

and the trace inequality [30]

$$\|v_h\|_{0,\partial K} \leq C_{\mathrm{Tr}}\, p\, h^{-1/2}\, \|v_h\|_{0,K}, \quad v_h \in V_h^{(p)}. \tag{12c}$$

We note that $V_h^{(p)} \not\subset H^2(\Omega)$ and hence, $V_h^{(p)}$ is a nonconforming finite element space for the approximation of the fourth order equation (6b). In particular, for a function z_h on $\overline{\Omega}$ that is elementwise polynomial, we define averages and jumps according to

$$\{z_h\}_E := \begin{cases} \frac{1}{2}(z_h|_{E\cap T_+} + z_h|_{E\cap T_-}), & E \in \mathscr{E}_h(\Omega), \\ z_h|_E, & E \in \mathscr{E}_h(\Gamma), \end{cases}$$

$$[z_h]_E := \begin{cases} z_h|_{E\cap T_+} - z_h|_{E\cap T_-}, & E \in \mathscr{E}_h(\Omega), \\ z_h|_E & E \in \mathscr{E}_h(\Gamma). \end{cases}$$

The general C^{0}DG approximation of (11a), (11b) reads: Given $w_h \in V_h^{(p)}$, find $(c_h, z_h) \in V_h^{(p)} \times Q_h^{(p)}$ such that for all $(\varphi_h, v_h) \in Q_h^{(p)} \times V_h^{(p)}$ it holds

$$\sum_{K\in\mathscr{T}_h(\Omega)} \Big((z_h, \varphi_h)_{0,K} + (\nabla c_h, \nabla\varphi_h)_{0,K}\Big) - \sum_{E\in\mathscr{E}_h(\bar{\Omega})} (\boldsymbol{n}_E \cdot \hat{\boldsymbol{c}}_E, \varphi_h)_{0,\partial K} = 0, \tag{13a}$$

$$\sum_{K\in\mathscr{T}_h(\Omega)} \Big((\kappa z_h, \Delta v_h)_{0,T} + (g(c_h), v_h)_{0,K}\Big) - \sum_{E\in\mathscr{E}_h(\bar{\Omega})} \Big((\hat{\boldsymbol{z}}_E, \nabla v_h)_{0,E} - (w_h, v_h)_{0,K}\Big) = 0, \tag{13b}$$

where $\hat{\boldsymbol{c}}_E$ and $\hat{\boldsymbol{z}}_E$ are suitably chosen numerical flux functions that determine the type of C^{0}DG approximation. In particular, for the C^{0}IPDG approximation we choose

$$\hat{\boldsymbol{c}}_E := \begin{cases} \{\nabla c_h\}_E, & E \in \mathscr{E}_h(\Omega), \\ 0, & E \in \mathscr{E}_h(\Gamma), \end{cases} \tag{13c}$$

$$\hat{\boldsymbol{z}}_E := \left(\{\Delta c_h\}_E - \frac{\alpha}{h_E} \left[\frac{\partial c_h}{\partial n} \right]_E \right) \boldsymbol{n}_E, \quad E \in \mathscr{E}_h(\bar{\Omega}), \tag{13d}$$

where $\alpha > 0$ is a penalization parameter. The choice (13c), (13d) has the advantage that for $\varphi_h = \kappa \Delta v_h$ in (13a) we may eliminate the dual variable z_h from the system and thus arrive at the following primal variational formulation of the C^{0}IPDG approximation: Find $c_h \in V_h^{(p)}$ such that for all $v_h \in V_h^{(p)}$ it holds

$$a_h^{DG}(c_h, v_h) + \sum_{K \in \mathscr{T}_h(\Omega)} (g(c_h), v_h)_{0,K} = (w_h, v_h)_{0,\Omega},$$

where $a_h^{DG}(\cdot, \cdot) : V_h^{(p)} \times V_h^{(p)} \to \mathbb{R}$ stands for the C^{0}IPDG bilinear form

$$a_h^{DG}(c_h, v_h) := \sum_{K \in \mathscr{T}_h(\Omega)} (\kappa \Delta c_h, \Delta v_h)_{0,K} - \sum_{E \in \mathscr{E}_h(\overline{\Omega})} \Big((\kappa \boldsymbol{n}_E \cdot \{\nabla c_h\}_E, [\Delta v_h]_E)_{0,E}$$
$$+ (\kappa [\Delta c_h]_E, \boldsymbol{n}_E \cdot \{\nabla v_h\}_E)_{0,E} \Big) + \sum_{E \in \mathscr{E}_h(\overline{\Omega})} \frac{\alpha}{h_E} (\boldsymbol{n}_E \cdot [\nabla c_h]_E, \boldsymbol{n}_E \cdot [\nabla v_h]_E)_{0,E}.$$

We note that the C^{0}IPDG bilinear form is not well-defined for functions $c \in Z$, since $\Delta c|_E$, $E \in \mathscr{E}_h(\bar{\Omega}$, does not live in $L^2(E)$. This can be cured by means of a lifting operator

$$L : V_h^{(p)} + Z \to V_h^{(p)}$$

which is defined according to

$$\int_\Omega L(c)\, v_h\, dx = - \sum_{E \in \mathscr{E}_h(\bar{\Omega})} \int_E \boldsymbol{n}_E \cdot [\nabla c]_E\, v_h\, ds.$$

We define an extension $\tilde{a}_h^{DG}(\cdot, \cdot) : (V_h^{(p)} + Z) \times (V_h^{(p)} + Z) \to \mathbb{R}$ as follows:

$$\tilde{a}_h^{DG}(c, v) := \sum_{K \in \mathscr{T}_h(\Omega)} \int_K \Big(\Delta c\, \Delta v + L(c)\, \Delta v + \Delta c\, L(v) \Big) dx$$
$$+ \sum_{E \in \mathscr{E}_h(\bar{\Omega})} \frac{\alpha}{h_E}\, \boldsymbol{n}_E \cdot [\nabla c]_E\, \boldsymbol{n}_E \cdot [\nabla v]_E\, ds.$$

On $V_h^{(p)} + Z$ we introduce the mesh-dependent IPDG semi-norm

$$|c|_{2,h,\Omega} := \left(\sum_{K\in\mathscr{T}_h(\Omega)} \|\Delta c\|_{0,K}^2 + \sum_{E\in\mathscr{E}_h(\bar{\Omega})} \frac{\alpha}{h_E} \|\boldsymbol{n}_E \cdot [\nabla c]_E\|_{0,E}^2 \right)^{1/2}$$

and the mesh-dependent IPDG norm

$$\|c\|_{2,h,\Omega} := \left(|c|_{2,h,\Omega}^2 + \|c\|_{0,\Omega}^2 \right)^{1/2}.$$

From the Poincaré-Friedrichs inequality for piecewise H^2 functions (cf., e.g., [6]) we deduce that there exists a constant $C_{\mathrm{PF}} > 0$ such that

$$\|\nabla v\|_{0,\Omega}^2 \le C_{\mathrm{PF}} |v|_{2,h,\Omega}^2, \quad v \in V_h^{(p)} + Z. \tag{14}$$

It is not difficult to show that for sufficiently large penalty parameter α there exist constants $\gamma > 0$ and $\beta > 0$ such that the C^{0}IPDG bilinear form $\tilde{a}_h^{DG}$ satisfies the Gårding-type inequality

$$\tilde{a}_h^{DG}(c, c) \ge \gamma \|c\|_{2,h,\Omega}^2 - \beta \|c\|_{0,\Omega}^2, \quad c \in V_h^{(p)} + Z. \tag{15}$$

Moreover, there exists a constant $\Gamma > 0$ such that

$$|\tilde{a}_h^{DG}(c, v)| \le \Gamma \|c\|_{2,h,\Omega} \|v\|_{2,h,\Omega}, \quad c, v \in V_h^{(p)} + Z. \tag{16}$$

The C^{0}IPDG method for the nonlinear fourth order elliptic equation has the advantage that we may approximate the dual variable w in the linear second order evolutionary equation by a function in $V_h^{(p)}$ as well. Hence, the C^{0}IPDG approximation of the initial-boundary value problem (6a)–(6d) for the sixth order Cahn-Hilliard equation reads:

Find $(c_h, w_h) \in H^1((0,T), V_h^{(p)}) \times L^2((0,T), V_h^{(p)})$ such that for all $v_h \in V_h^{(p)}$ it holds

$$\left(\sigma \frac{\partial c_h}{\partial t}, v_h \right)_{0,\Omega} - M(\nabla w_h, \nabla v_h)_{0,\Omega} = 0, \tag{17a}$$

$$a_h^{DG}(c_h, v_h) + \sum_{K\in\mathscr{T}_h(\Omega)} (g(c_h), v_h)_{0,K} - (w_h, v_h)_{0,\Omega} = 0, \tag{17b}$$

$$c_h(\cdot, 0) = I_h c_0. \tag{17c}$$

Remark 3 (i) The unique solvability of (17a)–(17c) can be shown by similar arguments as in [28].

(ii) The C^{0}IPDG approximation (17a)–(17c) is consistent with the weak formulation (9a)–(9c) of the initial-boundary value problem (6a)–(6d) in the sense that for all $v_h \in V_h^{(p)}$ it holds (cf., e.g., [5])

$$\left\langle \sigma \ \frac{\partial c}{\partial t}, v_h \right\rangle_{V,V^*} - M\ (\nabla w, \nabla v_h)_{0,\Omega} = 0,$$

$$\tilde{a}_h^{DG}(c, v_h) + \sum_{K \in \mathscr{T}_h(\Omega)} (g(c), v_h)_{0,K} - (w, v_h)_{0,\Omega} = 0.$$

5 Quasi-Optimal a Priori Error Estimates

We suppose that for some $r \geq 5$ the domain Ω has a boundary Γ of class C^{r+1}, the initial data satisfy $c_0 \in H^r(\Omega)$ as well as the compatibility condition (5) and that the unique solution (c, w) of (9a)–(9c) satisfies the regularity assumptions

$$c \in L^2((0,T), H^{r+1}(\Omega)) \cap H^1((0,T), H^{r-1}(\Omega)) \cap H^2((0,T), H^{r-3}(\Omega)), \tag{18a}$$

$$w \in L^2((0,T), H^{r-1}(\Omega)) \cap H^1((0,T), H^{r-3}(\Omega)) \cap H^2((0,T), H^{r-5}(\Omega)). \tag{18b}$$

Remark 4 It follows from (18a), (18b) that the pair (c, w) satisfies

$$c \in C([0,T], H^r(\Omega)) \cap C^1([0,T], H^{r-2}(\Omega)), \tag{19a}$$

$$w \in C([0,T], H^{r-2}(\Omega)) \cap C^1([0,T], H^{r-4}(\Omega)). \tag{19b}$$

The regularity assumptions (18a), (18b) imply the following interpolation estimates (cf., e.g., [4, 8])

$$\int_0^t \|c - I_h c\|_{m,\Omega}^2 \, d\tau \lesssim h^{2(\min(p+1,r+1)-m)} \int_0^t |c|_{\min(p+1,r+1),\Omega}^2 \, ds, \tag{20a}$$

$$\int_0^t \left\| \frac{\partial c}{\partial s} - I_h \frac{\partial c}{\partial s} \right\|_{0,\Omega}^2 ds \lesssim h^{2\min(p+1,r-1)} \int_0^t \left| \frac{\partial c}{\partial s} \right|_{\min(p+1,r-1),\Omega}^2 ds, \tag{20b}$$

$$\|(c - I_h c)(\cdot,t)\|_{m,\Omega}^2 \lesssim h^{2(\min(p+1,r)-m)} |c(\cdot,t)|_{\min(p+1,r),\Omega}^2, \tag{20c}$$

$$\int_0^t \|w - I_h w\|_{m,\Omega}^2 \, ds \lesssim h^{2(\min(p+1,r-1)-m)} \int_0^t |w|_{\min(p+1,r-1),\Omega}^2 \, ds, \tag{20d}$$

$$\int_0^t \left\| \frac{\partial w}{\partial s} - I_h \frac{\partial w}{\partial s} \right\|_{0,\Omega}^2 ds \lesssim h^{2\min(p+1,r-3)} \int_0^t \left| \frac{\partial w}{\partial s} \right|_{\min(p+1,r-3),\Omega}^2 ds, \tag{20e}$$

$$\|(w - I_h w)(\cdot,t)\|_{m,\Omega}^2 \lesssim h^{2(\min(p+1,r-2)-m)} |w(\cdot,t)|_{\min(p+1,r-2),\Omega}^2. \tag{20f}$$

For the interpolation error in the mesh-dependent IPDG-norm it follows from (20) that

$$\int_0^t \|c - I_h c\|^2_{2,h,\Omega}\, d\tau \lesssim h^{2(\min(p+1,r+1)-2)} \int_0^t |c|^2_{\min(p+1,r+1),\Omega}\, d\tau, \tag{21a}$$

$$\|(c - I_h c)(\cdot,t)\|^2_{2,h,\Omega} \lesssim h^{2(\min(p+1,r)-2)}\, |c(\cdot,t)|^2_{\min(p+1,r),\Omega}. \tag{21b}$$

Theorem 1 *Let* (c, w) *and* (c_h, w_h) *be the solutions of* (9a)–(9c) *and* (17a)–(17c). *Under the regularity assumptions* (18a), (18b), *and* (19a), (19b) *there exists a constant* $C > 0$, *independent of* h, *such that for all* $0 < t \leq T$ *it holds*

$$\begin{aligned}
&\|(c - c_h)(\cdot,t)\|^2_{2,h,\Omega} + \int_0^t \|c - c_h\|^2_{2,h,\Omega}\, ds + \int_0^t \|\nabla(w - w_h)\|^2_{0,\Omega}\, ds \\
&\quad \lesssim h^{2(p_{r+1}-2)} \int_0^t |c|^2_{p_{r+1},\Omega}\, ds + h^{2(p_{r-1}-2)} \int_0^t \left|\frac{\partial c}{\partial s}\right|^2_{p_{r-1},\Omega} ds \\
&\quad + h^{2(p_{r-1}-1)} \int_0^t |w|^2_{p_{r-1},\Omega}\, ds + h^{2p_{r-3}} \int_0^t \left|\frac{\partial w}{\partial s}\right|^2_{\min(p+1,r-3),\Omega} ds \\
&\quad + h^{2(p_r-2)} |c_0|^2_{\min(p+1,r),\Omega} + h^{2p_{r-2}} |w_0|^2_{p_{r-2},\Omega},
\end{aligned} \tag{22}$$

where $p_\ell := \min(p+1, \ell)$.

The proof of Theorem 1 will be given by a series of lemmas and propositions.

First of all, recalling that $\tilde{a}_h^{DG}(\cdot,\cdot)$ satisfies the Gårding-type inequality (15), we perform a scaling of the primal variable c and the dual variable w according to

$$c(x,t) := \exp(\tau t)\hat{c}(x,t), \quad w(x,t) := \exp(\tau t)\hat{w}(x,t), \quad \tau > 0. \tag{23}$$

In the new variables $(\hat{c}, \hat{w})$, the system (6a)–(6d) reads

$$\sigma \frac{\partial \hat{c}}{\partial t} + \sigma\tau\hat{c} - M\Delta\hat{w} = 0 \qquad \text{in } Q, \tag{24a}$$

$$\kappa\Delta^2\hat{c} + \hat{g}(\hat{c}) - \hat{w} = 0 \qquad \text{in } Q, \tag{24b}$$

with the boundary conditions

$$\boldsymbol{n}\cdot\nabla\hat{c} = \boldsymbol{n}\cdot\nabla\hat{w} = \boldsymbol{n}\cdot\nabla\Delta\hat{c} = 0 \quad \text{on } \Sigma, \tag{24c}$$

and the initial condition

$$\hat{c}(\cdot,0) = c_0 \quad \text{in } \Omega, \tag{24d}$$

where

$$\hat{g}(\hat{c}) := -\hat{a}(\hat{c})\Delta\hat{c} - a_2 \exp(2\tau t)\hat{c}|\nabla\hat{c}|^2 + \hat{f}_0(\hat{c}), \tag{24e}$$

$$\hat{a}(\hat{c}) := a_0 + a_2 \exp(2\tau t)\hat{c}^2, \tag{24f}$$

$$\hat{f}_0(\hat{c}) := \beta(\exp(\tau t)\hat{c} + 1)(\exp(\tau t)\hat{c} - 1)(\exp(2\tau t)\hat{c}^3 - (1 - 2h_0)\hat{c}). \tag{24g}$$

A pair (c, w) is said to be a weak solution of (6a)–(6d), if for all $v \in Z$ it holds

$$\sigma \left\langle \frac{\partial \hat{c}}{\partial t}, v \right\rangle_{V^*,V} + \sigma\tau(\hat{c}, v)_{0,\Omega} + M(\nabla\hat{w}, \nabla v)_{0,\Omega} = 0,$$
$$\kappa(\Delta\hat{c}, \Delta v)_{0,\Omega} + (\hat{g}(\hat{c}), v)_{0,\Omega} - (\hat{w}, v)_{0,\Omega} = 0,$$

and if the initial condition

$$\hat{c}(\cdot, 0) = c_0.$$

is satisfied. The semidiscrete variables (c_h, w_h) are scaled in the same way and hence, the semidiscrete approximation requires the computation of $(\hat{c}_h, \hat{w}_h) \in V_h^{(p)} \times V_h^{(p)}$ such that for all $v_h \in V_h^{(p)}$ it holds

$$\left(\sigma \frac{\partial \hat{c}_h}{\partial t}, v_h\right)_{0,\Omega} + \sigma\tau(\hat{c}_h, v_h)_{0,\Omega} - M(\nabla\hat{w}_h, \nabla v_h)_{0,\Omega} = 0, \tag{25a}$$

$$a_h^{DG}(\hat{c}_h, v_h) + \sum_{K \in \mathscr{T}_h(\Omega)} (\hat{g}(\hat{c}_h), v_h)_{0,K} - (\hat{w}_h, v_h)_{0,\Omega} = 0, \tag{25b}$$

$$\hat{c}_h(\cdot, 0) = c_{h,0}. \tag{25c}$$

Remark 5 If the regularity assumptions (18a), (18b) hold true for (c, w), they also apply to $(\hat{c}, \hat{w})$ and the interpolation estimates (20) are satisfied for $(\hat{c}, \hat{w})$ as well.

We will prove Theorem 1 based on an implicit time discretization of (24a)–(24d) and (25a)–(25c) by the backward Euler scheme with respect to an equidistant partition $\{t_m = m\Delta t \mid 0 \le m \le M\}$, $M \in \mathbb{N}$, of the time interval $[0, T]$ with step size $\Delta t = T/M$. Denoting by $(\hat{c}^m, \hat{w}^m)$ and $(\hat{c}_h^m, \hat{w}_h^m)$ approximations of $(\hat{c}, \hat{w})$ and $(\hat{c}_h, \hat{w}_h)$ at time t_m, $0 \le m \le M$, with $\hat{c}^0 = \hat{c}_0$ and $\hat{c}_h^0 = c_{h,0}$, the backward Euler scheme for (24a)–(24d) reads:

Find $(\hat{c}^m, \hat{w}^m)$ such that for all $v \in Z$ it holds

$$\sigma(\hat{c}^m, v)_{0,\Omega} + \sigma\tau\Delta t(\hat{c}^m, v)_{0,\Omega} + \Delta t(\nabla\hat{w}^m, \nabla v)_{0,\Omega} - \sigma(\hat{c}^{m-1}, v)_{0,\Omega} = 0, \tag{26a}$$

$$a_h^{DG}(\hat{c}^m, v) + (\hat{g}(\hat{c}^m, v)_{0,\Omega} - (\hat{w}^m, v)_{0,\Omega} = 0. \tag{26b}$$

The unique solvability of (26a), (26b) follows in the same way as that of (9a)–(9c).

Likewise, the backward Euler scheme for (25a)–(25c) is given by:

Find $(\hat{c}_h^m, \hat{w}_h^m)$ such that for all $v_h \in V_h^{(p)}$ it holds

$$\sigma\ (\hat{c}_h^m - \hat{c}_h^{m-1}, v_h)_{0,\Omega} + \sigma\tau\Delta t\ (\hat{c}_h^m, v_h)_{0,\Omega} + \Delta t\ (\nabla\hat{w}_h^m, \nabla v_h)_{0,\Omega} = 0, \tag{27a}$$

$$a_h^{DG}(\hat{c}_h^m, v_h) + \sum_{K\in\mathscr{T}_h(\Omega)} (\hat{g}(\hat{c}_h^m, v_h)_{0,K} - (\hat{w}_h^m, v_h)_{0,\Omega} = 0. \tag{27b}$$

Again, the unique solvability of (27a), (27b) follows in the same way as that of (17a)–(17c).

Remark 6 (i) The C^{0}IPDG approximation (27a), (27b) is consistent with (26a), (26b) in the sense that for all $v_h \in V_h^{(p)}$ it holds

$$\sigma\ (\hat{c}^m - \hat{c}^{m-1}, v_h)_{0,\Omega} + \sigma\tau\Delta t\ (\hat{c}^m, v_h)_{0,\Omega} + \Delta t\ (\nabla\hat{w}^m, \nabla v_h)_{0,\Omega} = 0,$$
$$\tilde{a}_h^{DG}(\hat{c}^m, v_h) + \sum_{K\in\mathscr{T}_h(\Omega)} (\hat{g}(\hat{c}^m, v_h)_{0,K} - (\hat{w}^m, v_h)_{0,\Omega} = 0.$$

(ii) Using similar arguments as in [25, 28] it can be shown that $\hat{c}_h^m$ is bounded in the C^{0}IPDG norm uniformly in h, i.e., there exists a constant $C_B^{(1)} > 0$, independent of h, such that

$$\|\hat{c}_h^m\|_{2,h,\Omega} \le C_B^{(1)}, \quad 0 \le m \le M. \tag{28}$$

Since $V_h^{(p)}$ is continuously embedded in $C(\bar{\Omega}$, there exists another constant $C_B^{(2)} > 0$, independent of h, such that

$$\max_{x\in\bar{\Omega}}\ |\hat{c}_h^m(x)| \le C_B^{(2)}, \quad 0 \le m \le M. \tag{29}$$

Lemma 1 *Let $\hat{g}$ be given by (24e). Then there exists a constant C_1, independent of h, such that for $\hat{c}^m \in H^r(\Omega)$, $r \ge 5$, $0 \le m \le M$, and $\hat{c}_h^m, v_h \in V_h^{(p)}$, $p \ge 2$, it holds*

$$|(\hat{g}(\hat{c}^m) - \hat{g}(\hat{c}_h^m), v_h)_{0,\Omega}| \le C_1\ \|\hat{c}^m - \hat{c}_h^m\|_{2,h,\Omega}\ \|v_h\|_{0,\Omega}. \tag{30}$$

Proof Observing (24e) we have

$$\sum_{K\in\mathscr{T}_h(\Omega)} (\hat{g}(\hat{c}^m) - \hat{g}(\hat{c}_h^m), v_h)_{0,K} = -\sum_{K\in\mathscr{T}_h(\Omega)} (\hat{a}(\hat{c}^m)\Delta\hat{c}^m - \hat{a}(\hat{c}_h^m)\Delta\hat{c}_h^m, v_h)_{0,K}$$
$$-\sum_{K\in\mathscr{T}_h(\Omega)} a_2\exp(2\tau t)(\hat{c}^m|\nabla\hat{c}^m|^2 - \hat{c}_h^m|\nabla\hat{c}_h^m|^2, v_h)_{0,K} + (\hat{f}_0(\hat{c}^m) - \hat{f}_0(\hat{c}_h^m), v_h)_{0,\Omega}. \tag{31}$$

In view of (3)

$$\hat{a}(\hat{c}_h^m)\Delta\hat{c}_h^m - \hat{a}(\hat{c}^m)\Delta\hat{c}^m = (\hat{a}(\hat{c}^m) - \hat{a}(\hat{c}_h^m))\Delta\hat{c}^m + \hat{a}(\hat{c}_h^m)(\Delta\hat{c}^m - \Delta\hat{c}_h^m)$$
$$= a_2\exp(2\tau t)(\hat{c}^m + \hat{c}_h^m)(\hat{c}^m - \hat{c}_h^m)\Delta\hat{c}^m + \hat{a}(\hat{c}_h^m)(\Delta\hat{c}^m - \Delta\hat{c}_h^m).$$

Then the first term on the right-hand side of (31) can be estimated according to

$$\begin{aligned}&\left|\sum_{K\in\mathscr{T}_h(\Omega)} (\hat{a}(\hat{c}^m)\Delta\hat{c}^m - \hat{a}(\hat{c}_h^m)\Delta\hat{c}_h^m, v_h)_{0,K}\right| \\ &\le D_1 \sum_{K\in\mathscr{T}_h(\Omega)} \|\hat{c}^m - \hat{c}_h^m\|_{0,K}\, \|v_h\|_{0,K} + D_2 \sum_{K\in\mathscr{T}_h(\Omega)} \|\Delta\hat{c}^m - \Delta\hat{c}_h^m\|_{0,K} \|v_h\|_{0,K},\end{aligned} \tag{32}$$

where the constants D_i, $1 \le i \le 2$, are given by

$$D_1 := \max_{x\in\bar{\Omega}} |(\hat{c}^m + \hat{c}_h^m)(x)\Delta\hat{c}^m(x)|, \quad D_2 := \max_{x\in\bar{\Omega}} |a_0 + a_2\exp(2\tau T)(\hat{c}_h^m(x))^2|.$$

We note that $\hat{c}^m$, $\Delta\hat{c}^m \in C(\bar{\Omega})$, since for $r \ge 5$ the spaces $Z \cap H^r(\Omega)$ and $H^{r-2}(\Omega)$ are continuously embedded in $C(\bar{\Omega})$. Moreover, due to (29) $\hat{c}_h^m$ is bounded in $C(\bar{\Omega})$ uniformly in h. Hence, the constants D_i, $1 \le i \le 2$, are well defined and bounded from above independent of h.

For the second term on the right-hand side of (31) we split

$$a_2\exp(2\tau t)(\hat{c}^m|\nabla\hat{c}^m|^2 - \hat{c}_h^m|\nabla\hat{c}_h^m|^2, v_h)_{0,K}, \quad K \in \mathscr{T}_h(\Omega),$$

by means of

$$\begin{aligned}&(a_2\exp(2\tau t)(\hat{c}^m\, |\nabla\hat{c}^m|^2 - \hat{c}_h^m\, |\nabla\hat{c}_h^m|^2, v_h)_{0,K} \\ &= a_2\exp(2\tau t)((\hat{c}^m - \hat{c}_h^m)|\nabla\hat{c}^m|^2, v_h)_{0,K} + a_2\exp(2\tau t)(\hat{c}_h^m\nabla\hat{c}^m\cdot(\nabla\hat{c}^m - \nabla\hat{c}_h^m), v_h)_{0,K} \\ &\quad + a_2\exp(2\tau t)(\hat{c}_h^m\nabla\hat{c}_h^m\cdot(\nabla\hat{c}^m - \nabla\hat{c}_h^m), v_h)_{0,K}.\end{aligned} \tag{33}$$

For the first term on the right-hand side of (33) we obtain

$$\begin{aligned}&\left|\sum_{K\in\mathscr{T}_h(\Omega)} a_2\exp(2\tau t)((\hat{c}^m - \hat{c}_h^m)|\nabla\hat{c}^m|^2, v_h)_{0,K}\right| \\ &\qquad\le D_3 \sum_{K\in\mathscr{T}_h(\Omega)} \|\hat{c}^m - \hat{c}_h^m\|_{0,K}\|v_h\|_{0,K},\end{aligned} \tag{34}$$

where

$$D_3 := a_2\exp(2\tau T)\max_{x\in\bar{\Omega}} |\nabla\hat{c}^m(x)|^2$$

which is well defined, since $\nabla\hat{c}^m \in C(\bar{\Omega})^2$.

Likewise, observing (14), the second term on the right-hand side of (33) can be estimated from above as follows:

$$\begin{aligned}\Bigg|\sum_{K\in\mathscr{T}_h(\Omega)} & a_2\exp(2\tau t)(\hat{c}_h^m\nabla\hat{c}^m\cdot(\nabla\hat{c}^m-\nabla\hat{c}_h^m),v_h)_{0,K}\Bigg|\\ &\le D_4\sum_{K\in\mathscr{T}_h(\Omega)}\|\nabla\hat{c}^m-\nabla\hat{c}_h^m\|_{0,K}\ \|v_h\|_{0,K}\le D_4\|\nabla(\hat{c}^m-\hat{c}_h^m)\|_{0,\Omega}\|v_h\|_{0,\Omega}\\ &\le C_{\mathrm{PF}}D_4|\hat{c}^m-\hat{c}_h^m|_{2,h,\Omega}\|v_h\|_{0,\Omega},\end{aligned}\tag{35}$$

where due to (29)

$$D_4:=a_2\exp(2\tau T)\max_{x\in\bar{\Omega}}|\hat{c}_h^m(x)\nabla\hat{c}^m(x)|\le a_2\exp(2\tau T)C_B^{(2)}\max_{x\in\bar{\Omega}}|\nabla\hat{c}^m(x)|.$$

Since $\nabla\hat{c}^m\in C(\bar{\Omega})^2$, we note that D_4 is well defined and independent of h.

For the third term on the right-hand side of (33) we use the generalized Hölder inequality (1) with $v_1=\nabla\hat{c}_h^m$, $v_2=\nabla\hat{c}^m-\nabla\hat{c}_h^m$, $v_3=v_h$, and $p_1=4/(1+2\varepsilon)$, $p_2=4/(1-2\varepsilon)$, $0<\varepsilon\ll 1$, and $p_3=2$.

$$\begin{aligned}|\sum_{K\in\mathscr{T}_h(\Omega)} & a_2\exp(2\tau t)(\hat{c}_h^m\nabla\hat{c}_h^m\cdot(\nabla\hat{c}^m-\nabla\hat{c}_h^m),v_h)_{0,K}|\\ &\le D_5\sum_{K\in\mathscr{T}_h(\Omega)}\int_K|\nabla\hat{c}_h^m|\,|\nabla\hat{c}^m-\nabla\hat{c}_h^m|\,|v_h|\,dx\\ &\le D_5\sum_{K\in\mathscr{T}_h(\Omega)}\|\hat{c}_h^m\|_{1,4/(1+2\varepsilon),K}\|\hat{c}^m-\hat{c}_h^m\|_{1,4/(1-2\varepsilon),K}\|v_h\|_{0,K}\\ &\le D_5\|\hat{c}_h^m\|_{1,4/(1+2\varepsilon),\Omega}\sum_{K\in\mathscr{T}_h(\Omega)}\|\hat{c}^m-\hat{c}_h^m\|_{1,4/(1-2\varepsilon),K}\|v_h\|_{0,K},\end{aligned}\tag{36}$$

where

$$D_5:=a_2\exp(2\tau T)\ \max_{x\in\bar{\Omega}}|\hat{c}_h^m(x)|\le a_2\exp(2\tau T)C_B^{(2)}.$$

Since $H^{3/2-\varepsilon}(\Omega)$ is continuously embedded in $W^{1,4/(1+2\varepsilon)}(\Omega)$ and $V_h^{(p)}$ is continuously embedded in $H^{3/2-\varepsilon}(\Omega)$ (cf., e.g., [5]), there exists a constant $D_6>0$ such that

$$\|\hat{c}_h^m\|_{1,4/(1+2\varepsilon),\Omega}\le D_6\ \|\hat{c}_h^m\|_{2,h,\Omega}.\tag{37}$$

Moreover, $H^2(K)$ is continuously embedded in $W^{1,4/(1-2\varepsilon)}(K)$ and hence, there exists a constant $D_7>0$, which can be chosen independent of h, such that for all $K\in\mathscr{T}_h(\Omega)$ it holds

$$\|\hat{c}^m-\hat{c}_h^m\|_{1,4/(1-2\varepsilon),K}\le D_7\|\hat{c}^m-\hat{c}_h^m\|_{2,K}.\tag{38}$$

Using (37) and (38) in (36), it follows that

$$\left| \sum_{K\in\mathscr{T}_h(\Omega)} a_2 \exp(2\tau t)(\hat{c}_h^m \nabla\hat{c}_h^m \cdot (\nabla\hat{c}^m - \nabla\hat{c}_h^m), v_h)_{0,K} \right| \\ \le D_8 \sum_{K\in\mathscr{T}_h(\Omega)} \|\hat{c}^m - \hat{c}_h^m\|_{2,K} \|v_h\|_{0,K}, \tag{39}$$

where due to (28)

$$D_8 := D_5\, D_6\, D_7\, \|\hat{c}_h^m\|_{2,h,\Omega} \le D_5\, D_6\, D_7\, C_B^{(1)}.$$

Finally, for the third term on the right-hand side of (31) we use that

$$\hat{f}_0(\hat{c}^m) - \hat{f}_0(\hat{c}_h^m) = \int_0^1 \hat{f}_0'(\hat{c}^m + s\,(\hat{c}_h^m - \hat{c}^m))\, ds\, (\hat{c}^m - \hat{c}_h^m)$$

to obtain

$$|(\hat{f}_0(\hat{c}^m) - \hat{f}_0(\hat{c}_h^m), v_h)_{0,\Omega}| \le D_9 \sum_{K\in\mathscr{T}_h(\Omega)} \|\hat{c}^m - \hat{c}_h^m\|_{0,K} \|v_h\|_{0,K}^2, \tag{40}$$

where

$$D_9 := \max_{x\in\bar{\Omega}} \int_0^1 |\hat{f}_0'(\hat{c}^m + s\,(\hat{c}_h^m - \hat{c}^m))|\, ds.$$

Now, (30) is a direct consequence of (32), (34), (35), (39), and (40).

Corollary 1 *Under the assumptions of Lemma 1 there exists a constant $C_2 > 0$, independent of h, such that for $0 \le m \le M$ it holds*

$$\|I_h\hat{w}^m - \hat{w}_h^m\|_{0,\Omega} \le C_2\, h^{-2}\, \|\hat{c}^m - \hat{c}_h^m\|_{2,h,\Omega} + \|\hat{w}^m - I_h\hat{w}^m\|_{0,\Omega}.$$

Proof Obviously, we have

$$\|I_h\hat{w}^m - \hat{w}_h^m\|_{0,\Omega} = \sup_{v_h\in V_h^{(p)}} \frac{|(I_h\hat{w}^m - \hat{w}_h^m, v_h)_{0,\Omega}|}{\|v_h\|_{0,\Omega}}.$$

Using (6b) and (17b) we find

$$(I_h\hat{w}^m - \hat{w}_h^m, v_h)_{0,\Omega} = (I_h\hat{w}^m - \hat{w}^m, v_h)_{0,\Omega} + (\hat{w}^m - \hat{w}_h^m, v_h)_{0,\Omega} = \tag{41}$$
$$(I_h\hat{w}^m - \hat{w}^m, v_h)_{0,\Omega} + a_h^{DG}(\hat{c}^m - \hat{c}_h^m, v_h) + (\hat{g}(\hat{c}^m) - \hat{g}(\hat{c}_h^m), v_h)_{0,\Omega}.$$

In view of (16), for the second term on the right-hand side of (41) we obtain

$$|a_h^{DG}(\hat{c}^m - \hat{c}_h^m, v_h)| \le \Gamma\ \|\hat{c}^m - \hat{c}_h^m\|_{2,h,\Omega}\ \|v_h\|_{2,h,\Omega}. \tag{42}$$

On the other hand, using (30) from Lemma 1 we find

$$|a_h^{DG}(\hat{c}^m - \hat{c}_h^m, v_h) + (\hat{g}(\hat{c}^m) - \hat{g}(\hat{c}_h^m), v_h)_{0,\Omega}| \le C_1\ \|\hat{c}^m - \hat{c}_h^m\|_{2,h,\Omega}\ \|v_h\|_{2,h,\Omega}. \tag{43}$$

The inverse inequalities (12a) and (12b) and the trace inequality (12c) imply the existence of a constant $D_{10} > 0$, independent of h, such that

$$\|v_h\|_{2,h,\Omega} \le D_{11}\ h^{-2}\ \|v_h\|_{0,\Omega}. \tag{44}$$

Using (43) and (44) in (41) gives the assertion.

We introduce the interpolation errors:

$$\begin{aligned}
&e_{int}^{(1)}(\hat{c}^\ell) := \Delta t\ \|\hat{c}^\ell - I_h\hat{c}^\ell\|_{0,\Omega}^2, \quad e_{int}^{(2)}(\hat{c}^\ell) :=\ \Delta t\ \|\nabla(\hat{c}^\ell - I_h\hat{c}^\ell)\|_{0,\Omega}^2, \quad 0 \le \ell \le m,\\
&e_{int}^{(3)}(\hat{c}^\ell) := \Delta t\ \|\frac{\hat{c}^\ell - \hat{c}^{\ell-1}}{\Delta t} - I_h(\frac{\hat{c}^\ell - \hat{c}^{\ell-1}}{\Delta t})\|_{0,\Omega}^2, \quad 1 \le \ell \le m,\\
&e_{int}^{(4)}(\hat{c}^\ell) := \Delta t\ \|\hat{c}^\ell - I_h\hat{c}^\ell\|_{2,h,\Omega}^2, \quad 0 \le \ell \le m,\\
&e_{int}^{(5)}(\hat{c}^\ell) := \Delta t\ \|\frac{\hat{c}^\ell - \hat{c}^{\ell-1}}{\Delta t} - I_h(\frac{\hat{c}^\ell - \hat{c}^{\ell-1}}{\Delta t})\|_{2,h,\Omega}^2, \quad 1 \le \ell \le m,\\
&e_{int}^{(1)}(\hat{w}^\ell) := \Delta t\ \|\hat{w}^\ell - I_h\hat{w}^\ell\|_{0,\Omega}^2, \quad e_{int}^{(2)}(\hat{w}^\ell) := \Delta t\ \|\nabla(\hat{w}^\ell - I_h\hat{w}^\ell)\|_{0,\Omega}^2, \quad 0 \le \ell \le m,\\
&e_{int}^{(3)}(\hat{w}^\ell) := \Delta t\ \|\frac{\hat{w}^\ell - \hat{w}^{\ell-1}}{\Delta t} - I_h(\frac{\hat{w}^\ell - \hat{w}^{\ell-1}}{\Delta t})\|_{0,\Omega}^2, \quad 1 \le \ell \le m.
\end{aligned} \tag{45}$$

Lemma 2 *Under the assumptions of Theorem 1, for $\eta, \xi > 0$ there exists a constant $C_3 > 0$, independent of h, such that it holds*

$$\begin{aligned}
&\frac{1}{2}\eta\sigma\ \|\hat{c}^m - \hat{c}_h^m\|_{0,\Omega}^2 + \frac{1}{2}\tau\eta\sigma\Delta t\ \|\hat{c}^m - \hat{c}_h^m\|_{0,\Omega}^2\ \le\\
&\frac{3}{2}\eta\xi^{-1}MC_{PF}\Delta t\ \|\hat{c}^m - \hat{c}_h^m\|_{2,h,\Omega}^2 + \frac{1}{3}\eta\xi M\Delta t\ \|\nabla(\hat{w}^m - \hat{w}_h^m)\|_{0,\Omega}^2 +\\
&C_3\Big((1+\Delta t)\ \|\hat{c}^{m-1} - \hat{c}_h^{m-1}\|_{0,\Omega}^2 + \|\hat{c}^{m-1} - I_h\hat{c}^{m-1}\|_{0,\Omega}^2 + \sum_{i=1}^{3} e_{int}^{(i)}(\hat{c}^m)\Big).
\end{aligned} \tag{46}$$

Proof We have

$$\begin{aligned}
&\eta\sigma\ \|\hat{c}^m - \hat{c}_h^m\|_{0,\Omega}^2 + \tau\eta\sigma\Delta t\ \|\hat{c}^m - \hat{c}_h^m\|_{0,\Omega}^2 \ = \\
&\eta\sigma\ (\hat{c}^m - \hat{c}_h^m, \hat{c}^m - I_h\hat{c}^m)_{0,\Omega} + \tau\eta\sigma\Delta t\ (\hat{c}^m - \hat{c}_h^m, \hat{c}^m - I_h\hat{c}^m)_{0,\Omega} + \\
&\eta\sigma\ (\hat{c}^m - \hat{c}_h^m, I_h\hat{c}^m - \hat{c}_h^m)_{0,\Omega} + \tau\eta\sigma\Delta t\ (\hat{c}^m - \hat{c}_h^m, I_h\hat{c}^m - \hat{c}_h^m)_{0,\Omega}.
\end{aligned} \tag{47}$$

By Young's inequality with $\varepsilon = 1/4$ the first two terms on the right-hand side of (47) can be estimated from above according to

$$\eta\sigma\ |(\hat{c}^m - \hat{c}_h^m, \hat{c}^m - I_h\hat{c}^m)_{0,\Omega}| \le \eta\sigma\ \|\hat{c}^m - \hat{c}_h^m\|_{0,\Omega}\ \|\hat{c}^m - I_h\hat{c}^m\|_{0,\Omega}\ \le \tag{48a}$$

$$\eta\sigma\ \|\hat{c}^m - \hat{c}_h^m\|_{0,\Omega}\ \Big(\Delta t\ \|\frac{\hat{c}^m - \hat{c}^{m-1}}{\Delta t} - I_h(\frac{\hat{c}^m - \hat{c}^{m-1}}{\Delta t})\|_{0,\Omega} + \|\hat{c}^{m-1} - I_h\hat{c}^{m-1}\|_{0,\Omega}\Big)$$

$$\le \frac{1}{4}\eta\sigma(1+\tau\Delta t)\ \|\hat{c}^m - \hat{c}_h^m\|_{0,\Omega}^2 + \eta\sigma\ \|\hat{c}^{m-1} - I_h\hat{c}^{m-1}\|_{0,\Omega}^2 + \eta\sigma\tau^{-1}\ e_{int}^{(3)}(\hat{c}^m),$$

$$\tau\eta\sigma\Delta t\ |(\hat{c}^m - \hat{c}_h^m, \hat{c}^m - I_h\hat{c}^m)_{0,\Omega}| \le \frac{1}{4}\tau\eta\sigma\Delta t\ \|\hat{c}^m - \hat{c}_h^m\|_{0,\Omega}^2 + \tau\eta\sigma\ e_{int}^{(1)}(\hat{c}^m). \tag{48b}$$

In view of (26a) and (27a), for the last two terms on the right-hand side of (47) we find

$$\begin{aligned}
&\eta\sigma\ (\hat{c}^m - \hat{c}_h^m, I_h\hat{c}^m - \hat{c}_h^m)_{0,\Omega} + \tau\eta\sigma\Delta t\ (\hat{c}^m - \hat{c}_h^m, I_h\hat{c}^m - \hat{c}_h^m)_{0,\Omega}\ = \\
&\eta\sigma\ (\hat{c}^{m-1} - \hat{c}_h^{m-1}, I_h\hat{c}^m - \hat{c}_h^m)_{0,\Omega} - \eta M\Delta t\ (\nabla(\hat{w}^m - \hat{w}_h^m), \nabla(I_h\hat{c}^m - \hat{c}_h^m))_{0,\Omega}.
\end{aligned} \tag{49}$$

The first term on the right-hand side of (49) can be estimated from above as follows:

$$\begin{aligned}
&\eta\sigma\ |(\hat{c}^{m-1} - \hat{c}_h^{m-1}, I_h\hat{c}^m - \hat{c}_h^m)_{0,\Omega}|\ \le \\
&\eta\sigma\ |(\hat{c}^{m-1} - \hat{c}_h^{m-1}, I_h\hat{c}^m - \hat{c}^m)_{0,\Omega}| + \eta\sigma\ |(\hat{c}^{m-1} - \hat{c}_h^{m-1}, \hat{c}^m - \hat{c}_h^m)_{0,\Omega}|.
\end{aligned} \tag{50}$$

As in (48a), for the first term on the right-hand side of (50) Young's inequality with $\varepsilon = 1/4$ yields

$$\begin{aligned}
&\eta\sigma\ |(\hat{c}^{m-1} - \hat{c}_h^{m-1}, \hat{c}^m - I_h\hat{c}^m)_{0,\Omega}|\ \le \\
&\frac{1}{4}\eta\sigma(1+\tau\Delta t)\ \|\hat{c}^{m-1} - \hat{c}_h^{m-1}\|_{0,\Omega}^2 + \eta\sigma\|\hat{c}^{m-1} - I_h\hat{c}^{m-1}\|_{0,\Omega}^2 + \tau^{-1}\eta\sigma\ e_{int}^{(3)}(\hat{c}^m).
\end{aligned} \tag{51}$$

For the second term on the right-hand side of (50) we obtain

$$\eta\sigma\ |(\hat{c}^{m-1} - \hat{c}_h^{m-1}, \hat{c}^m - \hat{c}_h^m)_{0,\Omega}| \le \eta\sigma\Big(\frac{1}{4}\|\hat{c}^m - \hat{c}_h^m\|_{0,\Omega}^2 + \|\hat{c}^{m-1} - \hat{c}_h^{m-1}\|_{0,\Omega}^2\Big). \tag{52}$$

For the second term on the right-hand side of (49) Young's inequality with $\varepsilon = 1/6$ and the Poincaré-Friedrichs inequality (14) yield

$$\begin{aligned}
&\eta M \Delta t\ |(\nabla(\hat{w}^m - \hat{w}_h^m), \nabla(I_h\hat{c}^m - \hat{c}_h^m))_{0,\Omega}| \ \le \\
&\eta M \Delta t\ \Big(|(\nabla(\hat{w}^m - \hat{w}_h^m), \nabla(I_h\hat{c}^m - \hat{c}^m))_{0,\Omega}| + |(\nabla(\hat{w}^m - \hat{w}_h^m), \nabla(\hat{c}^m - \hat{c}_h^m))_{0,\Omega}|\Big) \ \le \\
&\eta M \Delta t \Big(\frac{1}{3}\xi\ \|\nabla(\hat{w}^m - \hat{w}_h^m)\|_{0,\Omega}^2 + \frac{3}{2}\xi^{-1} C_{PF}\ \|\hat{c}^m - \hat{c}_h^m\|_{2,h,\Omega}^2 + \frac{3}{2}\eta\xi^{-1} M\ e_{int}^{(2)}(\hat{c}^m).
\end{aligned} \tag{53}$$

The assertion follows from (47)–(53).

Lemma 3 *Under the assumptions of Theorem 1, for $\lambda > 0$ there exist constants $C_i > 0, 4 \le i \le 6$, independent of h, such that it holds*

$$\begin{aligned}
&\frac{5}{6}\lambda M \Delta t\ \|\nabla(\hat{w}^m - \hat{w}_h^m)\|_{0,\Omega}^2 + \frac{1}{2}\lambda\sigma\gamma(1+\Delta t)\ \ \|\hat{c}^m - \hat{c}_h^m\|_{2,h,\Omega}^2 \ \le \\
&\lambda\sigma(C_4 + \tau C_5 \Delta t)\ \|\hat{c}^m - \hat{c}_h^m\|_{0,\Omega}^2 + C_6\Big((1+\Delta t)\ \|\hat{c}^{m-1} - \hat{c}_h^{m-1}\|_{0,\Omega}^2 + \\
&\|\hat{c}^{m-1} - \hat{c}_h^{m-1}\|_{2,h,\Omega}^2 + (1+h^{-4})(\|\hat{c}^{m-1} - I_h\hat{c}^{m-1}\|_{0,\Omega}^2 + \|\hat{c}^{m-1} - I_h\hat{c}^{m-1}\|_{2,h,\Omega}^2) + \\
&\|\hat{w}^{m-1} - I_h\hat{w}^{m-1}\|_{0,\Omega}^2 + (1+h^{-4})(e_{int}^{(1)}(\hat{c}^m) + e_{int}^{(3)}(\hat{c}^m)) + e_{int}^{(4)}(\hat{c}^m) + e_{int}^{(5)}(\hat{c}^m) + \\
&e_{int}^{(1)}(\hat{w}^m) + e_{int}^{(2)}(\hat{w}^m) + e_{int}^{(1)}(\hat{c}^{m-1}) + e_{int}^{(1)}(\hat{w}^{m-1})\Big).
\end{aligned} \tag{54}$$

Proof We have

$$\begin{aligned}
&\lambda M \Delta t\ \|\nabla(\hat{w}^m - \hat{w}_h^m)\|_{0,\Omega}^2 = \lambda M \Delta t\ (\nabla(\hat{w}^m - \hat{w}_h^m), \nabla(\hat{w}^m - I_h\hat{w}^m)_{0,\Omega} + \\
&\lambda M \Delta t\ (\nabla(\hat{w}^m - \hat{w}_h^m), \nabla(I_h\hat{w}^m - \hat{w}_h^m))_{0,\Omega}.
\end{aligned} \tag{55}$$

For the first term on the right-hand side of (55) Young's inequality with $\varepsilon = 1/6$ yields

$$\begin{aligned}
&\lambda M \Delta t\ |(\nabla(\hat{w}^m - \hat{w}_h^m), \nabla(\hat{w}^m - I_h\hat{w}^m))_{0,\Omega}| \ \le \\
&\frac{1}{6}\lambda M \Delta t\ \|\nabla(\hat{w}^m - \hat{w}_h^m)\|_{0,\Omega}^2 + \frac{3}{2}\lambda M\ e_{int}^{(2)}(\hat{w}^m).
\end{aligned} \tag{56}$$

Taking advantage of (26a) and (27a), for the second term on the right-hand side of (55) it follows that

$$\begin{aligned}
&\lambda M \Delta t\ (\nabla(\hat{w}^m - \hat{w}_h^m), \nabla(I_h\hat{w}^m - \hat{w}_h^m))_{0,\Omega} = \lambda\sigma\ (\hat{c}^{m-1} - \hat{c}_h^{m-1}, I_h\hat{w}^m - \hat{w}_h^m)_{0,\Omega} \\
&- \tau\lambda\sigma\Delta t\ (\hat{c}^m - \hat{c}_h^m, I_h\hat{w}^m - \hat{w}_h^m)_{0,\Omega} - \lambda\sigma\ (\hat{c}^m - \hat{c}_h^m, I_h\hat{w}^m - \hat{w}_h^m)_{0,\Omega} \ = \\
&\lambda\sigma\ (\hat{c}^{m-1} - \hat{c}_h^{m-1}, I_h\hat{w}^m - \hat{w}^m)_{0,\Omega} - \tau\lambda\sigma\Delta t\ (\hat{c}^m - \hat{c}_h^m, I_h\hat{w}^m - \hat{w}^m)_{0,\Omega} \\
&- \lambda\sigma\ (\hat{c}^m - \hat{c}_h^m, I_h\hat{w}^m - \hat{w}^m)_{0,\Omega} + \lambda\sigma\ (\hat{c}^{m-1} - \hat{c}_h^{m-1}, \hat{w}^m - \hat{w}_h^m)_{0,\Omega} \\
&- \tau\lambda\sigma\Delta t\ (\hat{c}^m - \hat{c}_h^m, \hat{w}^m - \hat{w}_h^m)_{0,\Omega} - \lambda\sigma\ (\hat{c}^m - \hat{c}_h^m, \hat{w}^m - \hat{w}_h^m)_{0,\Omega}.
\end{aligned} \tag{57}$$

The first and the third term on the right-hand side of (57) can be estimated from above as the corresponding terms in Lemma 2 using Young's inequality with $\varepsilon = 1$ and $\varepsilon = 1/6$:

$$\begin{aligned}&\lambda\sigma\ |(\hat{c}^{m-1} - \hat{c}_h^{m-1}, I_h\hat{w}^m - \hat{w}^m)_{0,\Omega}| \le \lambda\sigma(1+\Delta t)\ \|\hat{c}^{m-1} - \hat{c}_h^{m-1}\|_{0,\Omega}^2 + \\ &\frac{1}{4}\lambda\sigma\|\hat{w}^{m-1} - I_h\hat{w}^{m-1}\|_{0,\Omega}^2 + \frac{1}{4}\lambda\sigma\ e_{int}^{(3)}(\hat{w}^m),\end{aligned} \tag{58a}$$

$$\begin{aligned}&\lambda\sigma\ |(\hat{c}^m - \hat{c}_h^m, I_h\hat{w}^m - \hat{w}^m)_{0,\Omega}| \le \lambda\sigma(1+\frac{1}{6}\Delta t)\ \|\hat{c}^m - \hat{c}_h^m\|_{0,\Omega}^2 + \\ &\frac{1}{4}\lambda\sigma\ \|\hat{w}^{m-1} - I_h\hat{w}^{m-1}\|_{0,\Omega}^2 + \frac{3}{2}\lambda\sigma\ e_{int}^{(3)}(\hat{w}^m).\end{aligned} \tag{58b}$$

For the second term on the right-hand side of (57) Young's inequality with $\varepsilon = 1$ implies

$$\tau\lambda\sigma\Delta t\ |(\hat{c}^m - \hat{c}_h^m, I_h\hat{w}^m - \hat{w}^m)_{0,\Omega}| \le \tau\lambda\sigma\Big(\Delta t\ \|\hat{c}^m - \hat{c}_h^m\|_{0,\Omega}^2 + \frac{1}{4}\ e_{int}^{(1)}(\hat{w}^m)\Big). \tag{59}$$

For the last three terms on the right-hand side of (57) we obtain

$$\begin{aligned}&\lambda\sigma\ (\hat{c}^{m-1} - \hat{c}_h^{m-1}, \hat{w}^m - \hat{w}_h^m)_{0,\Omega} - \tau\lambda\sigma\Delta t\ (\hat{c}^m - \hat{c}_h^m, \hat{w}^m - \hat{w}_h^m)_{0,\Omega} \\ &-\lambda\sigma\ (\hat{c}^m - \hat{c}_h^m, \hat{w}^m - \hat{w}_h^m)_{0,\Omega} = \lambda\sigma\ (\hat{c}^{m-1} - I_h\hat{c}^{m-1}, \hat{w}^m - \hat{w}_h^m)_{0,\Omega} \\ &-\tau\lambda\sigma\Delta t\ (\hat{c}^m - I_h\hat{c}^m, \hat{w}^m - \hat{w}_h^m)_{0,\Omega} - \lambda\sigma\ (\hat{c}^m - I_h\hat{c}^m, \hat{w}^m - \hat{w}_h^m)_{0,\Omega} \\ &+\lambda\sigma\ (I_h\hat{c}^{m-1} - \hat{c}_h^{m-1}, \hat{w}^m - \hat{w}_h^m)_{0,\Omega} - \tau\lambda\sigma\Delta t\ (I_h\hat{c}^m - \hat{c}_h^m, \hat{w}^m - \hat{w}_h^m)_{0,\Omega} \\ &-\lambda\sigma\ (I_h\hat{c}^m - \hat{c}_h^m, \hat{w}^m - \hat{w}_h^m)_{0,\Omega}.\end{aligned} \tag{60}$$

Using Corollary (1) and Young's inequality with $\varepsilon = 1/18$ and $\varepsilon = 1/2$, the first term on the right-hand side of (60) can be estimated from above as follows:

$$\begin{aligned}&\lambda\sigma\ |(\hat{c}^{m-1} - I_h\hat{c}^{m-1}, \hat{w}^m - \hat{w}_h^m)_{0,\Omega}|\ \le \\ &\lambda\sigma\ \|\hat{c}^{m-1} - I_h\hat{c}^{m-1}\|_{0,\Omega}\ \Big(\|\hat{w}^m - I_h\hat{w}^m\|_{0,\Omega} + \|I_h\hat{w}^m - \hat{w}_h^m\|_{0,\Omega}\Big)\ \le \\ &\lambda\sigma\ \|\hat{c}^{m-1} - I_h\hat{c}^{m-1}\|_{0,\Omega}\ \Big(2\ \|\hat{w}^m - I_h\hat{w}^m\|_{0,\Omega} + C_2h^{-2}\ \|\hat{c}^m - \hat{c}_h^m\|_{2,h,\Omega}\Big)\ \le \\ &\lambda\sigma\ \|\hat{c}^{m-1} - I_h\hat{c}^{m-1}\|_{0,\Omega}\ \Big(2\Delta t\ \|\frac{\hat{w}^m - \hat{w}^{m-1}}{\Delta t} - I_h(\frac{\hat{w}^m - \hat{w}^{m-1}}{\Delta t})\|_{0,\Omega} + \\ &2\ \|\hat{w}^{m-1} - I_h\hat{w}^{m-1}\|_{0,\Omega} + C_2h^{-2}\ \|\hat{c}^m - \hat{c}_h^m\|_{2,h,\Omega}\Big)\ \le \\ &\frac{1}{18}\lambda\sigma\ \|\hat{c}^m - \hat{c}_h^m\|_{2,h,\Omega}^2 + \frac{9}{2}\lambda\sigma C_2^2h^{-4}\ \|\hat{c}^{m-1} - I_h\hat{c}^{m-1}\|_{0,\Omega}^2 + \\ &\lambda\sigma\Big(\|\hat{c}^{m-1} - I_h\hat{c}^{m-1}\|_{0,\Omega}^2 + \|\hat{w}^{m-1} - I_h\hat{w}^{m-1}\|_{0,\Omega}^2 + e_{int}^{(1)}(\hat{c}^{m-1}) + e_{int}^{(3)}(\hat{w}^m)\Big).\end{aligned} \tag{61}$$

Likewise, by Young's inequality with $\varepsilon = 1/14$, $\varepsilon = 1/18$, and $\varepsilon = 1/2$ and observing $\Delta t \le T$, for the third term on the right-hand side of (60) we get

$$\begin{aligned}
&\lambda\sigma\ |(\hat{c}^m - I_h\hat{c}^m, \hat{w}^m - \hat{w}_h^m)_{0,\Omega}|\ \le \qquad (62)\\
&\lambda\sigma\ \Big(\Delta t\ \|\frac{\hat{c}^m - \hat{c}^{m-1}}{\Delta t} - I_h(\frac{\hat{c}^m - \hat{c}^{m-1}}{\Delta t})\|_{0,\Omega} + \|\hat{c}^{m-1} - I_h\hat{c}^{m-1}\|_{0,\Omega}\Big)\cdot\\
&\Big(2\Delta t\ \|\frac{\hat{w}^m - \hat{w}^{m-1}}{\Delta t} - I_h(\frac{\hat{w}^m - \hat{w}^{m-1}}{\Delta t})\|_{0,\Omega} + 2\ \|\hat{w}^{m-1} - I_h\hat{w}^{m-1}\|_{0,\Omega}\\
&+ C_2h^{-2}\ \|\hat{c}^m - \hat{c}_h^m\|_{2,h,\Omega}\Big)\ \le\\
&\frac{1}{14}\lambda\sigma\,\Delta t\ \|\hat{c}^m - \hat{c}_h^m\|^2_{2,h,\Omega} + \lambda\sigma\Big((\frac{7}{2}C_2^2h^{-4} + T)\ e^{(3)}_{int}(\hat{c}^m) + T\ e^{(3)}_{int}(\hat{w}^m)\Big) +\\
&\lambda\sigma\Big(e^{(3)}_{int}(\hat{c}^m) + e^{(1)}_{int}(\hat{w}^m) + e^{(3)}_{int}(\hat{w}^m) + e^{(1)}_{int}(\hat{c}^{m-1})\Big) +\\
&\frac{1}{18}\lambda\sigma C_3\ \|\hat{c}^m - \hat{c}_h^m\|^2_{2,h,\Omega} + \frac{9}{2}\lambda\sigma C_2^2h^{-4}\ \|\hat{c}^{m-1} - I_h\hat{c}^{m-1}\|^2_{0,\Omega} +\\
&\lambda\sigma\ \|\hat{c}^{m-1} - I_h\hat{c}^{m-1}\|^2_{0,\Omega} + \lambda\sigma\,\|\hat{w}^{m-1} - I_h\hat{w}^{m-1}\|^2_{0,\Omega}.
\end{aligned}$$

Finally, for the second term on the right-hand side of (60) Young's inequality with $\varepsilon = 1/14$ and $\varepsilon = 1/2$ gives

$$\begin{aligned}
&\tau\lambda\sigma\,\Delta t\ |(\hat{c}^m - I_h\hat{c}^m, \hat{w}^m - \hat{w}_h^m)_{0,\Omega}|\ \le \qquad (63)\\
&\tau\lambda\sigma\,\Delta t\ \|\hat{c}^m - I_h\hat{c}^m\|_{0,\Omega}\Big(\|\hat{w}^m - I_h\hat{w}^m\|_{0,\Omega} + \|I_h\hat{w}^m - \hat{w}_h^m\|_{0,\Omega}\Big)\ \le\\
&\tau\lambda\sigma\,\Delta t\ \|\hat{c}^m - I_h\hat{c}^m\|_{0,\Omega}\Big(2\ \|\hat{w}^m - I_h\hat{w}^m\|_{0,\Omega} + C_2\ h^{-2}\ \|\hat{c}^m - \hat{c}_h^m\|_{2,h,\Omega}\Big)\ \le\\
&\frac{1}{14}\tau\lambda\sigma\gamma\,\Delta t\ \|\hat{c}^m - \hat{c}_h^m\|^2_{2,h,\Omega} + \frac{7}{2}\tau\lambda\sigma\Big(1 + \gamma^{-1}C_2^2h^{-4}\Big)\ e^{(1)}_{int}(\hat{c}^m) + \tau\lambda\sigma\ e^{(1)}_{int}(\hat{w}^m).
\end{aligned}$$

Using (26b) and (27b), for the first of the last three terms on the right-hand side of (60) we obtain

$$\begin{aligned}
&\lambda\sigma\ (I_h\hat{c}^{m-1} - \hat{c}_h^{m-1}, \hat{w}^m - \hat{w}_h^m)_{0,\Omega}\ = \qquad (64a)\\
&\lambda\sigma\ \Big(a_h^{DG}(\hat{c}^m - \hat{c}_h^m, I_h\hat{c}^{m-1} - \hat{c}_h^{m-1}) + (\hat{g}(\hat{c}^m) - \hat{g}(\hat{c}_h^m), I_h\hat{c}^{m-1} - \hat{c}_h^{m-1})_{0,\Omega}\Big)\ =\\
&\lambda\sigma\ \Big(a_h^{DG}(\hat{c}^m - \hat{c}_h^m, I_h\hat{c}^{m-1} - \hat{c}^{m-1}) + a_h^{DG}(\hat{c}^m - \hat{c}_h^m, \hat{c}^{m-1} - \hat{c}_h^{m-1})\Big) +\\
&\lambda\sigma\ \Big((\hat{g}(\hat{c}^m) - \hat{g}(\hat{c}_h^m), I_h\hat{c}^{m-1} - \hat{c}^{m-1}) + (\hat{g}(\hat{c}^m) - \hat{g}(\hat{c}_h^m), \hat{c}^{m-1} - \hat{c}_h^{m-1})_{0,\Omega}\Big).
\end{aligned}$$

Similarly, for the second term we get

$$\begin{aligned}
&\tau\lambda\sigma\,\Delta t\ (I_h\hat{c}^m - \hat{c}_h^m, \hat{w}^m - \hat{w}_h^m)_{0,\Omega}\ = \qquad\qquad (64b)\\
&\tau\lambda\sigma\,\Delta t\ \Big(a_h^{DG}(\hat{c}^m - \hat{c}_h^m, I_h\hat{c}^m - \hat{c}^m) + a_h^{DG}(\hat{c}^m - \hat{c}_h^m, \hat{c}^m - \hat{c}_h^m)\Big) +\\
&\tau\lambda\sigma\,\Delta t\ \Big((\hat{g}(\hat{c}^m) - \hat{g}(\hat{c}_h^m), I_h\hat{c}^m - \hat{c}^m) + (\hat{g}(\hat{c}^m) - \hat{g}(\hat{c}_h^m), \hat{c}^m - \hat{c}_h^m)_{0,\Omega}\Big),
\end{aligned}$$

whereas for the third term we obtain

$$\begin{aligned}
&-\lambda\sigma\ (I_h\hat{c}^m - \hat{c}_h^m, \hat{w}^m - \hat{w}_h^m)_{0,\Omega}\ = \qquad\qquad (64c)\\
&-\lambda\sigma\ \Big(\tilde{a}_h^{DG}(\hat{c}^m - \hat{c}_h^m, I_h\hat{c}^m - \hat{c}^m) + \tilde{a}_h^{DG}(\hat{c}^m - \hat{c}_h^m, \hat{c}^m - \hat{c}_h^m)\Big)\\
&-\lambda\sigma\ \Big((\hat{g}(\hat{c}^m) - \hat{g}(\hat{c}_h^m), I_h\hat{c}^m - \hat{c}^m)_{0,\Omega} + (\hat{g}(\hat{c}^m) - \hat{g}(\hat{c}_h^m), \hat{c}^m - \hat{c}_h^m)_{0,\Omega}\Big).
\end{aligned}$$

Taking advantage of (16) and (30) from Lemma 1 and using Young's inequality with $\varepsilon = 1/18$, for (64a) we can establish the upper bound

$$\begin{aligned}
&\lambda\sigma\ |(I_h\hat{c}^{m-1} - \hat{c}_h^{m-1}, \hat{w}^m - \hat{w}_h^m)_{0,\Omega}| \le \frac{2}{9}\lambda\sigma\ \|\hat{c}^m - \hat{c}_h^m\|_{2,h,\Omega}^2 + \qquad (65)\\
&\frac{9}{2}\lambda\sigma(\Gamma^2 + C_1^2)\ \|\hat{c}^{m-1} - \hat{c}_h^{m-1}\|_{2,h,\Omega}^2 + \frac{9}{2}\lambda\sigma(\Gamma^2 + C_1^2)\ \|\hat{c}^{m-1} - I_h\hat{c}_h^{m-1}\|_{2,h,\Omega}^2.
\end{aligned}$$

Similarly, for (64b) Gårding's inequality (15) and Young's inequality with $\varepsilon = 1/14$ yield

$$\begin{aligned}
&-\tau\lambda\sigma\,\Delta t\ (I_h\hat{c}^m - \hat{c}_h^m, \hat{w}^m - \hat{w}_h^m)_{0,\Omega} \le -\tau\lambda\sigma\gamma\,\Delta t\ \|\hat{c}^m - \hat{c}_h^m\|_{2,h,\Omega}^2 + \qquad (66)\\
&\tau\lambda\sigma\beta\,\Delta t\ \|\hat{c}^m - \hat{c}_h^m\|_{0,\Omega}^2 + \frac{3}{14}\tau\lambda\sigma\,\Delta t\ \|\hat{c}^m - \hat{c}_h^m\|_{2,h,\Omega}^2 +\\
&\frac{7}{2}\tau\lambda\sigma\gamma^{-1}\Big(C_1^2\,\Delta t\ \|\hat{c}^m - \hat{c}_h^m\|_{0,\Omega}^2 + \Big(C_1^2\ e_{int}^{(1)}(\hat{c}^m) + \Gamma^2\ e_{int}^{(4)}(\hat{c}^m)\Big)\Big).
\end{aligned}$$

Finally, for (64c) another application of Gårding's inequality (15) and Young's inequality with $\varepsilon = 1/14$ and $\varepsilon = 1/18$ we obtain

$$\begin{aligned}
&-\lambda\sigma\ (I_h\hat{c}^m - \hat{c}_h^m, \hat{w}^m - \hat{w}_h^m)_{0,\Omega} \le \lambda\sigma\Big(-\gamma\ \|\hat{c}^m - \hat{c}_h^m\|_{2,h,\Omega}^2 + \beta\ \|\hat{c}^m - \hat{c}_h^m\|_{0,\Omega}^2\Big) +\\
&\frac{3}{18}\lambda\sigma\ \|\hat{c}^m - \hat{c}_h^m\|_{2,h,\Omega}^2 + \frac{1}{7}\lambda\sigma\,\Delta t\ \|\hat{c}^m - \hat{c}_h^m\|_{2,h,\Omega}^2 + \frac{9}{2}\lambda\sigma C_1^2\ \|\hat{c}^m - \hat{c}_h^m\|_{0,\Omega}^2 +
\end{aligned}$$

$$\frac{9}{2}\lambda\sigma\Big(C_1^2\ \||\|\hat{c}^{m-1}-I_h\hat{c}^{m-1}\|_{0,\Omega}^2+\Gamma^2\ \|\hat{c}^{m-1}-I_h\hat{c}^{m-1}\|_{2,h,\Omega}^2\Big)+$$
$$\frac{7}{2}\lambda\sigma\Big(C_1^2\ e_{int}^{(3)}(\hat{c}^m)+\Gamma^2\ e_{int}^{(5)}(\hat{c}^m)\Big). \tag{67}$$

The assertion follows from (55)–(67).

Proposition 1 *Under the assumptions of Theorem 1 there exists a constant $C_7>0$, independent of h, such that it holds*

$$\|\hat{c}^m-\hat{c}_h^m\|_{0,\Omega}^2+\tau\Delta t\ \|\hat{c}^m-\hat{c}_h^m\|_{0,\Omega}^2+\frac{1}{2}\lambda\sigma\gamma\ \Big(\|\hat{c}^m-\hat{c}_h^m\|_{2,h,\Omega}^2 \tag{68}$$
$$+\ \Delta t\ \|\hat{c}^m-\hat{c}_h^m\|_{2,h,\Omega}^2\Big)+M\Delta t\ \|\nabla(\hat{w}^m-\hat{w}_h^m)\|_{0,\Omega}^2\ \le$$
$$C_7\Big(\|\hat{c}^{m-1}-\hat{c}_h^{m-1}\|_{0,\Omega}^2+\|\hat{c}^{m-1}-\hat{c}_h^{m-1}\|_{2,h,\Omega}^2+h^{-4}\ \|\hat{c}^{m-1}-I_h\hat{c}^{m-1}\|_{0,\Omega}^2+$$
$$\|\hat{c}^{m-1}-I_h\hat{c}^{m-1}\|_{2,h,\Omega}^2+\|\hat{w}^{m-1}-I_h\hat{w}^{m-1}\|_{0,\Omega}^2+h^{-4}(e_{int}^{(1)}(\hat{c}^m)+e_{int}^{(3)}(\hat{c}^m))+$$
$$\sum_{i=4}^{5}e_{int}^{(i)}(\hat{c}^m)+\sum_{i=1}^{3}e_{int}^{(i)}(\hat{w}^m)+e_{int}^{(1)}(\hat{c}^{m-1})+e_{int}^{(1)}(\hat{w}^{m-1})\Big).$$

Proof The estimates (46) from Lemma 2 and (54) from Lemma 3 imply the existence of a constant $D_{10}>0$, independent of h, such that

$$\sigma(\eta-\frac{3}{2}\lambda C_6)\ \|\hat{c}^m-\hat{c}_h^m\|_{0,\Omega}^2+\tau\sigma(\eta-\frac{1}{2}\lambda C_7)\Delta t\ \|\hat{c}^m-\hat{c}_h^m\|_{0,\Omega}^2+ \tag{69}$$
$$\frac{1}{2}\lambda\sigma\gamma\ \|\hat{c}^m-\hat{c}_h^m\|_{2,h,\Omega}^2+\frac{1}{2}\lambda\sigma(\tau\gamma-\frac{3}{2}\eta\xi^{-1}C_{PF})\Delta t\|\hat{c}^m-\hat{c}_h^m\|_{2,h,\Omega}^2+$$
$$M(\frac{5}{6}\lambda-\frac{1}{3}\eta\xi)\Delta t\ \|\nabla(\hat{w}^m-\hat{w}_h^m)\|_{0,\Omega}^2\le D_{10}\Big((1+\Delta t)\ \|\hat{c}^{m-1}-\hat{c}_h^{m-1}\|_{0,\Omega}^2+$$
$$+\ \|\hat{c}^{m-1}-\hat{c}_h^{m-1}\|_{2,h,\Omega}^2+(1+h^{-4})(\|\hat{c}^{m-1}-I_h\hat{c}^{m-1}\|_{0,\Omega}^2+\|\hat{c}^{m-1}-I_h\hat{c}^{m-1}\|_{2,h,\Omega}^2)$$
$$+\ \|\hat{w}^{m-1}-I_h\hat{w}^{m-1}\|_{0,\Omega}^2+(1+h^{-4})(e_{int}^{(1)}(\hat{c}^m)+e_{int}^{(3)}(\hat{c}^m))+e_{int}^{(2)}(\hat{c}^m)+$$
$$\sum_{i=4}^{5}e_{int}^{(i)}(\hat{c}^m)+\sum_{i=1}^{3}e_{int}^{(i)}(\hat{w}^m)+e_{int}^{(1)}(\hat{c}^{m-1})+e_{int}^{(1)}(\hat{w}^{m-1})\Big).$$

We choose $\frac{6}{5}<\lambda<2$ and $\eta>0$ such that

$$\eta-\max(\frac{3}{2}C_6,\frac{1}{2}C_7)\lambda\ge\sigma^{-1}.$$

Then, we choose $\xi>0$ by means of

$$\frac{5}{6}\lambda-\frac{1}{3}\eta\xi\ge 1\quad\Longleftrightarrow\quad\xi\le\frac{5\lambda-6}{2\eta}.$$

Finally, we choose $\tau > 0$ according to

$$\gamma\tau - \frac{3}{2}\eta\xi^{-1}C_{PF} \geq \gamma \quad \Longleftrightarrow \quad \tau \geq \frac{2\xi\gamma + 3\eta C_{PF}}{2\xi\gamma}.$$

For this choice of λ, η, ξ, and τ, the assertion follows from (69) observing that $e^{(1)}_{int}(\hat{c}^m) \leq e^{(4)}_{int}(\hat{c}^m)$, $e^{(2)}_{int}(\hat{c}^m) \leq C^2_{PF} e^{(4)}_{int}(\hat{c}^m)$, $e^{(3)}_{int}(\hat{c}^m) \leq e^{(5)}_{int}(\hat{c}^m)$, and $\|v\|_{0,\Omega} \leq \|v\|_{2,h,\Omega}$, $v \in V_h^{(p)} + Z$.

Proposition 2 *Under the assumptions of Theorem 1 there exists a constant $C_8 > 0$, independent of h, such that it holds*

$$\|\hat{c}^m - \hat{c}^m_h\|^2_{2,h,\Omega} + \Delta t \sum_{\ell=1}^{m} \|\hat{c}^\ell - \hat{c}^\ell_h\|^2_{2,h,\Omega} + \Delta t \sum_{\ell=1}^{m} \|\nabla(\hat{w}^\ell - \hat{w}^\ell_h)\|^2_{0,\Omega} \leq \tag{70}$$

$$C_8\Big(h^{-4}\sum_{\ell=1}^{m}(e^{(1)}_{int}(\hat{c}^\ell) + e^{(3)}_{int}(\hat{c}^\ell)) + \sum_{i=4}^{5}\sum_{\ell=1}^{m} e^{(i)}_{int}(\hat{c}^\ell) + \sum_{i=1}^{3}\sum_{\ell=1}^{m} e^{(i)}_{int}(\hat{w}^\ell) +$$

$$h^{-4}\,\|c_0 - I_h c_0\|^2_{0,\Omega} + (1 + h^{-4})\,\|c_0 - I_h c_0\|^2_{2,h,\Omega} + \|w_0 - I_h w_0\|^2_{0,\Omega}\Big),$$

where w_0 by (9b) with $c = c_0$ and $w = w_0$.

Proof The proof is by induction on m. For $m = 1$ the assertion follows from (68) taking into account that $\hat{c}^0 = c_0$ and $\hat{w}^0 = w_0$. Let us assume that (70) holds true for $m - 1$. Observing

$$\|\hat{c}^{m-1} - I_h\hat{c}^{m-1}\|_{0,\Omega} \leq \Delta t \sum_{\ell=1}^{m-1} \|\frac{\hat{c}^\ell - \hat{c}^{\ell-1}}{\Delta t} - I_h \frac{\hat{c}^\ell - \hat{c}^{\ell-1}}{\Delta t}\|_{0,\Omega} + \|c_0 - I_h c_0\|_{0,\Omega}$$

and the same for $\|\hat{c}^{m-1} - I_h\hat{c}^{m-1}\|_{2,h,\Omega}$ and $\|\hat{w}^{m-1} - I_h\hat{w}^{m-1}\|_{0,\Omega}$, it follows from (68) that (70) is satisfied for m as well.

Proof (*Theorem 1*) We have $t_m \to t$ as $\Delta t \to 0$. Due to the regularity assumptions (18a), (18b) for $\Delta t \to 0$ the left-hand side of (70) converges to

$$\|(\hat{c} - \hat{c}_h)(\cdot, t)\|^2_{2,h,\Omega} + \int_0^t \|\hat{c} - \hat{c}_h\|^2_{2,h,\Omega}\, ds + \int_0^t \|\nabla(\hat{w} - \hat{w}_h)\|^2_{0,\Omega}\, ds.$$

On the other hand, for the sum of the interpolation errors (45) it holds

$$\sum_{\ell=1}^{m} e^{(1)}_{int}(\hat{z}^\ell) \to \int_0^t \|\hat{z} - I_h\hat{z}\|^2_{0,\Omega}\, ds, \quad \hat{z} = \hat{c} \text{ and } \hat{z} = \hat{w},$$

$$\sum_{\ell=1}^{m} e_{int}^{(3)}(\hat{z}^{\ell}) \to \int\limits_0^t \|\frac{\partial \hat{z}}{\partial s} - I_h \frac{\partial \hat{z}}{\partial s}\|_{0,\Omega}^2 \, ds, \quad \hat{z} = \hat{c} \text{ and } \hat{z} = \hat{w},$$

$$\sum_{\ell=1}^{m} e_{int}^{(4)}(\hat{c}^{\ell}) \to \int\limits_0^t \|\hat{c} - I_h \hat{c}\|_{2,h,\Omega}^2 \, ds,$$

$$\sum_{\ell=1}^{m} e_{int}^{(5)}(\hat{c}^{\ell}) \to \int\limits_0^t \|\frac{\partial \hat{c}}{\partial s} - I_h \frac{\partial \hat{c}}{\partial s}\|_{2,h,\Omega}^2 \, ds,$$

$$\sum_{\ell=1}^{m} e_{int}^{(2)}(\hat{w}^{\ell}) \to \int\limits_0^t \|\nabla(\hat{w} - I_h \hat{w})\|_{0,\Omega}^2 \, ds.$$

Hence, taking (20),(21) into account, (22) holds true for $c = \hat{c}$. Finally, backtransformation according to (23) allows to conclude. □

6 Discretization in Time by Singly Diagonally Implicit Runge-Kutta Methods

For the discretization in time of the C^{0}IPDG approximation (17a)–(17c) we use (s, q) Singly Diagonally Implicit Runge-Kutta (SDIRK) methods of stage s and order q with respect to a partitioning of the time interval $[0, T]$ into subintervals $[t_{m-1}, t_m]$ of length $\tau_m := t_m - t_{m-1}$, $1 \le m \le M$ (cf., e.g., [1, 7, 19]). In particular, for polynomial order $p = 2$ of the C^{0}IPDG approximation we use a (2, 2) SDIRK method with coefficients given by the Butcher scheme in Table 1. If the polynomial degree is $p = 3$, we use a (3, 3) SDIRK method with Butcher scheme given by Table 2, and for $p = 4$ we use a (3, 4) SDIRK method with Butcher scheme given by Table 3.

The fully discrete approximation represents a parameter dependent nonlinear algebraic system with the time-step size as a parameter which is solved by a predictor-corrector continuation strategy with constant continuation as a predictor and Newton's method as a corrector [9, 20]. The predictor-corrector continuation strategy features an adaptive choice of the continuation parameter. For details we refer to [2].

Table 1 Butcher scheme of a 2-stage SDIRK method of order $q = 2$

κ	κ	0
1	$1-\kappa$	κ
	$1-\kappa$	κ

$\kappa = 1 \pm \frac{1}{2}\sqrt{2}$

Table 2 Butcher scheme of a 3-stage SDIRK method of order $q = 3$ (cf. [1])

$$\begin{array}{c|ccc} \alpha & \alpha & 0 & 0 \\ \frac{1+\alpha}{2} & \frac{1-\alpha}{2} & \alpha & 0 \\ 1 & b_0 & b_1 & \alpha \\ \hline & b_0 & b_1 & \alpha \end{array} \qquad \begin{aligned} &\alpha \approx 0.44 \text{ is the root of } p(x) = x^3 - 3x^2 + \frac{3}{2}x - \frac{1}{6}, \\ &b_0 = -\frac{6\alpha^2 - 16\alpha + 1}{4}, \\ &b_1 = \frac{6\alpha^2 - 20\alpha + 5}{4}. \end{aligned}$$

Table 3 Butcher scheme of a 3-stage SDIRK method of order $q = 4$

$$\begin{array}{c|ccc} (1+\kappa)/2 & (1+\kappa)/2 & 0 & 0 \\ \frac{1}{2} & -\kappa/2 & (1+\kappa)/2 & 0 \\ (1-\kappa)/2 & 1+\kappa & -(1+2\kappa) & (1+\kappa)/2 \\ \hline & 1/(6\kappa^2) & 1 - 1/(3\kappa^2) & 1/(6\kappa^2) \end{array} \qquad \kappa = 2\cos(\pi/18)/\sqrt{3}$$

7 Numerical Results

We consider the initial-boundary value problem (2a)–(2c) in $Q := \Omega \times (0, T]$ with $\Omega := (0, L)^2$, $L := 1.0 \times 10^{-4}$ m, and $T := 1.0 \times 10^{-1}$ s. The physical parameters β, κ, σ, and a_0, a_2, h_0, M are given in Table 4 in their physical units. We use the reference quantities

$$L_{\text{ref}} := 1.0 \times 10^{-5}\,\text{m}, \quad T_{\text{ref}} := 1.0 \times 10^{-2}\,\text{s}, \quad \sigma_{\text{ref}} := 1.0\,\text{Jm}^{-2} \tag{71}$$

and scale all independent variables and parameters to dimensionless form. Hence, the scaled domain and the scaled time interval become $\Omega = (0, 10)^2$ and $[0, 10]$. The values of the parameters in dimensionless form are also listed in Table 4. The initial concentration c_0 has been chosen as a smooth function $c_0 \in C^\infty(\Omega)$ satisfying the compatibility conditions (5).

Table 4 Physical parameters in the sixth order Cahn-Hilliard equation

Symbol	Value	Unit	Dimensionless v alue
σ	1.0	Jm^{-2}	1.0
β	5.0	Jm^{-2}	5.0
h_0	5.0×10^{-1}	1	5.0×10^{-1}
M	1.0×10^{-13}	m^2s^{-1}	1.0×10^{-3}
κ	1.0×10^{-25}	Jm^2	1.0×10^{-1}
a_0	-4.0×10^{-12}	J	-4.0
a_2	1.0×10^{-12}	J	1.0

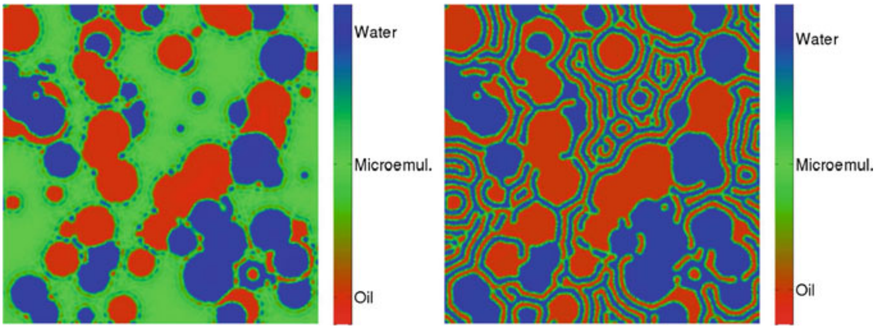


Fig. 1 Formation of oil-in-water and water-in-oil droplets at time instants $t = 0.60$ (left) and $t = 3.86$ (right). C^{0}IPDG approximation with $p = 2$ on a 128×128 grid and 2-stage SDIRK with $q = 2$ (from [2])

Figure 1 shows a visualization of the microemulsification process obtained by the numerical solution of the sixth order Cahn-Hilliard equation using a C^{0}IPDG approximation with $p = 2$ and penalization parameter $\alpha = 25.0$ and a 2-stage SDIRK with $q = 2$ at time instants $t = 0.60$ (left) and $t = 3.86$ (right). The pure water phase ($c = 1$) is depicted in dark blue, the pure oil phase ($c = -1$) in dark red, and the microemulsion phase ($c = 0$) in light green. In Fig. 1 (right), the formation of oil-in-water and water-in-oil droplets is clearly visible.

The underlying finite element mesh is a geometrically conforming, simplicial triangulation $\mathscr{T}_h(\Omega)$ of mesh size h. For $h = 1/24$, $1/48$ and at $t = 2.5$ we have computed the convergence rates in the mesh dependent C^{0}IPDG-norm. Obviously, the domain Ω does not have a boundary Γ of class C^{r+1}, $r \geq 5$, and hence, we cannot expect quasi-optimal convergence rates. Therefore, we also computed the convergence rates for a patch Ω of elements around the midpoint m_Ω of Ω given by

$$\omega := \bigcup \{K \in \mathscr{T}_{2h}(\Omega) \mid m_\Omega \in \mathscr{N}_{2h}(K)\},$$

where $\mathscr{N}_{2h}(K)$ is the set of nodal points in K. The convergence rates are as follows

$$\begin{aligned} \mathrm{err}_\omega(t) &:= \log_2 \frac{\|u_h(\cdot,t) - u_{2h}(\cdot,t)\|_{2,h/2,\omega}}{\|u_{h/2}(\cdot,t) - u_h(\cdot,t)\|_{2,h/2,\omega}}, \\ \mathrm{err}_\Omega(t) &:= \log_2 \frac{\|u_h(\cdot,t) - u_{2h}(\cdot,t)\|_{2,h/2,\Omega}}{\|u_{h/2}(\cdot,t) - u_h(\cdot,t)\|_{2,h/2,\Omega}}. \end{aligned}$$

In each case the time-step size has been chosen sufficiently small so that the error due to discretization in time do not affect the error due to spatial discretization. The convergence rates are shown in Table 5.

For domains Ω with boundary Γ of class C^{r+1}, $r \geq 5$, the quasi-optimal convergence rates are 1.0 for $p = 2$, 2.0 for $p = 3$, and 3.0 for $p = 4$ (cf. Theorem 5). We see that we get almost quasi-optimal convergence rates on the patch ω in the

Table 5 Patchwise and global convergence rates for the semidiscrete C^{0}IPDG approximation with $p = 2, 3, 4$

	$p = 2$		$p = 3$		$p = 4$	
	$\text{err}_\omega(2.5)$	$\text{err}_\Omega(2.5)$	$\text{err}_\omega(2.5)$	$\text{err}_\Omega(2.5)$	$\text{err}_\omega(2.5)$	$\text{err}_\Omega(2.5)$
$h = 1/24$	1.06	0.66	1.83	1.68	2.83	2.56
$h = 1/48$	1.02	0.91	1.91	1.79	2.90	2.67

$\| \cdot \|_{2,2h,\omega}$-norm, but as expected not quite as good convergence rates on the entire domain Ω in the $\| \cdot \|_{2,2h,\Omega}$-norm.

Acknowledgements Ronald H. W. Hoppe acknowledges support by the NSF grants DMS-1115658, DMS-1216857, DMS-1520886 and by the German National Science Foundation DFG within the Priority Program SPP 1506.

References

1. Alexander R (1977) Diagonally implicit Runge-Kutta methods for stiff o.d.e'.s. SIAM J Numer Anal 14(6):1006–1021
2. Boyarkin O, Hoppe RHW, Linsenmann C (2015) High order approximations in space and time of a sixth order Cahn-Hilliard equation. Russ J Numer Anal Math Model 30(6):313–328
3. Brenner SC, Gudi T, Sung L-Y (2010) An a posteriori error estimator for a quadratic C^0-interior penalty method for the biharmonic problem. IMA J Numer Anal 30(3):777–798
4. Brenner SC, Scott LR (2008) The mathematical theory of finite element methods, 3rd edn. Springer, New York
5. Brenner SC, Sung L-Y (2005) c^0 interior penalty methods for fourth order elliptic boundary value problems on polygonal domains. J Sci Comput 22(23):83–118
6. Brenner SC, Wang K, Zhao J (2004) Poincaré-Friedrichs inequalities for piecewise H^2 functions. Numer Funct Anal Optim 25(5–6):463–478
7. Butcher JC (2008) Numerical methods for ordinary differential equations, 2nd edn. Wiley, Chichester
8. Ciarlet PG (2002) The finite element method for elliptic problems. SIAM, Philadelphia, PA
9. Deuflhard P (2004) Newton methods for nonlinear problems: affine invariance and adaptive algorithms. Springer, Berlin
10. Engel G, Garikipati K, Hughes TJR, Larson MG, Mazzei L, Taylor RL (2002) Continuous/discontinuous finite element approximations of fourth-order elliptic problems in structural and continuum mechanics with applications to thin beams and plates, and strain gradient elasticity. Comput Methods Appl Mech Eng 191(34):3669–3750
11. Fraunholz T, Hoppe RHW, Peter M (2015) Convergence analysis of an adaptive interior penalty discontinuous Galerkin method for the biharmonic problem. J Numer Math 23(4):317–330
12. Georgoulis EH, Houston P (2009) Discontinuous Galerkin methods for the biharmonic problem. IMA J Numer Anal 29(3):573–594
13. Georgoulis EH, Houston P, Virtanen J (2011) An a posteriori error indicator for discontinuous Galerkin approximations of fourth order elliptic problems. IMA J Numer Anal 31(1):281–298
14. Ghosh PK, Murthy RS (2006) Microemulsions: a potential drug delivery system. Curr Drug Deliv 3(2):167–180
15. Gompper G, Goos J (1994) Fluctuating interfaces in microemulsion and sponge phases. Phys. Rev. E 50(2):1325–1335

16. Gompper G, Kraus M (1993) Ginzburg-Landau theory of ternary amphiphilic systems. I. Gaussian interface fluctuations. Phys Rev E 47(6):4289–4300
17. Gompper G, Kraus M (1993) Ginzburg-Landau theory of ternary amphiphilic systems. II. Monte Carlo simulations. Phys Rev E 47(6):4301–4312
18. Gompper G, Zschocke S (1992) Ginzburg-Landau theory of oil-water-surfactant mixtures. Phys Rev A 46(8):4836–4851
19. Hairer E, Wanner G (1996) Solving ordinary differential equations. II: stiff and differential-algebraic problems, 2nd edn. Springer, Berlin
20. Hoppe RHW, Linsenmann C (2012) An adaptive Newton continuation strategy for the fully implicit finite element immersed boundary method. J Comput Phys 231(14):4676–4693
21. Jha SK, Karki R, Venkatesh DP, Geethalakshami A (2011) Formulation development and characterization of microemulsion drug delivery systems containing antiulcer drug. Int J Drug Dev Res 3(4):336–343
22. Mehta SK, Kaur G (2011) Microemulsions: thermodynamics and dynamic properties. In: Tadashi M (ed) Thermodynamics. InTech, pp 381–406. http://www.intechopen.com/books/thermodynamics/microemulsions-thermodynamic-and-dynamic-properties
23. Moulik SP, Rakshit AK (2006) Physiochemistry and applications of micro-emulsions. J Surf Sci Tech 22(3–4):159–186
24. Pawlow I, Zajaczkowski WM (2011) A sixth order Cahn-Hilliard type equation arising in oil-water-surfactant mixtures. Commun Pure Appl Anal 10(6):1823–1847
25. Pawlow I, Zajaczkowski WM (2013) On a class of sixth order viscous Cahn-Hilliard type equations. Discrete Contin Dyn Syst Ser S 6(2):517–546
26. Prince LM (1977) Microemulsions: theory and practice. Academic Press, New York
27. Rosano HL, Clausse M (eds) (1987) Microemulsion systems. Marcel Dekker, New York
28. Schimperna G, Pawlow I (2013) On a class of Cahn-Hilliard models with nonlinear diffusion. SIAM J Math Anal 45(1):31–63
29. Tartar L (2007) Introduction to Sobolev spaces and interpolation spaces. UMI, Bologna, Springer, Berlin
30. Warburton T, Hesthaven JS (2003) On the constants in hp-finite element trace inverse inequalities. Comput Methods Appl Mech Eng 192(25):2765–2773
31. Wells GN, Kuhl E, Garikipati K (2006) A discontinuous Galerkin method for the Cahn-Hilliard equation. J Comput Phys 218(2):860–877

On Existence “In the Large” of a Solution to Modified Navier-Stokes Equations

George Kobelkov

Abstract The problem on existence “in the large” of a solution to the 3D Navier-Stokes equations is open up to now. Nevertheless, for some modifications of the Navier-Stokes equations describing practical problems this problem has been successfully solved. For instance, for the system of Primitive equations describing large-scale ocean dynamics, existence and uniqueness of a strong solution for any time interval and arbitrary initial conditions and viscosity coefficient was proved (Kobelkov J Math Fluid Mech 9(4):588–610, 2007) [1]. O.A. Ladyzhenskaya proposed (Trudy MIAN SSSR 102:85–104, 1967) [2] a modification of the Navier-Stokes equations allowing to prove existence of a solution “in the large”, but this modification was not “physical”. Here we improve the Ladyzhenskaya result modifying not all the three motion equations, but only two of them and only in two (horizontal) variables (not three). Such kind of problems arises in ocean dynamics models. We also consider the case when the viscosity coefficients in vertical and horizontal directions are different. For all these cases existence “in the large” of a solution is proved. Unfortunately, these results cannot be extended to the case of 3D Navier-Stokes equations as well as in the case of Ladyzhenskaya modification.

1 Case of Different Viscosity Coefficients

Let Ω be a bounded Lipshitz domain in $\mathbb{R}^3$. We denote independent variables by $x = (x_1, x_2, x_3)$ or x, y, z if it does not lead to misunderstandings. In the space of vector functions we shall use the norms and operators:

G. Kobelkov (✉)
Russian Academy of Sciences, Moscow, Russia
e-mail: kobelkov@dodo.inm.ras.ru

B. N. Chetverushkin et al. (eds.), *Contributions to Partial Differential Equations and Applications*, Computational Methods in Applied Sciences 47,
https://doi.org/10.1007/978-3-319-78325-3_17

$$\|\mathbf{f}\|^2 = \sum_{i=1}^{2} \int\limits_{\Omega} f_i^2(x)dx = \int\limits_{\Omega} |\mathbf{f}|^2\, dx, \quad \|\mathbf{f}_x\|^2 = \sum_{i=1}^{2}\sum_{j=1}^{3} \int\limits_{\Omega} \left(\frac{\partial f_i}{\partial x_j}\right)^2 dx,$$

$$\Delta = \frac{\partial^2}{\partial x_1^2} + \frac{\partial^2}{\partial x_2^2}, \quad \partial_{x_i} = \frac{\partial}{\partial x_i}, \quad \|\cdot\|_q = \|\cdot\|_{L_q},$$

$$\nabla = (\partial_x, \partial_y), \quad \operatorname{div} \mathbf{f} = \partial_x f_1 + \partial_y f_2, \quad |\nabla \mathbf{f}|^2 = \sum_{i,j=1}^{2} \left(\frac{\partial f_i}{\partial x_j}\right)^2,$$

$$|f|_{q,E_x}^q = \int\limits_{-\infty}^{\infty} |f|^q\, dx, \quad |f|_{q,E_{xy}}^q = \int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty} |f|^q\, dx\, dy.$$

In what follows, we assume summation over repeating indices in products. By c with and without indices we denote constants in inequalities not depending on the functions entering these inequalities but depending in general on initial data of the problem (a domain shape, constants from the embedding theorems, norms of the right-hand sides of equations, time interval, etc.).

The system of Navier-Stokes equations describing dynamics of incompressible viscous flow is of the form (see, e.g., [3, 4])

$$\begin{gathered}
\mathbf{u}_t - \nu\Delta\mathbf{u} - \nu\partial_z^2\mathbf{u} + \nabla p + (\mathbf{u}\cdot\nabla)\mathbf{u} + w\mathbf{u}_z = \mathbf{f}, \\
w_t - \nu\Delta w - \nu\partial_z^2 w + p_z + (\mathbf{u}\cdot\nabla)w + ww_z = g, \\
\operatorname{div}\mathbf{u} + w_z = 0, \\
(\mathbf{u}, w)(x, 0) = (\mathbf{u}_0, w_0)(x), \operatorname{div}\mathbf{u}_0 + \partial_z w_0 = 0, \quad (\mathbf{u}, w)\Big|_{\partial\Omega\times[0,T]} = \mathbf{0}.
\end{gathered} \tag{1}$$

In practice, there are problems when viscosity coefficients in vertical and horizontal directions are different. For instance, it takes place in simulation of ocean dynamics. So, it is natural to consider the case when the viscosity coefficient in horizontal direction equals ν, while in the vertical direction z it equals $\mu \geq \nu$. In this case Eq. (1) take the form

$$\begin{gathered}
\mathbf{u}_t - \nu\Delta\mathbf{u} - \mu\partial_z^2\mathbf{u} + \nabla p + (\mathbf{u}\cdot\nabla)\mathbf{u} + w\mathbf{u}_z = \mathbf{f}, \\
w_t - \nu\Delta w - \mu\partial_z^2 w + p_z + (\mathbf{u}\cdot\nabla)w + ww_z = g, \\
\operatorname{div}\mathbf{u} + w_z = 0, \\
(\mathbf{u}, w)(x, 0) = (\mathbf{u}_0, w_0)(x), \operatorname{div}\mathbf{u}_0 + \partial_z w_0 = 0, \quad (\mathbf{u}, w)\Big|_{\partial\Omega\times[0,T]} = (\mathbf{0}, 0).
\end{gathered} \tag{2}$$

For simplicity of consideration, put $\mathbf{f} = \mathbf{0}$, $g = 0$.

Let us study the solvability "in the large" of problem (2). The following theorem holds:

Theorem 1 *For any initial condition* $\mathbf{u}_0 \in \mathbf{H}_0^2$, $w_0 \in H_0^2$, *any* $\nu > 0$ *and arbitrary time interval* $[0, T]$ *there is* $\mu > 0$ *such that there exists a solution to (2) "in the* *large”, i.e. there exist* $\mathbf{u} \in \mathbf{H}^1(Q_T)$*,* $w \in H^1(Q_T)$ *satisfying* (2) *in a weak sense and the norm* $\|\mathbf{u}_x\| + \|w_x\|$ *is continuous in time on* $[0, T]$*. Moreover, in this case the following inequality holds*

$$\|\mathbf{u}_t(t)\|^2 + \|w_t(t)\|^2 \le \|\mathbf{u}_t(0)\|^2 + \|w_t(0)\|^2 \quad \forall t > 0.$$

Proof In what follows we shall use the technique proposed in [3]. To prove the theorem we need the Ladyzhenskaya inequality

$$\|f\|_4^4 \le c_1 \|f_{x_1}\| \, \|f_{x_2}\| \, \|f_{x_3}\| \, \|f\| \tag{3}$$

being valid for any $f \in H_0^1(\Omega)$; here the constant c_1 does not depend on Ω.

Take scalar product in $\mathbf{L}_2$ of the first equation of (2) and $\mathbf{u}$ and the second equation of (2) and w in L_2. Adding results, we have

$$\frac{1}{2}\frac{d}{dt}\left(\|\mathbf{u}\|^2 + \|w\|^2\right) + \nu(\|\nabla\mathbf{u}\|^2 + \|\nabla w\|^2) + \mu(\|\mathbf{u}_z\|^2 + \|w_z\|^2) = 0; \tag{4}$$

integration of (4) in time gives

$$\|\mathbf{u}(t)\|^2 + \|w(t)\|^2 \le \|\mathbf{u}_0\|^2 + \|w_0\|^2 \equiv M^2. \tag{5}$$

From (4) and (5) one gets

$$\nu(\|\nabla\mathbf{u}\|^2 + \|\nabla w\|^2) + \mu(\|\mathbf{u}_z\|^2 + \|w_z\|^2) \le M(\|\mathbf{u}_t\| + \|w_t\|). \tag{6}$$

Differentiate (2) in t:

$$\begin{aligned} \mathbf{u}_{tt} - \nu\Delta\mathbf{u}_t - \mu\partial_z^2\mathbf{u}_t + \nabla p_t + (\mathbf{u}_t\cdot\nabla)\mathbf{u} + w_t\mathbf{u}_z + (\mathbf{u}\cdot\nabla)\mathbf{u}_t + w_t\mathbf{u}_{tz} &= \mathbf{0},\\ w_{tt} - \nu\Delta w_t - \mu\partial_z^2 w_t + p_{tz} + (\mathbf{u}_t\cdot\nabla)w + w_t w_z + (\mathbf{u}\cdot\nabla)w_t + w w_{tz} &= 0,\\ \operatorname{div}\mathbf{u}_t + w_{tz} &= 0. \end{aligned} \tag{7}$$

Take now a scalar product of the first two equations of (7) and $(\mathbf{u}_t, w_t)$:

$$\begin{aligned} \frac{1}{2}\frac{d}{dt}(\|\mathbf{u}_t\|^2 + \|w_t\|^2) + \nu(\|\nabla\mathbf{u}_t\|^2 + \|\nabla w_t\|^2) + \mu(\|\mathbf{u}_{tz}\|^2 + \|w_{tz}\|^2)\\ + ((\mathbf{u}_t\cdot\nabla)\mathbf{u} + w_t\mathbf{u}_z, \mathbf{u}_t) + ((\mathbf{u}_t\cdot\nabla)w + w_t w_z, w_t) = 0. \end{aligned} \tag{8}$$

Estimate the scalar products of (8). Integration by parts gives

$$\begin{aligned} &|((\mathbf{u}_t\cdot\nabla)\mathbf{u} + w_t\mathbf{u}_z, \mathbf{u}_t) + ((\mathbf{u}_t\cdot\nabla)w + w_t w_z, w_t)|\\ &\qquad = |(u_{jt}\mathbf{u}, \partial_{x_j}\mathbf{u}_t) + (w_t\mathbf{u}, \mathbf{u}_{tz}) + (u_{jt}w, \partial_{x_j}w_t) + (w_t w, w_{tz})|. \end{aligned}$$

Estimate now each of these scalar products separately using the Hölder inequality and estimates (3), (5) and (6). We have

$$\begin{aligned}|(u_{jt}\mathbf{u}, \partial_{x_j}\mathbf{u}_t)| &\le \|\mathbf{u}_{tx}\|\,\|\mathbf{u}_t\|_4\|\mathbf{u}\|_4 \le c\|\mathbf{u}_{tx}\|^{7/4}\|\mathbf{u}_t\|^{1/4}\|\nabla\mathbf{u}\|^{1/2}\|\mathbf{u}_z\|^{1/4}\\ &\le \varepsilon\|\mathbf{u}_{tx}\|^2 + \frac{c}{\varepsilon^7}\|\nabla\mathbf{u}\|^4\|\mathbf{u}_z\|^2\|\mathbf{u}_t\|^2 \le \varepsilon\|\mathbf{u}_{tx}\|^2 + \frac{c}{\varepsilon^7\nu^2\mu}(\|\mathbf{u}_t\|^2 + \|w_t\|^2)^{5/2}.\end{aligned}$$

All other scalar products are estimated in the same way. Choosing proper ε, we finally obtain

$$\begin{aligned}|((\mathbf{u}_t\cdot\nabla)\mathbf{u} + w_t\mathbf{u}_z, \mathbf{u}_t)+&((\mathbf{u}_t\cdot\nabla)w + w_t w_z, w_t)|\\ &\le \frac{\nu}{2}(\|\mathbf{u}_{tx}\|^2 + \|w_{tx}\|^2) + \frac{c}{\nu^9\mu}(\|\mathbf{u}_t^2\| + \|w_t\|^2)^{5/2}. \qquad (9)\end{aligned}$$

Substituting (9) into (8), one gets

$$\begin{aligned}&\frac{d}{dt}(\|\mathbf{u}_t\|^2 + \|w_t\|^2) + \nu(\|\nabla\mathbf{u}_t\|^2 + \|\nabla w_t\|^2)\\ &\qquad + \mu(\|\mathbf{u}_{tz}\|^2 + \|w_{tz}\|^2) - \frac{c}{\nu^9\mu}(\|\mathbf{u}_t\|^2 + \|w_t\|^2)^{5/2} \le 0,\end{aligned}$$

from what follows

$$\begin{aligned}&\frac{d}{dt}\Big(\|\mathbf{u}_t\|^2 + \|w_t\|^2\Big) + \nu\Big(\|\mathbf{u}_{tx}\|^2 + \|w_{tx}\|^2\Big)\\ &\qquad + \left(\mu - \frac{c}{\nu^9\mu}(\|\mathbf{u}_t\|^2 + \|w_t\|^2)^{3/2}\right)(\|\mathbf{u}_{tz}\|^2 + \|w_{tz}\|^2) \le 0. \qquad (10)\end{aligned}$$

It is obvious that $\|\mathbf{u}_t(0)\|^2 + \|w_t(0)\|^2$ can be estimated from above by some expression depending on the norm $\|(\mathbf{u}_0, w_0)\|_{\mathbf{H}^2}$ only. Now, from (10) it follows that for any $\nu > 0$ and arbitrary $\|\mathbf{u}_t(0)\| + \|w_t(0)\|$ depending on the norm of initial condition $\|\mathbf{u}_0\|_{\mathbf{H}^2} + \|w_0\|_{H^2}$ there exists $\mu > 0$ such that

$$\mu - \frac{c}{\nu^9\mu}\left(\|\mathbf{u}_t(0)\|^2 + \|w_t(0)\|^2\right)^{3/2} \ge 0.$$

Then from (10) we conclude that the norm $\|\mathbf{u}_t(t)\|^2 + \|w_t(t)\|^2$ satisfies the inequality

$$\|\mathbf{u}_t(t)\|^2 + \|w_t(t)\|^2 \le \|\mathbf{u}_t(0)\|^2 + \|w_t(0)\|^2 \quad \forall t > 0. \qquad (11)$$

Existence and uniqueness of a solution “in the large” may be obtained with the help of estimate (11) in the same way as in [3]. The proof is completed.

2 Improvement of the Ladyzhenskaya Modification

Consider now another modification of the Navier-Stokes equations, when the viscosity coefficient is the same in all directions, but the elliptic operator is changed. O.A. Ladyzhenskaya proposed (see, e.g., [2]) a modification of the Navier-Stokes equations allowing to prove existence of a strong solution to (1) “in the large”:

$$\begin{gathered}
\mathbf{u}_t - \nu \Delta \mathbf{u} - \nu \partial_z^2 \mathbf{u} - \nu\varepsilon\Big[\operatorname{div}\left(D(\mathbf{u}, w)\nabla\mathbf{u}\right) + \partial_z\left(D(\mathbf{u}, w)\partial_z\mathbf{u}\right)\Big] \\
+ \nabla p + (\mathbf{u}\cdot\nabla)\mathbf{u} + w\mathbf{u}_z = \mathbf{f}, \\
w_t - \nu\Delta w - \nu\partial_z^2 w - \nu\varepsilon\left[\operatorname{div}\left(D(\mathbf{u}, w)\nabla w\right) + \partial_z\left(D(\mathbf{u}, w)\partial_z w\right)\right] \\
+ p_z + (\mathbf{u}\cdot\nabla)w + ww_z = g, \\
\operatorname{div}\mathbf{u} + w_z = 0, \\
(\mathbf{u}, w)(x, 0) = (\mathbf{u}_0, w_0)(x),\ \operatorname{div}\mathbf{u}_0 + \partial_z w_0 = 0, \quad (\mathbf{u}, w)\Big|_{\partial\Omega\times[0,T]} = 0;
\end{gathered} \tag{12}$$

here

$$D(\mathbf{u}, w) = |\nabla\mathbf{u}|^2 + |\partial_z\mathbf{u}|^2. \tag{13}$$

(As a matter of fact, Ladyzhenskaya used another form of D, but here, for simplicity, we use this form. For both forms of D all considerations are absolutely the same.)

Consider another modification of the Navier-Stokes equations arising in ocean dynamics, which may be considered as strengthening of the Ladyzhenskaya results. Namely, we consider (12) as modification of (1), but instead of (13) we use $D(\mathbf{u}, w) = |\nabla\mathbf{u}|^2$, remove the term $\partial_z\left(D(\mathbf{u}, w)\partial_z\mathbf{u}\right)$ from the first equation of (12), and do not change the equation for w. So, we consider the problem

$$\begin{gathered}
\mathbf{u}_t - \nu\Delta\mathbf{u} - \nu\partial_z^2\mathbf{u} - \nu\varepsilon\operatorname{div}\left(|\nabla\mathbf{u}|^2\nabla\mathbf{u}\right) + \nabla p + (\mathbf{u}\cdot\nabla)\mathbf{u} + w\mathbf{u}_z = \mathbf{f}, \\
w_t - \nu\Delta w - \nu\partial_z^2 w + p_z + (\mathbf{u}\cdot\nabla)w + ww_z = g, \\
\operatorname{div}\mathbf{u} + w_z = 0, \\
(\mathbf{u}, w)(x, 0) = (\mathbf{u}_0, w_0)(x),\ \operatorname{div}\mathbf{u}_0 + \partial_z w_0 = 0, \quad (\mathbf{u}, w)\Big|_{\partial\Omega\times[0,T]} = 0.
\end{gathered} \tag{14}$$

To study solvability of (14) “in the large”, we need the following lemmas.

Lemma 1 *Let $v \in H_0^1[0, l]$. Then the following estimate holds:*

$$\max_x v^2(x) \le 2\|v_x\|\,\|v\|.$$

Proof Extend v onto the whole axes by zero. Then

$$v^2(x) = 2\int_{-\infty}^{x} v_x(x)v(x)\,dx \le 2\|v_x\|\,\|v\|.$$

Q.E.D.

Lemma 2 *Let $f \in H_0^1(\Omega)$, $f_x,\ f_y \in L_4(\Omega)$, $\Omega \in \mathbb{R}^3$, then*

$$\|f\|_5^5 \le \frac{25}{2}\|f_x\|_4\|f_y\|_4\|f_z\|\,\|f\|^2. \tag{15}$$

Proof Extend by zero the function f onto the whole space $\mathbb{R}^3$. Then

$$|f(x,y,z)|^5 = \frac{25}{4}\int_{-\infty}^{x}\sqrt{|f(x,y,z)|}\,f(x,y,z)\,f_x(x,y,z)\,dx \\ \times \int_{-\infty}^{y}\sqrt{|f(x,y,z)|}\,f(x,y,z)\,f_y(x,y,z)\,dy.$$

Using the Hölder inequality, from the previous expression we have

$$|f(x,y,z)|^5 \le \frac{25}{4}|f|_{E_x}^{3/2}\,|f_x|_{4,E_x}\,|f|_{E_y}^{3/2}\,|f_y|_{4,E_y}.$$

Integration over $\mathbb{R}^3$ and further implementation of the Hölder inequality give

$$\begin{aligned}\int_{E_z}\int_{E_{xy}}|f|^5dx\,dy\,dz &\le \frac{25}{4}\int_{E_z}\left[\int_{E_y}|f|_{E_x}^{3/2}|f_x|_{4,E_x}dy\int_{E_x}|f|_{E_y}^{3/2}|f_y|_{4,E_y}dx\right]dz\\ &\le \frac{25}{4}\int_{E_z}|f|_{E_{xy}}^{3}\,|f_x|_{4,E_{xy}}\,|f_y|_{4,E_{xy}}dz\\ &\le \frac{25}{4}\max_z|f|_{E_{xy}}^2\int_{E_z}|f|_{E_{xy}}\,|f_x|_{4,E_{xy}}\,|f_y|_{4,E_{xy}}dz \qquad \text{(due to Lemma 1)}\\ &\le \frac{25}{2}\|f_x\|_4\,\|f_y\|_4\,\|f_z\|\,\|f\|^2dz.\end{aligned}$$

Q.E.D.

Corollary 1 *Since $abc \le 0.5a^2b^2 + 0.5c^2 \le 0.25(a^4+b^4) + 0.5c^2$, $a,b,c \ge 0$, then from (15) it follows*

$$\|f\|_5^5 \le \frac{25}{8}(\|f_x\|_4^4 + \|f_y\|_4^4 + 2\|f_z\|^2)\|f\|^2. \tag{16}$$

Let us obtain a proper a priori estimate for a solution to (14). Take a scalar product of (14) and $(\mathbf{u}, w)$:

$$\frac{1}{2}\frac{d}{dt}\left(\|\mathbf{u}\|^2 + \|w\|^2\right) + \nu\left(\|\mathbf{u}_x\|^2 + \|w_x\|^2\right) + \varepsilon\nu\|\nabla\mathbf{u}\|_4^4 = (\mathbf{f},\mathbf{u}) + (g,w). \tag{17}$$

Obvious estimation of the right-hand side of (17) and further integration in t from 0 to T give

$$\begin{aligned}\max_{0\le t\le T}\left(\|\mathbf{u}(t)\|^2+\|w(t)\|^2\right)+\nu\int\limits_0^T\left(\|\mathbf{u}_x\|^2+\|w_x\|^2+\varepsilon\|\nabla\mathbf{u}\|_4^4\right)dt\\ \le c_1\left(\|\mathbf{u}_0\|^2+\|w_0\|^2+\frac{1}{\nu}\int\limits_0^T\left(\|\mathbf{f}\|_{-1}^2+\|g\|_{-1}^2\right)dt\right)\equiv c_2.\end{aligned}\tag{18}$$

Rewrite (17) in another form:

$$\nu\left(\|\mathbf{u}_x\|^2+\|w_x\|^2\right)+\varepsilon\nu\|\nabla\mathbf{u}\|_4^4=(\mathbf{f},\mathbf{u})+(g,w)-(\mathbf{u}_t,\mathbf{u})-(w_t,w).$$

Estimating the right-hand side and using (18), we get

$$\|\mathbf{u}_x\|^2+\|w_x\|^2+\varepsilon\|\nabla\mathbf{u}\|_4^4\le c_3(\|\mathbf{u}_t\|+\|w_t\|+\|\mathbf{f}\|_{-1}^2+\|g\|_{-1}^2).$$

Now, differentiate (14) in t:

$$\begin{aligned}&\mathbf{u}_{tt}-\nu\mathrm{div}\left((1+\varepsilon|\nabla\mathbf{u}|^2)\nabla\mathbf{u}_t\right)-\nu\varepsilon\mathrm{div}\left([|\nabla\mathbf{u}|^2]_t\nabla\mathbf{u}\right)-\nu\partial_z^2\mathbf{u}_t\\ &\quad+\nabla p_t+(\mathbf{u}\cdot\nabla)\mathbf{u}_t+(\mathbf{u}_t\cdot\nabla)\mathbf{u}+w\mathbf{u}_{tz}+w_t\mathbf{u}_z=\mathbf{f}_t,\\ &\quad w_{tt}-\nu\Delta w_t-\nu\partial_z^2w_t+p_{tz}+(\mathbf{u}\cdot\nabla)w_t\\ &\qquad+(\mathbf{u}_t\cdot\nabla)w+ww_{tz}+w_tw_z=g_t,\\ &\mathrm{div}\,\mathbf{u}_t+w_{tz}=0,\quad(\mathbf{u}_t,w_t)\Big|_{\partial\Omega\times[0,T]}=(\mathbf{0},0).\end{aligned}\tag{19}$$

Taking scalar product of (19) and $(\mathbf{u}_t,w_t)$, one obtains

$$\begin{aligned}&\frac{1}{2}\frac{d}{dt}\|\mathbf{u}_t\|^2+\frac{1}{2}\frac{d}{dt}\|w_t\|^2+\nu\|\mathbf{u}_{tx}\|^2+\nu\|w_{tx}\|^2\\ &\quad+\varepsilon\nu\int\limits_\Omega(\nabla\mathbf{u})^2(\nabla\mathbf{u}_t)^2dx+\frac{\varepsilon\nu}{2}\|[(\nabla\mathbf{u})^2]_t\|^2+(u_{kt}\mathbf{u}_{x_k},\mathbf{u}_t)\\ &\quad+(w_t\mathbf{u}_z,\mathbf{u}_t)+(u_{kt}w_{x_k},w_t)+(w_tw_z,w_t)=(\mathbf{f}_t,\mathbf{u}_t)+(g_t,w_t).\end{aligned}\tag{20}$$

To estimate scalar products of (20) we need the following inequalities (see, e.g., [3])

$$\begin{aligned}\|v\|_{10/3}&\le(48)^{1/10}\|v_x\|^{3/5}\|v\|^{2/5},\\ \|v\|_3&\le(48)^{1/2}\|v_x\|^{1/2}\,\|v\|^{1/2},\\ \|v\|_{8/3}&\le(48)^{1/16}\|v_x\|^{3/8}\,\|v\|^{5/8},\end{aligned}\tag{21}$$

being valid for functions from the Sobolev space $H_0^1(\Omega)$, $\Omega \in \mathbb{R}^3$. Then we have

$$\begin{aligned}|I_1| = |(u_{kt}\mathbf{u}_{x_k}, \mathbf{u}_t)| &\le c\|\nabla\mathbf{u}\|_3\|\mathbf{u}_t\|_3^2 \le c\|\nabla\mathbf{u}\|_4\|\mathbf{u}_{tx}\|\|\mathbf{u}_t\| \\ &\le \delta\|\mathbf{u}_{tx}\|^2 + \frac{c}{\delta}\|\nabla\mathbf{u}\|_4^2\|\mathbf{u}_t\|^2.\end{aligned}$$

Estimate now the second scalar product. Integration by parts and the use of the incompressibility equation give

$$I_2 = (w_t\mathbf{u}_z, \mathbf{u}_t) = (\operatorname{div}\mathbf{u}_t\mathbf{u}, \mathbf{u}_t) - (w_t\mathbf{u}, \mathbf{u}_{tz}) = I_2' + I_2''.$$

Estimate each of these scalar products separately using the Hölder and Young inequalities. We get

$$\begin{aligned}|I_2'| &= |(\operatorname{div}\mathbf{u}_t\mathbf{u}, \mathbf{u}_t)| \quad \text{(use the Hölder inequality with the powers 2, 5, 10/3)} \\ &\le \|\mathbf{u}_{tx}\|\|\mathbf{u}\|_5\|\mathbf{u}_t\|_{10/3} \quad \text{(due to (21))} \\ &\le (48)^{1/10}\|\mathbf{u}_{tx}\|^{8/5}\|\mathbf{u}\|_5\|\mathbf{u}_t\|^{2/5} \quad \text{(due to the Young inequality)} \\ &\le \delta\|\mathbf{u}_{tx}\|^2 + c_\delta\|\mathbf{u}\|_5^5\|\mathbf{u}_t\|^2 \quad \text{(due to (16))} \\ &\le \delta\|\mathbf{u}_{tx}\|^2 + c_\delta(\|\nabla\mathbf{u}\|_4^4 + \|\mathbf{u}_z\|^2)\|\mathbf{u}_t\|^2.\end{aligned}$$

In the same way one gets

$$\begin{aligned}|I_2''| &= |(w_t\mathbf{u}, \mathbf{u}_{tz})| \quad \text{(use the Hölder inequality with the powers 10/3, 5, 2)} \\ &\le c\|\mathbf{u}_{tz}\|\|\mathbf{u}\|_5\|w_t\|_{10/3} \quad \text{(due to (21))} \\ &\le c\|\mathbf{u}_{tz}\|\|\mathbf{u}\|_5\|w_{tx}\|^{3/5}\|w_t\|^{2/5} \\ &\le \delta\|\mathbf{u}_{tz}\|^2 + c_\delta\|\mathbf{u}\|_5^2\|w_{tx}\|^{6/5}\|w_t\|^{4/5} \quad \text{(due to the Young inequality} \\ &\qquad\qquad \text{with the powers 5/3, 5/2)} \\ &\le \delta\|\mathbf{u}_{tz}\|^2 + \delta\|w_{tx}\|^2 + c_\delta\|\mathbf{u}\|_5^5\|w_t\|^2 \quad \text{(due to (16))} \\ &\le \delta\|\mathbf{u}_{tz}\|^2 + \delta\|w_{tx}\|^2 + c_\delta(\|\nabla\mathbf{u}\|_4^4 + \|\mathbf{u}_z\|^2)\|w_t\|^2.\end{aligned}$$

For estimation the two other scalar products, we need the following:

Lemma 3 *The estimate*

$$\max_z |w|_{4,E_{xy}} \le c\|\nabla\mathbf{u}\|_4$$

holds.

Proof As before, extend w and $\mathbf{u}$ onto the whole $\mathbb{R}^3$ by zero and denote the obtained functions by the same letters. Then, we get

$$|w|_{4,E_{xy}} = \left(\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left(\int_{-\infty}^{z} \operatorname{div}\mathbf{u}dz\right)^4 dxdy\right)^{1/4} \le \int_{-\infty}^{\infty} |\nabla\mathbf{u}|_{4,E_{xy}}dz \le c\|\nabla\mathbf{u}\|_4.$$

The statement of the lemma follows directly from the last inequality. Q.E.D.

Estimate now the scalar product I_3:

$$\begin{aligned} I_3 &= (u_{kt}w_{x_k}, w_t) = -(\operatorname{div}\mathbf{u}_t w, w_t) - (u_{kt}w, w_{tx_k}) \\ &= (w_{tz}w, w_t) - (u_{kt}w, w_{tx_k}) = I_3' + I_3''. \end{aligned}$$

Obtain estimates for I_3' and I_3'' separately. One has

$$\begin{aligned} |I_3''| &= |(u_{kt}w, w_{tx_k})| \le \int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |u_{kt}w, w_{tx_k}|\,dx\,dy\right)dz \\ &\le c\|\nabla\mathbf{u}\|_4 \int_{-\infty}^{\infty} |\mathbf{u}_t|_{4,E_{xy}}|w_{tx}|_{2,E_{xy}}dz \\ &\le c\|\nabla\mathbf{u}\|_4 \int_{-\infty}^{\infty} |\nabla\mathbf{u}_t|_{2,E_{xy}}^{1/2}|\mathbf{u}_t|_{2,E_{xy}}^{1/2}|w_{tx}|_{2,E_{xy}}dz \\ &\le c\|\nabla\mathbf{u}\|_4\|\nabla\mathbf{u}_t\|^{1/2}\|\mathbf{u}_t\|^{1/2}\|w_{tx}\| \le \delta\|w_{tx}\|^2 + c_\delta\|\nabla\mathbf{u}\|_4^2\|\nabla\mathbf{u}_t\|\|\mathbf{u}_t\| \\ &\le \delta\|w_{tx}\|^2 + \delta\|\nabla\mathbf{u}_t\|^2 + c_\delta\|\nabla\mathbf{u}\|_4^4\|\mathbf{u}_t\|^2, \\ |I_3'| &= |(w_t, w_z, w_t)| = |(\operatorname{div}\mathbf{u}, w_t^2)| \le \|\nabla\mathbf{u}\|_4\|w_t\|_{8/3}^2 \\ &\le c\|\nabla\mathbf{u}\|_4\|w_{tx}\|^{3/4}\|w_t\|^{5/4} \le \delta\|w_{tx}\|^2 + c_\delta\|\nabla\mathbf{u}\|_4^{8/5}\|w_t\|^2. \end{aligned}$$

Finally, $I_4 = I_3'$, so I_4 is estimated from above as I_3'. Substituting the above inequalities into (20) with appropriate δ and estimating the right-hand side of (20) in the obvious way, we get

$$\begin{aligned} \frac{d}{dt}(\|\mathbf{u}_t\|^2 + \|w_t\|^2) + \nu\|\mathbf{u}_{tx}\|^2 + \nu\|w_{tx}\|^2 + \varepsilon\nu\int_\Omega(\nabla\mathbf{u})^2(\nabla\mathbf{u}_t)^2dx + \frac{\varepsilon\nu}{2}\|[(\nabla\mathbf{u})^2]_t\|^2 \\ \le c\left(\|\mathbf{f}_t\|_{-1}^2 + \|g_t\|_{-1}^2 + (\|\nabla\mathbf{u}\|_4^4 + \|w_x\|^2)(\|\mathbf{u}_t\|^2 + \|w_t\|^2)\right). \qquad (22) \end{aligned}$$

Using the Gronwall inequality and (18), from (22) one obtains

$$\max_{0\le t\le T}(\|\mathbf{u}_t(t)\|^2+\|w_t(t)\|^2)+\int_0^T(\|\mathbf{u}_{tx}\|^2+\|w_{tx}\|^2)dt$$
$$\le\left(\|\mathbf{u}_t(0)\|^2+\|w_t(0)\|^2+\int_0^T(\|\mathbf{f}_t\|_{-1}^2+\|g_t\|_{-1}^2)dt\right)$$
$$\times\exp\left(c\int_0^T(\|\nabla\mathbf{u}\|_4^4+\|w_x\|^2)dt\right)$$
$$\le c_4\left(\|\mathbf{u}_t(0)\|^2+\|w_t(0)\|^2+\int_0^T(\|\mathbf{f}_t\|_{-1}^2+\|g_t\|_{-1}^2)dt\right).$$

Introduce the space $\mathbf{V}$ being the closure of divergence free vector functions $(\mathbf{u}, w)$ vanishing on $\partial\Omega\times[0,T]$ in the norm

$$\|(\mathbf{u},w)\|_V=\left(\int_0^T\left(\|\mathbf{u}_x\|^2+\|w_x\|^2+\|\mathbf{u}_t\|^2+\|w_t\|^2\right)dt\right)^{1/2}+\left(\int_0^T\|\nabla\mathbf{u}\|_4^4dt\right)^{1/4}$$

and define a solution to (14) as a vector function $(\mathbf{u}, w)\in\mathbf{V}$ being equal to $(\mathbf{u}_0, w_0)$ for $t=0$ and satisfying the following identity:

$$\int_0^T\Big((\mathbf{u}_t,\mathbf{v})+\nu(\mathbf{u}_x,\mathbf{v}_x)+\nu\varepsilon(|\nabla\mathbf{u}|^2\nabla\mathbf{u},\nabla\mathbf{v})+\nu(w_x,h_x)+((\mathbf{u}\cdot\nabla)\mathbf{u},\mathbf{v})$$
$$+((\mathbf{u}\cdot\nabla)w,h)+(ww_z,h)-(\mathbf{f},\mathbf{v})-(g,h)\Big)dt=0\quad\forall(\mathbf{v},h)\in\mathbf{V}. \tag{23}$$

Using the Galerkin method, estimate (26) and technique of [3, 4], it is not difficult to prove that a solution to (23) exists and is unique and the norm $\|\mathbf{u}_x\|+\|w_x\|$ is continuous in time.

Thus, we have proved the following:

Theorem 2 *Let* $(\mathbf{u}_0, w_0)\in\mathbf{H}_0^2$ *and*

$$\int_0^T(\|\mathbf{f}_t\|_{-1}^2+\|g_t\|_{-1}^2)dt<\infty.$$

Then for any $\varepsilon>0$, $\nu>0$, *and arbitrary time interval* T *there exists a unique solution to (14) satisfying (23), and the norm* $\|\mathbf{u}_x\|+\|w_x\|$ *is continuous in time on* $[0,T]$.

References

1. Kobelkov GM (2007) Existence of a solution "in the large" for ocean dynamics equations. J Math Fluid Mech 9(4):588–610
2. Ladyzhenskaya OA (1967) On modifications of the Navier-Stokes equations for large gradients. Trudy MIAN SSSR 102:85–104 (in Russian)
3. Ladyzhenskaya OA (1969) The mathematical theory of viscous incompressible flow. Gordon and Breach, New York
4. Temam R (1984) Navier-Stokes equations: theory and numerical analysis. North-Holland, Amsterdam

An Algebraic Solver for the Oseen Problem with Application to Hemodynamics

Igor N. Konshin, Maxim A. Olshanskii and Yuri V. Vassilevski

Abstract The paper studies an iterative method for algebraic problems arising in numerical simulation of blood flows. Here we focus on a numerical solver for the fluid part of otherwise coupled fluid-structure system of equations which models the hemodynamics in vessels. Application of the finite element method and semi-implicit time discretization leads to the discrete Oseen problem at every time step of the simulation. The problem challenges numerical methods by anisotropic geometry, open boundary conditions, small time steps and transient flow regimes. We review known theoretical results and study the performance of recently proposed preconditioners based on two-parameter threshold ILU factorization of non-symmetric saddle point problems. The preconditioner is applied to the linearized Navier–Stokes equations discretized by the stabilized Petrov–Galerkin finite element (FE) method. Careful consideration is given to the dependence of the solver on the stabilization parameters of the FE method. We model the blood flow in the digitally reconstructed right coronary artery under realistic physiological regimes. The paper discusses what is special in such flows for the iterative algebraic solvers, and shows how the two-parameter ILU preconditioner is able to meet these specifics.

Keywords Hemodynamics · Iterative methods · Threshold ILU factorization · Navier–Stokes equations · Finite element method · SUPG stabilization

I. N. Konshin · Y. V. Vassilevski (✉)
Institute of Numerical Mathematics of the Russian Academy of Sciences, Moscow, Russia
e-mail: yuri.vassilevski@gmail.com

M. A. Olshanskii
Department of Mathematics, University of Houston, Houston, USA
e-mail: molshan@math.uh.edu

I. N. Konshin
Dorodnicyn Computing Centre, FRC CSC RAS, Moscow, Russia
e-mail: igor.konshin@gmail.com

Y. V. Vassilevski
Moscow Institute of Physics and Technology, Moscow, Russia

Y. V. Vassilevski
I.M. Sechenov First Moscow State Medical University, Moscow, Russia

B. N. Chetverushkin et al. (eds.), *Contributions to Partial Differential Equations and Applications*, Computational Methods in Applied Sciences 47,
https://doi.org/10.1007/978-3-319-78325-3_18

1 Introduction

Numerical simulations play an increasing role in visualization, understanding and predictive modelling of many biological flows, including blood flow in arteries and the heart. The efficiency of a numerical approach depends on the right choice of mathematical model, its discretization and the algebraic solvers used to compute the solution to a discrete model. For the blood flow simulations, state-of-the-art methods are built on a fluid-structure interaction (FSI) model which typically includes equations describing the motion of Newtonian viscous fluid, equations for an elastic structure and coupling conditions [5]. In the process of numerical integration of the FSI system, however, one often decouples the fluid equations from the elasticity equations on every time step and hence applies segregated algebraic solvers for each of the decoupled problem, see, e.g., [12]. Furthermore, for the reason of time-sensitivity of simulations or the ambiguity of the information regarding the properties of the structure, hemodynamic simulations are often performed in a fixed geometries, i.e. the vessels wall is assumed to be rigid rather than elastic. In both cases, one is interested in an efficient numerical solve for the Navier–Stokes equations describing the motion of incompressible Newtonian fluids in a bounded domain $\Omega \subset \mathbb{R}^3$ and time interval $[0, T]$:

$$\begin{cases} \dfrac{\partial \mathbf{u}}{\partial t} - \nu \Delta \mathbf{u} + (\mathbf{u} \cdot \nabla)\mathbf{u} + \nabla p = \mathbf{f} & \text{in } \Omega \times (0, T], \\ \operatorname{div} \mathbf{u} = 0 & \text{in } \Omega \times [0, T], \\ \mathbf{u} = \mathbf{g} & \text{on } \Gamma_0 \times [0, T], \\ -\nu(\nabla \mathbf{u}) \cdot \boldsymbol{n} + p\boldsymbol{n} = \mathbf{h} & \text{on } \Gamma_{\mathrm{N}} \times [0, T], \\ \mathbf{u}(\mathbf{x}, 0) = \mathbf{u}_0(\mathbf{x}) & \text{in } \Omega. \end{cases} \tag{1}$$

The unknowns are the velocity vector field $\mathbf{u} = \mathbf{u}(\mathbf{x}, t)$ and the pressure field $p = p(\mathbf{x}, t)$. The volume forces $\mathbf{f}$, boundary and initial values $\mathbf{g}$, $\mathbf{h}$ and $\mathbf{u}_0$ are given. Parameter ν is the kinematic viscosity; the boundary of the domain is decomposed as $\partial\Omega = \overline{\Gamma}_0 \cup \overline{\Gamma}_{\mathrm{N}}$ with Dirichlet part $\Gamma_0 \neq \varnothing$ and Neumann part Γ_{N}. An important parameter of the flow is the dimensionless Reynolds number $\mathrm{Re} = UL/\nu$, where U and L are characteristic velocity and linear dimension.

The Navier–Stokes equations (1) are fundamental equations of fluid mechanics and are central for modelling of many physical phenomena. In hemodynamic applications, one may point to several special features of otherwise general fluid flow problem in (1):

(i) Anisotropic geometry. The domain Ω typically represents a blood vessel, which is a stretched branching object;

(ii) Open boundaries of mixed type. The computational domain has artificial (open) boundaries, where the vessel is cut. Depending on the stage of cardiac cycle, forward and reverse flows may happen through the same part of the open boundary, leading to the boundary changing type outflow/inflow;

(iii) Different flow regimes. Variable blood flux generated over one heartbeat may produce flows with varying Reynolds numbers from laminar to transitional;

(iv) Finite element method prevails. Due to complex geometry and coupling to elasticity equations, finite element method is the very common choice for discretization of (1) in hemodynamic applications. A regularization (in the form of least-square terms or a sub-grid model) is often added to stabilize the FE method for higher Reynolds numbers;

(v) Small time steps. The physics of the problem dictates small time steps of order 10^{-3} × cardiac cycle time for the numerical integration of (1).

Semi-implicit time discretization or an implicit one combined with the linearization of the Navier–Stokes system (1) by Picard fixed-point iteration result in a sequence of Oseen problems of the form

$$\begin{cases} \alpha\mathbf{u} - \nu\Delta\mathbf{u} + (\mathbf{w}\cdot\nabla)\mathbf{u} + \nabla p = \hat{\mathbf{f}} & \text{in } \Omega, \\ \mathrm{div}\mathbf{u} = \hat{g} & \text{in } \Omega, \\ \mathbf{u} = \mathbf{0} & \text{on } \Gamma_0, \\ -\nu(\nabla\mathbf{u})\cdot\boldsymbol{n} + p\boldsymbol{n} = \mathbf{0} & \text{on } \Gamma_{\mathrm{N}}, \end{cases} \tag{2}$$

where $\mathbf{w}$ is a known velocity field from a previous iteration or time step and α is proportional to the reciprocal of the time step. Non-homogeneous boundary conditions in the nonlinear problem are accounted in the right-hand side of (2). A finite element spatial discretization of (2) produces large sparse systems of the form

$$\begin{pmatrix} A & \widetilde{B}^T \\ B & -C \end{pmatrix}\begin{pmatrix} u \\ p \end{pmatrix} = \begin{pmatrix} f \\ g \end{pmatrix}, \tag{3}$$

where u and p represent the discrete velocity and pressure, respectively; $A \in \mathbb{R}^{n\times n}$ is the discretization of the diffusion, convection, and time-dependent terms. The matrix A accounts also for certain stabilization terms. Matrices B and $\widetilde{B}^T \in \mathbb{R}^{n\times m}$ are (negative) discrete divergence and gradient. These matrices may also be perturbed due to stabilization. It is typical for the stabilized methods that $B \neq \widetilde{B}$, while for a plain Galerkin method these two matrices are the same. Matrix $C \in \mathbb{R}^{m\times m}$ results from possible pressure stabilization terms, and f and g contain forcing and boundary terms. For the LBB stable finite elements, no pressure stabilization is required and so $C = 0$ holds. If the LBB condition is not satisfied, the stabilization matrix $C \neq 0$ is typically symmetric and positive semidefinite. For $B = \widetilde{B}$ of the full rank and positive definite $A = A^T$ the solution to (3) is a saddle point.

Considerable work has been done in developing efficient preconditioners for Krylov subspace methods applied to system (3) with $\widetilde{B} = B$; see the comprehensive studies in [4, 8, 19] of the preconditioning exploiting the block structure of the system. Several algebraic solvers were specifically designed or numerically tested for solving (3) resulting from hemodynamic applications. This includes incomplete block LU factorizations mimicking pressure correction splitting methods on the algebraic level [20], block-triangular preconditioners based on approximation of pressure advection–diffusion operator [18], additive Schwartz preconditioner [7], relaxed dimensional factorization block preconditioner [3], see also [7] for the numerical comparison of several preconditioners for the hemodynamic simulations.

The special features of blood flow problems discussed above impact the algebraic properties of the discrete system (3), and ideally, an efficient solver accounts for them. Thus, the inf-sup stability constants of velocity–pressure elements strongly depend on the anisotropy of domain Ω, see [6]. This may lead to poor performance of preconditioners based on pressure Schur complement approximations. Reversed flows through the open boundary is an energy increasing and de-stabilizing phenomenon, potentially resulting in the lost of ellipticity by the A block of (3). Next, different flow regimes require a robust preconditioner with respect to the variation of the Reynolds numbers. Finite element method leads, in general, to matrices with higher fill-in comparing to finite volumes or finite differences schemes. We note that hierarchical tetrahedral grids are rarely used to reconstruct blood vessels. This reduce the applicability of geometric multigrid methods. Furthermore, we shall see that additional terms added to stabilize finite element method for convection dominated flows often make algebraic problem harder to solve. Finally, small time steps suggest that reusable preconditioners and those benefiting from the diagonal dominance in the A-block should be preferred.

In the paper we study the properties of an algebraic solver for (3) based on a Krylov subspace iterative method and a two-parameter ILU preconditioner. The preconditioner results from a special incomplete elementwise LU factorization suggested and studied in [14] for symmetric positive definite matrices and further extended to non-symmetric saddle-point systems in [16, 17]. Here we review the available analysis and discuss how this algebraic solver addresses the challenges posed by hemodynamics applications. Further we simulate the blood flow in the digitally reconstructed part of the right coronary artery. Here we experiment with various grids, Reynolds numbers and finite element method stabilization parameters to assess the numerical properties for the iterative method.

The remainder of the paper is organized as follows. In Sect. 2 we give necessary details of the finite element method. Section 3 reviews known stability of the exact LU factorizations for (3). These results are formulated in terms of the properties of the (1,1)-block A, auxiliary Schur complement matrix $BA^{-1}B^T + C$, and the perturbation matrix $B - \widetilde{B}$. In Sect. 4, we formulate the properties of these matrices in terms of problem coefficients and parameters of the FE method. In Sect. 5, we briefly discuss the implication of these results on the stability of a two-parameter variant of the threshold ILU factorization for non-symmetric non-definite problems. In Sect. 6 we study the numerical performance of the method on the sequence of linear systems appearing in simulation of a blood flow in a right coronary artery. Conclusions are collected in the final Sect. 7.

2 Finite Element Method

We assume T_h to be a collection of tetrahedra forming a consistent subdivision of Ω. We also assume for T_h the shape-regularity condition,

$$\max_{\tau \in T_h} \operatorname{diam}(\tau)/\rho(\tau) \le C_T, \tag{4}$$

where $\rho(\tau)$ is the diameter of the inscribed ball in the tetrahedron τ. A constant C_T measures the maximum anisotropy ratio for T_h. Further we denote $h_\tau = \operatorname{diam}(\tau)$, $h_{\min} = \min_{\tau \in T_h} h_\tau$. Given conforming FE spaces $\mathbb{V}_h \subset (H^1_{\Gamma_0}(\Omega))^3$ and $\mathbb{Q}_h \subset L^2(\Omega)$, the Galerkin FE discretization of (2) is based on the weak formulation: Find $\{\boldsymbol{u}_h, p_h\} \in \mathbb{V}_h \times \mathbb{Q}_h$ such that

$$\begin{aligned} &\mathscr{L}(\boldsymbol{u}_h, p_h; \boldsymbol{v}_h, q_h) = (\hat{\mathbf{f}}, \boldsymbol{v}_h) + (\hat{g}, q_h) \qquad \forall \boldsymbol{v}_h \in \mathbb{V}_h,\ q_h \in \mathbb{Q}_h, \\ &\quad \mathscr{L}(\boldsymbol{u}, p; \boldsymbol{v}, q) := \alpha(\boldsymbol{u}, \boldsymbol{v}) + \nu(\nabla \boldsymbol{u}, \nabla \boldsymbol{v}) + ((\boldsymbol{w} \cdot \nabla)\, \boldsymbol{u}, \boldsymbol{v}) - (p,\ \operatorname{div} \boldsymbol{v}) + (q,\ \operatorname{div} \boldsymbol{u}), \end{aligned} \tag{5}$$

where $(\cdot, \cdot)$ denotes the $L^2(\Omega)$ inner product.

In experiments we use P2-P1 Taylor–Hood FE pair, which satisfies the LBB compatibility condition for $\mathbb{V}_h$ and $\mathbb{Q}_h$ [9] and hence ensures well-posedness and full approximation order for the FE linear problem.

The finite element method (5) needs stabilization or additional subgrid scale modelling if convection terms dominate over the diffusion. We consider one commonly used SUPG stabilization, while more details on the family of SUPG methods can be found in, e.g., [21]. Using (5) as the starting point, a weighted residual for the FE solution multiplied by an ‘advection’-depending test function is added:

$$\begin{aligned} \mathscr{L}(\boldsymbol{u}_h, p_h; \boldsymbol{v}_h, q_h) + \sum_{\tau \in T_h} \sigma_\tau (\alpha \boldsymbol{u}_h - \nu \Delta \boldsymbol{u}_h + \boldsymbol{w} \cdot \nabla \boldsymbol{u}_h + \nabla p_h - \boldsymbol{f}, \boldsymbol{w} \cdot \nabla \boldsymbol{v}_h)_\tau \\ = (\boldsymbol{f}, \boldsymbol{v}_h) + (\hat{g}, q_h) \quad \forall \boldsymbol{v}_h \in \mathbb{V}_h,\ q_h \in \mathbb{Q}_h, \end{aligned} \tag{6}$$

with $(f, g)_\tau := \int_\tau f g\, dx$. The second term in (6) is evaluated element-wise for each element $\tau \in T_h$. Parameters σ_τ are element- and problem-dependent. To define the parameters, we introduce mesh Reynolds numbers $\mathrm{Re}_\tau := \|\boldsymbol{w}\|_{L_\infty(\tau)} h_{\boldsymbol{w}}/\nu$ for all $\tau \in T_h$, where $h_{\boldsymbol{w}}$ is the diameter of τ in direction $\boldsymbol{w}$. Several recipes for the particular choice of the stabilization parameters can be found in the literature, see, e.g., [21].

We set

$$\sigma_\tau = \begin{cases} \bar{\sigma} \frac{h_{\boldsymbol{w}}}{2\|\boldsymbol{w}\|_{L_\infty(\tau)}} \left(1 - \frac{1}{\mathrm{Re}_\tau}\right), & \text{if } \mathrm{Re}_\tau > 1, \\ 0, & \text{if } \mathrm{Re}_\tau \le 1, \end{cases} \qquad \text{with } 0 \le \bar{\sigma} < 1. \tag{7}$$

Obviously, $\bar{\sigma} = 0$ means that no stabilization is added. The choice of σ_τ in (7) implies the following estimate which we need later in Sect. 6:

$$\sigma_\tau = \bar{\sigma}\frac{h_{\boldsymbol{w}}}{2\|\boldsymbol{w}\|_{L_\infty(\tau)}}\left(1 - \frac{1}{\mathrm{Re}_\tau}\right) \le \bar{\sigma}\frac{h_{\boldsymbol{w}}}{2\|\boldsymbol{w}\|_{L_\infty(\tau)}}\mathrm{Re}_\tau = \bar{\sigma}\frac{h_{\boldsymbol{w}}^2}{2\nu} \le \bar{\sigma}\frac{h_\tau^2}{2\nu}. \tag{8}$$

If one enumerates velocity unknowns first and pressure unknowns next, then the resulting discrete system has the 2 × 2-block form (3) with $C = 0$. The stabilization alters the (1,2)-block of the matrix making the latter not equal to the transpose of the (2,1)-block B. From the available analysis and results of numerical experiments we shall see that the perturbation of A caused by (6) affects the algebraic properties of (3).

3 Some Properties of LU Factorization for (3)

One can think about ILU factorization as a perturbation of exact LU factorization. Hence, it is instructive to have a first look at stability properties of the latter for non-symmetric saddle-point matrices as in (3). The results in this section summarize the analysis in [16, 17], where the reader can find full proofs and further details. The 2 × 2-block matrix from (3) is in general indefinite and if $C = 0$, its diagonal has zero entries. An LU factorization of such matrices often requires pivoting for stability reasons. However, exploiting the block structure and the properties of blocks A and C, one readily verifies that the LU factorization

$$\mathscr{A} = \begin{pmatrix} A & \widetilde{B}^T \\ B & -C \end{pmatrix} = \begin{pmatrix} L_{11} & 0 \\ L_{21} & L_{22} \end{pmatrix}\begin{pmatrix} U_{11} & U_{12} \\ 0 & -U_{22} \end{pmatrix} \tag{9}$$

with lower (upper) triangle matrices L_{11}, L_{22} (U_{11}, U_{22}) exists without pivoting, once $\det(A) \neq 0$ and there exist LU factorizations for the (1,1)-block

$$A = L_{11}U_{11}$$

and the Schur complement matrix $\widetilde{S} := BA^{-1}\widetilde{B}^T + C$ is factorized as

$$\widetilde{S} = L_{22}U_{22}.$$

Decomposition (9) then holds with $U_{12} = L_{11}^{-1}\widetilde{B}^T$ and $L_{21} = BU_{11}^{-1}$.

Assume A is positive definite. Then the LU factorization of A exists without pivoting. Its numerical stability (the relative size of entries in factors L_{11} and U_{11}) may depend on how large is the skew-symmetric part of A comparing to the symmetric part. More precisely, the following bound on the size of elements of L_{11} and U_{11} holds (see, e.g., (3.2) in [16]):

$$\frac{\||L_{11}||U_{11}|\|_F}{\|A\|} \le n\left(1 + C_A^2\right), \tag{10}$$

where $C_A := \|A_{\mathrm{S}}^{-\frac{1}{2}}A_{\mathrm{N}}A_{\mathrm{S}}^{-\frac{1}{2}}\|$, $A_{\mathrm{S}} = \frac{1}{2}(A + A^T)$, $A_{\mathrm{N}} = A - A_{\mathrm{S}}$. Here and further, $\|\cdot\|$ and $\|\cdot\|_F$ denote the matrix spectral norm and the Frobenius norm, respectively, and $|M|$ denotes the matrix of absolute values of M-entries.

If C is positive semi-definite, $\widetilde{B} = B$, and matrix B^T has the full column rank, then the positive definiteness of A implies that the Schur complement matrix $S := BA^{-1}B^T + C$ is also positive definite. However, this is not the case for a general block $\widetilde{B} \neq B$. The stabilization terms in the finite element method (6) produce the (1,2)-block $\widetilde{B}^T$ which is a *perturbation* of B^T. The positive definiteness of $\widetilde{S} := BA^{-1}\widetilde{B}^T + C$ and the stability of its LU factorization is guaranteed if the perturbation $E = \widetilde{B} - B$ is not too large [17]. In particular, $\widetilde{S}$ is positive definite if the perturbation matrix E is sufficiently small such that it holds

$$\kappa := (1 + C_A)\varepsilon_E c_S^{-\frac{1}{2}} < 1, \tag{11}$$

where $\varepsilon_E := \|A_{\mathrm{S}}^{-\frac{1}{2}} E^T\|$, $c_S := \frac{1}{2}\lambda_{\min}(S + S^T)$. Moreover, if $\widetilde{S}$ is positive definite, the factorization $\widetilde{S} = L_{22}U_{22}$ satisfies the stability bound similar to (10).

The following result about stability of LU factorization of (3) holds.

Theorem 1 *Assume matrix A is positive definite, C is positive semidefinite, and the inequality (11) holds with $\varepsilon_E = \|A_{\mathrm{S}}^{-\frac{1}{2}}(\widetilde{B} - B)^T\|$, $C_A = \|A_{\mathrm{S}}^{-\frac{1}{2}} A_{\mathrm{N}} A_{\mathrm{S}}^{-\frac{1}{2}}\|$, and $c_S = \frac{1}{2}\lambda_{\min}(S + S^T)$, then the LU factorization (9) exists without pivoting. The entries of the block factors satisfy (10) and the following bounds*

$$\frac{\||L_{22}||U_{22}|\|_F}{\|\widetilde{S}\|} \leq m \left(1 + \frac{(1 + \varepsilon_E c_S^{-\frac{1}{2}})C_A}{1 - \kappa}\right),$$

$$\frac{\|U_{12}\|_F + \|L_{21}\|_F}{\|U_{11}\|\|\widetilde{B}\|_F + \|L_{11}\|\|B\|_F} \leq \frac{m(1 + C_A)}{c_A}$$

with $c_A := \lambda_{\min}(A_{\mathrm{S}})$ and κ from (11).

The above analysis indicates that the LU factorization for (3) exists if the (1,1) block A is positive definite and the perturbation of the (1,2)-block is sufficiently small. The stability bounds depend on the constant C_A which measures the ratio of skew-symmetry for A, the ellipticity constant c_A, the perturbation measure ε_E and the minimal eigenvalue of the symmetric part of the unperturbed Schur complement matrix S. In Sect. 4, we show estimates of all these values for the finite element Oseen problem.

4 Properties of Matrices A and $\widetilde{S}$

The dependence of the critical constants c_A, C_A, ε_E and c_S from Theorem 1 on the problem and discretization parameters can be given explicitly. The analysis exploits the SUPG-FE origin of matrix A (matrix C is zero in the inf-sup FE method). Let

$\{\varphi_i\}_{1\le i\le n}$ and $\{\psi_j\}_{1\le j\le m}$ be bases of $\mathbb{V}_h$ and $\mathbb{Q}_h$, respectively. From the definition of matrix A and for arbitrary $v \in \mathbb{R}^n$ and corresponding $\boldsymbol{v}_h = \sum_{i=1}^{n} v_i\varphi_i$, one gets the following identity:

$$\begin{aligned}\langle Av, v\rangle = \alpha\|\boldsymbol{v}_h\|^2 + \nu\|\nabla\boldsymbol{v}_h\|^2 + \sum_{\tau\in T_h}\sigma_\tau\|\boldsymbol{w}\cdot\nabla\boldsymbol{v}_h\|_\tau^2 + \frac{1}{2}\int_{\Gamma_{\mathrm{N}}}(\boldsymbol{w}\cdot\boldsymbol{n})|\boldsymbol{v}_h|^2\,ds \\ -\frac{1}{2}\sum_{\tau\in T_h}((\operatorname{div}\boldsymbol{w})\boldsymbol{v}_h,\boldsymbol{v}_h)_\tau + \sum_{\tau\in T_h}\sigma_\tau(\alpha\boldsymbol{v}_h - \nu\Delta\boldsymbol{v}_h, \boldsymbol{w}\cdot\nabla\boldsymbol{v}_h)_\tau,\end{aligned} \tag{12}$$

where $\boldsymbol{n}$ is the outward normal on Γ_{N}. For a detailed discussion of the role each term from (12) plays in determining properties of matrix A, we refer to [16, 17]. Here we dwell on the last term in (12) due to the SUPG stabilization. The ν-dependent part of it vanishes for P1 finite element velocities, but not for most of inf-sup stable pressure–velocity pairs. Both analysis and numerical experiments below show that this term may significantly affect the properties of the matrix A, leading to unstable behavior of incomplete LU factorization unless the stabilization parameters are chosen sufficiently small.

The estimates for ellipticity and stability constants for A and $\widetilde{S}$ are summarized in Theorem 2. In order to formulate the theorem, we recall several well-known estimates. First, recall the Sobolev trace inequality

$$\int_{\Gamma_{\mathrm{N}}}|v|^2\,ds \le C_0\|\nabla v\|^2 \quad \forall\, v\in H^1(\Omega),\ v=0 \text{ on } \partial\Omega\setminus\Gamma_{\mathrm{N}}. \tag{13}$$

For any tetrahedron $\tau \in T_h$ and arbitrary $\boldsymbol{v}_h \in \mathbb{V}_h$, the following FE trace and inverse inequalities hold

$$\int_{\partial\tau}\boldsymbol{v}_h^2\,ds \le C_{\mathrm{tr}}h_\tau^{-1}\|\boldsymbol{v}_h\|_\tau^2,\ \|\nabla\boldsymbol{v}_h\|_\tau \le C_{\mathrm{in}}h_\tau^{-1}\|\boldsymbol{v}_h\|_\tau,\ \|\Delta\boldsymbol{v}_h\|_\tau \le \bar{C}_{\mathrm{in}}h_\tau^{-1}\|\nabla\boldsymbol{v}_h\|_\tau, \tag{14}$$

where the constants C_{tr}, C_{in}, $\bar{C}_{\mathrm{in}}$ depend only on the polynomial degree k and the shape regularity constant C_T from (4). In addition, denote by C_{f} the constant from the Friedrichs inequality:

$$\|\boldsymbol{v}_h\| \le C_{\mathrm{f}}\|\nabla\boldsymbol{v}_h\| \quad \forall\boldsymbol{v}_h \in \mathbb{V}_h, \tag{15}$$

and let $C_{\boldsymbol{w}} := \|(\boldsymbol{w}\cdot\boldsymbol{n})_-\|_{L^\infty(\Gamma_{\mathrm{N}})}$. We introduce the velocity mass and stiffness matrices M and K: $M_{ij} = (\varphi_i,\varphi_j)$, $K_{ij} = (\nabla\varphi_i,\nabla\varphi_j)$ and the pressure mass matrix M_p: $(M_p)_{ij} = (\psi_i,\psi_j)$.

Theorem 2 *Assume that $\boldsymbol{w} \in L^\infty(\Omega)$, problem and discretization parameters satisfy*

$$\begin{cases} C_{\boldsymbol{w}} C_{\mathrm{tr}} h_{\min}^{-1} \le \dfrac{\alpha}{4} \quad \textit{or} \quad C_{\boldsymbol{w}} C_0 \le \dfrac{\nu}{4}, \\ \|\operatorname{div} \boldsymbol{w}\|_{L^\infty(\Omega)} \le \dfrac{1}{4} \max\{\alpha, \nu C_f^{-1}\}, \\ \sigma_\tau \le \dfrac{h_\tau^2}{2\nu \bar{C}_{\mathrm{in}}^2} \left(1 + \dfrac{\alpha h_\tau^2}{\nu C_{\mathrm{in}}^2}\right) \quad \textit{and} \quad \sigma_\tau \le \dfrac{h_\tau}{4\|\boldsymbol{w}\|_{L^\infty(\tau)} C_{\mathrm{in}}} \quad \forall \tau \in T_h, \end{cases} \tag{16}$$

with constants defined in (13)–(15)*. Then the matrix A is positive definite and the constants c_A, C_A, c_S and ε_E can be estimated as follows:*

$$\begin{aligned} c_A &\ge \frac{1}{4} \lambda_{\min}(\alpha M + \nu K), \\ C_A &\le c\left(1 + \frac{\|\boldsymbol{w}\|_{L^\infty(\Omega)}}{\sqrt{\nu\alpha} + \nu + h_{\min}\alpha}\right), \\ c_S &\ge \frac{c\,\lambda_{\min}(M_p)}{(\nu + \alpha + \|\boldsymbol{w}\|_{L^\infty(\Omega)} + \|\operatorname{div}\boldsymbol{w}\|_{L^\infty(\Omega)})(1 + C_A^2)}, \\ \varepsilon_E &\le \left(\frac{\bar{\sigma}}{2\nu} \lambda_{\max}(M_p)\right)^{\frac{1}{2}}, \end{aligned} \tag{17}$$

where c is a generic constant independent of problem and discretization parameters.

Theorem 2 shows that matrices A and $\widetilde{S}$ are positive definite if conditions (16) on the parameters of the finite element method are satisfied. In this case, the matrix in (3) admits LU factorization without pivoting. The *first condition* in (16) is trivially satisfied with $C_{\boldsymbol{w}} = 0$ if $\Gamma_{\mathrm{N}} = \varnothing$ or the entire Γ_{N} is outflow boundary. However, we know that this is often not the case for the hemodynamics problems (see item (ii) in the introduction). On the other hand, small time step results in a large value of α which eases the first condition. The *second condition* is specific for finite element approximations. The given $\boldsymbol{w}$ approximates velocity field of an incompressible fluid and hence one intuitively expects $\|\operatorname{div}\boldsymbol{w}\|_{L^\infty(\Omega)}$ decreases for a refined grid (a rigorous proof may not be straightforward for lower order finite elements). However, the $\boldsymbol{w}$-divergence norm depends on fluid velocity field and may be large for ν small enough. Fortunately, for small Δt the second condition holds due to $\alpha \sim (\Delta t)^{-1}$. The *third condition* in (16) appears due to the stabilization included in the finite element formulation (6). The same or a similar condition on stabilization parameters appears in the literature on the analysis of SUPG stabilized methods for the linearized Navier–Stokes equations, see, e.g., [21]. The reason is that the positive definiteness of A is equivalent to the coercivity of the velocity part of the bilinear form from (6), which is crucial for deriving finite element method error estimates. Therefore, stabilization parameter design suggested in the literature typically satisfies $\sigma_\tau \lesssim h_\tau^2/\nu$ and $\sigma_\tau \lesssim h_\tau/\|\boldsymbol{w}\|_{L^\infty(\tau)}$ *asymptotically*, i.e. up to a scaling factor independent of discretization parameters. As follows from (8), the conditions (16) on the SUPG stabilization parameters (7) are valid if $\bar{\sigma} \le \min\{\bar{C}_{\mathrm{in}}^{-2}, \frac{1}{2}C_{\mathrm{in}}^{-1}\}$. Moreover, the value of the $\bar{\sigma}$ parameter from the SUPG term is crucial for the bound on ε_E which measures the discrepancy between B and $\widetilde{B}$. Thanks to (11) and Theorem 1 we see that ε_E has to be small enough to guarantee the stability of the factorization. Numerical results will support this observation. This puts additional implicit restrictions on $\bar{\sigma}$.

The domain anisotropy, see item (i) in the introduction, affects the lower bound for c_S in Theorem 2. The generic constant c in this bound depends on the inf-sup constant for $\mathbb{V}_h - \mathbb{Q}_h$ pair. Nevertheless, we shall see from experiments that the incomplete LU preconditioning in practice remains stable and efficient for stretched domains. Numerical experiments also show that the preconditioner has remarkable adaptivity properties with respect to different flow regimes, see item (iii) in the introduction. The bounds in Theorem 2 depend on $\boldsymbol{w}$ and ν, and hence on the Reynolds number. We observed in practice that the preconditioning remains stable over the range of Reynolds number and the fill-in adaptively increases or decreases in such a way that the number of iterations remains nearly the same.

5 Two-Parameter Threshold ILU Factorization

Incomplete LU factorizations of (3) can be written in the form $A = LU - E$ with an error matrix E. How small is the matrix E can be ruled by the choice of a threshold parameter $\tau > 0$. The error matrix E is responsible for the quality of preconditioning, see, for example, [15] for estimates on GMRES method convergence written in terms of $\|E\|$ and subject to a proper pre-scaling of A and the diagonalizability assumption. In general, the analysis of ILU factorization is based on the following arguments. For positive definite matrices A one can choose such a small τ that the product LU of its incomplete triangular factors L and U is also positive definite and so estimates from [11] can be applied to assess the numerical stability of the incomplete factorization: for $c_A = \lambda_{\min}(A_S)$, the sufficient condition is $\tau < c_A n^{-1}$. In practice, however, larger τ are used.

Theorem 2 shows that for certain flow regimes and for the choice of stabilization parameters the ellipticity constants c_A and c_S for A and S, respectively, approach zero. This may imply that the ILU factorization of (3) becomes unstable if possible at all. To ameliorate the performance of the preconditioning, we consider the two-parameter Tismenetsky–Kaporin variant of the threshold ILU factorization. The factorization was introduced and first studied in [14, 23, 24] for symmetric positive definite matrices and recently for non-symmetric matrices in [16, 17].

Given a matrix $A \in \mathbb{R}^{n\times n}$, the two-parameter factorization can be written as

$$A = LU + LR_u + R_\ell U - E, \tag{18}$$

where R_u and R_ℓ are strictly upper and lower triangular matrices, while U and L are upper and lower triangular matrices, respectively. Given two small parameters $0 < \tau_2 \le \tau_1$ the off-diagonal elements of U and L are either zero or have absolute values greater than τ_1, the absolute values of R_ℓ and R_u entries are either zero or belong to $(\tau_2, \tau_1]$; entries of the error matrix are of order $O(\tau_2)$. We refer to (18) as the ILU(τ_1, τ_2) factorization of A. In the particular case of $\tau_1 = \tau_2$, factorization ILU(τ_1,τ_2) is equivalent to the well-known ILUT(p, τ) dual parameter incomplete factorization [22] with $p = n$ (all elements passing the threshold criterion are kept in the factors). If no small pivots modification is done, the only differences between the algorithms (for $\tau_1 = \tau_2$ and $p = n$) are different scaling of pivots and row dependent scaling of threshold values. The two-parameter ILU factorization goes over a ILUT(n, τ) factorization: the fill-in of L and U is ruled by the first threshold parameter τ_1, while the quality of the resulting preconditioner is mainly defined by τ_2, once $\tau_1^2 \lesssim \tau_2$ holds. In other words, the choice $\tau_2 = \tau_1^2 := \tau^2$ may provide the fill-in of ILU(τ_1, τ_2) to be similar to that of ILUT(n, τ), while the convergence of preconditioned Krylov subspace method is better and asymptotically (for $\tau \to 0$) can be comparable to the one with ILUT(n, τ^2) preconditioner. For symmetric positive definite matrices this empirical advantages of ILU(τ_1, τ_2) are rigorously explained in [14], where estimates on the eigenvalues and K-condition number of $L^{-1}AU^{-1}$ were derived with $L^T = U$ and $R_\ell^T = R_u$. The price one pays is that computing L, U factors for ILU(τ_1, τ_2) is computationally more costly than for ILUT(n, τ_1), since intermediate calculations involve the entries of R_u. However, this factorization phase of ILU(τ_1, τ_2) is still less expensive than that of ILUT(n, τ_2). A pseudo-code of the row-wise ILU(τ_1, τ_2) factorization can be found in [16].

Analysis of the decomposition (18) of a general non-symmetric matrix is limited to simple estimate (2.5) from [10] applied to the matrix $(L + R_\ell)(U + R_u) = A + R_\ell R_u + E$. The lower bound for the pivots of the (18) factorization is the following:

$$|L_{ii}U_{ii}| \ge \min_{v\in\mathbb{R}^n} \frac{\langle (A + R_\ell R_u + E)v, v\rangle}{\|v\|^2} \ge c_A - \|R_\ell R_u\| - \|E\|, \qquad (19)$$

with the ellipticity constant c_A and the norms $\|R_\ell R_u\|$ and $\|E\|$ proportional to τ_1^2 and τ_2, respectively. Hence, we may conclude that the numerical stability of computing for $L^{-1}x$ and $U^{-1}x$ is ruled by the second parameter and the *square* of the first parameter, while the fill-in in both factors is defined by τ_1 rather than τ_1^2. The Oseen problem setup may be such that the estimates from Theorem 2 predict that the coercivity constant c_A and the ellipticity constant c_S are small. This increases the probability of the breakdown of ILUT(n, τ) factorization of the saddle-point matrix $\mathscr{A}$, and demonstrates the benefits of ILU(τ_1, τ_2) factorization.

The final important remark in this section is that in all computations we use the simple preprocessing of matrix $\mathscr{A}$ by the two-side scaling as described in [16].

6 Numerical Results

The model hemodynamic problem of interest is a blood flow in a right coronary artery. To set up the problem, we use the geometry recovered from a real patient coronary CT angiography. The 3D vessel is branching and is cut to embed in the box 6.5 cm × 6.8 cm × 5 cm, see Fig. 1. The diameter of the inlet cross-section is about

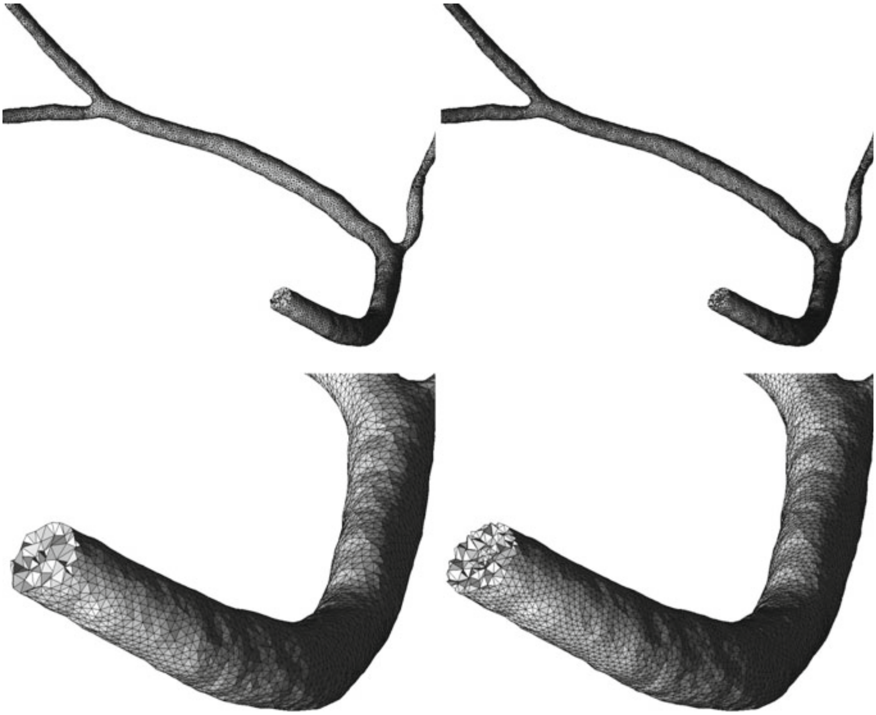


Fig. 1 The coarse (63k, left) and fine (120k, right) grids in the right coronary artery. The bottom figures zoom a part of the domain

0.27 cm. We generate two tetrahedral meshes using ANI3D package [2]. The meshes shown in Fig. 1 consist of 63k and 120k tetrahedra. The Navier–Stokes system (1) is integrated in time using a semi-implicit second order method with $\Delta t = 0.005$. This and the discretization with Taylor–Hood (P2-P1) finite elements result in a sequence of discrete Oseen problems (3). The algebraic systems have nearly 300k and 600k unknowns for the coarse and the fine meshes, respectively. Other model parameters are $\nu = 0.04\,\mathrm{cm}^2/\mathrm{s}$, $\rho = 1\,\mathrm{g/cm}$. We integrate the system over one cardiac cycle, which is 0.735 s. The inlet velocity waveform [13] shown in Fig. 2 defines the Poiseuille flow rate through the inflow cross-section. The figure shows the integral average of the normal velocity component over the inflow boundary. The vessel walls were treated as rigid and homogeneous Dirichlet boundary conditions for the velocity are imposed on the vessel walls. On all outflow boundaries we set the normal component of the stress tensor equal to zero. For the suitable choice of stabilization parameters, cf. below, the computed FE solutions are physically meaningful, see Fig. 3.

We study the performance of the ILU(τ) factorization for different values of discretization, stabilization, and threshold parameters. For numerical test we use the implementation of ILU(τ_1, τ_2) available in the open source software [1, 2]. The values

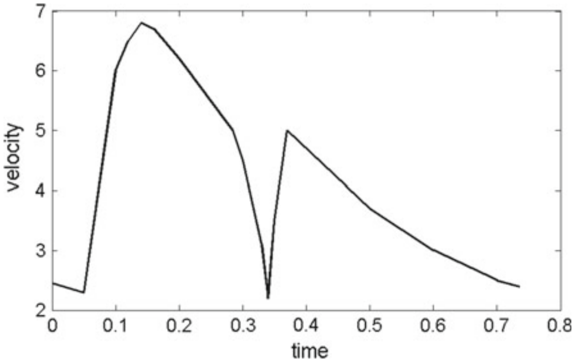


Fig. 2 The averaged velocity waveform on the inflow as a function of time in the right coronary artery

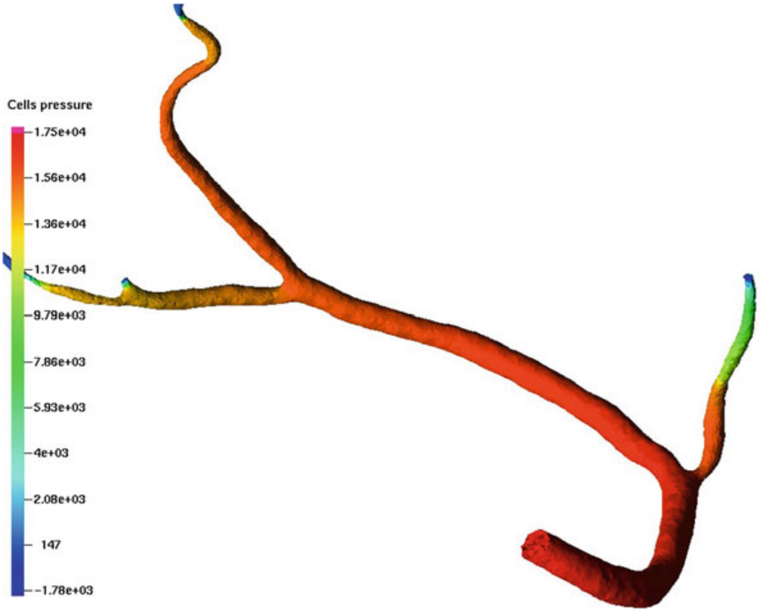


Fig. 3 The pressure distribution in the right coronary artery at time 0.15 s

of ILU thresholds $\tau_1 = 0.03$, $\tau_2 = 7\tau_1^2$ are taken from [16]. In that paper this design of threshold parameters was found to be close to optimal for a range of problems and fluid parameters. In all experiments we use BiCGstab method with the right preconditioner defined by the ILU(τ_1,τ_2) factorization.

Table 1 shows the total number of the preconditioned BiCGstab iterations #it, the total number of modifications of nearly zero pivots #pmod, the fill-in ratio and the CPU times (factorization time T_{build}, iteration time T_{it}, total solution CPU time

Table 1 The performance of ILU ($\tau_1 = 0.03$, $\tau_2 = 7\tau_1^2$) for right coronary artery. The number of iterations and pivot modifications and the solution stages times accumulated for 147 time steps

Mesh	$\bar{\sigma}$		fill$_{LU}$	pmod	#it	T_{build}	T_{it}	T_{CPU}
63k	0	Min	0.711	0	131	2.64	13.59	16.55
		Average	0.854	0	142.2	3.82	15.42	19.24
		Max	1.009	0	164	5.16	17.47	22.11
		Total	–	0	20908	562	2267	2829
63k	1/12	Min	0.711	0	125	2.63	13.03	16.10
		Average	0.838	0	138.0	3.65	14.84	18.49
		Max	0.980	0	156	4.85	22.62	26.42
		Total	–	0	20292	537	2182	2719
120k	0	Min	0.738	0	163	6.32	36.96	43.93
		Average	0.846	0	178.2	8.46	42.09	50.56
		Max	0.985	0	220	11.17	61.61	71.34
		Total	–	0	26209	1244	6188	7432
120k	1/12	Min	0.738	0	158	6.27	35.88	42.35
		Average	0.832	1	179.9	8.11	41.71	49.83
		Max	0.959	18	357	10.51	87.58	97.94
		Total	–	21	26446	1192	6132	7325

$T_{\text{CPU}} = T_{\text{build}} + T_{\text{it}}$) needed to perform 147 time steps. The fill-in ratio is defined by $\text{fill}_{LU} = (\text{nz}(L) + \text{nz}(U))/\text{nz}(A)$, where $\text{nz}(A) = \sum_{ij} \text{sign}|A_{ij}|$. On every time step, the Krylov subspace iterations are done until the initial residual is reduced by 10 orders of magnitude. The initial guess in the solver is the extrapolated solution from the previous time step. We generate sequences of the discrete Oseen problems (2) with ($\bar{\sigma} = 1/12$) and without ($\bar{\sigma} = 0$) SUPG-stabilization. In both cases, the 'quasi-optimal' choice of parameters τ_1, τ_2 leads to stable computations over the whole cardiac cycle. The total number of iterations depends on the mesh and appears to be very similar for both examples with and without stabilization. The total number of iterations is 20% larger for the fine grid, which should be expected for the preconditioner based on an incomplete factorization.

The time history of the statistics from Table 1 is shown in Figs. 4 and 5. It is interesting to note that the graph of the fill-in ratio for the LU-factors and the graph of the ILU factorization time repeat surprisingly well the waveform of the inflow velocity, see the two top plots in Figs. 4 and 5. This explains the rather modest variation of the iteration counts and CPU times per linear solve over the cardiac cycle, see the two bottom plots in Figs. 4 and 5. Note that the fill-in ratio $\text{fill}_{LU} < 1$ means that the number of non-zero elements in factors is less then in ILU(0), the commonly used ILU factorization by position. The fact that fill-in of the L and U blocks decreases or increases depending on the Reynolds number is the remarkable adaptive property of the two-parameter ILU preconditioner which makes it very competitive to other state-of-the-art preconditioners. The difference in otherwise similar performance of

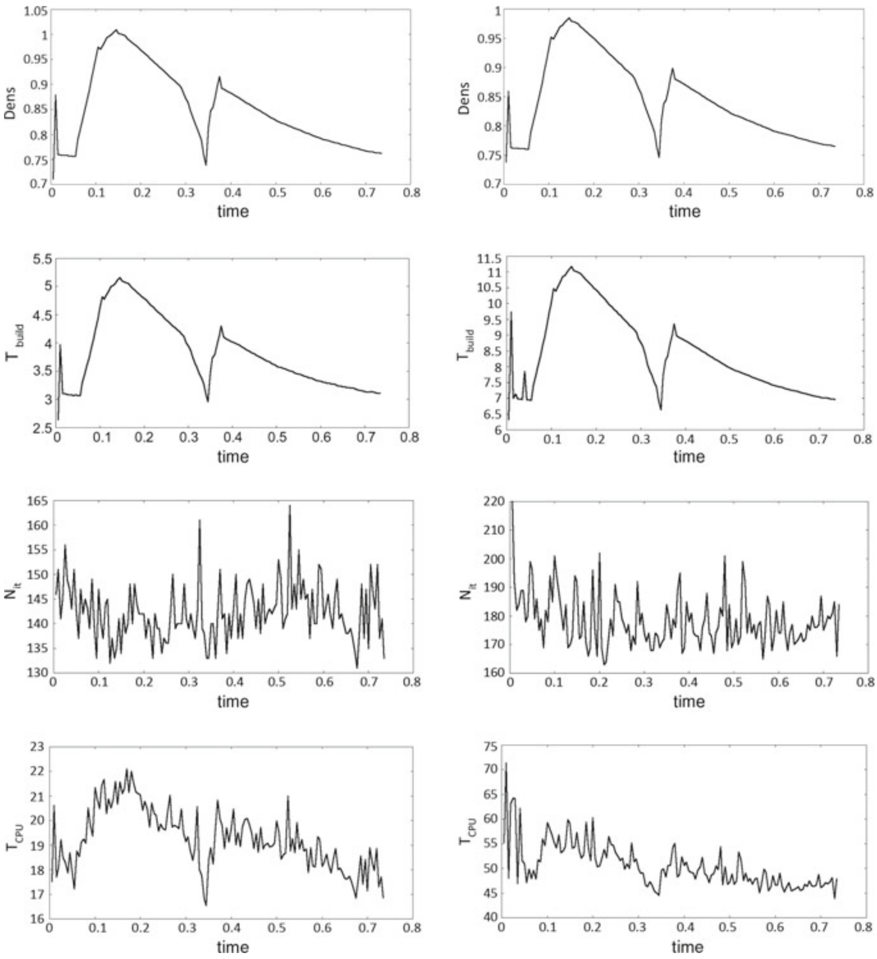


Fig. 4 Right coronary artery, computations on grid 63k (left) and grid 120k (right) without SUPG-stabilization and $\tau_1 = 0.03$: The plots (from top to bottom) show the density of the preconditioner (fill-in ratio), the time of ILU factorization, the number of BiCGStab iterations, the total CPU time of the linear system solution at each time step

linear solvers for the cases $\bar{\sigma} = 1/12$ and $\bar{\sigma} = 0$ is the following: For $\bar{\sigma} = 1/12$, when the maximum flow rate on the inlet is achieved, the number of iterations and times needed to build preconditioner increase essentially (approximately twice as much as average). This happens over a few time steps. In these cases when factorization is performed several small pivots occur and their modification is performed during the incomplete factorization.

In the second series of experiments, we demonstrate practical importance of restrictions (16) on σ_τ. The Theorems 1 and 2 state that the existence of exact stable LU factorization of $\mathscr{A}$ (almost) without pivoting is guaranteed for σ_τ small

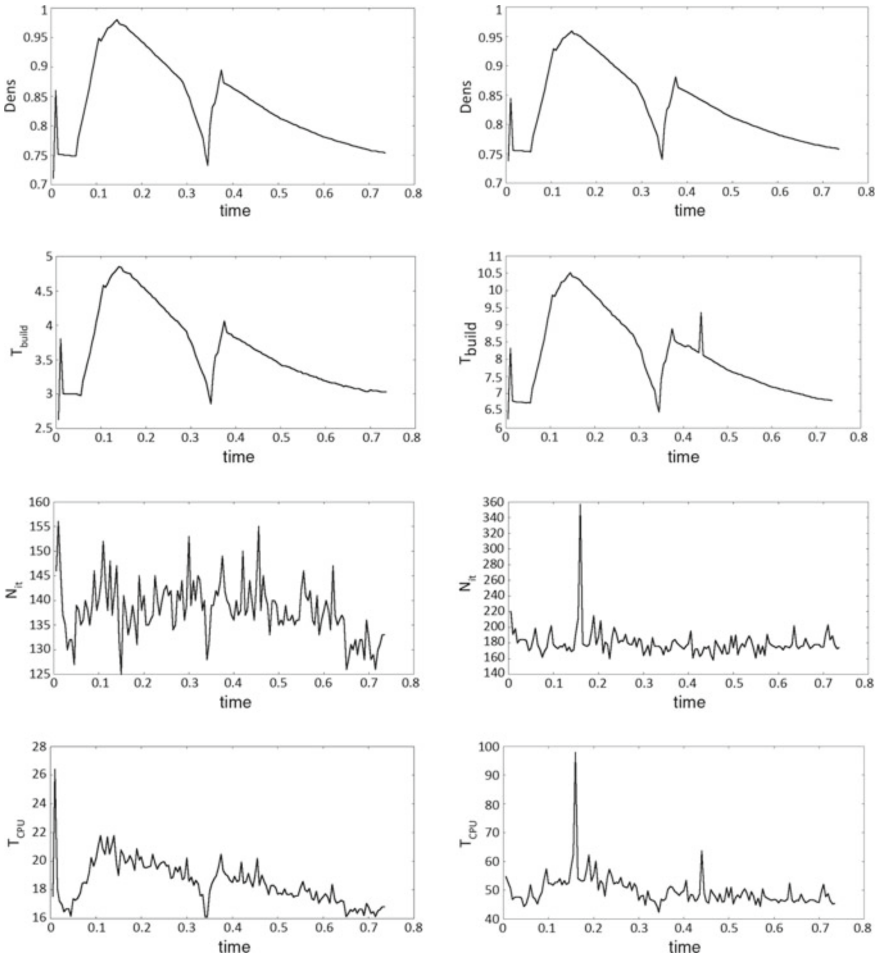


Fig. 5 Right coronary artery, computations on grid 63k (left) and grid 120k (right), SUPG-stabilization with $\bar{\sigma} = 1/12$ and $\tau_1 = 0.03$: The plots (from top to bottom) show the density of the preconditioner (fill-in ratio), the time of ILU factorization, the number of BiCGStab iterations, the total CPU time of the linear system solution at each time step

enough. The estimate (8) explains why σ_τ from (7) with $\bar{\sigma} \le \min\{\bar{C}_{\text{in}}^{-2}, \frac{1}{2}C_{\text{in}}^{-1}\}$ satisfies (16). The previous series of experiments show that for the stabilization parameter $\bar{\sigma} = 1/12$ the factorization is done on both meshes without pivot modifications even for the relatively large value of the threshold, $\tau_1 = 0.03$. Now we increase the value of the stabilization parameter and take $\bar{\sigma} = 1/6$. Table 2 reports on the performance of ILU(τ_1, $\tau_2 = 7\tau_1^2$) preconditioner for the sequence of the SUPG-stabilized Oseen systems generated on the coarse grid with $\bar{\sigma} = 1/6$. The choice of the threshold as small as $\tau_1 = 10^{-4}$ produces the factorization close to the exact one. Hence, the average number of BiCGstab iterations is only 8. Although no pivot modifications

Table 2 The performance of ILU ($\tau_1, \tau_2 = 7\tau_1^2$) for right coronary artery, $\bar{\sigma} = 1/6$, coarse mesh 63 k

τ_1		fill_{LU}	pmod	#it
0.0003	Min	5.978	0	7
	Average	8.466	1	12.2
	Max	11.206	12	135
	Total	–	16	1806
0.0001	Min	8.716	0	5
	Average	12.557	0	8.1
	Max	16.742	0	100
	Total	–	0	1198

Table 3 The performance of ILU ($\tau_1, \tau_2 = 7\tau_1^2$) for right coronary artery with different viscosities ν. The table shows values of τ_1 which allow to run the simulation for the complete cardiac cycle for different parameters $\bar{\sigma}$. '$\star$' means finite element solution blow-up, '–' means intractable systems for any possible τ_1

ν, \$\bar{\sigma}$ (cm^2/s)	0	1/96	1/48	1/24	1/12	1/6	1/3
0.040	0.03	0.03	0.03	0.03	0.03	0.03	0.003
0.025	$\star$	0.03	0.03	0.03	0.03	0.003	–

occurred, the fill-in ratio is unacceptably large and on some time steps the number of iterations may be large either. The observation that two-parameter ILU needs no pivoting with $\tau_1 = 10^{-4}$ suggests that the exact factorization is stable. For larger values of the threshold parameter, $\tau_1 = 3 \times 10^{-4}$, the fill-in ratio naturally decreases and the average number of BiCGstab iterations increases. Now, on two time steps the algorithm has to make 12 and 4 modifications of nearly zero pivots in order to avoid the breakdown. The pivot modifications causes the convergence slowdown, the maximum number of iterations in the Krylov subspace solver grows up to 135 iterations. Furthermore, on the *finer* grid certain Oseen systems with $\bar{\sigma} = 1/6$ can not be solved by the ILU-preconditioned BiCGstab iterations with any values of the threshold parameter which we tried.

We repeat the same simulations on the coarse grid, but for a smaller value of the viscosity coefficient, $\nu = 0.025\,\text{cm}^2/\text{s}$. For this viscosity, the simulation without SUPG stabilization fails (solution blows up at $t = 0.23$ s). Stabilization is necessary and adding it allows to obtain physiologically meaningful solution. At the same time, for larger parameter $\bar{\sigma}$ the linear systems are harder to solve. Indeed, $\bar{\sigma} = 1/6$ requires smaller threshold parameter τ_1, whereas $\bar{\sigma} = 1/3$ generates unsolvable systems, see Table 3. This experiment confirms that restrictions on $\bar{\sigma}$ come both from stability of the FE method and algebraic stability of the LU factorization. Both restrictions have to be taken into account when one decides about the choice of stabilization parameters.

Table 4 The performance of plain ILU (τ_1, τ_2) preconditioning versus reusing the same preconditioner over two time steps

	#it	T_{build}	T_{it}	T_{CPU}
Building preconditioner each time step	138	4.2	14.8	18.9
Building preconditioner every second time step	139	2.1	15.1	17.2

We also experiment with reusing ILU preconditioner over several time steps. This looks like a reasonable thing to try, since the time step is small and the system may not change too much from one time step to another one. Numerical results, however, show that the time cost of the setup phase of the preconditioner is small compared to the time needed by the Krylov subspace method to converge. Hence this strategy gives some time saving, but a moderate one. To illustrate this, we show in Table 4 the averaged data for the number of iterations per time step, the setup time needed to compute L and U factors, the time required by the Krylov subspace solver, and the total time, which is the sum of those two. The data is shown for the flow in the artery with the 63 K grid, $\nu = 0.04$, $\bar{\sigma} = 1/12$, $\tau_1 = 0.03$, $\tau_2 = 7\tau_1^2$. We see that reusing the same preconditioner over two time steps saves about 10% of the total computational time.

7 Conclusions

In this paper we studied the preconditioner based on elementwise incomplete two-parameter threshold ILU factorization of non-symmetric saddle-point matrices. The Krylov subspace solver with the preconditioner was used to simulate a blood flow in a right coronary artery reconstructed from a real patient coronary CT angiography. We tested the method for a range of physiological and discretization parameters. Several conclusions can be made: The solver efficiently handles typical features of hemodynamic applications such as geometrically stretched domains, variable flow regimes, and open boundary conditions with possible reversed flows. The preconditioner benefits from smaller time increments. One can reuse the preconditioner over several time steps, although for this particular application the benefit of doing this is modest, since the setup phase of the preconditioning is cheap compared to the time cost of iterations. A sequential version of the preconditioner is straightforward to implement for any type of finite elements and other discretizations once the matrix entries are available. For parallel computations it is natural to combine the ILU preconditioner with the additive Schwarz method. This is a subject of our further research.

Acknowledgements The experimental part of this work has been supported by the Russian Science Foundation through the grant 14-31-00024. The theoretical part (the contribution of the second author) was partially supported by NSF grant DMS1522252.

References

1. Advanced Numerical Instruments 2D. http://sourceforge.net/projects/ani2d
2. Advanced Numerical Instruments 3D. http://sourceforge.net/projects/ani3d
3. Benzi M, Deparis S, Grandperrin G, Quarteroni A (2016) Parameter estimates for the relaxed dimensional factorization preconditioner and application to hemodynamics. Comput Methods Appl Mech Engrg 300:129–145
4. Benzi M, Golub GH, Liesen J (2005) Numerical solution of saddle point problems. Acta Numer 14:1–137
5. Bodnár T, Galdi GP, Nečasová Š (eds) (2014) Fluid-structure interaction and biomedical applications. Birkhäuser, Basel
6. Chizhonkov EV, Olshanskii MA (2000) On the domain geometry dependence of the LBB condition. M2AN Math Model Numer Anal 34(5):935–951
7. Deparis S, Grandperrin G, Quarteroni A (2014) Parallel preconditioners for the unsteady Navier-Stokes equations and applications to hemodynamics simulations. Comput Fluids 92:253–273
8. Elman HC, Silvester DJ, Wathen AJ (2014) Finite elements and fast iterative solvers: with applications in incompressible fluid dynamics, 2nd edn. Oxford University Press, Oxford
9. Girault V, Raviart P-A (1979) Finite element approximation of the Navier-Stokes equations, vol 749. Lecture Notes in Mathematics. Springer, Berlin
10. Golub GH, Van Loan C (1979) Unsymmetric positive definite linear systems. Linear Algebra Appl 28:85–97
11. Golub GH, Van Loan CF (1996) Matrix computations, 3rd edn. Johns Hopkins University Press, Baltimore
12. Hou G, Wang J, Layton A (2012) Numerical methods for fluid-structure interaction—a review. Commun Comput Phys 12(2):337–377
13. Jung J, Hassanein A, Lyczkowski RW (2006) Hemodynamic computation using multiphase flow dynamics in a right coronary artery. Ann Biomed Engrg 34(3):393–407
14. Kaporin IE (1998) High quality preconditioning of a general symmetric positive definite matrix based on its $U^TU + U^TR + R^TU$-decomposition. Numer Linear Algebra Appl 5(6):483–509
15. Kaporin IE (2007) Scaling, reordering, and diagonal pivoting in ILU preconditionings. Russ J Numer Anal Math Model 22(4):341–376
16. Konshin IN, Olshanskii MA, Vassilevski YV (2015) ILU preconditioners for nonsymmetric saddle-point matrices with application to the incompressible Navier-Stokes equations. SIAM J Sci Comput 37(5):A2171–A2197
17. Konshin IN, Olshanskii MA, Vassilevski YV (2016) LU factorizations and ILU preconditioning for stabilized discretizations of incompressible Navier-Stokes equations. Numerical Analysis and Scientific Computing Preprint Seria 49, University of Houston
18. Nordsletten D, Smith N, Kay D (2010) A preconditioner for the finite element approximation to the arbitrary Lagrangian-Eulerian Navier-Stokes equations. SIAM J Sci Comput 32(2):521–543
19. Olshanskii MA, Tyrtyshnikov EE (2014) Iterative methods for linear systems: theory and applications. SIAM, Philadelphia
20. Passerini T, Quaini A, Villa U, Veneziani A, Canic S (2013) Validation of an open source framework for the simulation of blood flow in rigid and deformable vessels. Int J Numer Methods Biomed Engrg 29(11):1192–1213
21. Roos H-G, Stynes M, Tobiska L (1996) Numerical methods for singularly perturbed differential equations: convection-diffusion and flow problems. Springer, Berlin
22. Saad Y (2003) Iterative methods for sparse linear systems, 2nd edn. SIAM, Philadelphia
23. Suarjana M, Law KH (1995) A robust incomplete factorization based on value and space constraints. Int J Numer Methods Engrg 38(10):1703–1719
24. Tismenetsky M (1991) A new preconditioning technique for solving large sparse linear systems. Linear Algebra Appl 154(156):331–353

Martin's Problem for Volume-Surface Reaction-Diffusion Systems

Jeff Morgan and Vandana Sharma

Abstract We consider a question of global existence for two component volume-surface reaction-diffusion systems. The first of the components diffuses in a region, and then reacts on the boundary with the second component, which diffuses on the boundary. We show that if the first component is bounded a priori on any time interval, and the kinetic terms satisfy a generalized balancing condition, then both solutions exist globally. We also pose an open question in the opposite direction, and give some a priori estimates for associated m component systems.

Keywords Reaction-diffusion · Volume-surface · Systems · Global existence · A priori estimates

1 Introduction

We assume $n \geq 2$ and Ω is a bounded domain in R^n with smooth boundary M, such that Ω lies locally on one side of M. We denote η as the unit outward normal vector to Ω at points on M. Our initial interest is a volume-surface reaction-diffusion system having the form

J. Morgan (✉)
Department of Mathematics, University of Houston, Houston, TX, USA
e-mail: jjmorgan@central.uh.edu

V. Sharma
Department of Mathematics and Statistics, Arizona State University, Tempe, AZ, USA
e-mail: vsharm12@asu.edu

B. N. Chetverushkin et al. (eds.), *Contributions to Partial Differential Equations and Applications*, Computational Methods in Applied Sciences 47,
https://doi.org/10.1007/978-3-319-78325-3_19

$$\begin{aligned}
u_t &= d\triangle u && \text{on } \Omega \times (0,T),\\
d\frac{\partial u}{\partial \eta} &= f(u,v) && \text{on } M \times (0,T),\\
v_t &= e\triangle_M v + g(u,v) && \text{on } M \times (0,T),\\
u &= u_0 && \text{on } \overline{\Omega} \times \{0\}\\
v &= v_0 && \text{on } M \times \{0\},
\end{aligned} \tag{1}$$

where $d, e > 0$, $\triangle_M$ is the Laplace Beltrami operator, $f, g : R^2 \to R^2$ are smooth, and u_0 and v_0 are smooth, non negative and satisfy the compatibility condition

$$d\frac{\partial u_0}{\partial \eta} = f(u_0, v_0).$$

Systems of this type have appeared recently in the literature as so-called volume-surface reaction-diffusion systems (cf. [1, 2, 5, 8, 12]).

Throughout this work, we assume f and g are smooth and satisfy the quasi positivity condition $f(0,z), g(z,0) \geq 0$ for all $z \geq 0$. We have recently considered systems of this form in [9]. In that work, we obtained some estimates for solutions of linear scalar equations, and used them to prove that m component analogs of (1) have a unique, maximal, component-wise non negative solution, such that if the solution does not blow up in the sup norm in finite time, then the solution is a global solution. That is, $T = \infty$. We state this result below.

Theorem 1 *If f and g are smooth and satisfy the quasi positivity condition, then (1) has a unique, component-wise non negative, classical, maximal solution (u, v), and (u, v) is a global solution if u and v do not blow up in the sup-norm in finite time.*

Our primary interest in this work is a question that is analogous to a problem posed by Martin [3, 7] (referred to below). To this end, in additional to the smoothness and quasi positivity assumed above, we assume there are constants $a > 0$, $b \geq 0$, $L \in R$ and a natural number k so that

$$f(y,z), g(y,z) \leq L(y+z+1)^k \tag{2}$$

and

$$af(y,z) + g(y,z) \leq L(y+z) + b \tag{3}$$

for all $y, z \geq 0$. We refer to the former condition as a polynomial growth condition, and the latter condition as a balancing condition. We place no restrictions on the magnitude of L, k or b.

R. H. Martin posed a question for two component reaction-diffusion systems on bounded domains. The systems considered had the form

$$\begin{aligned}
u_t &= d\triangle u + f(u,v) && \text{on } \Omega \times (0,T),\\
v_t &= e\triangle v + g(u,v) && \text{on } \Omega \times (0,T),\\
\frac{\partial u}{\partial \eta} &= \frac{\partial v}{\partial \eta} = 0 && \text{on } M \times (0,T),\\
u &= u_0 && \text{on } \Omega \times \{0\},\\
v &= v_0 && \text{on } \Omega \times \{0\}.
\end{aligned} \tag{4}$$

In this setting, the initial data is only assumed to be bounded and non negative. But, the same assumptions are made on f and g as above. Martin asked whether solutions of (4) exist globally. Early work of Hollis et al. [3] proved that if (2) and (3) are satisfied, and one of the components of (4) is a priori sup norm bounded on finite time intervals, then the solution to (4) exists globally. This result made it a simple matter to obtain global existence results for a wealth of two component systems.

For example, consider the system

$$\begin{aligned}
u_t &= d\triangle u + \alpha v^2 - uv^4 && \text{on } \Omega \times (0,T),\\
v_t &= e\triangle v + \beta u + uv^4 - \alpha v^2 && \text{on } \Omega \times (0,T),\\
\frac{\partial u}{\partial \eta} &= \frac{\partial v}{\partial \eta} = 0 && \text{on } M \times (0,T),\\
u &= u_0 && \text{on } \Omega \times \{0\},\\
v &= v_0 && \text{on } \Omega \times \{0\},
\end{aligned} \tag{5}$$

where $\alpha, \beta \geq 0$ and $m \geq 2$, with $d, e > 0$ and u_0, v_0 are bounded and non negative. Well known local existence results (cf. [3]) imply the existence of a unique, maximal, component-wise non negative solution to (5). If we multiply the u_t equation by u, and note that $\alpha z - z^2 \leq \frac{\alpha^2}{4}$ for all $z \geq 0$, then

$$\begin{aligned}
\frac{\partial u^2}{\partial t} &\leq d\triangle u^2 + \frac{\alpha^2}{4} && \text{on } \Omega \times (0,T),\\
\frac{\partial u^2}{\partial \eta} &= 0 && \text{on } M \times (0,T),\\
u^2 &= (u_0)^2 && \text{on } \Omega \times \{0\}.
\end{aligned}$$

Therefore, from the comparison principle, u is bounded a priori in the sup norm on any finite time interval. In addition, it is a simple matter to see that f and g satisfy (2) and (3). Therefore, the results in [3] imply the solution to (5) exists globally.

A wonderful survey article of Pierre [7] outlines the development of a wealth of global existence work on systems of the form (4), under the assumptions of quasi positivity and (3), as well as analogous systems with more than two components.

Our results for (1) are somewhat different than those in [3], in that we can only obtain global existence for (1) when we know u is a priori sup norm bounded on any finite time interval, whereas the results in [3] can be applied with a priori sup norm bounds for either u or v. It our setting, it remains an open question whether a similar result can be obtained when we know v is a priori sup norm bounded on any finite time interval.

Our global existence result is given below.

Theorem 2 *Suppose f and g are smooth, quasi positive, and satisfy (2) and (3). If u is a priori sup norm bounded on any finite time interval, then (1) has a global solution.*

It is also possible to prove a boundedness result.

Theorem 3 *Suppose f and g satisfy the conditions of Theorem 2. If there exists $J > 0$ so that*

$$\|u(\cdot,\tau)\|_{\infty,\Omega},\ \|v\|_{1,M\times(\tau,\tau+1)} \le J \quad \textit{for all } \tau > 0,$$

then (1) has a global solution, and v is sup norm bounded on $M \times (0,\infty)$.

The proof of Theorem 2 is given in Sect. 2. The proof of Theorem 3 involves straight forward bootstrapping, and is similar in nature to the proof of Theorem 2. Similar arguments are given in [11], and we omit the proof in this work. Simple a priori estimates are give for m component systems in Sect. 3, and some examples and concluding remarks are given in Sect. 4. We extend the m component a priori estimates in Sect. 3 to global existence results in forthcoming work [6, 10].

2 Proof of Theorem 2

The work in [3] employed a duality argument, as does the work in [9, 11]. The duality arguments in [9, 11] involve the scalar boundary value problem

$$\begin{aligned} \phi_t &= -d\triangle\phi - \theta, && \text{on } \Omega\times(0,T),\\ \phi_t &= -e\triangle_M\phi - \hat{\theta}, && \text{on } M\times(0,T),\\ \phi &= 0 && \text{on } \overline{\Omega}\times\{T\}, \end{aligned} \tag{6}$$

where $d, e, T > 0$ are given in the previous section, and $\theta \in L_p(\Omega\times(0,T))$ and $\hat{\theta} \in L_p(M\times(0,T))$, with $p > 1$ and $\theta, \hat{\theta} \ge 0$. The following lemma is proved in [9] by utilizing fundamental results in [4], and is instrumental to our proof below.

Lemma 1 *Equation (6) has a unique, non-negative solution ϕ such that $\phi \in W_p^{2,1}(\Omega\times(0,T))$ and $\phi \in W_p^{2,1}(M\times(0,T))$. Furthermore, there exists a constant $C_{p,T} > 0$, independent of θ and $\hat{\theta}$, such that $\|\phi\|_{p,2,1,\Omega\times(0,T)}$, $\|\phi\|_{p,2,1,M\times(0,T)}$, $\|\frac{\partial\phi}{\partial\eta}\|_{p,M\times(0,T)}$, $\|\phi(\cdot,0)\|_{p,\Omega}$ and $\|\phi(\cdot,0)\|_{p,M}$ are all bounded by*

$$C_{p,T}\left[\|\theta\|_{p,\Omega\times(0,T)} + \|\hat{\theta}\|_{p,M\times(0,T)}\right].$$

Admittedly, much better estimates can be given for $\frac{\partial \phi}{\partial \eta}$ and $\phi(\cdot, 0)$, and those for $\frac{\partial \phi}{\partial \eta}$ can be found in [9]. However, the lemma above suffices for our purposes. We note that $C_{p,T}$ is non decreasing in T.

Lemma 2 *Suppose* f *and* g *satisfy quasi positivity conditions, (3), and* $0 < \tau_1 < \tau_2$ *so that the solution to (1) is defined for* $t < \tau_2$. *If* u is a priori sup norm bounded on $\Omega \times (\tau_1, \tau_2)$, $p > 1$, $\|\theta\|_{p,\Omega\times(0,T)}$, $\|\hat{\theta}\|_{p,M\times(0,T)} = 1$, and ϕ is the unique solution of *(6), then there is a constant* $N > 0$ *dependent on the sup norm bound for* u *on* $\Omega \times (\tau_1, \tau_2)$*, and* p*,* a*,* $\tau_2 - \tau_1$*,* L *and* b*, so that*

$$\int_M v(\cdot, t)\phi(\cdot, t) \le N \left(\int_M v(\cdot, \tau_1)\phi(\cdot, \tau_1) + 1 \right)$$

for all $\tau_1 < t < \tau_2$.

Proof Recall that $\phi, \theta, \hat{\theta}, u, v \ge 0$. Also,

$$\begin{aligned} \frac{d}{dt} \int_\Omega u\phi &= \int_\Omega (u_t\phi + u\phi_t) = \int_\Omega (\phi d \triangle u + u(-d\triangle\phi - \theta)) \\ &= \int_M \phi d \frac{\partial u}{\partial \eta} - \int_M ud\frac{\partial \phi}{\partial \eta} - \int_\Omega u\theta \\ &= \int_M \phi f(u,v) - \int_M ud\frac{\partial \phi}{\partial \eta} - \int_\Omega u\theta. \end{aligned}$$

Similarly,

$$\begin{aligned} \frac{d}{dt} \int_M v\phi &= \int_\Omega (v_t\phi + v\phi_t) = \int_M (\phi(e\triangle_M v + g(u,v)) + v(-e\triangle_M \phi - \hat{\theta})) \\ &= \int_M \phi g(u,v) - \int_M v\hat{\theta}. \end{aligned}$$

As a result,

$$\begin{aligned} &\int_\Omega au\theta + \int_M v\hat{\theta} + \frac{d}{dt}\left(\int_\Omega au\phi + \int_M v\phi \right) \\ &= \int_M \phi(af(u,v) + g(u,v)) - a\int_M ud\frac{\partial \phi}{\partial \eta} \\ &\le \int_M \phi\left[L(u+v) + b\right] - ad\int_M u\frac{\partial \phi}{\partial \eta} \\ &\le L\left(\int_\Omega au\phi + \int_M v\phi \right) + \int_M u\left(L\phi - ad\frac{\partial \phi}{\partial \eta} \right) + b\int_M \phi. \end{aligned}$$

Consequently, there is an $\hat{L} > 0$, dependent on the sup norm of u, L, d, b, a and $C_{p,T}$ so that

$$\frac{d}{dt}\left(\int_{\Omega} au\phi + \int_M v\phi\right) \le L\left(\int_{\Omega} au\phi + \int_M v\phi\right) + \hat{L}.$$

The result follows from Gronwall's inequality.

Now we are in a good position to prove Theorem 2.

Proof (of Theorem 2) From Theorem 1, (1) has a unique, maximal, component-wise non negative solution, such that if both components of the the solution do not blow up in the sup norm in finite time, then the solution is a global solution. Let $T > 0$ so that the solution of (1) exists. From our hypothesis, we know u is sup norm bounded on $\overline{\Omega} \times (0, T)$, so we will focus on showing that v is sup norm bounded.

We start by obtaining an $L_1(M)$ estimate for v. If we integrate and sum the first and third equations in (1), we obtain

$$\begin{aligned}\frac{d}{dt}\left(\int_{\Omega} au(\cdot,t) + \int_M v(\cdot,t)\right) &= \int_M (L(u(\cdot,t)+v(\cdot,t)) + b)\\ &\le K_1 + L\left(\int_{\Omega} au(\cdot,t) + \int_M v(\cdot,t)\right)\end{aligned}$$

for $0 < t < T$, where $K_1 > 0$ depends upon the sup norm bound for u, and the values L, b and $|M|$. As a result, Gronwall's inequality implies

$$\|v(\cdot,t)\|_{1,M} \le -\frac{K_1}{L} + \left(\frac{K_1}{L} + \int_{\Omega} au_0 + \int_M v_0\right)e^{Lt}. \tag{7}$$

Now we use duality to obtain $L_q(M \times (0,T))$ bounds on v, for $q > 1$. To this end, let $p > 1$, $\theta \in L_p(\Omega \times (0,T))$ and $\hat{\theta} \in L_p(M \times (0,T))$, with $\theta, \hat{\theta} \ge 0$, such that $\|\theta\|_{p,\Omega\times(0,T)}, \|\hat{\theta}\|_{p,M\times(0,T)} = 1$. If ϕ is the unique solution of (6) guaranteed by Lemma 1, and a, L and b are given in (3), then

$$\begin{aligned}\int_0^T\int_{\Omega} au\theta + \int_0^T\int_M v\hat{\theta} &= \int_0^T\int_{\Omega} au(-\phi_t - d\triangle\phi) + \int_0^T\int_M v(-\phi_t - e\triangle_M\phi)\\ &= \int_{\Omega} au_0\phi(\cdot,0) - d\int_0^T\int_M au\frac{\partial\phi}{\partial\eta} + \int_M v_0\phi(\cdot,0)\\ &\quad + \int_0^T\int_M \phi(af(u,v) + g(u,v))\\ &\le K_2C_{p,T} + \int_0^T\int_M \phi\,[L(u+v)+b]\\ &\le K_3C_{p,T} + L\int_0^T\int_M \phi v\\ &\le K_3C_{p,T} + LTN\left(\int_{\Omega} v_0\phi(\cdot,0) + 1\right),\end{aligned}$$

where $K_2, K_3 > 0$ depend upon a, d, T, $|M|$, $|\Omega|$, L, b and the sup norm bounds for u, u_0 and v_0, $C_{p,T}$ is given in Lemma 1, and N is given in Lemma 2. (Note, if T represents the upper bound for the maximal interval of existence, then the inequality above can also be obtained by working on $(0, \tau)$, with $0 < \tau < T$, and taking a limit as $\tau \to T^-$.) Therefore, from the non negativity of u and θ, we have

$$\int_0^T \int_M v\hat{\theta} \le K_3 C_{p,T} + LTN \left(\int_\Omega v_0 \phi(\cdot, 0) + 1 \right). \tag{8}$$

Consequently, if we note that $v, \hat{\theta} \ge 0$, $\hat{\theta}$ is arbitrary, and $\|\hat{\theta}\|_{p,M\times(0,T)} = 1$, then duality and the inequality above imply

$$\|v\|_{\frac{p}{p-1},M\times(0,T)} \le K_3 C_{p,T} + LTN \left(\int_M v_0 \phi(\cdot, 0) + 1 \right).$$

Since this inequality holds for every $p > 1$, we have a bound for $\|v\|_{q,M\times(0,T)}$ for each $q > 1$.

To finish, we use the polynomial bound on $g(u, v)$ to extend this estimate to a sup norm estimate for v. Since u is bounded, v is in every L_q space, and $g(u, v)$ satisfies (2), it follows that $g(u, v)$ is bounded above by a function in $L_p(M \times (0, T))$ for every $p > 1$. Consequently, we can conclude from the comparison principle and Lemma 1 that v is bounded above by a function that lives in $W_p^{2,1}(M \times (0, T))$ for every $p > 1$. By choosing p sufficiently large, and applying the Sobolev embedding theorem, we obtain a sup norm bound for v on $M \times (0, T)$. Therefore, (1) has a global solution.

3 *m*-Component Systems

In this section, we consider an m component version of (1), along with a generalization of the quasi positivity and balancing conditions given in Sect. 1. Let R_+^m denote the non negative orthant in R^m, and assume $\mathscr{K}, \mathscr{L} \subseteq \{1, \ldots, m\}$ such that

1. Either $\mathscr{K} = \emptyset$ or there exists $i_{\mathscr{J}} \in 1, \ldots, m$ so that $\mathscr{K} = \{1, \ldots, i_{\mathscr{J}}\}$;
2. $\mathscr{K} \cup \mathscr{L} = \{1, \ldots, m\}$;
3. $\mathscr{K} \cap \mathscr{L} = \emptyset$.

We consider the system

$$\begin{cases} \dfrac{\partial u_i}{\partial t} = d_i \triangle u_i, & \Omega \times (0, T), i \in \mathscr{K}, \\ d_i \dfrac{\partial u_i}{\partial \eta} = F_i(u), & M \times (0, T), i \in \mathscr{K}, \\ \dfrac{\partial u_i}{\partial t} = d_i \triangle_M u_i + F_i(u), & M \times (0, T), i \in \mathscr{L}, \\ u_i = u_{0_i}, & \overline{\Omega} \times \{0\}, i \in \mathscr{K}, \\ u_i = u_{0_i}, & M \times \{0\}, i \in \mathscr{L}, \end{cases} \tag{9}$$

where $F : R^m \to R^m$ is smooth, and satisfies the quasi positivity condition

$$F_i(z) \geq 0 \quad \text{for all } z \in R^m_+ \text{ with } z_i = 0. \tag{10}$$

In addition, we assume $d_i > 0$ for all i, and the initial data u_0 is smooth, component wise non negative, and satisfies the compatibility condition

$$d_i \frac{\partial u_{0_i}}{\partial \eta} = F_i(u_0) \quad \text{on } M \text{ for } i \in \mathscr{K}.$$

Finally, we assume a generalization of the balancing condition in (3). More specifically, we assume there are scalars $c_i > 0$ for $i = 1, \ldots, m$, and $\alpha, \beta > 0$ such that

$$\sum_{j=1}^{m} c_i F_i(z) \leq \alpha \sum_{j=1}^{m} z_i + \beta \quad \text{for all } z \in R^m_+. \tag{11}$$

Note that (9) can take different forms, depending upon the sets $\mathscr{K}$ and $\mathscr{L}$. For example, if $\mathscr{K} = \emptyset$, then (9) takes the form of a standard reaction diffusion system set on a manifold without boundary. In this setting, nearly all of the results in [7] for the case of homogeneous Neumann boundary conditions can be proved for (9). If $\mathscr{L} = \emptyset$, then (9) is a mass transport type system. We are currently completing global existence results for systems of this type in [10].

The result below is a consequence of our earlier work in [9], and only depends upon the quasi positivity condition (10).

Theorem 4 *If F is smooth and satisfies the quasi positivity condition (10), then (9) has a unique, componentwise nonnegative, classical, maximal solution u, and u is a global solution if u does not blow up in the sup-norm in finite time.*

In this section, we obtain L_1 a priori bounds for solutions of (9) under the assumption of quasi positivity and (11). In the remainder of this section, we assume $T_{max} > 0$ is the maximal time of existence of solutions of (9). Of course, it is possible that $T_{max} = \infty$.

Lemma 3 *There is a function $C_1 \in C(R_+, R_+)$ such that the unique maximal solution u of (9) satisfies*

$$\|u_i\|_{1,\Omega\times(0,\tau)}, \|u_j\|_{1,M\times(0,\tau)} \leq C_1(\tau)$$

for all $i \in \mathscr{K}$, $j \in \{1, \ldots, m\}$ and $0 \leq \tau < T_{max}$.

Proof For ease of exposition, we assume $c_i = 1$ for all $i \in 1, \ldots, m$. As a result, from (11), we have

$$\sum_{i=1}^{m} F_i(z) \leq \alpha \sum_{i=1}^{m} z_i + \beta$$

for all $z \in R^m_+$. Let $0 \leq \tau < T_{max}$. It is a simple matter to show that (9) and (11) imply there exists $L > 0$ dependent on the integral of the initial data, $|\Omega|$, $|M|$ and β, and independent of t, such that

$$\sum_{i \in \mathscr{K}} \int_\Omega u_i(\cdot, t) + \sum_{i \in \mathscr{L}} \int_M u_i(\cdot, t) \leq \alpha \int \left(\sum_{i \in \mathscr{K}} \int_\Omega u_i + \sum_{i=1}^m \int_M u_i \right) + L \qquad (12)$$

for all $0 < t < T_{\max}$. In order to get our result from Eq. (12), we need an estimate for $\|u_i\|_{1,M\times(0,t)}$ for $i \in \mathscr{K}$. To this end, let $\sigma > \frac{d_1 \alpha}{d_i}$ for all $i \in \mathscr{K}$, let ϕ_0 be smooth and positive on $\overline{\Omega}$ such that $d_1 \frac{\partial \phi_0}{\partial \eta} = 1 + \sigma\phi_0$, and let ϕ be the unique smooth positive solution to

$$\begin{aligned} \phi_t &= -d_1 \triangle \phi, && \text{on } \Omega \times (0, \tau), \\ d_1 \frac{\partial \phi}{\partial \eta} &= 1 + \sigma \phi, && \text{on } M \times (0, \tau), \\ \phi &= \phi_0, && \text{on } \Omega \times \tau. \end{aligned} \qquad (13)$$

Define

$$\theta_i = \phi_t + d_i \triangle \phi \quad \text{on } \Omega \times (0, \tau) \text{ for } i \in \mathscr{K}$$

and

$$\tilde{\theta}_j = \phi_t + d_i \triangle_M \phi \quad \text{on } M \times (0, \tau) \text{ for } j \in \mathscr{L}.$$

Note that $\theta_1 \equiv 0$, and θ_i and $\tilde{\theta}_j$ are not necessarily non negative. However, they are sup norm bounded. Straightforward (but tedious) integration leads to

$$\sum_{i \in \mathscr{K}} \int_0^\tau \int_\Omega u_i \theta_i + \sum_{i \in \mathscr{L}} \int_0^\tau \int_M u_j \tilde{\theta}_j \leq I + II,$$

where

$$I = \sum_{i=1}^m \int_0^\tau \int_M (\alpha u_i + \beta) \, \phi$$

and

$$II = -\sum_{i \in \mathscr{K}} \int_0^\tau \int_M u_i \frac{d_i}{d_1} (1 + \sigma \phi) \, .$$

Applying the boundedness of θ_i and $\tilde{\theta}_i$, and the choice of σ, we conclude there exists $\gamma > 0$ such that

$$\sum_{i \in \mathscr{K}} \int_0^\tau \int_M u_i \leq \gamma \left(\sum_{i \in \mathscr{K}} \int_0^\tau \int_\Omega u_i + \sum_{j \in \mathscr{L}} \int_0^\tau \int_M u_i + 1 \right) . \qquad (14)$$

Combining this with (12) gives the existence of $\tilde{\gamma} > 0$ such that

$$\sum_{i\in\mathscr{K}} \int_{\Omega} u_i(\cdot,t) + \sum_{i\in\mathscr{L}} \int_{M} u_i(\cdot,t) \le a \int_0^t \left(\sum_{i\in\mathscr{K}} \int_{\Omega} u_i + \sum_{i\in\mathscr{L}} \int_{M} u_i \right) + L.$$

Applying Gronwall's inequality guarantees

$$\int_0^t \left(\sum_{i\in\mathscr{K}} \int_{\Omega} u_i + \sum_{i\in\mathscr{L}} \int_{M} u_i \right)$$

is bounded by a continuous function of t. Combining this information with (14) gives the result.

4 Examples and Concluding Remarks

We start with a very elementary example. To this end, consider the system

$$\begin{aligned} u_t &= d\triangle u && \text{on } \Omega \times (0,T), \\ d\frac{\partial u}{\partial \eta} &= -uv^k && \text{on } M \times (0,T), \\ v_t &= e\triangle_M v + uv^k && \text{on } M \times (0,T), \\ u &= u_0 && \text{on } \overline{\Omega} \times \{0\}, \\ v &= v_0 && \text{on } M \times \{0\}, \end{aligned} \tag{15}$$

where k is a natural number, and $d, e > 0$. It is a simple matter to see that the system satisfies the conditions of Theorem 2. Furthermore, u is easily a priori sup norm bounded on any finite time interval. Consequently, it follows that solutions to (15) exist globally.

Interestingly, if the situation is changed only slightly, then there are no known results. To this end, consider the system

$$\begin{aligned} u_t &= d\triangle u && \text{on } \Omega \times (0,T), \\ d\frac{\partial u}{\partial \eta} &= u^k v && \text{on } M \times (0,T), \\ v_t &= e\triangle_M v - u^k v && \text{on } M \times (0,T), \\ u &= u_0 && \text{on } \overline{\Omega} \times \{0\}, \\ v &= v_0 && \text{on } M \times \{0\}. \end{aligned} \tag{16}$$

This time, we can see that the hypothesis of Theorem 2 is satisfied, but there is no apparent a priori sup norm bound for u. It is a simple matter to show that v is a priori bounded in the sup norm on any finite time interval, but there does not seem to be a way use this to obtain a sup norm bound for u, and global existence.

A natural question is whether there are nontrivial systems which satisfy the hypothesis of Theorem 2. One such example is the system

$$\begin{aligned} u_t &= d\triangle u && \text{on } \Omega \times (0,T), \\ d\frac{\partial u}{\partial \eta} &= \alpha v^2 - uv^4 && \text{on } M \times (0,T), \\ v_t &= e\triangle_M v + \beta u + uv^4 - \alpha v^2 && \text{on } M \times (0,T), \\ u &= u_0 && \text{on } \overline{\Omega} \times \{0\}, \\ v &= v_0 && \text{on } M \times \{0\}. \end{aligned} \tag{17}$$

This is similar to (5) given in the introduction. Note that $w = u^2$ satisfies

$$\begin{aligned} w_t &\le d\triangle w && \text{on } \Omega \times (0,T), \\ d\frac{\partial w}{\partial \eta} &\le \frac{\alpha^2}{4} && \text{on } M \times (0,T), \\ w &= (u_0)^2 && \text{on } \Omega \times (0,T). \end{aligned}$$

As a result, u is a priori sup norm bounded on any finite time interval. Therefore, global existence follows from Theorem 2.

References

1. Egger H, Fellner K, Pietschmann J-F, Tang BQ Analysis and numerical solution of coupled volume-surface reaction-diffusion systems with application to cell biology. arXiv:1511.00846 [math.NA], submitted
2. Fellner K, Rosenberger S, Tang BQ (2016) Quasi-steady-state approximation and numerical simulation for a volume-surface reaction-diffusion system. Commun Math Sci 14(6):1553–1580
3. Hollis SL, Martin RH, Pierre M (1987) Global existence and boundedness in reaction-diffusion systems. SIAM J Math Anal 18(3):744–761
4. Ladyzhenskaya OA, Solonnikov VA, Uraltseva NN (1968) Linear and quasi-linear equations of parabolic type, vol 23. Translations of mathematical monographs. AMS, Providence, RI
5. Madzvamuse A, Chung AHW, Venkataraman C (2015) Stability analysis and simulations of coupled bulk-surface reaction-diffusion systems. Proc R Soc A 471(2175):20140546
6. Morgan J, harma V Global existence for reaction-diffusion systems with dynamic and mass transport boundary conditions (in preparation)
7. Pierre M (2010) Global existence in reaction-diffusion systems with control of mass: a survey. Milan J Math 78(2):417–455
8. Rätz A, Röger M (2012) Turing instabilities in a mathematical model for signaling networks. J Math Biol 65(6–7):1215–1244

9. Sharma V, Morgan J Global existence of coupled reaction-diffusion systems with mass transport type boundary conditions. SIAM J Math Anal (submitted April 2015, revised December 2015)
10. Sharma V, Morgan J Global existence of solutions to reaction diffusion systems with Wentzell type boundary conditions (in preparation)
11. Sharma V, Morgan J Uniform bounds for solutions to volume-surface reaction diffusion systems. arXiv:1512.08765 [math.AP]
12. Tang BQ, Fellner K, Latos E Well-posedness and exponential equilibration of a volume-surface reaction-diffusion system with nonlinear boundary coupling. arXiv:1404.2809 [math.AP], submitted

A Posteriori Error Estimates for the Electric Field Integral Equation on Polyhedra

Ricardo H. Nochetto and Benjamin Stamm

Abstract We present a residual-based a posteriori error estimate for the Electric Field Integral Equation (EFIE) on a bounded polyhedron Ω with boundary Γ. The EFIE is a variational equation formulated in $\boldsymbol{H}_{\text{div}}^{-1/2}(\Gamma)$. We express the estimate in terms of L^2-computable quantities and derive global lower and upper bounds (up to oscillation terms).

Keywords Electric field integral equation · A posteriori error estimation

1 Introduction

The Electric Field Integral Equation (EFIE) describes the scattering of electromagnetic waves on a perfectly conducting obstacle Ω with surface Γ, in our case a polyhedron. Assuming a time-harmonic dependence, the Stratton-Chu representation formula expresses the electric field E in terms of a surface potential as

$$E(\boldsymbol{x}) = E^{inc}(\boldsymbol{x}) + \int_\Gamma \left(G_k(\boldsymbol{x},\boldsymbol{y})\boldsymbol{u}(\boldsymbol{y}) + \frac{1}{k^2}\, \mathbf{grad}_{\Gamma,\boldsymbol{x}}\, G_k(\boldsymbol{x},\boldsymbol{y})\, \operatorname{div} \boldsymbol{u}(\boldsymbol{y}) \right) \mathrm{d}\sigma(\boldsymbol{y}),$$

where k denotes the wave-number, $E^{inc}(\boldsymbol{x})$ is the given incident wave that is scattered on Γ and $G_k(\boldsymbol{x},\boldsymbol{y})$ denotes the fundamental solution of the Helmholtz operator. We assume that the wave-number is real, of moderate size (relative to the length of the scatterer Ω) and does not coincide with an interior eigenvalue. Then, invoking the boundary condition that the tangential component of the total electric field E vanishes

R. H. Nochetto
Department of Mathematics, University of Maryland, College Park, MD 20742, USA
e-mail: rhn@math.umd.edu

B. Stamm (✉)
Mathematics Department, Center for Computational Engineering,
RWTH Aachen University, Schinkelstr. 2, 52062 Aachen, Germany
e-mail: best@mathcces.rwth-aachen.de

B. N. Chetverushkin et al. (eds.), *Contributions to Partial Differential Equations and Applications*, Computational Methods in Applied Sciences 47,
https://doi.org/10.1007/978-3-319-78325-3_20

on the surface Γ, as corresponds to Ω being perfectly conducting, the EFIE consists of seeking the surface current $\boldsymbol{u} \in \boldsymbol{H}_{\text{div}}^{-\frac{1}{2}}(\Gamma)$ such that for all $\boldsymbol{x} \in \Gamma$

$$\int_{\Gamma} \left(G_k(\boldsymbol{x}, \boldsymbol{y}) \boldsymbol{u}(\boldsymbol{y}) + \frac{1}{k^2} \, \mathbf{grad}_{\Gamma,\boldsymbol{x}} \, G_k(\boldsymbol{x}, \boldsymbol{y}) \, \operatorname{div} \boldsymbol{u}(\boldsymbol{y}) \right) \mathrm{d}\sigma(\boldsymbol{y}) = -\boldsymbol{\gamma}_{\parallel}(E^{inc})(\boldsymbol{x}),$$

where $\boldsymbol{H}_{\text{div}}^{-\frac{1}{2}}(\Gamma)$ is the space of traces of $\boldsymbol{H}(\mathbf{curl}, \Omega)$ functions that are rotated by a right angle on the surface and $\boldsymbol{\gamma}_{\parallel}$ denotes the tangential trace onto Γ.

The Combined Field Integral Equation (CFIE) (see, e.g., [5, 13, 26, 38, 43]) can be used if the wave-number corresponds to an interior eigenvalue; the functional analysis is, however, not fully developed in this case so that a rigorous a posteriori error analysis is not possible for this formulation. Therefore, Regularized Combined Field Integral Equations have been proposed by [10, 11] which embed a robust formulation with respect to the wave-number in a well defined functional analysis setting. To keep things as simple as possible, we start with considering the EFIE formulation.

Computing approximations of the EFIE by means of the Boundary Element Method (BEM), namely using a Galerkin approach based on the variational formulation of the EFIE, is expensive due to the dense matrix structure of the ensuing linear system. Despite the existence of fast solvers for the Galerkin system such as the fast multipole method, see, e.g., [33, 34, 45], and the cluster methods by [35, 36, 42, 46], it is still crucial to locate the degrees of freedom efficiently, namely in regions of low regularity of the solution $\boldsymbol{u}$.

Since $\boldsymbol{u} \in \boldsymbol{H}_{\text{div}}^{-\frac{1}{2}}(\Gamma)$, $\boldsymbol{u}$ exhibits in general rather low regularity and, as a consequence, *a priori* estimates show extremely low convergence rates for quasi-uniform mesh refinements; see [24, 37]. In contrast, adaptive refinement techniques, based on *a posteriori* error estimates, exploit much weaker regularity of $\boldsymbol{u}$ in a nonlinear Sobolev scale and allow for optimal error decay in terms of degrees of freedom in situations where quasi-uniform meshes are suboptimal. The design and analysis of a posteriori error estimators is, however, problem dependent; we refer to [22, 40] for an account of the theory of adaptive finite element methods in the energy norm for linear second order elliptic partial differential equations in polyhedra.

The a posteriori error analysis for BEM started in 1995 [21] and has been developed ever since [14–19, 28, 41]. The corresponding theory of adaptivity is much more recent [2, 20, 30–32]. Contributions in the framework of FEM-BEM coupling can be found in [39, 48]. It seems that this paper presents the first a posteriori error analysis for electromagnetic scattering problems via EFIE.

For integral equations, additional difficulties arise since the residual typically lies in a Sobolev space with fractional index that is possibly also negative, as in the present case. Since such norms are difficult to compute in practice, this imposes additional challenges to the residual based approach of a posteriori error estimates.

In this paper we develop nevertheless a residual based a posteriori error estimator for the EFIE on polyhedra, and prove upper and lower global bounds. Residual based estimators are especially attractive due to their simplicity of derivation and computation, but they involve interpolation constants which can at best be estimated. Alternative estimators have been proposed, mostly for elliptic problems defined in Ω, at the expense of their simplicity; we believe that our approach can be extended to those estimators as well. We derive computable L^2-integrable quantities to estimate the error of the BEM measured in the $\boldsymbol{H}_{\mathrm{div}}^{-\frac{1}{2}}(\Gamma)$ norm, which is the natural norm for EFIE. We therefore avoid evaluating fractional Sobolev norms.

For proving well-posedness of the exact solution and developing a priori error estimates it is important to decompose both the exact solution and test function using a Helmholtz decomposition as has been shown in [12, 37]. In contrast, to derive *a posteriori* error estimates, it is crucial to decompose the *test* function according to a regular decomposition which extends the Helmholtz decomposition; see [23] for $H(\mathrm{div};\,\Omega)$. It is also worth mentioning that the idea of splitting the test function according to a Helmholtz decomposition goes back to [1] for mixed FEM and [3] for eddy current computations in $H(\mathrm{curl};\,\Omega)$.

This paper is organized as follows. In Sect. 2 we recall the necessary functional analysis from [6, 7, 9, 12] in order to derive a posteriori error estimates for the EFIE. We also present and study a Clément type interpolation operator for the Raviart-Thomas space, based on ideas from [4]. We discuss the EFIE integral equation in Sect. 3, and derive global upper and lower a posteriori error estimates in Sect. 4. Section 5 is finally left for conclusions.

2 Functional Spaces and Differential Operators

The functional analysis framework developed in [6, 7] will be used in this work. In this section we give a short introduction to the functional spaces and differential operators used in the following sections. However, for a detailed and thorough overview we refer to [6, 7, 9, 12]. Let us note that references [9, 12] deal with non-smooth Lipschitz surfaces, thus the theory is also valid for polyhedra, and covers therefore a more general framework. However, we restrict our theory to polyhedral surfaces.

2.1 Spaces, Norms and Trace Operators

Let Ω be a bounded polyhedron in $\mathbb{R}^3$, and denote its boundary by Γ and its different faces by Γ_j, $j = 1, \ldots, N_F$. The exterior part Ω^+ is defined by $\Omega^+ = \mathbb{R}^3 \setminus \overline{\Omega}$. Let $\boldsymbol{n}(\boldsymbol{x}), \boldsymbol{x} \in \Gamma$, denote the outer unit normal to the surface Γ, which is piecewise constant on Γ. We also indicate by $e_{ij} = \partial\Gamma_i \cap \partial\Gamma_j$ the *edges* of Γ and by $\boldsymbol{\tau}_{ij}$ the unit vectors parallel to e_{ij}, with its orientation fixed but arbitrary. If $\boldsymbol{n}_i = \boldsymbol{n}|_{\Gamma_i}$, we further define

$$\boldsymbol{\tau}_i = \boldsymbol{\tau}_{ij} \times \boldsymbol{n}_i, \quad \boldsymbol{\tau}_j = \boldsymbol{\tau}_{ij} \times \boldsymbol{n}_j$$

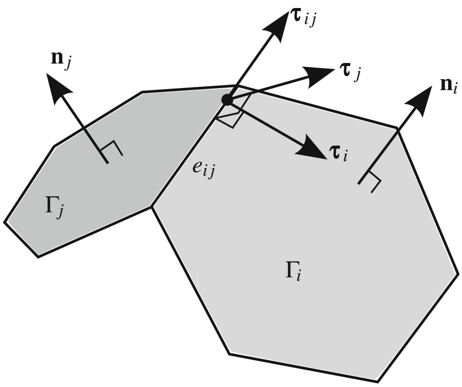


Fig. 1 Local coordinate systems around an edge $e_{ij} = \partial\Gamma_i \cap \partial\Gamma_j$

to be unit vectors lying on the supporting planes of Γ_i and Γ_j; see Fig. 1 for an illustration.

On Γ, we define the space of square integrable tangential fields

$$\boldsymbol{L}^2_t(\Gamma) = \{\boldsymbol{v} \in [L^2(\Gamma)]^3 \,|\, \boldsymbol{v}\cdot\boldsymbol{n} = 0 \quad \text{a.e.}\}.$$

Moreover, we let $H^s(\Gamma)$ and $\boldsymbol{H}^s(\Gamma) = [H^s(\Gamma)]^3$, with $s \in [-1, 1]$, denote the standard Sobolev spaces of complex-valued scalar and vector-valued functions on Γ and denote their norms by $\|\cdot\|_{H^s(\Gamma)}$ and $\|\cdot\|_{\boldsymbol{H}^s(\Gamma)}$, respectively; for negative Sobolev indices the norms are defined by duality. Furthermore, for $s \in (0, 1)$, we denote by

$$\gamma : H^{s+\frac{1}{2}}(\Omega) \to H^s(\Gamma), \quad \boldsymbol{\gamma} : [H^{s+\frac{1}{2}}(\Omega)]^3 \to \boldsymbol{H}^s(\Gamma)$$

the standard continuous trace operators, and by R_γ and $\boldsymbol{R}_{\boldsymbol{\gamma}}$ their continuous right inverses.

For complex-valued vector functions we introduce the facewise $\boldsymbol{H}^{\frac{1}{2}}$-broken space

$$\boldsymbol{H}^{\frac{1}{2}}_{-}(\Gamma) = \{\boldsymbol{v} \in \boldsymbol{L}^2_t(\Gamma) \,|\, \boldsymbol{v}_{|\Gamma_i} \in \boldsymbol{H}^{\frac{1}{2}}(\Gamma_i),\ 1 \le i \le N_F\}.$$

with corresponding norm

$$\|\boldsymbol{v}\|^2_{\boldsymbol{H}^{\frac{1}{2}}_{-}(\Gamma)} = \sum_{j=1}^{N_F} \|\boldsymbol{v}\|^2_{\boldsymbol{H}^{\frac{1}{2}}(\Gamma_j)}.$$

Moreover, we define the spaces

$$\begin{aligned} \boldsymbol{H}^{\frac{1}{2}}_{\|}(\Gamma) &= \left\{ \boldsymbol{v} \in \boldsymbol{H}^{\frac{1}{2}}_{-}(\Gamma) \,\middle|\, \boldsymbol{v}_{|\Gamma_i}\cdot\boldsymbol{\tau}_{ij} \overset{1/2}{=} \boldsymbol{v}_{|\Gamma_j}\cdot\boldsymbol{\tau}_{ij},\ \text{for every edge } e_{ij} \right\}, \\ \boldsymbol{H}^{\frac{1}{2}}_{\perp}(\Gamma) &= \left\{ \boldsymbol{v} \in \boldsymbol{H}^{\frac{1}{2}}_{-}(\Gamma) \,\middle|\, \boldsymbol{v}_{|\Gamma_i}\cdot\boldsymbol{\tau}_{i} \overset{1/2}{=} \boldsymbol{v}_{|\Gamma_j}\cdot\boldsymbol{\tau}_{j},\ \text{for every edge } e_{ij} \right\}, \end{aligned} \tag{1}$$

where the relation $\overset{1/2}{=}$ is understood in the sense that

$$v_i \overset{1/2}{=} v_j \quad \Leftrightarrow \quad \int_{\Gamma_i}\int_{\Gamma_j} \frac{|v_i(\boldsymbol{x}) - v_j(\boldsymbol{y})|^2}{\|\boldsymbol{x}-\boldsymbol{y}\|^3} \mathrm{d}\sigma(\boldsymbol{x})\mathrm{d}\sigma(\boldsymbol{y}) < \infty. \tag{2}$$

We further define

$$\mathcal{N}_{ij}^{\|}(\boldsymbol{v}) := \int_{\Gamma_i}\int_{\Gamma_j} \frac{|(\boldsymbol{v}_{|\Gamma_i} \cdot \boldsymbol{\tau}_{ij})(\boldsymbol{x}) - (\boldsymbol{v}_{|\Gamma_j} \cdot \boldsymbol{\tau}_{ij})(\boldsymbol{y})|^2}{\|\boldsymbol{x}-\boldsymbol{y}\|^3} \mathrm{d}\sigma(\boldsymbol{x})\mathrm{d}\sigma(\boldsymbol{y}),$$

$$\mathcal{N}_{ij}^{\perp}(\boldsymbol{v}) := \int_{\Gamma_i}\int_{\Gamma_j} \frac{|(\boldsymbol{v}_{|\Gamma_i} \cdot \boldsymbol{\tau}_{i})(\boldsymbol{x}) - (\boldsymbol{v}_{|\Gamma_j} \cdot \boldsymbol{\tau}_{j})(\boldsymbol{y})|^2}{\|\boldsymbol{x}-\boldsymbol{y}\|^3} \mathrm{d}\sigma(\boldsymbol{x})\mathrm{d}\sigma(\boldsymbol{y}),$$

for each edge e_{ij} of the polyhedron and denote by $\mathscr{I}_j$ the set of indices i such that Γ_j and Γ_i have a common edge e_{ij}.

Proposition 1 ([6, Prop. 2.6]) *The spaces $\boldsymbol{H}_{\|}^{\frac{1}{2}}(\Gamma)$ and $\boldsymbol{H}_{\perp}^{\frac{1}{2}}(\Gamma)$ are Hilbert spaces when endowed with the norms*

$$\|\boldsymbol{v}\|^2_{\boldsymbol{H}_{\|}^{\frac{1}{2}}(\Gamma)} := \|\boldsymbol{v}\|^2_{\boldsymbol{H}_{-}^{\frac{1}{2}}(\Gamma)} + \sum_{j=1}^{N_F} \sum_{i \in \mathscr{I}_j} \mathcal{N}_{ij}^{\|}(\boldsymbol{v}), \tag{3}$$

$$\|\boldsymbol{v}\|^2_{\boldsymbol{H}_{\perp}^{\frac{1}{2}}(\Gamma)} := \|\boldsymbol{v}\|^2_{\boldsymbol{H}_{-}^{\frac{1}{2}}(\Gamma)} + \sum_{j=1}^{N_F} \sum_{i \in \mathscr{I}_j} \mathcal{N}_{ij}^{\perp}(\boldsymbol{v}). \tag{4}$$

In other words, $\boldsymbol{v} \in \boldsymbol{H}_{\|}^{\frac{1}{2}}(\Gamma), \boldsymbol{H}_{\perp}^{\frac{1}{2}}(\Gamma)$ satisfies $\boldsymbol{v} \in \boldsymbol{H}^{\frac{1}{2}}(\Gamma_i)$ on the faces Γ_i of Γ, and the parallel resp. orthogonal component of the function $\boldsymbol{v}$ to edges e_{ij} of Γ are "$H^{\frac{1}{2}}$-continuous" in the sense of (2); see [6] for further details.

We denote by $\boldsymbol{H}_{\|}^{-\frac{1}{2}}(\Gamma)$, $\boldsymbol{H}_{\perp}^{-\frac{1}{2}}(\Gamma)$ the dual spaces of $\boldsymbol{H}_{\|}^{\frac{1}{2}}(\Gamma)$, $\boldsymbol{H}_{\perp}^{\frac{1}{2}}(\Gamma)$ with pivot space $\boldsymbol{L}_t^2(\Gamma)$. The corresponding duality pairing is denoted by $\langle\cdot,\cdot\rangle_{\|,\Gamma}$ resp. $\langle\cdot,\cdot\rangle_{\perp,\Gamma}$. The norms $\|\cdot\|_{\boldsymbol{H}_{\|}^{-\frac{1}{2}}(\Gamma)}$ and $\|\cdot\|_{\boldsymbol{H}_{\perp}^{-\frac{1}{2}}(\Gamma)}$ are defined by duality.

For complex-valued functions $\boldsymbol{v} \in [C^\infty(\overline{\Omega})]^3$ the tangential traces are defined by

$$\boldsymbol{\gamma}_{\|}(\boldsymbol{v}) := \boldsymbol{n} \times (\boldsymbol{v} \times \boldsymbol{n})_{|\Gamma}, \qquad \boldsymbol{\gamma}_{\perp}(\boldsymbol{v}) := (\boldsymbol{v} \times \boldsymbol{n})_{|\Gamma}. \tag{5}$$

We point out that $\boldsymbol{\gamma}_{\|}(\boldsymbol{v}) = \boldsymbol{v} - (\boldsymbol{v}\cdot\boldsymbol{n})\boldsymbol{n}$ gives the component of $\boldsymbol{v}$ tangential to Γ, whereas $\boldsymbol{\gamma}_{\perp}(\boldsymbol{v})$ provides a tangent vector field perpendicular to $\boldsymbol{\gamma}_{\|}(\boldsymbol{v})$. Since Γ is a polyhedron, for any edge e_{ij} of Γ the components of $\boldsymbol{\gamma}_{\|}(\boldsymbol{v})$ and $\boldsymbol{\gamma}_{\perp}(\boldsymbol{v})$ tangential and normal to e_{ij} are continuous, namely,

$$\boldsymbol{\gamma}_{\|}(\boldsymbol{v})|_{\Gamma_i} \cdot \boldsymbol{\tau}_{ij} - \boldsymbol{\gamma}_{\|}(\boldsymbol{v})|_{\Gamma_j} \cdot \boldsymbol{\tau}_{ij} = 0, \quad \boldsymbol{\gamma}_{\perp}(\boldsymbol{v})|_{\Gamma_i} \cdot \boldsymbol{\tau}_{i} - \boldsymbol{\gamma}_{\perp}(\boldsymbol{v})|_{\Gamma_j} \cdot \boldsymbol{\tau}_{j} = 0. \tag{6}$$

This means that both operators $\boldsymbol{\gamma}_{\|}$ and $\boldsymbol{\gamma}_{\perp}$ can be viewed as face-by-face projections; see [6, p. 36]. Combining this observation with definitions (1), we realize that $\boldsymbol{H}_{\|}^{\frac{1}{2}}(\Gamma)$ and $\boldsymbol{H}_{\perp}^{\frac{1}{2}}(\Gamma)$ are the trace spaces of $\boldsymbol{\gamma}_{\|}$, $\boldsymbol{\gamma}_{\perp}$ acting on $\boldsymbol{H}^1(\Omega)$. This is stated in the following proposition.

Proposition 2 ([6, Prop. 2.7]) *The trace operators*

$$\boldsymbol{\gamma}_{\|} : \boldsymbol{H}^1(\Omega) \to \boldsymbol{H}_{\|}^{\frac{1}{2}}(\Gamma), \quad \boldsymbol{\gamma}_{\perp} : \boldsymbol{H}^1(\Omega) \to \boldsymbol{H}_{\perp}^{\frac{1}{2}}(\Gamma)$$

are linear, surjective and continuous operators. In addition, there exists continuous right inverse maps $\boldsymbol{R}_{\|} : \boldsymbol{H}_{\|}^{\frac{1}{2}}(\Gamma) \to \boldsymbol{H}^1(\Omega)$ *and* $\boldsymbol{R}_{\perp} : \boldsymbol{H}_{\perp}^{\frac{1}{2}}(\Gamma) \to \boldsymbol{H}^1(\Omega)$.

We can now establish a critical result for the upcoming analysis. Note that $\boldsymbol{H}_{\perp}^{\frac{1}{2}}(\Gamma)$ consists of tangential vector fields whereas $\boldsymbol{H}^{\frac{1}{2}}(\Gamma)$ does not.

Lemma 1 *There exists a continuous map* $\boldsymbol{t}_{\perp} : \boldsymbol{H}^{\frac{1}{2}}(\Gamma) \to \boldsymbol{H}_{\perp}^{\frac{1}{2}}(\Gamma)$ *with right inverse* $\boldsymbol{t}_{\perp}^{-1} : \boldsymbol{H}_{\perp}^{\frac{1}{2}}(\Gamma) \to \boldsymbol{H}^{\frac{1}{2}}(\Gamma)$.

Proof We define $\boldsymbol{t}_{\perp} : \boldsymbol{H}^{\frac{1}{2}}(\Gamma) \to \boldsymbol{H}_{\perp}^{\frac{1}{2}}(\Gamma)$ and $\boldsymbol{t}_{\perp}^{-1} : \boldsymbol{H}_{\perp}^{\frac{1}{2}}(\Gamma) \to \boldsymbol{H}^{\frac{1}{2}}(\Gamma)$ by

$$\begin{aligned} \boldsymbol{t}_{\perp}(\boldsymbol{w}) &= \boldsymbol{\gamma}_{\perp}(\boldsymbol{R}_{\boldsymbol{\gamma}}(\boldsymbol{w})), \quad \forall \boldsymbol{w} \in \boldsymbol{H}^{\frac{1}{2}}(\Gamma), \\ \boldsymbol{t}_{\perp}^{-1}(\boldsymbol{v}) &= \boldsymbol{\gamma}(\boldsymbol{R}_{\perp}(\boldsymbol{v})), \quad \forall \boldsymbol{v} \in \boldsymbol{H}_{\perp}^{\frac{1}{2}}(\Gamma) \end{aligned}$$

where γ and $\boldsymbol{R}_{\boldsymbol{\gamma}}$ are the trace and its right inverse, whereas $\gamma_{\perp}$ and $\boldsymbol{R}_{\perp}$ are the operators of Proposition 2. The continuity of these operators implies the continuity of $\boldsymbol{t}_{\perp}$ and $\boldsymbol{t}_{\perp}^{-1}$.

To prove that $\boldsymbol{t}_{\perp}^{-1}$ is the right inverse of $\boldsymbol{t}_{\perp}$, we observe that

$$\boldsymbol{t}_{\perp}(\boldsymbol{t}_{\perp}^{-1}(\boldsymbol{v})) = \boldsymbol{\gamma}_{\perp}(\boldsymbol{R}_{\boldsymbol{\gamma}}(\boldsymbol{\gamma}(\boldsymbol{R}_{\perp}(\boldsymbol{v})))), \qquad \forall \boldsymbol{v} \in \boldsymbol{H}_{\perp}^{\frac{1}{2}}(\Gamma),$$

and that $\boldsymbol{\gamma}_{\perp}$ projects face by face on Γ, see [6, page 36]. If $\boldsymbol{w} = \boldsymbol{t}_{\perp}^{-1}(\boldsymbol{v}) \in \boldsymbol{H}^{\frac{1}{2}}(\Gamma)$ and $\boldsymbol{g} = \boldsymbol{R}_{\boldsymbol{\gamma}}\boldsymbol{w} \in \boldsymbol{H}^1(\Omega)$, then $\boldsymbol{w} = \gamma(\boldsymbol{R}_{\perp}(\boldsymbol{v}))$ and $\boldsymbol{t}_{\perp}(\boldsymbol{w}) = \boldsymbol{\gamma}_{\perp}(\boldsymbol{g})$. Since $\boldsymbol{\gamma}_{\perp}(\boldsymbol{g})|_{\Gamma_i} = \boldsymbol{\gamma}(\boldsymbol{g})_{|\Gamma_i} \times \boldsymbol{n}$ for each face Γ_i of Γ, we obtain on Γ_i

$$\begin{aligned} \boldsymbol{t}_{\perp}(\boldsymbol{t}_{\perp}^{-1}(\boldsymbol{v})) &= \boldsymbol{t}_{\perp}(\boldsymbol{w}) = \boldsymbol{\gamma}_{\perp}(\boldsymbol{g}) = \boldsymbol{\gamma}(\boldsymbol{g}) \times \boldsymbol{n} \\ &= \boldsymbol{\gamma}(\boldsymbol{R}_{\boldsymbol{\gamma}}\boldsymbol{w}) \times \boldsymbol{n} = \boldsymbol{w} \times \boldsymbol{n} = \boldsymbol{\gamma}(\boldsymbol{R}_{\perp}(\boldsymbol{v})) \times \boldsymbol{n} = \boldsymbol{\gamma}_{\perp}(\boldsymbol{R}_{\perp}(\boldsymbol{v})) = \boldsymbol{v}. \end{aligned}$$

Thus, $\boldsymbol{t}_{\perp}^{-1}$ is indeed the right inverse of $\boldsymbol{t}_{\perp}$.

2.2 Tangential Differential Operators

We set $H^{\frac{3}{2}}(\Gamma) := \gamma(H^2(\Omega))$, and define the tangential operators $\mathbf{grad}_\Gamma : H^{\frac{3}{2}}(\Gamma) \to \boldsymbol{H}_\parallel^{\frac{1}{2}}(\Gamma)$ and $\mathbf{curl}_\Gamma : H^{\frac{3}{2}}(\Gamma) \to \boldsymbol{H}_\perp^{\frac{1}{2}}(\Gamma)$ by

$$\mathbf{grad}_\Gamma\, \phi := \boldsymbol{\gamma}_\parallel(\mathbf{grad}\, \phi) \qquad \mathbf{curl}_\Gamma\, \phi := \boldsymbol{\gamma}_\perp(\mathbf{grad}\, \phi) \qquad \forall \phi \in H^2(\Omega), \qquad (7)$$

where $\mathbf{grad}$ denotes the standard gradient in $\mathbb{R}^3$. According to definitions (5), $\mathbf{grad}_\Gamma\, \phi$ is the orthogonal projection of $\mathbf{grad}\phi$ on each face Γ_i of Γ, whereas $\mathbf{curl}_\Gamma\, \phi$ is obtained from the former by a $\pi/2$ rotation. It can be shown that the maps $\mathbf{grad}_\Gamma$ and $\mathbf{curl}_\Gamma$ are linear and continuous.

The adjoint operators $\operatorname{div} : \boldsymbol{H}_\parallel^{-\frac{1}{2}}(\Gamma) \to H^{-\frac{3}{2}}(\Gamma)$ and $\operatorname{curl}_\Gamma : \boldsymbol{H}_\perp^{-\frac{1}{2}}(\Gamma) \to H^{-\frac{3}{2}}(\Gamma)$ can be defined as follows:

$$\langle \operatorname{div} \boldsymbol{v}, \phi \rangle_{\frac{3}{2},\Gamma} = -\langle \boldsymbol{v}, \mathbf{grad}_\Gamma\, \phi \rangle_{\parallel,\Gamma},$$
$$\langle \operatorname{curl}_\Gamma \boldsymbol{w}, \phi \rangle_{\frac{3}{2},\Gamma} = \langle \boldsymbol{w}, \mathbf{curl}_\Gamma\, \phi \rangle_{\perp,\Gamma},$$

for all $\phi \in H^{\frac{3}{2}}(\Gamma)$, $\boldsymbol{v} \in \boldsymbol{H}_\parallel^{-\frac{1}{2}}(\Gamma)$ and $\boldsymbol{w} \in \boldsymbol{H}_\perp^{-\frac{1}{2}}(\Gamma)$.

In view of these definitions we now introduce the spaces

$$\boldsymbol{H}_{\operatorname{div}}^{-\frac{1}{2}}(\Gamma) := \left\{ \boldsymbol{v} \in \boldsymbol{H}_\parallel^{-\frac{1}{2}}(\Gamma) \,\middle|\, \operatorname{div} \boldsymbol{v} \in H^{-\frac{1}{2}}(\Gamma) \right\},$$
$$\boldsymbol{H}_{\operatorname{curl}}^{-\frac{1}{2}}(\Gamma) := \left\{ \boldsymbol{v} \in \boldsymbol{H}_\perp^{-\frac{1}{2}}(\Gamma) \,\middle|\, \operatorname{curl}_\Gamma \boldsymbol{v} \in H^{-\frac{1}{2}}(\Gamma) \right\},$$

which will play a crucial role for the upcoming analysis and are endowed with the graph norms

$$\|\boldsymbol{v}\|^2_{\boldsymbol{H}_{\operatorname{div}}^{-\frac{1}{2}}(\Gamma)} := \|\boldsymbol{v}\|^2_{\boldsymbol{H}_\parallel^{-\frac{1}{2}}(\Gamma)} + \|\operatorname{div} \boldsymbol{v}\|^2_{H^{-\frac{1}{2}}(\Gamma)},$$
$$\|\boldsymbol{v}\|^2_{\boldsymbol{H}_{\operatorname{curl}}^{-\frac{1}{2}}(\Gamma)} := \|\boldsymbol{v}\|^2_{\boldsymbol{H}_\perp^{-\frac{1}{2}}(\Gamma)} + \|\operatorname{curl}_\Gamma \boldsymbol{v}\|^2_{H^{-\frac{1}{2}}(\Gamma)}.$$

Let the natural solution space of Maxwell's equations be denoted by

$$\boldsymbol{H}(\operatorname{curl}, \Omega) = \left\{ \boldsymbol{v} \in \boldsymbol{L}^2(\Omega) \,\middle|\, \operatorname{curl} \boldsymbol{v} \in \boldsymbol{L}^2(\Omega) \right\}.$$

Theorem 1 ([7, Theorem 4.6]) *The mappings $\boldsymbol{\gamma}_\parallel$ and $\boldsymbol{\gamma}_\perp$ admit linear and continuous extensions*

$$\boldsymbol{\gamma}_\parallel : \boldsymbol{H}(\operatorname{curl}, \Omega) \to \boldsymbol{H}_{\operatorname{curl}}^{-\frac{1}{2}}(\Gamma), \quad \boldsymbol{\gamma}_\perp : \boldsymbol{H}(\operatorname{curl}, \Omega) \to \boldsymbol{H}_{\operatorname{div}}^{-\frac{1}{2}}(\Gamma).$$

Moreover, the following integration by parts formula holds:

$$\int_{\Omega}\Big(\operatorname{curl}\boldsymbol{v}\cdot\boldsymbol{u}-\operatorname{curl}\boldsymbol{u}\cdot\boldsymbol{v}\Big)\mathrm{d}\Omega=\langle\boldsymbol{\gamma}_{\perp}\boldsymbol{u},\boldsymbol{\gamma}_{\|}\boldsymbol{v}\rangle_{\|,\Gamma},\quad\forall\boldsymbol{u}\in\boldsymbol{H}(\operatorname{curl},\Omega),\boldsymbol{v}\in\boldsymbol{H}^{1}(\Omega).$$

Furthermore, a duality pairing${}_{\perp}\langle\cdot,\cdot\rangle_{\|}$ between $\boldsymbol{H}_{\operatorname{div}}^{-\frac{1}{2}}(\Gamma)$ and $\boldsymbol{H}_{\operatorname{curl}}^{-\frac{1}{2}}(\Gamma)$ can be established by using an orthogonal decomposition of those spaces so that the following integration by parts formula still holds

$$\int_{\Omega}\Big(\operatorname{curl}\boldsymbol{v}\cdot\boldsymbol{u}-\operatorname{curl}\boldsymbol{u}\cdot\boldsymbol{v}\Big)\mathrm{d}\Omega={}_{\perp}\langle\boldsymbol{\gamma}_{\perp}\boldsymbol{u},\boldsymbol{\gamma}_{\|}\boldsymbol{v}\rangle_{\|},\qquad\forall\boldsymbol{u},\boldsymbol{v}\in\boldsymbol{H}(\operatorname{curl},\Omega).$$

For more details, we refer to [7].

The differential operators $\mathbf{grad}_{\Gamma}$ and $\mathbf{curl}_{\Gamma}$ can be further extended as follows.

Proposition 3 ([7, p. 39]) *The tangential gradient and curl operators introduced in* (7) *can be extended to linear and continuous operators defined on* $H^{\frac{1}{2}}(\Gamma)$

$$\mathbf{grad}_{\Gamma}:H^{\frac{1}{2}}(\Gamma)\to\boldsymbol{H}_{\perp}^{-\frac{1}{2}}(\Gamma),\quad\mathbf{curl}_{\Gamma}:H^{\frac{1}{2}}(\Gamma)\to\boldsymbol{H}_{\|}^{-\frac{1}{2}}(\Gamma).$$

Moreover their formal $\boldsymbol{L}_t^2(\Gamma)$-adjoints

$$\operatorname{div}:\boldsymbol{H}_{\perp}^{\frac{1}{2}}(\Gamma)\to H^{-\frac{1}{2}}(\Gamma)\quad\text{and}\quad\operatorname{curl}_{\Gamma}:\boldsymbol{H}_{\|}^{\frac{1}{2}}(\Gamma)\to H^{-\frac{1}{2}}(\Gamma)$$

can be defined by

$$\begin{aligned}\langle\operatorname{div}\boldsymbol{v},\phi\rangle_{\frac{1}{2},\Gamma}&=-\langle\boldsymbol{v},\mathbf{grad}_{\Gamma}\,\phi\rangle_{\perp,\Gamma},\\\langle\operatorname{curl}_{\Gamma}\boldsymbol{w},\phi\rangle_{\frac{1}{2},\Gamma}&=\langle\boldsymbol{w},\mathbf{curl}_{\Gamma}\,\phi\rangle_{\|,\Gamma},\end{aligned}\tag{8}$$

for all $\phi\in H^{\frac{1}{2}}(\Gamma)$, $\boldsymbol{v}\in\boldsymbol{H}_{\perp}^{\frac{1}{2}}(\Gamma)$ and $\boldsymbol{w}\in\boldsymbol{H}_{\|}^{\frac{1}{2}}(\Gamma)$.

2.3 Potentials

Let G_k denote the fundamental solution of the Helmholtz operator $\Delta+k^2$, which is given by

$$G_k(\boldsymbol{x},\boldsymbol{y}):=\frac{\exp(ik|\boldsymbol{x}-\boldsymbol{y}|)}{4\pi\,|\boldsymbol{x}-\boldsymbol{y}|}.$$

The scalar and vector single layer potential are then defined respectively by

$$\begin{aligned}\Psi_k^V&:H^{-\frac{1}{2}}(\Gamma)\to H^1_{\mathrm{loc}}(\mathbb{R}^3),\quad\Psi_k^V(v)(\boldsymbol{x}):=\int_{\Gamma}G_k(\boldsymbol{x},\boldsymbol{y})v(\boldsymbol{y})\mathrm{d}\sigma(\boldsymbol{y}),\\\boldsymbol{\Psi}_k^A&:\boldsymbol{H}_{\|}^{-\frac{1}{2}}(\Gamma)\to\boldsymbol{H}^1_{\mathrm{loc}}(\mathbb{R}^3),\quad\boldsymbol{\Psi}_k^A(\boldsymbol{v})(\boldsymbol{x}):=\int_{\Gamma}G_k(\boldsymbol{x},\boldsymbol{y})\boldsymbol{v}(\boldsymbol{y})\mathrm{d}\sigma(\boldsymbol{y}).\end{aligned}$$

These potentials are known to be continuous, see [12, Theorem 3.8]. Finally the scalar and vector single layer boundary operators are defined by

$$\begin{aligned} V_k &: H^{-\frac{1}{2}}(\Gamma) \to H^{\frac{1}{2}}(\Gamma), \quad V_k := \gamma \circ \Psi_k^V, \\ \boldsymbol{A}_k &: \boldsymbol{H}_{\|}^{-\frac{1}{2}}(\Gamma) \to \boldsymbol{H}_{\|}^{\frac{1}{2}}(\Gamma), \quad \boldsymbol{A}_k := \boldsymbol{\gamma}_{\|} \circ \boldsymbol{\Psi}_k^A. \end{aligned}$$

The simultaneous continuity of the trace operators $\gamma, \boldsymbol{\gamma}_{\|}$ and the single layer potentials yield then the continuity of the single layer boundary operators, namely,

$$\|V_k v\|_{H^{\frac{1}{2}}(\Gamma)} \preceq \|v\|_{H^{-\frac{1}{2}}(\Gamma)}, \qquad \|\boldsymbol{A}_k \boldsymbol{v}\|_{\boldsymbol{H}_{\|}^{\frac{1}{2}}(\Gamma)} \preceq \|\boldsymbol{v}\|_{\boldsymbol{H}_{\|}^{-\frac{1}{2}}(\Gamma)}, \tag{9}$$

for all $v \in H^{-\frac{1}{2}}(\Gamma)$ and $\boldsymbol{v} \in \boldsymbol{H}_{\|}^{-\frac{1}{2}}(\Gamma)$. In particular, the range of V_k lies in $H^1(\Gamma)$ if it is restricted to $L^2(\Gamma) \subset H^{-\frac{1}{2}}(\Gamma)$ (see [12, Theorem 3.8]), i.e.,

$$\operatorname{Im}(V_k(L^2(\Gamma)) \subset H^1(\Gamma). \tag{10}$$

Likewise, for the vector case the corresponding result reads

$$\operatorname{Im}(\boldsymbol{A}_k(\boldsymbol{L}_t^2(\Gamma)) \subset \boldsymbol{H}^1(\Gamma), \tag{11}$$

see [8, Prop. 2].

2.4 Interpolation of Weighted Spaces

In the following section we will be confronted with interpolation of weighted L^2-spaces. We thus recall in this section some basic results taken from [47], which are valid without regularity on the weights.

Let $\mathscr{T}$ be a family of shape-regular triangulations decomposing Γ into flat triangles such that the surface covered by the triangles coincides with Γ. Denote the set of edges of the mesh by $\mathscr{E}_{\mathscr{T}}$. For a fixed triangulation let h_T denote the diameter of any element $T \in \mathscr{T}$ and let $\mathtt{h}$ be the piecewise constant function such that $\mathtt{h}|_T = h_T$.

Throughout the entire paper, we use the notation $a \preceq b$ which needs to be understood in the sense that there exists a constant $\mathrm{C} > 0$ being independent of the mesh $\mathscr{T}$ such that $a \leq \mathrm{C}b$.

Lemma 2 ([47, Lemma 22.3, p. 110]) *If A is linear from $E_0 + E_1$ into $F_0 + F_1$ and maps E_0 into F_0 with $\|Ax\|_{F_0} \leq M_0 \|x\|_{E_0}$ for all $x \in E_0$, and maps E_1 into F_1 with $\|Ax\|_{F_1} \leq M_1 \|x\|_{E_1}$ for all $x \in E_1$, then A is linear continuous from $(E_0, E_1)_{\theta,p}$ into $(F_0, F_1)_{\theta,p}$ for all $0 < \theta < 1$ and $1 \leq p \leq \infty$ (or for $\theta = 0, 1$ with $p = \infty$). For $0 < \theta < 1$ one has*

$$\|Aa\|_{(F_0,F_1)_{\theta,p}} \leq M_0^{1-\theta} M_1^{\theta} \|a\|_{(E_0,E_1)_{\theta,p}} \quad \textit{for all } a \in (E_0, E_1)_{\theta,p}.$$

The space $(E_0, E_1)_{\theta,p}$ denotes the interpolation space between E_0 and E_1 based on the L^p-norm using the K-method, see [47, Definition 22.1, p. 109].

Lemma 3 ([47, Lemma 23.1, p. 115]) *For a measurable positive function w on Γ, let*

$$E(w) = \left\{ u \,\middle|\, \int_\Gamma |u(x)|^2 w(x)\, dx < \infty \right\} \quad \textit{with} \quad \|u\|_w = \left(\int_\Gamma |u(x)|^2 w(x)\, dx \right)^{\frac{1}{2}}.$$

If w_0, w_1 are two such functions, then for $0 < \theta < 1$ one has

$$(E(w_0), E(w_1))_{\theta,2} = E(w_\theta), \quad \textit{where } w_\theta = w_0^{1-\theta} w_1^{\theta},$$

with equivalent norms

$$\frac{\pi}{2\sin(\pi\theta)} \|u\|_{(E(w_0),E(w_1))_{\theta,2}} \leq \|u\|_{w_\theta} \leq \frac{\pi}{\sqrt{2}\sin(\pi\theta)} \|u\|_{(E(w_0),E(w_1))_{\theta,2}}.$$

Remark 1 The equivalence constants are not explicitly given in [47, Lemma 23.1, p. 115], but they follow from the proof. Notice further that the equivalence constants are independent of the weight functions.

Corollary 1 *Let $0 < s < 1$ be arbitrary. Let A be a linear continuous map from $L^2(\Gamma)$ into $L^2(\Gamma)$ and from $H^1(\Gamma)$ into $L^2(\Gamma)$ with*

$$\begin{aligned} \|Av\|_{L^2(\Gamma)} &\leq M_0 \|v\|_{L^2(\Gamma)} && \textit{for all } v \in L^2(\Gamma), \\ \|\mathtt{h}^{-1} Av\|_{L^2(\Gamma)} &\leq M_1 \|v\|_{H^1(\Gamma)} && \textit{for all } v \in H^1(\Gamma). \end{aligned}$$

Then A is a linear map from $H^s(\Gamma) = (H^1(\Gamma), L^2(\Gamma))_{s,2}$ into $L^2(\Gamma)$ with

$$\|\mathtt{h}^{-s} Av\|_{L^2(\Gamma)} \leq \frac{\pi}{\sqrt{2}\sin(\pi/2)} M_0^{1-s} M_1^{s} \|v\|_{H^s(\Gamma)} \quad \textit{for all } v \in H^{\frac{1}{2}}(\Gamma).$$

Proof Combine Lemmas 2 with 3.

2.5 Discrete Spaces and Interpolation

Let $\boldsymbol{RT}_0(T)$ denote the *local* Raviart-Thomas space of complex-valued functions on $T \in \mathscr{T}$ defined by (see [44] or an overview in the book by [29])

$$\boldsymbol{RT}_0(T) := \left\{ \boldsymbol{v}(\boldsymbol{x}) = \boldsymbol{\alpha} + \beta \boldsymbol{x} \,\middle|\, \boldsymbol{\alpha} \in \mathbb{C}^2, \beta \in \mathbb{C} \right\}.$$

The *global* Raviart-Thomas space is defined by

$$\boldsymbol{RT}_0 := \left\{ \boldsymbol{v} \in \boldsymbol{H}^0_{\mathrm{div}}(\Gamma) \,\middle|\, \boldsymbol{v}_{|T} \in \boldsymbol{RT}_0(T) \quad \forall T \in \mathscr{T} \right\},$$

where $\boldsymbol{H}^0_{\mathrm{div}}(\Gamma)$ is defined in a standard manner

$$\boldsymbol{H}^0_{\mathrm{div}}(\Gamma) := \left\{ \boldsymbol{v} \in \boldsymbol{L}^2_t(\Gamma) \,\middle|\, \mathrm{div}\, \boldsymbol{v} \in L^2(\Gamma) \right\}.$$

Further denote by $\mathbb{V}(\mathscr{T})$ the space of scalar complex-valued continuous functions that are piecewise linear, namely

$$\mathbb{V}(\mathscr{T}) = \left\{ v \in H^1(\Gamma) \,\middle|\, v|_T \in \mathbb{P}_1(T) \right\}, \tag{12}$$

where $\mathbb{P}_1(T)$ denotes the space of affine polynomials on T. Let $\mathscr{N}(\mathscr{T})$ denote the set of all nodes ν of $\mathscr{T}$ and $\{\varphi_\nu\}_{\nu \in \mathscr{N}(\mathscr{T})}$ be the family of nodal bases of $\mathbb{V}(\mathscr{T})$.

Definition 1 Let the Clément type interpolation operator $\mathrm{I}_{\mathscr{T}} : \mathrm{L}^2() \to \mathbb{V}(\mathscr{T})$ be

$$\mathrm{I}_{\mathscr{T}}\mathrm{v} := \sum_{\nu \in \mathscr{N}(\mathscr{T})} \phi_\nu(\mathrm{v})\varphi_\nu \quad \forall \mathrm{v} \in \mathrm{L}^2(),$$

where $\Gamma_\nu = \mathrm{supp}(\varphi_\nu)$ and the degrees of freedom are given by

$$\phi_\nu(v) := \frac{3}{|\Gamma_\nu|} \int_{\Gamma_\nu} v(\boldsymbol{x})\varphi_\nu(\boldsymbol{x})\, d\boldsymbol{x}.$$

Proposition 4 *If* $v \in H^s(\Gamma)$ *with* $0 < s < 1$, *then the interpolation operator* $\mathrm{I}_{\mathscr{T}}$ *satisfies the following interpolation properties:*

$$\|\mathtt{h}^{-s}(v - \mathrm{I}_{\mathscr{T}}\mathrm{v})\|_{\mathrm{L}^2()} \preceq \|\mathrm{v}\|_{\mathrm{H}^s()} \quad \textit{for all}\ \mathrm{v} \in \mathrm{H}^{\mathrm{s}}(). \tag{13}$$

Remark 2 A similar result has been developed in [48].

Proof This interpolation operator $\mathrm{I}_{\mathscr{T}}$ is also used in [27, (2.2.29)] and the following result is proven

$$\|\mathtt{h}^{-1}(v - \phi_\nu(v))\|_{L^2(\Gamma_\nu)} \preceq \|\,\mathbf{grad}_\Gamma\, v\|_{\boldsymbol{L}^2(\Gamma_\nu)}$$

for $v \in H^1(\Gamma)$ under the assumption of shape-regularity. (Note that in our case the mesh matches the surface and, therefore, the Eq. (2.2.29) can be simplified.). Following the arguments of the original paper of [25, Proof of Theorem 1], it is now straightforward to prove that

$$\|\mathtt{h}^{-1}(v - \mathrm{I}_{\mathscr{T}}\mathrm{v})\|_{\mathrm{L}^2(\mathrm{T})} \preceq \sum_{\nu \in \mathscr{N}(\mathrm{T})} \|\,\mathbf{grad}\ \mathrm{v}\|_{\boldsymbol{L}^2(_\nu)},$$

where $\mathscr{N}(T)$ denotes the set of nodes of the element T. Now, summing over all elements of the mesh $\mathscr{T}$ and using that the number of elements sharing a node is bounded, as a consequence of shape regularity of $\mathscr{T}$, we get

$$\|\mathtt{h}^{-1}(v - \mathrm{I}_{\mathscr{T}}\mathrm{v})\|_{\mathrm{L}^2()} \preceq |\mathrm{v}|_{\mathrm{H}^1()}.$$

Furthermore, the operator can also be shown to be L^2-stable, see [27, (2.2.33)].

Therefore, the linear continuous operator $A_{\mathscr{T}} = \mathrm{Id} - \mathrm{I}_{\mathscr{T}} : \mathrm{L}^2() \mapsto \mathrm{L}^2()$ satisfies

$$\begin{aligned} \|\mathtt{h}^{-1}(v - \mathrm{I}_{\mathscr{T}}\mathrm{v})\|_{\mathrm{L}^2()} &\preceq \|v\|_{H^1(\Gamma)}, \\ \|v - \mathrm{I}_{\mathscr{T}}\mathrm{v}\|_{\mathrm{L}^2()} &\preceq \|v\|_{L^2(\Gamma)}. \end{aligned}$$

The asserted estimate (13) follows from Corollary 1.

Besides this for $s = 1/2$, we will also need a Raviart-Thomas type interpolation operator for functions in $\boldsymbol{v} \in \boldsymbol{H}_{\perp}^{\frac{1}{2}}(\Gamma)$. Since $\operatorname{div} \boldsymbol{v} \notin L^2(\Gamma)$, the standard degrees of freedom are no longer well-defined. Therefore, we will utilize an interpolation operator similar to that introduced in [4] for the first type Nédélec elements.

For any edge $e \in \mathscr{E}_{\mathscr{T}}$ of the mesh we associate an arbitrary but fixed element T_e such that $e \subset \partial T_e$. On T_e we denote by $\boldsymbol{\pi}_e$ the $\boldsymbol{L}^2(T_e)$-projection onto constant functions. We let $\{\boldsymbol{\psi}_e\}_{e \in \mathscr{E}_{\mathscr{T}}}$ be the standard Raviart-Thomas basis of lowest order, sometimes also referred to as the Rao-Wilton-Glisson (RWG) basis in this context of electromagnetic scattering, such that

$$\int_e \boldsymbol{\psi}_e \cdot \boldsymbol{\nu}_e \mathrm{d}s = 1 \quad \text{and} \quad \int_e \boldsymbol{\psi}_{e'} \cdot \boldsymbol{\nu}_e \mathrm{d}s = 0$$

for any $e' \in \mathscr{E}_{\mathscr{T}}$ such that $e \neq e'$ and where $\boldsymbol{\nu}_e$ denotes the outer unit normal of T_e at the edge e which is coplanar with T_e; see Fig. 2.

Definition 2 Let the Clément type interpolation operator $\boldsymbol{I}_{\mathscr{T}} : \boldsymbol{L}_t^2(\Gamma) \to \boldsymbol{RT}_0$ for the Raviart-Thomas element of lowest order be given by

$$\boldsymbol{I}_{\mathscr{T}} \boldsymbol{v} := \sum_{e \in \mathscr{E}_{\mathscr{T}}} \alpha_e(\boldsymbol{v}) \boldsymbol{\psi}_e$$

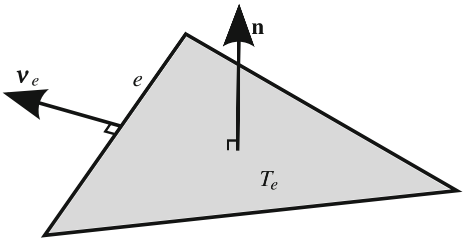


Fig. 2 Illustration of the normals on an element T_e

where the degrees of freedom are defined by

$$\alpha_e(\boldsymbol{v}) := \int_e \boldsymbol{\pi}_e(\boldsymbol{v}) \cdot \boldsymbol{v}_e \mathrm{d}s.$$

Remark 3 Note that $\boldsymbol{I}_{\mathscr{T}}$ does not satisfy the usual commutative property

$$\operatorname{div}(\boldsymbol{I}_{\mathscr{T}} \boldsymbol{v}) \neq \mathrm{P}_0(\operatorname{div} \boldsymbol{v}),$$

where P_0 denotes the element-wise L^2-projection of degree 0. This is important in the a priori analysis but not in the upcoming a posteriori error analysis.

Lemma 4 *The degrees of freedom of the interpolation operator* $\boldsymbol{I}_{\mathscr{T}} : \boldsymbol{L}_t^2(\Gamma) \to \boldsymbol{RT}_0$ *are well-defined and* $\boldsymbol{I}_{\mathscr{T}}$ *satisfies the local* L^2*-stability bound*

$$\|\boldsymbol{I}_{\mathscr{T}} \boldsymbol{v}\|_{\boldsymbol{L}^2(T)} \preceq \|\boldsymbol{v}\|_{\boldsymbol{L}^2(\Delta_T)} \quad \textit{for all } T \in \mathscr{T},$$

where Δ_T *denotes the set of elements that share at least one edge with* T.

Proof The argument is similar as in [4]. If $T \in \mathscr{T}$ is any element and $\mathscr{E}(T)$ denotes the three edges of T, then

$$\|\boldsymbol{I}_{\mathscr{T}} \boldsymbol{v}\|_{\boldsymbol{L}^2(T)} = \| \sum_{e \in \mathscr{E}(T)} \alpha_e \boldsymbol{\psi}_e \|_{\boldsymbol{L}^2(T)} \leq \sum_{e \in \mathscr{E}(T)} |\alpha_e| \, \|\boldsymbol{\psi}_e\|_{\boldsymbol{L}^2(T)}.$$

Invoking the Piola transformation, it can be shown that

$$\|\boldsymbol{\psi}_e\|_{\boldsymbol{L}^2(T)} \preceq \|\widehat{\boldsymbol{\psi}}_{\hat{e}}\|_{\boldsymbol{L}^2(\hat{T})} \preceq 1$$

since the basis functions $\widehat{\boldsymbol{\psi}}_{\hat{e}}$ on the reference element $\widehat{T}$ are bounded. Moreover, applying the Cauchy-Schwarz inequality, we get

$$|\alpha_e| = \left| \int_e \boldsymbol{\pi}_e(\boldsymbol{v}) \cdot \boldsymbol{v}_e \mathrm{d}s \right| \preceq h_T^{\frac{1}{2}} \, \|\boldsymbol{\pi}_e \boldsymbol{v}\|_{\boldsymbol{L}^2(e)} \preceq h_T \|\hat{\boldsymbol{\pi}}_e \hat{\boldsymbol{v}}\|_{\boldsymbol{L}^2(\hat{e})}$$

where $\hat{\boldsymbol{\pi}}_e$ denotes the $\boldsymbol{L}^2(\widehat{T})$-projection onto constant functions on the reference element $\widehat{T}$. Note that $\hat{\boldsymbol{\pi}}_e \hat{\boldsymbol{v}}$ is defined via the affine transformation from T_e (and not T) to $\hat{T}$. Norm equivalence of polynomials (constant functions in this case), the L^2-stability of $\hat{\boldsymbol{\pi}}_e$ and a scaling argument yield

$$|\alpha_e| \preceq h_T \|\hat{\boldsymbol{\pi}}_e \hat{\boldsymbol{v}}\|_{\boldsymbol{L}^2(\widehat{T})} \preceq h_T \|\hat{\boldsymbol{v}}\|_{\boldsymbol{L}^2(\widehat{T})} \preceq \|\boldsymbol{v}\|_{\boldsymbol{L}^2(T_e)}. \tag{14}$$

Combining the above estimates implies the asserted stability bound of $\boldsymbol{I}_{\mathscr{T}}$.

To explore the accuracy of the interpolant $\boldsymbol{I}_{\mathscr{T}}$, we need the following lemmas.

Lemma 5 (Local approximability) *The following estimate holds:*

$$\|v - \boldsymbol{I}_{\mathscr{T}}\boldsymbol{v}\|_{\boldsymbol{L}^2(T)} \preceq \inf_{\boldsymbol{c}\in\mathbb{R}^3} \|\boldsymbol{v} - \boldsymbol{\gamma}_\perp \boldsymbol{c}\|_{\boldsymbol{L}^2(\Delta_T)}, \quad \textit{for all } T \in \mathscr{T}.$$

Proof Let $\bar{\boldsymbol{v}} = \boldsymbol{\gamma}_\perp \boldsymbol{c}$ for any $\boldsymbol{c} \in \mathbb{R}^3$ which, in view of (5), is piecewise constant in $\mathscr{T}$. According to (6) the normal component of $\bar{\boldsymbol{v}}$ is continuous across all edges of the mesh, including those of the polyhedron Γ, whence $\bar{\boldsymbol{v}} \in \boldsymbol{RT}_0$.

We first observe that $\boldsymbol{I}_{\mathscr{T}}\bar{\boldsymbol{v}} = \bar{\boldsymbol{v}}$ because $\boldsymbol{\pi}_e\bar{\boldsymbol{v}}|_{T_e} = \bar{\boldsymbol{v}}|_{T_e}$ for any edge $e \subset \partial T$ and

$$\alpha_e(\bar{\boldsymbol{v}}) = \int_e \boldsymbol{\pi}_e\bar{\boldsymbol{v}} \cdot \boldsymbol{\nu}_e \mathrm{d}s = \int_e \bar{\boldsymbol{v}} \cdot \boldsymbol{\nu}_e \mathrm{d}s.$$

Since these three local degrees of freedom on $T \in \mathscr{T}$ are unisolvent and they coincide for both $\boldsymbol{I}_{\mathscr{T}}\bar{\boldsymbol{v}}|_T$ and $\bar{\boldsymbol{v}}|_T$, we deduce $\boldsymbol{I}_{\mathscr{T}}\bar{\boldsymbol{v}}|_T = \bar{\boldsymbol{v}}|_T$. Consequently

$$\|\boldsymbol{v} - \boldsymbol{I}_{\mathscr{T}}\boldsymbol{v}\|_{\boldsymbol{L}^2(T)} \leq \|\boldsymbol{v} - \bar{\boldsymbol{v}}\|_{\boldsymbol{L}^2(T)} + \|\boldsymbol{I}_{\mathscr{T}}(\boldsymbol{v} - \bar{\boldsymbol{v}})\|_{\boldsymbol{L}^2(T)}.$$

By the local $\boldsymbol{L}^2$-stability of Lemma 4 we conclude that

$$\|\boldsymbol{v} - \boldsymbol{I}_{\mathscr{T}}\boldsymbol{v}\|_{\boldsymbol{L}^2(T)} \preceq \|\boldsymbol{v} - \bar{\boldsymbol{v}}\|_{\boldsymbol{L}^2(\Delta_T)},$$

as asserted.

Lemma 6 *If $\boldsymbol{v} \in \boldsymbol{H}_\perp^{\frac{1}{2}}(\Gamma)$, then there exists $\boldsymbol{w} \in \boldsymbol{H}^{\frac{1}{2}}(\Gamma)$ such that $\boldsymbol{t}_\perp(\boldsymbol{w}) = \boldsymbol{v}$ and $\|\boldsymbol{w}\|_{\boldsymbol{H}^{\frac{1}{2}}(\Gamma)} \preceq \|\boldsymbol{v}\|_{\boldsymbol{H}_\perp^{\frac{1}{2}}(\Gamma)}$.*

Proof Simply set $\boldsymbol{w} = \boldsymbol{t}_\perp^{-1}(\boldsymbol{v})$, where $\boldsymbol{t}_\perp^{-1}$ is defined in Lemma 1, and use the facts that $\boldsymbol{t}_\perp^{-1}$ is the right inverse of $\boldsymbol{t}_\perp$ and the continuity of $\boldsymbol{t}_\perp^{-1}$.

Lemma 7 *If $\boldsymbol{w} \in \boldsymbol{H}^{\frac{1}{2}}(\Gamma)$, then $\|\boldsymbol{t}_\perp(\boldsymbol{w})\|_{\boldsymbol{L}^2(\Delta_T)} \preceq \|\boldsymbol{w}\|_{\boldsymbol{L}^2(\Delta_T)}$.*

Proof As in the proof of Lemma 1, let $\boldsymbol{g} \in \boldsymbol{H}^1(\Omega)$ be the function such that $\boldsymbol{g} = \boldsymbol{R}_\gamma \boldsymbol{w}$ and $\boldsymbol{t}_\perp(\boldsymbol{w}) = \boldsymbol{\gamma}_\perp(\boldsymbol{g})$. Therefore

$$\boldsymbol{\gamma}_\perp(\boldsymbol{g})|_T = \boldsymbol{\gamma}(\boldsymbol{g})|_T \times \boldsymbol{n}, \quad \text{for a.e. } \boldsymbol{x} \in T, \ \text{for all } T \in \mathscr{T},$$

whence

$$\begin{aligned}\|\boldsymbol{t}_\perp(\boldsymbol{w})\|^2_{\boldsymbol{L}^2(\Delta_T)} &= \sum_{T\subset\Delta_T} \|\boldsymbol{t}_\perp(\boldsymbol{w})\|^2_{\boldsymbol{L}^2(T)} = \sum_{T\subset\Delta_T} \|\boldsymbol{\gamma}(\boldsymbol{g}) \times \boldsymbol{n}\|^2_{\boldsymbol{L}^2(T)} \\ &\preceq \sum_{T\subset\Delta_T} \|\boldsymbol{\gamma}(\boldsymbol{g})\|^2_{\boldsymbol{L}^2(T)} = \|\boldsymbol{\gamma}(\boldsymbol{g})\|^2_{\boldsymbol{L}^2(\Delta_T)} = \|\boldsymbol{w}\|^2_{\boldsymbol{L}^2(\Delta_T)}\end{aligned}$$

because $\boldsymbol{R}_\gamma$ is the right inverse of $\boldsymbol{\gamma}$ and thus $\gamma(\boldsymbol{g}) = \boldsymbol{w}$.

Proposition 5 (Global approximability) *If $\boldsymbol{v} \in \boldsymbol{H}_{\perp}^{\frac{1}{2}}(\Gamma)$, then the interpolation operator $\boldsymbol{I}_{\mathscr{T}}$ satisfies the following global error estimate*

$$\|\mathtt{h}^{-\frac{1}{2}}(\boldsymbol{v} - \boldsymbol{I}_{\mathscr{T}}\boldsymbol{v})\|_{\boldsymbol{L}^2(\Gamma)} \preceq \|\boldsymbol{v}\|_{\boldsymbol{H}_{\perp}^{\frac{1}{2}}(\Gamma)} \quad \textit{for all}\ \boldsymbol{v} \in \boldsymbol{H}_{\perp}^{\frac{1}{2}}(\Gamma). \tag{15}$$

Proof Given $\boldsymbol{v} \in \boldsymbol{H}_{\perp}^{\frac{1}{2}}(\Gamma)$, there exists $\boldsymbol{w} \in \boldsymbol{H}^{\frac{1}{2}}(\Gamma)$ so that $\boldsymbol{t}_{\perp}(\boldsymbol{w}) = \boldsymbol{v}$ according to Lemma 6. For each $T \in \mathscr{T}$, we define $\overline{\boldsymbol{w}}_T = \int_{\Delta_T} \boldsymbol{w}(\boldsymbol{x})d\boldsymbol{x} \in \mathbb{R}^3$ and $\bar{\boldsymbol{v}}_T = \boldsymbol{t}_{\perp}(\overline{\boldsymbol{w}}_T) \in \boldsymbol{H}_{\perp}^{\frac{1}{2}}(\Gamma)$. Since the estimate of Lemma 5 is local, we have

$$\|\mathtt{h}^{\alpha}(\boldsymbol{v} - \boldsymbol{I}_{\mathscr{T}}\boldsymbol{v})\|^2_{\boldsymbol{L}^2(\Gamma)} = \sum_{T\in\mathscr{T}} \|\mathtt{h}^{\alpha}(\boldsymbol{v} - \boldsymbol{I}_{\mathscr{T}}\boldsymbol{v})\|^2_{\boldsymbol{L}^2(T)} \preceq \sum_{T\in\mathscr{T}} \|h_T^{\alpha}(\boldsymbol{v} - \bar{\boldsymbol{v}}_T)\|^2_{\boldsymbol{L}^2(\Delta_T)}$$

for $\alpha = -1, 0$. Using Lemma 7 yields

$$\|\boldsymbol{v} - \bar{\boldsymbol{v}}_T\|_{\boldsymbol{L}^2(\Delta_T)} = \|\boldsymbol{t}_{\perp}(\boldsymbol{w} - \overline{\boldsymbol{w}}_T)\|_{\boldsymbol{L}^2(\Delta_T)} \preceq \|\boldsymbol{w} - \overline{\boldsymbol{w}}_T\|_{\boldsymbol{L}^2(\Delta_T)}$$

and stability of the L^2-projection together with the definition of $\overline{\boldsymbol{w}}_T$ implies

$$\begin{aligned} \|\boldsymbol{w} - \overline{\boldsymbol{w}}_T\|_{\boldsymbol{L}^2(\Delta_T)} &\preceq \|\boldsymbol{w}\|_{\boldsymbol{L}^2(\Delta_T)}, \\ \|h_T^{-1}(\boldsymbol{w} - \overline{\boldsymbol{w}}_T)\|_{\boldsymbol{L}^2(\Delta_T)} &\preceq \|\boldsymbol{w}\|_{\boldsymbol{H}^1(\Delta_T)}, \end{aligned}$$

whence

$$\begin{aligned} \|\boldsymbol{v} - \boldsymbol{I}_{\mathscr{T}}\boldsymbol{v}\|^2_{\boldsymbol{L}^2(\Gamma)} &\preceq \sum_{T\in\mathscr{T}} \|\boldsymbol{w}\|^2_{\boldsymbol{L}^2(\Delta_T)} \preceq \|\boldsymbol{w}\|^2_{\boldsymbol{L}^2(\Gamma)}, \\ \|\mathtt{h}^{-1}(\boldsymbol{v} - \boldsymbol{I}_{\mathscr{T}}\boldsymbol{v})\|^2_{\boldsymbol{L}^2(\Gamma)} &\preceq \sum_{T\in\mathscr{T}} \|\boldsymbol{w}\|^2_{\boldsymbol{H}^1(\Delta_T)} \preceq \|\boldsymbol{w}\|^2_{\boldsymbol{H}^1(\Gamma)}. \end{aligned}$$

Applying Corollary 1 to vector-valued functions, we obtain

$$\|\mathtt{h}^{-\frac{1}{2}}(\boldsymbol{v} - \boldsymbol{I}_{\mathscr{T}}\boldsymbol{v})\|_{\boldsymbol{L}^2(\Gamma)} \preceq \|\boldsymbol{w}\|_{\boldsymbol{H}^{\frac{1}{2}}(\Gamma)} \preceq \|\boldsymbol{v}\|_{\boldsymbol{H}_{\perp}^{\frac{1}{2}}(\Gamma)},$$

where the last inequality results from Lemma 6. This concludes the proof.

3 Problem Setting

The variational formulation of the Electric Field Integral Equation (EFIE), also called Rumsey principle, consists of seeking $\boldsymbol{u} \in \boldsymbol{H}_{\mathrm{div}}^{-\frac{1}{2}}(\Gamma)$ such that

$$a(\boldsymbol{u}, \boldsymbol{v}) = {}_{\perp}\langle \boldsymbol{f}, \boldsymbol{v}\rangle_{\parallel} \qquad \text{for all}\ \boldsymbol{v} \in \boldsymbol{H}_{\mathrm{div}}^{-\frac{1}{2}}(\Gamma) \tag{16}$$

where$\boldsymbol{f} \in \boldsymbol{H}_{\mathrm{curl}}^{-\frac{1}{2}}(\Gamma)$, the sesquilinear form $a(\cdot,\cdot)$ is given by

$$a(\boldsymbol{u},\boldsymbol{v}) := \langle V_k \operatorname{div} \boldsymbol{u}, \operatorname{div} \boldsymbol{v}\rangle_{\frac{1}{2},\Gamma} - k^2 \langle \boldsymbol{A}_k \boldsymbol{u}, \boldsymbol{v}\rangle_{\parallel,\Gamma},$$

${}_{\perp}\langle\cdot,\cdot\rangle_{\parallel}$ is the duality pairing between $\boldsymbol{H}_{\mathrm{curl}}^{-\frac{1}{2}}(\Gamma)$ and $\boldsymbol{H}_{\mathrm{div}}^{-\frac{1}{2}}(\Gamma)$, $\langle\cdot,\cdot\rangle_{\frac{1}{2},\Gamma}$ is the duality pairing $H^{\frac{1}{2}}(\Gamma) - H^{-\frac{1}{2}}(\Gamma)$, $\langle\cdot,\cdot\rangle_{\parallel,\Gamma}$ is the duality pairing $\boldsymbol{H}_{\parallel}^{\frac{1}{2}}(\Gamma) - \boldsymbol{H}_{\parallel}^{-\frac{1}{2}}(\Gamma)$, and the integral operators $V_k, \boldsymbol{A}_k$ have been defined in Sect. 2.3.

The discrete formulation reads: find $\boldsymbol{U} \in \boldsymbol{RT}_0$ such that

$$a(\boldsymbol{U},\boldsymbol{V}) = {}_{\perp}\langle \boldsymbol{f}, \boldsymbol{V}\rangle_{\parallel} \qquad \text{for all } \boldsymbol{V} \in \boldsymbol{RT}_0. \tag{17}$$

The problem is well-posed for a sufficiently fine mesh, see [37]. As we want to quantify the approximation error a posteriori, we need to assume that $\boldsymbol{U}$ was computed, thus that the mesh satisfies the previous condition.

The continuous Eq. (16), on the other hand, is well-posed under the assumption that the wave number k does not correspond to an interior eigenmode of the Maxwell problem on Γ. As a consequence, the following continuous inf-sup condition holds (see also [37]):

$$\|\boldsymbol{u}\|_{\boldsymbol{H}_{\mathrm{div}}^{-\frac{1}{2}}(\Gamma)} \preceq \sup_{\boldsymbol{v}\in\boldsymbol{H}_{\mathrm{div}}^{-\frac{1}{2}}(\Gamma)} \frac{a(\boldsymbol{u},\boldsymbol{v})}{\|\boldsymbol{v}\|_{\boldsymbol{H}_{\mathrm{div}}^{-\frac{1}{2}}(\Gamma)}} \quad \text{for all } \boldsymbol{u} \in \boldsymbol{H}_{\mathrm{div}}^{-\frac{1}{2}}(\Gamma). \tag{18}$$

Since the boundary element discretization is conforming, i.e. $\boldsymbol{RT}_0 \subset \boldsymbol{H}_{\mathrm{div}}^{-\frac{1}{2}}(\Gamma)$, the following *Galerkin orthogonality* holds: If $\boldsymbol{u} \in \boldsymbol{H}_{\mathrm{div}}^{-\frac{1}{2}}(\Gamma)$ is the solution of (16) and $\boldsymbol{U} \in \boldsymbol{RT}_0$ is the solution of (17), then

$$a(\boldsymbol{u}-\boldsymbol{U},\boldsymbol{V}) = 0 \quad \text{for all } \boldsymbol{V} \in \boldsymbol{RT}_0. \tag{19}$$

In addition, as a direct consequence of the Cauchy-Schwarz inequality and the continuity (9) of the single layer boundary operators, the form $a(\cdot,\cdot)$ is continuous:

$$a(\boldsymbol{v},\boldsymbol{w}) \preceq \|\boldsymbol{v}\|_{\boldsymbol{H}_{\mathrm{div}}^{-\frac{1}{2}}(\Gamma)} \|\boldsymbol{w}\|_{\boldsymbol{H}_{\mathrm{div}}^{-\frac{1}{2}}(\Gamma)} \quad \text{for all } \boldsymbol{v},\boldsymbol{w} \in \boldsymbol{H}_{\mathrm{div}}^{-\frac{1}{2}}(\Gamma). \tag{20}$$

4 A Posteriori Error Analysis

As is customary in the theory of a posteriori error estimation, one has to assume a higher regularity of the right-hand side than it is needed for well-posedness in order to derive computable error bounds. Therefore we assume in this section that

$$\boldsymbol{f} \in \boldsymbol{H}_{\|}^{\frac{1}{2}}(\Gamma) \cap \boldsymbol{H}_{\mathrm{curl}}^{0}(\Gamma) \tag{21}$$

with $\boldsymbol{H}_{\|}^{\frac{1}{2}}(\Gamma)$ given in Proposition 1 and $H_{\mathrm{curl}}^{0}(\Gamma) = \left\{ v \in L^2(\Gamma) \,\middle|\, \mathbf{curl}_\Gamma\, v \in \boldsymbol{L}_t^2(\Gamma) \right\}$.

We proceed as in [23] for flat domains. To this end, we start with some auxiliary results that will be useful for our analysis later.

Lemma 8 (Regular decomposition [9, Theorem 5.5]) *The decomposition*

$$\boldsymbol{H}_{\mathrm{div}}^{-\frac{1}{2}}(\Gamma) = \mathbf{curl}_\Gamma (H^{\frac{1}{2}}(\Gamma)/\mathbb{C}) + \boldsymbol{H}_{\perp}^{\frac{1}{2}}(\Gamma)$$

is valid and is stable, i.e. for any $\boldsymbol{v} \in \boldsymbol{H}_{\mathrm{div}}^{-\frac{1}{2}}(\Gamma)$, *there exists* $\boldsymbol{\Psi} \in \boldsymbol{H}_{\perp}^{\frac{1}{2}}(\Gamma)$ *and* $\alpha \in H^{\frac{1}{2}}(\Gamma) \setminus \mathbb{C}$ *such that* $\boldsymbol{v} = \boldsymbol{\Psi} + \mathbf{curl}_\Gamma\, \alpha$ *and*

$$\|\boldsymbol{\Psi}\|_{\boldsymbol{H}_{\perp}^{\frac{1}{2}}(\Gamma)} + \|\alpha\|_{H^{\frac{1}{2}}(\Gamma)} \preceq \|\boldsymbol{v}\|_{\boldsymbol{H}_{\mathrm{div}}^{-\frac{1}{2}}(\Gamma)} \tag{22}$$

holds.

Lemma 9 *For* $\mathbb{V}(\mathscr{T})$ *given by* (12) *there holds* $\mathbf{curl}_\Gamma(\mathbb{V}(\mathscr{T})) \subset \boldsymbol{RT}_0$.

Proof By [9, Corollary 5.3] we have

$$\ker(\mathrm{div}) \cap \boldsymbol{L}_t^2(\Gamma) = \mathbf{curl}_\Gamma(H^1(\Gamma)).$$

Thus for all $\alpha \in \mathbb{V}(\mathscr{T}) \subset H^1(\Gamma)$ we infer that $\mathbf{curl}_\Gamma\, \alpha \in \boldsymbol{L}_t^2(\Gamma)$ is piecewise constant and that $\mathrm{div}\ \mathbf{curl}_\Gamma\, \alpha \equiv 0 \in L^2(\Gamma)$. This implies that $\mathbf{curl}_\Gamma\, \alpha \in \boldsymbol{H}_{\mathrm{div}}^{0}(\Gamma)$.

4.1 Upper Bound

Let $\boldsymbol{u} \in \boldsymbol{H}_{\mathrm{div}}^{-\frac{1}{2}}(\Gamma)$ be the exact solution of (16) and $\boldsymbol{U} \in \boldsymbol{RT}_0$ be its approximation defined by (17). By the Galerkin orthogonality we observe that

$$a(\boldsymbol{u} - \boldsymbol{U}, \boldsymbol{v}) = a(\boldsymbol{u} - \boldsymbol{U}, \boldsymbol{v} - \boldsymbol{V}) \quad \text{for all } \boldsymbol{v} \in \boldsymbol{H}_{\mathrm{div}}^{-\frac{1}{2}}(\Gamma),\ \boldsymbol{V} \in \boldsymbol{RT}_0.$$

Decompose $\boldsymbol{v}$ as $\boldsymbol{v} = \boldsymbol{\Psi} + \mathbf{curl}_\Gamma\, \alpha$, according to Lemma 8, and define

$$\delta\boldsymbol{\Psi} := \boldsymbol{\Psi} - \boldsymbol{\Psi}_{\mathscr{T}}, \quad \delta\alpha := \alpha - \alpha_{\mathscr{T}}$$

where $\boldsymbol{\Psi}_{\mathscr{T}} \in \boldsymbol{RT}_0$ and $\alpha_{\mathscr{T}} \in \mathbb{V}(\mathscr{T})$ can be arbitrarily chosen. Thus we can write $\boldsymbol{v} - \boldsymbol{V} = \delta\boldsymbol{\Psi} + \mathbf{curl}_\Gamma\, \delta\alpha$ and

$$
\begin{aligned}
a(\boldsymbol{u}-\boldsymbol{U}, \boldsymbol{v}-\boldsymbol{V}) &= {}_\perp\langle \boldsymbol{f}, \delta\boldsymbol{\Psi} + \mathbf{curl}_\Gamma\, \delta\alpha\rangle_\parallel - a(\boldsymbol{U}, \delta\boldsymbol{\Psi} + \mathbf{curl}_\Gamma\, \delta\alpha) \\
&= \underbrace{{}_\perp\langle \boldsymbol{f}, \delta\boldsymbol{\Psi}\rangle_\parallel + \langle k^2 \boldsymbol{A}_k \boldsymbol{U}, \delta\boldsymbol{\Psi}\rangle_{\parallel,\Gamma}}_{=\mathscr{I}_1} \\
&\quad + \underbrace{{}_\perp\langle \boldsymbol{f}, \mathbf{curl}_\Gamma\, \delta\alpha\rangle_\parallel + \langle k^2 \boldsymbol{A}_k \boldsymbol{U}, \mathbf{curl}_\Gamma\, \delta\alpha\rangle_{\parallel,\Gamma}}_{=\mathscr{I}_2} - \underbrace{\langle V_k \operatorname{div} \boldsymbol{U}, \operatorname{div} \delta\boldsymbol{\Psi}\rangle_{\frac{1}{2},\Gamma}}_{=\mathscr{I}_3}
\end{aligned}
$$

for any $\boldsymbol{v} \in \boldsymbol{H}_{\mathrm{div}}^{-\frac{1}{2}}(\Gamma)$, $\boldsymbol{\Psi}_{\mathscr{T}} \in \boldsymbol{RT}_0$ and $\alpha_{\mathscr{T}} \in \mathbb{V}(\mathscr{T})$. We proceed in four steps:

Step 1 We note that $\boldsymbol{f} \in \boldsymbol{L}_t^2(\Gamma)$, $k^2 \boldsymbol{A}_k \boldsymbol{U} \in \boldsymbol{H}_\parallel^{\frac{1}{2}}(\Gamma) \subset \boldsymbol{L}_t^2(\Gamma)$ and that $\boldsymbol{\Psi} \in \boldsymbol{H}_\perp^{\frac{1}{2}}(\Gamma) \subset \boldsymbol{L}_t^2(\Gamma)$ due to enhanced regularity of $\boldsymbol{\Psi}$ asserted in Lemma 8. Since $\boldsymbol{\Psi}_{\mathscr{T}} \in \boldsymbol{RT}_0 \subset \boldsymbol{L}_t^2(\Gamma)$, we can replace the duality pairing in $\mathscr{I}_1$ by an integral and thus write

$$
\mathscr{I}_1 = \int_\Gamma (\boldsymbol{f} + k^2 \boldsymbol{A}_k \boldsymbol{U}) \cdot \delta\boldsymbol{\Psi}\, \mathrm{d}\sigma. \tag{23}
$$

Step 2 Since $\boldsymbol{f} \in \boldsymbol{H}_\parallel^{\frac{1}{2}}(\Gamma)$ the duality pairing ${}_\perp\langle\cdot,\cdot\rangle_\parallel$ can be interpreted as

$$
{}_\perp\langle \boldsymbol{f}, \mathbf{curl}_\Gamma\, \delta\alpha\rangle_\parallel = \langle \boldsymbol{f}, \mathbf{curl}_\Gamma\, \delta\alpha\rangle_{\parallel,\Gamma},
$$

namely as a duality pairing in $\boldsymbol{H}_\parallel^{\frac{1}{2}}(\Gamma)$. The definition (8) of $\operatorname{curl}_\Gamma$ now yields

$$
\mathscr{I}_2 = \langle \boldsymbol{f} + k^2 \boldsymbol{A}_k \boldsymbol{U}, \mathbf{curl}_\Gamma\, \delta\alpha\rangle_{\parallel,\Gamma} = \langle \operatorname{curl}_\Gamma(\boldsymbol{f} + k^2 \boldsymbol{A}_k \boldsymbol{U}), \delta\alpha\rangle_{\frac{1}{2},\Gamma}.
$$

Since $\delta\alpha \in H^{\frac{1}{2}}(\Gamma)$ and $\operatorname{curl}_\Gamma(\boldsymbol{f} + k^2 \boldsymbol{A}_k \boldsymbol{U}) \in L^2(\Gamma)$ because of (11) and (21), we can also write $\mathscr{I}_2$ as an integral

$$
\mathscr{I}_2 = \int_\Gamma \operatorname{curl}_\Gamma(\boldsymbol{f} + k^2 \boldsymbol{A}_k \boldsymbol{U})\, \delta\alpha\, \mathrm{d}\sigma. \tag{24}
$$

Step 3 For the last term $\mathscr{I}_3$, we integrate by parts according to (8), whence

$$
\mathscr{I}_3 = -\langle \mathbf{grad}_\Gamma(V_k \operatorname{div} \boldsymbol{U}), \delta\boldsymbol{\Psi}\rangle_{\perp,\Gamma}.
$$

Since $\operatorname{div}(\boldsymbol{RT}_0) \subset L^2(\Gamma)$ we infer that $\mathbf{grad}_\Gamma(V_k \operatorname{div} \boldsymbol{U}) \in \boldsymbol{L}_t^2(\Gamma)$ in light of (10). This implies that $\mathscr{I}_3$ is also an integral

$$
\mathscr{I}_3 = -\int_\Gamma \mathbf{grad}_\Gamma(V_k \operatorname{div} \boldsymbol{U}) \cdot \delta\boldsymbol{\Psi}\, \mathrm{d}\sigma. \tag{25}
$$

Step 4 Inserting (23)–(25) back into the sesquilinear form a yields

$$
a(\boldsymbol{u}-\boldsymbol{U}, \boldsymbol{v}) = \int_\Gamma \boldsymbol{R} \cdot \delta\boldsymbol{\Psi}\, \mathrm{d}\sigma + \int_\Gamma r\, \delta\alpha\, \mathrm{d}\sigma \quad \text{for all } \boldsymbol{v} \in \boldsymbol{H}_{\mathrm{div}}^{-\frac{1}{2}}(\Gamma), \tag{26}
$$

where $\boldsymbol{R} \in \boldsymbol{L}_t^2(\Gamma)$ and $r \in L^2(\Gamma)$ are given element-by-element by

$$\begin{aligned} \boldsymbol{R}|_T &:= \boldsymbol{f} + k^2 \boldsymbol{A}_k \boldsymbol{U} + \mathbf{grad}_\Gamma (V_k \operatorname{div} \boldsymbol{U}) && \text{for all } T \in \mathscr{T}, \\ r|_T &:= \operatorname{curl}_\Gamma (\boldsymbol{f} + k^2 \boldsymbol{A}_k \boldsymbol{U}) && \text{for all } T \in \mathscr{T}. \end{aligned} \tag{27}$$

We now choose $\alpha_{\mathscr{T}} = \mathrm{I}_{\mathscr{T}} \alpha$ and $\boldsymbol{\Psi}_{\mathscr{T}} = \boldsymbol{I}_{\mathscr{T}} \boldsymbol{\Psi}$ where $\mathrm{I}_{\mathscr{T}}$ and $\boldsymbol{I}_{\mathscr{T}}$ are the interpolation operators of Definitions 1 and 2. Applying the Cauchy-Schwarz inequality and the interpolation estimates (13) and (15) yields

$$\begin{aligned} a(\boldsymbol{u} - \boldsymbol{U}, \boldsymbol{v}) &\leq \|\mathrm{h}^{\frac{1}{2}} \boldsymbol{R}\|_{L^2(\Gamma)} \|\mathrm{h}^{-\frac{1}{2}} \delta \boldsymbol{\Psi}\|_{L^2(\Gamma)} + \|\mathrm{h}^{\frac{1}{2}} r\|_{L^2(\Gamma)} \|\mathrm{h}^{-\frac{1}{2}} \delta \alpha\|_{L^2(\Gamma)} \\ &\preceq \|\mathrm{h}^{\frac{1}{2}} \boldsymbol{R}\|_{L^2(\Gamma)} \|\boldsymbol{\Psi}\|_{\boldsymbol{H}_\perp^{\frac{1}{2}}(\Gamma)} + \|\mathrm{h}^{\frac{1}{2}} r\|_{L^2(\Gamma)} \|\alpha\|_{H^{\frac{1}{2}}(\Gamma)}, \end{aligned}$$

which together with the stability (22) of the regular decomposition leads to

$$a(\boldsymbol{u} - \boldsymbol{U}, \boldsymbol{v}) \preceq \left(\|\mathrm{h}^{\frac{1}{2}} \boldsymbol{R}\|_{L^2(\Gamma)} + \|\mathrm{h}^{\frac{1}{2}} r\|_{L^2(\Gamma)} \right) \|\boldsymbol{v}\|_{\boldsymbol{H}_{\operatorname{div}}^{-\frac{1}{2}}(\Gamma)}.$$

Combining this with the inf-sup condition (18) finally implies

$$\|\boldsymbol{u} - \boldsymbol{U}\|_{\boldsymbol{H}_{\operatorname{div}}^{-\frac{1}{2}}(\Gamma)} \preceq \sup_{\boldsymbol{v} \in \boldsymbol{H}_{\operatorname{div}}^{-\frac{1}{2}}(\Gamma)} \frac{a(\boldsymbol{u} - \boldsymbol{U}, \boldsymbol{v})}{\|\boldsymbol{v}\|_{\boldsymbol{H}_{\operatorname{div}}^{-\frac{1}{2}}(\Gamma)}} \preceq \|\mathrm{h}^{\frac{1}{2}} \boldsymbol{R}\|_\Gamma + \|\mathrm{h}^{\frac{1}{2}} r\|_\Gamma.$$

We summarize this derivation in the following theorem.

Theorem 2 (Upper bound) *Let $\boldsymbol{f} \in \boldsymbol{H}_\parallel^{\frac{1}{2}}(\Gamma) \cap \boldsymbol{H}_{\operatorname{curl}}^0(\Gamma)$, $\boldsymbol{u} \in \boldsymbol{H}_{\operatorname{div}}^{-\frac{1}{2}}(\Gamma)$ be the exact solution of* (16) *and $\boldsymbol{U} \in \boldsymbol{RT}_0$ be its approximation defined by* (17)*. Then, there exists a constant $C_1 > 0$ depending on shape regularity of $\mathscr{T}$ such that the following bound holds*

$$\|\boldsymbol{u} - \boldsymbol{U}\|^2_{\boldsymbol{H}_{\operatorname{div}}^{-\frac{1}{2}}(\Gamma)} \leq C_1 \sum_{T \in \mathscr{T}} \eta^2_{\mathscr{T}}(T)$$

where the element indicators $\eta_{\mathscr{T}}(T)$ are defined as follows in terms of the residuals $\boldsymbol{R} \in \boldsymbol{L}_t^2(\Gamma)$ and $r \in L^2(\Gamma)$ given in (27)

$$\eta^2_{\mathscr{T}}(T) := h_T \|\boldsymbol{R}\|^2_{\boldsymbol{L}^2(T)} + h_T \|r\|^2_{L^2(T)}.$$

Remark 4 (Trace regularity of an incident plane wave) If the right-hand side $\boldsymbol{f}$ is the tangential trace of a plane wave $\boldsymbol{E}_{inc}$, then we conclude from the analyticity of the plane wave and of all its derivatives that

$$\boldsymbol{f} = \boldsymbol{\gamma}_\parallel(\boldsymbol{E}_{inc}) \in \boldsymbol{H}_\parallel^{\frac{1}{2}}(\Gamma), \quad \boldsymbol{\gamma}_\parallel(\partial_{x_i} \boldsymbol{E}_{inc}) \in \boldsymbol{H}_\parallel^{\frac{1}{2}}(\Gamma)$$

for $i = 1, 2, 3$. Therefore, $\boldsymbol{f}$ satisfies the stated regularity assumption (21).

4.2 Lower Bound

We next show *global* lower bounds for the error indicators $\eta^2_{\mathscr{T}}(T)$. Since $\boldsymbol{R} \in \boldsymbol{L}^2_t(\Gamma)$ and $r \in L^2(\Gamma)$ we define the local constants

$$\boldsymbol{R}_T = \int_T \boldsymbol{R}(\boldsymbol{x})\mathrm{d}\sigma(\boldsymbol{x}), \quad r_T = \int_T r(\boldsymbol{x})\mathrm{d}\sigma(\boldsymbol{x}), \quad \text{for all } T \in \mathscr{T},$$

and their global piecewise constant counterparts $\boldsymbol{R}_0|_T = \boldsymbol{R}_T$ and $r_0|_T = r_T$.

Theorem 3 (Global lower bound for the residual) *Let $\boldsymbol{u} \in \boldsymbol{H}^{-\frac{1}{2}}_{\mathrm{div}}(\Gamma)$ be the exact solution of* (16) *and $\boldsymbol{U} \in \boldsymbol{RT}_0$ be its approximation defined by* (17)*. Then, there exists a constant $C_2 > 0$, only depending on shape regularity of $\mathscr{T}$, such that the following bound holds*

$$C_2\|\mathrm{h}^{\frac{1}{2}}\boldsymbol{R}\|_{\boldsymbol{L}^2(\Gamma)} \leq \|\boldsymbol{u} - \boldsymbol{U}\|_{\boldsymbol{H}^{-\frac{1}{2}}_{\mathrm{div}}(\Gamma)} + \|\mathrm{h}^{\frac{1}{2}}(\boldsymbol{R} - \boldsymbol{R}_0)\|_{\boldsymbol{L}^2(\Gamma)}.$$

Proof The proof is based on the bubble-technique which was introduced by [49]. Let $b_T : \Gamma \to \mathbb{R}$ be a bubble function, namely a Lipschitz function so that

$$\text{supp } b_T \subset T, \qquad \int_T b_T\, d\boldsymbol{x} = |T| \approx \int_T b_T^2\, d\boldsymbol{x},$$

for a given $T \in \mathscr{T}$. Such a function can be given by a polynomial of degree three on T consisting of the product of all three barycentric coordinates times a real scaling factor. In consequence there holds

$$\int_T \mathrm{div}(\boldsymbol{\sigma}_T b_T)\mathrm{d}\sigma = \int_{\partial T} b_T \boldsymbol{\sigma}_T \cdot \boldsymbol{n}_T \mathrm{d}s = 0, \tag{28}$$

for any $\boldsymbol{\sigma}_T \in \mathbb{C}^2$. Let $\boldsymbol{\Psi}_T = h_T \boldsymbol{R}_T b_T$ and note that

$$\int_T \boldsymbol{R}_T \cdot \boldsymbol{\Psi}_T\, d\boldsymbol{x} = h_T\|\boldsymbol{R}_T\|^2_{\boldsymbol{L}^2(T)}$$

and

$$\|\boldsymbol{\Psi}_T\|_{\boldsymbol{L}^2(T)} \preceq h_T\|\boldsymbol{R}_T\|_{\boldsymbol{L}^2(T)} \preceq \|\boldsymbol{\Psi}_T\|_{\boldsymbol{L}^2(T)}.$$

Let $\boldsymbol{\Psi}$ denote the function that is defined element-wise by $\boldsymbol{\Psi}|_T = \boldsymbol{\Psi}_T$ for all $T \in \mathscr{T}$. We claim that $\boldsymbol{\Psi} \in \boldsymbol{H}^{\frac{1}{2}}_{\perp}(\Gamma)$ because it is made of piecewise polynomials with vanishing normal component on the interelement boundaries of $\mathscr{T}$. In view of Lemma 8, such a $\boldsymbol{\Psi}$ is an admissible test function in (26) and, together with the choices $\boldsymbol{\Psi}_{\mathscr{T}} = 0$ and $\alpha = \alpha_{\mathscr{T}} = 0$, yields

$$a(\boldsymbol{u}-\boldsymbol{U},\boldsymbol{\Psi}) = \int_{\Gamma} \boldsymbol{R}\cdot\boldsymbol{\Psi}\,\mathrm{d}\sigma = \int_{\Gamma} (\boldsymbol{R}-\boldsymbol{R}_0)\cdot\boldsymbol{\Psi}\,\mathrm{d}\sigma + \|\mathrm{h}^{\frac{1}{2}}\boldsymbol{R}_0\|^2_{\boldsymbol{L}^2(\Gamma)}.$$

By the continuity of the sesquilinear form $a(\cdot,\cdot)$, we have

$$\begin{aligned}\|\mathrm{h}^{\frac{1}{2}}\boldsymbol{R}_0\|^2_{\boldsymbol{L}^2(\Gamma)} &= a(\boldsymbol{u}-\boldsymbol{U},\boldsymbol{\Psi}) - \int_{\Gamma}(\boldsymbol{R}-\boldsymbol{R}_0)\cdot\boldsymbol{\Psi}\,d\boldsymbol{x}\\ &\preceq \|\boldsymbol{u}-\boldsymbol{U}\|_{\boldsymbol{H}_{\mathrm{div}}^{-\frac{1}{2}}(\Gamma)}\|\boldsymbol{\Psi}\|_{\boldsymbol{H}_{\mathrm{div}}^{-\frac{1}{2}}(\Gamma)} + \|\mathrm{h}^{\frac{1}{2}}(\boldsymbol{R}-\boldsymbol{R}_0)\|_{\boldsymbol{L}^2(\Gamma)}\|\mathrm{h}^{\frac{1}{2}}\boldsymbol{R}_0\|_{\boldsymbol{L}^2(\Gamma)}.\end{aligned}\tag{29}$$

It remains to estimate $\|\boldsymbol{\Psi}\|_{\boldsymbol{H}_{\mathrm{div}}^{-\frac{1}{2}}(\Gamma)}$. For $\varphi \in H^{\frac{1}{2}}(\Gamma)$, let φ_0 denote the elementwise average of φ. Applying Corollary 1 yields the interpolation estimate

$$\|\mathrm{h}^{-\frac{1}{2}}(\varphi-\varphi_0)\|_{\boldsymbol{L}^2(\Gamma)} \preceq |\varphi|_{H^{\frac{1}{2}}(\Gamma)}.$$

This, in conjunction with (28), implies

$$\begin{aligned}\langle \operatorname{div}\boldsymbol{\Psi}_T,\varphi\rangle_{\frac{1}{2},\Gamma} = \int_{\Gamma}\operatorname{div}\boldsymbol{\Psi}(\varphi-\varphi_0)\,d\boldsymbol{x} &\preceq \|\mathrm{h}^{\frac{1}{2}}\operatorname{div}\boldsymbol{\Psi}\|_{L^2(\Gamma)}|\varphi|_{H^{\frac{1}{2}}(\Gamma)}\\ &\preceq \|\mathrm{h}^{-\frac{1}{2}}\boldsymbol{\Psi}\|_{\boldsymbol{L}^2(\Gamma)}|\varphi|_{H^{\frac{1}{2}}(\Gamma)} \preceq \|\mathrm{h}^{\frac{1}{2}}\boldsymbol{R}_0\|_{\boldsymbol{L}^2(\Gamma)}|\varphi|_{H^{\frac{1}{2}}(\Gamma)}\end{aligned}$$

because of the norm equivalence for the discrete function $\boldsymbol{\Psi}$. Now, by definition

$$\|\operatorname{div}\boldsymbol{\Psi}\|_{H^{-\frac{1}{2}}(\Gamma)} = \sup_{\varphi\in H^{\frac{1}{2}}(\Gamma)}\frac{\langle\operatorname{div}\boldsymbol{\Psi},\varphi\rangle_{\frac{1}{2},\Gamma}}{|\varphi|_{H^{\frac{1}{2}}(\Gamma)}} \preceq \|\mathrm{h}^{\frac{1}{2}}\boldsymbol{R}_0\|_{\boldsymbol{L}^2(\Gamma)},$$

and

$$\|\boldsymbol{\Psi}\|_{\boldsymbol{H}_{\|}^{-\frac{1}{2}}(\Gamma)} \le \|\boldsymbol{\Psi}\|_{\boldsymbol{L}^2(\Gamma)} \preceq \|\mathrm{h}\boldsymbol{R}_0\|_{\boldsymbol{L}^2(\Gamma)}.$$

Consequently

$$\|\boldsymbol{\Psi}\|_{\boldsymbol{H}_{\mathrm{div}}^{-\frac{1}{2}}(\Gamma)} \preceq \|\mathrm{h}^{\frac{1}{2}}\boldsymbol{R}_0\|_{\boldsymbol{L}^2(\Gamma)}$$

which together with (29) implies that

$$\|\mathrm{h}^{\frac{1}{2}}\boldsymbol{R}_0\|_{\boldsymbol{L}^2(\Gamma)} \preceq \|\boldsymbol{u}-\boldsymbol{U}\|_{\boldsymbol{H}_{\mathrm{div}}^{-\frac{1}{2}}(\Gamma)} + \|\mathrm{h}^{\frac{1}{2}}(\boldsymbol{R}-\boldsymbol{R}_0)\|_{\boldsymbol{L}^2(\Gamma)}.$$

Invoking the triangle inequality finally finishes the proof.

It is important to realize the global nature of the above lower bound. This is due to the presence of integral operators $V_k, \boldsymbol{A}_k$ in the sesquilinear form $a(\cdot,\cdot)$ which lead to a global norm for the error in (29) regardless of the support of $\boldsymbol{\Psi}$.

In a very similar fashion, the following theorem can also be proven.

Theorem 4 (Global lower bound for the curl residual) *Let* $\boldsymbol{u} \in \boldsymbol{H}_{\mathrm{div}}^{-\frac{1}{2}}(\Gamma)$ *be the exact solution of* (16) *and* $\boldsymbol{U} \in \boldsymbol{RT}_0$ *be its approximation defined by* (17)*. Then, there exists a constant* $C_3 > 0$*, only depending on shape regularity of* $\mathscr{T}$*, such that the following bound holds*

$$C_3 \|\mathrm{h}^{\frac{1}{2}}\, r\|_{L^2(\Gamma)} \le \|\boldsymbol{u} - \boldsymbol{U}\|_{\boldsymbol{H}_{\mathrm{div}}^{-\frac{1}{2}}(\Gamma)} + \|\mathrm{h}^{\frac{1}{2}}(r - r_0)\|_{L^2(\Gamma)}.$$

5 Conclusions

In this paper we develop the first a posteriori error estimates for the electric field integral equation on polyhedra. We choose, for simplicity, to derive residual based error estimates but believe that our theory extends to other non-residual estimators. We also choose to develop the theory for polyhedra, the most interesting and useful case in practice, but we expect the results to extend to smooth surfaces. For scattering problems on polyhedra, the solution $\boldsymbol{u}$ of the integral equation, or surface current, is not smooth whereas the regularity of the right-hand side $\boldsymbol{f}$ is dictated by the surface Γ because the incident wave is always smooth. This justifies our additional regularity assumption (21) which, coupled with the properties $\mathbf{grad}_\Gamma(V_k \operatorname{div} \boldsymbol{U}) \in L^2(\Gamma)$, $\mathrm{curl}_\Gamma(\boldsymbol{A}_k \boldsymbol{U}) \in L^2(\Gamma)$, allows us to evaluate the residuals $\boldsymbol{R}$, r of (27) in $L^2(\Gamma)$ and thus avoid dealing with fractional Sobolev norms. We derive computable global upper and lower a posteriori bounds for the estimator (up to oscillation terms). In contrast to PDE, the estimator is global due to the presence of the potentials $V_k, \boldsymbol{A}_k$ in the definition of the sesquilinear form. However, the residuals $\boldsymbol{R}$, r being evaluated in $L^2(\Gamma)$ can be split elementwise and used to drive an adaptive boundary element method (ABEM). The actual implementation of ABEM for EFIE is rather delicate and is not part of the current discussion, which focusses on the derivation and properties of the estimators.

References

1. Alonso A (1996) Error estimators for a mixed method. Numer Math 74(4):385–395
2. Aurada M, Ferraz-Leite S, Praetorius D (2012) Estimator reduction and convergence of adaptive BEM. Appl Numer Math 62(6):787–801
3. Beck R, Hiptmair R, Hoppe R, Wohlmuth B (2000) Residual based a posteriori error estimators for eddy current computation. M2AN Math Model Numer Anal 34(1):159–182
4. Bernardi C, Hecht F (2007) Quelques propriétés d'approximation des éléments finis de Nédélec, application à l'analyse a posteriori. C R Math Acad Sci Paris 344(7):461–466
5. Brakhage H, Werner P (1965) Über das Dirichletsche Aussenraumproblem für die Helmholtzsche Schwingungsgleichung. Arch Math 16:325–329
6. Buffa A, Ciarlet P Jr (2001) On traces for functional spaces related to Maxwell's equations. I. An integration by parts formula in Lipschitz polyhedra. Math Methods Appl Sci 24(1):9–30
7. Buffa A, Ciarlet P Jr (2001) On traces for functional spaces related to Maxwell's equations. II. Hodge decompositions on the boundary of Lipschitz polyhedra and applications. Math Methods Appl Sci 24(1):31–48

8. Buffa A, Costabel M, Schwab C (2002) Boundary element methods for Maxwell's equations on non-smooth domains. Numer Math 92(4):679–710
9. Buffa A, Costabel M, Sheen D (2002) On traces for $H(\mathrm{curl}, \Omega)$ in Lipschitz domains. J Math Anal Appl 276(2):845–867
10. Buffa A, Hiptmair R (2004) A coercive combined field integral equation for electromagnetic scattering. SIAM J Numer Anal 42(2):621–640
11. Buffa A, Hiptmair R (2005) Regularized combined field integral equations. Numer Math 100(1):1–19
12. Buffa A, Hiptmair R, von Petersdorff T, Schwab C (2003) Boundary element methods for Maxwell transmission problems in Lipschitz domains. Numer Math 95(3):459–485
13. Burton AJ, Miller GF (1971) The application of integral equation methods to the numerical solution of some exterior boundary-value problems. Proc Roy Soc London Ser A 323(1553):201–210
14. Carstensen C (1996) Efficiency of a posteriori BEM-error estimates for first-kind integral equations on quasi-uniform meshes. Math Comp 65(213):69–84
15. Carstensen C (1997) An a posteriori error estimate for a first-kind integral equation. Math Comp 66(217):139–155
16. Carstensen C, Faermann B (2001) Mathematical foundation of a posteriori error estimates and adaptive mesh-refining algorithms for boundary integral equations of the first kind. Eng Anal Bound Elem 25(7):497–509
17. Carstensen C, Funken SA, Stephan EP (1996) A posteriori error estimates for *hp*-boundary element methods. Appl Anal 61(3–4):233–253
18. Carstensen C, Maischak M, Praetorius D, Stephan EP (2004) Residual-based a posteriori error estimate for hypersingular equation on surfaces. Numer Math 97(3):397–425
19. Carstensen C, Maischak M, Stephan EP (2001) A posteriori error estimate and h-adaptive algorithm on surfaces for Symm's integral equation. Numer Math 90(2):197–213
20. Carstensen C, Praetorius D (2012) Convergence of adaptive boundary element methods. J Integr Eqn Appl 24(1):1–23
21. Carstensen C, Stephan EP (1995) A posteriori error estimates for boundary element methods. Math Comp 64(210):483–500
22. Cascon JM, Kreuzer C, Nochetto RH, Siebert KG (2008) Quasi-optimal convergence rate for an adaptive finite element method. SIAM J Numer Anal 46(5):2524–2550
23. Cascon JM, Nochetto RH, Siebert KG (2007) Design and convergence of AFEM in $H(\mathrm{div})$. Math Models Methods Appl Sci 17(11):1849–1881
24. Christiansen SH (2004) Discrete Fredholm properties and convergence estimates for the electric field integral equation. Math Comp 73(245):143–167
25. Clément P (1975) Approximation by finite element functions using local regularization. Rev Française Automat Informat Recherche Opérationnelle Sér RAIRO Anal Numér 9(R–2):77–84
26. Colton D, Kress R (1998) Inverse acoustic and electromagnetic scattering theory, vol 93 of applied mathematical sciences, 2nd edn. Springer, Berlin
27. Demlow A, Dziuk G (2007) An adaptive finite element method for the Laplace-Beltrami operator on implicitly defined surfaces. SIAM J Numer Anal 45(1):421–442
28. Erath C, Ferraz-Leite S, Funken S, Praetorius D (2009) Energy norm based a posteriori error estimation for boundary element methods in two dimensions. Appl Numer Math 59(11):2713–2734
29. Ern A, Guermond J-L (2004) Theory and practice of finite elements, vol 159. Applied mathematical sciences. Springer, New York
30. Feischl M, Karkulik M, Melenk JM, Praetorius D (2013) Quasi-optimal convergence rate for an adaptive boundary element method. SIAM J Numer Anal 51(2):1327–1348
31. Ferraz-Leite S, Ortner C, Praetorius D (2010) Convergence of simple adaptive Galerkin schemes based on $h - h/2$ error estimators. Numer Math 116(2):291–316
32. Ferraz-Leite S, Praetorius D (2008) Simple a posteriori error estimators for the h-version of the boundary element method. Computing 83(4):135–162

33. Greengard L, Rokhlin V (1987) A fast algorithm for particle simulations. J Comput Phys 73(2):325–348
34. Greengard L, Rokhlin V (1997) A new version of the fast multipole method for the Laplace equation in three dimensions. Acta Numer 6:229–269
35. Hackbusch W, Nowak ZP (1989) On the fast matrix multiplication in the boundary element method by panel clustering. Numer Math 54(4):463–491
36. Hackbush W, Sauter SA (1993) On the efficient use of the Galerkin method to solve Fredholm integral equations. Appl Math 38(4–5):301–322 (Proceedings of ISNA '92 – International Symposium on Numerical Analysis, Part I (Prague, 1992))
37. Hiptmair R, Schwab C (2002) Natural boundary element methods for the electric field integral equation on polyhedra. SIAM J Numer Anal 40(1):66–86
38. Leis R (1965) Zur Dirichletschen Randwertaufgabe des Aussenraumes der Schwingungsgleichung. Math Z 90(3):205–211
39. Leydecker F, Maischak M, Stephan EP, Teltscher M (2010) Adaptive FE-BE coupling for an electromagnetic problem in $\mathbb{R}^3$ - A residual error estimator. Math Methods Appl Sci 33(18):2162–2186
40. Nochetto RH, Siebert KG, Veeser A (2009) Theory of adaptive finite element methods: an introduction. Multiscale. Nonlinear and Adaptive Approximation. Springer, Berlin, pp 409–542
41. Nochetto RH, von Petersdorff T, Zhang C-S (2010) A posteriori error analysis for a class of integral equations and variational inequalities. Numer Math 116(3):519–552
42. Nowak ZP, Hackbusch W (1986) Complexity of the method of panels. In: Marchuk G (ed) Computational processes and systems, No. 6. Nauka, Moscow, pp 233–244 (in Russian)
43. Panich OI (1965) On the solubility of exterior boundary-value problems for the wave equation and for a system of Maxwell's equations. Uspekhi Mat. Nauk 20:1(121):221–226
44. Raviart P-A, Thomas JM (1977) A mixed finite element method for 2nd order elliptic problems. In: Mathematical aspects of finite element methods (Proceedings of Conference, Consiglio Naz. delle Ricerche (C.N.R.), Rome, 1975), vol 606 of lecture notes in mathematics. Springer, Berlin, pp 292–315
45. Rokhlin V (1985) Rapid solution of integral equations of classical potential theory. J Comput Phys 60(2):187–207
46. Sauter SA, Schwab C (2011) Boundary element methods, vol 39. Springer series in computational mathematics. Springer, Berlin
47. Tartar L (2007) An introduction to Sobolev spaces and interpolation spaces, vol 3. Lecture Notes of the Unione Matematica Italiana. Springer, Berlin
48. Teltscher M, Stephan EP, Maischak M (2003) A residual error estimator for an electromagnetic FEM-BEM coupling problem in $\mathbb{R}^3$. Technical report, Institut für Angewandte Mathematik, Universität Hannover
49. Verfürth R (1998) A posteriori error estimators for convection-diffusion equations. Numer Math 80(4):641–663

On Some Weighted Stokes Problems: Applications on Smagorinsky Models

Jacques Rappaz and Jonathan Rochat

Abstract In this paper we study existence and uniqueness of weak solutions for some non-linear weighted Stokes problems using convex analysis. The characterization of these equations is the viscosity, which depends on the strain rate of the velocity field and in some cases is related with a weight being the distance to the boundary of the domain. Such non-linear relations can be seen as a first approach of mixing-length eddy viscosity from turbulent modeling. A well known model is von Karman's on which the viscosity depends on the square of the distance to the boundary of the domain. Numerical experiments conclude the work and show properties from the theory.

Keywords Stokes equations · Weighted Sobolev spaces · Finite element method

Mathematical Subject Classification 46E35 · 76F55 · 65N05

1 Introduction

Turbulent flows have an importance in many domains, including technology and industry. While measurements are sometimes difficult to make, the use of numerical simulations of such flows in industries can be very useful to optimize activities and reduce the cost of products and process development. The Navier-Stokes equations offer an accurate description of these flows, whose Reynolds number is large. The resolution of these equations is then challenging as the mesh required to obtain most of the structure of these flows should be very thin.

To overcome these difficulties, many turbulent models appear such as Large Eddy Simulation (LES), mainly described in [18], that assumes that the inertial scales of the

J. Rappaz (✉) · J. Rochat
Ecole Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland
e-mail: jacques.rappaz@epfl.ch

J. Rochat
e-mail: jonathan.rochat@alumni.epfl.ch

B. N. Chetverushkin et al. (eds.), *Contributions to Partial Differential Equations and Applications*, Computational Methods in Applied Sciences 47,
https://doi.org/10.1007/978-3-319-78325-3_21

flow have been captured by an sufficiently thin grid using low-pass filtering. Another simpler approach we consider here is the Reynold Averaged Navier-Stokes (RANS). This kind of model computes a time-averaged mean value for velocity field, which has a significantly larger period than the turbulent fluctuations. A type of simple modeling often used by engineering is a mixing-length model called "Smagorinsky Modeling" (see [18]). In practice, these models consist in changing the initial viscosity of the fluid by a turbulent viscosity depending of the velocity, transforming the initial linear elliptic term in the Navier-Stokes equations by a non-linear one.

If $\mathbf{u}$ and p are the velocity and the pressure of a stationary incompressible fluid of density ρ, submitted to a force $\mathbf{f}$, flowing in a cavity $\Omega \subset \mathbb{R}^n$, $n = 2, 3$, with a Lipschitz boundary $\partial\Omega$, stationary Navier-Stokes equations on Ω take on the form

$$\begin{cases} -\operatorname{div}(2\mu\boldsymbol{\varepsilon}(\mathbf{u})) + \nabla p = \mathbf{F}(\mathbf{u}) & \text{in } \Omega, \\ \operatorname{div}\mathbf{u} = 0 & \text{in } \Omega, \end{cases} \tag{1}$$

with $\mathbf{u} = \mathbf{0}$ on $\partial\Omega$, $\boldsymbol{\varepsilon}(\mathbf{u}) = \frac{1}{2}\left(\nabla\mathbf{u} + \nabla\mathbf{u}^T\right)$ and $\mathbf{F}(\mathbf{u}) = \mathbf{f} - \rho(\mathbf{u}\cdot\nabla)\mathbf{u}$. In this paper we will treat Smagorinsky models in which the viscosity depends on $|\boldsymbol{\varepsilon}(\mathbf{u})|$ and takes the form

$$\mu(|\boldsymbol{\varepsilon}(\mathbf{u})|) = \mu_L + \kappa^\alpha \rho l^{2-\alpha} d^\alpha_{\partial\Omega}|\boldsymbol{\varepsilon}(\mathbf{u})| \tag{2}$$

where $\mu_L > 0$ corresponds to a laminar viscosity, $\kappa = 0.41$ is the von Karman constant, $l > 0$ is a characteristic length of the domain, $\alpha \geq 0$ is a real number, $d_{\partial\Omega}(x)$ is the distance of a point $x \in \Omega$ to the boundary $\partial\Omega$ and $|\boldsymbol{\varepsilon}(\mathbf{u})| = (\sum_{i,j} \varepsilon_{ij}(\mathbf{u})^2)^{\frac{1}{2}}$.

The cases with $\alpha = 0$ can be treated in usual Sobolev spaces and their analysis can be found in several papers [2, 20]. At the opposite, the cases with $\alpha > 0$ have to be studied in weighted Sobolev spaces and present several difficulties. In particular, we will give some comments on a very popular model for a fluid flow in between to close plates (von Karman model) in which $\alpha = 2$.

We proceed in this paper to an analysis of the problem (1) with a viscosity given by (2) and $\alpha < 2$. To do it, we have to start by considering the simpler Stokes problem with a given $\mathbf{F}$ function. By using several known results concerning weighted Sobolev spaces [4, 14, 16, 17], we establish some theoretical results on the existence and uniqueness of a velocity field of Eq. (1) when $\mathbf{F}$ is given. We show how is important the role of the laminar viscosity μ_L when Von Karman model is used and its impact on numerical results when we use a finite element method to discretize Problem (1). The uniqueness of the pressure is sometimes an open question.

2 Main Existence Theorem

In this section we prove that the problem (1) for a given $\mathbf{F}$ function has a unique solution related to velocity in a space with free divergence. Let Ω be an bounded open subset of $\mathbb{R}^n$, $n = 2, 3$, with a Lipschitz boundary $\partial\Omega$. We first introduce some adequate weighted functional spaces on Ω to define a weak problem from the Eq. (1) concerning the velocity. We then prove by convex analysis the existence of such a velocity field. The section finishes with some results concerning the pressure.

2.1 Suitable Functional Spaces

Let $d_{\partial\Omega}(x) = \text{dist}(x, \partial\Omega) = \min_{y\in\partial\Omega} |x - y|$ the distance between $x \in \Omega$ and $\partial\Omega$. For $1 \leq p < \infty$ and $\alpha > 0$, we denote the weighted Sobolev space of order one as

$$W^{1,p}(\Omega, d^{\alpha}_{\partial\Omega}) = \left\{ v \in L^p(\Omega, d^{\alpha}_{\partial\Omega}) \mid \frac{\partial v}{\partial x_i} \in L^p(\Omega, d^{\alpha}_{\partial\Omega}), \quad \forall i = 1, \ldots, n \right\}$$

where $L^p(\Omega, d^{\alpha}_{\partial\Omega}) = \left\{ v : \Omega \to \mathbb{R} \mid \int_{\Omega} |v|^p d^{\alpha}_{\partial\Omega} dx < \infty \right\}$ provided with norm $\|v\|_{L^p(\Omega, d^{\alpha}_{\partial\Omega})} := \left(\int_{\Omega} |v|^p d^{\alpha}_{\partial\Omega} dx \right)^{\frac{1}{p}}$. We thus endowed $W^{1,p}(\Omega, d^{\alpha}_{\partial\Omega})$ with the norm

$$\|v\|_{W^{1,p}(\Omega, d^{\alpha}_{\partial\Omega})} := \left(\int_{\Omega} |v|^p d^{\alpha}_{\partial\Omega} dx + \int_{\Omega} |\nabla v|^p d^{\alpha}_{\partial\Omega} dx \right)^{\frac{1}{p}}. \qquad (3)$$

Lemma 1 *For all $1 < p < \infty$ and $\alpha \geq 0$, $W^{1,p}(\Omega, d^{\alpha}_{\partial\Omega})$ endowed with the norm (3) is a reflexive Banach space.*

Proof The properties of spaces $W^{1,p}(\Omega, d^{\alpha}_{\partial\Omega})$ are deduced from the ones of the spaces $L^p(\Omega, d^{\alpha}_{\partial\Omega})$ (see [16] or Theorem 1.3 in [10]). The reflexivity is due to the uniform convexity of these spaces [3, Theorem III.29].

For arbitrary weight ω, the books [10, 16] give a well overview of these spaces that found applications in a large scale of problems such as $p-$Laplacian [7] or degenerated elliptic problem [6]. Generally, the chosen weight ω belongs to the Muckenhoupt class A_p (see [1, 5]). For weights which are a positive power of the distance to the boundary, they belong to the Muckenhoupt class if $0 \leq \alpha < p - 1$ (see [6]). Publications on the space generated with such weights are less frequent but some papers and books treat many properties of these spaces, see for example [4, 6, 17]. One of the important property is that the embedding

$$W^{1,p}(\Omega, d^{\alpha}_{\partial\Omega}) \hookrightarrow L^p(\Omega, d^{\alpha}_{\partial\Omega}) \qquad (4)$$

is continuous and compact, as it is shown in [17, Theorem 3.8].

Another characterization is that we can define a continuous and bounded trace operator $\text{Tr} : W^{1,p}(\Omega, d^{\alpha}_{\partial\Omega}) \to L^p(\partial\Omega)$ if $1 < p < \infty$ and $0 \leq \alpha < p - 1$ (see [16, Theorem 9.15]). In that case, the space $W^{1,p}_0(\Omega, d^{\alpha}_{\partial\Omega})$ (the closure of $C^{\infty}_0(\Omega)$ in $W^{1,p}(\Omega, d^{\alpha}_{\partial\Omega})$) for the norm (3)) can be identified with the space of functions in $W^{1,p}(\Omega, d^{\alpha}_{\partial\Omega})$ vanishing on the boundary:

$$W_0^{1,p}(\Omega, d_{\partial\Omega}^\alpha) = \{v \in W^{1,p}(\Omega, d_{\partial\Omega}^\alpha) \mid \mathrm{Tr}(v) = 0\}.$$

Moreover, as the problem (1) involves vector fields $\mathbf{u} : \Omega \to \mathbb{R}^n$, $\mathbf{u} = (u_1, \ldots, u_n)$, we denote the following norms for $\mathbf{u} \in [W_0^{1,p}(\Omega, d_{\partial\Omega}^\alpha)]^n$:

$$\|\mathbf{u}\|_{W_0^{1,p}(\Omega,d_{\partial\Omega}^\alpha)} := \left(\sum_{i=1}^n \|u_i\|^p_{W_0^{1,p}(\Omega,d_{\partial\Omega}^\alpha)}\right)^{\frac{1}{p}},$$

$$\|\nabla\mathbf{u}\|_{L^p(\Omega,d_{\partial\Omega}^\alpha)} := \left(\sum_{i,j=1}^n \|\frac{\partial u_i}{\partial x_j}\|^p_{L^p(\Omega,d_{\partial\Omega}^\alpha)}\right)^{\frac{1}{p}}, \quad \text{and}$$

$$\|\boldsymbol{\varepsilon}(\mathbf{u})\|_{L^p(\Omega,d_{\partial\Omega}^\alpha)} := \left(\sum_{i,j=1}^n \|\varepsilon_{ij}(\mathbf{u})\|^p_{L^p(\Omega,d_{\partial\Omega}^\alpha)}\right)^{\frac{1}{p}}.$$

These definitions and characterizations allow us to prove an important result:

Proposition 1 (Korn Inequality) *Let* $0 \le \alpha < p - 1$. *There exists a generic constant* $C > 0$ *such that*

$$\|\nabla\boldsymbol{u}\|_{L^p(\Omega,d_{\partial\Omega}^\alpha)} \le C\|\boldsymbol{\varepsilon}(\boldsymbol{u})\|_{L^p(\Omega,d_{\partial\Omega}^\alpha)}, \quad \forall \boldsymbol{u} \in [W_0^{1,p}(\Omega, d_{\partial\Omega}^\alpha)]^n. \tag{5}$$

Proof The structure of the proof follows mainly the procedure developed in [15]. First of all, Theorem 6 in [14] states for $-1 \le \alpha < p - 1$ the existence of a constant $C > 0$ such that

$$\|\nabla\mathbf{u}\|_{L^p(\Omega,d_{\partial\Omega}^\alpha)} \le C\left\{\|\mathbf{u}\|_{L^p(\Omega,d_{\partial\Omega}^\alpha)} + \|\boldsymbol{\varepsilon}(\mathbf{u})\|_{L^p(\Omega,d_{\partial\Omega}^\alpha)}\right\}, \quad \forall \mathbf{u} \in [W^{1,p}(\Omega, d_{\partial\Omega}^\alpha)]^n. \tag{6}$$

Consequently, it remains to prove that there exists a generic constant $C > 0$ such that

$$\|\mathbf{u}\|_{L^p(\Omega,d_{\partial\Omega}^\alpha)} \le C\|\boldsymbol{\varepsilon}(\mathbf{u})\|_{L^p(\Omega,d_{\partial\Omega}^\alpha)}, \quad \forall \mathbf{u} \in [W^{1,p}(\Omega, d_{\partial\Omega}^\alpha)]^n.$$

By contradiction, we assume that there exists a sequence $(\mathbf{u}_l)_{l=1}^\infty \in [W_0^{1,p}(\Omega, d_{\partial\Omega}^\alpha)]^n$ satisfying

$$\|\mathbf{u}_l\|_{L^p(\Omega,d_{\partial\Omega}^\alpha)} = 1 \quad \text{and} \quad \lim_{l\to\infty} \|\boldsymbol{\varepsilon}(\mathbf{u}_l)\|_{L^p(\Omega,d_{\partial\Omega}^\alpha)} = 0. \tag{7}$$

By using (6) and (7), the sequence $\{\mathbf{u}_l\}_l^\infty$ is bounded in $[W_0^{1,p}(\Omega, d_{\partial\Omega}^\alpha)]^n$ and by compacity (4), it is not restrictive to assume there exists $\mathbf{u} \in [W_0^{1,p}(\Omega, d_{\partial\Omega}^\alpha)]^n$ such that

$$\lim_{l\to\infty} \|\mathbf{u}_l - \mathbf{u}\|_{L^p(\Omega,d_{\partial\Omega}^\alpha)} = 0 \quad \text{and} \quad \mathbf{u}_l \rightharpoonup \mathbf{u} \text{ weakly in } [W_0^{1,p}(\Omega, d_{\partial\Omega}^\alpha)]^n. \tag{8}$$

Relations (7) and (8) imply $\boldsymbol{\varepsilon}(\mathbf{u}) = 0$. Then from [15], the function $\mathbf{u}$ belongs to a class of polynomial of degree one. Since $\mathbf{u}$ is vanishing on the boundary, then $\mathbf{u} \equiv 0$. This contradicts the fact that $\|\mathbf{u}\|_{L^p(\Omega, d^\alpha_{\partial\Omega})} = 1$.

Remark 1 In order to study Stokes problem (1) with viscosity (2), we will see below that we need to work with weighted Sobolev spaces of order p = 3. In this particular case, inequality (5) takes the following form: for $0 \leq \alpha < 2$, there exists a generic constant $C > 0$ such that

$$\|\nabla \mathbf{u}\|_{L^3(\Omega, d^\alpha_{\partial\Omega})} \leq C \|\boldsymbol{\varepsilon}(\mathbf{u})\|_{L^3(\Omega, d^\alpha_{\partial\Omega})}, \quad \forall \mathbf{u} \in [W_0^{1,3}(\Omega, d^\alpha_{\partial\Omega})]^n.$$

In the following we consider $0 \leq \alpha < 2$. The problem (1) involves homogeneous Dirichlet conditions and takes into account two viscosity terms $\operatorname{div}(2\mu_0 \boldsymbol{\varepsilon}(\mathbf{u}))$ and $\operatorname{div}(2\kappa^\alpha \rho l^{2-\alpha} d^\alpha_{\partial\Omega} |\boldsymbol{\varepsilon}(\mathbf{u})| \boldsymbol{\varepsilon}(\mathbf{u}))$, see (2). Consequently, when we will consider a weak formulation of the problem (1) (see Sect. 2.2) we have to work in the two following Banach spaces $H_0^1(\Omega)$ and $W_0^{1,3}(\Omega, d^\alpha_{\partial\Omega})$. Let us remark that there exists α_0 with $0 \leq \alpha_0 < 2$ such that $W_0^{1,3}(\Omega, d^\alpha_{\partial\Omega}) \subset H_0^1(\Omega)$ when $0 \leq \alpha < \alpha_0$ (see [16]) but it is not the case when α is close to 2. Thus, if we want to analyse von Karman model corresponding to $\alpha = 2$, we have to define the space

$$X_\alpha = H_0^1(\Omega) \cap W_0^{1,3}(\Omega, d^\alpha_{\partial\Omega})$$

endowed with the following norm $\|v\|_{X_\alpha} = \|v\|_{H^1(\Omega)} + \|v\|_{W^{1,3}(\Omega, d^\alpha_{\partial\Omega})}$.

Lemma 2 *The normed space* $(X_\alpha, \|\cdot\|_{X_\alpha})$ *is a reflexive Banach space.*

Proof The proof is a consequence of the compact embedding

$$H_0^1(\Omega) \hookrightarrow L^3(\Omega) \subset L^3(\Omega, d^\alpha_{\partial\Omega})$$

and (4). In particular, we prove that each bounded sequence in X_α has a weakly convergent subsequence in X_α.

Lemma 3 *The space* X_α *endowed with the semi-norm*

$$|\boldsymbol{v}|_{X_\alpha} := \|\nabla v\|_{L^2(\Omega)} + \|\nabla v\|_{L^3(\Omega, d^\alpha_{\partial\Omega})}$$

is a reflexive Banach space.

Proof The proof is a consequence of Lemma 2 and from the equivalence of the norm $\|\cdot\|_{X_\alpha}$ and the semi-norm $|\cdot|_{X_\alpha}$ since $H_0^1(\Omega) \hookrightarrow L^3(\Omega, d^\alpha_{\partial\Omega})$ and by Poincaré's inequality.

2.2 On the Velocity of Stokes Problem

In this section we consider the non-linear Stokes problem (1) with viscosity (2) in which $\mathbf{F} \in [L^{\frac{4}{3}}(\Omega)]^n$ does not depend on $\mathbf{u}$. The index α verifies $0 \leq \alpha < 2$.

As in the problem (1) we are looking for a free divergence velocity field, we take now the space

$$\mathbf{X}_{\alpha,\mathrm{div}} = \{\mathbf{v} \in X_\alpha^n \mid \mathrm{div}\,\mathbf{v} = 0\}$$

endowed with the norm $|\mathbf{v}|_{\mathbf{X}_{\alpha,\mathrm{div}}} = \|\nabla \mathbf{v}\|_{L^2(\Omega)} + \|\nabla \mathbf{v}\|_{L^3(\Omega, d_{\partial\Omega}^\alpha)}$. By multiplying the Stokes equation in problem (1) by a test velocity field $\mathbf{v} \in \mathbf{X}_{\alpha,\mathrm{div}}$ and integrating by part, we obtain a weak formulation of the problem (1)–(2) for the velocity:

Find $\mathbf{u} \in \mathbf{X}_{\alpha,\mathrm{div}}$ such that

$$\int_\Omega (2\mu(|\boldsymbol{\varepsilon}(\mathbf{u})|)\boldsymbol{\varepsilon}(\mathbf{u}) : \boldsymbol{\varepsilon}(\mathbf{v}))dx = \int_\Omega (\mathbf{F} \cdot \mathbf{v})dx, \quad \forall \mathbf{v} \in \mathbf{X}_{\alpha,\mathrm{div}}. \tag{9}$$

We use convex arguments to show existence and uniqueness of a solution to (9). Let us define the functional $J : X_{\alpha,\mathrm{div}} \to \mathbb{R}$ by:

$$J(\mathbf{u}) = \int_\Omega [2A(x, |\boldsymbol{\varepsilon}(\mathbf{u}(x))|) - \mathbf{F}(x) \cdot \mathbf{u}(x)]dx,$$

where $A : (x, s) \in \Omega \times \mathbb{R} \to A(x, s) \in \mathbb{R}$ is given by

$$A(x, s) = \frac{\mu_L}{2} s^2 + \frac{1}{3} \kappa^\alpha \rho l^{2-\alpha} d_{\partial\Omega}^\alpha(x) s^3.$$

Lemma 4 *The functional J is Gâteaux-differentiable and its derivative at $\boldsymbol{u}$ in the direction $\boldsymbol{v}$ is*

$$DJ_{\boldsymbol{u}}(\boldsymbol{v}) = \int_\Omega (2\mu(|\boldsymbol{\varepsilon}(\boldsymbol{u})|)\boldsymbol{\varepsilon}(\boldsymbol{u}) : \boldsymbol{\varepsilon}(\boldsymbol{v}))dx - \int_\Omega \boldsymbol{F} \cdot \boldsymbol{v}\, dx.$$

Proof It is easy to verify that for $\beta \geq 2$:

$$\lim_{t \to 0} \frac{|\boldsymbol{\varepsilon}(\mathbf{u} + t\mathbf{v})|^\beta - |\boldsymbol{\varepsilon}(\mathbf{u})|^\beta}{t} = \beta |\boldsymbol{\varepsilon}(\mathbf{u})|^{\beta-2} \boldsymbol{\varepsilon}(\mathbf{u}) : \boldsymbol{\varepsilon}(\mathbf{v}).$$

Taking in account that

$$\frac{\partial}{\partial s} A(x, s) = \mu_L s + \kappa^\alpha \rho l^{2-\alpha} d_{\partial\Omega}^\alpha s^2,$$

we obtain

$$\lim_{t\to 0}\frac{J(\mathbf{u}+t\mathbf{v})-J(\mathbf{u})}{t}=\int_\Omega[2\boldsymbol{\varepsilon}(\mathbf{u}):\boldsymbol{\varepsilon}(\mathbf{v})+2\kappa^\alpha\rho l^{2-\alpha}d^\alpha_{\partial\Omega}|\boldsymbol{\varepsilon}(\mathbf{u})|\boldsymbol{\varepsilon}(\mathbf{u}):\boldsymbol{\varepsilon}(\mathbf{v})]dx$$
$$-\int_\Omega\mathbf{F}\cdot\mathbf{v}\,dx.$$

In the following we are going to prove that the functional J is continuous, strictly convex and coercive. Existence and uniqueness of a velocity field of the problem (9) will then follow from results in [9].

Lemma 5 *Let $f, g \in L^r(\Omega, d^\alpha_{\partial\Omega})$ with $1 \le r < \infty$. Then*

$$\int_\Omega |d^\alpha_{\partial\Omega}(|f|^r-|g|^r)|dx \le r\||f|+|g|\|^{r-1}_{L^r(\Omega,d^\alpha_{\partial\Omega})}\|f-g\|_{L^r(\Omega,d^\alpha_{\partial\Omega})}.$$

Proof The proof is similar from Lemma 4 in [8]. Generalization with weighted space is done using Holder inequality for weighted Lebesgue space: if p, q are such that $\frac{1}{p}+\frac{1}{q}=1$ and if $h \in L^p(\Omega, d^\alpha_{\partial\Omega})$, $l \in L^q(\Omega, d^\alpha_{\partial\Omega})$, then we have

$$\int_\Omega d^\alpha_{\partial\Omega}hl\,dx \le \|h\|_{L^p(\Omega,d^\alpha_{\partial\Omega})}\|l\|_{L^q(\Omega,d^\alpha_{\partial\Omega})}.$$

Lemma 6 *The functional J is continuous for the norm $|\cdot|_{X_{\alpha,\mathrm{div}}}$.*

Proof Taking $\mathbf{v} \in \mathbf{X}_{\alpha,\mathrm{div}}$ in a neighbourhood of a fixed $\mathbf{u} \in \mathbf{X}_{\alpha,\mathrm{div}}$ and using Lemma 5 with respectively $r = 2, \alpha = 0$ and $r = 3, \alpha > 0$, we have the existence of a constant $C > 0$ (depending of $\mathbf{u}$) such that

$$\int_\Omega 2|A(|\boldsymbol{\varepsilon}(\mathbf{u})|)-A(|\boldsymbol{\varepsilon}(\mathbf{v})|)|dx$$
$$=\int_\Omega 2\left|\frac{\mu_L}{2}(|\boldsymbol{\varepsilon}(\mathbf{u})|^2-|\boldsymbol{\varepsilon}(\mathbf{v})|^2)+\frac{\kappa^\alpha\rho l^{2-\alpha}d^\alpha_{\partial\Omega}}{3}(|\boldsymbol{\varepsilon}(\mathbf{u})|^3-|\boldsymbol{\varepsilon}(\mathbf{v})|^3)\right|dx \le C|\mathbf{u}-\mathbf{v}|_{\mathbf{X}_\alpha}.$$

Since $\mathbf{F} \in [L^{\frac{4}{3}}(\Omega)]^n$ and by Poincaré inequality, we have the existence of a constant $C_f > 0$ such that

$$|J(\mathbf{u})-J(\mathbf{v})| = \left|\int_\Omega 2|A(|\boldsymbol{\varepsilon}(\mathbf{u})|)-A(|\boldsymbol{\varepsilon}(\mathbf{v})|)|dx-\int_\Omega\mathbf{F}\cdot(\mathbf{u}-\mathbf{v})dx\right|$$
$$\le C|\mathbf{u}-\mathbf{v}|_{X_\alpha}+\|\mathbf{F}\|_{L^{\frac{4}{3}}}C_p\|\nabla(\mathbf{u}-\mathbf{v})\|_{L^2} \le C_f|\mathbf{u}-\mathbf{v}|_{\mathbf{X}_\alpha},$$

where $C_f = (C + \|\mathbf{F}\|_{L^{\frac{4}{3}}}C_p)$. If $\mathbf{u} \in \mathbf{X}_{\alpha,\mathrm{div}}$, then we have

$$\lim_{\mathbf{v}\in\mathbf{X}_{\alpha,\mathrm{div}},\mathbf{v}\to\mathbf{u}} J(\mathbf{v}) = J(\mathbf{u}),$$

which finishes the proof.

Lemma 7 *The functional J is strictly convex on $\boldsymbol{X}_{\alpha,\mathrm{div}}$.*

Proof For $x \in \Omega$, the function $A(x,s)$ is strictly convex on $\mathbb{R}^+$ in s variable since

$$\frac{\partial^2}{\partial s^2} A(x,s) \geq \mu_L > 0 \text{ when } s > 0.$$

For $0 < \eta < 1$ and $\boldsymbol{\xi} \neq \boldsymbol{\nu} \in \mathbb{R}^{n\times n}$, we have using triangle inequality

$$|\eta\boldsymbol{\xi} + (1-\eta)\boldsymbol{\nu}| \leq \eta|\boldsymbol{\xi}| + (1-\eta)|\boldsymbol{\nu}|.$$

Since $A(x,s)$ is strictly convex and monotone in s variable,

$$A(x, |\eta\boldsymbol{\xi} + (1-\eta)\boldsymbol{\nu}|) \leq A(x, \eta|\boldsymbol{\xi}| + (1-\eta)|\boldsymbol{\nu}|) < \eta A(x,|\boldsymbol{\xi}|) + (1-\eta)A(x,|\boldsymbol{\nu}|)$$

which proves that $A(x,|\cdot|)$ is strictly convex. Let $\mathbf{u}, \mathbf{v} \in \mathbf{X}_\alpha$ such that $\mathbf{u} \neq \mathbf{v}$ and $0 < \eta < 1$. Thus $\boldsymbol{\varepsilon}(\mathbf{u}) \neq \boldsymbol{\varepsilon}(\mathbf{v})$ since $\boldsymbol{\varepsilon}(\mathbf{u}) - \boldsymbol{\varepsilon}(\mathbf{v}) = \boldsymbol{\varepsilon}(\mathbf{u}-\mathbf{v}) \neq 0$, see [15]. Moreover,

$$\int_\Omega A(x, \eta|\boldsymbol{\varepsilon}(\mathbf{u})| + (1-\eta)|\boldsymbol{\varepsilon}(\mathbf{v})|)dx < \eta \int_\Omega A(x,|\boldsymbol{\varepsilon}(\mathbf{u})|)dx + (1-\eta)\int_\Omega A(x,|\boldsymbol{\varepsilon}(\mathbf{v})|)dx.$$

It follows that J is strictly convex on $\mathbf{X}_{\alpha,\mathrm{div}}$.

Lemma 8 *For $0 \leq \alpha < 2$, the functional J is coercive on $\boldsymbol{X}_{\alpha,\mathrm{div}}$ in the following sense:*

$$\lim_{\boldsymbol{u}\in \boldsymbol{X}_{\alpha,\mathrm{div}};\, |\boldsymbol{u}|_{X_{\alpha,\mathrm{div}}}\to\infty} \frac{J(\boldsymbol{u})}{|\boldsymbol{u}|_{X_{\alpha,\mathrm{div}}}} \to \infty.$$

Proof We have by definition of the function A and by Remark 1 the existence of $C_1, C_2 > 0$ such that

$$\begin{aligned}\int_\Omega A(x,|\boldsymbol{\varepsilon}(\mathbf{u})|)dx &= \int_\Omega \left(\frac{\mu_L}{2}|\boldsymbol{\varepsilon}(\mathbf{u})|^2 + \frac{\rho\kappa^\alpha l^{2-\alpha} d^\alpha_{\partial\Omega}(x)}{3}|\boldsymbol{\varepsilon}(\mathbf{u})|^3\right)dx \\ &\geq C_1\|\nabla\mathbf{u}\|^2_{L^2(\Omega)} + C_2\|\nabla\mathbf{u}\|^3_{L^3(\Omega,d^\alpha_{\partial\Omega})}.\end{aligned}$$

Since $\mathbf{F} \in [L^{\frac{4}{3}}(\Omega)]^n$, there exits $C_3 > 0$ such that

$$\int_\Omega |\mathbf{F}\cdot\mathbf{u}|dx \leq \|\mathbf{F}\|_{L^{\frac{4}{3}}}\|\mathbf{u}\|_{L^4} \leq C_3\|\mathbf{F}\|_{L^{\frac{4}{3}}}\|\mathbf{u}\|_{H^1} \leq C_3\|\mathbf{F}\|_{L^{\frac{4}{3}}}\|\nabla\mathbf{u}\|_{L^2}.$$

Consequently, we have

$$J(\mathbf{u}) := \int_\Omega 2A(|\boldsymbol{\varepsilon}(\mathbf{u})|)dx - \int_\Omega \mathbf{F}\cdot\mathbf{u}\,dx \geq \tilde{C}_1\|\nabla\mathbf{u}\|^2_{L^2(\Omega)} + \tilde{C}_2\|\nabla\mathbf{u}\|^3_{L^3(\Omega,d^\alpha_{\partial\Omega})} - D\|\nabla\mathbf{u}\|_{L^2},$$

where $\tilde{C}_1, \tilde{C}_2, D$ are constants independent of $\mathbf{u}$. Finally we easily obtain

$$\lim_{\mathbf{u}\in\mathbf{X}_{\alpha,\mathrm{div}};\,|\mathbf{u}|_{\mathbf{X}_{\alpha,\mathrm{div}}}\to\infty}\frac{J(\mathbf{u})}{|\mathbf{u}|_{\mathbf{X}_{\alpha,\mathrm{div}}}}\to\infty.$$

Proposition 2 *There exists a unique* $\boldsymbol{u}\in\boldsymbol{X}_{\alpha,\mathrm{div}}$ *such that*

$$J(\boldsymbol{u})=\inf\{J(\boldsymbol{v}):\boldsymbol{v}\in\boldsymbol{X}_{\alpha,\mathrm{div}}\}.$$

Moroever, $\boldsymbol{u}$ *is the unique solution of the problem* (9).

Proof Corollary III.8 in [3] shows that the functional J is weakly lowest semi-continuous. The proof then follows from [9] using the reflexivity of $\mathbf{X}_{\alpha,\mathrm{div}}$ and Lemmas 6, 7, and 8. In particular, uniqueness comes from the strict convexity of J.

2.3 On the Pressure of Stokes Problem

In the previous section we focus on the existence of a divergence free velocity field $\mathbf{u}$. As the problem (1) involves the pressure, we study now existence of a solution of the mixed problem: find $(\mathbf{u},p)\in\mathbf{X}_\alpha\times Y_\alpha$ such that

$$\begin{cases}\displaystyle\int_\Omega 2(\mu_L+\kappa^\alpha\rho l^{2-\alpha}d^\alpha_{\partial\Omega}|\boldsymbol{\varepsilon}(\mathbf{u})|)\boldsymbol{\varepsilon}(\mathbf{u}):\boldsymbol{\varepsilon}(\mathbf{v}))dx-\int_\Omega p\,\mathrm{div}(\mathbf{v})=\int_\Omega(\mathbf{F}\cdot\mathbf{v})dx,\\ \hfill\forall\mathbf{v}\in\mathbf{X}_\alpha,\\ \displaystyle\hfill\int_\Omega q\,\mathrm{div}(\mathbf{u})=0\quad\forall q\in Y_\alpha,\end{cases}\tag{10}$$

with $0\le\alpha<2$ and where Y_α is a space that should be defined. In particular, we investigate the existence of a pressure field $p\in Y_\alpha$ with Y_α an adequate functional space related to the velocity space $\mathbf{X}_\alpha$ that gives a sense of

$$\int_\Omega p\,\mathrm{div}\,\mathbf{v}\,dx,\quad\forall\mathbf{v}\in\mathbf{X}_\alpha.$$

We start with some useful results:

Proposition 3 *The dual of the space* $L^p(\Omega,d^\alpha_{\partial\Omega})$ *can be identified with* $L^q(\Omega,d^{-\alpha q/p}_{\partial\Omega})$*, for* $1<p,q<\infty$ *satisfying* $\frac{1}{p}+\frac{1}{q}=1$.

Proof Take a function $g\in L^p(\Omega,d^\alpha_{\partial\Omega})$ and define $\tilde{g}(x)=g(x)d^{\frac{\alpha}{p}}_{\partial\Omega}(x)$. Then $\tilde{g}$ is in $L^p(\Omega)$. We consider

$$B:L^p(\Omega,d^\alpha_{\partial\Omega})\to L^p(\Omega)\text{ given by: }B(g)=\tilde{g}.$$

The operator B is linear and invertible, with $B^{-1}(\tilde{g})=\tilde{g}d^{-\frac{\alpha}{p}}$. Suppose that K is in $L^p(\Omega,d^\alpha_{\partial\Omega})'$ (the dual space of $L^p(\Omega,d^\alpha_{\partial\Omega})$). We consider $\tilde{K}:L^p(\Omega)\to\mathbb{R}$ given by

$\tilde{K}(\tilde{g}) = K(g)$ for all $g \in L^p(\Omega, d^\alpha_{\partial\Omega})$. We easily see that $\tilde{K}$ is a linear and continuous functional and thus there exists a unique $\tilde{u} \in L^q(\Omega)$ with $\frac{1}{p} + \frac{1}{q} = 1$ such that

$$\int_\Omega \tilde{u} v \, dx = \tilde{K}(v), \quad \forall v \in L^p(\Omega).$$

If we define $u = \tilde{u} d_{\partial\Omega}^{\frac{\alpha}{p}}$, it means that $u \in L^q(\Omega, d_{\partial\Omega}^{-\alpha q/p})$ and we have

$$K(g) = \tilde{K}(\tilde{g}) = \int_\Omega \tilde{u}\tilde{g} \, dx = \int_\Omega ug \, dx, \quad \forall g \in L^p(\Omega, d^\alpha_{\partial\Omega}).$$

We have shown that for each K in $L^p(\Omega, d^\alpha_{\partial\Omega})'$, there exists $u \in L^q(\Omega, d_{\partial\Omega}^{-\alpha q/p})$ unique such that

$$\int_\Omega ug \, dx = K(g), \quad \forall g \in L^p(\Omega, d^\alpha_{\partial\Omega}).$$

Consequently $\mathscr{K} : L^p(\Omega, d^\alpha_{\partial\Omega})' \to L^q(\Omega, d_{\partial\Omega}^{-\frac{\alpha q}{p}})$ given by $\mathscr{K}(K) = u$ is an isomorphism and $L^p(\Omega, d^\alpha_{\partial\Omega})'$ can be identified with $L^q(\Omega, d_{\partial\Omega}^{-\alpha q/p})$ within the "L^2 scalar product".

Definition 1 For all $\alpha \geq 0$ and $1 < p, q < \infty$, we denote

$$L^q_0(\Omega, d_{\partial\Omega}^{-\alpha q/p}) := \left\{ q \in L^q(\Omega, d_{\partial\Omega}^{-\alpha q/p}) \mid \int_\Omega q = 0 \right\}.$$

Lemma 9 *The spaces* $[W^{1,p}_0(\Omega, d^\alpha_{\partial\Omega})]^n$ *and* $L^p_0(\Omega, d^\alpha_{\partial\Omega})' := L^q_0(\Omega, d_{\partial\Omega}^{-\alpha q/p})$ *satisfy the inf-sup condition: there exists* $\tilde{C} > 0$ *such that*

$$\inf_{q \in L^q_0(\Omega, d_{\partial\Omega}^{-\alpha q/p})} \sup_{\boldsymbol{v} \in [W^{1,p}_0(\Omega, d^\alpha_{\partial\Omega})]^n} \frac{\int_\Omega q \operatorname{div}(\boldsymbol{v}) dx}{\|q\|_{L^q_{d^{-\alpha q/p}}} \|\boldsymbol{v}\|_{W^{1,p}_{d^\alpha}}} > \tilde{C}.$$

Proof From Theorem 3.1 in [11], given $f \in L^p_0(\Omega, d^\alpha_{\partial\Omega})$, there exists a vector field $\mathbf{v} : \Omega \to \mathbb{R}$ such that

$$\begin{cases} \mathbf{v} \in [W^{1,p}_0(\Omega, d^\alpha_{\partial\Omega})]^n, \\ \operatorname{div} \mathbf{v} = f, \\ \|\nabla \mathbf{v}\|_{W^{1,p}_0(\Omega, d^\alpha_{\partial\Omega})} \leq c \|f\|_{L^p_0(\Omega, d^\alpha_{\partial\Omega})}. \end{cases}$$

In other words, this show that the operator div : $[W^{1,p}_0(\Omega, d^\alpha_{\partial\Omega})]^n \to L^p_0(\Omega, d^\alpha_{\partial\Omega})$ is surjective. Lemma A.42 in [12] and Proposition 3 conclude the proof.

We consider now the unique velocity field

$$\mathbf{u} \in \mathbf{X}_{\alpha,\mathrm{div}} := \{\mathbf{v} \in [H_0^1(\Omega) \cap W_0^{1,3}(\Omega, d_{\partial\Omega}^{\alpha})]^n \mid \mathrm{div}\, \mathbf{u} = 0\}$$

that solves

$$\int_{\Omega} (2\mu_L + \kappa^{\alpha} \rho l^{2-\alpha} d_{\partial\Omega}^{\alpha} |\boldsymbol{\varepsilon}(\mathbf{u})|)\boldsymbol{\varepsilon}(\mathbf{u}) : \boldsymbol{\varepsilon}(\mathbf{v}))dx = \int_{\Omega} (\mathbf{F} \cdot \mathbf{v})dx, \quad \forall \mathbf{v} \in \mathbf{X}_{\alpha,\mathrm{div}}$$

(see Proposition 2). Since the inf-sup conditions is satisfied for the couple of spaces $[H_0^1(\Omega)]^n$, $L_0^2(\Omega)$, there exists a unique function $p_1 \in L_0^2(\Omega)$ such that

$$\int_{\Omega} p_1 \,\mathrm{div}(\mathbf{v}) = \int_{\Omega} 2\mu_L \boldsymbol{\varepsilon}(\mathbf{u}) : \boldsymbol{\varepsilon}(\mathbf{v}))dx - \int_{\Omega} (\mathbf{F} \cdot \mathbf{v})dx, \quad \forall \mathbf{v} \in [H_0^1(\Omega)]^n.$$

On the other hand, using Lemma 9 with $p = 3$ and $q = \frac{3}{2}$, we can also obtain a unique function $p_2 \in L_0^{\frac{3}{2}}(\Omega, d_{\partial\Omega}^{-\frac{\alpha}{2}})$ such that

$$\int_{\Omega} p_2 \,\mathrm{div}(\mathbf{v}) = \int_{\Omega} 2\kappa^{\alpha} \rho l^{2-\alpha} d_{\partial\Omega}^{\alpha} |\boldsymbol{\varepsilon}(\mathbf{u})|)\boldsymbol{\varepsilon}(\mathbf{u}) : \boldsymbol{\varepsilon}(\mathbf{v}))dx, \quad \forall \mathbf{v} \in [W_0^{1,3}(\Omega, d_{\partial\Omega}^{\alpha})]^n.$$

Given now $\mathbf{X}_{\alpha} = [H_0^1(\Omega) \cap W_0^{1,3}(\Omega, d_{\partial\Omega}^{\alpha})]^n$ and $Y_{\alpha} := L_0^2(\Omega) \oplus L_0^{\frac{3}{2}}(\Omega, d_{\partial\Omega}^{-\frac{\alpha}{2}})$, we immediately deduce the following result for the problem (10):

Theorem 1 *There exists* $(\boldsymbol{u}, p = p_1 + p_2) \in \boldsymbol{X}_{\alpha} \times Y_{\alpha}$ *such that the relations* (10) *are satisfied.*

Remark 2 In Theorem 1, the pressure $p \in Y_{\alpha}$ is not necessary unique. In fact the second equation in (10) can be written as

$$\int_{\Omega} q \,\mathrm{div}(\mathbf{u}) = 0 \qquad \forall q \in L_0^2(\Omega),$$
$$\int_{\Omega} q \,\mathrm{div}(\mathbf{u}) = 0 \qquad \forall q \in L_0^{\frac{3}{2}}(\Omega, d_{\partial\Omega}^{-\frac{\alpha}{2}}).$$

These two relations imply $\mathrm{div}(\mathbf{u}) = 0$ a.e. in Ω and are redundant. Thus as we are looking for p under the form $p_1 + p_2 \in Y_{\alpha}$, the decomposition could not be unique since in general $L_0^2(\Omega)$ is not included in $L_0^{\frac{3}{2}}(\Omega, d_{\partial\Omega}^{-\frac{\alpha}{2}})$.

Nevertheless, uniqueness of the pressure is sometimes available. We start with a remark:

Remark 3 Consider $\Lambda = [0, 1]$ and the weight $\mathrm{dist}(x, \{0\}) = x$. Thus if we take $g \in L^3(\Lambda, d_{\{0\}}^{\alpha})$ then we have:

$$\int_0^1 |g|^2 dx = \int_0^1 (|g|^2 x^{\frac{2\alpha}{3}}) x^{-\frac{2\alpha}{3}} dx \le \left(\int_0^1 |g|^3 x^{\alpha} dx\right)^{\frac{2}{3}} \left(\int_0^1 x^{-2\alpha} dx\right)^{\frac{1}{3}}.$$

The second integral $\left(\int_0^1 x^{-2\alpha} dx\right)^{\frac{1}{3}}$ is bounded if $0 \leq \alpha < \frac{1}{2}$. When $\alpha \geq \frac{1}{2}$, this integral diverge. We then have $\|g\|_{L^2(\Lambda)} \leq C\|g\|_{L^3_{d^{\alpha}_{\{0\}}}}$ if $0 \leq \alpha < \frac{1}{2}$.

Remark 3 shows that if $0 \leq \alpha \leq \frac{1}{2}$, we have then $L^3(\Lambda, d^{\alpha}_{\{0\}}) \subset L^2(\Lambda)$. More generally, and using the proposition 6.5 in [16], we can show that there exists a number $\alpha_0 \leq \frac{1}{2}$ such that

$$W^{1,p}(\Omega, d^{\alpha}_{\partial\Omega}) \subset W^{1,p}(\Omega)$$

with continuous injection for $0 \leq \alpha < \alpha_0$. It means in particular that $L^2(\Omega) \subset L^{\frac{3}{2}}(\Omega, d^{-\frac{\alpha}{2}}_{\partial\Omega})$ and thus $Y_\alpha := L^2_0(\Omega) \oplus L^{\frac{3}{2}}_0(\Omega, d^{-\frac{\alpha}{2}}_{\partial\Omega})$ becomes $Y_\alpha := L^{\frac{3}{2}}_0(\Omega, d^{-\frac{\alpha}{2}}_{\partial\Omega})$. Consequently, we can obtain the following result when $\mathbf{F}$ belongs to the dual space of $[W^{1,3}_0(\Omega, d^{\alpha}_{\partial\Omega})]^n$:

Theorem 2 *There exists $0 < \alpha_0 \leq \frac{1}{2}$ such that for all $0 \leq \alpha < \alpha_0$, the problem* (10) *possesses a unique solution $(\boldsymbol{u}, p) \in \boldsymbol{X}_\alpha \times Y_\alpha$, with $\boldsymbol{X}_\alpha = [W^{1,3}_0(\Omega, d^{\alpha}_{\partial\Omega})]^n$ and $Y_\alpha := L^{\frac{3}{2}}_0(\Omega, d^{-\frac{\alpha}{2}}_{\partial\Omega})$.*

3 Some Comments on the von Karman Model

The turbulent viscosity of the popular von Karman model ($\alpha = 2$) for a fluid flow between to close plates [18] is given by:

$$\mu = \mu_L + \kappa^2 d^2_{\partial\Omega} |\boldsymbol{\varepsilon}(\mathbf{u})|.$$

In fact, the weight $d^{\alpha}_{\partial\Omega}$ does not belong to the Muckenhoupt class A_3 when $\alpha = 2$ [1, 5]. This has two major consequences:

1. The space $W^{1,3}(\Omega, d^2_{\partial\Omega})$ has in fact no trace on the boundary (an example is given in one dimension by $g(x) = \ln(|\ln(x)|)$, $x \in (0, \frac{1}{2})$, with $\operatorname{dist}(x, 0) = d_{\partial\Omega}$). Recall that in [16] a trace operator $\operatorname{Tr} : W^{1,p}(\Omega, d^{\alpha}_{\partial\Omega}) \to L^p(\partial\Omega)$ is defined if $1 < p < \infty$ and $0 \leq \alpha < p - 1$. In that case, the space $W^{1,p}_0(\Omega, d^{\alpha}_{\partial\Omega})$ (the closure of $C^{\infty}_0(\Omega)$ for the norm (3)) can be identified with the space of functions in $W^{1,p}(\Omega, d^{\alpha}_{\partial\Omega})$ whose $\operatorname{Tr}(u)$ is vanishing on the boundary. For $\alpha \geq p - 1$, the trace operator cannot be defined and for $\alpha > p - 1$ the closure of $C^{\infty}_0(\Omega)$ in $W^{1,p}(\Omega, d^{\alpha}_{\partial\Omega})$ is the space itself.
   Nevertheless, the space $W^{1,3}(\Omega, d^2_{\partial\Omega})$ does not correspond to any of these cases and its characterization is more complicated (see [16, Sect. 8]).
2. The second Korn inequality in [14] is valid only for $\mathbf{u} \in [W^{1,3}(\Omega, d^{\alpha}_{\partial\Omega})]^n$ with $-1 \leq \alpha < 2$. This is an open question when $\alpha = 2$ and thus we cannot prove the first Korn inequality. Counterexample is expected in that case.

The direct consequence of these remarks is that when $\mu_L = 0$, the von Karman model is ill-posed. In this case the boundary condition $u = 0$ on $\partial\Omega$ has no meaning.

A main consequence is when a numerical method is used for obtaining an approximation of von Karman model with μ_L small (with respect to the numerical viscosity), the obtained results depend strongly on the mesh of the method as shown in the following section.

4 Numerical Experiments

In this section, we provide some numerical experiments of the problem (1) with viscosity (2) using different values of α and μ_L. The following benchmark example in three dimensional case is considered: let $\Omega \subset \mathbb{R}^3$ be the rectangular parallelepiped with characteristic length $l = 0.1$ given by

$$\Omega = [0; 1] \times [0; 1] \times [0; 0.1].$$

For $N \in \mathbb{N}$ we discretize Ω by splitting each side of that rectangular parallelepiped with N nodes. It gives N^3 hexahedron, all of which are subdivided into five tetrahedron. We obtain then a triangulation $\mathscr{T}_h$ of Ω composed of $5N^3$ tetrahedron K with $h = \frac{1}{N}$ being the reference mesh size.

Let $\mathbb{P}^1(K)$ be the space of polynomial of degree one on K. We define the following finite dimensional spaces:

$$\boldsymbol{\chi}_h = \{\mathbf{v} \in C^0(\Omega_R)^3 \mid \mathbf{v}_{|K} \in (\mathbb{P}^1(K) \oplus B_K)^3 \text{ and } \mathbf{v}_{|\partial\Omega_R} = \mathbf{0}\},$$
$$\Upsilon_h = \left\{ q \in C^0(\Omega_R) \mid q_{|K} \in \mathbb{P}^1(K) \text{ and } \int_\Omega q\, dx = 0 \right\}.$$

Here B_K denote the Bubble function on K. A renormalized version of the Problem (1)–(2) is discretized with this Galerkin approximation to obtain approximate solutions $(\mathbf{u}_h, p_h) \in \boldsymbol{\chi}_h \times \Upsilon_h$. In particular we set $\nu = \frac{\mu}{\rho}$, with renormalized p and $\mathbf{f}$ divided by ρ ($p := p/\rho, \mathbf{f} := \mathbf{f}/\rho$). In that case the turbulent kinematic viscosity is given by

$$\nu = \nu_L + \kappa^\alpha l^{2-\alpha} d^\alpha_{\partial\Omega} |\boldsymbol{\varepsilon}(\mathbf{u})|.$$

The renormalized relation (1)–(2) is a non-linear problem which is solved by a Newton method based on the work [13]. The method is iterated to reach a velocity field $\mathbf{u}_h$ with a precision of $\mathrm{Tol}_{\mathrm{New}} = 1e^{-8}$.

Each iteration of that Newton method leads to solve a linear system given by the Galerkin matrix of the Stokes problem. This system is solved with GMRES algorithm [19] with ILU(2) preconditioner and a tolerance of $\mathrm{Tol}_{\mathrm{GMRES}} = 1e^{-8}$.

Table 1 Numerical resolution of the normalized non-linear stationary Stokes problem (1)–(2) and the stationary Navier-Stokes problem for different values of α and ν_L

N	Stokes						Navier-Stokes					
	$\nu_L = 1e^{-5}$			$\nu_L = 1e^{-7}$			$\nu_L = 1e^{-5}$			$\nu_L = 1e^{-7}$		
	ν_T	$u_{\max}$	Re_T	ν_T	$u_{\max}$	Re_T	ν_T	$u_{\max}$	Re_T	ν_T	$u_{\max}$	Re_T
(b) $\alpha = 0$												
20	4.0s5e-4	4.84e-2	12	4.08e-4	4.86e-2	12	4.01e-4	4.83e-2	12	4.06e-4	4.85e-2	12
40	4.22e-4	5.21e-2	12	4.29e-4	5.28e-2	12	4.21e-4	5.20e-2	12	4.27e-4	5.26e-2	12
80	4.32e-4	5.32e-2	12	4.36e-4	5.38e-2	12	4.30e-4	5.30e-2	12	4.34e-4	5.33e-2	12
(b) $\alpha = 1$												
20	2.54e-4	1.59e-1	62	2.57e-4	1.61e-1	63	2.46e-4	1.54e-1	63	2.48e-4	1.56e-1	63
40	2.55e-4	1.84e-1	72	2.58e-4	1.88e-1	73	2.47e-4	1.83e-1	73	2.50e-4	1.86e-1	74
80	2.55e-4	1.99e-1	78	2.58e-4	2.04e-1	79	2.49e-4	1.95e-1	78	2.51e-4	2.01e-1	80
(c) $\alpha = 2$												
20	1.79e-4	2.01e-1	112	1.85e-4	2.07e-1	112	1.61e-4	1.95e-1	121	1.68e-4	2.01e-1	121
40	1.88e-4	2.68e-1	142	1.95e-4	2.81e-1	144	1.64e-4	2.45e-1	149	1.66e-4	2.55e-1	152
80	1.97e-4	3.20e-1	162	2.10e-4	3.45e-2	173	1.62e-4	2.79e-1	172	1.69e-4	2.98e-1	177

In all the following computations, we consider the following force field which generates a velocity field composed of two axial symmetric vortex:

$$\mathbf{F}(x, y, z) = \begin{pmatrix} 0.3 * (y - 0.5)^2 \\ 0.3 * (-x + 0.5)^2 \\ 0 \end{pmatrix}.$$

In Table 1, we display for different values of α and ν_L the maximum of the Euclidian norm of the velocity field $u_{\max}$, the numerical kinematic viscosity ν_L (the numerical value of $l^{2-\alpha}\kappa^\alpha d^\alpha_{\partial\Omega}|\boldsymbol{\varepsilon}(\boldsymbol{u}_{\max})|$) and the resulting Reynolds number $\mathrm{Re}_T = \frac{u_{\max} l}{\nu_T}$. The domain is a rectangular parallelepiped $\Omega = [0, 1] \times [0, 1] \times [0, 0.1]$ with N nodes on each side for a total of $5N^3$ tetrahedra. The force is given by $\boldsymbol{f} = (0.3 * (y - 0.5)^2, 0.3 * (x - 0.5)^2, 0)$ and we set $l = 0.1$.

The main observations are the following:

- For $\alpha = 0$ and different values of ν_L, the maximum value of the velocity converges as the mesh decreases and does not depend of ν_L.
- When $\alpha \in \{1, 2\}$ the convergence is more difficult to obtain, especially in the case $\alpha = 2$. We observe that when N is increasing, the maximum value of the velocity increases too. Consequently, when the laminar viscosity ν_L is small with respect to the numerical viscosity ν_T, the obtained results depend strongly on the mesh. The same behavior is observed for the stationary Navier-Stokes equations corresponding to (1)–(2).

Acknowledgements The authors would like to thank Rio-Tinto Alcan Company for their financial support and Agnieska Kalamaskya for her input on Korn's Inequalities.

References

1. Aimar H, Carena M, Durán R, Toschi M (2014) Powers of distances to lower dimensional sets as Muckenhoupt weights. Acta Math Hungar 143(1):119–137
2. Baranger J, Najib K (1990) Analyse numérique des écoulements quasi-newtoniens dont la viscosité obéit à la loi puissance ou la loi de carreau. Numer Math 58(1):35–49
3. Brezis H (1983) Analyse fonctionnelle: Théorie et applications. Masson, Paris
4. Brown RC (1998) Some embeddings of weighted Sobolev spaces on finite measure and quasibounded domains. J Inequal Appl 2(4):325–356
5. Cavalheiro AC (2008) Existence of solutions for Dirichlet problem of some degenerate quasilinear elliptic equations. Complex Var Elliptic Equ 53(2):185–194
6. Cavalheiro AC (2008) Weighted Sobolev spaces and degenerate elliptic equations. Bol Soc Parana Mat (3) 26(1–2):117–132
7. Cavalheiro AC (2013) Existence results for Dirichlet problems with degenerated p-Laplacian. Opuscula Math 33(3):439–453
8. Colinge J, Rappaz J (1999) A strongly nonlinear problem arising in glaciology. M2AN Math Model Numer Anal 33(2):395–406
9. Dacorogna B (2004) Introduction to the calculus of variations. Imperial College Press, London
10. Drábek P, Kufner A, Nicolosi F (1997) Quasilinear elliptic equations with degenerations and singularities. Walter de Gruyter, Berlin

11. Durán RG, García FL (2010) Solutions of the divergence and Korn inequalities on domains with an external cusp. Ann Acad Sci Fenn Math 35(2):421–438
12. Ern A, Guermond J-L (2004) Theory and practice of finite elements. Springer, New York
13. Jouvet G (2010) Modélisation, analyse mathématique et simulation numérique de la dynamique des glaciers. PhD thesis, Ecole Polytechnique Fédérale de Lausanne. Thèse No 4677
14. Kałamajska A (1994) Coercive inequalities on weighted Sobolev spaces. Colloq Math 66(2):309–318
15. Kondrat'ev VA, Oleinik OA (1988) Boundary value problems for a system in elasticity theory in unbounded domains. Korn inequalities. Uspekhi Mat Nauk 43(5):55–98
16. Kufner A (1983) Weighted Sobolev spaces. Elsevier, Paris
17. Opic B, Gurka P (1989) Continuous and compact imbeddings of weighted Sobolev spaces. Czechoslovak Math J 39(1):78–94
18. Pope S (2000) Turbulent flows. Cambridge University Press, Cambridge
19. Quarteroni A, Valli A (1994) Numerical approximation of partial differential equations. Springer, Berlin
20. Sandri D (1993) Sur l'approximation numérique des écoulements quasi-newtoniens dont la viscosité suit la loi puissance ou la loi de Carreau. RAIRO Modél Math Anal Numér 27(2):131–155

Poincaré Type Inequalities for Vector Functions with Zero Mean Normal Traces on the Boundary and Applications to Interpolation Methods

Sergey Repin

Abstract We consider inequalities of the Poincaré–Steklov type for subspaces of H^1-functions defined in a bounded domain $\Omega \in \mathbb{R}^d$ with Lipschitz boundary $\partial\Omega$. For scalar valued functions, the subspaces are defined by zero mean condition on $\partial\Omega$ or on a part of $\partial\Omega$ having positive $d-1$ measure. For vector valued functions, zero mean conditions are applied to normal components on plane faces of $\partial\Omega$ (or to averaged normal components on curvilinear faces). We find explicit and simply computable bounds of constants in the respective Poincaré type inequalities for domains typically used in finite element methods (triangles, quadrilaterals, tetrahedrons, prisms, pyramids, and domains composed of them). The second part of the paper discusses applications of the estimates to interpolation of scalar and vector valued functions on macrocells and on meshes with non-overlapping and overlapping cells.

Keywords Poincaré type inequalities · Interpolation of functions · Estimates of constants in functional inequalities

1 Introduction

1.1 Classical Poincaré Inequality

Poincaré [29] proved that L^2 norms of functions with zero mean defined in a bounded domain Ω with smooth boundary $\partial\Omega$ are uniformly bounded by the L^2 norm of the

Dedicated to Professor Yuri Kuznetsov on the occasion of his 70th birthday.

S. Repin
University of Jyväskylä, P.O. Box 35, 40014 Jyväskylä, Finland

S. Repin (✉)
St. Petersburg Department of V.A. Steklov Institute of Mathematics
of Russian Academy of Sciences, Saint Petersburg, Russia
e-mail: sergey.s.repin@jyu.fi; repin@pdmi.ras.ru

B. N. Chetverushkin et al. (eds.), *Contributions to Partial Differential Equations and Applications*, Computational Methods in Applied Sciences 47,
https://doi.org/10.1007/978-3-319-78325-3_22

gradient, i.e.,

$$\|w\|_{2,\Omega} \leq C_{\mathrm{P}}(\Omega)\|\nabla w\|_{2,\Omega}, \qquad \forall w \in \widetilde{H}^1(\Omega), \tag{1}$$

where

$$\widetilde{H}^1(\Omega) := \left\{ w \in H^1(\Omega) \;\mid\; \{\!|0|\!\}\, w_\Omega := \frac{1}{|\Omega|} \int\limits_\Omega w\, dx = 0 \right\}.$$

Poincaré also deduced the very first estimates of C_{P}:

$$C_{\mathrm{P}}(\Omega) \leq \frac{3}{4}\mathrm{d}_\Omega, \quad \mathrm{d}_\Omega := \mathrm{diam}\,\Omega \qquad\qquad \text{for } d = 3, \tag{2}$$

$$C_{\mathrm{P}}(\Omega) \leq \sqrt{\frac{7}{24}}\mathrm{d}_\Omega \approx 0.5401\mathrm{d}_\Omega \qquad\qquad \text{for } d = 2. \tag{3}$$

For piecewise smooth domains the inequality (1) (and a similar inequality for functions vanishing on the boundary) was independently established by Steklov [34], who proved that $C_{\mathrm{P}}(\Omega) = \lambda^{-\frac{1}{2}}$, where λ is the smallest positive eigenvalue of the problem

$$\begin{aligned} -\Delta u &= \lambda u \quad && \text{in } \Omega, \\ \partial_{\mathbf{n}} u &= 0 && \text{on } \partial\Omega. \end{aligned}$$

Easily computable estimates of $C_{\mathrm{P}}(\Omega)$ are known for *convex domains* in $\mathbb{R}^d$. An upper bound

$$C_{\mathrm{P}}(\Omega) \;\leq\; \frac{\mathrm{d}_\Omega}{\pi} \approx 0.3183\,\mathrm{d}_\Omega \tag{4}$$

was established by Payne and Weinberger [28] (notice that for $d = 2$ the upper bound (3) is not far from (4)).

A lower bound of $C_{\mathrm{P}}(\Omega)$ was derived by Cheng [8] (for $d = 2$):

$$C_{\mathrm{P}}(\Omega) \geq \frac{\mathrm{d}_\Omega}{2 j_{0,1}} \approx 0.2079\,\mathrm{d}_\Omega. \tag{5}$$

Here $j_{0,1} \approx 2.4048$ is the smallest positive root of the Bessel function J_0.

For *isosceles triangles* an improvement of the upper bound (4) is presented in [23]

$$C_{\mathrm{P}}(\Omega) \leq \frac{d_\Omega}{j_{1,1}},$$

where $j_{1,1} \approx 3.8317$ is the smallest positive root of the Bessel function J_1. Poincaré type inequalities also hold for L^q norms if $1 \leq q < +\infty$. Acosta and Durán [1] have shown that for convex domains the constant in L^1 Poincaré type inequality satisfies the estimate

$$\inf_{c\in\mathbb{R}} \|w-c\|_{L^1} \leq \frac{\mathrm{d}_\Omega}{2}\|\nabla w\|_{L^1}. \tag{6}$$

Estimates of the constant for other q can be found in Chua and Wheeden [9, 10] (also for convex domains).

1.2 Boundary Poincaré Inequalities for Functions with Zero Mean Boundary Traces

Inequalities similar to (1) also hold for functions with zero mean traces on the boundary (or on a measurable part $\Gamma \subset \partial\Omega$) such that $|\Gamma| := \mathrm{meas}_{(d-1)}\,\Gamma > 0$. For any

$$w \in \widetilde{H}^1_\Gamma(\Omega) = \left\{ w \in H^1(\Omega) \,\middle|\, \{\!|0|\!\}\, w_\Gamma := \frac{1}{|\Gamma|}\int\limits_\Gamma w\,ds = 0 \right\},$$

we have two estimates for the $L^2(\Omega)$ norm of w

$$\|w\|_{2,\Omega} \leq C_\Gamma(\Omega)\|\nabla w\|_{2,\Omega} \tag{7}$$

and for its trace on Γ

$$\|w\|_{2,\Gamma} \leq C^{\mathrm{Tr}}_\Gamma(\Omega)\|\nabla w\|_{2,\Omega}. \tag{8}$$

Existence of positive constants $C_\Gamma(\Omega)$ and $C^{\mathrm{Tr}}_\Gamma(\Omega)$ is proved by standard compactness arguments. Inequality (7) arises in analysis of certain physical phenomena (the so-called "sloshing" frequencies, see [11, 15, 16] and references therein). In the paper by Babuška and Aziz [4] it was used in proving sufficiency of the maximal angle condition for finite element meshes with triangular elements. Inequalities (7) and (8) can be useful in many other cases, e.g., for nonconforming approximations, a posteriori error estimates (see [24, 25, 30, 31]), and advanced interpolation methods for scalar and vector valued functions. In this paper, we are mainly interested in the inequality (7) for functions with zero mean on Γ. For the sake of brevity, we will call it the *boundary Poincaré inequality*.

Exact constants C_Γ and C^{Tr}_Γ are known only for a restricted number of "simple" domains. Table 1 summarizes some of the results presented in [27], which are related to such domains as

Rectangle	$\Pi_{h_1\times h_2} := (0,h_1)\times(0,h_2)$,
Parallelepiped	$\Pi_{h_1\times h_2\times h_3} := (0,h_1)\times(0,h_2)\times(0,h_3)$,
Right triangle	$\overline{T}_h := \mathrm{conv}\{(0,0),(h,0),(0,h)\}$.

These results were used in [26], where sharp constants in Poincaré type inequalities were found for simplicial domains using the affine mappings technique.

Table 1 Sharp constants

d	Ω	Γ	$C_\Gamma(\Omega)$
2	$\Pi_{h_1\times h_2}$	Face $x_1 = 0$	$c_1 \max\{2h_1; h_2\}$, $c_1 = 1/\pi$
2	$\Pi_{h_1\times h_2}$	$\partial\Omega$	$c_1 \max\{h_1; h_2\}$
3	$\Pi_{h_1\times h_2\times h_3}$	Face $x_1 = 0$	$c_1 \max\{2h_1; h_2; h_3\}$
2	T_h	Leg	$c_2 h$, $c_2 = 1/\zeta$, $\zeta \approx 2.02876$
2	T_h	Two legs	$c_1 h$
2	T_h	Hypothenuse	$\sqrt{2}c_2 h$

In Sect. 2 we deduce easily computable majorants of C_Γ for *triangles, rectangles, tetrahedrons, polyhedrons, pyramides and prismatic type domains*. These results yield interpolation estimates (and respective constants) for interpolation of scalar valued functions on macrocells based on mean values on faces. As a result, we can deduce interpolation estimates for functions defined on meshes with very complicated (e.g., non-convex) cells.

Section 3 is concerned with boundary Poincaré inequalities for vector valued functions. Certainly, (7) admits a straightforward extension to vector fields. We consider more sophisticated forms where zero mean conditions are imposed on mean values of different components of a vector valued function $\mathbf{v}$ on different $d-1$-dimensional manifolds (which are assumed to be sufficiently regular). In particular, it suffices to impose zero mean conditions on normal components of $\mathbf{v}$ on d Lipschitz manifolds (e.g., on d faces lying on $\partial\Omega$). Then,

$$\|\mathbf{v}\|_\Omega \leq \mathbb{C}(\Omega, \Gamma_1, \ldots, \Gamma_d)\|\nabla\mathbf{v}\|_\Omega. \tag{9}$$

Theorem 1 proves (9) by compactness arguments. After that, we consider the case where the conditions are imposed on normal components of a vector field on d different faces of polygonal domains in $\mathbb{R}^d$ and deduce (9) directly by applying (7) to normal components of the vector field. This method also yields easily computable majorants of the constant $\mathbb{C}$.

The last part of the paper is devoted to interpolation of functions defined in a bounded Lipschitz domain $\Omega \in \mathbb{R}^d$, which are based on mean values of the function (or of mean values of normal components) on some set $\Gamma \in \mathbb{R}^{d-1}$. It should be noted that interpolation methods based on normal components of vector fields defined on edges of finite elements are widely used in numerical analysis of PDEs (see, e.g., [6, 33]). Raviart–Thomas (RT) type interpolation operators and their properties for approximations on polyhedral meshes has been deeply studied in the papers [2, 3, 5] and other publications. The respective interpolations belong to the space $H(\Omega, \mathrm{div})$. Approximations of this type are often used in mixed and hybrid finite element methods (see, e.g., [6, 12, 33]).

This paper is concerned with coarser interpolation methods, which provide L^2 approximation of fluxes (and H^{-1} approximation for the divergence what is sufficient for treating the balance equation in a weak sense!). Interpolation methods of such a type could be useful for numerical analysis of PDEs on highly irregular (distorted) meshes. For many years, this challenging problem has been studying by Yu. Kuznetsov and coauthors (see [17–22] and other publications cited therein). Smooth (high order) methods are probably too difficult for the interpolation of vector valued functions on distorted meshes. Moreover, in many cases smooth interpolations seem to be not really natural because exact solutions often have a very restricted regularity and because efficient numerical procedures (offered, e.g., by dual mixed and hybrid methods) operate with low order approximations for fluxes. If meshes are very irregular, then it is convenient to apply approximations of the lowest possible order and respective numerical methods with minimal regularity requirements. Boundary Poincaré inequalities for functions with zero mean conditions on manifolds of the dimension $d-1$ yield interpolations of exactly this type.

In Sect. 4, it is proved that the difference between u and its interpolation is controlled by the norm of ∇u. The respective interpolation constant is computable and depends on the maximal diameter of the cell (due to results of previous sections, realistic estimates of interpolation constants are known for “typical"cells). Finally, we shortly discuss interpolation on meshes. In this case, a (global) domain Ω is decomposed into a collection of local subdomains (cells) Ω_i. Using cell interpolation operators, we define the global interpolation operator $\mathbb{I}_{\mathcal{T}_h}$ and prove the respective interpolation estimates for scalar and vector valued functions with explicitly computable constants. The interpolation method operates with minimal amount of interpolation parameters related to mean values on a certain amount of faces and preserves mean values on faces (for scalar valued functions) and mean values of normal components (for vector valued functions).

2 Estimates of C_Γ for Typical Mesh Cells

2.1 Triangles

Consider a non-degenerated triangle ABC (Fig. 1 left) where Γ coincides with the side AC.

2.1.1 Majorant of C_Γ

Our analysis is based upon the estimate

$$C_\Gamma^2 \leq C_{\rm P}^2 + \frac{|\Omega|}{|\Gamma|^2} \inf_{\tau \in Q(\Omega)} \|\boldsymbol{\tau}\|_{2,\Omega}^2, \tag{10}$$

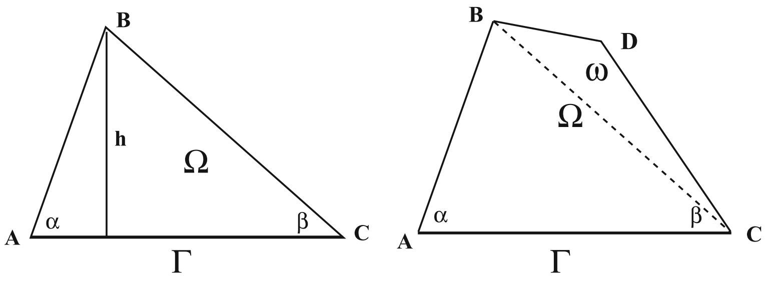


Fig. 1 Triangle and quadrilateral

which is a special form of the upper bound of C_Γ derived in [32]. Here $Q(\Omega)$ is a subset of $H(\Omega, \mathrm{div})$ containing vector functions such that $\mathrm{div}\,\boldsymbol{\tau} = \frac{|\Gamma|}{|\Omega|}$, $\boldsymbol{\tau} \cdot \mathbf{n} = 1$ on Γ, and $\boldsymbol{\tau} \cdot \mathbf{n} = 0$ on $\partial\Omega \setminus \Gamma$. We set $\boldsymbol{\tau}$ as an affine field with values at the nodes A, B, and C $(-\cot\alpha, -1)$, $(0, 0)$, and $(\cot\beta, -1)$, respectively. In this case,

$$\|\boldsymbol{\tau}\|^2_{2,\Omega} = \frac{1}{3}|\Omega| \left(\frac{3}{2} + \frac{1}{4}\cot^2\alpha + \frac{1}{4}\cot^2\beta + \frac{1}{4}(\cot\beta - \cot\alpha)^2 \right) = \frac{|\Omega|}{6}\Sigma_{\alpha\beta},$$

where

$$\Sigma_{\alpha\beta} = \cot^2\alpha + \cot^2\beta - \cot\alpha\cot\beta + 3.$$

Since $|\Omega| = \frac{1}{2}h|\Gamma|$, we see that $\frac{|\Omega|^2}{|\Gamma|^2} = \frac{h^2}{4}$. In view of (4), the constant C_P is bounded from above by $\frac{d_\Omega}{\pi}$, where $d_\Omega = \max\{|AB|, |BC|, |CD|\}$, and we deduce an easily computable bound

$$C^2_\Gamma \le C^2_\mathrm{P} + \frac{h^2\Sigma_{\alpha\beta}}{24} \le \frac{\mathrm{d}^2_\Omega}{\pi^2} + \frac{h^2\Sigma_{\alpha\beta}}{24}. \tag{11}$$

We can represent $\Sigma_{\alpha\beta}$ in a somewhat different form

$$\Sigma_{\alpha\beta} = \frac{|AB|^2 + |BC|^2 + \overrightarrow{AB} \cdot \overrightarrow{BC}}{h^2},$$

which yields the estimate

$$C^2_\Gamma \le \frac{\mathrm{d}^2_\Omega}{\pi^2} + \frac{|AB|^2 + |BC|^2 + \overrightarrow{BA} \cdot \overrightarrow{BC}}{24}. \tag{12}$$

Example 1 If $\alpha = \frac{\pi}{2}$, then $\mathrm{d}^2_\Omega = h^2 + |\Gamma|^2$, $|\Gamma| = h\cot\beta$, $\mathrm{d}^2_\Omega = h^2(1 + \cot^2\beta)$ and we obtain

$$C_\Gamma \le h\sqrt{\frac{1}{\pi^2} + \frac{1}{8} + \cot^2\beta\left(\frac{1}{\pi^2} + \frac{1}{24}\right)} \approx 0.4757\,h\sqrt{1 + 0.6354\cot^2\beta}.$$

In particular, for $\beta = \frac{\pi}{4}$, we obtain $C_\Gamma \leq 0.6083h$ (exact constant for the right triangle is $0.4929h$).

2.1.2 Minorant of C_Γ

A lower bound for C_Γ follows from (5) and relations between $C_{\mathrm{P}}(\Omega)$ and $C_\Gamma(\Omega)$. Any function in $\widetilde{H}^1_\Gamma(\Omega)$ can be represented as $w - \{\!|0|\!\}\, w_\Gamma$, where $w \in H^1(\Omega)$. Hence,

$$(C_\Gamma(\Omega))^{-2} = \inf_{w \in H^1(\Omega)} \frac{\int_\Omega |\nabla w|^2 \, dx}{\int_\Omega |w - \{\!|0|\!\}\, w_\Gamma|^2 \, dx}$$

and the constant $C_\Gamma(\Omega)$ can be defined as maximum of $\|w - \{\!|0|\!\}\, w_\Gamma\|_{2,\Omega}$ for all $w \in H^1(\Omega)$ such that $\|\nabla w\|_{2,\Omega} = 1$. Analogously, C_{P} can be defined as maximum of $\|w - \{\!|0|\!\}\, w_\Omega\|_{2,\Omega}$ over the same set of functions. Since

$$\|w - \{\!|0|\!\}\, w_\Gamma\|_{2,\Omega} \geq \inf_{c \in \mathbb{R}} \|w - c\|_{2,\Omega} = \|w - \{\!|0|\!\}\, w_\Omega\|_{2,\Omega},$$

we conclude that for any selection of Γ

$$C_{\mathrm{P}}(\Omega) \leq C_\Gamma(\Omega). \tag{13}$$

From (5) and (13), it follows that $C_\Gamma \geq \frac{1}{2} \frac{\mathrm{d}_\Omega}{j_{0,1}}$. In particular, for $\alpha = \frac{\pi}{2}$ we have $C_\Gamma \geq 0.2079\, h\sqrt{1 + \cot^2 \beta}$.

2.2 *Quadrilaterals*

Using previous results, we deduce an estimate of C_Ω for a quadrilateral ABCD (Fig. 1 right). On Ω_1 we set the same field τ as in the previous case and set $\tau = 0$ on Ω_2. Let $\kappa^2 = \frac{|\Omega_2|}{|\Omega_1|}$. Then,

$$C^2_\Gamma \leq C^2_{\mathrm{P}} + \left(\kappa C_{\mathrm{P}} + \frac{\Sigma^{1/2}_{\alpha\beta} |\Omega|}{\sqrt{6}\, |\Gamma|} \right)^2. \tag{14}$$

Note that (14) also holds for more general cases in which Ω_2 is a bounded Lipschitz domain having only one common boundary with Ω_1, which is BC.

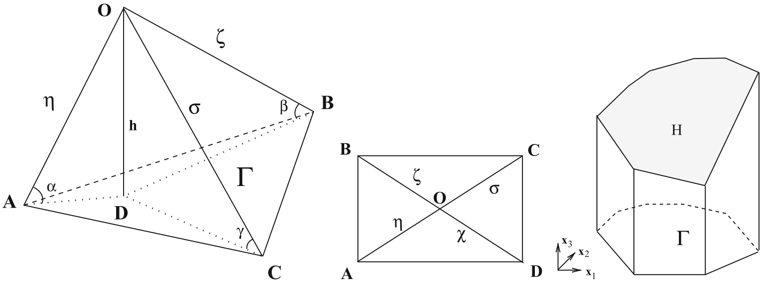


Fig. 2 Tetrahedron, pyramid, and prism

2.3 Tetrahedrons

Consider a tetrahedron OABC (Fig. 2 left), where Γ is the triangle ABC which lies in the plane Ox_1x_2. Let $\overrightarrow{OA} = \boldsymbol{\eta}$, $\overrightarrow{OB} = \boldsymbol{\zeta}$, and $\overrightarrow{OC} = \boldsymbol{\sigma}$. At vertexes A, B, and C, we define three vectors

$$\widehat{\boldsymbol{\tau}}_A = \frac{\boldsymbol{\eta}}{|\boldsymbol{\eta}|\sin\alpha}, \quad \widehat{\boldsymbol{\tau}}_A = \frac{\boldsymbol{\zeta}}{|\boldsymbol{\zeta}|\sin\beta}, \quad \text{and} \quad \widehat{\boldsymbol{\tau}}_A = \frac{\boldsymbol{\sigma}}{|\boldsymbol{\sigma}|\sin\gamma}.$$

The vector field $\boldsymbol{\tau}(x_1, x_2, x_3)$ is the affine field in Ω with zero value at the vertex O. We compute

$$\int\limits_{\Omega} |\tau|^2\, dx = \int\limits_0^h \left(\int\limits_{\omega(x_3)} |\tau(x_1, x_2, x_3)|^2 dx_1 dx_2 \right) dx_3.$$

Notice that the horizontal cross section $\omega(x_3)$ associated with the height x_3 has the measure $|\omega(x_3)| = \left(1 - \frac{x_3}{h}\right)^2 |\Gamma|$ and at the respective point A' on OA (which third coordinate is x_3) by linear proportion we have $\boldsymbol{\tau}_{A'} = \left(1 - \frac{x_3}{h}\right) \widehat{\boldsymbol{\tau}}_A$. Similar relations hold for the points B' and C' associated with the cross section on the height x_3. For the internal integral we apply the Gaussian quadrature for $|\tau|^2 = \tau_1^2 + \tau_2^2 + \tau_3^3$ and obtain

$$C_\Gamma^2 \leq \frac{\mathrm{d}_\Omega^2}{\pi^2} + \frac{|\boldsymbol{\eta}|^2 + |\boldsymbol{\zeta}|^2 + |\boldsymbol{\sigma}|^2 + \boldsymbol{\eta}\cdot\boldsymbol{\zeta} + \boldsymbol{\eta}\cdot\boldsymbol{\sigma} + \boldsymbol{\zeta}\cdot\boldsymbol{\sigma}}{90}. \tag{15}$$

In particular, for the equilateral tetrahedron with all edges equal to h we have

$$\boldsymbol{\eta}\cdot\boldsymbol{\zeta} = \boldsymbol{\eta}\cdot\boldsymbol{\sigma} = \boldsymbol{\zeta}\cdot\boldsymbol{\sigma} = \frac{1}{2}h^2, \quad \mathrm{d}_\Omega = h,$$

and, therefore, $C_\Gamma \leq h\sqrt{\frac{1}{\pi^2} + \frac{1}{20}} \approx 0.39h$.

For the right tetrahedron with nodes $(0,0,0)$, $(h,0,0)$, $(0,h,0)$, $(0,0,h)$ and face $\mathit{\Gamma} = \{x \in \overline{\Omega},\ x_3 = 0\}$, we have $\mathrm{d}_\Omega = h\sqrt{2}$, $|\eta| = h$, $|\zeta| = |\sigma| = h\sqrt{2}$, scalar products are equal to h^2 and (15) yields $C_\Gamma \le h\sqrt{\frac{2}{\pi^2} + \frac{4}{45}} \approx 0.54\,h$. Sharp constants C_Γ for triangle and tetrahedrons has been recently evaluated in [26]. For the right tetrahedron, the constant computed in [26] is $C_\Gamma \approx 0.3756\,h$.

2.4 Pyramid

We can apply (15) in order to evaluate C_Γ for a pyramid OABCD, which can be divided into two tetrahedrons OABC and OACD (Fig. 2 middle, view from above). Assume that the triangles ABC and ACD have equal areas and Γ is the pyramid basement ABCD. Then, we can use (10) with τ defined in each tetrahedron as in 2.3. We obtain

$$C_\Gamma^2 \le \frac{\mathrm{d}_\Omega^2}{\pi^2} + \frac{2|\boldsymbol{\eta}|^2 + |\boldsymbol{\zeta}|^2 + 2|\boldsymbol{\sigma}|^2 + |\boldsymbol{\chi}|^2 + 2\boldsymbol{\eta}\cdot\boldsymbol{\sigma} + (\boldsymbol{\eta}+\boldsymbol{\sigma})\cdot(\boldsymbol{\chi}+\boldsymbol{\zeta})}{180}. \tag{16}$$

2.5 Prismatic Cells

Consider a prismatic type domain (Fig. 2 right)

$$\Omega = \{x \in \mathbb{R}^3 \mid (x_1, x_2) \in \Gamma,\ 0 \le x_3 \le H(x_1, x_2),\ H(x_1, x_2) \ge H_{\min}\}.$$

By the same method as in 2.1 we find that

$$C_\Gamma^2 \le \overline{C}_\Gamma^2 := C_\mathrm{P}^2 + \left(\frac{\{\!|0|\!\}\, H_\Gamma}{\sqrt{3}} + C_\mathrm{P}\,\kappa\right)^2. \tag{17}$$

where $\kappa = \left(\frac{\{\!|0|\!\} H_\Gamma}{H_{\min}} - 1\right)^{1/2}$ characterizes variations of the mean height.

In particular, if $H = const$ (so that $\kappa = 0$) and Γ is a convex domain in $\mathbb{R}^{d-1}$, then

$$C_\Gamma^2 \le \frac{\mathrm{d}_\Gamma^2 + H^2}{\pi^2} + \frac{H^2}{3} = \frac{1}{\pi^2}\left(\mathrm{d}_\Gamma^2 + \left(1 + \frac{\pi^2}{3}\right)H^2\right). \tag{18}$$

For a parallelepiped with $\Gamma = (0, a) \times (0, b)$, we know that the exact value of C_Γ is $\frac{1}{\pi}\max\{2H, a, b\}$. In this case $\mathrm{d}_\Gamma^2 = a^2 + b^2$ and we can compare it with the upper bound that follows from (17):

$$\frac{\overline{C}_\Gamma}{C_\Gamma} = \frac{\sqrt{a^2 + b^2 + 4.29H^2}}{\max\{2H, a, b\}} \ge 1. \tag{19}$$

For the cases where one dimension of Ω dominates, $\overline{C}_\Gamma$ is a good approximation of C_Γ^2. If $a = b = H$ (cube), then we have $\frac{\overline{C}_\Gamma}{C_\Gamma} = \frac{\sqrt{6.29}}{2} \approx 1.25$. The largest ratio is for $a = b = 2H$ (≈ 1.75).

3 Boundary Poincaré Inequalities for Vector Valued Functions

Estimates (7) and (8) yield analogous estimates for vector valued functions in $H^1(\Omega, \mathbb{R}^d)$. Let $\Omega \in \mathbb{R}^d$ $(d \geq 1)$ be a connected domain with N plane faces $\Gamma_i \in \mathbb{R}^{d-1}$. Assume that we have d unit vectors $\mathbf{n}^{(k)}$, (associated with some faces) that form a linearly independent system in $\mathbb{R}^d$, i.e.,

$$\det \boldsymbol{N} \neq 0, \qquad \boldsymbol{N} := \left\{ n_j^{(i)} \right\} \in \mathbb{M}^{d \times d}, \quad i, j = 1, 2, \ldots, d, \tag{20}$$

where $n_j^{(i)} = \mathbf{n}^{(i)} \cdot \mathbf{e}_j$ and $\mathbf{e}_i$ denote the Cartesian basis. Then, $\mathbf{v} \in H^1(\Omega, \mathbb{R}^d)$ satisfies a Poincaré type estimate provided that it satisfies zero mean conditions (21).

Theorem 1 *If (20) holds and*

$$\{\!|0|\!\}\, \mathbf{v} \cdot \mathbf{n}^{(i)}{}_{\Gamma_i} = 0 \quad i = 1, 2, \ldots, d, \tag{21}$$

then

$$\|\mathbf{v}\|_\Omega \leq \mathbb{C}(\Omega, \Gamma_1, \ldots, \Gamma_d) \|\nabla \mathbf{v}\|_\Omega, \tag{22}$$

where $\mathbb{C} > 0$ depends only on geometrical properties of the cell.

Proof Assume the opposite. Then, there exists a sequence $\{\mathbf{v}_k\}$ such that $\{\!|0|\!\}\, \mathbf{v}_k \cdot \mathbf{n}^{(i)}$ $_{\Gamma_i} = 0$ and

$$\|\mathbf{v}_k\| \geq k \|\nabla \mathbf{v}_k\|. \tag{23}$$

Without a loss of generality we can operate with a sequence of normalized functions, so that

$$\|\mathbf{v}_k\| = 1. \tag{24}$$

Hence,

$$\|\nabla \mathbf{v}_k\| \leq \frac{1}{k} \to 0 \quad \text{as } k \to +\infty. \tag{25}$$

We conclude that there exists a subsequence (for simplicity we omit additional subindexes and keep the notation $\{\mathbf{v}_k\}$) such that

$$\mathbf{v}_k \rightharpoonup \mathbf{w} \quad \text{in } H^1(\Omega, \mathbb{R}^d), \tag{26}$$

$$\mathbf{v}_k \to \mathbf{w} \quad \text{in } L^2(\Omega, \mathbb{R}^d). \tag{27}$$

In view of (26),

$$0 = \liminf_{k\to+\infty} \|\nabla \mathbf{v}_k\| \geq \|\nabla \mathbf{w}\|,$$

we see that $\mathbf{w} \in P^0(\Omega, \mathbb{R}^d)$. For any face Γ_i we have (in view of the trace theorem)

$$\|\mathbf{v}_k - \mathbf{w}\|_{2,\Gamma_i} \leq C \left(\|\mathbf{v}_k - \mathbf{w}\|_{2,\Omega} + \|\nabla \mathbf{v}_k\|_{2,\Omega} \right). \tag{28}$$

We recall (25) and (27) and conclude that the traces of $\mathbf{v}_k$ on Γ_i converge to the trace of $\mathbf{w}$. Since $\mathbf{v}_k \cdot \mathbf{n}^{(i)}$ have zero means,

$$\mathbf{w} \cdot \mathbf{n}_i |\Gamma_i| = \int\limits_{\Gamma_i} \mathbf{w} \cdot \mathbf{n}_i \, d\Gamma = 0 \tag{29}$$

and $\mathbf{w}$ is orthogonal to d linearly independent vectors, i.e., $\mathbf{w} = 0$. On the other hand, $\|\mathbf{w}\| = 1$. We obtain a contradiction, which shows that the assumption is not true.

We notice that conditions of the Theorem are very flexible with respect to choosing Γ_i and vectors $\mathbf{n}^{(i)}$ entering the integral type conditions (21). Probably the most interesting case is where $\mathbf{n}^{(i)}$ are defined as unit outward normals to faces Γ_s. If $d = 2$, then we can also define $\mathbf{n}^{(i)}$ as unit tangential vectors. Moreover, in the proof it is not essential that $\mathbf{n}^{(i)}$ is strictly related to one face Γ_i (only the condition (20) is essential). For example, if $d = 3$ then we can define two vectors as two mutually orthogonal tangential vectors of one face and the third one as a normal vector to another face. Theorem holds for this case as well. Henceforth, for the sake of definiteness we assume that $\mathbf{n}^{(i)}$ are normal vectors or mean normal vectors (for curvilinear faces) associated with faces Γ_i, $i = 1, 2, \ldots, d$. Possible modifications of the results to other cases are rather obvious.

3.1 Value of the Interpolation Constant for d = 2

Estimates of the constant $\mathbb{C}(\Omega, \Gamma_1, \Gamma_2)$ follow from (7) and depend on the constants $C_{\Gamma_i}(\Omega)$. Below we deduce explicit and easily computable bounds of $\mathbb{C}(\Omega, \Gamma_1, \Gamma_2)$.

First, we consider a special, but important case where Ω is a polygonal domain in $\mathbb{R}^2$. Let Γ_1 and Γ_2 be two faces selected for the interpolation of $\mathbf{v}$. The respective normals $\mathbf{n}^{(1)} = (n_1^{(1)}, n_2^{(1)})$ and $\mathbf{n}^{(2)} = (n_1^{(2)}, n_2^{(2)})$ must satisfy the condition (20), which means that

$$\angle(\mathbf{n}^{(1)}, \mathbf{n}^{(2)}) = \beta \in (0, \pi).$$

Let the conditions (21) hold. Then

$$\|n_1^{(1)} v_1 + n_2^{(1)} v_2\|^2 \leq C_{\Gamma_1}^2(\Omega) \|n_1^{(1)} \nabla v_1 + n_2^{(1)} \nabla v_2\|^2, \tag{30}$$

$$\|n_1^{(2)} v_1 + n_2^{(2)} v_2\|^2 \leq C_{\Gamma_2}^2(\Omega) \|n_1^{(2)} \nabla v_1 + n_2^{(2)} \nabla v_2\|^2. \tag{31}$$

Introduce the matrix

$$\boldsymbol{T} := \mathbf{n}^{(1)} \otimes \mathbf{n}^{(1)} + \mathbf{n}^{(2)} \otimes \mathbf{n}^{(2)} = \begin{pmatrix} (n_1^{(1)})^2 + (n_1^{(2)})^2 & n_1^{(1)} n_2^{(1)} + n_1^{(2)} n_2^{(2)} \\ n_1^{(1)} n_2^{(1)} + n_1^{(2)} n_2^{(2)} & (n_2^{(1)})^2 + (n_2^{(2)})^2 \end{pmatrix}.$$

Here and later on $\otimes$ denotes the diadic product of vectors. Summation of (30) and (31) yields

$$\int\limits_{\Omega} \boldsymbol{T}\mathbf{v} \cdot \mathbf{v} dx_1 dx_2 \leq C^2 \int\limits_{\Omega} (T_{11} |\nabla v_1|^2 + 2T_{12} \nabla v_1 \cdot \nabla v_2 + T_{22} |\nabla v_2|^2) dx_1 dx_2, \quad (32)$$

where

$$C = \max\{C_{\Gamma_1}(\Omega); C_{\Gamma_2}(\Omega)\}.$$

It is easy to see that $\boldsymbol{T}$ is a positive definite matrix. Indeed,

$$\begin{aligned} \det(\boldsymbol{T} - \lambda \boldsymbol{E}) &= ((n_1^{(1)})^2 + (n_1^{(2)})^2 - \lambda)((n_2^{(1)})^2 + (n_2^{(2)})^2 - \lambda) - \left(n_1^{(1)} n_2^{(1)} + n_1^{(2)} n_2^{(2)}\right)^2 \\ &= \lambda^2 - 2\lambda + \left(n_1^{(1)} n_2^{(2)} - n_1^{(2)} n_2^{(1)}\right)^2 = \lambda^2 - 2\lambda + (\det \boldsymbol{N})^2, \end{aligned}$$

where

$$\boldsymbol{N} := \begin{pmatrix} \mathbf{n}^{(1)} \\ \mathbf{n}^{(2)} \end{pmatrix}.$$

Hence for any vector $\mathbf{b}$, we have $\lambda_1 |\mathbf{b}|^2 \leq \boldsymbol{T}\mathbf{b} \cdot \mathbf{b} \leq \lambda_2 |\mathbf{b}|^2$, and

$$\lambda_{1,2} = 1 \mp \sqrt{1 - (\det \boldsymbol{N})^2}.$$

If $\mathbf{n}^{(1)}$ and $\mathbf{n}^{(1)}$ are orthogonal, then $\det \boldsymbol{N} = 1$ and the unique eigenvalue of $\boldsymbol{N}$ is $\lambda = 1$. In this case, the left hand side of (32) coincides with $\|\mathbf{v}\|^2$. In all other cases $\det \boldsymbol{N} < 1$ and $\lambda_1 < \lambda_2$.

We can always select the coordinate system such that

$$n_1^{(1)} = 1, \quad n_2^{(1)} = 0, \quad n_1^{(2)} = -\cos\beta, \quad n_2^{(2)} = \sin\beta.$$

Then,

$$T_{11} = 1 + \cos^2\beta, \quad T_{22} = 1 - \cos^2\beta, \quad T_{12} = -\sin\beta\cos\beta,$$

and the matrix is

$$\boldsymbol{N} := \begin{pmatrix} 1 & 0 \\ -\cos\beta & \sin\beta \end{pmatrix}.$$

We see that $\det \boldsymbol{N} = \sin\beta$, and $\lambda_1 = 1 - |\cos\beta|$.

Consider the right-hand side of (32). It is bounded from above by the quantity

$$I(\mathbf{v}) := C^2 \int\limits_{\Omega} \left((T_{11} + \gamma |T_{12}|) |\nabla v_1|^2 + (T_{22} + \gamma^{-1} |T_{12}|) |\nabla v_2|^2 \right) dx,$$

where γ is any positive number. We define γ by means of the relation $T_{11} - T_{22} = (\gamma^{-1} - \gamma)|T_{12}|$, which yields $\gamma = \frac{1-|\cos\beta|}{\sin\beta}$. Then,

$$I(\mathbf{v}) \ \leq \ (1 + |\cos\beta|)\|\nabla \mathbf{v}\|^2. \tag{33}$$

From (32) and (33), we find that

$$\boxed{\|\mathbf{v}\| \leq \max_{i=1,2} \left\{ C_{\Gamma_i}(\Omega) \right\} \sqrt{\frac{1 + |\cos\beta|}{1 - |\cos\beta|}}\, \|\nabla \mathbf{v}\|.} \tag{34}$$

This is the Poincaré type inequality for the vector valued function $\mathbf{v}$ with zero mean normal traces on Γ_1 and Γ_2. It is worth noting that for small β (and for β close to π) the constant blows up. Therefore, interpolation operators (considered in Sect. 4) should avoid such situations.

3.2 Value of the Interpolation Constant for d ≥ 3

Now we are concerned with the general case and deduce the estimate valid for any dimension d. In view of (21) we have

$$\sum_{k=1}^{d} \|\mathbf{n}^{(k)} \cdot \mathbf{v}\|^2_{2,\Omega} \leq C^2 \sum_{k=1}^{d} \int\limits_{\Omega} \left(\sum_{i=1}^{d} n_i^{(k)} \nabla v_i \right)^2 d\mathbf{x}, \tag{35}$$

where

$$C = \max_{k=1,2,\ldots,d} \left\{ C_{\Gamma_k}(\Omega) \right\}.$$

In view of the relation

$$(\mathbf{n}^{(k)} \otimes \mathbf{n}^{(k)})\mathbf{v} \cdot \mathbf{v} = (\mathbf{n}^{(k)}(\mathbf{n}^{(k)} \cdot \mathbf{v})) \cdot \mathbf{v} = (\mathbf{n}^{(k)} \cdot \mathbf{v})^2,$$

the left-hand side of (35) is $\int\limits_{\Omega} \boldsymbol{T}\mathbf{v} \cdot \mathbf{v}$, where

$$\boldsymbol{T} := \sum_{k=1}^{d} \mathbf{n}^{(k)} \otimes \mathbf{n}^{(k)}. \tag{36}$$

If $\mathbf{n}^{(k)}$ form a linearly independent system, then $\boldsymbol{T}$ is a positive definite matrix. Indeed, $\boldsymbol{T}\mathbf{b}\cdot\mathbf{b}=\sum_{k=1}^{d}(\mathbf{n}^{(k)}\cdot\mathbf{b})^2$. Hence, $\boldsymbol{T}\mathbf{b}\cdot\mathbf{b}=0$ if and only if $\mathbf{b}$ has zero projections to d linearly independent vectors $\mathbf{n}^{(k)}$, i.e., $\boldsymbol{T}\mathbf{b}\cdot\mathbf{b}=0$ if and only if $\mathbf{b}=0$. Therefore,

$$\lambda_1\,\|\mathbf{v}\|^2\leq\int\limits_{\Omega}\boldsymbol{T}\mathbf{v}\cdot\mathbf{v}\,d\mathbf{x},\tag{37}$$

where $\lambda_1>0$ is the minimal eigenvalue of $\boldsymbol{T}$.

Consider the right hand side of (35). We have

$$\begin{aligned}\int\limits_{\Omega}\left(\sum_{i=1}^{d}n_i^{(k)}\nabla v_i\right)^2 d\mathbf{x}&=\int\limits_{\Omega}\sum_{i,j=1}^{d}n_i^{(k)}n_j^{(k)}\nabla v_i\cdot\nabla v_j d\mathbf{x}\\&=\sum_{i,j=1}^{d}n_i^{(k)}n_j^{(k)}\int\limits_{\Omega}\nabla v_i\cdot\nabla v_j d\mathbf{x}=\mathbf{n}^{(k)}\otimes\mathbf{n}^{(k)}:\boldsymbol{G},\end{aligned}$$

where

$$\boldsymbol{G}(\mathbf{v}):=\{G_{ij}\},\quad G_{ij}(\mathbf{v})=\int\limits_{\Omega}\nabla v_i\cdot\nabla v_j d\mathbf{x}.$$

Hence,

$$\sum_{k=1}^{d}\int\limits_{\Omega}\left(\sum_{i=1}^{d}n_i^{(k)}\nabla v_i\right)^2 d\mathbf{x}=\boldsymbol{T}:\boldsymbol{G}(\mathbf{v})\leq|\boldsymbol{T}|\,|\boldsymbol{G}(\mathbf{v})|.\tag{38}$$

Now (35), (36), (37), and (38) yield the estimate

$$\|\mathbf{v}\|^2\leq C^2\frac{1}{\lambda_1}|\boldsymbol{T}|\,|\boldsymbol{G}(\mathbf{v})|\leq C^2\frac{d}{\lambda_1}\,|\boldsymbol{G}(\mathbf{v})|.$$

Since $|\boldsymbol{G}(\mathbf{v})|\leq\|\nabla\mathbf{v}\|^2$, for any $\mathbf{v}\in H^1(\Omega,\mathbb{R}^d)$ satisfying (21) we have

$$\boxed{\|\mathbf{v}\|\leq C\sqrt{\frac{d}{\lambda_1}}\|\nabla\mathbf{v}\|.}\tag{39}$$

In other words, the constant in (39) can be defined as follows:

$$\mathbb{C}(\Omega,\Gamma_1,\Gamma_2,\ldots,\Gamma_d)=\max_{k=1,2,\ldots,d}\{C_{\Gamma_k}(\Omega)\}\sqrt{\frac{d}{\lambda_1}},$$

where λ_1 is the minimal eigenvalue of $\boldsymbol{T}$.

For $d = 2$ this estimate exposes a slightly worse constant than (34) with the factor $\sqrt{\frac{2}{1-|\cos\beta|}}$ instead of $\sqrt{\frac{1+|\cos\beta|}{1-|\cos\beta|}}$.

4 Interpolation of Functions

The classical Poincaré inequality (1) yields a simple interpolation operator $\mathbb{I}_{\Omega} : H^1(\Omega) \to P^0(\Omega)$ defined by the relation $\mathbb{I}_{\Omega} w := \{\!|0|\!\}\, w_{\Omega}$. In view of (1), we know that

$$\|w - \mathbb{I}_{\Omega} w\|_{2,\Omega} \leq C_{\mathrm{P}}(\Omega) \|\nabla w\|_{2,\Omega},$$

which means that the interpolation operator is stable and $C_{\mathrm{P}}(\Omega)$ is the respective constant.

Above discussed estimates for functions with zero mean traces yield somewhat different interpolation operators for scalar and vector valued functions. For a scalar valued function $w \in H^1(\Omega)$, we set $\mathbb{I}_{\Gamma}(w) := \{\!|0|\!\}\, w_{\Gamma}$, i.e., the interpolation operator uses mean values of w a $d-1$-dimensional set Γ. Since $\{\!|0|\!\}\, w - \mathbb{I}_{\Gamma} w_{\Gamma} = 0$, we use (7) and obtain the interpolation estimate

$$\|w - \mathbb{I}_{\Gamma} w\|_{2,\Omega} \;\leq\; C_{\Gamma}(\Omega) \|\nabla w\|_{2,\Omega},$$

where the constant C_{Γ} appears as the interpolation constant. Analogously, (8) yields an interpolation estimate for the boundary trace

$$\|w - \mathbb{I}_{\Gamma} w\|_{2,\Gamma} \;\leq\; C_{\Gamma}^{\mathrm{Tr}} \|\nabla w\|_{2,\Omega}.$$

Applying these estimates to cells of meshes we obtain analogous interpolation estimates for mesh interpolation of scalar functions with explicit constants depending on character diameter of cells.

For the interpolation of vector valued functions we use (39) and generalise this idea.

4.1 Cells with Plane Faces

Define the operator

$$\mathbb{I}_{\Gamma_1,\Gamma_2,\ldots,\Gamma_d} : H^1(\Omega, \mathbb{R}^d) \to P^0(\Omega, \mathbb{R}^d)$$

that performs zero order interpolation of a vector valued function $\mathbf{v}$ using mean values of normal components on the faces Γ_i, $i = 1, 2, \ldots, d$. In this case, we set

$$\int\limits_{\Gamma_i} \left(\mathbb{I}_{\Gamma_1,\Gamma_2,\ldots,\Gamma_d}\mathbf{v}\right)\cdot \mathbf{n}^{(i)}\, d\Gamma = \int\limits_{\Gamma_i} \mathbf{v}\cdot\mathbf{n}^{(i)}\, d\Gamma \quad i = 1, 2, \ldots, d. \tag{40}$$

This condition means that *the interpolation must preserve integral values of normal fluxes on d selected faces*. In general, we may define several different operators associated with different collections of faces. However, once the set of $\Gamma_1, \Gamma_2, \ldots, \Gamma_d$ satisfying (20) has been defined, the operator $\mathbb{I}_{\Gamma_1,\Gamma_2,\ldots,\Gamma_d}$ uniquely defines the vector $\mathbb{I}_{\Gamma_1,\Gamma_2,\ldots,\Gamma_d}\, \mathbf{v}$. In view of (40) and the identity

$$\left(\mathbb{I}_{\Gamma_1,\Gamma_2,\ldots,\Gamma_d}\mathbf{v}\right)\cdot\mathbf{n}^{(i)} = (\mathbb{I}_{\Gamma_1,\Gamma_2,\ldots,\Gamma_d}\mathbf{v})_j\mathbf{e}_j\cdot\mathbf{n}^{(i)},$$

we conclude that the components of the interpolated field are uniquely defined by the system

$$\sum_{j=1}^{d} n_j^{(i)}(\mathbb{I}_{\Gamma_1,\Gamma_2,\ldots,\Gamma_d}\,\mathbf{v})_j = \frac{1}{|\Gamma_i|}\int\limits_{\Gamma_i}\mathbf{v}\cdot\mathbf{n}^{(i)}\, d\Gamma \quad i = 1, 2, \ldots, d.$$

Define $\mathbf{w} := \mathbf{v} - \mathbb{I}_{\Gamma_1,\Gamma_2,\ldots,\Gamma_s}\mathbf{v}$. From (40), it follows that

$$\{\!|0|\!\}\,\mathbf{w}\cdot\mathbf{n}^{(i)}{}_{\Gamma_i} = 0 \qquad i = 1, 2, \ldots, d.$$

Therefore, we can apply Theorem 1 to $\mathbf{w}$ and find that

$$\|\mathbf{w}\|_\Omega \leq \mathbb{C}(\Omega, \Gamma_1, \ldots, \Gamma_d)\|\nabla\mathbf{w}\|_\Omega. \tag{41}$$

Since $\nabla\mathbf{w} = \nabla\mathbf{v}$, (41) yields the estimate

$$\boxed{\|\mathbf{v} - \mathbb{I}_{\Gamma_1,\Gamma_2,\ldots,\Gamma_d}\mathbf{v}\|_\Omega \leq \mathbb{C}(\Omega, \Gamma_1, \ldots, \Gamma_d)\|\nabla\mathbf{v}\|_\Omega,} \tag{42}$$

where $\mathbb{C}(\Omega, \Gamma_1, \ldots, \Gamma_d)$ depends on the constants C_{Γ_i} (see Sect. 3.2).

4.2 Cells with Curvilinear Faces

Let Ω be a Lipschitz domain with a piecewise smooth boundary consisting of faces $\Gamma_1, \Gamma_2, \ldots, \Gamma_N$ (see Fig. 3). In order to avoid complicated topological structures (which may lead to difficulties with definitions of "mean normals"), we assume that all the faces are such that normal vectors are defined at all points and impose an additional condition

$$\mathbf{n}_i(x^{(1)})\cdot\mathbf{n}_i(x^{(2)}) > 0 \qquad \forall x^{(1)}, x^{(2)} \in \Gamma_i, \quad i = 1, 2, \ldots, d.$$

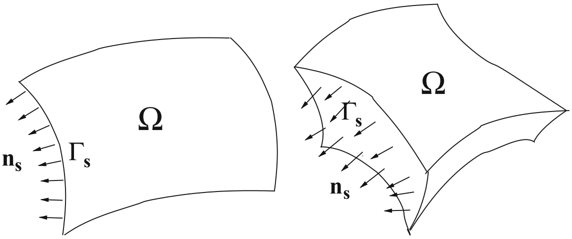


Fig. 3 Cells with curvilinear faces in 2D and 3D

Then, we can define the mean normal vector associated with Γ_i:

$$\widehat{\mathbf{n}}(i) := \left\{ \frac{1}{|\Gamma_i|} \int\limits_{\Gamma_i} n_1^{(i)}\, d\Gamma,\ \frac{1}{|\Gamma_i|} \int\limits_{\Gamma_i} n_2^{(i)}\, d\Gamma, \dots,\ \frac{1}{|\Gamma_i|} \int\limits_{\Gamma_i} n_d^{(i)}\, d\Gamma \right\}.$$

It is not difficult to verify that Theorem 1 holds if $\boldsymbol{N}$ is replaced by $\widehat{\boldsymbol{N}}$ formed by mean normal vectors, i.e.,

$$\det \widehat{\boldsymbol{N}} \neq 0, \quad \text{where } \widehat{n}_j^{(i)} := \widehat{\mathbf{n}}^{(i)} \cdot \mathbf{e}_j, \tag{43}$$

and (21) is replaced by the condition

$$\{\!|0|\!\}\, \mathbf{v} \cdot \widehat{\mathbf{n}}^{(i)}{}_{\Gamma_i} = 0 \quad i = 1, 2, \dots, d.$$

In other words, for cells with curvilinear faces the necessary interpolation condition reads as follows: *mean values of normal vectors averaged on faces must form a linearly independent system satisfying* (43).

The operator $\mathbb{I}_{\Gamma_1,\Gamma_2,\dots,\Gamma_d}\, \mathbf{v}$ is defined by modifying the condition (40). Since

$$\int\limits_{\Gamma_i} \mathbb{I}_{\Gamma_1,\Gamma_2,\dots,\Gamma_d} \mathbf{v} \cdot \mathbf{n}^{(i)}\, d\Gamma = \mathbb{I}_{\Gamma_1,\Gamma_2,\dots,\Gamma_d} \mathbf{v} \cdot \widehat{\mathbf{n}}^{(i)}\, \big|\Gamma_i\big|,$$

the function $\mathbb{I}_{\Gamma_1,\Gamma_2,\dots,\Gamma_d}\, \mathbf{v}$ is defined by the system

$$\sum_{j=1}^{d} \widehat{n}_j^{(i)} (\mathbb{I}_{\Gamma_1,\Gamma_2,\dots,\Gamma_d}\, \mathbf{v})_j = \frac{1}{|\Gamma_i|} \int\limits_{\Gamma_i} \mathbf{v} \cdot \mathbf{n}^{(i)}\, d\Gamma \quad i = 1, 2, \dots, d.$$

By repeating the same arguments, we obtain the estimate (42) for the function $\mathbb{I}_{\Gamma_1,\Gamma_2,\dots,\Gamma_d} \mathbf{v}$.

4.3 Comparison of Interpolation Constants for $\mathbb{I}_\Omega$ and $\mathbb{I}_\Gamma$

4.3.1 Triangles

First, we compare five different interpolation operators for the right triangle with equal legs (see Fig. 4). For the interpolation operator $\mathbb{I}_\Omega$ (Fig. 4a) we have (6), where (4) yields the upper bound of the respective interpolation constant

$$C_{\rm P}(\Omega) \leq \sqrt{2}\frac{h}{\pi} \approx 0.4502h.$$

Four different operators $\mathbb{I}_\Gamma$ are generated by setting zero mean values on one leg (Fig. 4b), two legs (Fig. 4c), median (Fig. 4d), and hypothenuse (Fig. 4e)

$$\|w - \mathbb{I}_\Gamma(w)\|_{2,\Omega} \leq C_\Gamma(\Omega)h\|\nabla w\|_{2,\Omega}.$$

The respective constants follow from Table 1. For Fig. 4b, $C_\Gamma(\Omega) = \frac{h}{\zeta} \approx 0.4929h$, for Fig. 4c $C_\Gamma(\Omega) = \frac{h}{\pi} \approx 0.3183h$, for Fig. 4d, e $C_\Gamma(\Omega) = \frac{h}{\zeta\sqrt{2}} \approx 0.3485h$.

We can use these data and compare the efficiency of $\mathbb{I}_\Gamma$ and $\mathbb{I}_\Omega$ for uniform meshes which cells are right equilateral triangles (Fig. 4f). For a mesh with 2 nm cells, the operator $\mathbb{I}_\Omega$ uses 2 nm parameters (mean values on triangles) and provides interpolation with the constant $C_{\rm P}$. The operator $\mathbb{I}_\Gamma$ using mean values on diagonals (see Fig. 4e) has almost the same constant but needs only nm parameters.

4.3.2 Squares

Similar results hold for square cells. For the interpolation operator $\mathbb{I}_\Omega$ (Fig. 5a) we have the exact constant $C_{\rm P} = \frac{\pi}{h}$. The constants for $\mathbb{I}_\Gamma$ are as follows. For Fig. 5b,

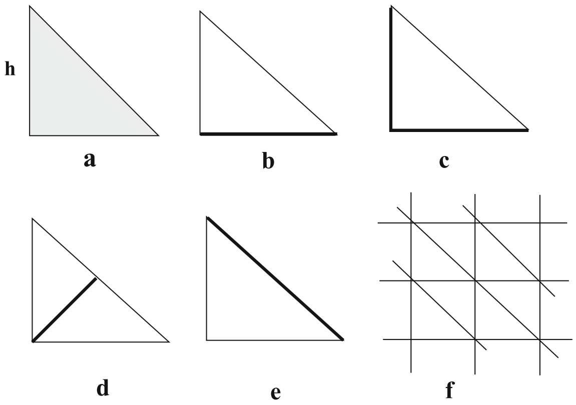


Fig. 4 Triangular cells

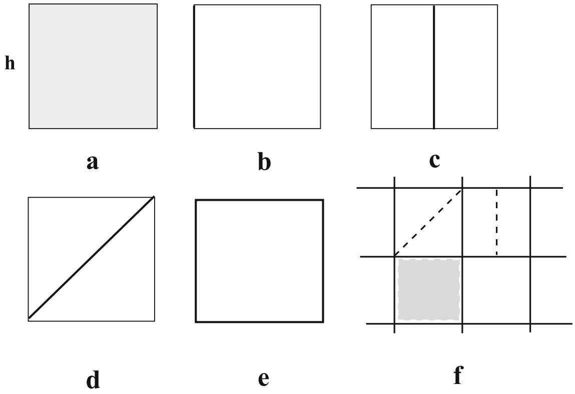


Fig. 5 Square cells

$C_\Gamma = \frac{h}{\pi}$, for Fig. 5c, d $C_\Gamma = \frac{2h}{\pi}$, and for Fig. 5e $C_\Gamma = \frac{h}{2.869}$. We see that for a uniform mesh with square cells $\mathbb{I}_\Gamma$ and $\mathbb{I}_\Omega$ have the same efficiency if Γ is selected as on Fig. 5d or e.

4.4 Interpolation on Macrocells

Advanced numerical methods and respective computer programs often operate with macrocells (see, e.g., [6, 7, 13, 14, 17, 20, 22] and references cited therein). Let Ω be a macrocell consisting of N simple subdomains ω_i (e.g., simplexes). Let the boundary Γ consist of faces Γ_i (each Γ_i is a part of some subdomain boundary $\partial\omega_i$). For $w \in H^1(\Omega)$ we define $\mathbb{I}_\Gamma w$ as a piecewise constant function that satisfies the conditions

$$\{\!|0|\!\}\, w - \mathbb{I}_\Gamma w_{\Gamma_i} = 0 \quad i = 1, 2, \ldots, N.$$

Then, we can apply interpolation operators $\mathbb{I}_{\gamma_i}$ to any subdomain ω_i and find that for the whole cell

$$\|w - \mathbb{I}_\Gamma w\|^2_{2,\Omega} = \sum_{i=1}^{N} \|w - \mathbb{I}_\Gamma w\|^2_{2,\omega_i} \le \sum_{i=1}^{N} C^2_{\gamma_i} \|\nabla w\|^2_{2,\omega_i} \le C^2_\Gamma \|\nabla w\|^2_{2,\Omega}, \quad (44)$$

where $C_\Gamma = \max_i\{C_{\Gamma_i}\}$.

Estimates for vector valued functions are derived quite similarly. For example, let $d = 2$ and Ω be a polygonal domain with N faces. If N is an odd number, then we form out of Γ_i a set of K pairs $\{\Gamma_1^{(l)}, \Gamma_2^{(l)}\}$, $l = 1, 2, \ldots, K$ such that the respective subdomains cover Ω and for each pair $\mathbf{n}_1^{(l)}$ and $\mathbf{n}_2^{(l)}$ satisfy (20). Then, $\mathbb{I}_\Gamma \mathbf{v}$ can be defined as a piecewise constant field in each pair of subdomains $\omega_1^{(l)} \cup \omega_2^{(l)}$ that satisfies

$$\{\!|0|\!\}\, (\mathbf{v} - \mathbb{I}_\Gamma \mathbf{v}) \cdot \mathbf{n}_{i\,\Gamma_i} = 0 \qquad i = 1, 2, \ldots, N.$$

Analogously to (44), we obtain

$$\|\mathbf{v} - \mathbb{I}_\Gamma \mathbf{v}\|_{2,\Omega} \leq \mathbb{C}\|\nabla \mathbf{v}\|_{2,\Omega} \qquad \mathbf{v} \in H^1(\Omega, \mathbb{R}^2), \tag{45}$$

where $\mathbb{C} = \max\limits_{l=1,2,\ldots,K} \mathbb{C}_{\Gamma_1^{(l)},\Gamma_2^{(l)}}(\omega_1^{(l)} \cup \omega_2^{(l)})$.

4.5 Interpolation on Meshes

Finally, we shortly discuss applications to mesh interpolation. It is clear that analogous operators $\mathbb{I}_\Gamma$ can be constructed for scalar and vector valued functions defined in a bounded Lipschitz domain Ω, which is covered by a mesh $\mathscr{T}_h$ with sells Ω_i, $i = 1, 2, \ldots, M_h$.

Let Ω_i be Lipschitz domains such that $\Omega_i \cup \Omega_j = \emptyset$ if $i \neq j$ and

$$\overline{\Omega} = \bigcup_{i=1}^{M_h} \overline{\Omega}_i.$$

We assume that $c_1 h \leq \operatorname{diam} \Omega_i \leq c_2 h$ for all $i = 1, 2, \ldots M_h$, where $c_2 \geq c_1 > 0$ and h is a small parameter. The intersection of $\overline{\Omega}_i$ and $\overline{\Omega}_j$ is either empty or a face Γ_{ij} (which is a Lipschitz domain in $\mathbb{R}^{d-1}$). By $\mathscr{E}_h$ we denote the collection of all faces in $\mathcal{T}_h$.

It is easy to see that a function $w \in H^1(\mathscr{D})$ can be interpolated by a piecewise constant function on cells of $\mathcal{T}_h$ if we set

$$\mathbb{I}_{\mathcal{T}_h}(w)(x) = \mathbb{I}_{\Gamma_i} w(x) = \{\!|0|\!\}\, w_{\Gamma_i} \qquad \text{if} \quad x \in \Omega_i.$$

Here Γ_i is a face of Ω_i selected for the local interpolation operator. Then,

$$\|w - \mathbb{I}_{\mathcal{T}_h}(w)\|_{2,\Omega} \leq C(\mathcal{T}_h)\, \|\nabla w\|_{2,\Omega}, \tag{46}$$

where $C(\mathcal{T}_h)$ is the maximal constant in inequalities (7) associated with Ω_i, $i = 1, 2, \ldots, M_h$. We note that the amount of parameters used in such type interpolation is essentially smaller than the amount of faces in $\mathcal{T}_h$.

If $\mathbb{I}_{\mathcal{T}_h}$ is constructed by means of averaging on each face Γ_{ij} then (46) holds with a better constant and $\mathbb{I}_{\mathcal{T}_h} w$ possesses an important property: *it preserves mean values of w.*

Similar consideration is valid for vector valued functions. If we define the interpolation operator $\mathbb{I}_{\mathcal{T}_h}(\mathbf{v})(x)$ on $\mathcal{T}_h$ by the conditions

$$\mathbb{I}_{\mathcal{T}_h} \mathbf{v} \cdot \mathbf{n}_{ij} = \{\!|0|\!\}\, \mathbf{v} \cdot \mathbf{n}_{ij\,\Gamma_{ij}} \quad \forall \Gamma_{ij} \in \mathcal{E}_h,$$

then

$$\|\mathbf{v} - \mathbb{I}_{\mathcal{T}_h}\mathbf{v}\|_{2,\Omega} \leq \mathbb{C}(\mathcal{T}_h)\, \|\nabla \mathbf{v}\|_{2,\Omega},$$

where $\mathbb{C}(\mathcal{T}_h)$ is the maximal constant in the inequalities (45) used for Ω_i, $i = 1, 2, \ldots, N(\mathcal{T}_h)$. The function $\mathbb{I}_{\mathcal{T}_h}\mathbf{v}$ possesses an important property: *it preserves mean values of* $\mathbf{v} \cdot \mathbf{n}_{ij}$ on all the faces of $\mathcal{T}_h$.

References

1. Acosta G, Durán RG (2004) An optimal Poincaré inequality in L^1 for convex domains. Proc Amer Math Soc 132(1):195–202
2. Arnold D, Boffi D, Falk R (2002) Approximation by quadrilateral finite elements. Math Comp 71(239):909–922
3. Arnold D, Boffi D, Falk R (2005) Quadrilateral H(div) finite elements. SIAM J Numer Anal 42(6):2429–2451
4. Babuška I, Aziz A (1976) On the angle condition in the finite element method. SIAM J Numer Anal 13(2):214–226
5. Bermúdez A, Gamallo P, Nogueiras MR, Rodríguez R (2005) Approximation properties of lowest-order hexahedral Raviart-Thomas finite elements. C R Math Acad Sci Paris 340(9):687–692
6. Brezzi F, Fortin M (1991) Mixed and hybrid finite element methods. Springer, New York
7. Brezzi F, Lipnikov K, Shashkov M, Simoncini V (2007) A new discretization methodology for diffusion problems on generalized polyhedral meshes. Comput Methods Appl Mech Engrg 196(37–40):3682–3692
8. Cheng SY (1975) Eigenvalue comparison theorems and its geometric applications. Math Z 143(3):289–297
9. Chua S-K, Wheeden RL (2006) Estimates of best constants for weighted Poincaré inequalities on convex domains. Proc London Math Soc (3), 93(1):197–226
10. Chua S-K, Wheeden RL (2010) Weighted Poincaré inequalities on convex domains. Math Res Lett 17(5):993–1011
11. Fox DW, Kuttler JR (1983) Sloshing frequencies. Z Angew Math Phys 34(5):668–696
12. Girault V, Raviart PA (1986) Finite element methods for Navier-Stokes equations: theory and algorithms. Springer, Berlin
13. Hackbusch W, Löhndorf M, Sauter SA (2006) Coarsening of boundary-element spaces. Computing 77(3):253–273
14. Hecht F (2012) New development in FreeFem++. J Numer Math 20(3–4):251–265
15. Kozlov V, Kuznetsov N (2004) The ice-fishing problem: The fundamental sloshing frequency versus geometry of holes. Math Methods Appl Sci 27(3):289–312
16. Kozlov V, Kuznetsov N, Motygin O (2004) On the two-dimensional sloshing problem. Proc R Soc Lond Ser A Math Phys Eng Sci 460(2049):2587–2603
17. Kuznetsov Yu (2006) Mixed finite element method for diffusion equations on polygonal meshes with mixed cells. J Numer Math 14(4):305–315
18. Kuznetsov Yu (2011) Approximations with piece-wise constant fluxes for diffusion equations. J Numer Math 19(4):309–328
19. Kuznetsov Yu (2014) Mixed FE method with piece-wise constant fluxes on polyhedral meshes. Russian J Numer Anal Math Modelling 29(4):231–237
20. Kuznetsov Yu (2015) Error estimates for the RT_0 and PWCF methods for the diffusion equations on triangular and tetrahedral meshes. Russian J Numer Anal Math Modelling 30(2):95–102
21. Kuznetsov Yu, Prokopenko A (2010) A new multilevel algebraic preconditioner for the diffusion equation in heterogeneous media. Numer Linear Algebra Appl 17(5):759–769

22. Kuznetsov Yu, Repin S (2003) New mixed finite element method on polygonal and polyhedral meshes. Russian J Numer Anal Math Modelling 18(3):261–278
23. Laugesen RS, Siudeja BA (2010) Minimizing Neumann fundamental tones of triangles: an optimal Poincaré inequality. J Differen Equat 249(1):118–135
24. Mali O, Neittaanmäki P, Repin S (2014) Accuracy verification methods: Theory and algorithms, vol 32. Computational Methods in Applied Sciences. Springer, Dordrecht
25. Matculevich S, Neittaanmäki P, Repin S (2015) A posteriori error estimates for time-dependent reaction-diffusion problems based on the Payne-Weinberger inequality. Discrete Contin Dyn Syst 35(6):2659–2677
26. Matculevich S, Repin S (2016) Explicit constants in Poincaré-type inequalities for simplicial domains and application to a posteriori estimates. Comput Methods Appl Math 16(2):277–298
27. Nazarov A, Repin S (2015) Exact constants in Poincaré type inequalities for functions with zero mean boundary traces. Math Methods Appl Sci 38(15):3195–3207
28. Payne LE, Weinberger HF (1960) An optimal Poincaré inequality for convex domains. Arch Rational Mech Anal 5:286–292
29. Poincaré H (1894) Sur les équations de la physique mathématique. Rend Circ Mat Palermo 8:57–155
30. Repin S (2008) A posteriori estimates for partial differential equations. Walter de Gruyter, Berlin
31. Repin S (2015) Estimates of constants in boundary-mean trace inequalities and applications to error analysis. In: Abdulle A, Deparis S, Kressner D, Nobile F, Picasso M, (eds) Numerical Mathematics and Advanced Applications – ENUMATH2013, volume 103 of Lecture Notes in Computational Science and Engineering, pp 215–223
32. Repin S (2015) Interpolation of functions based on Poincaré type inequalities for functions with zero mean boundary traces. Russian J Numer Anal Math Modelling 30(2):111–120
33. Roberts JE, Thomas J-M (1991) Mixed and hybrid methods. In: Handbook of Numerical Analysis, Vol II, pp 523–639. North-Holland, Amsterdam,
34. Steklov VA (1896) On the expansion of a given function into a series of harmonic functions. Commun Kharkov Math Soc Ser 2(5):60–73 (in Russian)

Ensemble Interpretation of Quantum Mechanics and the Two-Slit Experiment

Glenn F. Webb

Abstract An evolution equation model is provided for the two-slit experiment of quantum mechanics. The state variable of the equation is the probability density function of particle positions. The equation has a local diffusion term corresponding to stochastic variation of particles, and a nonlocal dispersion term corresponding to oscillation of particles in the transverse direction perpendicular to their forward motion. The model supports the ensemble interpretation of quantum mechanics and gives descriptive agreement with the Schrödinger equation model of the experiment.

1 Introduction

In the two-slit experiment of quantum mechanics electrons or other quantum particles are randomly directed toward two slits one at a time, and then detected on a screen downstream. If only one slit is open, a roughly Gaussian pattern is observed corresponding to the width of the slit. But, if both slits are open, an interference pattern of regularly spaced intensities is observed, which is not the sum of the patterns observed for the slits separately [3, 6, 19, 21, 26, 37–39]. A detailed description of the two-slit experiment in given in [10]. Feynman [14] observed that the interference pattern arising from the two-slit experiment is

> ... impossible, absolutely impossible to explain in any classical way, and has in it the heart of quantum mechanics.

It is possible to formulate this experiment in terms of the time-dependent complex-valued non-relativistic Schrödinger equation, which is the foundational model of quantum mechanics:

G. F. Webb (✉)
Vanderbilt University, Nashville, TN, USA
e-mail: glenn.f.webb@vanderbilt.edu

B. N. Chetverushkin et al. (eds.), *Contributions to Partial Differential Equations and Applications*, Computational Methods in Applied Sciences 47,
https://doi.org/10.1007/978-3-319-78325-3_23

$$\frac{\partial}{\partial t}\psi(x,t) = i\frac{\hbar}{2m}\frac{\partial^2}{\partial x^2}\psi(x,t), \quad t>0,\ \psi(x,0)=\psi_0(x),\ -\infty<x<\infty. \quad (1)$$

Here $\hbar$ is the reduced Planck's constant and m is the particle mass, which without loss of generality can be assumed to satisfy $\frac{\hbar}{m}=1$. The interpretation of the solution is that $\int_{x_1}^{x_2}\rho(x,t)dx$ is the probability of finding the particle in the interval (x_1,x_2) at time t, where $\rho(x,t)=|\psi(x,t)|^2$, and $\rho(x,0)=|\psi_0(x)|^2$ is normalized so that $1=\int_{-\infty}^{\infty}\rho(x,0)dx$. This formulation, however, is ambiguous in its interpretation of the state variable ψ, time t, and the initial condition ψ_0. What do the real and imaginary parts of ψ represent? What is time t in an experiment with randomly separated independent temporal events? What does the initial condition ψ_0 correspond to for single particles emitted one at a time?

In current paradigms of quantum mechanics this ambiguity is answered by specifying the Schrödinger equation solutions to individual particle behavior, rather than to aggregate multi-particle behavior. Although it is always possible to apply a probabilistic population model to a single individual, these paradigms extend their probabilistic interpretation to an assumption of multi-state existences of single individuals. In these interpretations a single particle may exist in indefinite super-positioned multi-states simultaneously, traverse all possible paths from source to detection, collapse to one state with detection by a conscious observer, and bifurcate to separate existences in parallel universes. An alternative to these metaphysical interpretations of quantum mechanics is the statistical or ensemble interpretation of quantum mechanics advanced by Einstein [13, 31], Born [8], Popper [29], Lande [22, 23], and Ballentine [4, 5]. In this view of quantum mechanics, the Schrödinger equation is only a mathematical probabilistic description of ensemble behavior. In the view of Einstein [31]

> The attempt to conceive the quantum-theoretical description as the complete description of the individual systems leads to unnatural theoretical interpretations, which become immediately unnecessary if one accepts the interpretation that the description refers to ensembles of systems and not to individual.

Our objective here is to support the ensemble interpretation of quantum mechanics by providing an alternative differential equation model for the two-slit experiment, which has an evident ensemble probabilistic interpretation, and which gives descriptive agreement with the Schrödinger equation model.

2 Schrödinger Equation for the Two-Slit Experiment

We interpret the initial condition of the Schrödinger equation as a probabilistic distribution of possible starting points of individual particles in relation to their arrival at the two slits. We first consider the initial condition $\psi(x,0)=\text{Re}\psi(x,0)$ in (1) corresponding to $\rho(x,0,s)$ as a single Gaussian distribution with mean s and standard deviation a:

$$\psi(x,0,s) = \frac{\exp\left(-\frac{(s-x)^2}{4a^2}\right)}{(2\pi)^{\frac{1}{4}}\sqrt{a}}, \quad \rho(x,0,s) = \frac{\exp\left(-\frac{(s-x)^2}{2a^2}\right)}{\sqrt{2\pi}a}. \tag{2}$$

In this case the solution of (1) for the single slit is given by the formulas

$$\psi(x,t,s) = \frac{2^{\frac{1}{4}}\sqrt{a}\exp\left(-\frac{(s-x)^2}{2(2a^2+it)}\right)}{\pi^{\frac{1}{4}}\sqrt{2a^2+it}}, \quad \rho(x,t,s) = \frac{\sqrt{2}a\exp\left(-\frac{\left(\sqrt{2}a(s-x)\right)^2}{4a^4+t^2}\right)}{\sqrt{\pi}\sqrt{4a^2+t^2}} \tag{3}$$

For the case of two slits the initial data $\psi(x,0) = \mathrm{Re}\psi(x,0)$ in (1) corresponds to $\rho(x,0)$ as two Gaussian distributions, both with standard deviation a, and means at $\pm s$ from the origin. We note that by scaling with respect to s, we can assume without loss of generality that $s = 1$. We assume that the standard deviation parameter a is very much less than the slit separation distance $2s = 2$. Let

$$\begin{aligned} \psi(x,0) &= \frac{\exp\left(\frac{-(1+x)^2}{2a^2}\right)\left(1+\exp(\frac{2x}{a^2})\right)}{\pi^{\frac{1}{4}}\sqrt{2a}}, \\ \rho(x,0) &= \frac{\exp\left(\frac{-(1+x)^2}{a^2}\right)\left(1+\exp(\frac{2x}{a^2})\right)^2}{\sqrt{\pi}2a}. \end{aligned} \tag{4}$$

Then $\psi(x,t)$ and $\rho(x,t) = |\psi(x,t)|^2$ for the two-slit case satisfy (1), where

$$\begin{aligned} \psi(x,t) &= \frac{\sqrt{a}\exp\left(-\frac{(1+x^2)}{a^2+it}\right)\left(\exp\left(\frac{(x-1)^2}{2(a^2+it)}\right)+\exp\left(\frac{(x+1)^2}{2(a^2+it)}\right)\right)}{\sqrt{2}\pi^{\frac{1}{4}}\sqrt{a^2+it}} \\ \rho(x,t) &= \frac{a\exp\left(-\frac{a^2+a^2x^2}{a^4+t^2}\right)\left(\cosh\left(\frac{2a^2x}{a^4+t^2}\right)+\cos\left(\frac{2tx}{a^4+t^2}\right)\right)}{\sqrt{\pi}\sqrt{a^4+t^2}\left(e^{-\frac{1}{a^2}}+1\right)} \end{aligned} \tag{5}$$

(these formulas are found in [24, 25, 40]). The initial data (4) for the two-slit case is illustrated in Fig. 1.

The solutions of (1) exhibit a two phase pattern as time advances. In the first phase the initial information $\rho_0(x)$ evolves to an established pattern and in the second phase the pattern established in the first phase undergoes a space-time dilation [35, 40]. The first phase of $\rho(x,t)$ (Fig. 2) exhibits an extremely elaborate transition from the initial condition $\rho_0(x)$ to the formed interference pattern completed at time $t \approx 1/\pi$ for a sufficiently small. The established pattern has the property that the bottom peaks touch-down (approximately) to the x-axis.

In the second phase the pattern established in the first phase undergoes a space-time dilation. Specifically, the probability amplitude $\rho(x,t)$ satisfies as T increases and $t \geq 1$

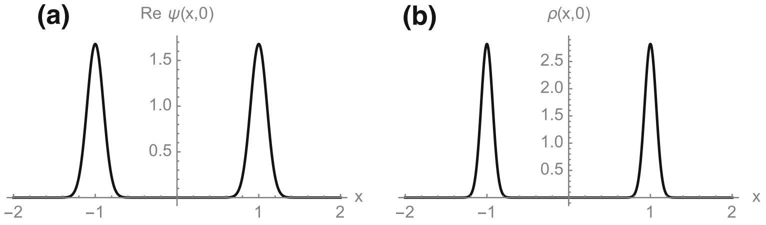


Fig. 1 Initial data for the two-slit experiment corresponding to two Gaussian distributions as in (4), with $a = 0.1$. **a** $\psi(x, 0) = \text{Re}\psi(x, 0)$; **b** the initial probability amplitude $\rho(x, 0) = |\psi(x, 0)|^2$. The imaginary part of $\psi(x, 0)$ is 0

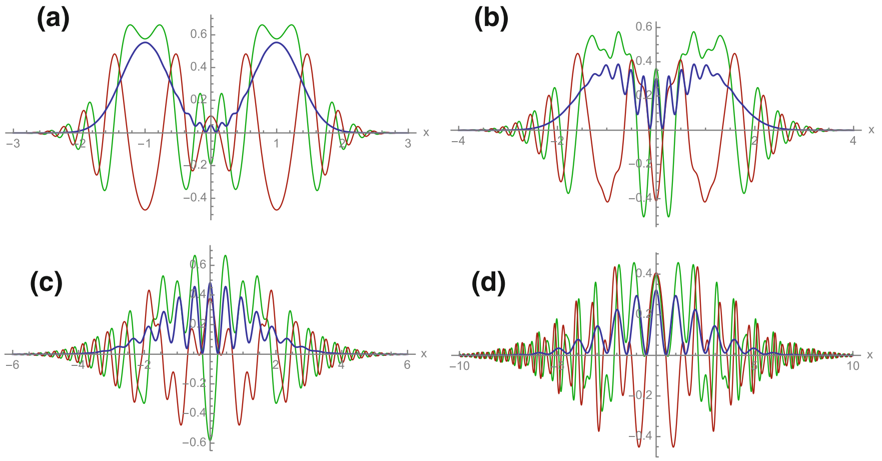


Fig. 2 The first phase of $\rho(x, t)$ and $\psi(x, t)$ with initial condition as in Fig. 1 with $a = 0.1$ at four different times. $\rho(x, t)$ (blue), $\text{Re}\psi(x, t)$ (green), $Im\psi(x, t)$ (red). **a** $t = 0.05$; **b** $t = 0.08$; **c** $t = 0.15$; **d** $t = 1.0/\pi$. The first phase is completed at time $\approx 1.0/\pi$ for a sufficiently small

$$\rho(x, tT) \approx \frac{1}{t}\rho\left(\frac{x}{t}, T\right). \tag{6}$$

In the second phase of $\rho(x, t)$ the information in $\rho(x, t)$ is (essentially) unchanging in time *ad infinitum*. Indeed, the profiles of $\rho(x, tT)$ and $\frac{1}{t}\rho(\frac{x}{t}, T)$ become almost indistinguishable as T increases (Fig. 3). In the second phase the peaks of the wave profile $\rho(x, t)$ propagated in the traverse x-direction at constant velocity x/π, provided a is sufficiently small. That is, the first right peak occurs at t at time t/π, the second right peak is at $2t/\pi$ at time t/π, etc. In [40] it is proved that for general initial data $\psi_0 \in L^1(((-\infty, \infty); C), y^2dy) \cap L^2((-\infty, \infty); C)$ in (1) the probability distribution $\rho(x, t)$ satisfies the asymptotic space-time dilation property: for $x \in (-\infty, \infty)$, $T > 0$, $t \geq 1$

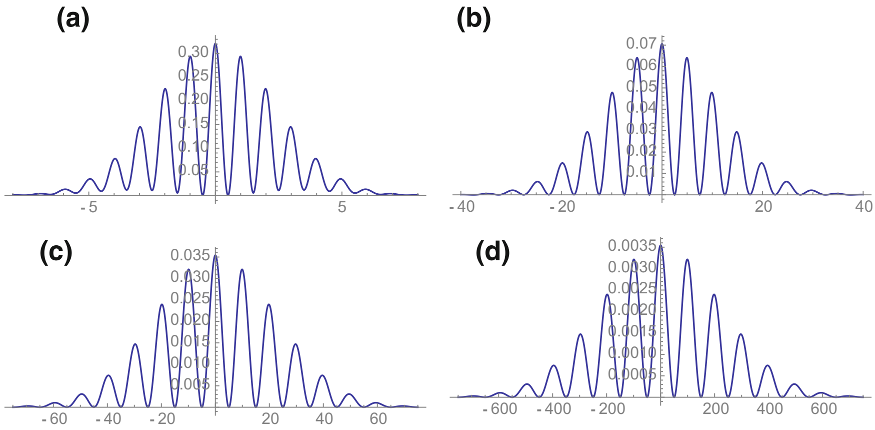


Fig. 3 The second phase of $\rho(x, t)$ with initial condition as in Fig. 1 with $a = 0.1$ at four different times. **a** $\rho(x, 1.0/\pi)$; **b** $\rho(x, 5.0/\pi)$; **c** $\rho(x, 10.0/\pi)$; **d** $\rho(x, 100.0/\pi)$

$$\left|\rho(x, tT) - \frac{1}{t}\rho\left(\frac{x}{t}, T\right)\right| \leq \frac{\sqrt{2}}{\pi t T^2}\left(\int_{-\infty}^{\infty} y^2|\psi_0(y)|dy\right)\left(\int_{-\infty}^{\infty} |\psi_0(y)|dy\right). \quad (7)$$

In [40] it is proved that the relationship of $u(x, tT)$ and $\frac{1}{t}u(\frac{x}{t}, T)$ is an exact equality for the solution of the partial differential equation

$$\frac{\partial}{\partial t}u(x, t) = -\frac{1}{t}\frac{\partial}{\partial x}\left(xu(x, t)\right), \quad u(x, 1) = u_1(x),\ u_1 \in L^1(-\infty, \infty),\ x \in (-\infty, \infty),\ t > 1, \quad (8)$$

where $u(x, t) = \frac{1}{t}u_1(\frac{x}{t})$, with $u_1(x)$ sufficiently smooth.

3 Ensemble Equation for the Two-Slit Experiment

3.1 *First Phase of the Ensemble Equation*

The ensemble equation for the two-slit experiment has as independent variables the transverse direction parallel to the slit openings and the detection plate (x-coordinate), and the forward direction perpendicular to the slit openings and detection plate (z-coordinate). It is assumed that particle positions vary relative to the experimental apparatus and are not uniformly absolute with respect to any reference point. For simplicity, it is assumed that the position of particles is independent of the vertical height of the slit openings.

The dependent variable $\omega(x, z)$ in the ensemble equation is the probability density function for the distribution of particle positions at the detection distance z. Time in the ensemble model of the two-slit experiment has no role, since quantum particles are sent toward the slits independently in time. The time variable t in the Schrödinger Equation (1) can be correlated to the downstream distance z in the ensemble equation. The ensemble equation is a phenomenological model for aggregate macro-behavior of particle movement, rather than a mechanistic model for individual micro-behavior of particle movement.

The ensemble equation incorporates a local diffusion term (with parameter α) and a nonlocal dispersion term (with parameter β). The local diffusion term represents stochastic variation in the x-coordinate. The nonlocal dispersion term represents environmental signaling in an x-coordinate signaling range. The nonlocal dispersion term corresponds to an x-direction movement through an environment with outermost reach equal to the discrete slit separation constant s. The equations of the model are

$$\begin{aligned} \frac{\partial}{\partial z}\omega(x,z) &= \alpha\frac{\partial^2}{\partial x^2}\omega(x,z) + \beta\frac{\partial}{\partial x}\left(\int_{-s}^{s}\omega(x+\hat{x},z)\frac{\hat{x}}{|\hat{x}|}d\hat{x}\right), \\ &= \alpha\frac{\partial^2}{\partial x^2}\omega(x,z) + \beta\left(\omega(x+s,z) - 2\omega(x,z) + \omega(x-s,z)\right), \\ & \qquad z > 0,\ -\infty < x < \infty, \\ \omega(x,0) &= \omega_0(x),\ \omega_0 \in L^1_+(-\infty,\infty),\ \int_{-\infty}^{\infty}\omega_0(x)dx = 1. \end{aligned} \tag{9}$$

The solution of (9) has the properties that $\omega(x,z) \geq 0, z \geq 0$, and $\int_{-\infty}^{\infty}\omega(x,z)dx = 1$ for $z \geq 0$, and may thus be viewed as probability density functions. The nonlocal term in (9) is similar to models of biological cell movement in which individual cells have a sensing radius for reaction to their environment [2, 12, 27, 32]. Such terms model aggregate cell population behavior, such as contact-mediated dispersal, cell-cell adhesion, and self-organization of spatial patterning. The simulations of cell population patterns in such models bear a remarkable similarity to quantum interference patterns [27, 32].

As with the solutions of the Schrödinger Equation (1), the solutions $\omega(x,z)$ of (9), with initial data corresponding to the two-slit experiment, have a two-phase behavior—the first phase in which the initial data $\omega(x,0) = \rho(x,0)$ as in (5) is transitioned in the z coordinate to an interference pattern in the x coordinate, and the second phase in which the bottom peaks of $\omega(x,z)$ undergo (approximately) a lift-off from the x-axis. We illustrate the first phase in Fig. 4 with the initial condition $\rho(x,0)$ as in Fig. 1 ($a = 0.1$). The diffusion parameter is $\alpha = 1/8\pi$ and the nonlocal dispersion parameter is $\beta = 1/(4a)^2$. The initial data transitions to the formed interference pattern in a simple way. As with the solutions of (1), where the first phase is completed at time $t \approx 1/\pi$, the first phase of the solutions of (9) are completed in the z-direction at $z \approx 1/\pi$ (a sufficiently small). The choice of the parameters α and β assures that $\rho(x,t)$ and $\omega(x,z)$ align (approximately) at the end of the first phase, with peak spacing at $x \approx 0, \pm1, \pm2, \ldots$, independently of a, for a sufficiently small. In Fig. 5 we compare $\rho(x,t)$ and $\omega(x,z)$ at the end of the first phase for four different values of the slit parameter a.

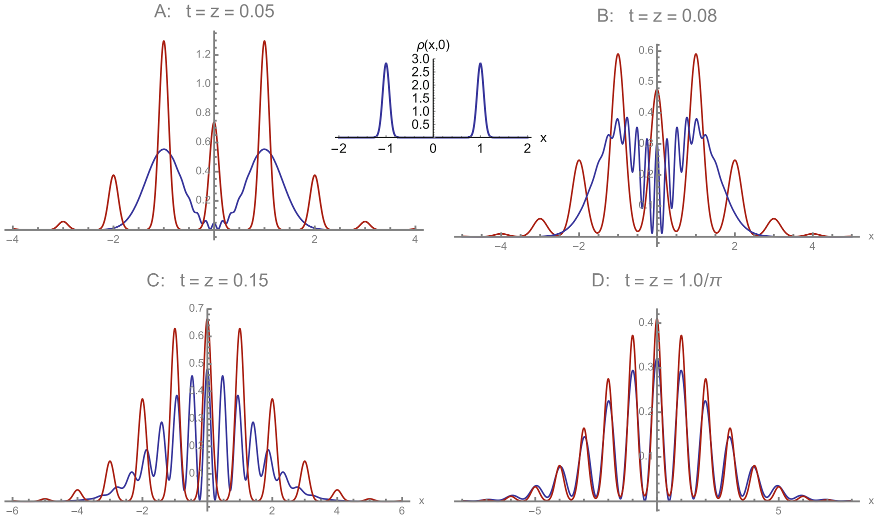


Fig. 4 The first phase of $\omega(x, t)$ (red) and $\rho(x, t)$ (blue) with initial condition $\rho(x, 0)$ as in Fig. 1 at four different times. **a** $t = 0.05$; **b** $t = 0.08$; **c** $t = 0.15$; **d** $t = 1/\pi$. The parameters are $\alpha = 1/(8\pi)$, $\beta = 1.0/(4.0a)^2$, $a = 0.1$. The areas under all the graphs are ≈ 1.0

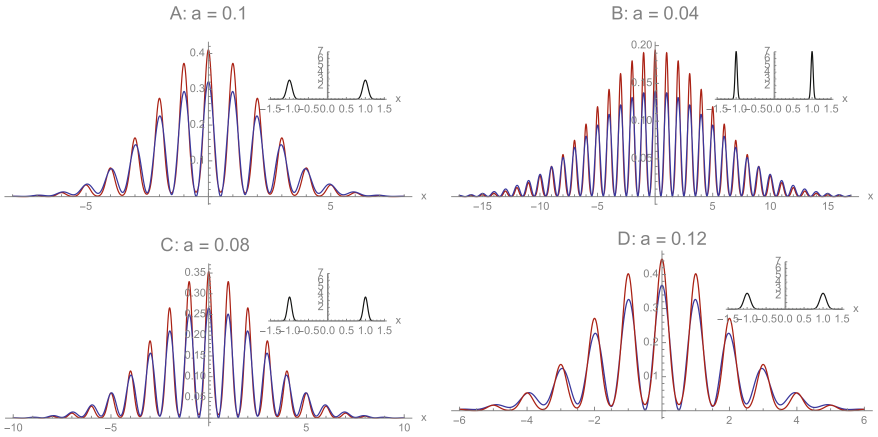


Fig. 5 $\rho(x, 1/\pi)$ (blue) and $\omega(x, 1/\pi)$ (red) at the end of the first phase with initial condition $\rho(x, 0)$ as in (5) with four different values of the standard deviation parameter a. The parameters are $\alpha = 1/(8\pi)$, $\beta = 1.0/(4.0a)^2$. **a** $a = 0.1$; **b** $a = 0.04$; **c** $a = 0.08$; **d** $a = 0.12$. The peak spacing occurs at $x \approx 0, \pm1, \pm2, \ldots$, independently of a, for a sufficiently small

3.2 Second Phase of the Ensemble Equation

In the second phase of the ensemble equation, after the interference pattern is established, the solutions of the Schrödinger Equation (1) and the ensemble Equation (5) are very different. The solutions $\rho(x,t)) = |\psi(x,t)|^2$ of the Schrödinger Equation (1) undergo space-time dilation in the traverse x-direction. In the Appendix we prove that the solutions $\omega(x,z)$ of the ensemble Equation (9), for any initial condition ω_0, are asymptotic as $z \to \infty$ to the Gaussian distribution $\exp\left(-\frac{x^2}{2\sigma^2}\right)/\sqrt{2\pi}\sigma$ with mean 0 and standard deviation $\sigma = \sqrt{2.0(\alpha+\beta)z}$. The information in the Schrödinger equation is conserved for all time, whereas the information in the ensemble equation is dispersed over increasing distance from the plane of the 2-slits. In a letter from H. A. Lorentz to Schrödinger, May 27, 1926, Lorentz questioned Schrödinger's recently proposed wave equation as a model of a moving wave packet for a representation of a particle [30]:

> …But a wave packet can never stay together and remain confined to a small volume in the long run. The slightest dispersion in the medium will pull it apart in the direction of propagation, and even without that dispersion it will always spread more and more in the transverse direction. Because of this unavoidable blurring a wave packet does not seem to be very suitable for representing things to which we want to ascribe a rather permanent individual existence ….

In Fig. 6 we illustrate the dispersion of solutions of the ensemble equation in the second phase. In Fig. 7 we compare the Schrödinger equation solutions and the ensemble equation solutions in the second phase.

4 Schrödinger and Ensemble Equations with 1-Slit

The Schrödinger equation and the ensemble equation are comparable if only one slit is open. We take as the initial condition for the Schrödinger equation $\psi(x,0,s)$ with $s = 1$ as in (2), which corresponds to a single slit on the right-hand side one spatial unit distance from the origin. We take as the initial condition for the ensemble equation $\rho(x,0,1)$ as in (2). We note that the integral of $\rho(x,0,1)$ from $-\infty$ to $+\infty$ is 1.

We take as the ensemble equation for this one slit case equation (9) with $\alpha = 3/(4a)^2$ and $\beta = 1/(4a)^2$. Since this value of α is different from the value of α for the two-slit ensemble Equations (9), the solution of the ensemble equation for two slits is not the sum of the solutions of the ensemble equation for the right-hand slit and the left-hand slit.

There is again a two-phase behavior of the probability amplitudes $\rho(x,t)$. The first phase of $\rho(x,t)$ is completed at time $t \approx 2.0$ for $s = 1.0$ and a sufficiently small, and in the second phase, $\rho(x,t)$ undergoes a space-time dilation as in Eq. 7 (Fig. 8). In Fig. 9 both $\rho(x,t)$ and $\omega(x,z)$ are graphed at $t = z = 2.0$ (the end of the first phase of $\rho(x,t)$) for different values of the parameter a, where it is seen that the two graphs

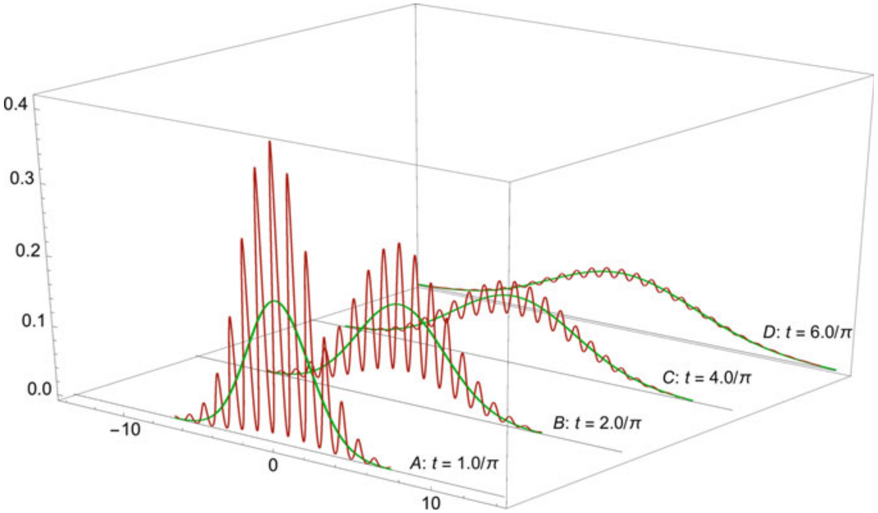


Fig. 6 The second phase of the solutions $\omega(x, t)$ of the ensemble Equation (9) (red graphs) with initial condition $\omega(x, 0) = \rho_0(x)$ as in Fig. 1 at four different values of the propagation coordinate $z = t$. **a** $\omega(x, 1.0/\pi)$; **b** $\omega(x, 2.0/\pi)$; **c** $\omega(x, 4.0/\pi)$; **d** $\omega(x, 6.0/\pi)$. The parameters are $\alpha = 1/(8\pi)$, $\beta = 1.0/(4.0a)^2$, $a = 0.1$. The green graphs are the Gaussian distributions $\exp\left(-\frac{x^2}{2\sigma^2}\right)/\sqrt{2\pi}\sigma$ with $\sigma = \sqrt{2.0(\alpha + \beta)z} = \sqrt{2.0(\alpha + \beta)t}$

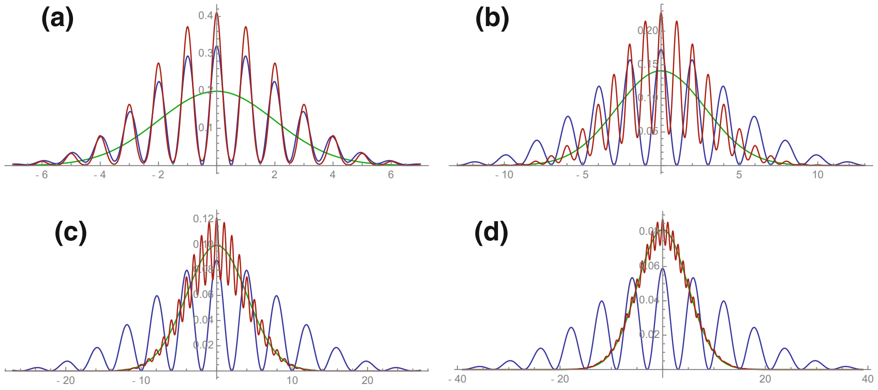


Fig. 7 The second phase of the solutions $\rho(x, t)$ of the Schrödinger equation (blue) and the ensemble equation $\omega(x, t)$ (red) with initial condition $\rho(x, 0) = \omega(x, 0) = \rho_0(x)$ as in Fig. 1 at four different values of the propagation coordinate $z = t$. **a** $\omega(x, 1.0/\pi)$; **b** $\omega(x, 2.0/\pi)$; **c** $\omega(x, 4.0/\pi)$, **d** $\omega(x, 6.0/\pi)$. The parameters are $\alpha = 1/(8\pi)$, $\beta = 1.0/(4.0a)^2$, $a = 0.1$. The green graphs are the Gaussian distributions as in Fig. 6

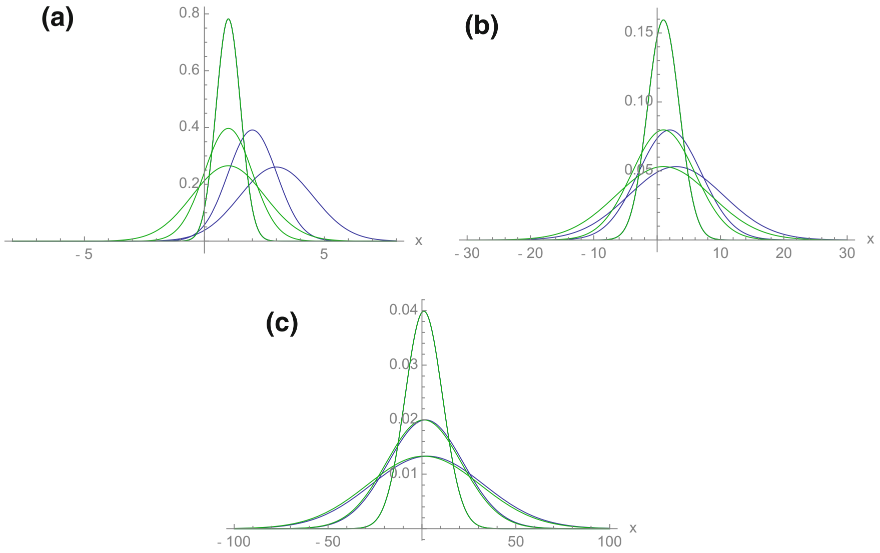


Fig. 8 The probability amplitudes $\rho(x, t)$ of the Schrödinger equation with only the right slit open at various times. The green graphs are $\rho(x, tT)$ and the blue graphs are $\rho(x/t, T)/t$. **a** $T = 0.1$, $t = 1, 2, 3$; **b** $T = 0.5$, $t = 1, 2, 3$; **c** $T = 2.0$, $t = 1, 2, 3$. The solutions exhibit space-time dilation for $T > 2.0$. The standard deviation parameter is $a = 0.1$

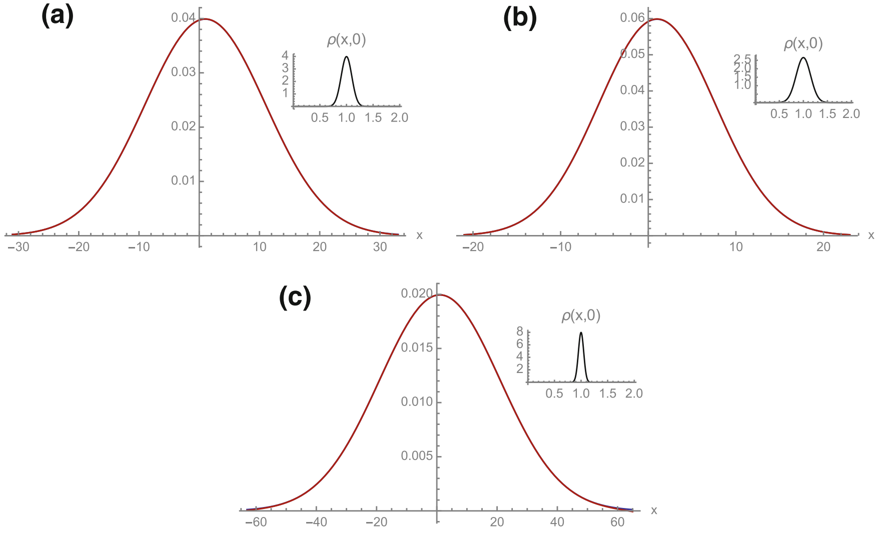


Fig. 9 The graphs of $\rho(x, 2.0)$ (blue) and $\omega(x, 2.0)$ (red) for **a** $a = 0.1$; **b** $a = 0.15$; **c** $a = 0.05$. The parameters in (9) are $\alpha = 3/(4a)^2$ and $\beta = 1\ (4a^2)$. The graphs are almost identical. The parameters for the ensemble Equation (9) are $a = 0.1$, $\alpha = 3/(4a)^2$, and $\beta = 1/(4a)^2$

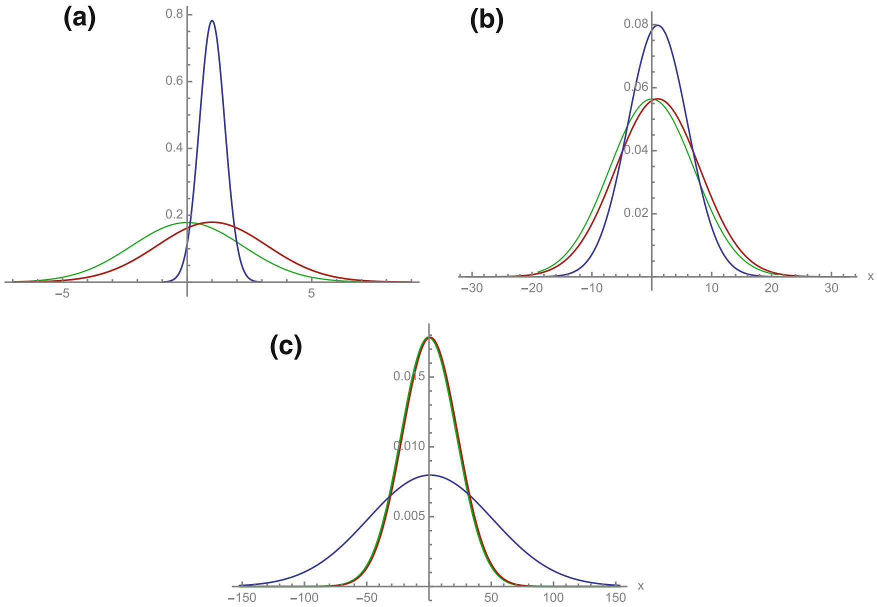


Fig. 10 The graphs of $\rho(x,t)$ (blue) and $\omega(x,z)$ (red) for **a** $t=z=0.1$; **b** $t=z=1.0$; **c** $t=z=10.0$. The parameter $a=0.1$. The parameters for the ensemble Equation (9) are $\alpha = 3/(4a)^2$, and $\beta = 1/(4a)^2$. The green graphs are Gaussians with mean 0 and standard deviation $\sigma = \sqrt{2.0(\alpha+\beta)z} = \sqrt{z/2}/a$. The graphs of $\rho(x,t)$ disperse much faster than the graphs of $\omega(x,z)$ as $t = z \to \infty$

essentially agree. The behavior of $\rho(x,t)$ and $\omega(x,z)$ as $z=t\to\infty$ is illustrated in Fig. 10, where it is seen that $\rho(x,t)$ disperses in the transverse x-direction much faster than $\omega(x,z)$ as $t=z\to\infty$.

5 Asymptotic Behavior of the Schrödinger and Ensemble Equations

For arbitrary initial data ψ_0 the probability amplitude $\rho(x,t)$ of the Schrödinger Equation (1) satisfies [18, 20]

$$\rho(x,t) = \frac{1}{2\pi t}\left|\int_{-\infty}^{\infty} e^{i(\frac{y^2-2xy}{2t})}\psi_0(y)dy\right|^2, \quad -\infty < x < \infty,\ t>0,$$

which implies for $x \in$ bounded intervals of $(-\infty,\infty)$

$$\lim_{t\to\infty} t\rho(x,t) = \frac{1}{2\pi}\left|\int_{-\infty}^{\infty}\psi_0(y)dy\right|^2.$$

For arbitrary initial data ρ_0 the solution $\omega(x,z)$ of the ensemble Equation (9) satisfies (see Appendix)

$$\lim_{t\to\infty}\sqrt{z}\omega(x,z)=\frac{1}{2\sqrt{\alpha+\beta}}\quad\text{uniformly in bounded intervals of } x.$$

For the one slit and the two slit examples, the asymptotic behavior of $\rho(x,t)$ and $\omega(x,z)$ as $t=z\to\infty$ is as follows: For one slit initial data $\psi(x,0,1)$ and $\rho(x,0,1)$ as in (2), (3) implies

$$\lim_{t\to\infty} t\rho(x,t)=a\sqrt{2/\pi}.$$

For one slit and initial data $\omega(x,0)=\rho(x,0,1)$ as in (2), $\alpha=3.0/(4.0a)^2$, $\beta=1.0/(4.0a)^2$ as in (9)

$$\lim_{z\to\infty}\sqrt{z}\omega(x,z)=\frac{1}{2\sqrt{\pi(\alpha+\beta)}}=\frac{2a^2}{\sqrt{\pi}}.$$

For two slits and initial data $\psi(x,0,1)$ and $\rho(x,0,1)$ as in (3), (5) implies

$$\lim_{t\to\infty} t\rho(x,t)=\frac{2a}{\sqrt{\pi}}(1+e^{-1/a^2})\approx\frac{2a}{\sqrt{\pi}},\quad\text{for } a \text{ sufficiently small.}$$

For two slits and initial data $\omega(x,0)=\rho(x,0,1)$ as in (3), $\alpha=1.0/(8.0\pi)$, $\beta=1.0/(4.0a)^2$ in (9)

$$\lim_{z\to\infty}\sqrt{z}\omega(x,z)=\frac{1}{2\sqrt{\pi(\alpha+\beta)}}=\frac{2a}{\sqrt{\pi+a^2}}\approx\frac{2a}{\sqrt{\pi}}\quad\text{for } a \text{ sufficiently small.}$$

These asymptotic behaviors are illustrated in Fig. 11.

The probability amplitude $\rho(x,t)$ of the Schrödinger equation possesses another asymptotic behavior in addition to space-time dilation, namely asymptotic symmetrization of the initial data. For arbitrary initial data $\psi_0(x,0)$ in (1), with the imaginary part $Im\psi_0(x,0)=0$, $\rho(x,t)$ becomes symmetric about the origin as $t\to\infty$ no matter how asymmetric $\mathrm{Re}\psi_0(x,0)$ is at time 0. In [40] the following property of the solutions of (1) is proved:

$$|\rho(x,t)-\rho(-x,t)|\le\frac{1}{\pi t^2}\left(\int_{-\infty}^{\infty}y^2|\psi_0(y)|dy\right)\left(\int_{-\infty}^{\infty}|\psi_0(y)|dy\right).\qquad(10)$$

The solutions $\omega(x,z)$ of Eq. (9) are also asymptotically symmetric for arbitrary initial data, since they are asymptotic as $z\to\infty$ to Gaussian distributions with mean 0 (Appendix). An illustration of asymmetric initial data symmetrizing for both the Schrödinger equation and the ensemble equation is illustrated in Fig. 12, where the initial data is

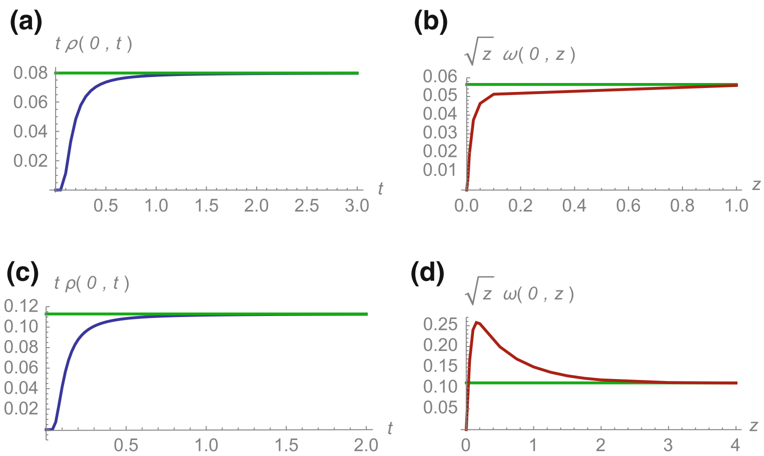


Fig. 11 The graphs of $t\rho(0, t)$ (blue) and $\sqrt{z}\omega(0, z)$ (red) for **a** one slit, $\lim_{t\to\infty} t\rho(0, t) \approx 0.798$; **b** one slit, $\lim_{z\to\infty} \sqrt{z}\omega(0, z) \approx 0.565$; **c** two slits, $\lim_{t\to\infty} t\rho(0, t) \approx 0.1128$. **d** two slits, $\lim_{z\to\infty} \sqrt{z}\omega(0, z) \approx 0.1125$ The parameter $a = 0.1$. The green lines are the limiting values

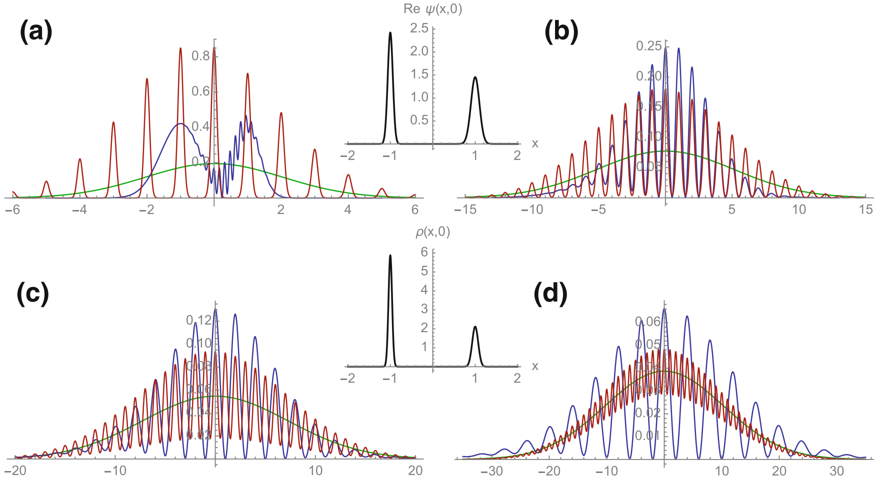


Fig. 12 The graphs of the solution $\rho(x, t)$ of (1) (blue) and the solution $\omega(x, z)$ (red) of (9) for **a** $t = z = 0.05$; **b** $t = z = 1.0/\pi$; **c** $t = z = 2.0/\pi$; **d** $t = z = 4.0/\pi$. The initial conditions are as in (11). The parameters are $\alpha = 1/(8\pi)$, $\beta = 1.0/(2.0a_1 \times 2.0a_2)$, $a_1 = 0.1$, $a_2 = 0.06$. The green graphs are the Gaussian distributions $\exp\left(-\frac{x^2}{2\sigma^2}\right)/\sqrt{2\pi}\sigma$ with $\sigma = \sqrt{2.0(\alpha + \beta)z}$

$$\psi(x,0) = \sqrt{\frac{a_1 a_2}{(a_1+a_2)\sqrt{\pi}}}\left(\frac{1}{a_1}\exp\left(-\frac{(x-1)^2}{2{a_1}^2}\right)+\frac{1}{a_2}\exp\left(-\frac{(x+1)^2}{2{a_2}^2}\right)\right),$$

$$\rho(x,0) = \frac{\exp\left(-\frac{(x-1)^2}{a_1^2}-\frac{(x+1)^2}{a_2^2}\right)}{a_1 a_2(a_1+a_2)\sqrt{\pi}}\left(a_1\exp\left(\frac{(x-1)^2}{2a_1^2}\right)+a_2\exp\left(\frac{(x+1)^2}{2a_2^2}\right)\right)^2, \tag{11}$$

$\omega(x,0) = \rho(x,0)$, and the probability amplitude is

$$\begin{aligned}\rho(x,t) = {} & \frac{a_1 a_2}{(a_1+a_2)\sqrt{\pi}}\left(\frac{\exp\left(-\frac{a_1^2(x-1)^2}{a_1^4+t^2}\right)}{\sqrt{a_1^4+t^2}}+\frac{\exp\left(-\frac{a_2^2(x+1)^2}{a_2^4+t^2}\right)}{\sqrt{a_2^4+t^2}}\right. \\ & + 2\frac{\exp\left(-\frac{a_1^2(x-1)^2}{2(a_1^4+t^2)}-\frac{a_2^2(x+1)^2}{2(a_2^4+t^2)}\right)}{(a_1^4+t^2)^{1/4}(a_2^4+t^2)^{1/4}}\cos\left[\frac{1}{2}\left(\frac{t(x-1)^2}{a_1^4+t^2}-\frac{t(x+1)^2}{a_2^4+t^2}\right.\right. \\ & \left.\left.\left. -\arctan\left(\frac{1}{a_1^2}\right)-\arctan\left(\frac{1}{a_2^2}\right)\right)\right].\end{aligned} \tag{12}$$

6 Discussion

Ensemble and statistical models of quantum mechanics have many formulations (reviews may be found in [1, 7, 9, 11, 33]). Nonlocal terms in models of quantum mechanics also have many formulations, such as Wigner phase-space distributions [34], dynamic and kinematic nonlocalities [36], and quantum balance Equations [11]. Diffusion terms as models of stochastic behavior in quantum mechanics have many formulations, as well [17, 28]. Such terms model quantum decoherence, which accounts for the dissipation of formed patterns generated by nonlocal terms and are relevant for quantum teleportation and quantum computing [15].

We have developed a model for the two-slit experiment based on a partial differential equation for the probability density function of ensemble particle behavior. The solutions of this equation align with the probability amplitude function obtained from the Schrödinger equation for this experiment in the formation of characteristic interference patterns. The equation is deterministic, but time irreversible. This equation contains a local diffusion term, which accounts for stochastic variation in the movement of particles. The equation contains a nonlocal term that accounts for the transverse movements of particles in the direction perpendicular to their forward motion. The term nonlocal is used in the convention of differential equations, namely, a rate of change dependent on a translated independent variable.

The interpretation of time in the nonlocal diffusion equation can be exchanged for the downstream distance from the plane of the slits. Thus, there is no ambiguity in the meaning of time for particles emitted randomly, separately, and independently one-at-a-time. Although the solutions of this equation align very well with the solutions obtained from the Schrödinger equation in the formed interference pattern, their behavior is very different before and after the interference pattern is established. The transition from the initial data is very simple for the nonlocal diffusion equation, but extremely complex for the Schrödinger equation. After the interference pattern is established, the Schrödinger equation solution preserves the pattern almost perfectly in a space-time dilation that propagates in the transverse direction with constant speed at each point x. The solution of the nonlocal diffusion equation, in contrast, disperses the interference pattern in the transverse direction by a typical diffusion process. The ensemble equation provides an aggregate behavioral model for the description of quantum particle interference phenomena.

Acknowledgements This work is dedicated to William Fitzgibbon and Yuri Kuznetsov in honor of their most valuable contributions to mathematical research and the community of mathematical researchers.

Appendix

Let $X = C[-\infty, \infty]$, the space of bounded uniformly continuous functions on $(-\infty, \infty)$ with norm $\|f\| = \sup_{-\infty<x<\infty} |f(x)|$. For $\sigma > 0$ let

$$(T_\sigma(t)f)(x) = \frac{1}{2\sqrt{\sigma t}} \int_{-\infty}^{\infty} \exp\left(- \frac{(x-y)^2}{4\sigma t} \right) f(y)\, dy, \quad f \in X,\ t > 0,\ -\infty < x < \infty. \tag{13}$$

Then $T_\sigma(t), t \geq 0$ is a strongly continuous holomorphic semigroup of positive linear operators in X with infinitesimal generator $(A_\sigma f)(x) = \sigma d^2 f(x)/dx^2$ satisfying $|T_\sigma(t)| \leq 1, t \geq 0$ [41, Chap. IX]. For $t > 0$, $A_\sigma T_\sigma(t)$ is bounded in X, and there exists $M_\sigma > 0$ such that $|A_\sigma T_\sigma(t)| \leq M_\sigma/t, t > 0$ [16, Part 2]. Further, $(T_\sigma(t)f)(x)$ is the strong solution in X to the diffusion equation

$$\frac{\partial}{\partial x} u(x,t) = \sigma \frac{\partial^2}{\partial x^2} u(x,t), \quad u(x,0) = f(x),\ t > 0,\ -\infty < x < \infty,\ f \in X. \tag{14}$$

For $\beta > 0$ define the bounded linear operator B in X by

$$(Bf)(x) = \beta\left(f(x+1) - 2f(x) + f(x-1)\right), \quad f \in X,\ -\infty < x < \infty.$$

Define

$$(\exp(tB)f)(x) = \left(\sum_{n=0}^{\infty} \frac{(tB)^n}{n!} f \right)(x) = e^{-2t\beta} \sum_{n=0}^{\infty} \sum_{m=0}^{\infty} \frac{(t\beta)^{n+m}}{n!m!} f(x+m-n) \tag{15}$$

for $f \in X, t \geq 0, -\infty < x < \infty$. Then $\exp(tB), t \geq 0$, is a group of positive linear operators in X with infinitesimal generator B satisfying $|\exp(tB)| \leq 1, t \geq 0$ (see [41, p. 244]).

Theorem 1 *Let $\beta > 0,\ s = 1,\ f \in X \cap L^1_+(-\infty, \infty)$, and let $\int_{-\infty}^{\infty} f(x)dx = 1$. Then for $t \geq 0$,*

$$\int_{-\infty}^{\infty} (\exp(tB)f)(x)dx = \int_{-\infty}^{\infty} f(x)dx = 1. \tag{16}$$

If also, $x^2 f(x) \in L^1(-\infty, \infty)$, then the mean of $\exp(tB)f$ *= the mean of f and the variance of* $\exp(tB)f$ *= the variance of $f + 2t\beta$.*

Proof From (15), we obtain (16), since for $t \geq 0$

$$\begin{aligned}\int_{-\infty}^{\infty} (\exp(tB)f)(x)dx &= \int_{-\infty}^{\infty} e^{-2t\beta} \sum_{n=0}^{\infty}\sum_{m=0}^{\infty} \frac{(t\beta)^{n+m}}{n!m!} f(x+m-n)dx \\ &= e^{-2t\beta} \sum_{n=0}^{\infty}\sum_{m=0}^{\infty} \frac{(t\beta)^{n+m}}{n!m!} \int_{-\infty}^{\infty} f(x)dx = \int_{-\infty}^{\infty} f(x)dx.\end{aligned}$$

Let μ_f be the mean of f and $\mu_{\exp(tB)f}$ the mean of $\exp(tB)f$. Then

$$\begin{aligned}\mu_{\exp(tB)f} &= \int_{-\infty}^{\infty} x(\exp(tB)f)(x)dx \\ &= \int_{-\infty}^{\infty} e^{-2t\beta} \sum_{n=0}^{\infty}\sum_{m=0}^{\infty} \frac{(t\beta)^{n+m}}{n!m!} xf(x+m-n)dx \\ &= e^{-2t\beta} \sum_{n=0}^{\infty}\sum_{m=0}^{\infty} \frac{(t\beta)^{n+m}}{n!m!} \int_{-\infty}^{\infty} (x+m-n)f(x)dx \\ &= e^{-2t\beta} \sum_{n=0}^{\infty}\sum_{m=0}^{\infty} \frac{(t\beta)^{n+m}}{n!m!} \left(\int_{-\infty}^{\infty} xf(x)dx + \int_{-\infty}^{\infty} (m-n)f(x)dx \right) \\ &= \mu_f + e^{-2t\beta} \left(\sum_{n=0}^{\infty}\sum_{m=0}^{\infty} \frac{(t\beta)^{n+m}}{n!m!} m - \sum_{n=0}^{\infty}\sum_{m=0}^{\infty} \frac{(t\beta)^{n+m}}{n!m!} n \right) \\ &= \mu_f + e^{-2t\beta} \left(\sum_{n=0}^{\infty}\sum_{m=1}^{\infty} \frac{(t\beta)^{n+m}}{n!(m-1)!} - \sum_{n=1}^{\infty}\sum_{m=0}^{\infty} \frac{(t\beta)^{n+m}}{(n-1)!m!} \right) = \mu_f.\end{aligned} \tag{17}$$

Let ν_f be the variance of f and $\nu_{\exp(tB)f}$ the variance of $\exp(tB)f$. A calculation similar to (17) yields

$$\begin{aligned}
\nu_{\exp(tB)f} &= \int_{-\infty}^{\infty} x^2(\exp(tB)f)(x)dx \\
&= \int_{-\infty}^{\infty} e^{-2t\beta}\sum_{n=0}^{\infty}\sum_{m=0}^{\infty}\frac{(t\beta)^{n+m}}{n!m!}x^2 f(x+m-n)dx \\
&= e^{-2t\beta}\sum_{n=0}^{\infty}\sum_{m=0}^{\infty}\frac{(t\beta)^{n+m}}{n!m!}\int_{-\infty}^{\infty}(x+m-n)^2 f(x)dx \\
&= e^{-2t\beta}\sum_{n=0}^{\infty}\sum_{m=0}^{\infty}\frac{(t\beta)^{n+m}}{n!m!}\left(\int_{-\infty}^{\infty}x^2 f(x)dx + 2(m-n)\int_{-\infty}^{\infty}xf(x)dx + (m-n)^2\right) \\
&= \nu_f + e^{-2t\beta}\sum_{n=0}^{\infty}\sum_{m=0}^{\infty}\frac{(t\beta)^{n+m}}{n!m!}\left(2(m-n)\mu_f + (m-n)^2\right) \\
&= \nu_f + e^{-2t\beta}\sum_{n=0}^{\infty}\sum_{m=0}^{\infty}\frac{(t\beta)^{n+m}}{n!m!}(m^2-2mn+n^2) = \nu_f + 2t\beta, \qquad (18)
\end{aligned}$$

since

$$e^{-2t\beta}\sum_{n=0}^{\infty}\sum_{m=0}^{\infty}\frac{(t\beta)^{n+m}}{n!m!}m^2 = \beta t(1+\beta t), \quad e^{-2t\beta}\sum_{n=0}^{\infty}\sum_{m=0}^{\infty}\frac{(t\beta)^{n+m}}{n!m!}2mn = 2(\beta t)^2.$$

□

Theorem 2 *Let $\alpha > 0$, $\beta > 0$, and $s = 1$. For $\omega_0 \in X \cap L^1_+(-\infty,\infty)$ such that $\int_{-\infty}^{\infty}\omega_0(x)dx = 1$, the unique solution $\omega(x,z) = \omega(x,t)$ of (9) is*

$$(T_\alpha(t)\exp(tB)\omega_0)(x), \quad t > 0, \ -\infty < x < \infty,$$

(where we have identified $z = t$ in (9)). Further, there exists a constant C such that

$$|\omega(x,t) - (T_{\alpha+\beta}(t)\omega_0)(x)| \le \frac{C}{t}\|\omega_0\|, \quad t > M_\alpha, \ -\infty < x < \infty. \qquad (19)$$

Remark 1 If ω_0 also has compact support, then (19) implies that uniformly on bounded sets of $x \in (-\infty,\infty)$

$$\lim_{t\to\infty}\sqrt{t}\omega(x,t) = \frac{1}{2\sqrt{\alpha+\beta}}, \quad \text{since } \lim_{t\to\infty}\sqrt{t}(T_{\alpha+\beta}(t)\omega_0)(x) = \frac{1}{2\sqrt{\alpha+\beta}}.$$

Proof Since A_σ and B commute, $T_\sigma(t)$ and $\exp(tB)$ commute for $t \ge 0$, and $T_\sigma(t)\exp(tB)$, $t \ge 0$ is the semigroup of operators for the solutions of (9). Since $T_\alpha(t)$, $t \ge 0$ is a holomorphic semigroup in X, $T_\alpha(t)\exp(tB)\omega_0$ is the unique strong solution of (9) in X for $t > 0$, $\omega_0 \in X$. If $f \in X$ is analytic, then

$$f(x \pm 1) = f(x) + \sum_{n=1}^{\infty} \frac{(\pm 1)^n}{n!} f^{(n)}(x)$$

and

$$f(x+1) - 2f(x) + f(x-1) = f^{(2)}(x) + 2 \sum_{n=2}^{\infty} \frac{1}{(2n)!} f^{(2n)}(x).$$

For $g \in X$, $t > 0$, $T_\alpha(t)g$ is analytic, and so

$$BT_\alpha(t)g = \beta A_1 T_\alpha(t)g + 2\beta \sum_{n=2}^{\infty} \frac{1}{(2n)!} A_1^n T_\alpha(t)g$$

Then,

$$\begin{aligned} \frac{d}{dt} T_\alpha(t)\exp(tB)\omega_0 &= (A_\alpha + B)T_\alpha(t)\exp(tB)\omega_0 \\ &= (\alpha+\beta)A_1 T_\alpha(t)\exp(tB)\omega_0 + 2\beta \sum_{n=2}^{\infty} \frac{1}{(2n)!} A_1^n T_\alpha(t)\exp(tB)\omega_0. \end{aligned}$$

From [20, Chap. 9] the solution of this nonhomogeneous equation satisfies

$$\begin{aligned} &T_\alpha(t)\exp(tB)\omega_0 \\ &= T_{\alpha+\beta}(t)\omega_0 + 2\beta \int_0^t T_{\alpha+\beta}(s) \sum_{n=2}^{\infty} \frac{1}{(2n)!} A_1^n T_\alpha(t-s)\exp((t-s)B)\omega_0 ds \\ &= T_{\alpha+\beta}(t)\omega_0 + 2\beta \int_0^t T_\beta(s) \sum_{n=2}^{\infty} \frac{1}{(2n)!\alpha^n} A_\alpha^n T_\alpha(t)\exp((t-s)B)\omega_0 ds. \end{aligned}$$

Thus, (19) follows, since for $t \geq M_\alpha$,

$$\begin{aligned} \|T_\alpha(t)\exp(tB)\omega_0 - T_{\alpha+\beta}(t)\omega_0\| &\leq 2\beta t \sum_{n=2}^{\infty} \frac{(M_\alpha/(\alpha t))^n}{(2n!)} \|\omega_0\| \\ &= 2\beta t \left(\cosh(\sqrt{M_\alpha/(\alpha t)}) - 1 - M_\alpha/(2\alpha t)\right) \|\omega_0\| \end{aligned}$$

and by L'Hospital's Rule

$$\lim_{t\to\infty} \frac{\cosh(\sqrt{M_\alpha/(\alpha t)}) - 1 - M_\alpha/(2\alpha t)}{(M_\alpha/(\alpha t))^2} = \frac{1}{4!}.$$

□

References

1. Allahverdyan AE, Balian R, Nieuwenhuizen TM (2013) Understanding quantum measurement from the solution of dynamical models. Phys Rep 525(1):1–166
2. Armstrong NJ, Painter KJ, Sherratt JA (2006) A continuum approach to modelling cell-cell adhesion. J Theoret Biol 243(1):98–113
3. Arndt M, Nairz O, Vos-Andreae J, Keller C, van der Zouw G, Zeilinger A (1999) Wave-particle duality of C_{60} molecules. Nature 401:680–682
4. Ballentine LE (1970) The statistical interpretation of quantum mechanics. Rev Modern Phys 42(4):358–381
5. Ballentine LE (1998) Quantum mechanics: a modern development. World Scientific, River Edge, NJ
6. Barrachina RO, Frémont F, Fossez K, Gruyer D, Helaine V, Lepailleur A, Leredde A, Maclot S, Scamps G, Chesnel J-Y (2010) Linewidth oscillations in a nanometer-size double-slit interference experiment with single electrons. Phys Rev A 81:060702
7. Bassi A, Ghirardi G (2003) Dynamical reduction models. Phys Rep 379(5–6):257–426
8. Born M (1926) Quantenmechanik der Stoßvorgänge (Quantum mechanics of collision). Z Phys A 38(11–12):803–827
9. Cetto AM, de la Peña L, Valdés-Hernández A (2015) Specificity of the Schrödinger equation. Quantum Stud Math Found 2(3):275–287
10. Cresser JD (2011) Quantum physics notes. Macquarie University. http://physics.mq.edu.au/~jcresser/Phys304/Handouts/QuantumPhysicsNotes.pdf
11. de la Peña L, Cetto AM, Valdés Hernández A (2015) The emerging quantum: the physics behind quantum mechanics, chapter The phenomenological stochastic approach: A short route to quantum mechanics, pp 33–66. Springer, Cham
12. Dyson J, Gourley SA, Villella-Bressan R, Webb GF (2010) Existence and asymptotic properties of solutions of a nonlocal evolution equation modeling cell-cell adhesion. SIAM J Math Anal 42(4):1784–1804
13. Einstein A, Podolsky B, Rosen N (1935) Can quantum-mechanical description of physical reality be considered complete? Phys Rev 41(10):777–780
14. Feynman RP, Leighton RB, Sands M (1965) The Feynman lectures on physics vol 3: Quantum mechanics. Addison-Wesley, Reading, MA
15. Ford GW, O'Connell RF (2002) Wave packet spreading: Temperature and squeezing effects with applications to quantum measurement and decoherence. Amer J Phys 70(3):319–324
16. Friedman A (1969) Partial differential equations. Holt, Rinehart and Winston, New York
17. Gisin N, Percival IC (1993) The quantum state diffusion picture of physical processes. J Phys A 26(9):2245–2260
18. Goldstein JA (1985) Semigroups of linear operators and applications. Oxford University Press, Oxford
19. Hongo H, Miyamoto Y, Furuya K, Suhara M (1997) A 40-nm-pitch double-slit experiment of hot electrons in a semiconductor under a magnetic field. Appl Phys Lett 70(1):93–95
20. Kato T (1966) Perturbation theory for linear operators. Springer, New York
21. Kocsis S, Braverman B, Ravets S, Stevens MJ, Mirin RP, Shalm LK, Steinberg AM (2011) Observing the average trajectories of single photons in a two-slit interferometer. Science 332(6034):1170–1173
22. Landé A (1965) New foundations of quantum mechanics. Cambridge University Press, New York
23. Landé A (1973) Quantum mechanics in a new key. Exposition Press, New York
24. Marcella TV (2002) Quantum interference with slits. Eur J Phys 23(6):615–621
25. McClendon M, Rabitz H (1988) Numerical simulations in stochastic mechanics. Phys Rev A 37(9):3479–3492
26. Nairz O, Arndt M, Zeilinger A (2003) Quantum interference experiments with large molecules. Amer J Phys 71(4):319–325

27. Painter KJ, Bloomfield JM, Sherratt JA, Gerisch A (2015) A nonlocal model for contact attraction and repulsion in heterogeneous cell populations. Bull Math Biol 77(6):1132–1165
28. Percival I (1998) Quantum state diffusion. Cambridge University Press, Cambridge
29. Popper KR (1967) Quantum mechanics without “the observer”. In: Bunge M (ed) Quantum Theory and Reality. pp. Springer, Berlin, pp 7–44
30. Przibram K (ed) (1967) Letters on wave mechanics. Philosophical Library, New York
31. Schilpp PA (ed) (1970) Albert Einstein: philosopher-scientist. Cambridge University Press, London
32. Sherratt JA, Gourley SA, Armstrong NJ, Painter KJ (2009) Boundedness of solutions of a nonlocal reaction-diffusion model for adhesion in cell aggregation and cancer invasion. European J Appl Math 20(1):123–144
33. Skála J, Čižek L, Kapsa V (2011) Quantum mechanics as applied mathematical statistics. Ann Phys 326(5):1174–1188
34. Styer DF, Balkin MS, Becker KM, Burns MR, Dudley CE, Forth ST, Gaumer JS, Kramer MA, Oertel DC, Park LH, Rinkoski MT, Smith CT, Wotherspoon TD (2002) Nine formulations of quantum mechanics. Amer J Phys 70(3):288–296
35. Thaller B (2000) Visual quantum mechanics: selected topics with computer-generated animations of quantum-mechanical phenomena. Springer, New York
36. Tollaksen J, Aharonov Y, Casher A, Kaufherr T, Nussinov S (2010) Quantum interference experiments, modular variables and weak measurements. New J Phys 12:013023
37. Tonomura A (1998) The quantum world unveiled by electron waves. World Scientific, Singapore
38. Tonomura A (2005) Direct observation of thitherto unobservable quantum phenomena by using electrons. PNAS 102(42):14952–14959
39. Tonomura A, Endo J, Matsuda T, Kawasaki T, Ezawa H (1989) Demonstration of single-electron buildup of an interference pattern. Amer J Phys 57(2):117–120
40. Webb GF (2011) Event based interpretation of Schrödinger’s equation for the two-slit experiment. Int J Theor Phys 50:3571–3601
41. Yosida K (1968) Functional analysis, 2nd edn. Springer, New York

Correction to: Remarks About Spatially Structured SI Model Systems with Cross Diffusion

Verónica Anaya, Mostafa Bendahmane, Michel Langlais
and Mauricio Sepúlveda

Correction to:
Chapter "Remarks About Spatially Structured SI Model Systems with Cross Diffusion" in: B. N. Chetverushkin et al. (eds.), *Contributions to Partial Differential Equations and Applications*, Computational Methods in Applied Sciences 47, https://doi.org/10.1007/978-3-319-78325-3_5

The book was inadvertently published with chapter author's incorrect given name. This information has been updated from "Vanaya Anaya" to "Verónica Anaya" in the initially published version of chapter "Remarks About Spatially Structured SI Model Systems with Cross Diffusion".

The updated version of this chapter can be found at
https://doi.org/10.1007/978-3-319-78325-3_5

B. N. Chetverushkin et al. (eds.), *Contributions to Partial Differential Equations and Applications*, Computational Methods in Applied Sciences 47,
https://doi.org/10.1007/978-3-319-78325-3_24