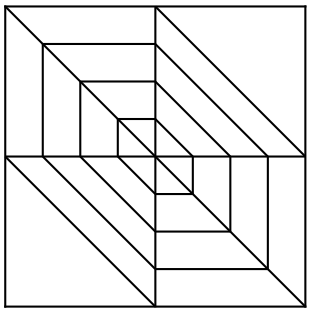
Proceedings of the International Conference

Applications of Mathematics 2015

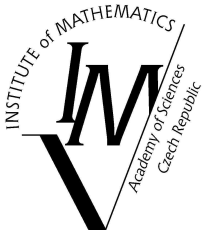
Prague, November 18–21, 2015

In honor of the birthday anniversaries of
Ivo Babuška (90), Milan Práger (85), and Emil Vitásek (85)

Edited by
J. Brandts, S. Korotov, M. Křížek,
K. Segeth, J. Šístek, T. Vejchodský



Institute of Mathematics
Czech Academy of Sciences
Prague 2015



ISBN 978-80-85823-65-3
Institute of Mathematics
Czech Academy of Sciences
Prague 2015
LaTeX typesetting prepared by Hana Bílková

Conference Applications of Mathematics 2015, in honor of the birthday anniversaries of Ivo Babuška (90), Milan Práger (85), and Emil Vitásek (85)
J. Brandts, S. Korotov, M. Křížek, K. Segeth, J. Šístek, T. Vejchodský (Eds.)
Institute of Mathematics CAS, Prague 2015

PREFACE

"Will you sign the blueprint?" is the favorite question of Professor Ivo Babuška, the question well known among the participants of many conferences and seminars. This question expresses the philosophy of Ivo Babuška's professional life. His experience tells him that it is not enough to model real processes on a computer but that it is extremely important to assess the quality of the result obtained. Only then you can sign the blueprint. Unfortunately, there are examples of blueprints signed without care of the reliability of the results, blueprints that caused fatal failures and huge damage in practice.

Ivo Babuška

Ivo Babuška, a mathematician recognized all over the world, was born on March 22, 1926 in Praha (Prague). One of the aims of the conference Applications of Mathematics 2015, organized by the Institute of Mathematics of the Czech Academy of Sciences, was to remember all the achievements Ivo Babuška has realized so far. We appreciate not only his theoretical results in the finite element method and in computational mathematics in general, but also his role of mentor of several dozens of PhD students and his position of a wise man who can predict the future of computational mathematics, present his visions to his colleagues, and successfully lead them to progress in this field.

Many papers have already been written about the life and work of Ivo Babuška. He is still very active in mathematics and publishes new results. Although his curriculum vitae has been published in various journals and proceedings many times, let me provide you with at least some principal biographical data and some of Babuška's outstanding mathematical results. More information can be found on the site `users.ices.utexas.edu/~babuska/` or in biographical papers.

Ivo Babuška studied civil engineering at the Czech Technical University in Prague, received his MS (Ing.) degree in 1949 and the PhD degree in Technical Science (Dr. tech.) in 1951. Then he studied mathematics at the Central Mathematical Institute in Prague as a graduate student of Professor V. Knichal. From 1951 he was a research fellow at the Institute. The Institute changed its name to the Mathematical Institute of the Czechoslovak Academy of Sciences in 1953 (now the Institute of Mathematics of the Czech Academy of Sciences).

In 1955 Ivo Babuška received the PhD (CSc.) degree in Mathematics and in 1960 the D.Sc. (DrSc.) degree which was in Czechoslovakia (as well as is now in the Czech Republic) awarded for the highest scientific achievements. From 1955 to 1968 he was the Head of Department of Constructive Methods of Mathematical Analysis of the Mathematical Institute. It was my privilege to work on my MS thesis at this Department during my studies at Charles University in Prague and to become a member of the Department in 1964. Later I also became Ivo Babuška's graduate student.

All Ivo Babuška's biographies mention his first important computational achievement in 1953–1956 when he was the leader of a numerical group that analyzed the technology of constructing the 91 meter high gravitational Orlík Dam on the Vltava River in Bohemia. The mathematical problem was to solve a nonlinear partial differential equation. Let me stress that all the computations were carried out by a team of people on mechanical desk calculators since no better devices were available in Czechoslovakia that time. The mathematical and numerical problems treated in the project provided many fruitful topics for investigation and initiated the research in a general theory of numerical stability of algorithms.

Ivo Babuška is the Honorary Editor of the journal Applications of Mathematics (formerly Aplikace matematiky) that he established in Prague in 1956. He was one of the founders of the EQUADIFF international scientific meetings that are still taking place. The first international EQUADIFF Conference on Differential

Equations was held in Prague in 1962. Later this series of conferences merged with another European series bearing the same name.

Ivo Babuška was appointed professor at Charles University in Prague in 1968. The same year he arrived in the United States and became a professor at the Institute for Physical Science and Technology and the Department of Mathematics of the University of Maryland at College Park. His interest in applied and numerical analysis brought him to the finite element method. He has achieved numerous bright results in the method itself, in its *hp*-version, in its reliability, a priori and a posteriori estimations, and adaptive approaches. These are recognized all over the world and belong to the fundamentals of the method. Moreover, Ivo Babuška has accomplished excellent results in several other branches of computational mathematics.

Ivo Babuška belongs among the founders of the Finite Element Circus, an informal meeting which, for more than 40 years, takes place twice a year.

From 1989, when the political situation in Czechoslovakia changed, he could resume visiting Prague. In 1994, he established the Prize for Young Czech Scientists in the field of numerical analysis and computational mechanics that is funded by his own means and awarded annually.

In 1995, Ivo Babuška became a senior research scientist and Robert Trull Professor at the Institute for Computational Engineering and Sciences at the University of Texas at Austin.

Along with his other activities, he has been involved in mentoring several dozens of graduate students, see `genealogy.math.uni-bielefeld.de/genealogy`. He is a member of editorial boards of numerous mathematical and engineering journals.

Ivo Babuška has received recognition and various awards for his scientific work. A brief supplement to the long list of his honors obtained before 2005 includes the following: Member of the U.S. National Academy of Engineering (2005), Member of the Academy of Medicine, Engineering, and Science of Texas (2005), Honorary Diploma of the Czech Society of Mechanics (2005), Honorary Medal De scientia et humanitate optime meritis, the highest award provided by the Czech Academy of Sciences (2006), Congress Medal of the 7th World Congress of Computational Mechanics in Los Angeles, International Association for Computational Mechanics (2006), Honorary Doctor of Science at the Czech Technical University in Prague (2007), Leroy P. Steele Prize for Lifetime Achievement, American Mathematical Society (2012), Neuron Award for Contribution to Science, Neuron Fund, Prague (2014).

Ivo Babuška's name is inseparably connected with the development of the finite element method. His theoretical results are widely used, directly or indirectly, in engineering practice. He has been invited for numerous lectures at conferences all over the world. The list of Ivo Babuška's monographs and papers in the Mathematical Reviews contains more than 300 items.

We must not omit a particular source of Ivo Babuška's scientific success, the family background provided by his wife Renata. They have a daughter and a son and four grandchildren.

* * *

To commemorate the significant life jubilees of Ivo Babuška, Milan Práger, and Emil Vitásek, the Institute of Mathematics of the Czech Academy of Sciences organized the International Conference *Applications of Mathematics 2015* on the premises of the Institute in Žitná 25, Prague 1 from November 18 to 21, 2015 (see website `am2015.math.cas.cz`).

The Scientific Committee consisted of

Mark Ainsworth (Brown University, Providence, RI, U.S.A.)

Jan Brandts (University of Amsterdam, the Netherlands)

Jan Chleboun (Czech Technical University, Prague, Czech Republic)

Miloslav Feistauer (Charles University, Prague, Czech Republic)

Jaroslav Haslinger (Charles University, Prague, Czech Republic)

Sergey Korotov (Basque Center for Applied Mathematics, Bilbao, Spain)

Qun Lin (Academy of Mathematics and System Science, Beijing, China)

Hans-Goerg Roos (Technical University, Dresden, Germany)

Theofanis Strouboulis (Texas A&M University, College Station, TX, U.S.A.)

Martin Stynes (National University of Ireland, Cork, Ireland)

Takuya Tsuchiya (Ehime University, Matsuyama, Japan)

Shuhua Zhang (Tianjin University of Finance and Economics, China)

Zhiming Zhang (Wayne State University, Detroit, MI, U.S.A.)

The Local Organizing Committee (Academy of Sciences) consisted of
Hana Bílková
Michal Křížek
Karel Segeth (Chair)
Jakub Šístek
Tomáš Vejchodský

The Organizing Committee is grateful to all authors for their contributions, to Project RVO 67985840 (Institute of Mathematics, Czech Academy of Sciences), and to Grant MTM2011-24766 (MICINN, Spain).

* * *

Ivo Babuška, Milan Práger, and Emil Vitásek deserve our congratulations and our sincere wishes of good health, optimistic mind, family happiness, and yet more scientific achievements.

Karel Segeth, on behalf of the Organizing Committee

Conference Applications of Mathematics 2015, in honor of the birthday anniversaries of Ivo Babuška (90), Milan Práger (85), and Emil Vitásek (85)
J. Brandts, S. Korotov, M. Křížek, K. Segeth, J. Šístek, T. Vejchodský (Eds.)
Institute of Mathematics CAS, Prague 2015

MY WONDERFUL NUMERICAL ANALYSIS TEACHERS — MILAN PRÁGER AND EMIL VITÁSEK

Michal Křížek

Institute of Mathematics, Academy of Sciences
Žitná 25, CZ – 115 67 Prague 1, Czech Republic
krizek@math.cas.cz

1. Numerical analysis at the Faculty of Mathematics and Physics

In 1970 I began to study mathematics at the Faculty of Mathematics and Physics at Charles University in Prague. In the third year, we had to choose one of the following specializations: algebra, mathematical analysis, applied mathematics, probability theory and statistics, topology, geometry, and numerical mathematics. My mother advised me at that time to choose numerical mathematics, since this was apparently a very new and modern discipline. I obeyed her suggestion, although I had absolutely no idea what this branch of science dealt with.

At the first numerical mathematics lecture Dr. Milan Práger showed us how to calculate the integral

$$I_n = \frac{1}{\mathrm{e}} \int_0^1 x^n \mathrm{e}^x \,\mathrm{d}x > 0. \tag{1}$$

First, using integration by parts, he derived the recurrence formula (cf. [14, p. 505])

$$I_n = 1 - nI_{n-1}, \quad n = 1, 2, \ldots, \tag{2}$$

and then he said that for simplicity we will evaluate the individual integrals only to three decimal places. Gradually he calculated on the blackboard the following values:

$$I_0 = 1 - \mathrm{e}^{-1} = 0.632, \qquad I_1 = 1 - 0.632 = 0.368, \qquad I_2 = 1 - 2 \cdot 0.368 = 0.264,$$
$$I_3 = 1 - 3 \cdot 0.264 = 0.208, \quad I_4 = 1 - 4 \cdot 0.208 = 0.168, \quad I_5 = 1 - 5 \cdot 0.168 = 0.16,$$
$$I_6 = 1 - 6 \cdot 0.16 = 0.04, \qquad I_7 = 1 - 7 \cdot 0.04 = 0.72.$$

Slowly I began to get bored and in my mind I wondered: *That, that is the modern mathematical discipline?* Then a big surprise came. Dr. Práger calculated

$$I_8 = 1 - 8 \cdot 0.72 = -4.76$$

and said: *Notice, dear students, that we have got a negative value, while the integral* (1) *is certainly positive. This is a completely unacceptable numerical result.* I immediately thought that the absurd negative number must be just a result of rounding errors, and I began to suspect what numerical analysis is about. At that time, of course, I had no idea about the instability of scheme (2) that was examined by Renata Babušková in her 1964 paper [5] (cf. also [1, p. 102]).

The above numerical phenomenon happens due to the fact that at each step we subtract two numbers of almost the same size. Then the difference contains only a few nonzero significant digits in computer arithmetic that necessarily leads to loss of accuracy. A very similar recurrence to (2) was examined by Muller [7].

I do also remember very well my first seminar on numerical mathematics. With Dr. Jitka Segethová we calculated the values of polynomials using Horner's scheme on large and heavy mechanical calculators that were powered by an electrical engine. Nevertheless, on the ground floor of our building on the Lesser Town Square there already was a big mainframe electronic computer Minsk 22. Here I used ALGOL 60 (Algorithmic Language) to program simple numerical algorithms that Dr. Práger taught us. Minsk 22 had 64 KB of memory, input via punched tape, and was very slow. Moreover, approximately every 20 minutes computer calculations crashed due to MACHINE ERROR. So basically it was not possible to perform any longer calculation. We learned also the machine code to speed up computations.

In the fourth year of my studies, the numerical mathematics was taught by Dr. Emil Vitásek. In fact, the recurrence (2) was invented by him (see [1]). He lectured by heart using no written notes and with great enthusiasm. His performance was truly wonderful, logically assembled, and understandable. He concentrated on solving partial differential equations by the finite difference method, which is a sort of forerunner of my favorite finite element method. In particular, I was charmed by the convergence proof of the finite difference method that he presented to us.

2. Department of Constructive Methods of Mathematical Analysis

During my military service in 1975–1976, I received a letter initiated by Dr. Práger, whether I wanted to start postgraduate studies at the Mathematical Institute of the Czechoslovak Academy of Sciences. Because I had not negotiated any further job after completing my military service, I agreed, and certainly at present I do not regret that decision. Therefore, in September 1976 I began postgraduate studies with Dr. Práger at the Department of Constructive Methods of Mathematical Analysis, where he was the Head during the period 1969–1994. His Deputy was Dr. Vitásek. The Department was located at the rear of the Opletalova street no. 45. Dr. Práger and Dr. Vitásek shared the front room, where also our Numerical Analysis Seminars

were held. I did not understand the first several lectures there and I have to admit that it took me quite a long time to follow the issues that were investigated in our Department. I started to read at that time the recent paper [12] on overimplicit multistep methods for ordinary differential equations written by M. Práger, J. Taufer, and E. Vitásek.

For my Candidate of Sciences examination I studied the classical 1966 monograph *Numerical processes in differential equations* [3] by Ivo Babuška, Milan Práger, and Emil Vitásek. It already contained the description of the finite element method for elliptic boundary value problems — my favorite topic. Contour lines of the standard piecewise linear finite element basis functions are illustrated in [3, p. 305]. This picture serves, in fact, as the LOGO of our Numerical Analysis Seminar and also of this Conference. Some other linear and bilinear finite element basis functions were already sketched in their previous book [2] published in the Czech language.

Both the monographs [2] and [3] begin with the recurrence (2). However, there are other nice and illustrative numerical examples — for instance, the investigation of numerical instability of successive performance of the following arithmetic operations

$$\ldots(((((1:2)\cdot 2):3)\cdot 3):4)\cdot 4\ldots$$

Various numerical results of this expression were obtained by Karel Segeth on different computers involving thousands of divisions and multiplications [3, p. 6]. I liked such examples very much. Later I wrote the article [6] jointly with M. Práger and E. Vitásek on the reliability of numerical computations. We systematically collected many other pathological examples, where the numerical solution behaves in an unpredictable manner. This resulted in another article [19] with Dr. Vitásek and I continue with this activity ever today. The main reason is that programmers should not always believe their computer outputs, in particular, if they are not familiar with numerical analysis topics like finite precision arithmetic, theory of rounding errors, ill-conditioned problems, and so on.

In our Department there has always been a great friendly and creative atmosphere. I can begin to describe what I have learned from my numerical analysis teachers in over 40 years. Dr. Milan Práger significantly contributed to the issue of numerical modelling in electrical engineering. Together we dealt with several real-life technical problems for the Research Institute VÚSE Běchovice. In particular, Dr. Práger numerically calculated the magnetic field inside large oil-immersed transformers. Then I used his results on the density of heat sources to calculate the temperature distribution in the magnetic core. With Dr. Vitásek I discussed the various theoretical aspects of numerical methods that I applied. He wrote a whole series of monographs (see [2], [3], [4], [16], [17]) devoted to numerical methods. They originated from our Department in Opletalova street without any personal computers and internet.

Allow me to finish this section by a funny story. A Vietnamese aspirant once visited our Department and was looking for Dr. Vitásek. Dr. Práger told him that

Milan Práger

Dr. Vitásek is lecturing in Italy and will return after three weeks. However, the Vietnamese aspirant did not understand well, he leaned in a large armchair and said: *Never mind, I'll wait.*

3. Milan Práger — Curriculum vitae

RNDr. Milan Práger, CSc., was born on April 21, 1930 in Prague. After grammar school in Smíchov in 1940–1948 he became a student of mathematics at the Faculty of Science of Charles University in Prague. His studies ended in 1952 when he passed the leaving State Examination. During the period 1952–1954 he worked at the Faculty of Mechanical Engineering of the Czech Technical University in Prague as an assistant in the Mathematical Department, and then became a postgraduate student of Ing. Dr. Ivo Babuška at the Mathematical Institute of the Czechoslovak Academy of Sciences in Prague (1954–1957). He received the scientific title Candidate of Sciences (CSc. $\approx$ PhD.) in the year 1959, and stayed to work at the Mathematical Institute as a researcher, senior researcher (1965), and chief researcher (1977). From 1971 to 1992 he was the Head of the Department of Constructive Methods of Mathematical Analysis. He retired in 1996, but still worked part-time at the Mathematical Institute until 2005. During this period he published several valuable papers, see e.g. [9], [10], [11]. He is still interested and participates in our regular Friday seminar *Current problems in numerical analysis.*

The main subject of scientific interest of Dr. Prager is the theory of numerical methods for solving differential equations. He has published about 40 original mathematical papers, conference articles, co-authored the 1964 monograph[1] *Numerical*

[1]The logo from the front page of this Proceedings was taken from [3, p. 305]. It shows the support of a linear finite element basis function with its contour lines.

Dr. Milan Práger lecturing at our Numerical Analysis Seminar

solutions of differential equations [2] and the 1966 monograph *Numerical processes in differential equations* [3], which was translated into Russian in 1969, see [4]. He also wrote a chapter in the world wide known Rektorys' *Survey of applicable mathematics* [14], which was published in two English and six Czech edition series. Another important accomplishment is his textbook *Numerical mathematics I* [8].

Dr. Práger participated in numerous domestic and international scientific meetings and research visits. Let us name, for instance, a few series of Equadiff Conferences. He was also a lecturer at the postgraduate course at the University of Zagreb, the Istituto per le Applicazioni del Calcolo in Rome, at Chalmers University of Technology in Gothenburg, at the Royal Institute of Technology in Stockholm and in many other places in former Czechoslovakia.

In addition, Dr. Práger was always intensely concentrated on educational activities and organization of scientific meetings. During the period 1967–1990 he taught fundamentals of numerical methods at the Faculty of Mathematics and Physics of Charles University in Prague. He is the author of lecture notes on this topic and translated with Dr. Emil Vitásek the comprehensive Ralston's guide *A first course of numerical analysis* [13]. Dr. Práger was the advisor of my Candidate of Sciences thesis *An equilibrium finite element method in three-dimensional elasticity* defended in 1980. Its co-advisor was his colleague Ivan Hlaváček. Dr. Práger led theses of other four scientific aspirants: Michal Kočvara, Stanislav Míka, Karel Višňák, Jan Vlček, and successfully trained several master students.

Milan Práger was a member of the final state examination committee at Charles

Emil Vitásek

University in Prague, a member of the committee for candidate and doctoral dissertations, a member of the National Commission on issues of information technology, and many others. He was also a co-organizer of more than ten years of popular summer school "Programs and Numerical Algorithms" traditionally held at various locations of Jizera Mountains.

It would take a long time to enumerate what Milan Práger has done for mathematics and for the Institute of Mathematics. Let us mention also his interests that go far beyond mathematics. For example, he has a deep knowledge about history, cartography and music, he likes to solve various puzzles and cross-words, he is a very good chess player and played for many years in the chess section of Prague universities.

4. Emil Vitásek — Curriculum vitae

RNDr. Emil Vitásek, CSc., was born on May 29, 1931 in České Budějovice. After high school in Přerov, where he graduated in 1950, he began to study mathematics at the Faculty of Science of Charles University, which in 1953 changed to the Faculty of Mathematics and Physics. After his studies he joined the Mathematical Institute of the Czechoslovak Academy of Sciences in 1954 as a research assistant in the department of Ing. Dr. Ivo Babuška. Under Dr. Babuška's leadership, he attained the Candidate of Sciences degree CSc. in 1960. Dr. Vitásek became a researcher and later a senior researcher. He is still actively working at the Institute of Mathematics (see e.g. [18]).

Dr. Emil Vitásek lecturing at our Numerical Analysis Seminar

His mathematical research is associated with the numerical solution of differential equations, in particular, numerical methods for time-dependent equations, i.e. ordinary and parabolic equations. There he employed his deep knowledge of mathematical and functional analysis. His first papers are associated with the calculations of the Dam Orlík on the Vltava river. Then he dedicated himself to the study of numerical stability. He was one of those who developed the theory of transfer boundary conditions for boundary value problems for ordinary differential equations. At the same time he dealt with problems associated with engineering practice. He published about 60 original scientific papers and held lecture courses in Croatia, Sweden, and Italy. He was invited to give plenary lectures at several national and international conferences.

Dr. Vitásek is a member of the Editorial Board of Applications of Mathematics since 1971. He is the author of a chapter in the *Survey of applied mathematics*. He contributed three chapters to its last edition [14] and was the Associate Editor of its two volumes. He is also a co-author of the book *Numerical solutions of differential*

equations (1964), which has been revised and expanded to the English version Numerical processes in differential equations (1966) and was published in 1969 and in Russian translation (see [2], [3], [4]).

We should also mention the long-term pedagogical activity of Dr. Vitásek. He lectured on Numerical mathematics at the Faculty of Mathematics and Physics for more than 20 years, and then at the University of West Bohemia in Pilsen. He was the advisor of several Master students and four PhD students: Ľubor Malina, Hassan Nasr, Jan Šafář, Jiří Taufer. His fifth student Marian Brezina defended PhD in USA. In connection with these activities two of his monographs in the field of numerical mathematics appeared: *Numerical methods* [16] and *Foundations of the theory of numerical methods for solving differential equations* [17]. Another important accomplishment are his three textbooks. The first one *Numerical mathematics II — Numerical solution of differential equations* [15] was published by Charles University. The other two were published by the University of West Bohemia: *Selected chapters from the theory of numerical methods for the solution of differential equations* and *Introduction to the theory of generalized functions* that discusses the foundations of the theory of distributions. He also translated with Dr. Práger the famous monograph by A. Ralston: *A first course in numerical analysis* [13]. Dr. Vitásek was a member of the board of examiners for the Final State Exams and the board for the Rigorous Exams.

Emil Vitásek is a researcher with wide interests connected mainly with technical problems. He has a deep knowledge of aviation, but also of modern history and literature. He won the Czechoslovak national championship in correspondence chess. Anyone who comes to him with any problem, whether mathematical or generally human, finds that he is always a patient and attentive listener. Finally, we would like to mention his continuous aversion against the communist regime.

5. Felicitations

We all wish to Dr. Milan Práger and Dr. Emil Vitásek to their jubilees a good health and great satisfaction for a number of happy years.

Acknowledgement

The author would like to thank Jan Brandts and Lawrence Somer for their help with preparation of this paper and the journal Pokroky Mat. Fyz. Astronom. for providing certain materials. The paper was supported by Project RVO 67985840.

References

[1] Babuška, I., Práger, M., and Vitásek, E.: Numerische Stabilität der Rechenprozessen. Wiss. Z. der Tech. Univ. Dresden **12** (1963), H. 1, 101–110.

[2] Babuška, I., Práger, M., and Vitásek, E.: *Numerické řešení diferenciálních rovnic*. SNTL, Praha, 1964.

[3] Babuška, I., Práger, M., and Vitásek, E.: *Numerical processes in differential equations*. John Willey & Sons, London, New York, Sydney, 1966.

[4] Babuška, I., Práger, M., and Vitásek, E.: *Čislennye processy rešenia differencial'nych uravněnij*. Mir, Moscow, 1969.

[5] Babušková, R.: Über numerische Stabilität einiger Rekursionsformeln. Apl. Mat. **9** (1964), 186–193.

[6] Křížek, M., Práger, M., Vitásek, E.: Spolehlivost numerických výpočtů. Pokroky Mat. Fyz. Astronom. **42** (1997), 8–23.

[7] Muller J.-M. et al.: *Handbook of floating-point arithmetic*. Birkhäuser, 2009.

[8] Práger, M., *Numerická matematika I (skripta)*. SPN, Praha, 1981, 217 pp.

[9] Práger, M.: *Eigenvalues and eigenfunctions of the Laplace operator on an equilateral triangle for the discrete case*. Appl. Math. **46** (2001), 231–239.

[10] Práger, M.: On a construction of fast direct solvers. Appl. Math. **48** (2003), 225–236.

[11] Práger, M., Sýkorová, I., Jak počítače počítají. Pokroky Mat. Fyz. Astronom. **49** (2004), 32–45.

[12] Práger, M., Taufer, J., Vitásek, E.: Overimplicit multistep methods. Apl. Mat. **18** (1973), 399–421.

[13] Ralston, A.: *A first course in numerical analysis*. McGraw-Hill, 1965; Czech translation: *Základy numerické matematiky*. Academia, Prague, 1978.

[14] Rektorys, K.: *Survey of applicable mathematics, Vol. I and II*. Kluwer Acad. Publ., Dordrecht, 1994.

[15] Vitásek, E., *Numerická matematika II — Numerické řešení diferenciálních rovnic (skripta)*. SPN, Praha, 1981, 236 pp.

[16] Vitásek, E.: *Numerické metody*. SNTL, Praha, 1987.

[17] Vitásek, E.: *Základy teorie numerických metod pro řešení diferenciálních rovnic*. Academia, Praha, 1994.

[18] Vitásek, E.: *Approximate solution of an inhomogeneous abstract differential equation*. Appl. Math. **57** (2012), 31–41.

[19] Vitásek, E., Křížek, M.: *(Ne)spolehlivost numerických výpočtů. Jaká nebezpečí skrývá numerické počítání?* Sborník semináře: Programy a algoritmy numerické matematiky 9, Kořenov, MÚ AV ČR Praha, 1998, 139–150.

Contents

Conference Applications of Mathematics 2015, in honor of the birthday anniversaries of Ivo Babuška (90), Milan Práger (85), and Emil Vitásek (85)
J. Brandts, S. Korotov, M. Křížek, K. Segeth, J. Šístek, T. Vejchodský (Eds.)
Institute of Mathematics CAS, Prague 2015

DIFFERENTIAL ALGEBRAIC EQUATIONS OF FILIPPOV TYPE

Martin Biák, Drahoslava Janovská

University of Chemical Technology, Prague
Technická 5, 166 28 Prague 6 Dejvice, Czech Republic
biakm.mobil@gmail.com, janovskd@vscht.cz

Abstract: We will study discontinuous dynamical systems of Filippov-type. Mathematically, Filippov-type systems are defined as a set of first-order differential equations with discontinuous right-hand side. These systems arise in various applications, e.g. in control theory (so called relay feedback systems), in chemical engineering (an ideal gas–liquid system), or in biology (predator-prey models). We will show the way how to extend these models by a set of algebraic equations and then study the resulting system of differential-algebraic equations. All MATLAB simulations are performed in modified version of the program developed by Petri T. Piiroinen and Yuri A. Kuznetsov published in ACM Trans. Math. Software, 2008.

Keywords: Filippov systems, differential algebraic equations (DAEs), Filippov systems with DAEs, soft drink process

MSC: 34C60, 37N30, 65P30

1. Introduction

There are a variety of engineering problems involving dynamical systems. In recent years, the need to describe systems with a discontinuity in the state variables has emerged. The theory of the non-smooth systems has been introduced and thoroughly studied in [9]. From recent years, let us mention the book [5].

In addition to dynamical systems described by ordinary differential equations there are also models that require the use of differential equations along with algebraic ones. These are so-called differential algebraic equations (DAEs).

From the dynamical point of view, the essential differences between differential-algebraic equations (DAEs) and explicit ordinary differential equations (ODEs) arise in so-called singular problems, which lead to new dynamic phenomena such as those displayed at impasse points or singularity-induced bifurcations.

The origins of DAEs theory can be traced back to the work of K. Weierstrass and L. Kronecker on parameterized families of bilinear forms [20, 14]. In terms of

matrices, pencils were applied to the analysis of linear systems of ordinary differential equations with a possibly singular leading coefficient matrix by F. R. Gantmacher [10, 11]. Another milestone is the work of P. Dirac on generalized Hamiltonian systems [6, 7, 8]. The key ideas supporting what nowadays is known as the differentiation index of a semi-explicit DAEs can be found in these references. The work of Dirac was mainly motivated by applications in mechanics. A large amount of research on differential-algebraic equations has also been motivated by applications in circuit theory. The differential-algebraic form of circuit equations is naturally due to the combination of differential equations coming from reactive elements with algebraic (non-differential) relations modeling Kirchhoff laws and device characteristics.

To "measure" how difficult is to solve a DAEs system, the concept of indices has been introduced. There are different indices (Kronecker index, strangeness index, differentiation index, perturbation index, etc.), and the choice of the index depends on the DAEs and on the application, for which it is used (see [13, 19]).

If the model with DAEs features a discontinuity, then we have to modify the non-smooth dynamical systems theory to include DAEs. We will extend the theory of the non-smooth systems, namely the theory of Filippov systems, to the systems with DAEs. Finally, we will apply this theory to some application from chemical engineering.

2. Filippov systems

Let φ be a continuous and differentiable scalar function, $\varphi : \mathcal{D} \subseteq \mathbb{R}^n \to \mathbb{R}$, $n \geq 2$. The function φ divides the region $\mathcal{D}$ into three parts:

$$S_1 = \{\mathbf{x} \in \mathcal{D} \subseteq \mathbb{R}^n : \varphi(\mathbf{x}) > 0\},$$

$$S_2 = \{\mathbf{x} \in \mathcal{D} \subseteq \mathbb{R}^n : \varphi(\mathbf{x}) < 0\},$$

$$\Sigma = \{\mathbf{x} \in \mathcal{D} \subseteq \mathbb{R}^n : \varphi(\mathbf{x}) = 0\}.$$

Let us assume that the function φ has a non-vanishing gradient $\nabla\varphi$ on the boundary Σ. We define the Filippov system $\mathcal{F}$ on $\mathcal{D} = S_1 \cup S_2 \cup \Sigma$ as

$$\mathcal{F} : \dot{x} = \begin{cases} \mathbf{f}^{(1)}(\mathbf{x}) \ , & \mathbf{x} \in S_1 \, , \\ \mathbf{f}^{(0)}(\mathbf{x}) \ , & \mathbf{x} \in \Sigma \, , \\ \mathbf{f}^{(2)}(\mathbf{x}) \ , & \mathbf{x} \in S_2 \, , \end{cases} \tag{1}$$

where $\mathbf{x}(t) \in \mathbb{R}^n$, $\mathbf{f}^{(i)} : \mathbb{R}^n \to \mathbb{R}^n$, $i = 0, 1, 2$, are sufficiently smooth functions in all arguments, and $t \in \mathbb{R}$. We suppose that the state space $\mathcal{D} = S_1 \cup S_2 \cup \Sigma$, $\mathcal{D} \subset \mathbb{R}^n$, the vector fields $\mathbf{f}^{(1)}$ on S_1 and $\mathbf{f}^{(2)}$ on S_2 are given.

We have to define the vector field $\mathbf{f}^{(0)}$ that determines the behavior of the system (1) on the boundary Σ. There are several possible scenarios that occur if the trajectory with an initial condition $\mathbf{x}_0 \notin \Sigma$ reaches the boundary Σ. Let for example $\mathbf{x}_0 \in S_1$. The trajectory can cross the boundary from S_1 to S_2, turn back to S_1,

or it can even slide along the boundary Σ. The direction in which the trajectory continues after a contact with Σ is affected by both vector fields $\mathbf{f}^{(1)}$ and $\mathbf{f}^{(2)}$.

Let us define a scalar function $\sigma(\mathbf{x})$, $\mathbf{x} \in \Sigma$, as the product of dot products in $\mathbb{R}^n$

$$\sigma(\mathbf{x}) = \langle \nabla\varphi(\mathbf{x}), \mathbf{f}^{(1)}(\mathbf{x}) \rangle \cdot \langle \nabla\varphi(\mathbf{x}), \mathbf{f}^{(2)}(\mathbf{x}) \rangle. \tag{2}$$

The sign of the function $\sigma(\mathbf{x})$ determines the behavior of the trajectory after a contact with the boundary Σ. Let us use this sign as a criterion for the identification of two types of sets on the boundary Σ, a crossing set Σ_c and a sliding set Σ_s,

$$\Sigma_c \subseteq \Sigma = \{\mathbf{x} \in \Sigma : \sigma(\mathbf{x}) > 0\},$$

$$\Sigma_s \subseteq \Sigma = \{\mathbf{x} \in \Sigma : \sigma(\mathbf{x}) \leq 0\}.$$

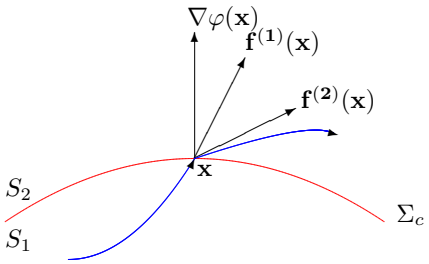
The vector field $f^{(0)}$ on the boundary Σ is defined as follows:

• on Σ_c,

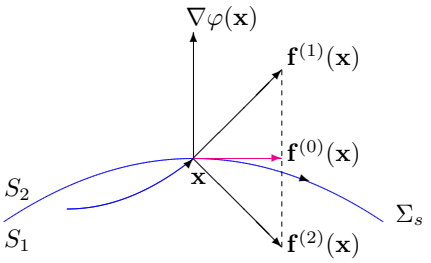
$$\mathbf{f}^{(0)} = \frac{1}{2}\left(\mathbf{f}^{(1)} + \mathbf{f}^{(2)}\right), \tag{3}$$

• on Σ_s, the vector field $f^{(0)}$ is defined as a convex combination

$$\mathbf{f}^{(0)} = (1-\lambda)\,\mathbf{f}^{(1)} + \lambda\,\mathbf{f}^{(2)}, \quad \lambda = \frac{\langle \nabla\varphi, \mathbf{f}^{(1)} \rangle}{\langle \nabla\varphi, \mathbf{f}^{(1)} - \mathbf{f}^{(2)} \rangle}, \quad 0 \leq \lambda \leq 1\,. \tag{4}$$



$\Sigma_c = \{\mathbf{x} \in \Sigma : \sigma(\mathbf{x}) > 0\}$



$\Sigma_s = \{\mathbf{x} \in \Sigma : \sigma(\mathbf{x}) \leq 0\}$

Let us note that Σ_c contains those points $\mathbf{x} \in \Sigma$ in which both vectors $\mathbf{f}^{(1)}(\mathbf{x})$ and $\mathbf{f}^{(2)}(\mathbf{x})$ head to the same region. The set $\Sigma_s = \{\mathbf{x} \in \Sigma : \sigma(\mathbf{x}) \leq 0\}$ contains those points $\mathbf{x} \in \Sigma$ in which all other cases of configuration occur.

The equation (4) is called the Filippov convex combination. Let us note that it is not the only possibility how to define the vector field on the boundary Σ. Another possibility is for example to apply the so-called Utkin's equivalent control method, see e.g. [5].

Remark 2.1 Formula (4) follows from the fact that the trajectory slides along the sliding set, i.e., the vector field $\mathbf{f}^{(0)}(\mathbf{x})$ must be tangent to Σ_s,

$$\langle \nabla\varphi(\mathbf{x}), \mathbf{f}^{(0)}(\mathbf{x}) \rangle = 0, \quad \forall \mathbf{x} \in \Sigma_s. \tag{5}$$

On the sliding boundary Σ_s special points, so called sliding points, can be detected. Let us classify some of them.

- Singular sliding point is a point $\mathbf{x} \in \Sigma_s$ such that

$$\langle \nabla\varphi(\mathbf{x}), \mathbf{f}^{(1)}(\mathbf{x})\rangle = 0 \quad \text{and also} \quad \langle \nabla\varphi(\mathbf{x}), \mathbf{f}^{(2)}(\mathbf{x})\rangle = 0\,.$$

  At these points, both vectors $\mathbf{f}^{(1)}(\mathbf{x})$ and $\mathbf{f}^{(2)}(\mathbf{x})$ are tangent to Σ_s.

- The point $\mathbf{x} \in \Sigma_s$ is a generic pseudo-equilibrium if

$$\mathbf{f}^{(0)}(\mathbf{x}) = \mathbf{0}\,, \quad \mathbf{f}^{(1)}(\mathbf{x}) \neq \mathbf{0}\,, \quad \mathbf{f}^{(2)}(\mathbf{x}) \neq \mathbf{0}\,.$$

  At these points, the vectors $\mathbf{f}^{(1)}(\mathbf{x})$ and $\mathbf{f}^{(2)}(\mathbf{x})$ are anti-collinear.

- In a boundary equilibrium $\mathbf{x} \in \Sigma_s$, one of the vectors $\mathbf{f}^{(i)}(\mathbf{x})$ vanishes,

$$\mathbf{f}^{(1)}(\mathbf{x}) = \mathbf{0} \quad \text{or} \quad \mathbf{f}^{(2)}(\mathbf{x}) = \mathbf{0}\,.$$

- The point $\mathbf{x} \in \Sigma_s$ is a tangent point if both $\mathbf{f}^{(1)}(\mathbf{x}) \neq \mathbf{0}\,, \mathbf{f}^{(2)}(\mathbf{x}) \neq \mathbf{0}$ and

$$\langle \nabla\varphi(\mathbf{x}),\ \mathbf{f}^{(1)}(\mathbf{x})\rangle = 0 \quad \text{or} \quad \langle \nabla\varphi(\mathbf{x}),\ \mathbf{f}^{(2)}(\mathbf{x})\rangle = 0\,.$$

  In this case, both vectors $\mathbf{f}^{(1)}(\mathbf{x}),\ \mathbf{f}^{(2)}(\mathbf{x})$ are nonzero, but one of them is tangent to Σ. The tangent point terminates Σ_s in Σ, i.e., the sliding set Σ_s can be delimited solely by computing all tangent points.

3. Filippov systems with DAEs

Differential algebraic equations have become a widely accepted tool for the modeling and simulation of constrained dynamical systems in numerous applications, such as mechanical multibody systems, electrical circuit simulation, chemical engineering, control theory, fluid dynamics, and many other areas.

Let us have a general nonlinear system of differential-algebraic equations

$$\mathbf{F}(t, \mathbf{z}, \dot{\mathbf{z}}) = 0, \tag{6}$$

where $\mathbf{F} : I \times U \times V \to \mathbb{R}^n$, $t \in I$, $\mathbf{z}(t) \in U$, $\dot{\mathbf{z}}(t) \in V$, $\mathbf{z} : I \to \mathbb{R}^n$ is an unknown function, $\mathbf{z} \in \mathcal{C}^1(I, \mathbb{R}^n)$, $I \subseteq \mathbb{R}$ is a compact interval, $U, V \subseteq \mathbb{R}^n$ are open regions.

Let the equation (6) be equipped with the initial condition

$$\mathbf{z}(t_0) = \mathbf{z}_0,\ t_0 \in I,\ \mathbf{z}_0 \in \mathbb{R}^n. \tag{7}$$

Definition 3.1 Let the system of differential-algebraic equations (6), (7) be uniquely solvable. We define the so-called derivative array equations as

$$\mathbf{F}_\ell(t, \mathbf{z}, \dot{\mathbf{z}}, \ldots, \mathbf{z}^{(\ell+1)}) := \begin{bmatrix} \mathbf{F}(t, \mathbf{z}, \dot{\mathbf{z}}) \\ \frac{\mathrm{d}}{\mathrm{d}\,t}\mathbf{F}(t, \mathbf{z}, \dot{\mathbf{z}}) \\ \vdots \\ \frac{\mathrm{d}^{\,\ell}}{\mathrm{d}\,t^\ell}\mathbf{F}(t, \mathbf{z}, \dot{\mathbf{z}}) \end{bmatrix}, \tag{8}$$

where we can expand the term $\frac{\mathrm{d}}{\mathrm{d}\,t}\mathbf{F}(t, \mathbf{z}, \dot{\mathbf{z}})$ using the chain rule:

$$\frac{\mathrm{d}}{\mathrm{d}\,t}\mathbf{F}(t, \mathbf{z}, \dot{\mathbf{z}}) = \mathbf{F}_t(t, \mathbf{z}, \dot{\mathbf{z}}) + \mathbf{F}_{\mathbf{z}}(t, \mathbf{z}, \dot{\mathbf{z}})\dot{\mathbf{z}} + \mathbf{F}_{\dot{\mathbf{z}}}(t, \mathbf{z}, \dot{\mathbf{z}})\ddot{\mathbf{z}}.$$

Other terms can be treated similarly.

In derivative array equations (8), let us formally replace $\dot{\mathbf{z}}(t)$ by $\mathbf{v}(t) \in \mathbb{R}^n$ and $(\ddot{\mathbf{z}}(t), \ldots, \mathbf{z}^{(\ell+1)}(t))$ by $\mathbf{w}(t) \in W$, $W \subseteq \mathbb{R}^{\ell n}$. In this setting, a given $(t, \mathbf{z})$ is said to be consistent if there exists a $(t, \mathbf{z}, \mathbf{v}, \mathbf{w}) \in I \times U \times V \times W$ for which $\mathbf{F}_\ell(t, \mathbf{z}, \mathbf{v}, \mathbf{w}) = 0$.

Definition 3.2 The smallest number $\nu \in \mathbb{N}_0$ for which $\mathbf{F}_\nu(t, \mathbf{z}, \mathbf{v}, \mathbf{w}) = 0$ holds for every consistent $(t, \mathbf{z})$, is called the differentiation index of (6).

The idea behind the differentiation index framework is, roughly speaking, to define the index of (6) as the number of differentiations needed to write $\dot{\mathbf{z}}$ in terms of (t, $\mathbf{z}$). Further details can be found in [13] or in [19].

In many technical applications a very common form of DAEs is the so called semi-explicit DAEs that provides a significant simplification of the fully nonlinear system. Therefore, in what follows we will explore this particular type of DAEs.

Let us consider DAEs (6). In $\mathbf{z}(t) = (\mathbf{x}(t), \mathbf{y}(t)) \in \mathbb{R}^{m+k}$ we distinguish two types of variables, in particular $\mathbf{x}(t) \in \mathbb{R}^m$ are called differential variables, and $\mathbf{y}(t) \in \mathbb{R}^k$, $k = n - m$, are called algebraic variables.

We rewrite (6) with the new variables $\mathbf{x}(t)$, $\mathbf{y}(t)$ as the semi-explicit DAEs:

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, \mathbf{y}), \tag{9}$$

$$\mathbf{0} = \mathbf{g}(\mathbf{x}, \mathbf{y}), \tag{10}$$

where $\mathbf{f} : U \times V \to \mathbb{R}^m$, $\mathbf{g} : U \times V \to \mathbb{R}^k$, $\mathbf{x} : I \to U$, $\mathbf{y} : I \to V$, $\mathbf{x} \in \mathcal{C}^1(I, \mathbb{R}^m)$ $I \subseteq \mathbb{R}$ is a compact interval, $U \subseteq \mathbb{R}^m$ and $V \subseteq \mathbb{R}^k$ are open regions, [18]

The proof of the following Theorem and more information can be found in e.g. [19, 13].

Theorem 3.1 Consider the semi-explicit differential algebraic equation (9)–(10). Then (9)–(10) has the differentiation index $\nu = 1$ if and only if the Jacobi matrix $\mathbf{g}_\mathbf{y}(\mathbf{x}, \mathbf{y})$ is regular for all consistent points $(\mathbf{x}, \mathbf{y}) \in U \times V$.

Remark 3.1 In (9),(10) the differential part of DAEs is denoted by $\mathbf{f}$, the algebraic part by $\mathbf{g}$.

Let us suppose that our system of DAEs (9),(10) has differentiation index $\nu = 1$. It implies that the Jacobi matrix $\mathbf{g}_{\mathbf{y}}(\mathbf{x}, \mathbf{y})$ is regular for all consistent points $(\mathbf{x}, \mathbf{y}) \in U \times V$. Thus according to the Implicit Function Theorem, there exists a function $\mathbf{h} : \mathbb{R}^m \to \mathbb{R}^k$, such that $\mathbf{y} = \mathbf{h}(\mathbf{x})$, and

$$\mathbf{g}(\mathbf{x},\, \mathbf{h}(\mathbf{x})) = \mathbf{0}, \quad \forall \mathbf{x} \in U \subseteq \mathbb{R}^m \,.$$

We substitute $\mathbf{y} = \mathbf{h}(\mathbf{x})$, $\mathbf{x} \in \mathbb{R}^m$, into (9) and obtain

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, \mathbf{h}(\mathbf{x})), \tag{11}$$

where $\mathbf{x} \in U \subseteq \mathbb{R}^m$.

The equation (11) is a system of ODEs on the $(n-k)$–dimensional manifold

$$\mathbb{M} = \{(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^{m+k} : \mathbf{g}(\mathbf{x}, \mathbf{y}) = \mathbf{0}\}, \quad m + k = n. \tag{12}$$

Let again a continuous and differentiable scalar function $\varphi : \mathcal{D} \subseteq \mathbb{R}^{m+k} \to \mathbb{R}$ divide the region $\mathcal{D} \subseteq \mathbb{R}^{m+k}$ into three parts:

$$\begin{aligned} S_1 &= \{(\mathbf{x},\, \mathbf{y}) \in \mathcal{D} \subseteq \mathbb{R}^{m+k} : \varphi(\mathbf{x},\, \mathbf{y}) > 0\}, \\ S_2 &= \{(\mathbf{x},\, \mathbf{y}) \in \mathcal{D} \subseteq \mathbb{R}^{m+k} : \varphi(\mathbf{x},\, \mathbf{y}) < 0\}, \\ \Sigma &= \{(\mathbf{x},\, \mathbf{y}) \in \mathcal{D} \subseteq \mathbb{R}^{m+k} : \varphi(\mathbf{x},\, \mathbf{y}) = 0\}. \end{aligned}$$

We define a Filippov system $\mathcal{F}$ on $\mathcal{D} = S_1 \cup S_2 \cup \Sigma$ as

$$\mathcal{F} : \begin{bmatrix} \dot{\mathbf{x}} \\ \mathbf{0} \end{bmatrix} = \begin{cases} \mathbf{F}^{(1)}(\mathbf{x},\, \mathbf{y}), & (\mathbf{x},\, \mathbf{y}) \in S_1 \\ \mathbf{F}^{(0)}(\mathbf{x},\, \mathbf{y}), & (\mathbf{x},\, \mathbf{y}) \in \Sigma \\ \mathbf{F}^{(2)}(\mathbf{x},\, \mathbf{y}), & (\mathbf{x},\, \mathbf{y}) \in S_2 \end{cases} \quad \mathbf{F}^{(i)} = \begin{bmatrix} \mathbf{f}^{(i)} \\ \mathbf{g}^{(i)} \end{bmatrix}, \quad i = 0, 1, 2\,,$$

where $\mathbf{x}(t) \in \mathbb{R}^m$, $\mathbf{y}(t) \in \mathbb{R}^k$, $t \in \mathbb{R}$, $\mathbf{f}^{(i)} : \mathbb{R}^m \times \mathbb{R}^k \to \mathbb{R}^m$, $\mathbf{g}^{(i)} : \mathbb{R}^m \times \mathbb{R}^k \to \mathbb{R}^k$, $i = 0, 1, 2$, are sufficiently smooth functions in all arguments.

Similarly as in generic Filippov systems, we define the function

$$\sigma(\mathbf{x},\, \mathbf{y}) = \langle \nabla\varphi(\mathbf{x},\, \mathbf{y}), \mathbf{F}^{(1)}(\mathbf{x},\, \mathbf{y}) \rangle \cdot \langle \nabla\varphi(\mathbf{x},\, \mathbf{y}), \mathbf{F}^{(2)}(\mathbf{x},\, \mathbf{y}) \rangle.$$

that divides the boundary Σ into a crossing set Σ_c and a sliding set Σ_s,

$$\Sigma_c \subseteq \Sigma = \{(\mathbf{x},\, \mathbf{y}) \in \Sigma : \sigma(\mathbf{x},\, \mathbf{y}) > 0\},$$

$$\Sigma_s \subseteq \Sigma = \{(\mathbf{x},\, \mathbf{y}) \in \Sigma : \sigma(\mathbf{x},\, \mathbf{y}) \leq 0\}.$$

On Σ_c, we set $\mathbf{F}^{(0)} = \frac{1}{2}\left(\mathbf{F}^{(1)} + \mathbf{F}^{(2)}\right)$, on Σ_s, we define the vector field $\mathbf{F}^{(0)}$ as a convex combination

$$\mathbf{F}^{(0)} = (1-\lambda)\,\mathbf{F}^{(1)} + \lambda\,\mathbf{F}^{(2)}, \quad \lambda = \frac{\langle \nabla\varphi, \mathbf{F}^{(1)}\rangle}{\langle \nabla\varphi, \mathbf{F}^{(1)} - \mathbf{F}^{(2)}\rangle}, \quad 0 \le \lambda \le 1\,. \tag{13}$$

According to the convex combination (13), we can couple the differential parts of DAEs given by $\mathbf{f}^{(1)}$, $\mathbf{f}^{(2)}$, and separate them from the coupling of the algebraic parts given by $\mathbf{g}^{(1)}$, $\mathbf{g}^{(2)}$, i.e.,

$$\mathbf{f}^{(0)} = (1-\lambda)\,\mathbf{f}^{(1)} + \lambda\,\mathbf{f}^{(2)}, \tag{14}$$

$$\mathbf{g}^{(0)} = (1-\lambda)\,\mathbf{g}^{(1)} + \lambda\,\mathbf{g}^{(2)}. \tag{15}$$

The coupling of the differential equations of DAEs (14) is the same as in Section 2, but the coupling of the algebraic equations (15) is much more difficult. We don't a priori know which equations couple together, because here we don't have derivatives on the left side of the equations.

There are different ways to deal with this problem. Some authors prefer to pair only differential equations of DAEs and then add to them all algebraic equations.

We prefer to pair algebraic equations, too. This, however, requires more information about the system $\mathcal{F}$. Usually, we model some real applications and therefore each equation (differential or algebraic) has a physical meaning. In that case, we couple together the algebraic equations with the same physical meaning. Otherwise we could obtain unreasonable results. For more details and examples of coupling, see [13].

Let

$$\mathbb{M}_i = \{(\mathbf{x},\,\mathbf{y}) \in \mathbb{R}^{m+k} :\, \mathbf{g}^{(i)}(\mathbf{x},\,\mathbf{y}) = \mathbf{0}\}, \quad i = 1, 2, \tag{16}$$

be $(n-k)$-manifolds, where $n = m+k$. In Figure 1, the evolution of the trajectory on the manifolds $\mathbb{M}_1$ and $\mathbb{M}_2$ is shown. The trajectory starts with the initial condition $(\mathbf{x}(t_0),\,\mathbf{y}(t_0)) = (\mathbf{x}_0,\,\mathbf{y}_0) \in \mathbb{M}_1$ and crosses the boundary Σ to the manifold $\mathbb{M}_2$ at the crossing point $(\mathbf{x}(t_e),\,\mathbf{y}(t_e)) = (\mathbf{x}_e,\,\mathbf{y}_e)$. The subscript e denotes the so-called *event*, here the event is the contact of the trajectory with the boundary. In the following example, we illustrate the behavior of trajectories on manifolds $\mathbb{M}_1$ and $\mathbb{M}_2$.

Example 3.1 Let us have the Filippov system

$$\mathcal{F}: \quad \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ 0 \end{bmatrix} = \begin{cases} \mathbf{F}^{(1)}(x_1,\,x_2,\,y)\ , & \varphi(x_1,\,x_2,\,y) < 0, \\ \\ \mathbf{F}^{(2)}(x_1,\,x_2,\,y)\ , & \varphi(x_1,\,x_2,\,y) > 0, \end{cases} \tag{17}$$

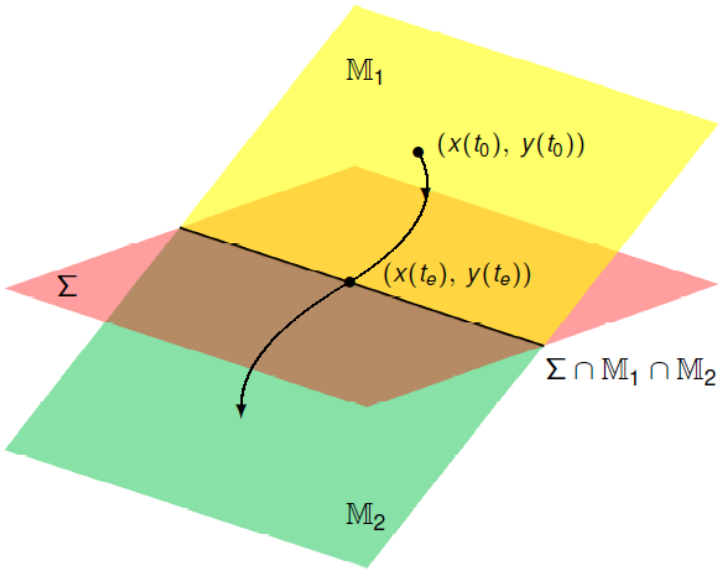


Figure 1: Evolution of the trajectory on the manifolds $\mathbb{M}_1$ and $\mathbb{M}_2$.

where

$$\mathbf{F}_1(x_1,\, x_2,\, y) \;=\; \left[\begin{array}{rcrcr} -x_1 & - & 3x_2 & + & y+15 \\ 3x_1 & - & x_2 & - & 2y \\ x_1 & & & - & y \end{array}\right], \qquad \left.\begin{array}{l} \\ \\ \end{array}\right\} \mathbf{f}^{(1)} \;\; \left.\begin{array}{l} \\ \\ \end{array}\right\} \mathbf{g}^{(1)} \tag{18}$$

$$\mathbf{F}_2(x_1,\, x_2,\, y) \;=\; \left[\begin{array}{rcrcr} x_1 & + & 3x_2 & + & 2y-1 \\ 3x_1 & + & x_2 & - & 3y \\ x_1 & & & + & y \end{array}\right], \qquad \left.\begin{array}{l} \\ \\ \end{array}\right\} \mathbf{f}^{(2)} \;\; \left.\begin{array}{l} \\ \\ \end{array}\right\} \mathbf{g}^{(2)} \tag{19}$$

Let the function $\varphi : \mathbb{R}^{2+1} \to \mathbb{R}$ be defined as

$$\varphi(x_1,\, x_2,\, y) = x_1. \tag{20}$$

Because $\nabla\varphi(x_1,\, x_2,\, y) = (1,\, 0,\, 0)$ and $x_1 = 0$ for $(x_1,\, x_2,\, y) \in \Sigma$, the scalar function $\sigma(x_1,\, x_2,\, y)$ has the form

$$\sigma(x_1,\, x_2,\, y) = (-3x_2 + y + 15)(3x_2 + 2y - 1).$$

The function σ divides the boundary Σ into two sets:

$$\Sigma_c \subseteq \Sigma = \{(x_1,\, x_2,\, y) \in \Sigma : \sigma(x_1,\, x_2,\, y) > 0\},$$

$$\Sigma_s \subseteq \Sigma = \{(x_1,\, x_2,\, y) \in \Sigma : \sigma(x_1,\, x_2,\, y) \leq 0\}.$$

On Σ_s, we set

$$\mathbf{F}^{(0)} = (1-\lambda)\,\mathbf{F}^{(1)} + \lambda\,\mathbf{F}^{(2)},$$

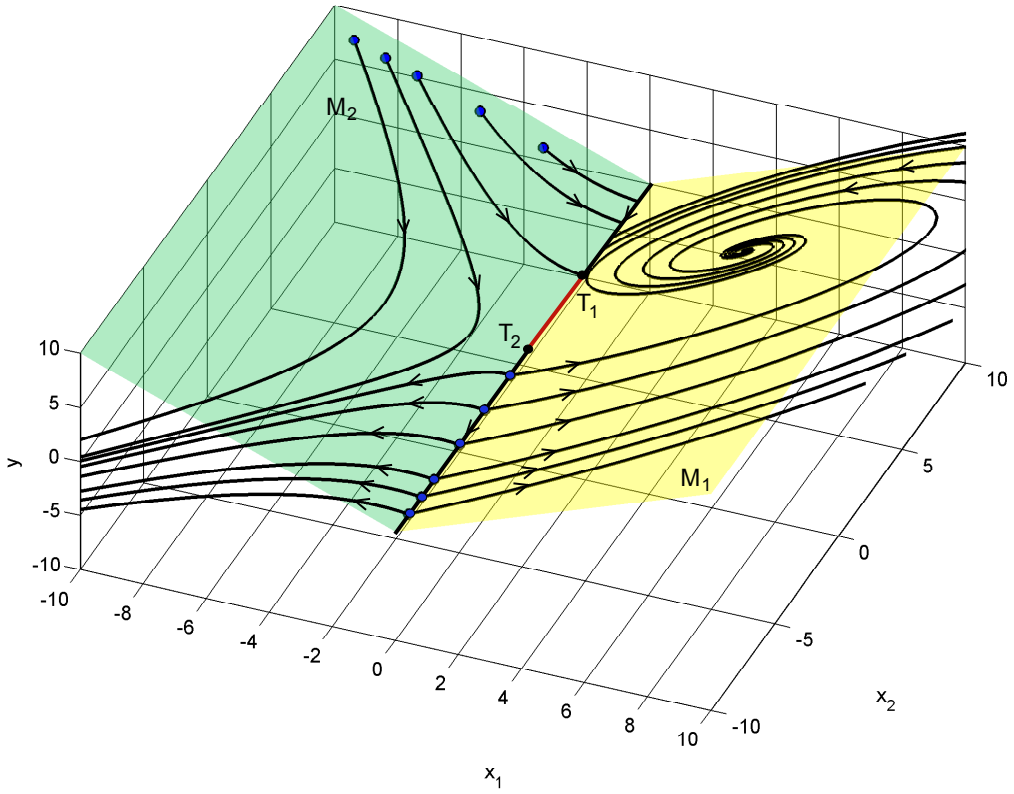


Figure 2: The phase portrait of the Filippov system in Example.

where

$$\lambda = \frac{-3x_2 + y + 15}{-6x_2 - y + 16}.$$

In Figure 2, the initial condition for each trajectory is depicted with the small blue circle. The yellow and green planes are the $(n-k)$–dimensional manifolds $\mathbb{M}_1$ and $\mathbb{M}_2$, $n = 3$, $k = 1$,

$$\mathbb{M}_1 = \{(x, y) \in \mathbb{R}^3 : \ y = x_1\}, \quad \mathbb{M}_2 = \{(x, y) \in \mathbb{R}^3 : \ y = -x_1\}. \tag{21}$$

The boundary Σ is depicted as the intersection of manifolds $\mathbb{M}_1$ and $\mathbb{M}_2$. On the boundary Σ, there are two tangent points T_1 and T_2 that delimit the set of sliding.

4. Soft drink process

The process of manufacturing soft-drink depicted in Figure 3 is based on the reaction between CO_2 and water:

$$CO_2 + H_2O \rightarrow H_2CO_3. \tag{22}$$

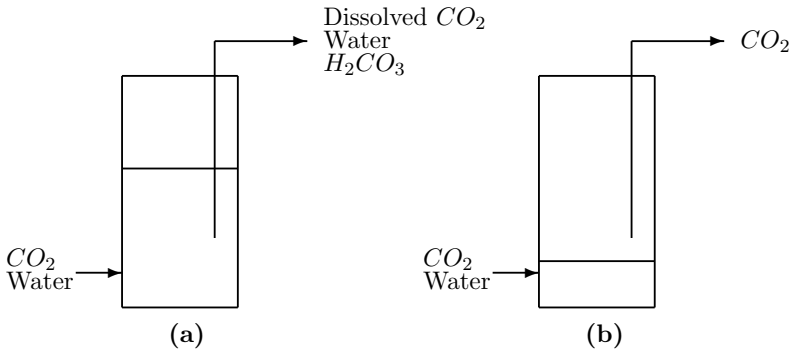


Figure 3: The soft-drink process.

To simplify the model, we will suppose that

- The system contains only components CO_2, H_2O and H_2CO_3 (denoted by indices 1, 2 and 3, respectively).
- Intermediate ionisation reactions and dissociation of H_2CO_3 are ignored.
- In the liquid there are no gas bubbles.
- The valve dynamics is ignored.
- The flow rate through the valve is proportional to the difference of the tank pressure P and the outlet pressure P_{out}.
- The temperature T, the molar inflow rates F_1 and F_2, the outlet pressure, valve coefficients k_G and k_L and the valve opening X are all constant.

Let

$$M_1 = M_1(t)\,, \qquad M_2 = M_2(t)\,, \qquad M_3 = M_3(t)\,,$$

be the total molar hold-ups of CO_2, H_2O and H_2CO_3, respectively. For a fixed t, let us define a scalar function $\varphi = \varphi(M_1, M_2, M_3)$,

$$\varphi(M_1, M_2, M_3) = \frac{M_2}{\rho_L} + \frac{M_3}{\rho_a} - V_d\,, \tag{23}$$

where ρ_L, ρ_a are molar densities of water and acid, respectively. The volume of the whole tank is equal to V and the part of the volume that is below the opening of the dip tube is denoted as V_d, $0 < V_d < V$.

Similarly as in [4] and [2], in the tank two systems take place: the liquid model (the liquid leaves the tank) if $\varphi(M_1, M_2, M_3) > 0$ or the gas model (the gas leaves the tank) for $\varphi(M_1, M_2, M_3) < 0$. The acid phase consists of H_2CO_3, H_2O and dissolved CO_2 while the gas phase contains only CO_2. As a consequence, the liquid model is described by 3 ODEs and 6 algebraic equations, the gas model needs also 3 ODEs but only 4 algebraic equations. Let us give the list of these equations.

Differential equations:

Liquid model : $\varphi(M_1, M_2, M_3) > 0$	Gas model : $\varphi(M_1, M_2, M_3) < 0$
$\dfrac{\mathrm{d}M_1}{\mathrm{d}t} = F_1 - L_1 - rV\,,$	$\dfrac{\mathrm{d}M_1}{\mathrm{d}t} = F_1 - G - rV\,,$
$\dfrac{\mathrm{d}M_2}{\mathrm{d}t} = F_2 - L_2 - rV\,,$	$\dfrac{\mathrm{d}M_2}{\mathrm{d}t} = F_2 - rV\,,$
$\dfrac{\mathrm{d}M_3}{\mathrm{d}t} = -L_3 + rV\,,$	$\dfrac{\mathrm{d}M_3}{\mathrm{d}t} = rV\,,$

The molar flow rates of the components through the valve are denoted L_1, L_2 and L_3 in the liquid model and G in the gas model. The rate r of the reaction (22) is given by

$$r = \kappa_c \frac{M_1 M_2}{V^2}, \qquad \text{where} \quad \kappa_c \quad \text{is the rate constant.} \tag{24}$$

Algebraic equations:

Liquid model : $\varphi(M_1, M_2, M_3) > 0$	Gas model : $\varphi(M_1, M_2, M_3) < 0$
$0 = M_1 - (M_\ell + M_g)\,,$	$0 = M_1 - (M_\ell + M_g)\,,$
$0 = P - \dfrac{\sigma M_\ell}{M_\ell + M_2 + M_3}\,,$	$0 = P - \dfrac{\sigma M_\ell}{M_\ell + M_2 + M_3}\,,$
$0 = V - \left(\dfrac{M_1 \mathrm{R} T}{P} + \dfrac{M_2}{\rho_L} + \dfrac{M_3}{\rho_a}\right)\,,$	$0 = V - \left(\dfrac{M_1 \mathrm{R} T}{P} + \dfrac{M_2}{\rho_L} + \dfrac{M_3}{\rho_a}\right)\,,$
$0 = \dfrac{M_2}{M_\ell + M_2 + M_3} - \dfrac{L_2}{L_1 + L_2 + L_3}\,,$	$0 = G - k_G X (P - P_{\mathrm{out}})\,,$
$0 = \dfrac{M_3}{M_\ell + M_2 + M_3} - \dfrac{L_3}{L_1 + L_2 + L_3}\,,$	
$0 = L_1 + L_2 + L_3 - k_L X (P - P_{\mathrm{out}})\,,$	

P and T means pressure and temperature in the tank, the hold-ups of CO_2 in liquid and gas are denoted M_ℓ and M_g, the constant X is a valve opening, R is a gas constant and σ is Henry's constant for CO_2.

The straightforward computation shows that both the system of DAEs for the gas mode and the system of DAEs for the liquid model have differentiation index $\nu = 1$, [13].

Let us denote $\mathbf{x} = (x_1, x_2, x_3)$ the differential variables, $\mathbf{y} = (y_1, y_2, y_3, y_4, y_5, y_6)$ the algebraic ones.

For differential variables in both models we set

$$x_1 := M_1, \quad x_2 := M_2, \quad \text{and} \quad x_3 := M_3\,.$$

As the algebraic variables are concerned, we have to distinguish the models. In the gas model the algebraic variables are $\mathbf{y} = (y_1, y_4, y_5, y_6)$ and we substitute

$$y_1 := G, \quad y_4 := M_g, \quad y_5 := M_\ell, \quad \text{and} \quad y_6 := P\,.$$

In the liquid model $\mathbf{y} = (y_1, y_2, y_3, y_4, y_5, y_6)$, where we substitute

$$y_1 := L_1, \quad y_2 := L_2, \quad y_3 := L_3, \quad y_4 := M_g, \quad y_5 := M_\ell, \quad \text{and} \quad y_6 := P\,.$$

We extend the functions $\mathbf{f}^{(1)}$, $\mathbf{f}^{(2)}$, $\mathbf{g}^{(1)}$, $\mathbf{g}^{(2)}$ to all variables from liquid and gas model $(\mathbf{x}, \mathbf{y}) = (x_1, x_2, x_3, y_1, y_2, y_3, y_4, y_5, y_6)$. Then we can define the Filippov system $\mathcal{F}$

$$\mathcal{F}: \quad \begin{bmatrix} \dot{\mathbf{x}} \\ \mathbf{0} \end{bmatrix} = \begin{cases} \mathbf{F}^{(1)}(\mathbf{x},\, \mathbf{y})\ , & (\mathbf{x},\, \mathbf{y}) \in S_1, \\ \mathbf{F}^{(0)}(\mathbf{x},\, \mathbf{y})\ , & (\mathbf{x},\, \mathbf{y}) \in \Sigma, \\ \mathbf{F}^{(2)}(\mathbf{x},\, \mathbf{y})\ , & (\mathbf{x},\, \mathbf{y}) \in S_2 \end{cases} \qquad \mathbf{F}^{(i)} = \begin{bmatrix} \mathbf{f}^{(i)} \\ \mathbf{g}^{(i)} \end{bmatrix}, \quad i = 0, 1, 2, \tag{25}$$

where we set $\mathbf{x} = (x_1, x_2, x_3)$ and $\mathbf{y} = (y_1, y_2, y_3, y_4, y_5, y_6)$, and

$$\mathbf{f}^{(1)}(\mathbf{x}, \mathbf{y}) = \begin{bmatrix} F_1 - G - \kappa_c \dfrac{M_1 M_2}{V} \\ F_2 - \kappa_c \dfrac{M_1 M_2}{V} \\ \kappa_c \dfrac{M_1 M_2}{V} \end{bmatrix}, \quad \mathbf{f}^{(2)}(\mathbf{x}, \mathbf{y}) = \begin{bmatrix} F_1 - L_1 - \kappa_c \dfrac{M_1 M_2}{V} \\ F_2 - L_2 - \kappa_c \dfrac{M_1 M_2}{V} \\ -L_3 + \kappa_c \dfrac{M_1 M_2}{V} \end{bmatrix}. \tag{26}$$

$$\mathbf{g}^{(1)}(x, y) = \begin{bmatrix} y_1 - k_G X (y_6 - P_{\text{out}}) \\ x_1 - (y_5 + y_4) \\ y_6 - \dfrac{\sigma y_5}{y_5 + x_2 + x_3} \\ V - \left(\dfrac{x_1 RT}{y_6} + \dfrac{x_2}{\rho_L} + \dfrac{x_3}{\rho_a} \right) \end{bmatrix}, \tag{27}$$

$$\mathbf{g}^{(2)}(x, y) = \begin{bmatrix} \dfrac{x_2}{y_5 + x_2 + x_3} - \dfrac{y_2}{y_1 + y_2 + y_3} \\ \dfrac{x_3}{y_5 + x_2 + x_3} - \dfrac{y_3}{y_1 + y_2 + y_3} \\ y_1 + y_2 + y_3 - k_L X (y_6 - P_{\text{out}}) \\ x_1 - (y_5 + y_4) \\ y_6 - \dfrac{\sigma y_5}{y_5 + x_2 + x_3} \\ V - \left(\dfrac{x_1 RT}{y_6} + \dfrac{x_2}{\rho_L} + \dfrac{x_3}{\rho_a} \right) \end{bmatrix}. \tag{28}$$

We apply the routine described in Section 3 to our system and obtain the convex combination of the differential part:

$$\dot{x_1} = F_1 - y_1 - \kappa_c \frac{x_1 x_2}{V}, \tag{29}$$

$$\dot{x_2} = -\lambda y_2 + F_2 - \kappa_c \frac{x_1 x_2}{V}, \tag{30}$$

$$\dot{x_3} = -\lambda y_3 + \kappa_c \frac{x_1 x_2}{V}, \tag{31}$$

where

$$\lambda = \frac{F_2 \rho_a + \kappa_c \frac{x_1 x_2}{V^2}(\rho_L - \rho_a)}{y_2 \rho_a + y_3 \rho_L}. \tag{32}$$

The convex combination of the algebraic part is

$$\begin{aligned}
0 &= \frac{x_2}{y_5 + x_2 + x_3} - \frac{y_2}{y_1 + y_2 + y_3}, \\
0 &= \frac{x_3}{y_5 + x_2 + x_3} - \frac{y_3}{y_1 + y_2 + y_3}, \\
0 &= (1 - \lambda)\left(y_1 - k_G X(y_6 - P_{\text{out}})\right) + \lambda\left(y_1 + y_2 + y_3 - k_L X(y_6 - P_{\text{out}})\right), \\
0 &= x_1 - (y_5 + y_4), \\
0 &= y_6 - \frac{\sigma y_5}{y_5 + x_2 + x_3} \\
0 &= V - \left(\frac{x_1 \mathrm{R} T}{y_6} + \frac{x_2}{\rho_L} + \frac{x_3}{\rho_a}\right)
\end{aligned}$$

Parameter	Value	Meaning
F_1 (mol/s)	0.5	molar inflow of CO_2
F_2 (mol/s)	7.5	molar inflow of water
ρ_L (mol/ℓ)	50	molar density of water
ρ_a (mol/ℓ)	16	molar density of acid
V (ℓ)	10	volume of the tank
V_d (ℓ)	2.25	volume below the outlet tube
$T(K)$	293	absolute temperature
P_{out} (atm)	1	pressures in the outlet
X	1.0	valve opening
k_L (mol/atm/s)	2.5	valve coef. for the liquid flow
k_G (mol/atm/s)	3.0	valve coef. for the gas flow
κ_c (ℓ/mol/s)	0.433/4000	rate constant
σ (atm)	1640	Henry's constant for CO_2
R (ℓ atm/mol/K)	0.0820574587	gas constant

Table 1: The parameters used for the simulation of the system.

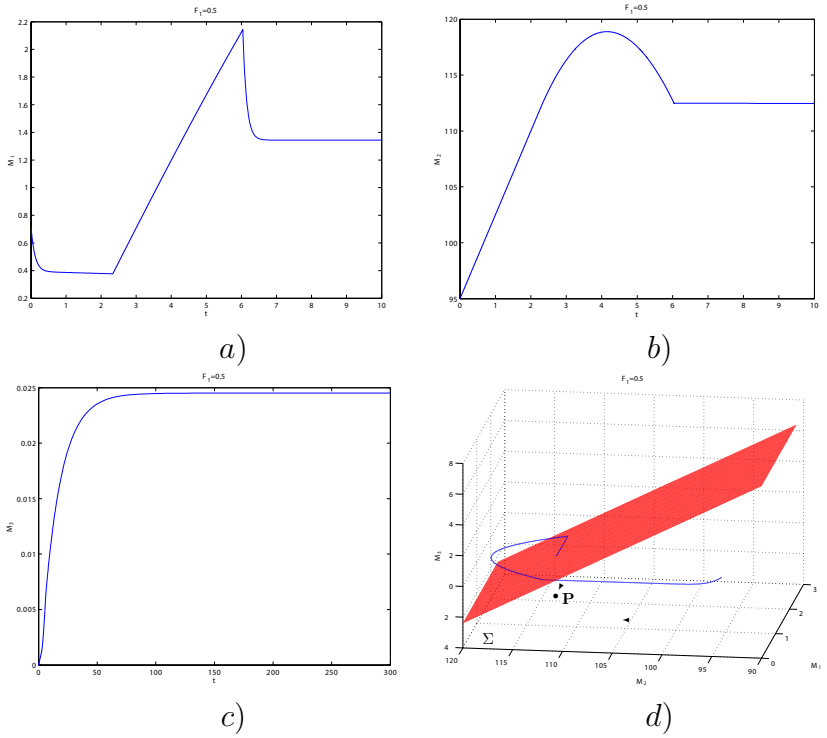


Figure 4: Soft-drink process: a)–c) The integral curves of the state variables M_1, M_2 and M_3. d) The trajectory of the system (25) starting at the point (0.72, 95, 0).

The behavior of the solution of the Filippov system (25) depends on thirteen parameters $F_1, F_2, \rho_L, \rho_a, V, V_d, T, P_{\text{out}}, X, k_L, k_G, \kappa_c, \sigma$, for a particular values used in simulations, see Table 1.

In Figure 4 a)–c), the integral curves of the state variables M_1, M_2 and M_3 are depicted. In Figure 4 d), the trajectory in coordinates (M_1, M_2, M_3) starting at the point (0.72, 95, 0) is drown, and the boundary Σ (red plane) is shown. On the boundary Σ, the generic pseudo-equilibrium P was detected.

5. Conclusions

In the paper, we gave a brief overview of the theory of Filippov dynamical systems for ordinary differential equations. Many specific applications for example in chemical engineering are based on models of differential algebraic equations, i.e., the problem formulation contains both differential equations and algebraic equations. We show that also in this case the system can be seen as a dynamical system of Filippov type.

As a practical example, a model of the gas-liquid system with a reaction is pre-

sented. This system can't be formulated as a Filippov system with ODEs only. An extension of the Filippov systems theory is necessary. By using a modified Filippov convex method, the integral curves of both differential and algebraic variables can be obtained.

Let us remark that the study of the gas-liquid system is just the first step towards modeling of the real HDPE (High Density Polyethylene) reactor.

In the future, we intend to perform additional studies of Filippov systems with DAEs. Till now, there are assumptions that are too restrictive. Deeper understanding of the behavior of non-smooth dynamical systems defined by DAEs is required.

In simplified model, the generic pseudo-equilibrium P on the boundary Σ acted as an attractor for the whole state space, see [2, 3]. We want to find out whether this also applies in a more general model.

All MATLAB simulations were performed in a modified version of the program developed by Petri T. Piiroinen and Yuri A. Kuznetsov [17].

References

[1] Agrawal, J., Moudgalya, K. M., and Pani, A. K.: Sliding motion of discontinuous dynamical systems described by semi-implicit index one differential algebraic equations. Chemical Engineering Science **61** (2006), 4722–4731.

[2] Biák, M.: *Piecewise smooth dynamical systems*. Ph.D. thesis, University of Chemistry and Technology, Prague, 2015.

[3] Biák, M. and Janovská, D.: Filippov dynamical systems. In: R. Blaheta and J. Starý (Eds.), *Seminar on Numerical Analysis & Winter School, Proceedings of the Conference SNA'09*, Institute of Geonics AS CR, Ostrava, 2009, Appendix, pp. 1–4.

[4] Biák, M. and Janovská, D.: Filippov systems with DAE. In: R. Blaheta, J. Starý, and D. Sysalová (Eds.), *Seminar on Numerical Analysis & Winter School, Proceedings of the Conference SNA'15*, Institute of Geonics AS CR, Ostrava, 2015, 13–16.

[5] di Bernardo, M., Budd, C. J., Champneys, A. R., and Kowalczyk, P.: *Piecewise-smooth dynamical systems: theory and applications*. Springer-Verlag, London, 2008.

[6] Dirac, P. A. M.: Generalized Hamiltonian dynamics. Can. J. Math. **2** (1950), 129–148.

[7] Dirac, P. A. M.: Generalized Hamiltonian dynamics. Proc. Royal Soc. London A 246 (1958), 326–332.

[8] Dirac, P. A. M.: *Lectures on Quantum Mechanics*. Yeshiva University, Dover, 1964.

[9] Filippov, A. F.: *Differential equations with discontinuous righthand sides*. Kluwer Academic Publishers, Dordrecht, 1988.

[10] Gantmacher, F. R.: *The theory of matrices I*. Chelsea Publishing Company, New York, 1959.

[11] Gantmacher, F. R.: *The theory of matrices II*. Chelsea Publishing Company, New York, 1959.

[12] Hirsch, M. W., Smale, S., and Devaney, R. L.: *Differential equations, dynamical systems, and an introduction to chaos*. Academic Press, London, 2004.

[13] Kunkel, P. and Mehrmann, V.: *Differential-algebraic equations: analysis and numerical solution*. European Mathematical Society, Zürich, 2006.

[14] Kronecker, L.: *Algebraische Reduction der Schaaren bilinearer Formen*. Sitzungsberichte Akad. Wiss. Berlin (1890), 1225–1237.

[15] Kuznetsov, Yu. A., Rinaldi, S., and Gragnani, A.: One-parameter bifurcations in planar Filippov systems. Int. J. Bifurcation & Chaos **13** (2003), 2157–2188.

[16] Meiss, J. D.: *Differencial dynamical systems*. Society for Industrial and Applied Mathematics SIAM, Philadelphia, 2007.

[17] Piiroinen, P. T. and Kuznetsov, Yu. A.: ACM Trans. Math. Software **34** (2008) 124.

[18] Steinbrecher, A.: *Numerical solution of quasi-linear differential-algebraic equations and industrial simulation of multibody systems*. Thesis, TU Berlin, 2006.

[19] Riaza, R.: *Differential-algebraic systems: analytical aspects and circuit applications*. World Scientific Publishing Company, Incorporated, 2008.

[20] Weierstrass, K.: Zur Theorie der bilinearen und quadratischen Formen. Monatsberichte Akad. Wiss. Berlin (1868) 310–338.

Conference Applications of Mathematics 2015, in honor of the birthday anniversaries of Ivo Babuška (90), Milan Práger (85), and Emil Vitásek (85)
J. Brandts, S. Korotov, M. Křížek, K. Segeth, J. Šístek, T. Vejchodský (Eds.)
Institute of Mathematics CAS, Prague 2015

DIMENSION REDUCTION FOR INCOMPRESSIBLE PIPE AND OPEN CHANNEL FLOW INCLUDING FRICTION

Mehmet Ersoy

Université de Toulon
IMATH, EA 2134, 83957 La Garde, France
Mehmet.Ersoy@univ-tln.fr

Abstract: We present the full derivation of a one-dimensional free surface pipe or open channel flow model including friction with non constant geometry. The free surface model is obtained from the three-dimensional incompressible Navier-Stokes equations under shallow water assumptions with prescribed "well-suited" boundary conditions.

Keywords: free surface flow, incompressible Navier-Stokes, shallow water approximation, hydrostatic approximation, closed water pipe, open channel, friction

MSC: 65M08, 65M75, 76B07, 76M12, 76M28, 76N15

1. Introduction

Simulation of free surface pipe or open channel flow plays an important role in many engineering applications such as storm sewers, waste or supply pipes in hydroelectric installations, etc.

The free surface flows are described by a newtonian, viscous and incompressible fluid through the three-dimensional incompressible Navier-Stokes equations. The use of the full three-dimensional equations leads to time-consuming simulations. Therefore, for specific applications such as shallow water, one can proceed to a model reduction preserving some of the main physical features of the flow leading to the so-called shallow water equations. This is one of the most challenging issues that we address with the obvious consequence to decrease the computational time. During these last years, many efforts were devoted to the modelling and the simulation of free surface water flows (see for instance [14, 13, 6, 5, 10, 9, 8, 11, 1, 7, 2, 3] and the reference therein).

The classical shallow water equations are usually derived from the three-dimensional Navier-Stokes equations (or the two-dimensional Navier-Stokes equations) by vertical averaging. It leads to a two-dimensional or a one-dimensional

shallow water model. For instance, Gerbeau and Perthame [10] study the full derivation of the one-dimensional shallow water equations from the two-dimensional Navier-Stokes equations while [11] considers the two-dimensional equations obtained from the three-dimensional one. In both cases, the so-called "motion by slices" is obtained. This property ensures that the horizontal velocity does not depend upon the vertical coordinate. As a consequence, one can perform the model reduction by vertical averaging. Following the applications under consideration, one can take into account as a source term the Coriolis effects, the topography, the friction, the capillary effects, the geometry, etc.

Unlike the previous works, we propose to study the full derivation of a **one-dimensional** free surface flows for pipe and open channel from the **three-dimensional** Navier-Stokes equations. In particular, we propose to revisit the work by Bourdarias et al. [3] done in the context of the three-dimensional Euler equations. The use of the Navier-Stokes equations with suitable boundary conditions allows first to establish the crucial "motion by slices" property, and second to include the friction (linear or non-linear) into the derivation. Let us emphasize that it was not possible to deal with in the framework of Bourdarias et al. [3]. More precisely, this property was assumed from the beginning and the friction was added to the obtained averaged equations.

The paper is organized as follows. In Section 2, we recall the full incompressible Navier-Stokes equations defining the boundary conditions including a general friction law, and we fix the notations. The "motion by slices" property under large Reynolds number flows is obtained through the hydrostatic equations (approximation) in Section 3. Next, these equations are averaged through the pipe or open channel section assumed to be orthogonal to the main flow direction. Finally, we obtain the one-dimensional free surface model. Since the constructed model is similar to the one by Bourdarias et al. [3], the issues of the numerical approximation is not addressed here. Please, refer to [1] or [4].

2. The incompressible Navier-Stokes equation and its closure

In this section, we fix the notations of the geometrical quantities involved to describe the thin domain representing a pipe or an open channel. In particular, without loss of generality (see Remark 2.1), we consider the case of pipe with circular section.

2.1. Geometrical settings

Let us consider an incompressible fluid confined in a three-dimensional rigid domain $\mathcal{P}$ representing a pipe or a channel, of length L:

$$\mathcal{P} := \left\{(x, y, z) \in \mathbb{R}^3;\ x \in [0, L],\ (y, z) \in \Omega_p(x)\right\}$$

where the section $\Omega_p(x)$, $x \in [0, L]$, is

$$\Omega_p(x) = \{(y, z) \in \mathbb{R}^2;\ y \in [\alpha(x, z), \beta(x, z)],\ z \in [0, 2R(x)]\}$$

as displayed on figure 1(a). Both flows and pipe are assumed to be oriented in the **i**-direction.

With these settings, we define the free surface section by

$$\Omega(t,x) = \Omega_p(x) \cap \{(y,z) \in \mathbb{R}^2; 0 \leqslant z \leqslant H(t,x,y)\}, \quad t > 0,\ x \in [0,L]$$

assumed to be orthogonal to the main flow direction. $H(t,x,y)$ is the local water elevation from the surface $z = 0$ in the $\Omega_p(x)$-plane. $R(x)$ stands for the radius of the pipe section $S(x) = \pi R^2(x)$, $\alpha(x,z)$ (resp. $\beta(x,z)$) is the left (resp. the right) boundary point at elevation $0 \leqslant z \leqslant 2R(x)$ as displayed on figure 1(b).

On the wet boundary (part of the boundary in contact with water), we define the coordinate of a point $\mathbf{m} \in \partial\Omega(t,x) := \Gamma_b(t,x)$, $t > 0$, $x \in [0,L]$, by $(y, \varphi(x,y))$ where

$$\Gamma_b(t,x) = \{(y,z) \in \mathbb{R}^2;\ z = \varphi(x,y) \leqslant H(t,x,y)\}\ .$$

Then, we note $\mathbf{n} = \dfrac{\mathbf{m}}{|\mathbf{m}|}$ the outward unit vector at the point $\mathbf{m} \in \partial\Omega(t,x)$, $x \in [0,L]$ as represented on figure 1(b). The point $\mathbf{m}$ also stands for the vector $\omega\mathbf{m}$ where $\omega(x,0,b(x))$ defines the main slope elevation of the pipe with $b'(x) = \sin\theta(x)$.

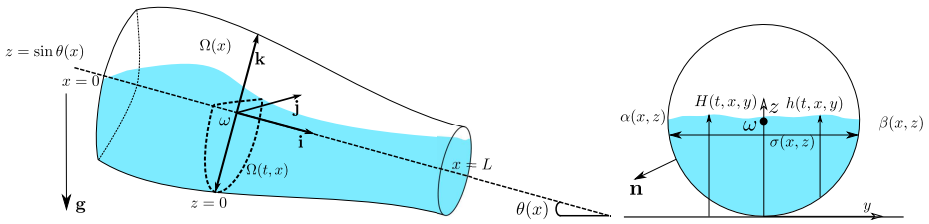
On the free surface, we define the coordinate of a point $\mathbf{m} \in \partial\Omega(t,x) := \Gamma_{fs}(t,x)$, $t > 0$, $x \in [0,L]$, by $(y, H(t,x,y))$ where

$$\Gamma_{fs}(t,x) = \{(y,z) \in \mathbb{R}^2;\ z = H(t,x,y)\}\ .$$

Finally, we note

$$h(t,x,y) = H(t,x,y) - \varphi(x,y)$$

the local elevation of the water.



(a) Configuration (b) Ω-plane

Figure 1: Geometric characteristics of the pipe

Remark 2.1 One can easily adapt this work to any realistic pipe or open channel by defining appropriately the previous quantities. For instance, in the case of "horseshoe" section (see figure 2(a)), the section $\Omega_p(x)$, $x \in [0,L]$, is given by

$$\Omega_p(x) = \Omega_H(x) \cap \Omega_R(x)$$

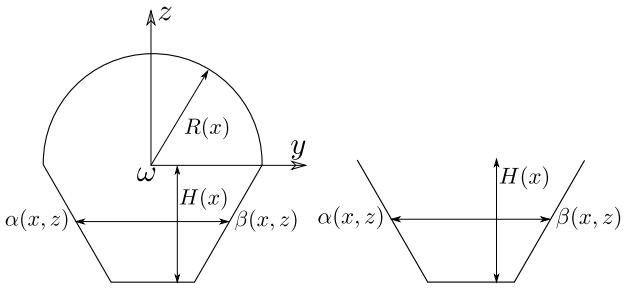
where

$$\Omega_H(x) = \left\{(y,z) \in \mathbb{R}^2; y \in [\alpha(x,z), \beta(x,z)],\ z \in [0, H(x)]\right\}$$

and

$$\Omega_R(x) = \{(y,z) \in \mathbb{R}^2;\ y \in [\alpha(x,z), \beta(x,z)],\ z \in [H(x), R(x)]\}\ .$$

H is the height of the trapezoidal basis and R is the radius of the upper part of the "horseshoe". A second example is represented on figure 2(b).



(a) "horseshoe" section (b) open channel

Figure 2: Example of a pipe and a open channel geometry

2.2. The water flow model

In the domain $\mathcal{P}$, we assume that the flow is incompressible and the pipe is always partially filled (otherwise we have to deal with pressurized flows that we omit here, please see [3] for details). Thus, we consider the incompressible Navier-Stokes equations with a prescribed general wall law conditions including friction on the wet boundary and a no stress one on the free surface. We complete the system with inflows and outflows conditions at the upstream and downstream ends.

The governing equations for the motion of an incompressible fluid in $[0,T] \times \mathcal{P}$, $T > 0$ are given by

$$\left\{ \begin{array}{rcl} \mathrm{div}(\rho_0 \mathbf{u}) & = & 0\ , \\ \partial_t(\rho_0 \mathbf{u}) + \mathrm{div}(\rho_0 \mathbf{u} \otimes \mathbf{u}) - \mathrm{div}\sigma - \rho_0 F & = & 0\ , \end{array} \right. \tag{1}$$

where $\mathbf{u} = \begin{pmatrix} u \\ \mathbf{v} \end{pmatrix}$ is the velocity fields with u the $\mathbf{i}$-component and $\mathbf{v} = \begin{pmatrix} v \\ w \end{pmatrix}$ the Ω-component, ρ_0 is the density of the fluid at atmospheric pressure and $F = -g \begin{pmatrix} -\sin\theta(x) \\ 0 \\ \cos\theta(x) \end{pmatrix}$ is the external gravity force of constant g. The total stress tensor can be written:

$$\sigma = \begin{pmatrix} -p + 2\mu\partial_x u & \mathcal{R}(\mathbf{u})^t \\ \mathcal{R}(\mathbf{u}) & -pI_2 + 2\mu D_{y,z}(\mathbf{v}) \end{pmatrix} \tag{2}$$

where I_2 is the identity matrix, μ is the dynamical viscosity and $\mathcal{R}(\mathbf{u})$ is defined by $\mathcal{R}(\mathbf{u}) = \mu\left(\nabla_{y,z}u + \partial_x \mathbf{v}\right)$. $\nabla_{y,z}u = \begin{pmatrix} \partial_y u \\ \partial_z u \end{pmatrix}$ is the gradient of u with respect to (y,z). Noting X^t the transpose of X, we define the strain tensor $D_{y,z}(\mathbf{v})$ with respect to the variable (y,z):

$$2D_{y,z}(\mathbf{u}) = \nabla_{y,z}\mathbf{v} + \nabla_{y,z}^t \mathbf{v} \ .$$

2.3. The boundary conditions

The Navier-Stokes system (1)–(2) is completed with suitable boundary conditions to introduce the border friction term on the wet boundary. On the free surface, we prescribe a no-stress condition.

On the wet boundary

For pipe flow calculations, the Darcy-Weisbach equation, valid for laminar as well as turbulent flows, is generally adopted. Roughly speaking, such formula relates losses h occurred during flows and it reads:

$$h = C_f \frac{L}{D}\frac{U^2}{2g}$$

where L, D, U are the pipe length, the pipe diameter and the velocity. The friction factor C_f, rather being a simple constant, turns out to be a factor that depends upon several parameters such as the Reynolds number R_e, the relative roughness δ, the Froude number F_r, the Mach number M_a, geometrical parameters, etc., and cannot be set as a constant. Following the type of the material, rough or smooth pipe, leaves C_f depend upon less quantities and lead to several expressions. An empirical transition function for the region between smooth pipes and the complete turbulence zone has been proposed by Colebrook:

$$\frac{1}{\sqrt{C_f}} = -0.86 \ln\left(\frac{\delta}{3.7D} + \frac{2.51}{R_e\sqrt{C_f}}\right)$$

where δ is the roughness of the material.

Because of the extreme complexity of the rough surfaces, most of the advances in understanding have been developed around experiments leading to charts such as the Moody-Stanton diagram, expressing C_f as a function of the Reynolds number R_e, the relative roughness and some geometrical parameters depending on the material. This yields to several formula depending on the modelling, for instance Chézy and Manning which are well-known by the engineers community, see for instance [16, 15].

For laminar flow, the effects of the material roughness can be ignored due to a presence of a thin laminar film at the pipe wall. Then, it can be shown that the Darcy-Weisbach equation reduces to $C_f = \dfrac{64}{R_e}$ that we note $C_f = C_l$ in the sequel. And, the losses are directly proportional to the velocity. When increasing the

Reynolds number R_e, the thin laminar film becomes unstable and causes turbulence increasing the head loss. Thus, the dependence on the Reynolds number R_e can be neglected and the head loss is almost directly proportional to U^2. The value of the friction factor, that we note $C_f = C_t$ in the sequel, can be read on diagrams.

In particular, this motivates the use of the following general friction law:

$$k(\mathbf{u})\mathbf{u} = C_f(|\mathbf{u}|)\mathbf{u} = C_l\mathbf{u} + C_t|\mathbf{u}|\mathbf{u},\ C_l \geqslant 0, C_t > 0 \tag{3}$$

where C_f stands for the friction factor. We do not intend in this work to define precisely the friction law but instead, we want to directly include it in its general form to explicitly show its dependency on physical parameters in the present model reduction.

Thus, on the inner wall $\partial\Omega_p(x)$, $\forall x \in (0, L)$, we assume a wall-law condition including a general friction law:

$$(\sigma(\mathbf{u})\mathbf{n}_b) \cdot \boldsymbol{\tau}_{b_i} = \rho_0 k(\mathbf{u})\mathbf{u} \cdot \boldsymbol{\tau}_{b_i},\ \ x \in (0, L),\ (y, z) \in \Gamma_b(x),\ i = 1, 2$$

where $\boldsymbol{\tau}_{b_i}$ is the i^{th} vector of the tangential basis and $\mathbf{n}_b$ stands for the unit outward normal vector:

$$\mathbf{n}_b = \frac{1}{\sqrt{(\partial_x\varphi)^2 + \mathbf{n}\cdot\mathbf{n}}} \begin{pmatrix} -\partial_x\varphi \\ \mathbf{n} \end{pmatrix}$$

with $\mathbf{n} = \begin{pmatrix} -\partial_y\varphi \\ 1 \end{pmatrix}$ the outward normal vector in the Ω_p-plane. Writing the wall-law condition in its vectorial form (i.e. the tangential constraints),

$$\sigma(\mathbf{u})\mathbf{n}_b - (\sigma(\mathbf{u})\mathbf{n}_b \cdot \mathbf{n}_b)\,\mathbf{n}_b = \rho_0 k(\mathbf{u})\mathbf{u},\ \ t > 0,\ x \in (0, L),\ (y, z) \in \Gamma_b(t, x)\ ,$$

one can split up the $\mathbf{i}-$component and the $(\mathbf{j}, \mathbf{k})-$components. Thus, the wall-law boundary conditions are

$$\begin{aligned} \mathcal{R}(\mathbf{u}) \cdot \mathbf{n}\ (\mathbf{n}\cdot\mathbf{n} - (\partial_x\varphi)^2) + 2\mu\partial_x\varphi\,(D_{y,z}(\mathbf{v})\mathbf{n}\cdot\mathbf{n} - \partial_x u\ (\mathbf{n}\cdot\mathbf{n})) \\ = \left(\mathbf{n}\cdot\mathbf{n} + (\partial_x\varphi)^2\right)^{3/2} \rho_0 k(u)u\ , \end{aligned} \tag{4}$$

$$\begin{aligned} 2\mu(\partial_x\varphi)^2\,(D_{y,z}(\mathbf{v})\mathbf{n} - \mathbf{n}) + \partial_x\varphi\mathcal{R}(\mathbf{u})\ (\mathbf{n}\cdot\mathbf{n} - (\partial_x\varphi)^2) \\ = \left(\mathbf{n}\cdot\mathbf{n} + (\partial_x\varphi)^2\right)^{3/2} \rho_0 k(\mathbf{v})\mathbf{v}\ . \end{aligned} \tag{5}$$

supplemented with a no-penetration condition:

$$\mathbf{u}\cdot\mathbf{n}_b = 0,\ \ t > 0,\ x \in (0, L),\ (y, z) \in \Gamma_b(t, x)$$

i.e.

$$u\partial_x\varphi = \mathbf{v}\cdot\mathbf{n},\ \ t > 0,\ x \in (0, L),\ (y, z) \in \Gamma_b(t, x)\ . \tag{6}$$

On the free surface boundary

For the sake of simplicity, on the free surface we assume a no-stress condition:

$$\sigma(\mathbf{u})\boldsymbol{N}^{fs} = 0, \;\; t > 0, \;\; x \in (0, L), \;\; (y, z) \in \Gamma_{fs}(t, x)$$

where

$$\boldsymbol{N}^{fs} = \frac{1}{\sqrt{(\partial_x H)^2 + \mathbf{n}_{fs} \cdot \mathbf{n}_{fs}}} \begin{pmatrix} -\partial_x H \\ \mathbf{n}_{fs} \end{pmatrix} \text{ where } \mathbf{n}_{fs} = \begin{pmatrix} -\partial_y H \\ 1 \end{pmatrix}$$

is the outward normal vector to the free surface.

Finally, as done before, splitting up the horizontal and the Ω_p-component, the free surface boundary conditions read

$$(p - 2\mu\partial_x u)\partial_x H + R(u) \cdot \mathbf{n}_{fs} = 0 \; , \tag{7}$$

$$R(u)\partial_x H + (p - 2\mu D_{y,z}(\mathbf{v}))\mathbf{n}_{fs} = 0 \; . \tag{8}$$

Introducing the indicator function Φ of the fluid region

$$\Phi(t, x, y, z) = \begin{cases} 1 & \text{if } \varphi(x, y) \leqslant z \leqslant H(t, x, y) \; , \\ 0 & \text{otherwise} \end{cases}$$

and because of the incompressibility condition, the divergence equation can be expressed as follows:

$$\partial_t \Phi + \partial_x(\Phi u) + \mathrm{div}_{y,z}(\Phi \mathbf{v}) = 0 \; . \tag{9}$$

3. The averaged model

The technique presented in this section is the one introduced by Gerbeau and Perthame [10] in the context of the reduction of the two-dimensional incompressible Navier-Stokes model to the one-dimensional shallow water equations. Here, instead, we proceed to the reduction of the three-dimensional incompressible Navier-Stokes equations to a one-dimensional shallow water equations.

3.1. Dimensionless Navier-Stokes equations

Thus, in the sequel we consider the non-dimensional form of the Navier-Stokes system using the shallow water assumption by introducing a "small" parameter so that

$$\varepsilon = \frac{D}{L} = \frac{W}{U} = \frac{V}{U} \ll 1$$

where U, $\boldsymbol{V} = (V, W)$ are the characteristic speeds in the $\mathbf{i}-$direction and the $(\mathbf{j}, \mathbf{k})-$direction.

We introduce a characteristic time T and a characteristic pressure P such that $T = \dfrac{L}{U}$ and $P = \rho_0 U^2$. The dimensionless quantities of time $\tilde{t}$, coordinate $(\tilde{x}, \tilde{y}, \tilde{z})$ and velocity field $(\tilde{u}, \tilde{v}, \tilde{w})$, noted temporarily by a $\tilde{\cdot}$, are defined by

$$\tilde{t} = \frac{t}{T}, \quad (\tilde{x}, \tilde{y}, \tilde{z}) = \left(\frac{x}{L}, \frac{y}{D}, \frac{z}{D}\right), \quad (\tilde{u}, \tilde{v}, \tilde{w}) = \left(\frac{u}{U}, \frac{v}{W}, \frac{w}{W}\right)$$

with the modified friction factor C_f/U that we write in the sequel C_f.

Let us define the following non-dimensional numbers:

$$\begin{array}{lll} F_r & \text{Froude number following the } \Omega\text{-plane} & : \ F_r = U/\sqrt{gD}\,, \\ F_L & \text{Froude number following the } \mathbf{i}\text{-direction} & : \ F_L = U/\sqrt{gL}\,, \\ R_e & \text{Reynolds number with respect to } \mu & : \ R_e = \rho_0 UL/\mu\,. \end{array}$$

Using these new variables in Equations (1), dropping the $\tilde{\cdot}$, ordering the terms with respect to ε, the dimensionless incompressible Navier-Stokes system becomes:

$$\begin{aligned} \operatorname{div}(\mathbf{u}) &= 0 && (10) \\ \partial_t(u) + \partial_x(u^2) + \operatorname{div}_{y,z}(u\mathbf{v}) + \partial_x p &= -\frac{\sin\theta(x)}{F_L^2} + \operatorname{div}_{y,z}\left(\frac{R_e^{-1}}{\varepsilon^2}\nabla_{y,z}u\right) && (11) \\ &\quad + R_{\varepsilon,1}(\mathbf{u}) \\ \nabla_{y,z}p &= \begin{pmatrix} 0 \\ -\dfrac{\cos\theta(x)}{F_r^2} \end{pmatrix} + R_{\varepsilon,2}(\mathbf{u}) && (12) \end{aligned}$$

where

$$R_{\varepsilon,1}(\mathbf{u}) = R_e^{-1}\left(\partial_x\left(2\partial_x u\right) + \operatorname{div}_{y,z}\left(\partial_x \mathbf{v}\right)\right) = O(R_e^{-1})$$

and

$$\begin{aligned} R_{\varepsilon,2}(\mathbf{u}) &= R_e^{-1}\left(\partial_x\left(\nabla_{y,z}u + \varepsilon^2\partial_x\mathbf{v}\right) + \operatorname{div}_{y,z}\left(2D_{y,z}(\mathbf{v})\right)\right) \\ &\quad -\varepsilon^2\left(\partial_t(\mathbf{v}) + \partial_x(u\mathbf{v}) + \operatorname{div}_{y,z}(\mathbf{v}\otimes\mathbf{v})\right)\,, \\ &= R_e^{-1}\left(\partial_x\left(\nabla_{y,z}u\right) + \operatorname{div}_{y,z}\left(2D_{y,z}(\mathbf{v})\right)\right) + \boldsymbol{O}(\varepsilon^2)\,, \\ &= \boldsymbol{O}(R_e^{-1}) + \boldsymbol{O}(\varepsilon^2)\,. \end{aligned}$$

The first component of the wall-law boundary condition (4) becomes:

$$\begin{aligned} \frac{R_e^{-1}}{\varepsilon}\nabla_{y,z}u\cdot\mathbf{n} &= \frac{\left(\mathbf{n}\cdot\mathbf{n} + \varepsilon^2(\partial_x\varphi)^2\right)^{3/2}\frac{k(u)}{U}u}{\left(\mathbf{n}\cdot\mathbf{n} - \varepsilon^2(\partial_x\varphi)^2\right)} \\ &\quad \varepsilon R_e^{-1}\left(\frac{2\partial_x\varphi\left(D_{y,z}(\mathbf{v})\mathbf{n}\cdot\mathbf{n} - \partial_x u\ (\mathbf{n}\cdot\mathbf{n})\right)}{\left(\mathbf{n}\cdot\mathbf{n} - \varepsilon^2(\partial_x\varphi)^2\right)} + \partial_x\mathbf{v}\cdot\mathbf{n}\right)\,, \qquad (13) \\ &= -K(u) + O(\varepsilon) + O(\varepsilon R_e^{-1}) \end{aligned}$$

where we make use of the notations

$$K(u) = \sqrt{\mathbf{n}\cdot\mathbf{n}}\frac{k(u)}{U}u \quad \text{and} \quad \nabla_{y,z}u\cdot\mathbf{n} := \partial_{\mathbf{n}}u$$

which are respectively the friction term and the normal derivative of u in the Ω_p-plane.

The second component of the wall-law boundary condition (5) becomes:

$$\begin{aligned} R_e^{-1}\nabla_{y,z}u &= \frac{\varepsilon^2\left(\mathbf{n}\cdot\mathbf{n}+\varepsilon^2(\partial_x\varphi)^2\right)^{3/2}\rho_0\frac{k(\mathbf{v})}{U}\mathbf{v}}{\partial_x\varphi(\mathbf{n}\cdot\mathbf{n}-\varepsilon^2(\partial_x\varphi)^2)} \\ &\quad -\frac{2\varepsilon^3R_e^{-1}\partial_x\varphi^2\left(D_{y,z}(\mathbf{v})\mathbf{n}-\mathbf{n}\right)}{\partial_x\varphi(\mathbf{n}\cdot\mathbf{n}-\varepsilon^2(\partial_x\varphi)^2)}-\varepsilon^2\partial_x\mathbf{v}\cdot\mathbf{n}\ , \\ &= \boldsymbol{O}(\varepsilon^2)+\boldsymbol{O}(\varepsilon^3R_e^{-1}) \end{aligned} \tag{14}$$

On the free surface, the boundary conditions (7)-(8) are now

$$\begin{aligned} R_e^{-1}\nabla_{y,z}u\cdot\mathbf{n}_{fs} &= -\varepsilon^2\left((p-2R_e^{-1}\partial_x u)\partial_x H+R_e^{-1}\partial_x\mathbf{v}\cdot\mathbf{n}_{fs}\right) \qquad (15)\\ &= O(\varepsilon^2)\ , \\ (p-2R_e^{-1}D_{y,z}(\mathbf{v}))\mathbf{n}_{fs} &= -\left(R_e^{-1}\nabla_{y,z}u+\varepsilon^2R_e^{-1}\partial_x\mathbf{v}\right)\partial_x H\ . \qquad (16) \end{aligned}$$

Thanks to the relations (15) and (16), the pressure on the free surface satisfies the following equality

$$p\ (\mathbf{n}_{fs}\cdot\mathbf{n}_{fs})-2R_e^{-1}D_{y,z}(\mathbf{v})\mathbf{n}_{fs}\cdot\mathbf{n}_{fs}=\varepsilon^2\left(\partial_x H\right)^2(p-2R_e^{-1}\partial_x u)=O(\varepsilon^2)\ . \tag{17}$$

3.2. First order approximation

As emphasized before in Section 2.3, when increasing the Reynolds number R_e, we observe instabilities at the pipe wall leading to turbulent flows. Assuming the characteristic length of the thin unstable film is larger than the relative roughness of the pipe, one can always assume some smallness of the friction law (see for instance [16, 15]). In particular, it motivates, for large Reynolds number R_e, the following asymptotic assumptions:

$$R_e^{-1}=\varepsilon\mu_0,\quad K=\varepsilon K_0 \tag{18}$$

where μ_0 is some viscosity constant and K_0 is the asymptotic friction law

$$K_0(u)=\sqrt{\mathbf{n}\cdot\mathbf{n}}\ k(u)u\ . \tag{19}$$

Under these conditions, the Archimedes principle is applicable and induces small vertical accelerations. As a consequence, one can drop all terms of order $O(\varepsilon^2)$ in Equations (10)–(12). Then, taking the formal limit as ε goes to 0, we deduce the hydrostatic equations

$$\partial_x(u_\varepsilon)+\mathrm{div}_{y,z}(\boldsymbol{v}_\varepsilon) = 0 \tag{20}$$

$$\partial_t(u_\varepsilon)+\partial_x(u_\varepsilon^2)+\mathrm{div}_{y,z}(u_\varepsilon\boldsymbol{v}_\varepsilon)+\partial_x p_\varepsilon = -\frac{\sin\theta(x)}{F_L^2}+\mathrm{div}_{y,z}\left(\frac{\mu_0}{\varepsilon}\nabla_{y,z}u_\varepsilon\right) \tag{21}$$

$$\nabla_{y,z}p_\varepsilon = \begin{pmatrix}0\\ -\dfrac{\cos\theta(x)}{F_r^2}\end{pmatrix} \tag{22}$$

Let us emphasize that even if this system results from a formal limit, we note its solution $(p_\varepsilon, u_\varepsilon, \boldsymbol{v}_\varepsilon)$ due to the explicit dependency on ε in the term $\mathrm{div}_{y,z}\left(\frac{\mu_0}{\varepsilon}\nabla_{y,z}u_\varepsilon\right)$ in Equation (21). At zero order, this term will be precisely the friction at the wet boundary through the condition (13). In particular, the boundary conditions write

- on the wet boundary; conditions (13)-(14) are

$$\frac{\mu_0}{\varepsilon}\nabla_{y,z}u_\varepsilon \cdot \mathbf{n} = K_0(u_\varepsilon) + O(\varepsilon), \quad t>0,\ x\in(0,L),\ (y,z)\in\Gamma_b(t,x)\ . \tag{23}$$

- on the free surface boundary; conditions (15)-(16) and (17) are

$$\frac{\mu_0}{\varepsilon}\nabla_{y,z}u_\varepsilon \cdot \boldsymbol{n}_\varepsilon^{fs} = O(\varepsilon), \quad t>0,\ x\in(0,L),\ (y,z)\in\Gamma_{fs}(t,x)\ . \tag{24}$$

Next, identifying terms at order $\dfrac{1}{\varepsilon}$ in Equations (20)–(22), thanks to Equations (23) and (24), we obtain the so-called "motion by slices"

$$u_\varepsilon(t,x,y,z) \;=\; u_0(t,x) + O(\varepsilon) \tag{25}$$

for some function $u_0 = u_0(t,x)$, by solving formally the Neumann problem for $t>0$, $x\in(0,L)$

$$\left\{\begin{array}{lcl} \mathrm{div}_{y,z}\left(\mu_0\nabla_{y,z}u_\varepsilon\right) &=& O(\varepsilon)\ ,\quad (y,z)\in\Omega(t,x) \\ \mu_0\partial_{\mathbf{n}}u_\varepsilon &=& O(\varepsilon)\ ,\quad (y,z)\in\partial\Omega(t,x) \end{array}\right.$$

One one hand, the following approximation at first order holds

$$u_\varepsilon(t,x,y,z) \;\approx\; \overline{u_\varepsilon}(t,x)$$

where $\overline{u_\varepsilon}(t,x) = \dfrac{1}{|\Omega_\varepsilon(t,x)|}\displaystyle\int_{\Omega_\varepsilon(t,x)} u_\varepsilon(t,x,y,z)\,dy\,dz$ is the mean speed of the fluid over the wet section. Consequently, one can approximate at first order the non-linear term as follows

$$\overline{u_\varepsilon^2} \approx \overline{u_\varepsilon}^2\ . \tag{26}$$

On the other hand, using the second component of Equations (22), we may write

$$\partial_z p_\varepsilon(t,x,y,z) = -\frac{\cos\theta(x)}{F_r^2} + O(\varepsilon)\ .$$

Then, fixing y and integrating this equation for $\xi\in[z,H(t,x,y)]$, keeping in mind the identity (17), we obtain

$$p_\varepsilon(t,x,y,z) = \frac{\cos\theta}{F_r^2}(H_\varepsilon(t,x,y)-z) + O(\varepsilon)\ .$$

Moreover, using the first component of Equations (22) leads to

$$H_\varepsilon(t,x,y) = H_\varepsilon(t,x,0) + O(\varepsilon)\ . \tag{27}$$

As a consequence, we recover the classical hydrostatic pressure

$$p_\varepsilon(t,x,y,z) \ \approx \ \frac{\cos\theta}{F_r^2}(H_\varepsilon(t,x,0) - z)\ , \tag{28}$$

Finally, in view of the the definition of the water elevation H_ε (27), the wet section is approximated at first order as follows, $t > 0, x \in [0, L]$:

$$\Omega_\varepsilon(t,x) = \{(y,z) \in \mathbb{R}^2; \alpha(x,z) \leqslant y \leqslant \beta(x,z) \text{ and } 0 \leqslant z \leqslant H_\varepsilon(t,x,0)\} \tag{29}$$

and the outward unit normal vector to the free surface $\mathbf{n}_{fs}$ is now $\boldsymbol{n}_\varepsilon^{fs} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$ as displayed on figure 3.

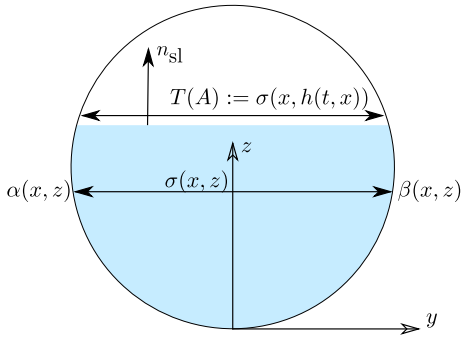


Figure 3: First order approximation of the wet area

In the sequel, due to its dependency at first order, we write $H_\varepsilon(t,x,y)$ by $H_\varepsilon(t,x)$.

3.3. The free surface model

By virtue of the relations (25)–(29), integrating Equations (20)–(22) over the cross-section $\Omega_p(t,x)$, the free surface model immediately follows.

First, let us recall that $\mathbf{m} = (y, \varphi(x,y)) \in \partial\Omega_p(x)$ stands for the vector $\omega\mathbf{m}$ and $\mathbf{n} = \dfrac{\mathbf{m}}{|\mathbf{m}|}$ for the outward unit normal vector to the boundary Γ_b at the point $\mathbf{m}$ in the Ω_p-plane as displayed on figure 1(b).

Second, let us introduce $A(t,x)$ and $Q(t,x)$ the conservative variables of wet area and discharge defined by the following relations:

$$A(t,x) = \int_{\Omega_\varepsilon(t,x)} dydz \tag{30}$$

and

$$Q(t,x) = A(t,x)\overline{u_\varepsilon}(t,x) \tag{31}$$

where

$$\overline{u_\varepsilon}(t,x) = \frac{1}{A(t,x)} \int_{\Omega_\varepsilon(t,x)} u(t,x,y,z)\, dydz$$

is the mean speed of the fluid over the section $\Omega_\varepsilon(t,x)$.

Equation of the conservation of the momentum and the kinematic boundary condition

Let $\mathbf{v}$ be the vector field $\begin{pmatrix} v \\ w \end{pmatrix}$. Integrating the equation of conservation of the mass (9) on the set:

$$\overline{\Omega}(x) = \{(y,z);\, \alpha(x,z) \leqslant y \leqslant \beta(x,z),\, 0 \leqslant z \leqslant \infty\},$$

we get the following equation:

$$\int_{\overline{\Omega}(x)} \partial_t(\phi) + \partial_x(\phi u_\varepsilon) + \mathrm{div}_{y,z}(\phi \boldsymbol{v}_\varepsilon)\, dydz = \partial_t A + \partial_x Q - \int_{\partial\Omega_\varepsilon(t,x)} (u_\varepsilon \partial_x \mathbf{m} - \boldsymbol{v}_\varepsilon)\cdot \mathbf{n}\; ds\ . \tag{32}$$

Now, integrating Equation (9) on $\Omega_\varepsilon(t,x)$, we get:

$$\int_0^{H_\varepsilon(t,x)} \partial_t \int_{\alpha(x,z)}^{\beta(x,z)} dydz + \partial_x Q + \int_{\partial\Omega_\varepsilon(t,x)} (\boldsymbol{v}_\varepsilon - u_\varepsilon \partial_x \mathbf{m})\cdot \mathbf{n}\, ds = 0$$

where

$$\int_0^{H_\varepsilon(t,x)} \partial_t \int_{\alpha(x,z)}^{\beta(x,z)} dydz = \partial_t A - \sigma(x, H_\varepsilon(t,x))\partial_t h$$

with $\sigma(x, H_\varepsilon(t,x))$ is the width at the free surface elevation as displayed on figure 3.

Then, one has:

$$\begin{aligned} &\partial_t(A) + \partial_x(Q) - \int_{\Gamma_\varepsilon^{fs}(t,x)} (\partial_t \mathbf{m} + u_\varepsilon \partial_x \mathbf{m} - \boldsymbol{v}_\varepsilon)\cdot \boldsymbol{n}_\varepsilon^{fs}\, ds \\ &- \int_{\Gamma_b(t,x)} (u_\varepsilon \partial_x \mathbf{m} - \boldsymbol{v}_\varepsilon)\cdot \mathbf{n}\, ds = 0. \end{aligned} \tag{33}$$

Keeping in mind the no penetration condition (6) and comparing Equations (32) and (33), we finally derive the kinematic boundary condition at the free surface:

$$\int_{\Gamma_\varepsilon^{fs}(t,x)} (\partial_t \mathbf{m} + u_\varepsilon \partial_x \mathbf{m} - \boldsymbol{v}_\varepsilon)\cdot \boldsymbol{n}_\varepsilon^{fs}\, ds = 0 \tag{34}$$

i.e.

$$\partial_t H_\varepsilon + u_\varepsilon(z = H_\varepsilon)\partial_x H_\varepsilon - w_\varepsilon(z = H_\varepsilon) = 0\ .$$

Finally, gathering Equations (33) and (34), we get the equation of the conservation of the mass:

$$\partial_t(A) + \partial_x(Q) = 0. \tag{35}$$

Equation of the conservation of the momentum

In order to get the equation of the conservation of the momentum of the free surface model, we integrate each term of Equation (21) over sections $\Omega_\varepsilon(t,x)$ as follows:

$$\int_{\Omega_\varepsilon(t,x)} \underbrace{\partial_t(u_\varepsilon)}_{a_1} + \underbrace{\partial_x(u_\varepsilon^2)}_{a_2} + \underbrace{\mathrm{div}_{y,z}\left(u_\varepsilon \boldsymbol{v}_\varepsilon\right)}_{a_3} + \underbrace{\partial_x p_\varepsilon}_{a_4}\, dydz = \int_{\Omega_\varepsilon(t,x)} \underbrace{-\frac{\sin\theta}{F_L^2}}_{a_5}\, dydz +$$

$$\int_{\Omega_\varepsilon(t,x)} \underbrace{\mathrm{div}_{y,z}\left(\frac{\mu_0}{\varepsilon}\nabla_{y,z}u_\varepsilon\right)}_{a_6}\, dydz\ .$$

By virtue of relations (25), (26) and (28), we successively get:

Computation of the term $\int_{\boldsymbol{\Omega_\varepsilon(t,x)}} \boldsymbol{a_1}\, \boldsymbol{dydz}$

The pipe being non-deformable, only the integral at the free surface is non zero since

$$\int_{\Gamma_b(t,x)} u_\varepsilon\, \partial_t \mathbf{m}\cdot\mathbf{n}\, ds = 0.$$

Thus, we get:

$$\int_{\Omega_\varepsilon(t,x)} \partial_t(u_\varepsilon)\, dydz \;=\; \partial_t \int_{\Omega_\varepsilon(t,x)} u_\varepsilon\, dydz - \int_{\Gamma_\varepsilon^{fs}(t,x)} u_\varepsilon\, \partial_t\mathbf{m}\cdot\boldsymbol{n}_\varepsilon^{fs}\, ds.$$

Computation of the term $\int_{\boldsymbol{\Omega_\varepsilon(t,x)}} \boldsymbol{a_2}\, \boldsymbol{dydz}$

$$\begin{aligned}\int_{\Omega_\varepsilon(t,x)} \partial_x(u_\varepsilon^2)\, dydz \;&=\; \partial_x \int_{\Omega_\varepsilon(t,x)} u_\varepsilon^2\, dydz - \int_{\Gamma_\varepsilon^{fs}(t,x)} u_\varepsilon^2 \partial_x\mathbf{m}\cdot\boldsymbol{n}_\varepsilon^{fs}\, ds \\ &\quad - \int_{\Gamma_b(t,x)} u_\varepsilon^2 \partial_x\mathbf{m}\cdot\mathbf{n}\, ds.\end{aligned}$$

Computation of the term $\int_{\boldsymbol{\Omega_\varepsilon(t,x)}} \boldsymbol{a_3}\, \boldsymbol{dydz}$

$$\int_{\Omega_\varepsilon(t,x)} \mathrm{div}_{y,z}\left(u_\varepsilon \boldsymbol{v}_\varepsilon\right)\, dydz \;=\; \int_{\Gamma_\varepsilon^{fs}(t,x)} u_\varepsilon \mathbf{v}\cdot\boldsymbol{n}_\varepsilon^{fs}\, ds + \int_{\Gamma_b(t,x)} u_\varepsilon \boldsymbol{v}_\varepsilon\cdot\mathbf{n}\, ds.$$

Summing the result of the previous step $a_1 + a_2 + a_3$, we get:

$$\int_{\Omega_\varepsilon(t,x)} a_1 + a_2 + a_3\, dydz = \partial_t(Q) + \partial_x\left(\frac{Q^2}{A}\right) \tag{36}$$

where A and Q are given by (30) and (31).

Computation of the term $\int_{\Omega_\varepsilon(t,x)} a_4\, dydz$

For the pressure term p_ε given by the relation (28), (t,x) fixed, we have:

$$\begin{aligned}
\int_{\Omega_\varepsilon(t,x)} \partial_x p_\varepsilon \, dydz &= \int_0^{H_\varepsilon(t,x)} \int_{\alpha(x,z)}^{\beta(x,z)} \partial_x p_\varepsilon \, dydz \\
&= \int_0^{H_\varepsilon(t,x)} \sigma(x,z) \partial_x p_\varepsilon \, dydz \\
&= \int_0^{H_\varepsilon(t,x)} \partial_x \left(p_\varepsilon \sigma(x,z)\right) dz - \int_0^{H_\varepsilon(t,x)} p_\varepsilon \partial_x \sigma(x,z)\, dz \\
&= \partial_x \int_{\Omega_\varepsilon(t,x)} p_\varepsilon \sigma(x,z)\, dydz \\
&\quad - \int_0^{H_\varepsilon(t,x)} p_\varepsilon \partial_x \sigma(x,z)\, dz - \partial_x H_\varepsilon(t,x) p_{\varepsilon|z=H_\varepsilon(t,x)}
\end{aligned}$$

Finally, we have:

$$\int_{\Omega_\varepsilon(t,x)} \partial_x p_\varepsilon \, dydz = \partial_x \left(g I_1(x,A) \frac{\cos\theta(x)}{F_r^2} \right) - g I_2(x,A) \frac{\cos\theta(x)}{F_r^2} \tag{37}$$

where I_1 is the hydrostatic pressure:

$$I_1(x,A) = \int_0^{H_\varepsilon(A)} (H_\varepsilon(A) - z) \sigma(x,z)\, dz.$$

The term I_2 is the pressure source term:

$$I_2(x,A) = \int_0^{H_\varepsilon(A)} (H_\varepsilon(A) - z) \partial_x \sigma(x,z)\, dz.$$

which takes into account of the section variation through the term $\partial_x \sigma(x,\cdot)$.

Computation of the term $\int_{\Omega_\varepsilon(t,x)} a_5\, dydz$

We have:

$$\int_{\Omega_\varepsilon(t,x)} g \sin\theta \, dydz = gA \sin\theta. \tag{38}$$

Computation of the term $\int_{\Omega_\varepsilon(t,x)} a_6\, dydz$

We have:

$$\int_{\Omega_\varepsilon(t,x)} \mathrm{div}_{y,z} \left(\frac{\mu_0}{\varepsilon} \nabla_{y,z} u_\varepsilon \right) dydz = \int_{\Gamma_\varepsilon^{fs}(t,x)} \frac{\mu_0}{\varepsilon} \nabla_{y,z} u_\varepsilon \cdot \boldsymbol{n}_\varepsilon^{fs}\, ds + \int_{\Gamma_b(t,x)} \frac{\mu_0}{\varepsilon} \nabla_{y,z} u_\varepsilon \cdot \mathbf{n}\, ds \tag{39}$$

where $\displaystyle\int_{\Gamma_\varepsilon^{fs}(t,x)} \frac{\mu_0}{\varepsilon}\nabla_{y,z}u_\varepsilon \cdot \boldsymbol{n}_\varepsilon^{fs}\, ds = 0$ due to the boundary condition (24). Using the boundary conditions (23) and the approximation (25), the second integral writes

$$\int_{\Gamma_b(t,x)} \frac{\mu_0}{\varepsilon}\nabla_{y,z}u_\varepsilon \cdot \mathbf{n}\, ds = \int_{\Gamma_b(t,x)} K_0(u_\varepsilon)\, ds = AK(\overline{u_\varepsilon})$$

where

$$K(x,\overline{u_\varepsilon}) = K_0(\overline{u_\varepsilon})\frac{\int_{\Gamma_b(t,x)} ds}{A}ds$$

with $\displaystyle\int_{\Gamma_b(t,x)} ds$ is the wet perimeter P_m (i.e. the portion of the perimeter where the wall is in contact with the fluid) and thus $\displaystyle\frac{A}{\int_{\Gamma_b(t,x)} ds}$ is nothing but the so-called hydraulic radius. This quantity was introduced by engineers as a length scale for non-circular ducts in order to use the analysis derived for the circular pipes (see for instance [16, 17]). Let us outline that this factor is naturally obtained in the derivation of the averaged model and holds for any realistic pipe or open channel (see Remark 2.1).

Then, gathering results (35) and (36)–(39), we get the equation of the conservation of the momentum. Finally, multiplying by $\rho_0 U^2/L$, the shallow water equations for free surface flows are:

$$\left\{\begin{array}{lcl} \partial_t(A) + \partial_x(Q) & = & 0 \\ \partial_t(Q) + \partial_x\left(\dfrac{Q^2}{A} + gI_1\cos\theta\right) & = & -gA\sin\theta + gI_2\cos\theta - gAK(x,Q/A) \end{array}\right. \tag{40}$$

This model takes into account the slope variation, change of section and the friction due to roughness on the inner wall of the pipe. This system was formally introduced by the author in [7] and [3] in the context of unsteady mixed flows in closed water pipes assuming the motion by slices that we have now justified here with the friction term.

We have proposed a finite volume discretisation of the free surface model introducing a new kinetic solver in [2, 4] based on the kinetic scheme of Perthame and Simeoni [12]. We have also proposed a new well-balanced VFRoe scheme [1]. These numerical schemes have been validated in [4] in a channel with varying width on a trans-critical steady state with shock. Several test cases have been passed with success through comparison with an exact solution or a code to code comparison, see for instance [1, 2].

4. Conclusions and perspectives

Finally, we have performed an asymptotic analysis of the three-dimensional incompressible Navier-Stokes equation with a general wall-law conditions including

friction and free surface boundary conditions in the shallow water limit. We have considered the three-dimensional incompressible hydrostatic approximation with friction boundary conditions and free surface boundary conditions and we have integrated these equations along the Ω sections to get the one-dimensional free surface model. In particular, we have shown that the free surface model (40) is an approximation of $O(\varepsilon)$ of the hydrostatic approximation (20)–(22) and therefore of the three-dimensional incompressible Navier-Stokes equations (10)-(12). Except the three-dimensional model reduction to a one-dimensional one, we have shown how to integrate correctly a general friction law into the model derivation. The next step and the work in progress will consist in studying the rigorous limit.

Acknowledgements

This work is supported by the ModTerCom project within the APEX program of the region Provence-Alpe-Côte d'Azur and the Project MTM2011-29306-C01-01 from the MICINN (Spain).

The author wish to thank the referees for their careful reading of the previous version of the manuscript and useful remarks.

References

[1] Bourdarias, C., Ersoy, M., and Gerbi, S.: A model for unsteady mixed flows in non uniform closed water pipes and a well-balanced finite volume scheme. Int. J. Finite **6** (2009), 1–47.

[2] Bourdarias, C., Ersoy, M., and Gerbi, S.: A kinetic scheme for transient mixed flows in non uniform closed pipes: a global manner to upwind all the source terms. J. Sci. Comput. **48** (2011), 89–104.

[3] Bourdarias, C., Ersoy, M., and Gerbi, S.: A mathematical model for unsteady mixed flows in closed water pipes. Sci. China Math. **55** (2012), 221–244.

[4] Bourdarias, C., Ersoy, M., and Gerbi, S.: Unsteady mixed flows in non uniform closed water pipes: a full kinetic approach. Numer. Math. **128** (2014), 217–263.

[5] Capart, H., Sillen, X., and Zech, Y.: Numerical and experimental water transients in sewer pipes. J. of Hydr. Res. **35** (1997), 659–672.

[6] Dong, N. T.: Sur une méthode numérique de calcul des écoulements non permanents soit à surface libre, soit en charge, soit partiellement à surface libre et partiellement en charge. La Houille Blanche (1990), 149–158.

[7] Ersoy, M.: *Modélisation, analyse mathématique et numérique de divers écoulements compressibles ou incompressibles en couche mince.* Ph.D. thesis, Université de Savoie, 2010.

[8] Ferrari, S. and Saleri, F.: A new two-dimensional shallow water model including pressure effects and slow varying bottom topography. ESAIM: M2AN **38** (2004), 211–234.

[9] Fuamba, M.: Contribution on transient flow modelling in storm sewers. J. of Hydr. Res. **40** (2002), 685–693.

[10] Gerbeau, J.F. and Perthame, B.: Derivation of viscous Saint-Venant system for laminar shallow water; numerical validation. Discrete Cont. Dyn. Syst. Ser. B **1** (2001), 89–102.

[11] Marche, F.: Derivation of a new two-dimensional viscous shallow water model with varying topography, bottom friction and capillary effects. Eur. J. Mech. B/ Fluids **26** (2007), 49–63.

[12] Perthame, B. and Simeoni, C.: A kinetic scheme for the saint-venant system with a source term. Calcolo **38** (2001), 201–231.

[13] Roe, P. L.: Some contributions to the modelling of discontinuous flows. In: B. E. Engquist, S. Osher, and R. C. J. Somerville (Eds.), *Large-Scale Computations in Fluid Mechanics*, *Lectures in Applied Mathematics*, vol. 22, pp. 163–193. American Mathematical Society, Providence, RI, 1985.

[14] Song, C. C., Cardie, J. A., and Leung, K. S.: Transient mixed-flow models for storm sewers. J. of Hydr. Eng. **109** (1983), 1487–1504.

[15] Streeter, V. L., Wylie, E. B., and Bedford, K. W.: Fluid Mechanics, wcb, 1998.

[16] Wylie, E. B. and Streeter, V. L.: *Fluid transients*. New York, McGraw-Hill International Book Co., 1978, 401 p.

[17] Wylie, E. B., Streeter, V. L., and Suo, L.: *Fluid transients in systems*. Prentice Hall Englewood Cliffs, NJ, 1993.

Conference Applications of Mathematics 2015, in honor of the birthday anniversaries of Ivo Babuška (90), Milan Práger (85), and Emil Vitásek (85)
J. Brandts, S. Korotov, M. Křížek, K. Segeth, J. Šístek, T. Vejchodský (Eds.)
Institute of Mathematics CAS, Prague 2015

ON CONTINUOUS AND DISCRETE MAXIMUM/MINIMUM PRINCIPLES FOR REACTION-DIFFUSION PROBLEMS WITH THE NEUMANN BOUNDARY CONDITION

István Faragó[1], Sergey Korotov[2], Tamás Szabó[3]

[1] Department of Applied Analysis and Computational Mathematics
Eötvös Loránd University
H–1117, Budapest, Pázmány P. s. 1/c., Hungary
& MTA-ELTE NumNet Research Group
faragois@cs.elte.hu
[2] Department of Computing, Mathematics and Physics
Bergen University College
Inndalsveien 28, 5020 Bergen, Norway
sergey.korotov@hib.no
[3] CAE Engineering Kft.
H–1118, Budapest, Kelenhegyi ut 43, Hungary
klaymen1984@gmail.com

Abstract: In this work, we present and discuss continuous and discrete maximum/minimum principles for reaction-diffusion problems with the Neumann boundary condition solved by the finite element and finite difference methods.

Keywords: elliptic problem, Neumann boundary condition, maximum/minimum principle, discrete maximum/minimum principle

MSC: 35B50, 65N06, 65N30, 65N50

1. Continuous maximum/minimum principles

Consider the following boundary-value problem of elliptic type: Find a function $u \in C^2(\overline{\Omega})$ such that

$$-\Delta u + cu = f \quad \text{in} \quad \Omega \quad \text{and} \quad \frac{\partial u}{\partial n} = g \quad \text{on} \quad \partial\Omega, \tag{1}$$

where $\Omega \subset \mathbf{R}^d$ is a bounded domain with Lipschitz continuous boundary $\partial\Omega$, n is the unit outward normal to $\partial\Omega$, and the reactive coefficient $c(x) \geq 0$ for all $x \in \overline{\Omega}$. The boundary condition in (1) is commonly called the *the Neumann boundary condition.*

The additional assumptions on the data of the problem will be given in appropriate places of the paper later on.

First, we prove the continuous *maximum/minimum principles* for problem (1) in the following form.

Theorem 1. *Assume that in* (1) *the functions* $c, f \in C(\overline{\Omega}), g \in C(\partial\Omega)$, *and* $c(x) \geq c_\star > 0$ *for all* $x \in \overline{\Omega}$, *where* $c_\star$ *is a positive constant. Let*

$$g(s) \leq -g_\star < 0 \quad \textit{for all} \quad s \in \partial\Omega, \tag{2}$$

where $g_\star$ *is a positive constant. Then the following a priori upper estimate (continuous maximum principle) for the classical solution of problem* (1) *is valid for any* $x \in \overline{\Omega}$*:*

$$u(x) \leq \max_{\bar{x}\in\overline{\Omega}} \frac{f(\bar{x})}{c(\bar{x})}. \tag{3}$$

Now, let

$$g(s) \geq g_\star > 0 \quad \textit{for all} \quad s \in \partial\Omega, \tag{4}$$

where $g_\star$ *is a positive constant. Then the following a priori lower estimate (continuous minimum principle) for the classical solution of problem* (1) *is valid for any* $x \in \overline{\Omega}$*:*

$$u(x) \geq \min_{\bar{x}\in\overline{\Omega}} \frac{f(\bar{x})}{c(\bar{x})}. \tag{5}$$

Proof. First, we prove estimate (3). If u attains its maximum at some interior point $x_0 \in \Omega$, then all the first order partial derivatives $u_{x_i}(x_0) = 0$, and all the second order partial derivatives $u_{x_ix_i}(x_0) \leq 0$ for $i = 1, 2, \ldots, d$. Therefore, from the equation in (1) and the positivity of c we observe that $u(x_0) \leq f(x_0)/c(x_0)$. Now we claim that under the assumptions of the theorem the maximum of u cannot be attained on the boundary. Indeed, if u attains its maximum at some boundary point $s_0 \in \partial\Omega$, then, unavoidably, $0 \leq \frac{\partial u}{\partial n}(s_0) = g(s_0)$, which contradicts the assumption on g in (2).

Obviously, estimate (5) can be proved in a similar way under conditions in (4). □

In what follows we will always assume that the following condition on the coefficient c holds

$$c(x) \geq c_\star > 0 \quad \text{for all} \quad x \in \overline{\Omega}, \tag{6}$$

where $c_\star$ is a positive constant.

The main goal of the paper is to construct suitable discrete analogues of (3) and (5), called the *discrete maximum/minimum principles*, and find practical conditions on the numerical schemes, namely the finite element method (FEM) and the finite difference method (FDM), providing their validity.

In most of available papers devoted to maximum principles for elliptic problems, see e.g. [9, 11] and references therein, continuous maximum (and minimum) principles usually take a form of implications involving certain sign-conditions. For example, for the equation from (1) combined with vanishing Dirichlet boundary condition, the maximum principle reads as follows:

$$f(x) \leq 0 \quad \text{in} \quad \overline{\Omega} \implies \max_{x \in \overline{\Omega}} u(x) \leq 0. \tag{7}$$

However, the implications with sign-conditions (like in (7)) have been recently generalized in [6, 7] to more general situations for problems with Dirichlet and Robin boundary condition. In this work we consider the case of Neumann problem and perform an analysis of some FE and FD schemes in the context of discrete maximum/minimum principles.

Remark 1. We mention that discrete maximum principles, besides their practical importance for imitating the nonnegativity of nonnegative physical quantities in numerical simulations, have been often used for proving stability and finding the rate of convergence of FD approximations, see e.g. [1, 2, 4], and for proving the convergence of FE approximations in the maximum norm, see e.g. [1, 5].

2. Discrete maximum principle

After discretization of problem (1) by many popular numerical techniques (e.g. by FEM and FDM) we arrive at the problem of solving $N \times N$ system of linear algebraic equations

$$\mathbf{A}\mathbf{u} = \mathbf{F}, \tag{8}$$

where the vector of unknowns $\mathbf{u} = [u_1, \ldots, u_N]^T$ approximates the unknown solution u at certain selected points $B_1, \ldots, B_N$ of the solution domain Ω and its boundary $\partial\Omega$, and the vector $\mathbf{F} = [F_1, \ldots, F_N]^T$ approximates (in the sense depending on the nature of the actual numerical method used) the values $f(B_i)$ and $g(B_j)$.

In what follows, the entries of matrix $\mathbf{A}$ are denoted by a_{ij}, and all matrix and vector inequalities appearing in the text are always understood component-wise.

Definition 1. The square $N \times N$ matrix $\mathbf{M}$ is called *monotone* if

$$\mathbf{M}\mathbf{z} \geq 0 \quad \implies \quad \mathbf{z} \geq 0. \tag{9}$$

Equivalently, monotone matrices are characterized as follows (see e.g. [2, p. 119]).

Theorem 2. *The square $N \times N$ matrix $\mathbf{M}$ is monotone if and only if $\mathbf{M}$ is nonsingular and $\mathbf{M}^{-1} \geq 0$.*

Definition 2. The square $N \times N$ matrix $\mathbf{M}$ is called *M-matrix* if it is monotone and its entries $m_{ij} \leq 0$ for $i \neq j$.

It is obvious that for M-matrix $\mathbf{M} = (m_{ij})$, we have $m_{ii} > 0$ for all $i = 1, \ldots, N$.

Definition 3. The square $N \times N$ matrix $\mathbf{M}$ (with entries m_{ij}) is called *strictly diagonally dominant* (or SDD in short) if the values

$$\delta_i(\mathbf{M}) := |m_{ii}| - \sum_{j=1,\, j\neq i}^{N} |m_{ij}| > 0 \quad \text{for all} \quad i = 1, \ldots, N. \tag{10}$$

In [17] the following result is proved.

Theorem 3. *Let matrix* $\mathbf{A}$ *in system* (8) *be SDD and M-matrix. Then the following two-sided estimates for the entries of the solution* $\mathbf{u}$ *are valid*

$$\min_{j=1,\ldots,N} \frac{F_j}{\delta_j(\mathbf{A})} \leq u_i \leq \max_{j=1,\ldots,N} \frac{F_j}{\delta_j(\mathbf{A})}, \quad i = 1, \ldots, N. \tag{11}$$

As the estimates in (11) resemble the estimates in (3) and (5), it is natural to give the following definition.

Definition 4. We say that the solution $\mathbf{u}$ of system (8) with an SDD matrix $\mathbf{A}$ satisfies the *discrete maximum principle* corresponding to continuous maximum principle (3), if the upper estimate in (11) is valid, and, in addition, the following inequality

$$\max_{j=1,\ldots,N} \frac{F_j}{\delta_j(\mathbf{A})} \leq \max_{\bar{x}\in\overline{\Omega}} \frac{f(\bar{x})}{c(\bar{x})} \tag{12}$$

holds. Similarly, we say that the solution $\mathbf{u}$ of system (8) with an SDD matrix $\mathbf{A}$ satisfies the *discrete minimum principle* corresponding to continuous minimum principle (5), if the lower estimate in (11) is valid, and, in addition, the following inequality

$$\min_{j=1,\ldots,N} \frac{F_j}{\delta_j(\mathbf{A})} \geq \min_{\bar{x}\in\overline{\Omega}} \frac{f(\bar{x})}{c(\bar{x})} \tag{13}$$

holds.

Remark 2. In case of earlier versions of continuous and discrete maximum principles no estimates like (12) and (13) were, in fact, needed as one dealt there with various implications involving the sign-conditions only (cf. [4, 5, 13, 9]).

Remark 3. The validity of relations (12) and (13) is important for producing controllable numerical approximations, because under these conditions the approximate solutions (obtained by the FEM or the FDM for example) stay within the same bounds as the exact solutions and these bounds are a priori known from the continuous problem.

3. DMPs for the finite element (FE) schemes

The standard FE scheme is based on the so-called variational formulation of (1), which reads: Find $u \in H^1(\Omega)$ such that

$$a(u,v) = \mathcal{F}(v) \qquad \forall v \in H^1(\Omega), \tag{14}$$

where

$$a(u,v) = \int_\Omega \nabla u \cdot \nabla v dx + \int_\Omega cuvdx\,,\ \mathcal{F}(v) = \int_\Omega fvdx + \int_{\partial\Omega} gvds. \tag{15}$$

The existence and uniqueness of the weak solution u is provided by the Lax-Milgram lemma, the Friedrichs-type inequalities, and the assumption on c (6) (cf. [14, Chapt. 2]). (Actually, for the well-posedness in above, one can require less smoothness from the problem data, e.g. that $c \in L^\infty(\Omega), f \in L^2(\Omega), g \in L^2(\partial\Omega)$ only.)

Let $\mathcal{T}_h$ be a FE mesh of $\overline{\Omega}$ with interior nodes $B_1, \ldots, B_n$ lying in Ω and boundary nodes $B_{n+1}, \ldots, B_{n+n^\partial}$ lying on $\partial\Omega$. The elements of $\mathcal{T}_h$ will be denoted by the symbol T, possibly with subindices. Further, let the basis functions $\phi_1, \phi_2, \ldots, \phi_{n+n^\partial}$, associated with these nodes, have the following properties

$$\phi_i(B_j) = \delta_{ij},\ i,j = 1,\ldots,n+n^\partial,\ \phi_i \geq 0 \text{ in } \overline{\Omega},\ i = 1,\ldots,n+n^\partial,$$

$$\sum_{i=1}^{n+n^\partial} \phi_i \equiv 1 \text{ in } \overline{\Omega}, \tag{16}$$

where δ_{ij} is the Kronecker delta. Note that these properties are easily met for example for the lowest-order simplicial, block, and prismatic finite elements. The basis functions $\phi_1, \phi_2, \ldots, \phi_{n+n^\partial}$ are spanning a finite-dimensional subspace V_h of $H^1(\Omega)$.

The FE approximation of u is defined to be a function $u_h \in V_h$ such that

$$a(u_h, v_h) = \mathcal{F}(v_h) \qquad \forall v_h \in V_h, \tag{17}$$

whose existence and uniqueness are also provided by the Lax-Milgram lemma.

Remark 4. Algorithmically, $u_h = \sum_{i=1}^{n+n^\partial} u_i\phi_i$, where the coefficients u_i are the entries of the solution $\mathbf{u}$ of system (8) with $a_{ij} = a(\phi_i, \phi_j)$, $F_i = \mathcal{F}(\phi_i)$, and $N = n + n^\partial$. It is clear that, if properties (16) hold, the FE approximation u_h satisfies the bounds from (11) at each point of $\overline{\Omega}$ if all its nodal values u_i do satisfy them.

Lemma 1. *Assume that problem* (1) *under condition* (2) *is solved by the FEM with basis functions satisfying* (16). *In addition, let the matrix* $\mathbf{A}$ *in the resulting system* $\mathbf{Au} = \mathbf{F}$ *be such that* $a_{ij} \leq 0$ *for* $i \neq j$. *Then* $\mathbf{A}$ *is SDD and estimates* (11) *are valid.*

Proof. From (15) and (2), it clearly follows that $a_{ii} = a(\phi_i, \phi_i) > 0$ for all $i = 1, \ldots, n + n^{\partial}$. If $a_{ij} \leq 0$ $(i \neq j)$, we observe for $i = 1, \ldots, n + n^{\partial}$ that

$$\delta_i(\mathbf{A}) = \sum_{j=1}^{n+n^{\partial}} a_{ij} = a(\phi_i, \sum_{j=1}^{n+n^{\partial}} \phi_j) = a(\phi_i, 1) = \int_{\Omega} c\phi_i dx > 0, \tag{18}$$

where the last (strict) inequality holds due to (2). Thus, the matrix $\mathbf{A}$ is always SDD for our type of problems. Moreover $\mathbf{A}$ is the Minkowski matrix, i.e. M-matrix (cf. [2, pp. 119–120]). Hence, estimates (11) are valid, due to Theorem 3, with $\delta_i(\mathbf{A})$ computed as in (18). □

The proofs of further estimates (12) and (13) strongly depend on the way we compute a_{ij} and F_j in real calculations. Below we consider in detail only the following representative case.

Theorem 4. *Assume that the coefficient c is a positive constant and that all entries a_{ij} and F_j in system* (8) *are computed exactly. Then estimates* (12) *and* (13)*, and therefore discrete maximum and minimum principles, corresponding to* (3) *and* (5)*, correspondingly, are valid provided $a_{ij} \leq 0$ for $i \neq j$, and the relevant sign condition on g holds.*

Proof. Let us prove first (12) under condition $g(s) \leq -g_\star < 0$. In view of (18), (15), (2), and the first mean value theorem for integration, we get

$$\frac{F_i}{\delta_i(\mathbf{A})} = \frac{\int_\Omega f\phi_i dx + \int_{\partial\Omega} g\phi_i ds}{\int_\Omega c\phi_i dx} \leq \frac{\int_\Omega f\phi_i dx}{c\int_\Omega \phi_i dx} =$$

$$= \frac{f(x^\star)\int_\Omega \phi_i dx}{c\int_\Omega \phi_i dx} \leq \max_{\xi\in\overline{\Omega}} \frac{f(\xi)}{c},$$

where $x^\star$ is some point from $\overline{\Omega}$ and i is an arbitrary index from the set $\{1, \ldots, n+n^{\partial}\}$.

Similarly, we can prove (13) under condition $g(s) \geq g_\star > 0$. □

Remark 5. In fact, the entries a_{ij} can always be computed exactly if c is a positive constant, and the entries F_j can be computed exactly if the functions f and g are piecewise polynomials for example. If c is not constant, and f and g are general functions, then for computations of entries (which are sums of integrals over Ω and its boundary $\partial\Omega$) in system (8), we should use, in practice, special quadrature rules, and, thus, each such a case requires a separate analysis in the context of discrete maximum/minimum principles (cf. [10]).

Remark 6. Various geometric conditions on FE meshes guaranteeing the validity of the requirement $a_{ij} \leq 0$ for $i \neq j$ are presented e.g. in [3, 8, 12].

4. DMPs for some finite difference (FD) schemes

On the base of several representative FD schemes, we shall demonstrate how the discrete maximum/minimum principles from Definition 4 can be proved also for finite difference approximations.

First, consider problem (1) posed in one-dimensional domain $\Omega = (0, 1)$. For the governing equation at the interior nodes we shall employ the following standard FD discretization:

$$\frac{-y_{i-1} + 2y_i - y_{i+1}}{h^2} + c_i y_i = f_i, \tag{19}$$

where $i = 1, ..., \hat{n} - 1,\ h = 1/\hat{n}$, c_i and f_i denote the values of functions c and f, respectively, at the node ih. The Neumann boundary condition is discretized as follows:

$$\frac{y_0 - y_1}{h} = g_0,\ \frac{y_{\hat{n}} - y_{\hat{n}-1}}{h} = g_{\hat{n}}. \tag{20}$$

The resulting FD system of linear equations is of size $(\hat{n} + 1) \times (\hat{n} + 1)$. However, its matrix is not SDD as, due to equations (20), the corresponding sums of entries of the matrix in the first and the last rows are zeros, so we cannot immediately use Theorem 3.

However, we notice that, e.g. under the sign-condition $g \le -g_\star < 0$ (used to prove the continuous maximum principle), it follows from (20) that $y_0 < y_1$ and $y_{\hat{n}} < y_{\hat{n}-1}$, and it is thus sufficient to get a suitable upper estimate only for the entries $y_1, \ldots, y_{\hat{n}-1}$. Further, we form the reduced system of equations of the size $(\hat{n}-1)\times(\hat{n}-1)$ for finding (and estimating) $y_1, \ldots, y_{\hat{n}-1}$ using discretization (19)–(20). This reduced system will consist of $\hat{n} - 3$ equations (19), for $i = 2, \ldots, \hat{n} - 2$, and two following equations

$$\frac{y_1 - y_2}{h^2} + c_1 y_1 = f_1 + \frac{g_0}{h},$$

$$\frac{-y_{\hat{n}-2} + y_{\hat{n}-1}}{h^2} + c_{\hat{n}-1} y_{\hat{n}-1} = f_{\hat{n}-1} + \frac{g_{\hat{n}}}{h},$$

obtained by combining (20) and (19) for $i = 1$ and $i = \hat{n} - 1$. It is clear that the corresponding sums $\delta_i(\mathbf{A}) = c_i,\ i = 1, \ldots, \hat{n} - 1$, and, therefore, the matrix of the reduced system is SDD and it is also M-matrix. Further, due to the sign-condition on g we observe that the entries of the right-hand side of the reduced system $F_i \le f_i,\ i = 1, \ldots, \hat{n} - 1$. Therefore, estimates (11) and (12) are valid, i.e. the discrete maximum principle holds. The discrete minimum principle can be proved similarly under the condition $g \ge g_\star > 0$.

Consider now the two-dimensional case. Let, for simplicity, the solution domain be a square, i.e. $\Omega = (0, 1) \times (0, 1)$. Using the same step-size $h = 1/\hat{n}$ in both directions and the classical 5-point FD stencil, we arrive at the following interior equations inside of $\overline{\Omega}$

$$\frac{-y_{i-1,j} - y_{i+1,j} - y_{i,j-1} - y_{i,j+1} + 4y_{i,j}}{h^2} + c_{i,j} y_{i,j} = f_{i,j}, \tag{21}$$

where now $i, j = 1, ..., \hat{n} - 1$ and $c_{i,j}$ and $f_{i,j}$ denote the values of functions c and f, respectively, at the node (ih, jh).

The first order accurate FD discretization of the Neumann boundary condition on $\partial\Omega$ (consisting of four intervals in this case) reads as follows:

$$\frac{y_{i,0} - y_{i,1}}{h} = g_{i,0}, \ \frac{y_{i,\hat{n}} - y_{i,\hat{n}-1}}{h} = g_{i,\hat{n}}, \ i = 1, 2, ..., \hat{n} - 1, \tag{22}$$

$$\frac{y_{0,j} - y_{1,j}}{h} = g_{0,j}, \ \frac{y_{\hat{n},j} - y_{\hat{n}-1,j}}{h} = g_{\hat{n},j}, \ j = 1, 2, ..., \hat{n} - 1, \tag{23}$$

where $g_{i,j}$ denotes the value of g at the node (ih, jh). We notice that we do not deal with the corner points of Ω in our case as the normal vectors are not well defined at these points.

We see again, that the matrix of the full system is not SDD, however, just the same trick as in the one-dimensional case can be used. And the following results can be easily proved.

Theorem 5. *The FD discretization* (21)–(23) *has the following properties:*

a) it approximates a sufficiently smooth solution u with the first order of accuracy,
b) the reduced FE matrix is SDD and is M-matrix,
c) the discrete maximum/minimum principles are valid provided the relevant conditions on g hold.

The approximation (22)–(23) (and (20)) of the Neumann boundary condition has only the first order of accuracy, which is not consistent with the second order of accuracy of the FD discretization for the governing differential equation. Therefore, we shall present and analyse another FD scheme, now with an increased accuracy of approximation for the Neumann boundary condition. We discuss in detail only the more complicated 2D case, because the analysis of 1D case is similar. So, let us approximate the Neumann boundary condition on the boundary of $\Omega = (0, 1) \times (0, 1)$ in the following manner:

- on the part of the boundary with $x = 0$ as

$$\begin{aligned} &\frac{y_{0,j} - y_{1,j}}{h} - \frac{h}{2}\left(\frac{y_{0,j+1} - 2y_{0,j} + y_{0,j-1}}{h^2}\right) + \frac{h}{2}c_{0,j}y_{0,j} = \\ &= g_{0,j} + \frac{h}{2}f_{0,j}, \quad j = 1, 2, \ldots, \hat{n} - 1. \end{aligned} \tag{24}$$

- on the part of the boundary with $x = 1$ as

$$\begin{aligned} &\frac{y_{\hat{n},j} - y_{\hat{n}-1,j}}{h} - \frac{h}{2}\left(\frac{y_{\hat{n},j+1} - 2y_{\hat{n},j} + y_{\hat{n},j-1}}{h^2}\right) + \frac{h}{2}c_{\hat{n},j}y_{\hat{n},j} = \\ &= g_{\hat{n},j} + \frac{h}{2}f_{\hat{n},j}, \quad j = 1, 2, \ldots, \hat{n} - 1. \end{aligned} \tag{25}$$

- on the part of the boundary with $y = 0$ as

$$\begin{aligned} &\frac{y_{i,0} - y_{i,1}}{h} - \frac{h}{2}\left(\frac{y_{i+1,0} - 2y_{i,0} + y_{i-1,0}}{h^2}\right) + \frac{h}{2}c_{i,0}y_{i,0} = \\ &= g_{i,0} + \frac{h}{2}f_{i,0}, \quad i = 1, 2, \ldots, \hat{n} - 1. \end{aligned} \tag{26}$$

- on the part of the boundary with $y = 1$ as

$$\begin{aligned} &\frac{y_{i,\hat{n}} - y_{i,\hat{n}-1}}{h} - \frac{h}{2}\left(\frac{y_{i+1,\hat{n}} - 2y_{i,\hat{n}} + y_{i-1,\hat{n}}}{h^2}\right) + \frac{h}{2}c_{i,\hat{n}}y_{i,\hat{n}} = \\ &= g_{i,\hat{n}} + \frac{h}{2}f_{i,\hat{n}}, \quad i = 1, 2, \ldots, \hat{n} - 1. \end{aligned} \tag{27}$$

Theorem 6. *The FD discretization* (21), (24)–(27) *has the following properties:*

a) it approximates a sufficiently smooth solution u with the second order of accuracy,
b) the resulting FD matrix $\mathbf{A}$ is SDD and M-matrix,
c) the discrete maximum/minimum principles are valid provided the relevant conditions on g hold.

Proof. We shall prove the statement a) only for the case of the part of the boundary with $x = 1$, because the proofs for the other cases are similar. Clearly, it is sufficient to show the second order of accuracy at the boundary nodes only. Let us define

$$\begin{aligned} \Psi_j = &\frac{u(1, jh) - u(1 - h, jh)}{h} - \\ &- \frac{h}{2}\left(\frac{u(1, (j+1)h) - 2u(1, jh) + u(1, (j-1)h)}{h^2}\right) + \\ &+ \frac{h}{2}c(1, jh)u(1, jh) - g(1, jh) - \frac{h}{2}f(1, jh). \end{aligned} \tag{28}$$

Using the Taylor expansion, we get

$$\frac{u(1, jh) - u(1 - h, jh)}{h} = \left(\partial_1 u - \frac{h}{2}\partial_{11}^2 u\right)\Big|_{(1,jh)} + \mathcal{O}(h^2), \tag{29}$$

$$\frac{u(1, (j+1)h) - 2u(1, jh) + u(1, (j-1)h)}{h^2} = \left(\partial_{22}^2 u\right)\big|_{(1,jh)} + \mathcal{O}(h^2), \tag{30}$$

where symbols like $\partial_i u$ and $\partial_{ij} u$ denote the partial derivatives of u as usual. Hence, putting (29) and (30) into (28), we obtain

$$\Psi_j = \left(\partial_1 u - g\right)\big|_{(1,jh)} - \frac{h}{2}\left(\partial_{11}^2 u + \partial_{22}^2 u - cu + f\right)\big|_{(1,jh)} + \mathcal{O}(h^2). \tag{31}$$

Due to the boundary condition in (1), and the relation $\frac{\partial u}{\partial n}(1,y) = \partial_1 u(1,y)$, the first term in the right-hand side of (31) vanishes. The second term is also equal to zero. This shows the validity of the statement a).

To prove the statement b), it is enough to show the diagonal dominance at the boundary nodes only. For convenience, we introduce the index k to have the single-index numbering of all the nodes of our domain (in order to keep the consistency with the "single-index" definition of $\delta_k(\mathbf{A})$) in which the indices $1, 2, \ldots, n^*$ are preserved for n^* interior nodes and the indices $n^*+1, \ldots, n^*+n^0$ are used for n^0 boundary nodes. Then we have that

$$\delta_k(\mathbf{A}) = \frac{h}{2} c_k > 0 \quad \text{for} \quad k = n^*+1, \ldots, n^*+n^0. \tag{32}$$

Therefore, under our assumptions $\mathbf{A}$ is SDD matrix and M-matrix.

To prove the statement c), one observes that for the right-hand side of the resulting FD system we have

$$F_k = f_k \quad \text{for} \quad k = 1, \ldots, n^*, \text{ and } F_k = g_k + \frac{h}{2} f_k \quad \text{for} \quad k = n^*+1, \ldots, n^*+n^0. \tag{33}$$

Due to the property b), Theorem 3 can now be used. To get estimates (11) and (12), we use the corresponding sign-conditions on g. □

5. Final remarks

It would be interesting to obtain suitable practical conditions guaranteeing the validity of our variant of discrete maximum/minimum principles also for various *hp*-versions of FEM (see [16]), and analyse the case of elliptic problems with full diffusive tensors (cf. [15]).

Acknowledgements

The first author was supported by the Hungarian Research Fund OTKA under grant no. K112157. The second author was partially supported by MINECO under Grant MTM2011-24766.

References

[1] Axelsson, O. and Kolotilina, L.: Monotonicity and discretization error estimates. SIAM J. Numer. Anal. **27** (1990), 1591–1611.

[2] Bramble, J. H. and Hubbard, B. E.: On a finite difference analogue of an elliptic boundary problem which is neither diagonally dominant nor of non-negative type. J. Math. and Phys. **43** (1964), 117–132.

[3] Brandts, J., Korotov, S., Křížek, M., and Šolc, J.: On nonobtuse simplicial partitions. SIAM Rev. **51** (2009), 317–335.

[4] Ciarlet, P. G.: Discrete maximum principle for finite-difference operators. Aequationes Math. **4** (1970), 338–352.

[5] Ciarlet, P. G. and Raviart, P. -A.: Maximum principle and uniform convergence for the finite element method. Comput. Methods Appl. Mech. Engrg. **2** (1973), 17–31.

[6] Faragó, I., Korotov, S., and Szabó, T.: On modifications of continuous and discrete maximum principles for reaction-diffusion problems. Adv. Appl. Math. Mech. **3** (2011), 109–120.

[7] Faragó, I., Korotov, S., and Szabó, T.: On continuous and discrete maximum principles for elliptic problems with the third boundary condition. Appl. Math. Comput. **219** (2013), 7215–7224.

[8] Hannukainen, A., Korotov, S., and Vejchodský, T.: Discrete maximum principle for FE solutions of the diffusion-reaction problem on prismatic meshes. J. Comput. Appl. Math. **226** (2009), 275–287.

[9] Karátson, J. and Korotov, S.: Discrete maximum principles for finite element solutions of nonlinear elliptic problems with mixed boundary conditions. Numer. Math. **99** (2005), 669–698.

[10] Karátson, J. and Korotov, S.: Discrete maximum principles for finite element solutions of some mixed nonlinear elliptic problems using quadratures. J. Comput. Appl. Math. **192** (2006), 75–88.

[11] Karátson, J., Korotov, S., and Křížek, M.: On discrete maximum principles for nonlinear elliptic problems. Math. Comput. Simulation **76** (2007), 99–108.

[12] Korotov, S., Křížek, M., and Neittaanmäki, P.: Weakened acute type condition for tetrahedral triangulations and the discrete maximum principle. Math. Comp. **70** (2001), 107–119.

[13] Křížek, M. and Qun Lin: On diagonal dominance of stiffness matrices in 3D. East-West J. Numer. Math. **3** (1995), 59–69.

[14] Křížek, M. and Neittaanmäki, P.: *Finite element approximation of variational problems and applications*, Longman Scientific & Technical, Harlow, 1990.

[15] Kuzmin, D., Shashkov, M. J., and Svyatskiy, D.: A constrained finite element method satisfying the discrete maximum principle for anisotropic diffusion problems. J. Comput. Phys. **228** (2009), 3448–3463.

[16] Vejchodský, T. and Šolín, P.: Discrete maximum principle for higher-order finite elements in 1D. Math. Comp. **76** (2007), 1833–1846.

[17] Windisch, G.: A maximum principle for systems with diagonally dominant M-matrices. In: E. Adams, R. Ansorge, C. Grossman, and H.-G. Roos (Eds.), *Discretization in Differential Equations and Enclosures*, *Math. Res.*, vol. 36, pp. 243–250, Akademie-Verlag, Berlin, 1987.

Conference Applications of Mathematics 2015, in honor of the birthday anniversaries of Ivo Babuška (90), Milan Práger (85), and Emil Vitásek (85)
J. Brandts, S. Korotov, M. Křížek, K. Segeth, J. Šístek, T. Vejchodský (Eds.)
Institute of Mathematics CAS, Prague 2015

SHOALING OF NONLINEAR STEADY WAVES: MAXIMUM HEIGHT AND ANGLE OF BREAKING

Sebastião Romero Franco[1], Leandro Farina[2,3]

[1] Departamento de Matemática, Universidade Estadual do Centro-Oeste,
Irati, PR, Brazil
romero@irati.unicentro.br
[2]Instituto de Matemática, Universidade Federal do Rio Grande do Sul,
Porto Alegre, RS, 91509-900, Brazil
farina@mat.ufrgs.br
[3]BCAM - Basque Center for Applied Mathematics,
Mazarredo 14, 48009 Bilbao, Basque Country, Spain
lfarina@bcamath.org

Abstract: A Fourier approximation method is used for modeling and simulation of fully nonlinear steady waves. The set of resulting nonlinear equations are solved by Newton's method. The shoaling of waves is simulated based on comparisons with experimental data. The wave heights and the angles of breaking are analysed until the limit of inadequacy of the numerical method. The results appear quite close to those criteria predicted by the theory of completely nonlinear surface waves and contribute to provide information on the study of the relationship between computational modeling and the theory of steady waves.

Keywords: nonlinear water waves, steady waves, wave shoaling, angle of wave breaking, maximum wave height, spectral method

MSC: 74J30, 76B15, 74S25

1. Introduction

Waves in water are natural phenomena which have been extensively studied. The knowledge about its properties is of fundamental importance in several socio-economical activities, such as coastal environment protection, industrial activities in deep waters, where an analysis of the impact and force of waves is of extreme importance. Other not less important activities are applications to sailing, sediment transport prediction and conversion of waves energy into electrical energy.

The study of wave shoaling and breaking has a deserved remarkable place in this context, given that the energy of waves is intrinsically associated to the wave's

height. Experimental, analytical and computational methods have been used for investigation of these phenomena. A certain amount of experimental data about wave shoaling is known. Among these, the field data in [11] and the laboratory measurements in [6] are classical and used for validation of numerical methods. More recently Tsai et al. [17] examined criteria used in wave breaking via experimental results. In their work, steeper bottoms have been studied.

From the analytical and computational points of view, the paper by Rienecker and Fenton [14] has been one of the first work to propose a method for the simulation of steady completely nonlinear water waves. Denominated as Fourier methods, this technique does not assume analytical approximations and the solution of the nonlinear equations for the dynamics of waves is expressed by a Fourier series. The nonlinear equations obtained are resolved numerically by Newton's method. A great number of subsequent papers propose improvements and extension of Fourier methods to the study of nonlinear free surface waves. Gimenez-Curto and Corniero Lera [10] present procedures to reduce the computation time of Fourier methods for very long waves. Assuming Fourier's expansions of superior orders and including nonlinear interactions of arbitrary order, Dommermuth and Yue [3, 18] expanded Fourier methods via a spectral method of superior order and calculated the evolution of nonlinear waves in several cases, including the interaction between two waves.

Approaches with analytical approximations for the calculation of nonlinear waves have also been used [19]. Freilich and Guza [9] use variants of Boussinesq equations to study the shoaling of waves. Fenton [7] deduced expressions of fifth order based on Stoke's theory and presented numerical results, comparing them to experimental data. In this same context, Pihl et. al. [13] examined the shoaling of waves described by an approximation of sixth order in the presence of a current.

For studies of nonlinear waves dynamics with a more computational emphasis, we cite Drimer and Agnon [4], which uses the boundary element method and the work of Bingham and Zhang [1] for an approach of the problem through finite differences of higher order. Finally, Ducrozet et. al. [5] make a comparative study of two fast methods for the problem of nonlinear surface waves: the higher order spectral method and the higher order method of finite differences.

In this paper, we solved the problem of steady completely nonlinear surface waves by Fourier methods combined with Newton's method [2]. No analytical approximation is done and we assume that in a bottom with declivity, waves in any depth behave as if the bottom were horizontal. The approximation by Fourier series showed to be a very powerful tool since it allows the direct calculation of accurate solutions, even for high waves and for every wavelength, except for a soliton's limit. We explored this characteristics to study with a certain level of detail, the phenomenon of wave shoaling.

The mathematical model and the non-dimensionalisation are presented with details in section 2. The approximation used for nonlinear steady waves is described in section 3, where we also present the computational approach. On section 4, the additional modeling and the method to examine the shoaling of waves are examined. In

subsection 5.1, the numerical results obtained are compared with experimental data. The maximum height and angle of wave breaking are examined computationally in subsections 5.2 and 5.3, respectively. Final conclusions are given in section 6.

2. Mathematical model

The mathematical description of the propagation of gravity waves on the water surface usually requires some assumptions about the water properties and the motion performed by it. Thus, we consider a homogeneous, incompressible fluid with non-rotational motion, where the main restoring force is due to the gravitational acceleration. Additionally, the viscosity and the surface tension are neglected. Moreover, we will not consider wind forcing.

We consider two-dimensional steady waves in water of finite depth and formulate the problem in terms of the stream function ψ. In what follows, we will use the same framework as in [14].

We will use symbol $*$ will denote dimensional variables and all variables are non-dimensionalised with respect to the acceleration of gravity g^*, and to the average depth, $\overline{\eta}^*$. Thus, consider the changes of variables $x = \frac{x^*}{\overline{\eta}^*}$, $y = \frac{y^*}{\overline{\eta}^*}$, $\eta = \frac{\eta^*}{\overline{\eta}^*}$, $\psi = \frac{\psi^*}{\sqrt{g^*\overline{\eta}^{*3}}}$, $Q = \frac{Q^*}{\sqrt{g^*\overline{\eta}^{*3}}}$, $R = \frac{R^*}{g^*\overline{\eta}^*}$. The spatial coordinates x and y indicate the horizontal and vertical direction with the origin of the Cartesian system lying at the water bottom. Here, η is the water surface, ψ is the stream function, Q is the volumetric flow rate per unit wavelength normal to the plane xy and R is the total energy of the system.

Other non-dimensional variables relevant to the problem are the wave velocity $c = \frac{c^*}{\sqrt{g^*\overline{\eta}^*}}$, the wavenumber k is defined by $k = k^*\overline{\eta}^* = \frac{2\pi}{\lambda^*}\overline{\eta}^*$, where λ^* is the wavelength, the wave period is given by $\tau = \tau^*\sqrt{\frac{g^*}{\overline{\eta}^*}}$, and the so called arbitrary reference level D, is non-dimensionalised by $D = \frac{D^*}{\overline{\eta}^*}$.

We denote by (u, v) the components of the velocity vector $\mathbf{u}$ and the stream function ψ is defined such that $u = \frac{\partial\psi}{\partial y}$ and $v = -\frac{\partial\psi}{\partial x}$. $\psi(x, y)$ satisfies Laplace's equation

$$\frac{\partial^2\psi}{\partial x^2} + \frac{\partial^2\psi}{\partial y^2} = 0 \qquad \text{in } \ 0 < y < \eta(x). \tag{1}$$

The boundary conditions that must be satisfied by the stream function are

$$\psi(x, 0) = 0, \tag{2}$$

at the origin (background) and

$$\psi(x, \eta(x)) = -Q, \tag{3}$$

on the free surface $y = \eta(x)$.

In equation (3), it is assumed that water flow of moving from right to left is in the negative direction. On the free surface, the pressure is constant so that Bernoulli's equation gives:

$$\frac{1}{2}\left[\left(\frac{\partial\psi}{\partial x}\right)^2+\left(\frac{\partial\psi}{\partial y}\right)^2\right]+\eta=R. \tag{4}$$

The boundary conditions involving the wave periodicity are given by:

$$\lambda=\frac{2\pi}{k}, \tag{5}$$

and

$$\lambda=c\tau. \tag{6}$$

Next we define the contours of conditions, the condition of periodicity and some additional equations involving wave height, volume flow and wave speed, it is possible to obtain a closed system of variables that can be solved by Newton's method. We will describe, in the next section, how to to accomplish this, essentially by expanding the stream function ψ, in Fourier series.

3. Approximation of fully nonlinear steady waves

We present now the problem of fully nonlinear steady waves. The approximation of the solution is obtained by a spectral method combined with Newton's method [2, 14].

We expand $\psi(x,y)$ as

$$\psi(x,y)=B_0y+\sum_{j=1}^{N}B_j\frac{\sinh jky}{\cosh jkD}\cos jkx \tag{7}$$

for the Fourier coefficients B_j. This representation of the stream function assumes symmetry about the wave crest. The description below, in this section, is essentially the one given in [14]. We present some of the details here for completeness.

Note that the above expansion satisfies the Laplace's equation (1) and the boundary condition (2). The boundary condition (3) requires that

$$B_0\eta+\sum_{j=1}^{N}B_j\frac{\sinh jk\eta}{\cosh jkD}\cos jkx=-Q, \tag{8}$$

and the equation (4) takes the form

$$\frac{1}{2}\left[k\sum_{j=1}^{N}jB_j\frac{\sinh jk\eta}{\cosh jkD}\ \sin jkx\right]^2+\frac{1}{2}\left[B_0+k\sum_{j=1}^{N}jB_j\frac{\cosh jk\eta}{\cosh jkD}\ \cos jkx\right]^2+\eta=R, \tag{9}$$

for all x.

In these approximations, we observe that the arguments of $\sinh jk\eta$, $\cosh jk\eta$ and $\cosh jkD$ grow up rapidly with j. To avoid instability and numerical errors in the divisions in (9), we use the approximation

$$\frac{\cosh jk\eta}{\cosh jkD} \sim \frac{\sinh jk\eta}{\cosh jkD} \sim \exp\left[jk(\eta - D)\right], \tag{10}$$

for sufficiently large values of j.

The choice of an appropriate value for the parameter D is important. We will adopt the non-dimensional value $D = 1$, suggested by Rienecker & Fenton [14], which corresponds to a value of relative water depth and characterises a regime of of water intermediate.

We will now impose equations (8) and (9) on $2N$ collocation points over one wavelength. This allows a discretisation of the problem. By symmetry, we can work with only $N+1$ points from the wave crest to the trough. Thus, we use the discretization $x_m = \frac{m\lambda}{2N}, \quad m = 0, 1, \ldots, N$. From $\lambda = \frac{2\pi}{k}$ it follows that $kx_m = \frac{m\pi}{N}$. Moreover, we abbreviate the notation of $\eta(x_m), u(x_m, y_m)$ e $v(x_m, y_m)$ to η_m, u_m and v_m. Thus, from (8) and (9), we have:

$$B_0\eta_m + \sum_{j=1}^{N} B_j \frac{\sinh jk\eta_m}{\cosh jkD} \cos\left(\frac{jm\pi}{N}\right) + Q = 0, \tag{11}$$

$$\frac{1}{2}u_m^2 + \frac{1}{2}v_m^2 + \eta_m - R = 0, \tag{12}$$

for $m = 0, 1, \ldots, N$, where

$$u_m = B_0 + k\sum_{j=1}^{N} jB_j \frac{\cosh jk\eta_m}{\cosh jkD} \cos\left(\frac{jm\pi}{N}\right),$$

$$v_m = k\sum_{j=1}^{N} jB_j \frac{\sinh jk\eta_m}{\cosh jkD} \sin\left(\frac{jm\pi}{N}\right).$$

We now have $2N+2$ nonlinear equations. However, these involve $2N+5$ variables, which are $\eta_j, B_j, (j = 0, 1, \ldots, N)$, k, Q and R. Thus, in principle, we need three further equations.

As the mean non-dimensionalised wave height is unity, we can write

$$\int_S \eta \, dS = 1, \tag{13}$$

where S is the horizontal distance from the crest to the trough of the wave. Discretisation x_0 and x_N represent the abscissas of these extremities and using the trapezoidal rule in (13), we have

$$\frac{1}{2N}\left[\eta_0 + \eta_N + 2\sum_{j=1}^{N-1} \eta_j\right] - 1 = 0. \tag{14}$$

In certain situations can solve the problem of nonlinear waves for prescribed values of the height H and the wave period τ. The height H is merely the difference between the elevation of the crest η_0 and the height of the wave trough η_N. Hence,

$$\eta_0 - \eta_N - H = 0. \tag{15}$$

By combining equations (5) and (6), which involve the wave period, we have,

$$kc\tau - 2\pi = 0. \tag{16}$$

We obtained then, from (14)– (16), three new equations. We introduced, however, a new variable to the system; the wave velocity c. Therefore, let us analyse in more detail this quantity.

Let c_E the Eulerian mean velocity c_E of fluid. For steady waves, we have the relation [14]

$$c - c_E + B_0 = 0. \tag{17}$$

Alternatively, one can consider the drift velocity c_s of the fluid particle, which is the mass transport velocity. In steady wave regime, with unit mean depth, volume flow Q is equal to the mean velocity by which the fluid particle moves. Therefore, the speed of mass transport can be given by

$$c - c_s - Q = 0. \tag{18}$$

Finally, the $2N + 6$ equations (11), (12),(14)–(16), (17) or (18) form a closed system for the variables $(\eta_j, B_j (j = 0, 1, \ldots, N), k, Q, R, c)$.

4. Wave shoaling

The shoaling of waves occurs when they propagate in intermediary waters in a variable depth zone, gradually decreasing. In this study, it is assumed that the changes in depth occur in a smooth way. Thus, it can be assumed that the wave does not reflect and can adapt to the new depth. Due to energy conservation, when the group velocity, C_g, decreases, the wave tends to increase its height, or to shoal, until the subsequent wave break.

By using the wave refraction theory, it can be shown that the wave period is also constant during the process of shoaling. This follows from the conservation of crests for steady waves.

The phenomenon of incident waves shoaling on a coastal region has been well approximated by Rienecker& Fenton [14], assuming that if the bottom's inclination is less than $4,5^\circ$ the wave acts as if it is steady and with a constant local depth. Employing this hypothesis, a simple approximation neglects the dissipation by friction with the bottom and assumes that the wave period and the energy flux remain constant from a depth to another. That is, we assume that the conservation of crests occurs and there is no reflection of energy with decreasing of depth.

4.1. Method

To describe the shoaling of waves, due to the reduction of depth, the system of equations presented in section 3 must be extended to include the additional variables; the wave height H and the average flux of energy F, of which the non-dimensional value can be written as [14]:

$$F = \frac{1}{2}c^3 - \frac{3}{2}c^2Q + c\left(2R - 1 - \frac{1}{2}QB_0 - \overline{\eta^2}\right) - Q(R-1), \tag{19}$$

where

$$\overline{\eta^2} = \frac{1}{2N}\left[\eta_0^2 + \eta_N^2 + 2\sum_{j=1}^{N-1}\eta_j^2\right].$$

The solution of the system from the starting depth provides the flux of energy according to equation (19). We will model the shoaling of waves by using a discrete and finite number of depths. For successive depths, the period and the flux of energy will be preserved, while the wave height H will be the variable of the problem. Thus, we must include in the system, additional equations to specify the wave height for the starting depth and the flux of energy for subsequent depths.

The additional equations are

$$f_{2N+7} = \frac{1}{2}c^3 - \frac{3}{2}c^2Q + c\left(2R - 1 - \frac{1}{2}QB_0 - \overline{\eta^2}\right) - Q(R-1) - F = 0$$

and

$$\begin{aligned} f_{2N+8} &= H - \frac{H_0^*}{\overline{\eta}_0^*} = 0 \quad \text{for the initial depth, and} \\ f_{2N+8} &= F - F_0 = 0 \quad \text{for the subsequent depths,} \end{aligned}$$

where F_0 is the non-dimensional energy flux. We use Newton's method to solve the resulting discrete system.

An initial estimate for the energy flux is given as a function of the other variables. From Stokes approximation, we have

$$F = \frac{\pi}{8}\frac{c^2H^2}{\tau}\frac{\sinh k\cosh k + k}{\sinh^2 k}.$$

This estimate is necessary only for the first depth. Subsequently, for small changes in the depth, the following solution can be used as a good starting approximation for the problem, provided that the change in depth is calculated in the new non-dimensionalisation.

Suppose that the sub-index 1 is the solution for a certain depth and sub-index 2, the starting approximation of the next depth. Thus, the change of depth occurs as follows, $\overline{\eta}_2^* = \overline{\eta}_1^* \cdot r$. That is, $r = \overline{\eta}_2^*/\overline{\eta}_1^*$ is the ratio between the successive depths.

To obtain a satisfactory starting estimation of the variables to be used in the new depth, we assume the change in depths to be smooth. With this, we neglect the reflection of waves and we can use the last solution obtained as a good starting approximation for the next depth, as long as it is non-dimensionalised according to the new depth. That is,

$$H_2 = \frac{H_2^*}{\overline{\eta}_2^*} = \frac{H_1^*}{\overline{\eta}_2^*} = \frac{H_1\overline{\eta}_1^*}{\overline{\eta}_2^*} \implies H_2 = \frac{H_1}{r},$$

$$c_2 = \frac{c_2^*}{(g\overline{\eta}_2^*)^{\frac{1}{2}}} = \frac{c_1^*}{(g\overline{\eta}_2^*)^{\frac{1}{2}}} = \frac{c_1\,(g\overline{\eta}_1^*)^{\frac{1}{2}}}{(g\overline{\eta}_2^*)^{\frac{1}{2}}} \implies c_2 = \frac{c_1}{r^{\frac{1}{2}}},$$

$$k_2 = k_2^*\overline{\eta}_2^* = k_1^*\overline{\eta}_2^* = \frac{k_1}{\overline{\eta}_1^*}\overline{\eta}_2^* \implies k_2 = k_1 r,$$

$$Q_2 = \frac{Q_2^*}{[g(\overline{\eta}_2^*)^3]^{\frac{1}{2}}} = \frac{Q_1^*}{[g(\overline{\eta}_2^*)^3]^{\frac{1}{2}}} = \frac{Q_1\,[g(\overline{\eta}_1^*)^3]^{\frac{1}{2}}}{[g(\overline{\eta}_2^*)^3]^{\frac{1}{2}}} \implies Q_2 = \frac{Q_1}{r^{\frac{1}{2}}},$$

$$R_2 = 1 + \frac{R_2^*}{g\overline{\eta}_2^*} = 1 + \frac{R_1^*}{g\overline{\eta}_2^*} = 1 + \frac{(R_1-1)g\overline{\eta}_1^*}{g\overline{\eta}_2^*} \implies R_2 = 1 + \frac{R_1-1}{r}.$$

For the next non-dimensionalisations, the spatial discretisation is required, indicated by j, for $j = 1, 2, \ldots, N$. We will have

$$(B_j)_2 = \frac{(B_j^*)_2}{g\overline{\eta}_2^*} = \frac{(B_j^*)_1}{g\overline{\eta}_2^*} = \frac{(B_j^*)_1 g\overline{\eta}_1^*}{g\overline{\eta}_2^*} \implies (B_j)_2 = \frac{(B_j)_1}{r^{\frac{1}{2}}}.$$

As the origin of the system is at the water bottom, the non-dimensional form of the free surface elevation is expressed as

$$(\eta_j)_2 = 1 + \frac{(\eta_j^*)_2}{\overline{\eta}_2^*} = 1 + \frac{(\eta_j^*)_1}{\overline{\eta}_2^*} = 1 + \frac{[(\eta_j)_1 - 1]\overline{\eta}_1^*}{\overline{\eta}_2^*} \implies (\eta_j)_2 = 1 + \frac{(\eta_j)_1 - 1}{r}.$$

Yet, despite these remain constant, the non-dimensional energy flow and wave period are

$$F_2 = \frac{F_2^*}{\rho\,[g^3(\overline{\eta}_2^*)^5]^{\frac{1}{2}}} = \frac{F_1^*}{\rho\,[g^3(\overline{\eta}_2^*)^5]^{\frac{1}{2}}} = \frac{F_1\,\rho\,[g^3(\overline{\eta}_1^*)^5]^{\frac{1}{2}}}{\rho\,[g^3(\overline{\eta}_2^*)^5]^{\frac{1}{2}}} \implies F_2 = \frac{F_1}{r^{\frac{5}{2}}},$$

$$\tau_2 = \tau_2^*\left(\frac{g}{\overline{\eta}_2^*}\right)^{\frac{1}{2}} = \tau_1^*\left(\frac{g}{\overline{\eta}_2^*}\right)^{\frac{1}{2}} = \tau_1\left(\frac{g\overline{\eta}_1^*}{g\overline{\eta}_2^*}\right)^{\frac{1}{2}} \implies \tau_2 = \frac{\tau_1}{r^{\frac{1}{2}}}.$$

5. Results

On this section we will show the results obtained. These are organised in 10 cases on which we have experimental data for comparison. These cases, described in detail below, are referenced as waves 1 to 3 and from 4(a) to 4(g).

5.1. Comparison with experiments

For comparison with the experiments, we will use wave data obtained in beaches and in testing tanks. These experiments were originally related in Hansen e Svendsen [11] and in Eagleson [6] and were used in comparisons with other methods and theories. See, for example [14] and [16].

The experimental data were obtained from uniform beach slope of 1/35 [11] and from slope of 1/15 [6] in laboratory tanks. The data were collected until the wave breaking, point in which the modeling presented in this paper is no longer applicable.

Wave	H_0	τ_0	H_0^* (mm)	τ_0^* (s)
1	0,31	5,72	93	1
2	0,13	9,55	39	1,67
3	0,14	19,04	42	3,33

Table 1: Initial values of heights and periods for waves 1 to 3. Experimental data of Hansen e Svendsen [11].

For the first simulations we used the parameters $D = 1$, the number N of terms for the Fourier's expansions in (11) and (12) equals to 16 and $r = 0,999$. In table 1, we summarised the cases of waves 1 to 3 which we are now going to examine with the simulations done with the present method.

In figures 1(a) and 1(b) we show, respectively, the wave height and its phase velocity as a function of depth, for wave 1. With non-dimensional values of the initial height and the initial period, given respectively, by $H_0 = 0,31$ and $\tau_0 = 5,72$, an excellent agreement between the simulation and the experimental data is visually observed, before the wave breaks.

Wave 2 is shorter and has a greater initial period. The comparison for this case is represented in figure 2.

In figure 3, similar comparisons are done for wave 3 which presents a significantly greater period than the former ones. Again, we observe that the simulations present a good agreement with the experimental data. Particularly, the shoaling of waves is remarkable, with the decrease of the depth, in all cases until values very close to the point of break of the wave.

In the following cases, we show the comparison with experimental data [6], obtained in a wave tank with uniform slope of 1/15. Seven simulations are reported, where the dimensional parameters that define them are in table 2. In this table it is also shown the values of initial waves steepness, which is given by $\varepsilon = \frac{H^*}{\lambda^*}$, according to Eagleson [6] and according to our numerical simulations. The reference level is $D = 1$ in all cases except in case (f), where $D = 0,9$. This difference is due to the wave height being a little smaller, in this case. The number of terms on the Fourier expansions used for the next simulations is $N = 32$.

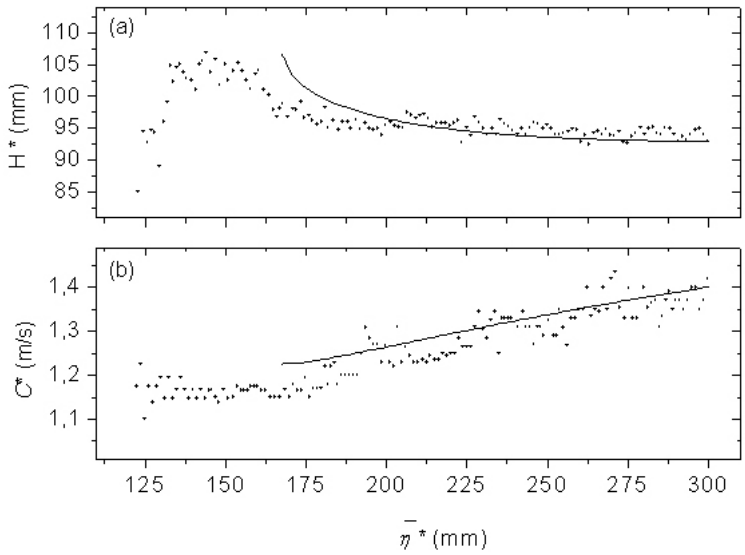


Figure 1: (a) Wave height as a function of water depth. (b) Phase velocity of the wave function of depth for wave 1. The solid line indicates the data obtained from numerical simulations with the present method and the points indicate the data obtained experimentally by [11].

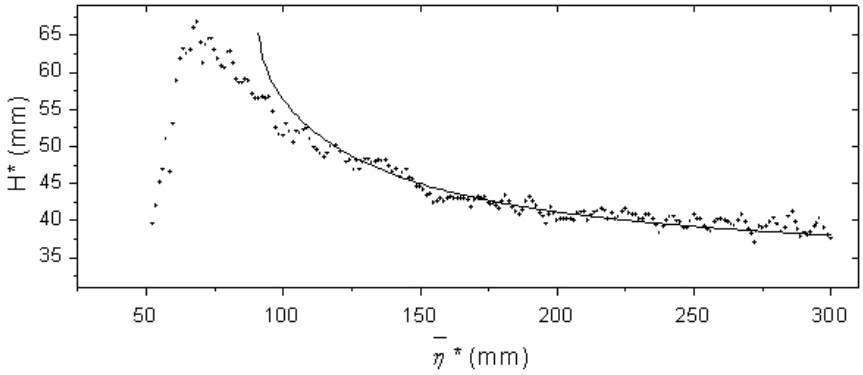


Figure 2: Wave height as a function of water depth for wave 2. The solid line indicates the data obtained from numerical simulations with the present method and the points indicate the data obtained experimentally by [11].

Figure 4 shows the wave shoaling coefficient, given by $\frac{H}{H_0}$, as a function of the respective relative water depth for waves 4(a) to 4(g). This coefficient represents only the relation of the wave height with the decrease of the depth, while the relative depth, indicates if the wave is in shallow, intermediate or deep waters. In all 7 cases, we had an intermediate water regime according to the ratio $0,05 > \frac{\overline{\eta}^*}{\lambda^*} < 0,5$. We observed on this regime a simulated shoaling very close to reality.

On next subsections we analyses with detail the height and the shape of the waves close to their break, using the computational tool we developed and validated here.

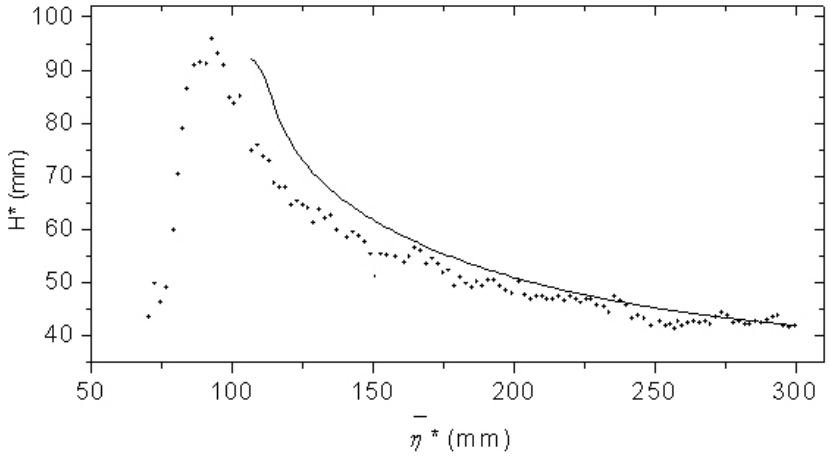


Figure 3: Wave height as a function of water depth for wave 3. The solid line indicates the data obtained from numerical simulations with the present method and the points indicate the data obtained experimentally by [11].

Wave	$\overline{\eta}_0^*$ (feet)	H_0^* (feet)	τ_0^* (s)	H_0^*/λ_0^* (according to [6])	H_0^*/λ_0^* (simulated)
4(a)	1,75	0,230	0,938	0,0528	0,051739
4(b)	1,75	0,234	1,101	0,0396	0,039545
4(c)	1,75	0,357	1,105	0,0598	0,059963
4(d)	1,75	0,440	1,235	0,0611	0,061695
4(e)	1,75	0,354	1,389	0,0420	0,041634
4(f)	1,75	0,186	1,428	0,0209	0,021037
4(g)	1,75	0,265	1,684	0,0237	0,023999

Table 2: Initial values of water depth, the height, the period and the slope according to the experimental data of [6] and the slope calculated by the present method.

5.2. Breaking height

Waves propagating in the shoaling zone, in intermediate waters, become unstable and break when the velocity of the water particle on the wave crest becomes equal or greater than the phase velocity of the wave. At breaking, the wave height is limited by the depth and the wavelength. For a given depth and wave period, there is a maximum limit for the wave height, called *wave breaking height*. According to Stoke's theory, in intermediate waters, the breaking height is $\frac{H}{\overline{\eta}} = 0,78$ [15, p. 06].

In our model, $\eta(x)$ is by definition only defined for each x. Therefore, the method used for the solution will not apply until the physical limit of the wave break. Before the break, the wave surface becomes multivalued and thus not modelled by a function. In the specific case of Newton's method, it will diverge.

We defined as *computational wave breaking height* and denoted by $\left(\frac{H^*}{\overline{\eta}^*}\right)_b$, the

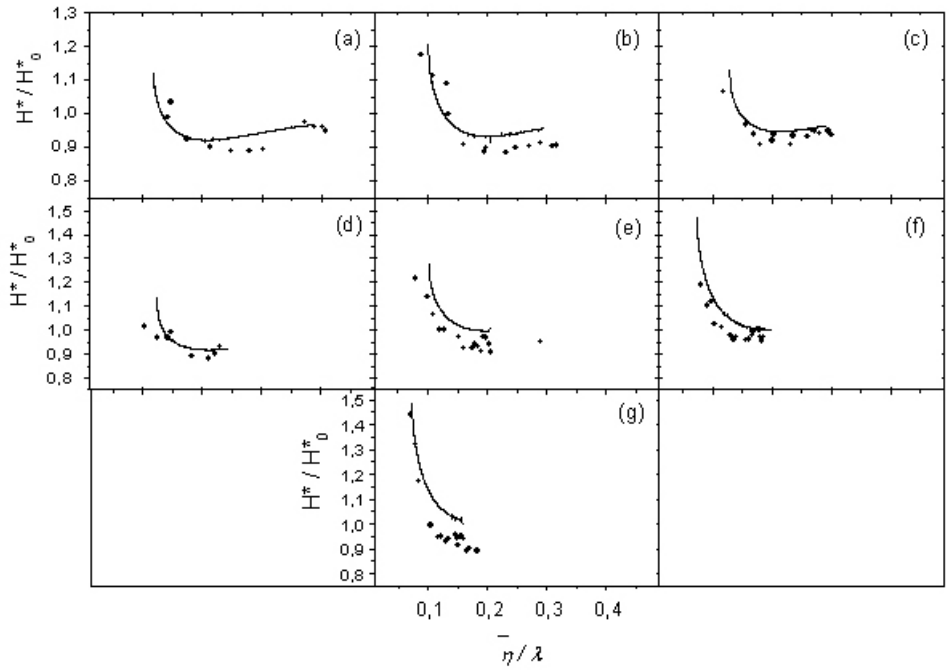


Figure 4: Wave's shoaling coefficient a function on the depth of water to the waves 4 (a) to 4 (g) (see table 2). The solid line indicates the data obtained from numerical simulations with the present method and the points indicate the data obtained experimentally by [6].

last height for which there has been convergence of Newton's method described in subsection 4.1.

Figure 5 shows the evolution of parameter height by depth, given by $H^*/\overline{\eta}^*$, the computation wave breaking height $\left(\frac{H^*}{\overline{\eta}^*}\right)_b$ as a function of parameter $\overline{\eta}/\lambda$, and the depth relative to the wavelength. In this figure, cases of waves 1, 2 and 3 are shown.

On curve 5(a) with height and initial period $H_0 = 0,31$ and $\tau_0 = 5,72$ respectively, the initial relative depth is $\frac{\overline{\eta}_0}{\lambda_0} = 0,214$ and the relative depth on the wave break is $\frac{\overline{\eta}}{\lambda} = 0,13256$. This indicates that the whole shoaling process until the break of the wave happened in intermediate waters.

On curve 5(b), representing wave 2, the initial relative depth is $\frac{\overline{\eta}_0}{\lambda_0} = 0,1128$ and the relative depth on the wave break is $\frac{\overline{\eta}}{\lambda} = 0,0516$. This wave also had the process of shoaling and breaking in intermediate waters, but it breaks practically in shallow waters and with a greater height.

The wave represented on curve 5(c) is significantly longer and presents practically all of its shoaling process in shallow waters, breaking with $\left(\frac{H^*}{\overline{\eta}^*}\right)_b = 0,755$ on a relative depth of $\frac{\overline{\eta}}{\lambda} \approx 0,027$. We observed that the model and the computational method predict a breaking height very close to the observed experimentally.

The cases referred to waves 4(a) to 4(g) on table 2 are represented and summarised

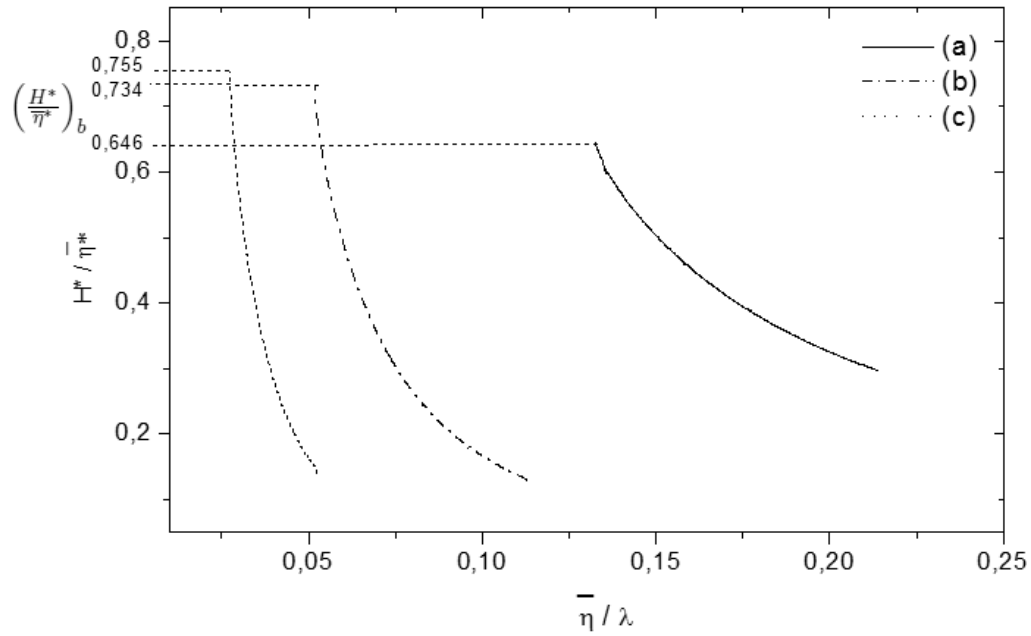


Figure 5: The evolution of the parameter $H^*/\overline{\eta}^*$ and computational wave's breaking height, $\left(\frac{H^*}{\overline{\eta}^*}\right)_b$ a function of depth relative to the wavelength, $\overline{\eta}/\lambda$. In figure (a), we show the results referring to figure 1, in figure (b), referring to figure 2 e in figure (c), referring to figure 3.

in figure 6. We can verify that all the shoaling processes until the break of the waves, occur in intermediate waters. Furthermore, we see that the computational breaking heights are so that $\frac{H}{\overline{\eta}} \approx 0,7$. This is a value that reaffirms the good performance of the method to model the phenomenon of wave shoaling.

5.3. Waves profile and their breaking angles

By using Stokes theory, it can be shown [12] that the breaking angle of a wave is 120°. We are going to use the spectral method described in this paper to estimate the *computational breaking angle* α, i.e., the one until when we can obtain convergence of Newton's method used for solving the system of nonlinear equations which governs the water waves.

Figures 7 to 9 represent the cases of waves 1 to 3, respectively.

Figures are double: part (I) shows the wave profile on the initial instant and part (II), at the moment of the computational wave break. A horizontal straight line is included in all figures to represent the average depth. To estimate the value α, we used a straight line passing by three points next to the crest, using symmetry, and calculated the line's angular coefficient.

It can be verified that the way the wave shoals depends directly on the wavelength. For waves of greater length, we note that the wave's trough becomes horizontally longer. With this, the crest has a more pronounced increase on the wavelength.

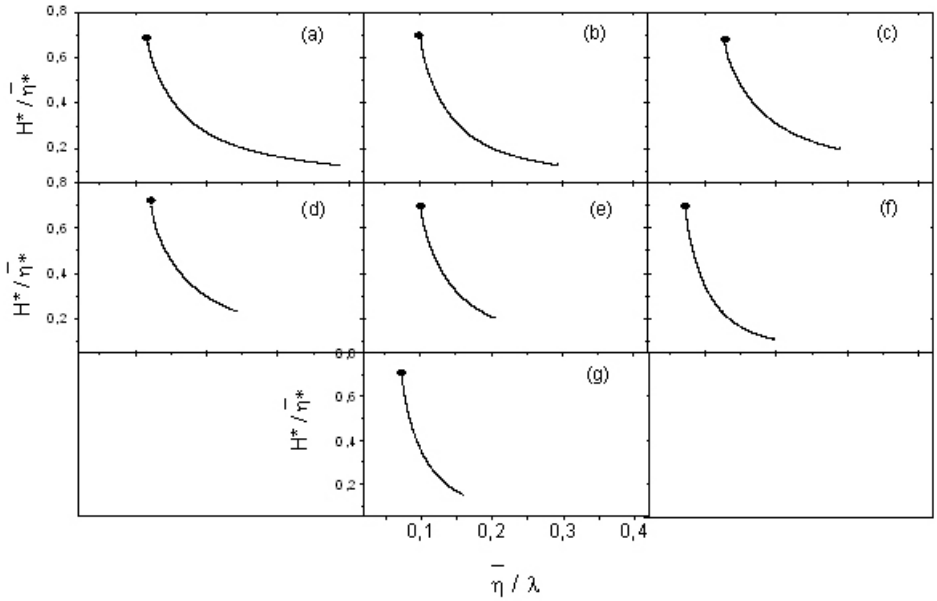


Figure 6: The evolution of the parameter $H^*/\overline{\eta}^*$ and computational wave's breaking height, $\left(\frac{H^*}{\overline{\eta}^*}\right)_b$ a function of the relative depth of water referring to initial data contained in table 2 and also shown in figure 4.

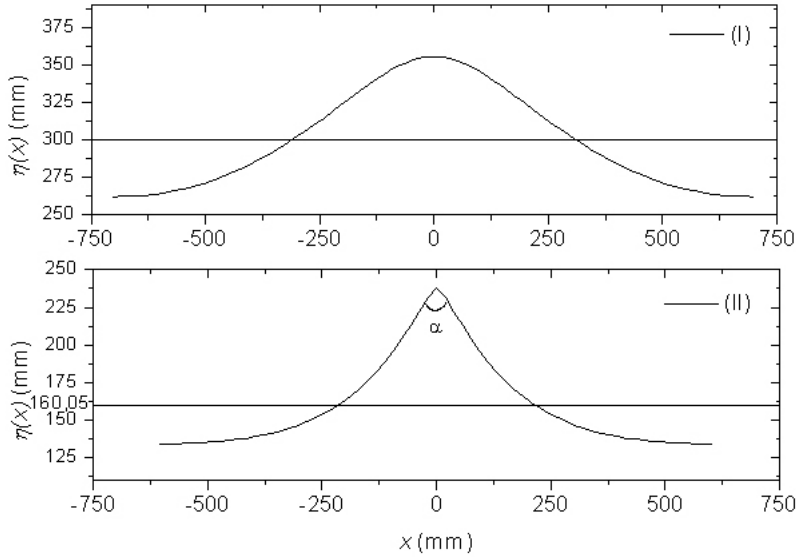


Figure 7: Wave profile: The figure (I) shows the wave profile at the initial moment, with $\overline{\eta}_0^* = 300$ mm, $H_0^* = 93$ mm and $\tau_0^* = 1,0$ s. The figure (II) shows the wave profile at the moment of the break, In this case, the depth of water is indicated by straight line $\overline{\eta}^* = 160,05$ mm.

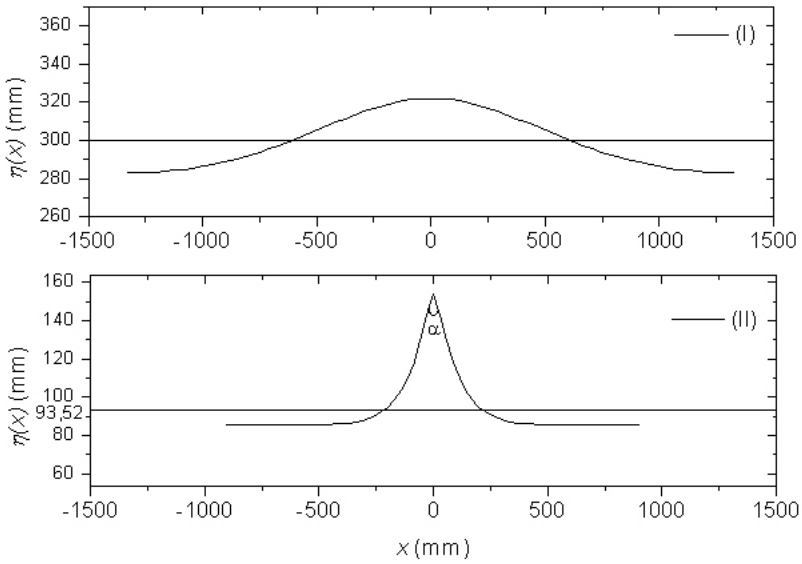


Figure 8: Wave profile: Figure (I) shows the wave profile at the initial moment, with $\overline{\eta}_0^* = 300$ mm, $H_0^* = 39$ mm and $\tau_0^* = 1,67$ s. The figure (II) shows the wave profile at the moment of breaking, In this case, the depth of water is indicated by straight line is $\overline{\eta}^* = 93,52$ mm.

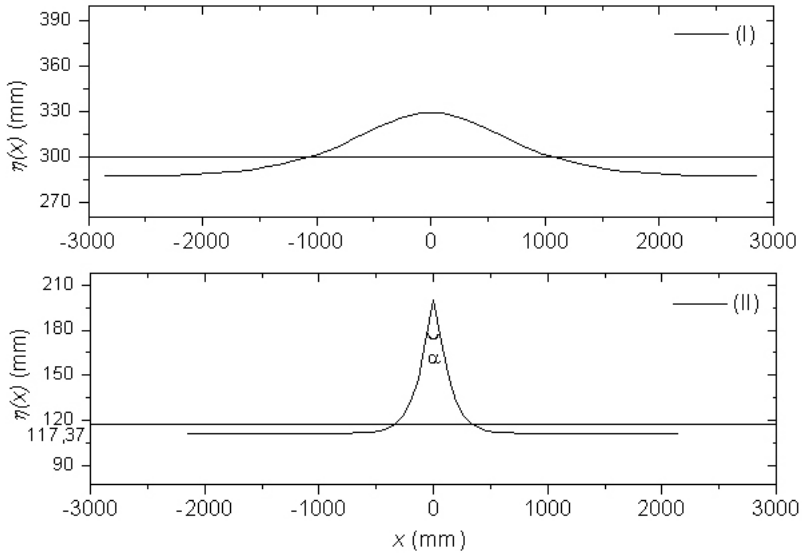


Figure 9: Wave profile: Figure (I) shows the wave profile at the initial moment, with $\overline{\eta}_0^* = 300$ mm, $H_0^* = 42$ mm and $\tau_0^* = 3,33$ s. The figure (II) shows the wave profile at the moment of breaking, In this case, the depth of water is indicated by straight line is $\overline{\eta}^* = 117,37$ mm.

We can observe that the computational breaking angles of waves 1, 2 and 3 were $134,12°, 142,64°$ and $126,20°$ respectively. The first two are in intermediate waters, while wave 3, which propagates in shallow waters, has the smaller breaking angle.

We then revisited the cases of waves 4(a)-4(g), presented in subsection 5.1 and showed at table 2. Table 3 shows the heights and periods of waves at the initial moment, apart from the water depths and computational breaking angles. The initial depth is equal to $\overline{\eta}_0^* = 1,75$ feet, in all cases.

Wave	H_0^* (feet)	τ_0^* (s)	$\overline{\eta}_f^*$(feet)	α (degrees)
4(a)	0,230	0,938	0,3817	124,14
4(b)	0,234	1,101	0,4110	123,60
4(c)	0,357	1,105	0.6066	122,48
4(d)	0,440	1,235	0,7140	119,30
4(e)	0,354	1,389	0,6591	124,48
4(f)	0,186	1,428	0,3989	130,49
4(g)	0,265	1,684	0,5644	128,58

Table 3: Data for the wave profiles. The values of the initial heights and initial wave periods as well as the water depths and angles formed on the crests of the waves at the moment when they break.

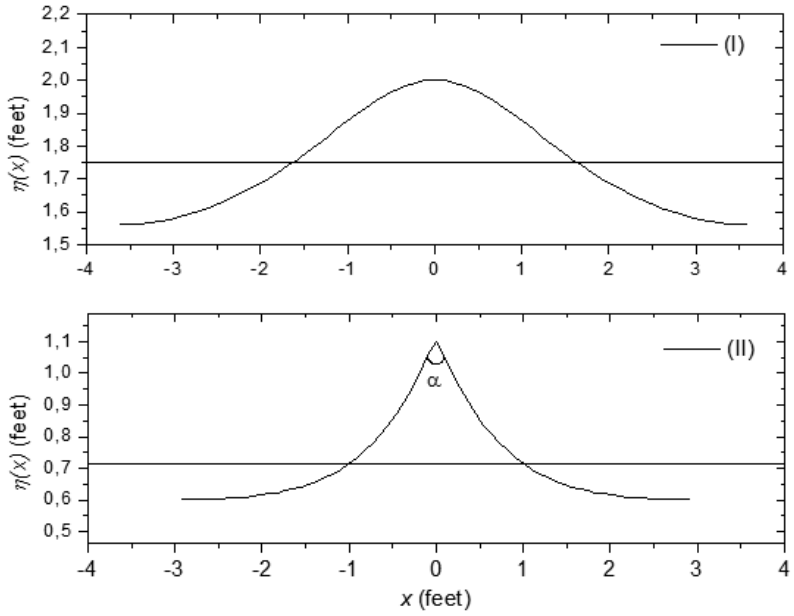


Figure 10: Profile of waves of the case (d) of table 3: Figure (I) shows the wave profile at the initial moment and the figure (II) shows the wave profile at the moment of the break.

To calculate the breaking angles in all cases given at table 3, tree points subsequent to the wave crest were used. We notice that all breaking angles were practically identical and slightly greater than the limit for the breaking angle given in the literature.

Figure 10 shows the profile of the waves of case 4(d) at table 3. The remaining cases have a similar graphical aspect.

Thus as waves 1, 2 and 3, waves 4(a)-4(g) present a common characteristic of shoaling which is the decrease in the wavelength and an increase in its height, with the trough becoming horizontally longer. This is the eminent and favourable aspect to the wave break.

6. Conclusion

A Fourier approximation method was employed for modeling and simulating fully nonlinear steady water waves. The resulting set of nonlinear equations was solved by Newton's method. After a careful non-dimensionality, we assumed that in an inclined bottom, the waves, in any depth, behave as in horizontal bottoms. An iterative method was described for the study of wave shoaling.

A set of experimental data was used to define the initial states in 10 study cases. From those, we could validate the method which presented excellent agreement with the measurements. An analysis of the so called wave breaking height and computational breaking angle was done and values were obtained for comparison between simulations and the theoretical criteria of breaking height and angle. These results, therefore contribute to the knowledge of existing relationships between analytical-computational approximation methods and the theory of nonlinear surface waves.

Acknowledgements

The first author carried out part of work with a grant by CAPES and the second author, as a member of the EU project FP7-295217 - HPC-GA. His research was supported by Grant MTM2011-24766 of the MICINN, Spain and also by the Basque Government through the BERC 2014-2017 program and by Spanish Ministry of Economy and Competitiveness MINECO: BCAM Severo Ochoa excellence accreditation SEV-2013-0323.

References

[1] Bingham, H. and Zhang, H.: On the accuracy of finite-difference solutions for nonlinear water waves. J. of Engng Math. **58**, (2007), 211–228.

[2] Burden, R. L. and Faires, J. D.: *Numerical analysis*. Brooks-Cole Publishing, 2004.

[3] Dommermuth, D. G. and Yue, D. K. P.: A high-order spectral method for the study of nonlinear gravity waves. J. Fluid Mech. **184** (1987), 267–288.

[4] Drimer, N. and Agnon, Y.: An improved low-order bounday element method for breaking surface waves. Journal of Wave Motion **43** (2006), 241–258.

[5] Ducrozet, G., Bingham, H. B., Engsig-Karup, A. P., Bonnefoy, F., and Ferrant, P.: A comparative study of two fast nonlinear free-surface water waves models. Int. J. Num. Methods in Fluids **69** (2012), 1818–1834.

[6] Eagleson, P. S.: Properties of shoaling waves by theory and experiment. Transactions, American Geophysical Union **37** (1956), 565–572.

[7] Fenton, J. D.: A fifth-order Stokes theory for stead waves. J. Waterway, Port, Coastal, Ocean Engng. **111** (1985), 216–234.

[8] Fenton, J. D.: The numerical solution of stedy water wave problems. Computers and Geosciences **14** (3) (1988), 357–368.

[9] Freilich, M. H. and Guza, R. T.: Nonlinear effects on shoaling surface gravity waves. Phil. Trans. R. Soc. Lond. A **311** (1984), 1–41.

[10] Giménez-Curto, L. A. and Corniero Lera, M. A.: Application of Fourier methods to water waves in small depths. Applied Ocean Res. **18** (1995), 275–281.

[11] Hansen, J. B. and Svendsen, I. A.: Regular wave in shoaling water: experimental data. *Inst. of Hydrodynamics and Hydraulic Engng. Tech. Univ. of Denmark*, series paper No. 21 , 1979.

[12] Kinsman, B.: *Wind waves - their generation and propagation on the ocean surface*. Prentice-Hall, Inc.: New Jersey, 1965.

[13] Pihl, J. H., Bredmose H., and Larsen, J.: Shoaling of sixth-order waves on a current. Ocean Engng. **28** (2001), 667–687.

[14] Rienecker, M. M. and Fenton, J. D.: A Fourier approximation method for steady water waves. J. Fluid Mech. **104** (1981), 119–137.

[15] Secretariat of the World Meteorological Organization. Guide to Wave Analysis and Forecasting, No. 702, Geneva, Switzerland, 1998.

[16] Stiassnie, M. and Peregrine, D. H.: Shoaling of finite-amplitude surface waves on water of slowly-varying depth. J. Fluid Mech. **97** (1980), 783–805.

[17] Tsai, C., Chen, H., Hwung, H., and Huang, M.: Examination of empirical formulas for wave shoaling and braking on steep slopes. Ocean Engineering **32** (2005), 469–483.

[18] Tsai, W. and Yue, D. K. P.: Computation of nonlinear free-surface flows. Annual Reviews Fluids Mechanic **28** (1996), 249–278.

[19] Whitham, G. B.: *Linear and nonlinear waves*. John Wiley, New York, 1974.

Conference Applications of Mathematics 2015, in honor of the birthday anniversaries of Ivo Babuška (90), Milan Práger (85), and Emil Vitásek (85)
J. Brandts, S. Korotov, M. Křížek, K. Segeth, J. Šístek, T. Vejchodský (Eds.)
Institute of Mathematics CAS, Prague 2015

NUMERICAL ANALYSIS OF A LUMPED PARAMETER FRICTION MODEL

Vladimír Janovský

Department of Numerical Mathematics, Charles University, Prague,
Sokolovská 83, 186 75 Prague 8, Czech Republic
janovsky@karlin.mff.cuni.cz

Abstract: We consider a contact problem of planar elastic bodies. We adopt Coulomb friction as (an implicitly defined) constitutive law. We will investigate highly simplified lumped parameter models where the contact boundary consists of just one point. In particular, we consider the relevant *static* and *dynamic* problems. We are interested in numerical solution of both problems. Even though the static and dynamic problems are qualitatively different, they can be solved by similar piecewise-smooth *continuation* techniques. We will discuss possible generalizations in order to tackle more complex structures.

Keywords: lumped parameter systems, nonlinear vibrations, Filippov systems, Coulomb friction, impact mechanics

MSC: 65P40, 37M05, 74H15

1. Introduction

Let us consider elastic two-dimensional bodies in mutual contact. The relevant mathematical description consists in modeling of both non-penetration conditions and a friction law. The widely accepted Coulomb friction law represents a serious mathematical and numerical problem. We adopt a discretization via (mixed) Finite Element Method (FEM). The key parameters are degrees of freedom and the number of nodes on the contact boundary. The problems depend on a positive parameter called friction coefficient $\mathcal{F}$.

We have in mind numerical solution of both

1. the *static*, parameter dependent contact problems with Coulomb friction, see e.g. [7, 6, 4, 11, 5],
2. the *dynamic* (i.e. time dependent) contact problems with a friction, see e.g. [9] and with Coulomb friction, [10].

The dynamic solvers use time-stepping schemes (with a fixed stepsize). As a rule, the schemes have to be stabilized. The above authors advocate the stabilization via a *mass redistribution*.

In this contribution we consider a case-study problem with just *one point* on the contact boundary. We analyze both static and dynamic formulations, see [7] and [10]. You may think of toy-problems (lumped parameter models) which reflect the reality qualitatively.

The plan is as follows: In Section 2, we consider the static problem (both the case-study and the example of a real structure). The problem is parameter-dependent in order to model a continuous evolution. The natural numerical tools are continuation (path-following) techniques. The underlying message is: If we learn to solve the toy-problem we get important clues for solving large scale problems. In Section 3 we formulate the dynamic case-study problem. We discuss two numerical techniques: An event-driven algorithm (Section 4) and a time-stepping algorithm (Section 5). In Conclusions (Section 6), we hint at the fact that continuation techniques (Section 2) and, because time is also a parameter, event-driven algorithms and time-stepping algorithms (Section 4 and Section 5) are closely related.

2. The static problem

As a case study, we consider a *static* finite element model of Coulomb friction with one contact point, see [7]: Find $(u_\nu, u_\tau, \lambda_\nu, \lambda_\tau)^T \in \mathbb{R}^4$

$$\left\{\begin{array}{rcl} bu_\nu - cu_\tau - f_\nu - \lambda_\nu &=& 0\,, \\ -cu_\nu + bu_\tau - f_\tau - \lambda_\tau &=& 0\,, \\ \lambda_\nu - P_{(-\infty,0]}(\lambda_\nu - ru_\nu) &=& 0\,, \\ \lambda_\tau - P_{[-\mathcal{F}|\lambda_\nu|,\mathcal{F}|\lambda_\nu|]}(\lambda_\tau - ru_\tau) &=& 0\,. \end{array}\right. \tag{1}$$

Parameters of the model are as follows: The nonnegative friction coefficient $\mathcal{F}$, and the stiffness matrix $\mathbf{A}$,

$$\mathbf{A} = \begin{bmatrix} b & c \\ c & b \end{bmatrix}, \quad b = -\frac{\lambda + 3\nu}{2}\,, \quad c = \frac{\lambda + \nu}{2}\,,$$

where λ and ν are positive parameters (Lamé coefficients). The operators $P_{(-\infty,0]}$ and $P_{[-\mathcal{F}|\lambda_\nu|,\mathcal{F}|\lambda_\nu|]}$ are piecewise linear projectors, see Figure 1. The arguments of both projectors depend on a positive parameter r, that can be arbitrary but fixed.

The system (1) models one linear finite element which rests on a rigid foundation, see Figure 2. The problem is as follows: Given a load $\mathbf{f} = (f_\nu, f_\tau)^\top \in \mathbb{R}^2$, the normal and the tangential *load* components, find

- u_ν and u_τ i.e., the normal and the tangential *displacement*
- λ_ν and λ_τ i.e., the normal and the tangential *stress* components.

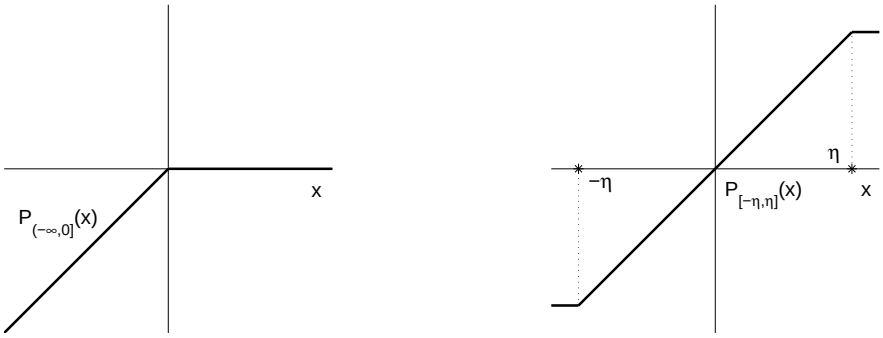


Figure 1: Projectors $x \longmapsto P_{(-\infty,0]}(x)$, $x \longmapsto P_{[-\eta,\eta]}(x)$, $\eta = \mathcal{F}|\lambda_\nu|$.

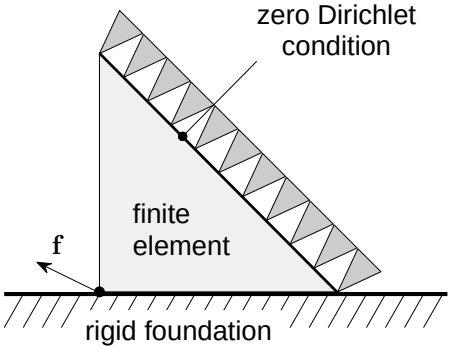


Figure 2: FEM interpretation.

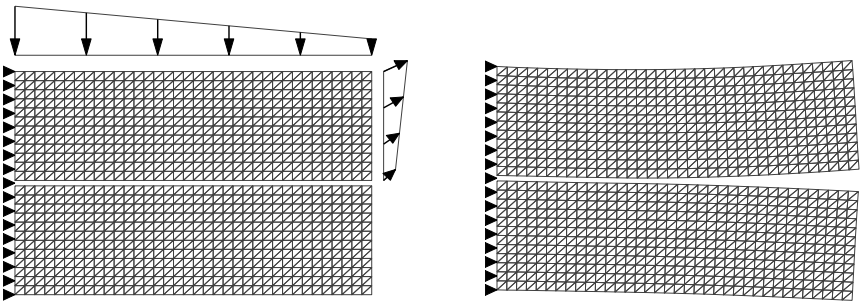


Figure 3: Contact of two elastic bodies Ω^1 (the upper body) and Ω^2, along the contact boundary. The loading is due to the surface traction. Discretization: $n = 1320$ (degrees od freedom), $m = 30$ (number of nodes on the contact boundary). On the right: Resulting deformation.

The system (1) is solvable for any given load $\mathbf{f} \in \mathbb{R}^2$ nevertheless the solution may not be unique. In [6], we proposed path following techniques to find non-unique solutions. The aim was to investigate the model (1) subject to a parameter-dependent force i.e., $\alpha \mapsto f_\nu(\alpha)$ and $\alpha \mapsto f_\tau(\alpha)$. We developed a numerical technique based on *piecewise-smooth continuation*. Starting from this comparatively simple model (1) we generalized the continuation technique for problems of practical interest that involve several thousands elements, see [4, 5]. We also refer to [11] for an alternative approach.

Just to illustrate the technique, we consider the example formulated in [4], see Figure 3. The aim is to investigate dependence of this particular contact problem on the friction coefficient $\mathcal{F}$. The relevant continuation technique is described in [5]. For an illustration of this new technique see Figure 4 and Figure 5. Note that there are three basic contact modes: `no contact`, `contact-stick` and `contact-slip`, see e.g. [6, 4].

3. The dynamic problem

As a case study, we consider a *dynamic* finite element model of Coulomb friction with one contact point, see [10] and Figure 2: We seek for time-dependent functions u_ν, u_τ, λ_ν, $\lambda_\tau : [0, T] \to \mathbb{R}$ such that

$$\mathbf{M} \begin{bmatrix} u_\nu''(t) \\ u_\tau''(t) \end{bmatrix} = \mathbf{A} \begin{bmatrix} u_\nu(t) \\ u_\tau(t) \end{bmatrix} + \begin{bmatrix} f_\nu(t) \\ f_\tau(t) \end{bmatrix} + \begin{bmatrix} \lambda_\nu(t) \\ \lambda_\tau(t) \end{bmatrix} \tag{2}$$

$$-\lambda_\nu(t) \in \mathrm{N}_{\mathbb{R}^1_-}\, u_\nu(t) \tag{3}$$

$$\lambda_\tau(t) \in \mathcal{F}\, \lambda_\nu(t)\, \mathrm{Sign}\, u_\tau'(t) \tag{4}$$

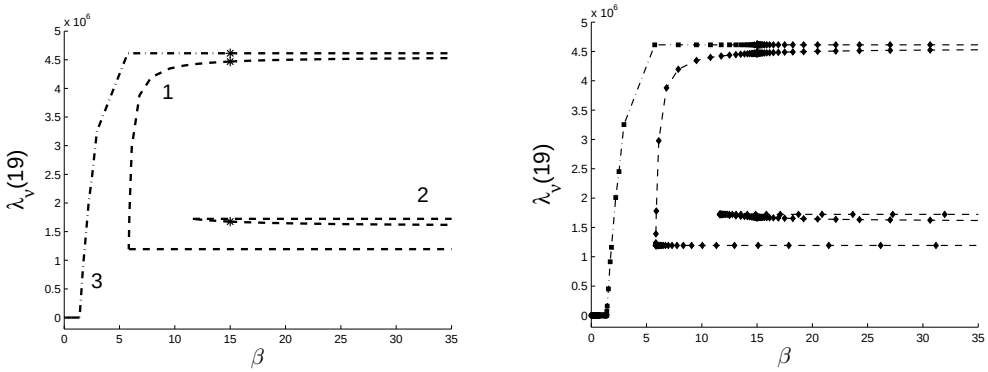


Figure 4: The solution path related to the nodal point No19 consists of three branches. They are initialized by points marked by asterisks. Parameter is $\beta = \mathcal{F}$, the friction coefficient. On the right: An illustrations of the adaptive stepsize refinement of the algorithm. The curves interpretations: solid (no contact), dashed (contact-stick) and dash-dotted (contact-slip).

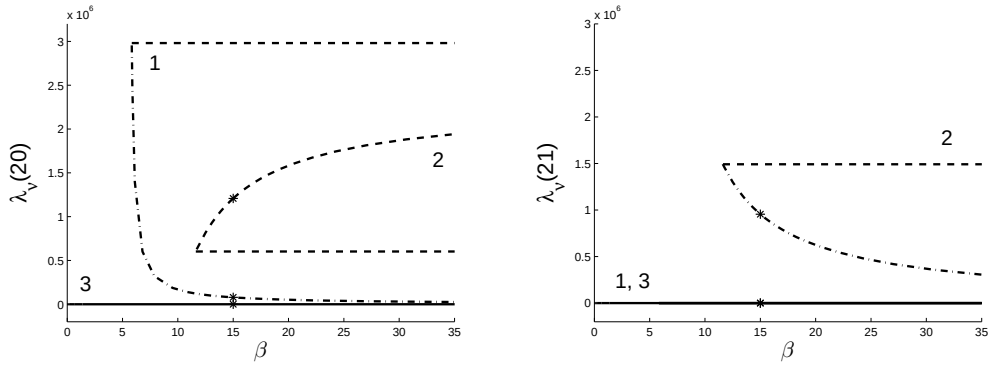


Figure 5: The solution path related to the nodal point No20 (on the left) and the nodal point No21 (on the right). Parameter: $\beta = \mathcal{F}$. The curves interpretations: solid (no contact), dashed (contact-stick) and dash-dotted (contact-slip).

almost everywhere (a.e.) in $[0, T]$. The initial value condition

$$\begin{bmatrix} u_\nu(0) \\ u_\tau(0) \end{bmatrix} = \mathbf{u}^0\,, \quad \begin{bmatrix} u'_\nu(0) \\ u'_\tau(0) \end{bmatrix} = \mathbf{v}^0 \tag{5}$$

is satisfied for any given $\mathbf{u}^0 \in \mathbb{R}^2$, $\mathbf{v}^0 \in \mathbb{R}^2$. The unknowns of the model are

- $u_\nu(t)$ and $u_\tau(t)$ i.e., the normal and the tangential *displacement*
- $\lambda_\nu(t)$ and $\lambda_\tau(t)$ i.e., the normal and the tangential *stress components.*

The data are the given $f_\nu(t)$ and $f_\tau(t)$ i.e., normal and tangential *load* components.

Parameters of the model: The nonnegative friction coefficient $\mathcal{F}$, and the mass and stiffness matrices

$$\mathbf{M} = \begin{bmatrix} a & 0 \\ 0 & a \end{bmatrix}\,, \quad \mathbf{A} = \begin{bmatrix} b & c \\ c & b \end{bmatrix}\,,$$

$$a = \frac{\rho l^2}{12}\,, \quad b = -\frac{\lambda + 3\nu}{2}\,, \quad c = \frac{\lambda + \nu}{2}\,,$$

where ρ, l, λ and ν are positive parameters (the density, the diameter of the element, and two Lamé coefficients).

The symbols Sign and $\mathrm{N}_{\mathbb{R}^1_-}$ denote *multivalued mappings* Sign : $\mathbb{R} \rightrightarrows \mathbb{R}$ and $\mathrm{N}_{\mathbb{R}^1_-} : \mathbb{R} \rightrightarrows \mathbb{R}$ called *signum* and *normal cone*, respectively, see e.g. [1]. We skip formal definitions. Instead, we introduce equivalent formulations via variational inequalities:

The condition (3) is called the *complementarity condition.* It can be interpreted as the `no contact` or the `contact`

$$\begin{cases} \lambda_\nu(t) = 0 & \text{for} \quad u_\nu(t) < 0 \quad \ldots\texttt{no contact} \\ \lambda_\nu(t) \le 0 & \text{for} \quad u_\nu(t) = 0 \quad \ldots\texttt{contact} \end{cases} \tag{6}$$

with the rigid foundation. The condition (4) reads as

$$\begin{cases} \lambda_\tau(t) = \mathcal{F}\,\lambda_\nu(t) & \text{for} \quad u'_\tau(t) > 0 \\ \lambda_\tau(t) = -\mathcal{F}\,\lambda_\nu(t) & \text{for} \quad u'_\tau(t) < 0 \\ |\lambda_\tau(t)| \le -\mathcal{F}\,\lambda_\nu(t) & \text{for} \quad u'_\tau(t) = 0 \end{cases} \tag{7}$$

One can easily conclude that

1. In the case of `no contact` in (6), the condition (7) yields $\lambda_\nu(t) = \lambda_\tau(t) = 0$
2. In the case of `contact` in (6), the condition (7) can be interpreted as

$$\begin{cases} \lambda_\tau(t) = \mathcal{F}\,\lambda_\nu(t) & \text{for} \quad u'_\tau(t) > 0 \quad \ldots\texttt{contact-slip} \\ \lambda_\tau(t) = -\mathcal{F}\,\lambda_\nu(t) & \text{for} \quad u'_\tau(t) < 0 \quad \ldots\texttt{contact-slip} \\ |\lambda_\tau(t)| \le -\mathcal{F}\,\lambda_\nu(t) & \text{for} \quad u'_\tau(t) = 0 \quad \ldots\texttt{contact-stick} \end{cases} \tag{8}$$

The aim is to solve the *initial value problem* (2)–(5). We consider two kinds of algorithms: In Section 4, we introduce an event driven algorithm and in Section 5 we sketch a time-stepping algorithm.

In the following, let us relabel the state variables $x_1 = u_\nu$, $x_2 = u'_\nu$, $x_3 = u_\tau$, $x_4 = u'_\tau$.

4. The event-driven algorithm

The idea is a *dynamical simulation* of the particular solution modes `contact` and `no contact`. They are defined by different systems of ordinary differential equation (i.e., different *vector fields*). Then the solution modes should be concatenated according certain rules (continuity of displacements).

The mode `contact` is modeled as a *Filippov system*, see e.g. [3, 1]. Details can be found in Supplement 7, see the system (12). In this solution mode we have $\lambda_\nu(t) < 0$ on an open time interval $t \ge 0$. It can be shown that $x_1(t) = x_2(t) = 0$, and $\lambda_\nu(t) = -c\,x_3(t) - f_\nu(t) \le 0$. We distinguish two cases:

- If $x_1(t) = x_2(t) = 0$ and $x_4(t) = 0$ then the body is in `contact-stick` regime,
- If $x_1(t) = x_2(t) = 0$ and $x_4(t) \neq 0$ then the body is in `contact-slip` regime.

The dynamical simulation of the contact mode is bases on the *Filippov convex method* and its modifications, [3, 1]. In forthcoming experiments we used the open-source software [12] which is based on the MATLAB ODE suit [15] with an adaptive stepsize.

The mode `no contact` is modeled as two coupled linear oscillators where $\lambda_\nu(t) = \lambda_\tau(t) = 0$, $x_1(t) < 0$ on an open time-interval $t \ge 0$, see Supplement 7, the system (14)&(15).

The coupling of the modes `contact` and `no contact` can be viewed as an *hybrid impact model*, [1]. Why do we call the algorithm an *event-driven* algorithm?

Changing particular modes is linked to the sign-changes of functions $t \mapsto x_1(t)$, $t \mapsto \lambda_\nu(t) \equiv -c\, x_3(t) - f_\nu(t)$ and $t \mapsto x_4(t)$. The MATLAB ODE suit [15] provides an efficient tool called *event location* to localize sign-changes of functionals in space and time.

The given acting force f_ν and f_τ in (2) may be time dependent. In following examples we let the tangential component $f_\tau = f_\tau(t)$ to be periodic and the normal component f_ν to be fixed. We model the action of the craftsman instrument called 'Jack plane'.

Example 4.1 `Contact` *only*
Data*:* $a = 1$, $b = -1.2$, $c = 1$, $\mathcal{F} = 0.4$,
a periodic forcing: $f_\tau(t) = \sin(\omega t)$, $\omega = 1/6$, $f_\nu(t) \equiv f_\nu = 1.3$, *a 'Jack plane' model. The initial condition:* $[0, 0, 0, 0.1]$. *The time-span:* $[0, T]$, $T = 10 \cdot \frac{2\pi}{\omega}$.

The relevant results are shown in Figure 6 and Figure 7. The value of f_ν is sufficiently large and the instrument rests on the foundation for all time.

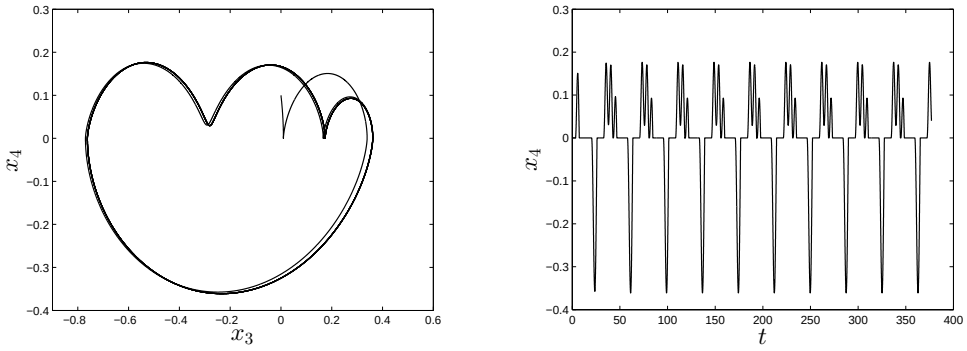


Figure 6: $f_\nu = 1.3$. On the left: A phase plot of x_4 versus x_3. On the right: A plot of x_4 versus time t. Contact regime: If $x_4(t) = 0$ then `contact-stick`. If $x_4(t) \neq 0$ then `contact-slip`.

Example 4.2 *Coupling of the modes* `contact` *and* `no contact`
Data*:* $a = 1$, $b = -1.2$, $c = 1$, $\mathcal{F} = 0.3$,
a periodic forcing: $f_\tau(t) = \sin(\omega t)$, $\omega = 1/6$, $f_\nu(t) \equiv f_\nu = 0.5$, *a 'Jack plane' model. The initial condition:* $[0, 0, 0, 0.1]$. *The time-span:* $[0, T]$, $T = 10 \cdot \frac{2\pi}{\omega}$.

The relevant results are shown in Figure 8 and Figure 9. This time f_ν is small enough and the instrument is lifted from the foundation for particular time periods. The 'Jack plane' is bouncing on the foundation.

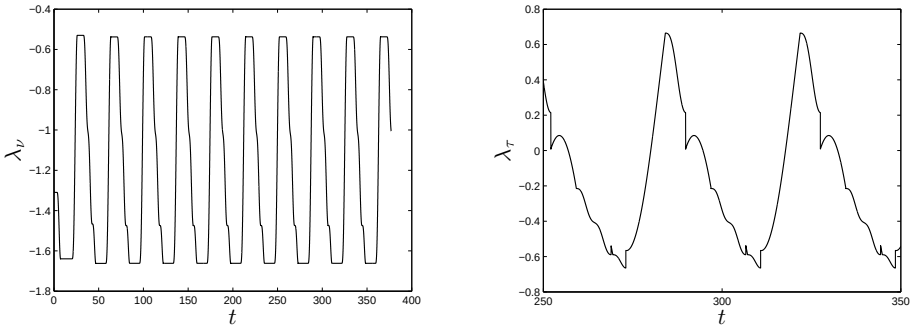


Figure 7: $f_\nu = 1.3$. A plot of λ_ν versus time t. Note that $\lambda_\nu(t) < 0$ characterizes the `contact` mode. On the right: A plot of λ_τ versus time t, a zoom.

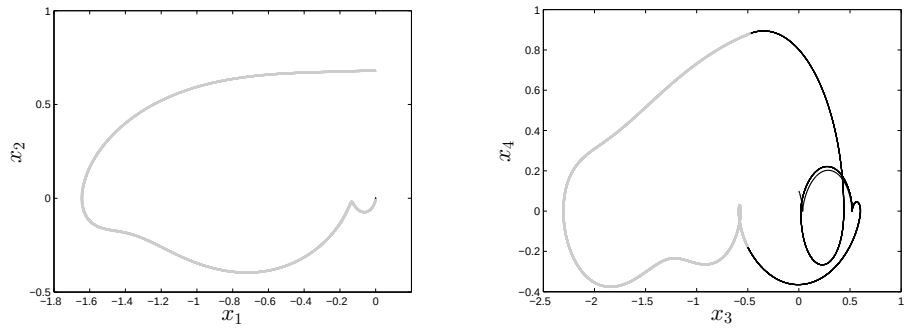


Figure 8: $f_\nu = 0.5$. On the left: A phase plot of x_1 versus x_2. Observe that $x_1 \leq 0$, an impact at $x_1 = 0$. On the right: A phase plot of x_3 versus x_4. Legend: `contact` ... black, `no contact` ... gray.

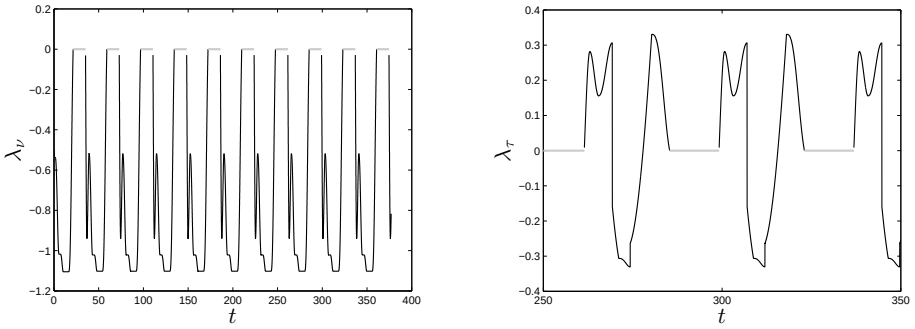


Figure 9: $f_\nu = 0.5$. On the left: A plot of λ_ν versus time t. On the right: A plot of λ_τ versus time t, a zoom. Legend: `contact` ... black, `no contact` ... gray.

5. The time-stepping algorithm

Consider the initial value problem (2)–(5). In [10], there was proposed a natural time discretization of this problem via *mid-point rule* with a fixed stepsize dt. At each time step, the algorithm identifies the solution mode (namely, the options `contact`, `contact − stick` and `contact − slip`) and propose the solution update. The identification is unique provided that the stepsize dt is sufficiently small. (Note that we used the scheme without *mass-redistribution*, [10]). Let us run the mid-point algorithm using the same data as in Example 4.1. We expect qualitatively similar plots as in Figure 6 and Figure 7.

Example 5.3 `Contact` *only, see Example 4.1*
Data*: $a = 1$, $b = -1.2$, $c = 1$, $\mathcal{F} = 0.4$, time increment $dt = 0.001$,*
a periodic forcing: $f_\tau(t) = \sin(\omega t)$, $\omega = 1/6$, $f_\nu(t) \equiv f_\nu = 1.3$, a 'Jack plane' model.
The initial condition: $[0, 0, 0, 0.1]$. The time-span: $[0, T]$, $T = 10 \cdot \frac{2\pi}{\omega}$.

In Figure 10, on the left, there is a plot of initial stages of x_4 computed via the mid-point rule. Note that corresponding zoom in Figure 6, on the right, computed via the event-driven algorithm would look much the same. Remarkable are the run-time differences: 2495.6 seconds (the mid-point rule) vs 2.3 seconds (the event-driven algorithm). The zoom in Figure 10 reveals that the numerical solution oscillates between the stages `contact-slip` and `contact-stick` (see the isolated dots). In that case, the remedy is to guide the solution to remain in regime `contact-stick`. It can be done by adapting slightly the original code in [10] e.g., in case `contact-stick` we set directly $x_4 = 0$. We call the resulting algorithm the *stabilized* mid-point rule. In Figure 11, we plot x_4 versus t computed via stabilized mid-point rule. Due to the setting of Example 5.3, i.e. `contact` only, we have just two competing modes namely `contact-slip` and `contact-stick` depicted by dashed and solid curves. Elapsed time was 64.841882 seconds (stabilized mid-point rule, $0 \le t \le 380$, $dt = 0.001$).

The above stabilization technique can be related to the approach by [2, 14].

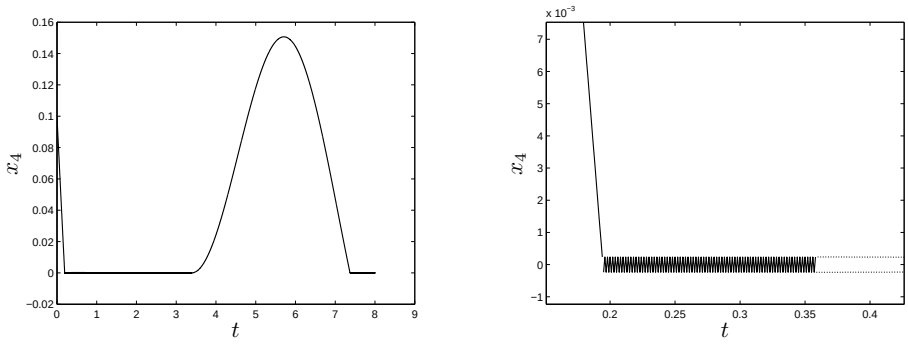


Figure 10: $f_\nu = 1.3$, time increment $dt = 0.001$. On the left: The solution via mid-point rule. A plot of x_4 versus t as $0 \le t \le 8$. On the right: a zoom.

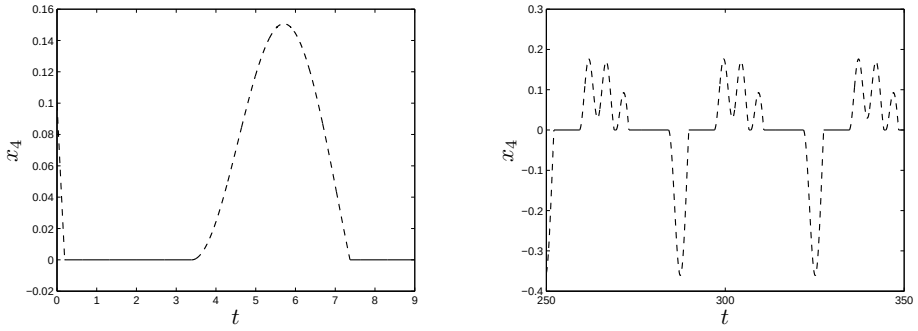


Figure 11: $f_\nu = 1.3$, time increment $dt = 0.001$. The solution via *stabilized* midpoint rule. Legend: `contact-stick` ... solid, `contact-slip` ... dashed curves. On the left: the initial stages $0 \leq t \leq 9$. On the right: The periodic pattern of the limit set, $250 \leq t \leq 350$.

6. Conclusions

We considered simplified models (i.e., lumped parameter models) for both the static, see (1), and the dynamic friction model, see (2)–(5).

The static model (1) is piecewise smooth, parameter dependent. It can be solved by continuation techniques. The dynamic model (2)–(5) is piecewise smooth dynamical system where time t is a parameter. The approaches to numerical solution (the event-driven algorithm in Section 4 and the time-stepping algorithm in Section 5) can be viewed as approximations of discrete time, piecewise-smooth dynamical systems. Both the static and dynamic problems can be solved by similar (continuation) techniques in spite of the fact that both models are qualitatively different. The continuation techniques for solving the static case-study model (1) can be extended to higher dimensions. We hope for such an extension for dynamic contact problems which would deal with structures as in Figure 3.

Comparison of the event-driven algorithm and the time-stepping algorithm: In [8], we compared an event-driven algorithm (based on the software in [12]) and a time-stepping algorithm (based on implicitly defined law of Coulomb friction, [14, 2, 13]) for the Dry-friction model (in 2-D) i.e., the model of a slide fastener. The comparison in [8] argue strongly for an event-driven algorithm:

1. In [12], there is implemented an adaptive stepsize refinement. As a consequence, the solver reduces the computational costs.
2. The solution modes are clearly distinguished and precisely localized (in case of Dry friction we distinguish just $\texttt{contact} - \texttt{slip}$ and $\texttt{contact} - \texttt{stick}$ modes).

Coming back to the algorithms formulated in Section 4 and Section 5, respectively: The event-driven algorithm seems to be superior to the time-stepping algorithm. The argument for this statement is the same as the above. Mind you the

failure in Figure 10, on the left. It can be fixed, see Figure 11. Nevertheless, there is a space for improvements as the mode identification is concerned.

On the other hand, the event-driven algorithm uses built-in MATLAB routines namely, the routines concerning the stepsize control, see [15]. When thinking about possible generalizations of event-driven algorithms in order to deal with real structures as in Figure 3, one has to programm adaptive step refinement or event-location routines himself. In principle, it is possible. In the continuation context, the key algorithms are already developed, see Figure 4, on the right.

7. Supplement: Modelling the modes `contact` and `no contact`

This supplement pertains to Section 4, giving particular details. Basically, we shall follow [8].

7.1. The `contact` mode

Assume that the body is in contact with the rigid foundation at a particular time $t^0 \geq 0$ and on an open non-empty time interval $\mathcal{I}(t^0)$. It means that the equations (2) together with the conditions $\{\lambda_\nu(t) \leq 0\,,\ u_\nu(t) = 0\}$ and (8) are satisfied for $t \in \mathcal{I}(t^0)$.

The system (2) consists of two equations:

$$a u''_\nu(t) = b u_\nu(t) + c u_\tau(t) + f_\nu(t) + \lambda_\nu(t) \tag{9}$$

$$a u''_\tau(t) = c u_\nu(t) + b u_\tau(t) + f_\tau(t) + \lambda_\tau(t) \tag{10}$$

Since $u_\nu(t) = 0$ for all $t \in \mathcal{I}(t_0)$ then $u''_\nu(t) = 0$ for all $t \in \mathcal{I}(t_0)$. The equation (9) reduces to an algebraic constraint:

$$\lambda_\nu(t) = -c\, u_\tau(t) - f_\nu(t)\,, \quad \lambda_\nu(t) \leq 0 \tag{11}$$

for $t \in \mathcal{I}(t^0)$. From (10) and (8), we conclude that

1. If $u'_\tau > 0$ then $\lambda_\tau = \mathcal{F}\lambda_\nu$, see (8). The equations (10)&(11) yield

$$u''_\tau = \frac{b - \mathcal{F}c}{a} u_\tau + \frac{1}{a}\,(f_\tau - \mathcal{F} f_\nu)$$

2. If $u'_\tau < 0$ then $\lambda_\tau = -\mathcal{F}\lambda_\nu$, see (8). Due to the equations (10)&(11)

$$u''_\tau = \frac{b + \mathcal{F}c}{a} u_\tau + \frac{1}{a}\,(f_\tau + \mathcal{F} f_\nu)$$

3. If $u'_\tau = 0$ then $|\lambda_\tau| \leq -\mathcal{F}\lambda_\nu$, see (8). In a spirit of the Filippov convex method [3, 1] we consider the convex combination of the right-hand sides of the above equations

$$u''_\tau = \frac{(1-2\lambda)\,\mathcal{F}c + b}{a}\, u_\tau + \frac{1}{a} f_\tau + \frac{1-2\lambda}{a}\,\mathcal{F} f_\nu\,, \quad \lambda \in [0,1]\ .$$

Let us relabel the state variables $x_1 = u_\nu$, $x_2 = u'_\nu$, $x_3 = u_\tau$ and $x_4 = u'_\tau$. Accordingly, we introduce vector fields $F_1 : \mathbb{R}^5 \to \mathbb{R}^5$ and $F_2 : \mathbb{R}^5 \to \mathbb{R}^5$ as

$$F_1 = \begin{bmatrix} 0 \\ 0 \\ x_4 \\ \dfrac{b - \mathcal{F}c}{a}\, x_3 + \dfrac{1}{a}\,(f_\tau - \mathcal{F} f_\nu) \\ 1 \end{bmatrix}, \quad F_2 = \begin{bmatrix} 0 \\ 0 \\ x_4 \\ \dfrac{b + \mathcal{F}c}{a}\, x_3 + \dfrac{1}{a}\,(f_\tau + \mathcal{F} f_\nu) \\ 1 \end{bmatrix}$$

where $f_\tau = f_\tau(t) = f_\tau(x_5)$, $f_\nu = f_\nu(t) = f_\nu(x_5)$. The vector fields F_1 and F_2 are autonomous (which was the condition to use the ready-made software [12]). Nevertheless, we can recover time t easily.

Moreover, we define the level-set operator $H_{12} : \mathbb{R}^5 \to \mathbb{R}$,

$$H_{12}(x) = x_4\,.$$

The fields F_1 and F_2, respectively, are defined on

$$S_1 = \left\{x \in \mathbb{R}^5 :\, H_{12}(x) > 0\right\} \quad \text{end} \quad S_2 = \left\{x \in \mathbb{R}^5 :\, H_{12}(x) < 0\right\}\,.$$

The set $\Sigma_{12} = \{x \in \mathbb{R}^5 :\, H_{12}(x) = 0\}$ is the discontinuity surface. We consider the Filippov system

$$x' = \begin{cases} F_1(x) & \text{for} \quad x \in S_1 \\ F_2(x) & \text{for} \quad x \in S_2 \end{cases} \tag{12}$$

For a given initial condition $x^0 \in \mathbb{R}^5$, the Filippov's convex method, e.g. [3, 1, 12], gives the solution of the system (12) on a time span for which the body stays in contact with the rigid obstacle i.e.,

$$\lambda_\nu(t) = -c\, x_3(t) - f_\nu(t) \le 0\,.$$

It means that the initial condition $x^0 \in \mathbb{R}^5$ has to satisfy

$$x^0 = \left[0, 0, x^0_3, x^0_4, t^0\right]^\top, \quad -c\, x^0_3(t^0) - f_\nu(t^0) < 0\,. \tag{13}$$

7.2. The `no contact` mode

Recall the original meaning of the state variables $x_1 = u_\nu$, $x_2 = u'_\nu$, $x_3 = u_\tau$ and $x_4 = u'_\tau$. Assume that the body is not in contact with the rigid foundations at a particular time $t^0 \ge 0$ and on an open non-empty time interval $\mathcal{I}(t^0)$. Due to (6) (the option `no contact`) we can claim that $\{\lambda_\nu(t) = 0, u_\nu(t) < 0\}$ for $t \in \mathcal{I}(t_0)$. We

already noted that $\lambda_\nu(t) = \lambda_\tau(t) = 0$ for $t \in \mathcal{I}(t^0)$, as a consequence of (7). Hence, the system (2) reduces to equations

$$au''_\nu(t) = bu_\nu(t) + cu_\tau(t) + f_\nu(t) \tag{14}$$

$$au''_\tau(t) = cu_\nu(t) + bu_\tau(t) + f_\tau(t) \tag{15}$$

for $t \in \mathcal{I}(t^0)$ provided that $u_\nu(t) < 0$. We formulate (14)&(15) as an autonomous system adding an extra equation $t' = 1$. Coming back to the variable $x \in \mathbb{R}^5$ we introduce the vector field $F_3 : \mathbb{R}^5 \to \mathbb{R}^5$ as

$$F_3 = \begin{bmatrix} x_2 \\ \frac{b}{a}x_1 + \frac{c}{a}x_3 + \frac{1}{a}f_\nu(x_5) \\ x_4 \\ \frac{c}{a}x_1 + \frac{b}{a}x_3 + \frac{1}{a}f_\tau(x_5) \\ 1 \end{bmatrix} .$$

The field F_3 is defined on

$$S_3 = \left\{ x \in \mathbb{R}^5 : \, x_1 < 0 \right\} .$$

Acknowledgements

This work was supported by the grant GAČR P201/12/0671.

References

[1] di Bernardo, M., Budd, C. J., Champneys, A. R., and Kowalczyk, P.: *Piecewise-smooth Dynamical Systems. Theory and Applications.* Springer Verlag, New York, 2008.

[2] Darbha, S., Nakshatrala, K., and Rajagopal, K.: On the vibrations of lumped parameter systems governed by differential-algebraic equations. Journal of the Franklin Institute **347** (2010), 87–101.

[3] Filippov, A. F.: *Differential equations with discontinuous righthand sides.* Kluwer Academic Publishers, Dordrecht, 1988.

[4] Haslinger, J., Janovský, V., and Kučera, R.: Path-following the static contact problem with coulomb friction. In: J. Brandts, S. Korotov, M. Křížek, J. Šístek, and T. Vejchodský (Eds.), *Proceedings of the International Conference Application of Mathematics 2013, Prague, May 15-17, 2013*, pp. 104–116. Institute of Mathematics, Academy of Sciences of the Czech Republic, 2013.

[5] Haslinger, J., Janovský, V., Kučera, R., and Motyčková, K.: Nonsmooth continuation of parameter dependent static contact problems with Coulomb friction. Mathematics and Computers in Simulation (submitted).

[6] Haslinger, J., Janovský, V., and Ligurský, T.: Qualitative analysis of solutions to discrete static contact problems with Coulomb friction. Comp. Meth. Appl. Mech. Engrg. **205–208** (2012), 149–161.

[7] Hild, P. and Renard, Y.: Local uniqueness and continuation of solutions for the discrete Coulomb friction problem in elastostatics. Quart. Appl. Math. **63** (2005), 553–573.

[8] Janovský, V.: Lumped parameter friction models. In: *Proceedings of 4th Scientific Colloquium, Prague, June 24-26, 2014*, pp. 132–149. Institute of Chemical Technology, Czech Republic, 2014.

[9] Khenous, H., Laborde, P., and Renard, I.: Mass redistribution method for finite element contact problems in elastodynamics. Eur. J. Mech., A/Solids **27** (2008), 918–932.

[10] Ligurský, T. and Renard, I.: A well-posed semi-discretization of elastodynamic contact problems with friction. Quart. J. Mech. Appl. Math. **64** (2011), 215–238.

[11] Ligurský, T. and Renard, I.: A continuation problem for computing solutions of discretised evolution problems with application to plane quasi-static contact problems with friction. Comp. Meth. Appl. Mech. Engrg. **280** (2014), 222–262.

[12] Piiroinen, P. and Kuzetsov, Y. A.: An event-driven method to simulate Filippov systems with accurate computing of sliding motions. ACM Transactions on Mathematical Software **34** (2008).

[13] Pražák, D. and Rajagopal, K.: Mechanical oscillators described by a system of differencial-algebraic equation. Appl. Math. **57** (2012), 129–142.

[14] Rajagopal, K.: A generalized framework for studying vibrations of lumped parameter systems. Mech. Res. Comm. **37** (2010), 463–466.

[15] Shampine, L. and Reichelt, M.: The matlab ode suit. SIAM J. Sci. Comput. **18** (1997), 1–19.

Conference Applications of Mathematics 2015, in honor of the birthday anniversaries of Ivo Babuška (90), Milan Práger (85), and Emil Vitásek (85)
J. Brandts, S. Korotov, M. Křížek, K. Segeth, J. Šístek, T. Vejchodský (Eds.)
Institute of Mathematics CAS, Prague 2015

MESSAGE DOUBLING AND ERROR DETECTION IN THE BINARY SYMMETRICAL CHANNEL

Lucie Kárná[1,2], Štěpán Klapka[2]

[1] Faculty of Transportation Sciences CTU
Na Florenci 25, Praha 1, Czech Republic
karna@fd.cvut.cz
[2] AŽD Praha s.r.o., Research and Development
Žirovnická 2, Praha 10, Czech Republic
klapka.stepan@azd.cz

Abstract: The error correcting codes are a common tool to ensure safety in various safety-related systems. The usual technique, employed in the past, is to use two independent transmission systems and to send the safety relevant message two times. This article focuses on analysis of the detection properties of this strategy in the binary symmetrical channel (BSC) model.
Besides, various modifications of the mentioned technique can be used. Their impact on the detection properties can be significant, positively or negatively. This article demonstrates one of these modifications.

Keywords: error correcting code, undetected error, message repetition
MSC: 62P30, 94A40, 94B70

1. Introduction

Communication safety is a small, but important part of the safety of every electronics-based system, particularly in railway interlocking systems. A special position in this issue has the safety code, because it is the unique tool to protect messages against corruption.

The basic motivation for this paper was the cooperation on design of interlocking systems. The communication protocol, used by our partner, includes sending the safety relevant messages twice using two transmission lines. It turns up, that safety analysis of this simple approach is not quite simple.

The first part of the article describes some basic terms of coding theory. The second part introduces the concept of probability of undetected error in the binary symmetrical channel as a basic tool for evaluating detection quality of the code. The next part investigates the main approaches to message doubling. The problem of calculating the probability of an undetected error in these cases is studied.

2. Coding theory

This section defines the basic terminology for linear binary codes and the related binary symmetrical channel (BSC) model. The "code-related" terminology in this paper is based on terms used in the mathematical coding theory (see for example [2]).

2.1. Linear binary codes

A linear binary $(n, k)-$code K is any $k-$dimensional subspace of the space $(\mathbf{Z}_2)^n$. Traditionally, binary vectors from $(\mathbf{Z}_2)^n$ are called *words*; the words from the code K are the *code words*. In an $(n, k)-$code the code word length is n, the number of information bits is equal to k and the number of redundant bits is equal to $c = n - k$. Any linear $(n, k)-$code K can be described by its *generator matrix*, whose rows are exactly the words forming a basis of the subspace K.

In practice, usually the code word of an $(n, k)-$code is created by the addition of c bits (the *redundant* or *control part* of the code word) to a word of length k (the *information part* of the code word). This technique is called a *systematic encoding*, the code is a *systematic code*. A generator matrix of the systematic code has the form $G = (E|B)$, where E denotes the identity matrix of the order k and B is some $k \times c$ matrix.

2.2. Error detection

During the transfer of a message unwanted modifications can occur. Usually, it is supposed that a number of bits is preserved and these modifications are manifested by altered bit(s). The adverse situation occurs, when the modification during transfer unfortunately creates another code word, different from the sent one. The receiver has no possibility to recognize this state.

This scenario is dangerous and results in an undetected error. The probability of such an undetected error of the detection codes used in safety relevant applications (including transportation control) is a very important safety parameter.

We define the *Hamming weight* of a word as the count of non-zero bits in the word. Then we define the *minimal distance* of a linear code as the smallest non-zero Hamming weight of its code word.

The minimal distance of a linear code sets the ability of the code to detect some classes of transmission errors. A code with a minimal distance d will detect all errors with at most $d - 1$ modified bits in the transmitted code word (see [2, 3]).

For a more detailed description of the code, a *weight structure* of the code is defined as a vector $A = (A_0, A_1, A_2, \ldots, A_n)$, where A_i denotes the number of code words with Hamming weight equal to i. For linear codes, the weight structure is fully sufficient for the description of its ability to detect errors.

3. Probability of undetected error

The most useful approach for measuring the detection properties of a code uses its maximal value of the probability of undetected error in a binary symmetrical channel.

3.1. Description of the BSC model

The binary symmetrical channel (BSC) is a simple probabilistic model based on a bit (binary symbol) transmission. The BSC model does not describe the reality completely, but it is an appropriate tool for comparison of the detection properties of the codes.

In this model the probability of an error is supposed to be independent from one bit to the next one. The probability p_e that the bit changes its value during the transmission (*bit error rate*) is the same for both possibilities ($0 \to 1$ and $1 \to 0$). The probability that the code word with n symbols is corrupted exactly in i symbols is then equal to

$$p_e^i \, (1 - p_e)^{n-i} \, . \tag{1}$$

The probability of an undetected error in the BSC model for a linear binary code K with code words of length n and with minimal Hamming distance d is given by the following formula

$$P_{ud}(K, p_e) = \sum_{i=d}^{n} p_e^i \, (1 - p_e)^{n-i} \, A_i, \tag{2}$$

where A_i is the number of code words with exactly i nonzero symbols and p_e is the bit error rate in the BSC channel.

For every linear $(n, k)-$code the value of the function $P_{ud}(K, .)$ for $p_e = 1/2$ is equal to $(2^k - 1)/2^n$ and this is a local maximum of this function. Although the use of a transmission channel with bit error rate near to $1/2$ is virtually excluded, the standard EN 50159 for safety-related communication in railway applications [1] recommends not to use a better detection estimate than this value for calculations in a safety model.

Actually, for the codes used in safety relevant applications it is necessary to know (or, at least, estimate) an upper bound of the function $P_{ud}(K, .)$ on the entire interval $[0, 1/2]$. In particular, it is recommended to use codes with a monotone function $P_{ud}(K, .)$ or, at least, this function should not exceed the value $P_{ud}(K, 1/2)$ (see [1]).

3.2. Indirect calculation using dual code

The formula (2) for the probability of an undetected error of a code is quite simple in principle. However, the coefficients A_i (the number of code words with i nonzero symbols) cannot be expressed by some elegant formula (with exception of rare family of codes). They have to be calculated by counting the weight of every

individual code word. As the number of code words is equal to 2^k, these calculations are not feasible for long code words.

To get more effective calculations, it is useful to apply the MacWilliams Identity, which links the weight structure of the given code and its dual code. These computations use another representation of the weight structure by the weight enumerator $\mathbf{pw}(x, K)$. It is the following formal polynomial:

$$\mathbf{pw}(K, x) = \sum_{i=0}^{n} A_i x^i. \tag{3}$$

3.2.1. Dual code

We define for the binary words $u = u_1 u_2 \dots u_n$ and $v = v_1 v_2 \dots v_n$

$$u \cdot v = \sum_{i=1}^{n} u_i \cdot v_i. \tag{4}$$

This bilinear form is usually referred as *inner product*, despite it does not satisfy condition that from $u \cdot u = 0$ follows $u = (0, 0, \dots, 0)$. This is a consequence of the fact that in the space $\mathbf{Z}_2$ it is $1 + 1 = 0$.

A *dual code* to the linear binary $(n, k)-$code K is a linear binary $(n, n-k)-$code $K^\perp$ consisting from all words $u \in (\mathbf{Z}_2)^n$, whose inner product with every code word from the code K is equal to zero:

$$u \in K^\perp \iff u \cdot v = 0 \text{ for every } v \in K. \tag{5}$$

If the code K is a systematic code with generator matrix $G = (E|B)$, where E is the identity matrix and B is some $k \times c$ matrix, then the dual code $K^\perp$ has a generator matrix $G^\perp = (B^{\mathrm{T}}|E)$, where B^{T} is the transposed of the matrix B.

3.2.2. MacWilliams Identity

The following formula is the MacWilliams Identity for binary codes:

$$2^k \mathbf{pw}(K^\perp, x) = (1 + x)^n \mathbf{pw}\left(K, \frac{1 - x}{1 + x}\right). \tag{6}$$

The advantage of this formula is that the dual code has much fewer code words ($2^{n-k} << 2^k$, because typically, $n - k = c << k$) and then it is significantly easier to compute the weight distribution for a dual code.

4. Message doubling

A natural procedure to ensure authenticity of the message is to use two independent transmission systems and to send the safety relevant message twice. The received message is considered undamaged only if both copies are delivered and theirs contents are matching.

The situation with a missing message is trivial, so we focus only on the case when both copies arrived and their length is preserved (verification of the correct length of the message is done by other techniques). In the BSC model (independent transmission of single symbols – bits), it is equivalent to a serial transmission using a single transmission channel.

4.1. Repetition of the message

A plain repetition of the message with length equal to k is represented by the linear binary $(2k, k)-$code with binomial weight structure, where

$$A_{2j} = \binom{n}{j} \quad \text{for } j = 0, \ldots, k \tag{7}$$

$$A_{2j-1} = 0 \quad \text{for } j = 1, \ldots, k. \tag{8}$$

The minimal distance of the code is equal to 2, which is insufficient for most purposes.

More useful is a repetition of the message already protected by some linear code. Consider a binary message of length k. This message we protect by a linear binary (n, k)-code K_A with minimal distance d and with known weight structure $A = (A_0, A_1, A_2, \ldots, A_n)$. Then we send this message twice.

This procedure corresponds to the protection of the message with linear binary $(2n, k)-$code K_D. The minimal distance of this code is equal to $2d$ and its weight structure, denoted as $D = (D_0, D_1, D_2, \ldots, D_n)$, is given by the weight structure of the code K_A:

$$D_{2j} = A_j \quad \text{for } j = 0, \ldots, n \tag{9}$$

$$D_{2j-1} = 0 \quad \text{for } j = 1, \ldots, n. \tag{10}$$

The probability of undetected error in the BSC of the code K_D is then

$$P_{ud}(K_D, p_e) = \sum_{i=2d}^{2n} p_e^i \, (1 - p_e)^{n-i} \, D_i = \sum_{i=d}^{n} \left(p_e^i \, (1 - p_e)^{n-i} \right)^2 A_i. \tag{11}$$

Obviously, we have

$$P_{ud}(K_D, .) < P_{ud}(K_A, .). \tag{12}$$

The following graph illustrates the situation for one sample code with length $n = 32$ and with $c = 8$ control bits. (Note: it is a shortened cyclic code generated by the polynomial $x^8 + x^7 + x^2 + 1$ – for explanation see e. g. [3].) The upper curve represents the probability of an undetected error for the sample code, the lower curve represents the corresponding probability with repetition of the message. The vertical axis is in logarithmic scale.

Let us consider the lower bound of the function $P_{ud}(K_A, .)$ as a $A_d \, p_e^d (1 - p_e)^{n-d}$. The ratio between the lower bounds for the codes K_D and K_A is $p_e^d (1 - p_e)^{n-d}$, and the minimal improvement is obtained for $p_e = d/n$. Hence with increased length n the maximal value of the lower bound decreases significantly slower than the value $P_{ud}(K_D, p_e)$. From this it is evident that the minimal distance is a very important parameter, which has a dominant influence to the detection properties of doubling messages.

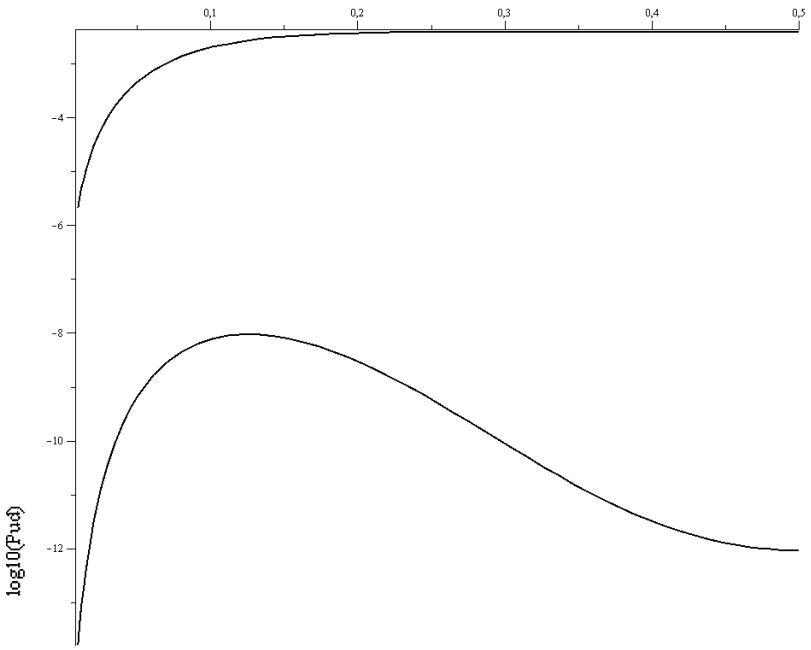


Figure 1: The probability of an undetected error for the sample code (upper curve) and for the same code combined with repetition of the message. Horizontal axis: bit error rate p_e, vertical axis: logarithm of probability of undetected error $P_{ud}(p_e)$.

4.2. Double encoding of the message

In some situations a more sophisticated approach can be useful. We protect a binary message M of length k by a linear binary $(n, k)-$code K_A with known weight structure $A = (A_0, A_1, A_2, \ldots, A_n)$; we denote this encoded message by M_A. Then we repeat this procedure with the original message M and with another linear binary $(n, k)-$code K_B with weight structure $B = (B_0, B_1, B_2, \ldots, B_n)$; denote the encoded message by M_B. Finally we send both messages M_A and M_B using two separate transmission lines.

One advantage of this approach is that the received messages are "signed" – if one of the messages M_A and M_B is wrong, we know on which transmission line (or in which encoder) the failure occurred. More important, this technique protects against the situation, when two copies of one received message are handled as two independent messages.

4.2.1. Weight structure

The two-transmission-lines configuration is in the BSC model equivalent with transmission of concatenated messages M_A and M_B. This corresponds with some linear binary $(2n, k)-$code K_{AB}. Unfortunately, the weight structure of the code K_{AB} cannot be derived from the weight structures of the codes K_A and K_B. However,

the number of information bits k is equal for all three codes K_A, K_B and K_{AB} and therefore if the calculation of the weight structure of the codes K_A, K_B is manageable, then for the code K_{AB} the computation is practicable as well.

The questionable situation occurs, when the number of information bits k is too high and it is impossible to generate 2^k code words in a reasonable time. The dual codes to the K_A and K_B are $(n, n-k)-$codes, and if the number of the redundant bits $c = n - k$ is acceptably small, it is possible to compute the weight structures of these duals and then use the MacWilliams identity (6) to compute the weight structures of the codes K_A and K_B.

However, the dual code to the code K_{AB} is a $(2n, n+c)-$code and generation of the 2^{n+c} code words may be impossible, as in a typical case the number of information bits k is considerably greater than the number of control bits $c = n-k$. This problem can be solved by utilization of the special form of the code dual to K_{AB}.

Let us assume that the codes K_A and K_B are systematic codes. This is a reasonable assumption, because every linear code is equivalent with a systematic code. Then the codes K_A and K_B have generator matrices in the form $G_A = (E|A)$ and $G_B = (E|B)$, respectively. A generator matrix of the code K_{AB} is $G_{AB} = (E|A|E|B)$, and there exists an equivalent generator matrix $(E|E|A|B)$. Then a generator matrix of the dual code $K_{AB}^{\perp}$ has the following form:

$$G_{AB}^{\perp} = \left(\begin{array}{c|ccc} E & E & \mathbf{0} & \mathbf{0} \\ A^T & \mathbf{0} & E & \mathbf{0} \\ B^T & \mathbf{0} & \mathbf{0} & E \end{array} \right), \tag{13}$$

where $\mathbf{0}$ denotes a zero matrix.

The matrix

$$G^* = \left(\begin{array}{c|cc} A^T & E & \mathbf{0} \\ B^T & \mathbf{0} & E \end{array} \right), \tag{14}$$

derived from the $G_{AB}^{\perp}$, is a generator matrix of some $(k+2c, 2c)-$code K^*. In the favourable case it is acceptable to generate 2^{2c} code words and enumerate their weights.

Computation of the weight structure of the code K_{AB} is based on more detailed information about weights of the code words of the code K^*. Rather than the weight structure we compute a matrix of weight structures. We split a code word into two parts: the information part of length $2c$ and the control part of length k. Then we construct a matrix $N = (n_{ij})$, where n_{ij} is the number of code words of the code K^* with weight of the information part equal to i and weight of the control part equal to j.

Every code word of the code $K_{AB}^{\perp}$ is the sum of two words $v + w$:

- $v = (u, u, o)$, where u is an arbitrary binary word of length k and o is a zero vector of length $2c$, and
- $w = (w_1, o, w_2)$, where (w_1, w_2) is a code word of the code K^* (w_1 consists of its first k bits, w_2 is the rest) and o is a zero vector of length k.

Consider a word w with weight of w_1 equal to i and weight of w_2 equal to j. We add to this word every possible word of the type $v = (u, u, o)$. For every position, where it is one in the word w_1 and zero in the word u, the weight of the sum $v + w$ increases by 2. Then, for given w there exist $\binom{k-i}{m}$ words with weight $i + j + 2m$. The number of these words w is $2^j n_{ij}$. Adding these contributions for all indices i and j we obtain the desired weight structure of the code $K_{AB}^{\perp}$ and finally by means of the MacWilliams Identity (6) the weight structure of the code K_{AB}.

This procedure is quite complicated, nevertheless, our computations show, that for a code with 16 control bits it is fully manageable on ordinary personal computer.

4.2.2. Upper estimate of $P_{ud}(K_{AB}, .)$

In case the enumeration of the 2^{2c} code words of the code K^* is computationally too difficult, but 2^c code words of the codes K_A and K_B is still computationally accessible, we can estimate the maximal value of $P_{ud}(K_{AB}, .)$ by the following construction.

We use the known weight structures $A = (A_0, A_1, \ldots, A_n)$ of the code K_A and $B = (B_0, B_1, \ldots, B_n)$ of the code K_B to create a new weight structure $C = (C_0, C_1, \ldots, C_n)$ of the fictive code K_f. The value of C_i we define as the maximum value of A_i, B_i. Then we consider doubling of the message with this fictive code K_f as described in Section 4.1 and enumerate the upper bound of the $P_{ud}(K_f, .)$. This is the upper bound for the function $P_{ud}(K_{AB}, .)$ as well.

5. Conclusions

Repetition of the message is a natural and undemanding method of protecting its content. In the safety relevant applications it is not a sufficient technique. Therefore, more sophisticated variations of this principle can be useful as additional defence.

Providing the probabilistic analysis of the code using some of these variants of message doubling is surprisingly complicated. Nevertheless, an effective, though not elegant, method for necessary computations was developed.

References

[1] EN 50159 Railway applications – Communication, signalling and processing systems – Safety-related communication in transmission systems. European standard, CENELEC, September 2010.

[2] Huffman, W. C. and Pless, V.: *Fundamentals of error-correcting codes*. Cambridge University Press, Cambridge, 2003.

[3] Sweeney, P.: *Error control coding. From theory to practice*. John Wiley & Sons, 2002.

Conference Applications of Mathematics 2015, in honor of the birthday anniversaries of Ivo Babuška (90), Milan Práger (85), and Emil Vitásek (85)
J. Brandts, S. Korotov, M. Křížek, K. Segeth, J. Šístek, T. Vejchodský (Eds.)
Institute of Mathematics AS CR, Prague 2015

WILDLAND FIRE PROPAGATION MODELLING: A NOVEL APPROACH RECONCILING MODELS BASED ON MOVING INTERFACE METHODS AND ON REACTION-DIFFUSION EQUATIONS

Inderpreet Kaur[1], Andrea Mentrelli[1,2], Frederic Bosseur[3], Jean Baptiste Filippi[3], Gianni Pagnini[1,4]

[1] BCAM - Basque Center for Applied Mathematics
Alameda de Mazarredo 14, 48009 Bilbao, Basque Country – Spain
ikaur@bcamath.org
[2] Department of Mathematics and AM^2, University of Bologna
Via Saragozza 8, 40123 Bologna, Italy
andrea.mentrelli@unibo.it
[3] SPE–CNRS/University of Corsica, Corte, Corsica – France
fbosseur@gmail.com; filippi@univ-corse.fr
[4] Ikerbasque
Calle de María Díaz de Haro 3, 48013 Bilbao, Basque Country – Spain
gpagnini@bcamath.org

Abstract: A novel approach to study the propagation of fronts with random motion is presented. This approach is based on the idea to consider the motion of the front, split into a drifting part and a fluctuating part; the front position is also split correspondingly. In particular, the drifting part can be related to existing methods for moving interfaces, for example, the Eulerian level set method and the Lagrangian discrete event system specification. The fluctuating part is the result of a comprehensive statistical description of the system which includes the random effects in agreement with the physical properties of the system. The resulting averaged process emerges to be governed by an evolution equation of the reaction-diffusion type. Hence, following the proposed approach, when fronts propagate with a random motion, models based on methods for moving interfaces and those based on reaction-diffusion equations can indeed be considered complementary and reconciled. This approach turns out to be useful to simulate random effects in wildland fire propagation as those due to turbulent heat convection and fire spotting phenomena.

Keywords: random front, wildland fire propagation, turbulence, fire spotting
MSC: 60K37, 62P12, 62P35, 82Dxx

1. Introduction

Modelling moving interfaces is an important issue in many research fields and in several real world applications. In many natural phenomena the front propagates into systems characterized by randomness and therefore the motion of the front gets a random character. Here a novel formulation for modelling random front motion is presented and its application to wildland fire propagation discussed.

Wildland fire propagation is a complex multi-scale, as well as a multi-physics and multi-discipline process, strongly influenced by the atmospheric wind. Wildland fire is fed by the fuel on the ground and displaced, beside meteorological and orographical factors, also by the hot air that pre-heats the fuel and aids the fire propagation. Heat transfer is turbulent due to the heat release in the Atmospheric Boundary Layer and the fire-induced flow. Moreover, fire generates firebrands which after landing on the ground act as new sources of fire. Both turbulence and jump-length of firebrands are random processes that affect the fireline propagation.

Fire propagation has been mainly modelled in the literature by using methods for simulating moving interfaces as the Eulerian level set method (LSM) [17], see e.g. [6, 7], or the Lagrangian discrete event system specification (DEVS) [4, 11] with the fire propagation solver ForeFire, see e.g. [3, 2], and reaction-diffusion type equations, see e.g. [1, 8].

These two approaches, namely that based on moving interface methods and that based on reaction-diffusion equations, are considered alternatives to each other because the solution of the reaction-diffusion equation is generally a continuous smooth function that has an exponential decay, and it is not zero in an infinite domain, while methods for simulating moving interfaces are associated to an indicator function that is 1 in the inner domain and 0 outside. However, when random processes (as for example hot air turbulent convection and fire spotting) are taken into account according to the proposed formulation, these two approaches can indeed be considered complementary and reconciled.

2. Random front model formulation

The proposed approach is based on the idea to consider the motion of the front split into a drifting part and a fluctuating part and the front position is split correspondingly. This splitting allows specific numerical and physical choices that can improve the algorithms and the models. In particular, the drifting part can be related to existing methods for moving interfaces, for example, the Eulerian LSM [17] or the Lagrangian DEVS [4, 11], and this permits the choice of the best method for any specific application. The fluctuating part is the result of a comprehensive statistical description of the system which includes the random effects in agreement with the physical properties of the system.

The resulting averaged process emerges to be governed by an evolution equation of the reaction-diffusion type. Hence, following the proposed approach, when fronts propagate with a random motion, models based on methods for moving interfaces and

those based on reaction-diffusion equations can indeed be considered complementary and reconciled.

Let Γ be a simple closed curve, or an ensemble of simple non-intersecting closed curves, representing a propagating interface in two dimensions, and let S be the domain of interest $S \subseteq R^2$. In the case of Γ being an ensemble of n curves, the ensemble of the n interfaces is considered to be an *interface*.

The subset of the domain S corresponding to the region Ω enclosed by Γ may be conveniently identified by an indicator function $I_\Omega : S \times [0, +\infty[\rightarrow \{0, 1\}$ defined as follows

$$I_\Omega(\boldsymbol{x}, t) = \begin{cases} 1\,, & \boldsymbol{x} \in \Omega\,, \\ \\ 0\,, & \text{elsewhere}\,. \end{cases} \tag{1}$$

In the case of a front line Γ made of more than one closed curve, the domain Ω is not simply connected, resulting in more than one sourrounded area evolving independently.

The indicator function I_Ω at time $t = 0$, i.e. $I_\Omega(\boldsymbol{x}, t = 0)$, describing the initial topology of the front, is indicated in the following as $I_{\Omega_0}(\boldsymbol{x}_0)$.

Let $\boldsymbol{X}^\omega(t, \overline{\boldsymbol{x}}_0) = \overline{\boldsymbol{x}}(t, \overline{\boldsymbol{x}}_0) + \eta^\omega$ be the ω-realization of a random trajectory driven by the random noise η. For every realization, the initial condition is stated to be $\boldsymbol{X}^\omega(0, \overline{\boldsymbol{x}}_0) = \overline{\boldsymbol{x}}_0$. Using the sifting property of δ-function, i.e. $g(\boldsymbol{x}) = \int g(\overline{\boldsymbol{x}})\, \delta(\boldsymbol{x} - \overline{\boldsymbol{x}})\, d\overline{\boldsymbol{x}}$, the evolution in time of the ω-realization of a random front contour $\gamma^\omega(\boldsymbol{x}, t)$ is given by

$$\gamma^\omega(\boldsymbol{x}, t) = \int_S \gamma(\overline{\boldsymbol{x}}_0)\, \delta(\boldsymbol{x} - \boldsymbol{X}^\omega(t, \overline{\boldsymbol{x}}_0))\, d\overline{\boldsymbol{x}}_0\,, \tag{2}$$

which in terms of the random indicator $I_{\Omega^\omega}(\boldsymbol{x}, t)$ reads

$$\begin{aligned} I_{\Omega^\omega}(\boldsymbol{x}, t) &= \int_S I_{\Omega_0}(\overline{\boldsymbol{x}}_0)\, \delta(\boldsymbol{x} - \boldsymbol{X}^\omega(t, \overline{\boldsymbol{x}}_0))\, d\overline{\boldsymbol{x}}_0 \\ &= \int_{\Omega_0} \delta(\boldsymbol{x} - \boldsymbol{X}^\omega(t, \overline{\boldsymbol{x}}_0))\, d\overline{\boldsymbol{x}}_0 = \int_{\Omega(t)} \delta(\boldsymbol{x} - \boldsymbol{X}^\omega(t, \overline{\boldsymbol{x}}))\, d\overline{\boldsymbol{x}}\,, \end{aligned} \tag{3}$$

where an incompressibility-like condition $\dfrac{d\overline{\boldsymbol{x}}_0}{d\overline{\boldsymbol{x}}} = 1$ is assumed.

Let $\varphi_e(\boldsymbol{x}, t) : S \times [0, +\infty[\rightarrow [0, 1]$ be an *effective indicator*. It may be defined as

$$\begin{aligned} \varphi_e(\boldsymbol{x}, t) = \langle I_{\Omega^\omega}(\boldsymbol{x}, t) \rangle &= \left\langle \int_{\Omega(t)} \delta(\boldsymbol{x} - \boldsymbol{X}^\omega(t, \overline{\boldsymbol{x}}))\, d\overline{\boldsymbol{x}} \right\rangle = \int_{\Omega(t)} \langle \delta(\boldsymbol{x} - \boldsymbol{X}^\omega(t, \overline{\boldsymbol{x}})) \rangle\, d\overline{\boldsymbol{x}} \\ &= \int_{\Omega(t)} f(\boldsymbol{x}; t | \overline{\boldsymbol{x}})\, d\overline{\boldsymbol{x}} = \int_S I_\Omega(\overline{\boldsymbol{x}}, t) f(\boldsymbol{x}; t | \overline{\boldsymbol{x}})\, d\overline{\boldsymbol{x}}\,, \end{aligned} \tag{4}$$

where $\langle\cdot\rangle$ is the ensemble average and $f(\boldsymbol{x};t|\overline{\boldsymbol{x}}) = \langle\delta(\boldsymbol{x}-\boldsymbol{X}^{\omega}(t,\overline{\boldsymbol{x}}))\rangle$ is the probability density function (PDF) of fluctuations of the perimeter around the contour $\Gamma(t)$.

Since the present approach is formulated to study the effects of an underlying diffusive process in front propagation, according to classical properties of diffusion, the resulting PDF $f(\boldsymbol{x};t|\overline{\boldsymbol{x}})$ of the stochastic process $\boldsymbol{X}^{\omega}$ is considered to be unimodal and its mean and median are coincident. This means that $f(\boldsymbol{x};t|\overline{\boldsymbol{x}})$ is a symmetric probability distribution which normalizes after integration both over $\boldsymbol{x}$ and $\overline{\boldsymbol{x}}$. Consequently, values of the effective indicator $\varphi_e(\boldsymbol{x},t)$ range in the compact interval $[0,1]$.

The front line $\Gamma(t)$ can be obtained by existing methods for moving interfaces, as for example the already mentioned LSM or DEVS. For a deterministic motion, it holds $f(\boldsymbol{x};t|\overline{\boldsymbol{x}}) = \delta(\boldsymbol{x}-\overline{\boldsymbol{x}})$ and the result reduces to that of the chosen moving interface method, i.e. $I_{\Omega}(\boldsymbol{x},t)$.

The evolution of the effective indicator $\varphi_e(\boldsymbol{x},t)$ can be estimated by applying in (4) the Reynolds transport theorem [12]

$$\frac{\partial\varphi_e}{\partial t} = \int_{\Omega(t)} \frac{\partial f}{\partial t}\, d\overline{\boldsymbol{x}} + \int_{\Omega(t)} \nabla_{\overline{\boldsymbol{x}}}\cdot[\boldsymbol{V}(\overline{\boldsymbol{x}},t) f(\boldsymbol{x};t|\overline{\boldsymbol{x}})]\, d\overline{\boldsymbol{x}}\,. \tag{5}$$

Let $f(\boldsymbol{x};t|\overline{\boldsymbol{x}})$ be the solution of the evolution equation,

$$\frac{\partial f}{\partial t} = \mathcal{E} f\,, \quad f(\boldsymbol{x};0|\overline{\boldsymbol{x}}_0) = \delta(\boldsymbol{x}-\overline{\boldsymbol{x}}_0)\,, \tag{6}$$

with $\mathcal{E} = \mathcal{E}(\boldsymbol{x})$ a generic evolution operator not acting on both $\overline{\boldsymbol{x}}$ and t, then equation (5) becomes the following reaction-diffusion type equation

$$\frac{\partial\varphi_e}{\partial t} = \mathcal{E}\varphi_e + \int_{\Omega(t)} \nabla_{\overline{\boldsymbol{x}}}\cdot[\boldsymbol{V}(\overline{\boldsymbol{x}},t) f(\boldsymbol{x};t|\overline{\boldsymbol{x}})]\, d\overline{\boldsymbol{x}}\,, \tag{7}$$

where $\boldsymbol{V}(\boldsymbol{x},t)$ is the expansion velocity of the domain $\Omega(t)$ determined by $d\overline{\boldsymbol{x}}/dt = \boldsymbol{V}(\overline{\boldsymbol{x}},t) = \mathcal{V}(\overline{\boldsymbol{x}},t)\,\widehat{\boldsymbol{n}}$ and $\widehat{\boldsymbol{n}}$ is the normal to the front contour.

Finally, the front line is obtained by choosing an arbitrary threshold value φ_e^{th} which serves as the criterion to mark the inner region $\Omega_e(t) = \left\{\boldsymbol{x}\in S|\varphi_e(\boldsymbol{x},t) > \varphi_e^{th}\right\}$.

The above formulation has been considered for applications to diffusive media governed by fractional differential equations [9, 10]. In the following section, the application to wildland fire propagation is discussed.

3. Application to wildland fire propagation

In wildland fire propagation modelling, both the LSM and DEVS are adopted to simulate the evolution of the burning area, see e.g. [6, 7] and [3, 2], respectively. The

present approach can be used with both the methods to include random processes such as turbulence and fire spotting.

In particular, let $\boldsymbol{X}^{\omega}(t, \overline{\boldsymbol{x}}_0) = \overline{\boldsymbol{x}}(t, \overline{\boldsymbol{x}}_0) + \chi^{\omega} + \xi^{\omega}$ be the ω-realization of a random trajectory driven by the random noises χ and ξ corresponding to turbulence and fire spotting, respectively. For every realization, the initial condition is stated to be $\boldsymbol{X}^{\omega}(0, \overline{\boldsymbol{x}}_0) = \overline{\boldsymbol{x}}_0$. Average turbulent fluctuations are zero, i.e. $\langle \chi \rangle = 0$, and fire spotting is assumed to be independent of turbulence and to be a downwind phenomenon such that $\xi^{\omega} = \ell^{\omega}\, \widehat{\boldsymbol{n}}_U$, where ℓ is the landing distance from the main fireline such that $\langle \ell \rangle > 0$ and $\widehat{\boldsymbol{n}}_U$ is the mean wind direction.

The modelling of the random processes is handled by the PDF $f(\boldsymbol{x}; t|\overline{\boldsymbol{x}})$, accounting for the sum of the two independent random variables $(\overline{\boldsymbol{x}} + \chi)$ and ξ, representing turbulence and fire spotting respectively. This means that $f(\boldsymbol{x}; t|\overline{\boldsymbol{x}})$ is determined by the convolution between the PDF corresponding to $(\overline{\boldsymbol{x}} + \chi)$, hereinafter labeled as G, and the PDF corresponding to ξ, hereinafter labeled as q.

Since fire spotting is assumed to be an independent downwind phenomenon, the effect of fire spotting is accounted for only the leeward part of the fireline. Taking into account previous assumptions $f(\boldsymbol{x}; t|\overline{\boldsymbol{x}})$ results in

$$f(\boldsymbol{x}; t|\overline{\boldsymbol{x}}) = \begin{cases} \displaystyle\int_0^{\infty} G(\boldsymbol{x} - \overline{\boldsymbol{x}} - \ell \widehat{\boldsymbol{n}}_U; t)\, q(\ell; t)\, d\ell\,, & \text{if} \quad \widehat{\boldsymbol{n}} \cdot \widehat{\boldsymbol{n}}_U \geq 0\,, \\ G(\boldsymbol{x} - \overline{\boldsymbol{x}}; t)\,, & \text{otherwise}\,. \end{cases} \tag{8}$$

Since the effective fireline contour $\varphi_e(\boldsymbol{x}, t)$ is a smooth function continuously ranging from 0 to 1, a criterion to mark burned points have to be stated. For example, points $\boldsymbol{x}$ such that $\varphi_e(\boldsymbol{x}, t) > \varphi_e^{th} = 0.5$ are marked as burned and the effective burned area emerges to be $\Omega_e(t) = \{\boldsymbol{x} \in S | \varphi_e(\boldsymbol{x}, t) > \varphi_e^{th} = 0.5\}$. However, beside this criterion, a further criterion associated to an ignition delay due to the pre-heating action of the hot air or to the landing of firebrands is introduced. Hence, in the proposed modelling approach, an unburned point $\boldsymbol{x}$ will be marked as burned when one of these two criteria is met.

This ignition delay, due to a certain *heating-before-burning mechanism*, can be depicted as an accumulation in time of heat [13, 14], i.e.

$$\psi(\boldsymbol{x}, t) = \int_0^t \varphi_e(\boldsymbol{x}, s)\, \frac{ds}{\tau}\,, \tag{9}$$

where $\psi(\boldsymbol{x}, 0) = 0$ corresponds to the unburned initial condition and τ is a characteristic ignition delay that can be understood as an electrical resistance. Since the fuel can burn because of two independent pathways, i.e. hot-air heating and firebrand

landing, the resistance analogy suggests that τ can be approximatively computed as resistances acting in parallel, i.e.

$$\frac{1}{\tau} = \frac{1}{\tau_h} + \frac{1}{\tau_f} = \frac{\tau_f + \tau_h}{\tau_h \tau_f}\,, \tag{10}$$

where τ_h and τ_f are the ignition delays due to hot air and firebrands, respectively.

The amount of heat is proportional to the increasing of the fuel temperature $T(\boldsymbol{x}, t)$, then

$$\psi(\boldsymbol{x}, t) \propto \frac{T(\boldsymbol{x}, t) - T(\boldsymbol{x}, 0)}{T_{ign} - T(\boldsymbol{x}, 0)}\,, \quad T(\boldsymbol{x}, t) \leq T_{ign}\,, \tag{11}$$

where T_{ign} is the ignition temperature.

Finally, when $\psi(\boldsymbol{x}, t) = 1$ the ignition temperature is assumed to be reached, so that a new ignition occurs in $(\boldsymbol{x}, t)$ and, with reference to (4), the modelled fire goes on by setting $I_\Omega(\boldsymbol{x}, t) = 1$. Then, as a consequence of the heating-before-burning mechanism described in (9), the domain $\Omega(t)$ is established as $\Omega(t) = \{\boldsymbol{x} \in S | I_\Omega(\boldsymbol{x}, t) = 1\}$ which is hard to be analytically evaluated but numerically computed only. The expansion velocity of the domain $\Omega(t)$ in the normal direction is stated by means of the prescription of the so-called Rate Of Spread (ROS).

4. Numerical simulations

To estimate the performance of the LSM based model and DEVS based model coupled with the random processes a series of simulation experiments are conducted. For LSM, a formulation developed in References [7, 6] is followed, while for DEVS, ForeFire fire simulator [3] is used. Both these models have a different formulation to incorporate the nature of vegetation and slope hence, it is tried to parametrise both models in an identical setup.

In the present study, for brevity no particular type of vegetation is defined and simulations are carried out with a pre-defined value of ROS. It is assumed that the ROS remains constant for a particular terrain. It is emphasised that these are simplified and idealised test cases and no attempt is made to choose the parameters for a realistic setup. The present scope of this work is to provide a first look into the investigation of comparing LSM and DEVS based fire simulations with random processes.

A flat area of hypothetical homogeneous vegetation spread over a domain size of $5000\,\mathrm{m} \times 5000\,\mathrm{m}$ is selected for simulations. Different values of the ROS are utilised for different test cases. The ROS is assumed to be $0.05\,\mathrm{ms}^{-1}$ in no wind conditions while, in the presence of wind, it is estimated by the 3% ROS model [2]:

$$ROS = 0.03\,\boldsymbol{U} \cdot \widehat{\boldsymbol{n}}\,, \tag{12}$$

where, $\boldsymbol{U}$ is the mean wind velocity. Since, 3% ROS considers the propagation only towards the mean wind direction, in order to study the evolution of the fireline towards the flank and rear the following ROS is also considered [6]:

$$ROS(U,\theta) = \begin{cases} \varepsilon_o + a\sqrt{U\cos^n\theta}, & \text{if } |\theta| \leq \frac{\pi}{2}\,, \\ \varepsilon_o(\alpha + (1-\alpha)|\sin\theta|), & \text{if } |\theta| > \frac{\pi}{2}\,, \end{cases} \tag{13}$$

where, ε_o is the flank velocity and $(\alpha\varepsilon_o)$ is the rear velocity with $\alpha \in [0,1]$, and θ is defined as the angle between the normal to the front $\widehat{\boldsymbol{n}}$ and the mean wind direction $\widehat{\boldsymbol{n}}_U$. For the present setup, we assume the values $\alpha = 0.8$, $n = 3$, $a = 0.5\,\mathrm{m}^{1/2}\mathrm{s}^{-1/2}$, $\varepsilon_o = 0.2\,\mathrm{ms}^{-1}$.

In LSM, the domain is discretised with a Cartesian grid of 20 m in both x and y directions, while for DEVS the resolution of the simulation is defined in the terms of quantum distance Δq and perimeter resolution Δc [3]. The quantum distance Δq is defined as the maximum allowable distance to be covered by a particle at each advance, while a measure of Δc is used to decompose/regenerate/coalesce two particles on propagation. The choice of Δq and Δc is *dependent on the type of problem*, and in the present study two sets of values are used. The simulation is performed with $\Delta q = 4\,\mathrm{m}$, $\Delta c = 18\,\mathrm{m}$ for zero wind; and $\Delta q = 0.3\,\mathrm{m}$, $\Delta c = 8\,\mathrm{m}$ in the presence of wind. To avoid instability in the presence of wind, Δq is chosen to be of a much higher resolution than the wind data (20 m in this setup). The time is advanced according to the events in ForeFire, and the simulation can move ahead according to a user defined time. To facilitate a comparison between the two models, the simulation in DEVS model is advanced by the time step computed according to the Courant–Friedrichs–Lewy (CFL) criteria in LSM.

The mean wind, wherever used, is assumed to be constant in magnitude and direction. The turbulence is modelled according to a bi-variate Gaussian PDF

$$G(\boldsymbol{x} - \overline{\boldsymbol{x}}; t) = \frac{1}{2\pi\sigma^2(t)} \exp\left\{ -\frac{(x-\overline{x})^2 + (y-\overline{y})^2}{2\,\sigma^2(t)} \right\}, \tag{14}$$

where σ^2 is the particle displacement variance related to the turbulent diffusion coefficient $\mathcal{D}$, such that $\langle (x-\overline{x})^2 \rangle = \langle (y-\overline{y})^2 \rangle = \sigma^2(t) = 2\mathcal{D}t$. In the present model, the whole effect of the turbulent processes over different scales is assumed to be parametrised by the turbulent diffusion coefficient only.

The phenomenon of fire spotting is included according to the discussion provided in References [16] and [5]. In Reference [16] it is shown that the firebrand distribution followes a bimodal distribution but only the contribution of the firebrands with short landing distance is significant because the ones with long-distance landing reach the ground in charred oxidation state. According to this, the distribution of the firebrands follows a log-normal distribution [16]

$$q(\ell; t) = \frac{1}{\sqrt{2\pi}\, s(t)\ell} \exp\left\{ -\frac{(\ln\ell - \mu(t))^2}{2\, s^2(t)} \right\}, \tag{15}$$

where, $\mu(t) = \langle \ln \ell \rangle$ and $s(t) = \langle (\ln \ell - \mu(t))^2 \rangle$ are the mean and the standard deviation of $\ln \ell$ respectively. They are stated to be [15]

$$\mu = 1.32 I_f^{0.26} U^{0.11} - 0.02 \tag{16}$$

$$s = 4.95 I_f^{-0.01} U^{-0.02} - 3.48 \tag{17}$$

where U is the magnitude of the mean wind and $I_f = I + I_t$ represents fire intensity I enriched by the tree torching intensity I_t. The turbulent diffusion coefficient $\mathcal{D}$ and ignition delay τ are also assumed to be constant throughout the simulations. The value of $\mathcal{D}$ is assumed to be $0.15\,\mathrm{m^2 s^{-1}}$ and the ignition delay for hot air and firebrand is fixed at 600 s and 60 s respectively. The initial fire intensity is assumed to be $10\,000\,\mathrm{kW m^{-1}}$ and the tree torching intensity is fixed at $0.015\,\mathrm{kW m^{-1}}$.

A series of idealised simulation tests are made to investigate the behaviour of the two models with identical initial conditions. Different simulations are performed both in the presence and the absence of wind by neglecting and considering the random processes. The first case evaluates an isotropic growth of the fireline for zero wind in both the models by neglecting all the random processes. In the second test, the spread of fireline for different ROS in different directions is studied. The third test discusses the propagation of the fireline with wind when the ROS is defined according to the 3% formula (12) and to formula (13). The random processes are neglected for the first three test cases. The fourth test presents the evolution of fireline when turbulent processes are included both in the presence and absence of wind. The last test evaluates the performance when fire spotting also included along with turbulence. Firebreak lines are also introduced in the last two tests to highlight the propagation of the fireline while encountering areas of null fuel. It should be noted that for brevity fire spotting is assumed to be an independent downwind phenomenon. Hence, the effect of fire spotting is accounted for only the leeward part of the fireline. Also, to simplify the simulation, the region across and behind the centre of the initial fireline is demarcated as the leeward side and the windward side respectively.

5. Discussion

Figure 1 presents the evolution of the fireline for a circular initial condition of radius 300 m for both LSM based model and DEVS based model. In the absence of the wind, the initial circular fireline is transported into a isotropic growth, and the circular contours correspond to 40 min isochronous fronts. It can very well appreciated that with the same value of ROS and initial conditions, the two different tracking schemes provide an identical evolution of the fireline. Modelling real situations of fire involves presence of zones without fuel, where the ROS is zero. In case of firebreak zones, pure LSM and DEVS are inherently unable to simulate the realistic situations of fire overcoming a fire break. Figure 2 shows that the fireline fails to propagate across the firebreak when no random processes are included. The

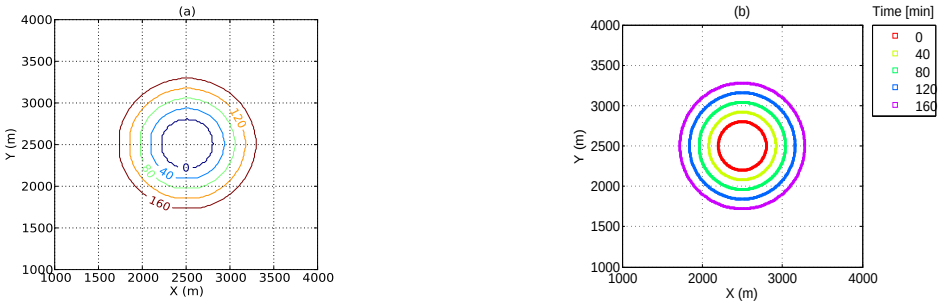


Figure 1: Evolution in time of the fireline contour without random processes and zero wind for a) LSM and b) ForeFire. The initial fireline is a circle of radius 300 m.

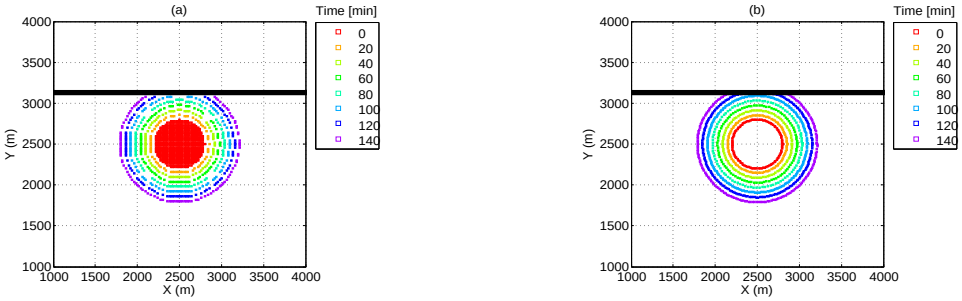


Figure 2: Same as Figure 1, but in the presence of a firebreak zone (60 m wide).

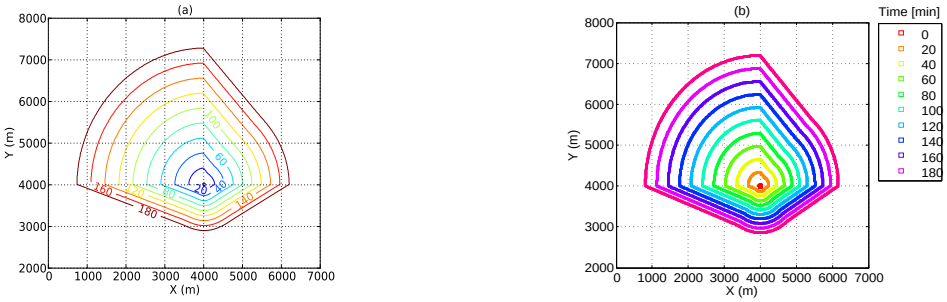


Figure 3: Same as Figure 1, but with a non-homogeneous ROS. The ROS is 0.3 ms^{-1}, 0.2 ms^{-1}, 0.1 ms^{-1} in upper-left, upper right and bottom quadrants respectively.

evolution is shown only upto 140 min, but an extended run upto 250 min indicates the limitation of the models to simulate the fire jump across the break zone.

Figure 3 presents the growth of an initial spot fire but with a non-homogenous ROS in absence of any wind. The different values of the ROS can be attributed to different fuel types. The fireline propagates with different speed in the three directions. In the absence of wind, the two tracking methods show an identical behaviour in simulating situations with constant ROS. This paves way for a comparison of situations with higher variability and complexity.

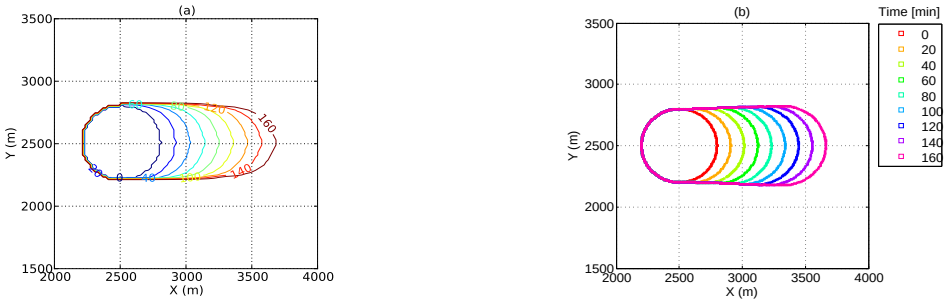


Figure 4: Same as Figure 1, but when the mean wind velocity is $3\,\mathrm{ms}^{-1}$ in the positive x-direction.

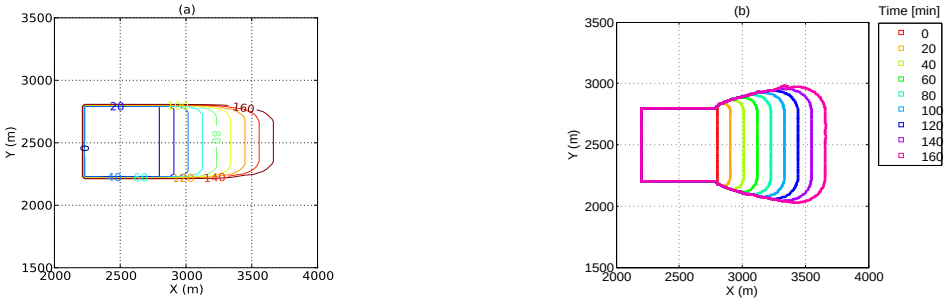


Figure 5: Same as Figure 4, but when the initial profile is square with side 600 m.

Figure 4 shows the evolution of the firefront with a circular initial profile (with radius 300 m) in case of a light wind of $3\,\mathrm{ms}^{-1}$ directed in the positive x direction. The isochronous fronts are plotted at every 20 min and follow an oval shape for both the models. The fire contours in DEVS based model diverge slightly from the mean wind field and an increasing flanking fire develops over time. This divergence in the evolution of firefront occurs due to differences in the computation of the normal for both the models. The computation of normal for an active fire point in DEVS model is approximated as the measure of the bisector of the angle between the point and its left and right neighbours. This fact can be very well appreciated when an initial square profile is considered.

Figure 5 shows the evolution of the fire spread with square initial profile of side 600 m. Under the effect of the constant zonal wind, the evolution in LSM strictly follows the initial square shape, while in ForeFire the initial angular points are transported to provide a flanking spread. The 3% ROS does not model the rear and back fire but DEVS generates spurious flanking fire that gives a realistic behaviour to the fire spread even if due to the approximate construction of the front normal. The differences in the evolution of flank fireline are also studied by introducing different ROS for the head, flank and rear directions according to (13). Here θ is defined as the angle between the normal to the front and the mean wind direction. Since the normal computation in DEVS approach is approximate, two

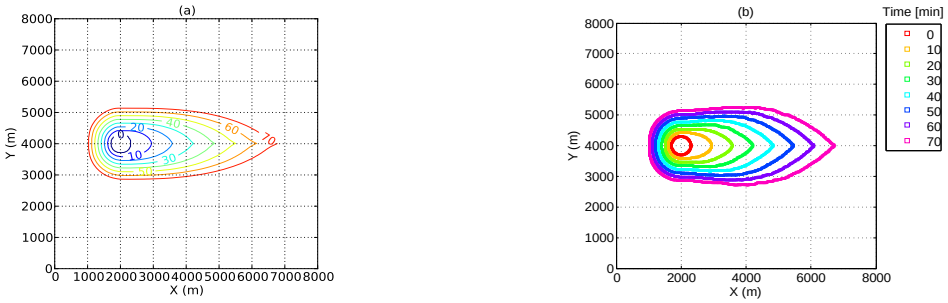


Figure 6: Evolution in time of the fireline contour without random processes with ROS given by formula (13) where θ is the angle between the outward normal in a contour point and the mean wind direction for a) LSM and b) ForeFire. The mean wind velocity is $3\,\mathrm{ms}^{-1}$ in the positive x-direction.

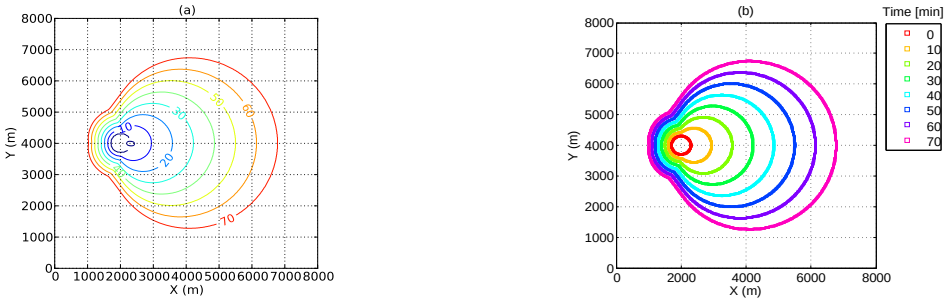


Figure 7: Same as Figure 6, but where θ is the angle between the line joining a contour point and the centre of the initial burned area.

separate tests are performed to evaluate effect of the normal on the spread: firstly when θ is computed according to the definition, and secondly, to ensure identical angle for both methods, when θ is assumed to be the angle between the line joining a contour point and the centre of the initial burned area. Figure 6 shows that when θ is computed in accordance to the definition, the simulations for head and rear fires are identical, but spurious flanks are observed for DEVS based model. On the other hand, it is evident from Figure 7 that identical values of θ shows an identical propagation of the fireline in all directions.

As shown above, in case of firebreak zones pure LSM and ForeFire are inherently unable to simulate the realistic situations of fire overcoming a fire break. But Figure 8 shows that with the introduction of turbulence the models can simulate the effect of hot air to overcome firebreak zone. The value of turbulent diffusion coefficient is assumed to be $0.15\,\mathrm{m}^2\mathrm{s}^{-1}$. The evolution of the fireline is almost similar for both the models, though a slight underestimation is visible in ForeFire with respect to the LSM based model. Stronger turbulence causes a more rapid propagation of the fireline and an earlier ignition across the firebreak zone. A detailed analysis of the effect of varying turbulence over *long-term propagation* with the LSM can be found in [13, 14].

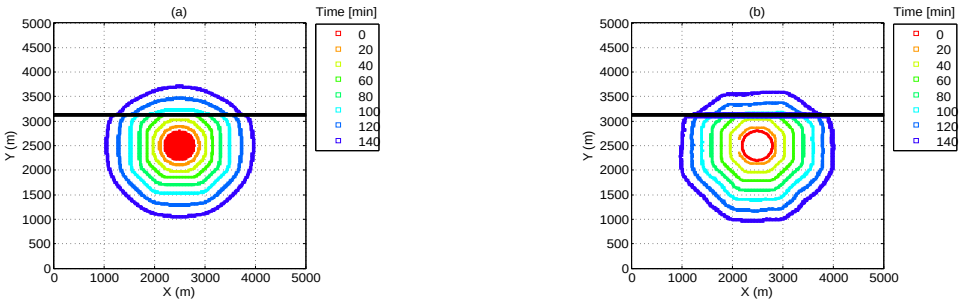


Figure 8: Evolution in time of the fireline contour with turbulence in zero wind for a) LSM and b) ForeFire. The initial fireline is a circle of radius 300 m. The turbulent diffusion coefficient is $0.15\,\mathrm{m^2 s^{-1}}$.

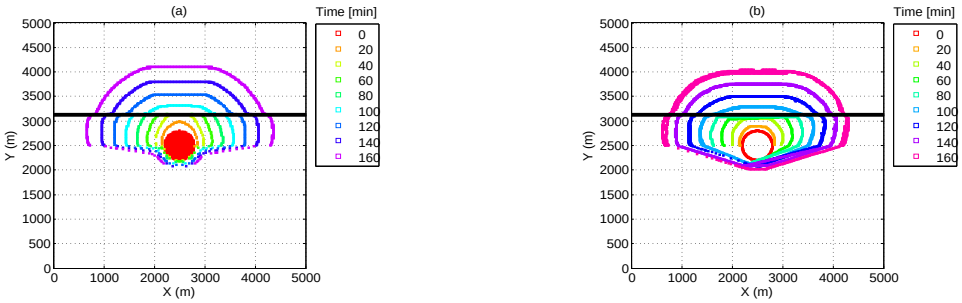


Figure 9: Same as Figure 8, but when the mean wind velocity is $3\,\mathrm{ms^{-1}}$ in positive y-direction.

Figure 9 presents the effect of inclusion of turbulence with a non-zero wind profile and 3%-wind ROS. The effect of turbulence is most pronounced in the direction of the mean wind and it clearly shows that randomisation of the fireline permits the fire to overcome the obstacle without fuel along with an increased growth in the flank-fire, back fire and head fire. Both models show almost similar characteristics in modelling the spread of fire, though the flank fire has a slightly larger spread in ForeFire. This is due to the fact the particle transportation in the direction of the front normal is computed with an approximated method.

Another aspect contributing towards the increase in the fire spread and allowing new fire ignitions across obstacles due to fire spotting is presented in Figure 10. With inclusion of fire spotting along with turbulence, the evolution of the fire front is faster in comparison to the effect of turbulence alone as seen in Figure 9. The region across the fire break has a quick ignition pertaining to the embers flowing and landing with the effect of wind. It should be noted that within the considered parametrisation (15) the phenomenon of fire spotting can only be observed in the presence of the wind. The flank fire and the head fire are also well simulated in both the models, and again a larger spread out the flanking fires is observed for DEVS. Fire spotting along with turbulence has a remarkable effect on enhancing the fireline and igniting secondary fires across the fire break zones.

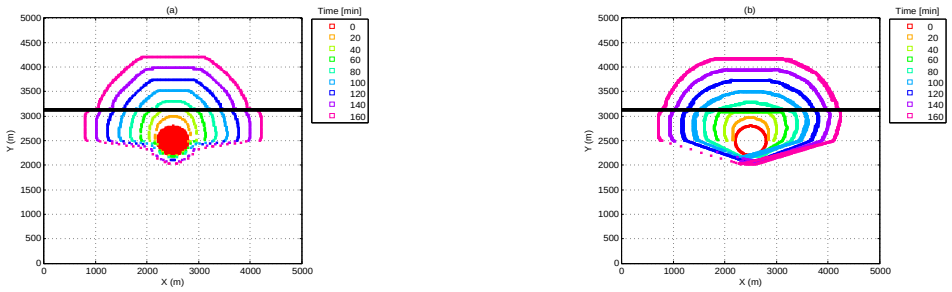


Figure 10: Same as Figure 9, but when phenomenon of fire-spotting is also included.

6. Conclusions

This paper describes an approach to model the effects of the random processes in the propagation of the wildland fires. The propagation of the fire can be split into a drifting part and a fluctuating part. The fluctuating part is generated by a comprehensive statistical description of the system and includes the effects of random processes in agreement with the physical properties of the system.

The drifting part is modelled in terms of a deterministic position determined by Eulerian LSM or Lagrangian DEVS with a certain ROS, and the fluctuating part according to the PDF of the random displacement of points marked as active burning points. Numerical simulations show that this formulation emerges to be suitable for both LSM and DEVS based models to manage the real world situations related to random character of fire e.g., increase in ROS due to pre-heating of the fuel by hot air and vertical lofting and transporting of firebrands and fire overcoming no fuel zones. DEVS computes an approximated outward normal of the fire perimeter that generates differences with the LSM. Such differences result in spurious flanking fires, which however provide a more realistic fire contour. The two models perform in agreement with each other and can be complementary to each other for simple situations, but for increasing complexity the introduction of random processes amplifies differences between DEVS and LSM which are mainly due to the approximate computation of normal.

Acknowledgements

This research is supported by MINECO under Grant MTM2013-40824-P, by Bizkaia Talent and European Commission through COFUND programme under Grant AYD-000-226, and also by the Basque Government through the BERC 2014–2017 program and by the Spanish Ministry of Economy and Competitiveness MINECO: BCAM Severo Ochoa accreditation SEV-2013-0323.

References

[1] Asensio, M. I. and Ferragut, L.: On a wildland fire model with radiation. Int. J. Numer. Meth. Engng. **54** (2002), 137–157.

[2] Filippi, J. B., Mallet, V., and Nader, B.: Evaluation of forest fire models on a large observation database. Nat. Hazards Earth Syst. Sci. **14** (2014), 3077–3091.

[3] Filippi, J. B., Morandini, F., Balbi, J. H., and Hill, D.: Discrete event front tracking simulator of a physical fire spread model. Simulation **86** (2010), 629–646.

[4] Karimabadi, H., Driscoll, J., Omelchenko, Y. A., and Omidi, N.: A new asynchronous methodology for modeling of physical systems: breaking the curse of Courant condition. J. Comp. Phys. **205** (2005), 755–775.

[5] Kortas, S., Mindykowski, P., Consalvi, J. L., Mhiri, H., and Porterie, B.: Experimental validation of a numerical model for the transport of firebrands. Fire Safety J. **44** (2009), 1095–1102.

[6] Mallet, V., Keyes, D. E., and Fendell, F. E.: Modeling wildland fire propagation with level set methods. Comput. Math. Appl. **57** (2009), 1089–1101.

[7] Mandel, J., Beezley, J. D., and Kochanski, A. K.: Coupled atmosphere-wildland fire modeling with WRF 3.3 and SFIRE 2011. Geosci. Model. Dev. **4** (2011), 591–610.

[8] Mandel, J. et al.: A wildland fire model with data assimilation. Math. Comput. Simulat. **79** (2008), 584–606.

[9] Mentrelli, A. and Pagnini, G.: Random front propagation in fractional diffusive systems. Communications in Applied and Industrial Mathematics (2014). http://dx.doi.org/10.1685/journal.caim.504.

[10] Mentrelli, A. and Pagnini, G.: Front propagation in anomalous diffusive media governed by time-fractional diffusion. J. Comp. Phys. **293** (2015), 427–441.

[11] Omelchenko, Y. A. and Karimabadi, H.: HYPERS: A unidimensional asynchronous framework for multiscale hybrid simulations. J. Comp. Phys. **231** (2012), 1766–1780.

[12] Pagnini, G. and Bonomi, E.: Lagrangian formulation of turbulent premixed combustion. Phys. Rev. Lett. **107** (2011), 044 503.

[13] Pagnini, G. and Massidda, L.: The randomized level-set method to model turbulence effects in wildland fire propagation. In: D. Spano, V. Bacciu, M. Salis, and C. Sirca (Eds.), *Modelling Fire Behaviour and Risk. Proceedings of the International Conference on Fire Behaviour and Risk. ICFBR 2011, Alghero, Italy, October 4–6, 2011*, 126–131, 2012. ISBN 978-88-904409-7-7.

[14] Pagnini, G. and Massidda, L.: Modelling turbulence effects in wildland fire propagation by the randomized level-set method. Tech. Rep. 2012/PM12a, CRS4, 2012. Revised version August 2014, arXiv:1408.6129.

[15] Perryman, H. A., Dugaw, C. J., Varner, J. M., and Johnson, D. L.: A cellular automata model to link surface fires to firebrand lift-off and dispersal. Int. J. Wildland Fire **22** (2013), 428–439.

[16] Sardoy, N., Consalvi, J. L., Kaiss, A., Fernandez-Pello, A. C., and Porterie, B.: Numerical study of ground-level distribution of firebrands generated by line fires. Combust. Flame **154** (2008), 478–488.

[17] Sethian, J. A. and Smereka, P.: Level set methods for fluid interfaces. Ann. Rev. Fluid Mech. **35** (2003), 341–372.

Conference Applications of Mathematics 2015, in honor of the birthday anniversaries of Ivo Babuška (90), Milan Práger (85), and Emil Vitásek (85)
J. Brandts, S. Korotov, M. Křížek, K. Segeth, J. Šístek, T. Vejchodský (Eds.)
Institute of Mathematics CAS, Prague 2015

FACTORIZATION MAKES FAST WALSH, PONS AND OTHER HADAMARD-LIKE TRANSFORMS EASY

Jaroslav Kautsky

Flinders University
CSEM, GPO Box 2100, Adelaide 5001, Australia
and
Prometheus, Inc., Newport, RI USA
jardakau@internode.on.net

Abstract: A simple device, based on the factorization of invertible matrix polynomials, enabling to identify the possibility of fast implementation of linear transforms is presented. Its applicability is demonstrated in the case of Hadamard matrices and their generalization, Hadamard matrix polynomials.

Keywords: Hadamard matrices, matrix polynomials, fast implementation, in-place implementation

MSC: 11C08, 15A23, 15B34, 65Y20

1. Introduction

Performing a linear transformation $\boldsymbol{y} = A\boldsymbol{x}$ with an $n \times n$ matrix A requires n^2 operations. We talk about a *fast* implementation of such a transformation if we can lower the number of operations, such as by using the Fast Fourier Transform (FFT) [3], which caused a revolution in signal processing by bringing the cost down to $n \log n$.

Hand-in-hand with the operation cost go memory requirements. In a straightforward implementation we need additional n memory locations. It may be desirable to perform the transformation *in place*, that is, the output $\boldsymbol{y}$ is stored directly into the locations occupied by the input $\boldsymbol{x}$, requiring possibly some small number, independent of n, of memory locations. Fast implementations usually allow such memory savings.

Not long after the FFT technique for reducing the cost similar results appeared for Walsh-Hadamard transforms [5, 12]; this development has continued to the present [13] and now includes Hadamard transforms other than those based on Walsh matrices [1]. Surprisingly, all of these results appear to be based on considerations following those in the development of the FFT.

In this paper we offer a different way to derive fast and in-place algorithms, not only for Hadamard matrices but also for their generalization, *Hadamard matrix polynomials* introduced in Section 3. Our approach is based on the factorization of invertible matrix polynomials discussed in Section 2 and allows not only the derivation of theoretical results but also, being based on a simple deterministic algorithm, the capacity to determine by computational means if a fast implementation exists in particular cases. That is presented in Section 4 and Section 5 for fast and in-place implementations, respectively. We conclude in Section 6 by factorizing a degree three 8×8 Hadamard matrix polynomial into five sparse factors.

The concept, properties and the construction of Hadamard matrix polynomials have been developed by the author and Radka Tezaur/Turcajova. They are documented in unpublished reports for Prometheus, Inc. (some related work can be found in `http://www.prometheus-us.com/PONS-papers/`) and have been used in signal processing applications such as [11]. A special case (size 2×2) has been presented in a Departmental seminar for which an abstract is available [10]. They are introduced here only for the purpose of demonstrating the applicability of the idea of finding fast implementations by factorization of matrix polynomials.

2. IMP — invertible matrix polynomials and their factorization

A matrix polynomial of size m and of order p (or degree $p-1$), $m > 1$, $p > 0$, is given by

$$A(z) = \sum_{k=1}^{p} A_k z^{k-1}$$

where A_k are real $m \times m$ matrices. The product of matrix polynomials is again a matrix polynomial which, unlike scalar polynomials ($m = 1$), may have degree less than the sum of the exact degrees of the multiplicands. This leads to the concepts of invertible matrix polynomials (IMPs) and, as a special case, orthogonal (or unitary) matrix polynomials.

We state these concepts formally as follows:

Definition 2.1 *1. A matrix polynomial $A(z)$ of order p is called invertible if there exists a polynomial $B(z)$ of order q such that $B(0) \neq 0$ and*

$$A(z)B(z) = \beta z^s I$$

(I is the identity matrix) for some scalar $\beta > 0$ and some s, $0 \leq s \leq p+q-2$.

2. An invertible matrix polynomial $A(z)$ of order p is called orthogonal if

$$B(z) = \sum_{k=1}^{p} A_{p-k+1}^T z^{k-1} = z^{p-1} A^T \left(\frac{1}{z}\right) , \tag{1}$$

$\beta = \|\boldsymbol{e}_1^T A(1)\|_2^2$ ($\boldsymbol{e}_1$ is the first column of the identity matrix) and $s = p-1$.

We have introduced the parameter β into the concept of inverse for later convenience and also include the parameter s for larger applicability in some situations.

Our results depend on the factorization of matrix polynomials. An orthogonal matrix polynomial of degree p can always be factorized into a product of exactly p linear factors [7, 9], a strong result a weaker form of which has not yet been established for IMPs. It has been shown in [6, p. 112], though, that matrix polynomials can not be generally factorized as demonstrated by the following example:

Example 2.2

$$A(z) = \begin{pmatrix} 0 & -1 \\ 0 & 0 \end{pmatrix} + z^2 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

is an IMP of size 2 *and degree* 2 *and its inverse is*

$$B(z) = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} + z^2 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

with $s = 4$ *and* $\beta = 1$.

What has actually been proved in [6] is that this IMP can not be expressed as a product of *two* IMPs, *both* of degree less than two (it is, in fact, factorizable into a product of three IMPs of degree one). The phrase "can not be factorized" here has a very restricted meaning — the product of factors of lower degrees and not more of them than the current degree. The word "factorization" in this paper (and in its title) is taken to mean simply "expressing as a product of factors".

Due to their invertibility IMPs can be factorized in many ways; we can actually prescribe all the partial products in the factorization.

Theorem 2.3 *Let* $B_j(z)$, $j = 0, 1, \ldots, n$, *be invertible matrix polynomials of the same size. Then there exist IMPs* $F_j(z)$, $j = 0, 1, \ldots, n$, *such that*

$$B_j(z) = F_j(z)F_{j-1}(z)\ldots F_0(z)\ , \qquad j = 0, 1, \ldots, n\ .$$

Proof. Take $F_0(z) = B_0(z)$ and $F_j(z) = B_j(z)B_{j-1}^{-1}(z)$, $j = 1, 2, \ldots, n$. □

The degrees of the factors depend, of course, on the chosen partial products. Regardless of how trivial this observation is, it is in fact a basis for what follows because some factorings may have certain desirable properties that others do not, such as particularly sparseness. We can demonstrate an application of this idea—using factorization into *sparse* factors to reduce complexity—in the case of Walsh-Hadamard matrices, which are given by the recurrence

$$W_0 = 1\ , \qquad W_{k+1} = \begin{pmatrix} W_k & W_k \\ W_k & -W_k \end{pmatrix}\ , \qquad k = 0, 1, \ldots \quad .$$

The steps in this construction are the source of partial products; the only difficulty in applying Theorem 2.3 is the "same size" requirement because W_{k+1} is a square

matrix of size 2^{k+1}, double that of W_k. That is overcome by doubling the size of W_k; if we choose

$$B_0 = \begin{pmatrix} W_k & 0 \\ 0 & W_k \end{pmatrix} \quad \text{and} \quad B_1 = W_{k+1}$$

we get

$$B_1 = FB_0 \quad \text{where} \quad F = \begin{pmatrix} W_k & W_k \\ W_k & -W_k \end{pmatrix} \begin{pmatrix} W_k^{-1} & 0 \\ 0 & W_k^{-1} \end{pmatrix} = \begin{pmatrix} I & I \\ I & -I \end{pmatrix}$$

leading immediately to the fast application of W_n costing $N \log_2 N$ operations, $N = 2^n$. A suitable same permutation of both rows and columns changes this factor F into a block diagonal matrix with 2×2 blocks W_1 in the diagonal, leading to an in-place implementation needing only one temporary memory location. Alternatively, this structure can be exploited for parallel processing.

We conclude this section by a comment on the apparently obvious statement that if $A(z) = F(z)B(z)$ then the application of $A(z)$ is equivalent to the application of $B(z)$ followed by that of $F(z)$. While this is straightforward for order $p = 1$ (multiplication of matrices and vectors), for $p > 1$ a rigorous definition of the meaning of "application" involves a lengthy exposition. The difference between applying matrix polynomials of order $p = 1$ and orders $p > 1$ is similar to moving from discrete Fourier transform to discrete wavelet transforms (originally called *lapped* transforms because of the need to use input data from the adjacent blocks). One approach is to express this application using block Toeplitz matrices (made circulant for invertibility) as in [8].

However, for the purpose of this paper it is sufficient to state that by application of an $N \times N$ matrix polynomial $A(z)$ of order p to a vector $\boldsymbol{x} = (\, \boldsymbol{x}_1^T \quad \boldsymbol{x}_2^T \quad \dots \quad \boldsymbol{x}_{K+p}^T \,)^T$ with $N \times 1$ components $\boldsymbol{x}_j$ we mean the evaluation of output

$$\boldsymbol{y}_j = \sum_{k=1}^{p} A_k \boldsymbol{x}_{j+k-1} \,, \quad j = 1, 2, \dots, K \,, \tag{2}$$

from which the statement about application of factorized matrix polynomials is easily justified.

It is important to note that when $\boldsymbol{y}_j$, for some j, has been evaluated in (2) then the vector $\boldsymbol{x}_j$ will not be needed to evaluate $\boldsymbol{y}_k$ for $k > j$.

3. HMP — Hadamard matrix polynomials and some constructions

Hadamard matrices are, up to a scalar multiple, orthogonal matrices the elements of which are restricted to have values 1 and -1. We extend this notion to matrix polynomials as follows:

Definition 3.1 *An orthogonal matrix polynomial* $A(z) = \sum_{j=1}^{p} A_j z^{j-1}$ *with elements* 1 *and* -1 *in all matrices* A_j *will be called a Hadamard matrix polynomial (HMP).*

We have introduced the parameter β into the definition of inverse so that we can say that the inverse of an HMP is again an HMP. For an HMP of size m and order p, $\beta = mp$.

Example 3.2

$$A(z) = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} + z \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$$

is an HMP of size 2 *and order* 2 *and its inverse is*

$$B(z) = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} + z \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$$

with $s = 1$ *and* $\beta = 4$.

In principle, construction of HMPs can be done by exhaustive search. However, this approach quickly becomes infeasible as the dimension of the problem grows.

As for Walsh-Hadamard matrices, there are constructions which allow doubling either the size or the order of an existing HMP. Starting with the simplest HMP (size 2, order 1)

$$H_0(z) = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$$

we can thus construct HMPs for which both size and order are powers of 2.

In this paper we will consider only three types of constructions: two which double the size of an HMP and one which doubles the order. Let us define them formally.

Denote by E the per-identity, that is the matrix such that XE reverses the order of columns of matrix X while EX does the same for rows. Also, let D_{S} be a diagonal matrix with $(\,1 \quad -1 \quad 1 \quad -1 \quad \dots\,)$ in the diagonal, the matrix post-multiplication by which will change the sign of every second column.

Definition 3.3 *Let* $A(z) = \sum_{k=1}^{p} A_k z^{k-1}$ *be an HMP of even size.*

1. *We define the Walsh-Sylvester size extension* $S_W(A, z)$ *by*

$$S_W(A, z) = \begin{pmatrix} A(z) & A(z) \\ A(z) & -A(z) \end{pmatrix} .$$

2. *We define the PONS-like size extension* $S_P(A, z)$ *by*

$$S_P(A, z) = \begin{pmatrix} A(z) & D_S E A(z) \\ D_S E A(z) & A(z) \end{pmatrix} .$$

3. Block the matrices A_k into equal parts

$$A_k = \begin{pmatrix} U_k^T \\ V_k^T \end{pmatrix} .$$

We define the order extension $O_p(A, z)$ by

$$O_p(A, z) = \sum_{k=1}^{p} \begin{pmatrix} U_k^T \\ U_k^T \end{pmatrix} z^{2k-2} + \begin{pmatrix} V_k^T \\ -V_k^T \end{pmatrix} z^{2k-1} .$$

Proposition 3.4 *The extended matrix polynomials are again HMPs.*

Proof. Straightforward using the second characterization of the orthogonal matrix polynomial in (1). □

Note that the HMP in Example 3.2 is obtained by $O_p(H_0, z)$.

There are other constructions which double either the size or the order of an HMP and also some which preserve both size and order: these include transposing the coefficient matrices, changing sign or permuting rows and columns or reversing the polynomials. It is thus possible to create a large variety of HMPs of increasing size and order. The type S_P is one of similar constructions resulting in so called PONS Hadamard matrices with many properties which are desirable in signal processing [2].

4. HMP — Fast implementation

The constructions of HMPs introduced in the previous section are the basis for applying Theorem 2.3, as was already demonstrated for Walsh-Hadamard matrices in Section 2 (order $p = 1$). We now show the extension to a degree one HMP (order $p = 2$).

Proposition 4.1 *Let W_n be a Walsh-Hadamard matrix of size $N = 2^n$. Then*

$$O_p(W_n, z) = F(z)W_n \quad \textit{where} \quad F(z) = \begin{pmatrix} I & 0 \\ I & 0 \end{pmatrix} + z \begin{pmatrix} 0 & I \\ 0 & -I \end{pmatrix} ,$$

here the identity matrices (as well as the zero blocks) are of size $N/2$. The factor's first coefficient $F_1 = \begin{pmatrix} I & 0 \\ I & 0 \end{pmatrix}$ is permutable to a block diagonal matrix with blocks $\begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix}$ of size 2.

Proof. Denote, for brevity, $W = W_{n-1}$. Calculate

$$\left(\begin{pmatrix} I & 0 \\ I & 0 \end{pmatrix} + z \begin{pmatrix} 0 & I \\ 0 & -I \end{pmatrix} \right) \begin{pmatrix} W & W \\ W & -W \end{pmatrix} = \begin{pmatrix} W & W \\ W & W \end{pmatrix} + z \begin{pmatrix} W & -W \\ -W & W \end{pmatrix} = O_p(W_n, z) .$$

The permutation $(\,1 \quad N/2+1 \quad 2 \quad N/2+2 \quad \dots \quad N/2 \quad N\,)$ applied to both rows and columns of F_1 achieves the required diagonalization. □

Further doubling of the order leads to factors of the same structure but of correspondingly higher degrees, with only the first and last coefficient matrices non-zero. The computational complexity is two per item of the output (i. e., two nonzero elements in each row of the linear transformation) in each step.

A more complicated factor is obtained when we reverse the order of extensions (results of this kind are best obtained by computer software and can be checked by calculations similar to those proving the Proposition 4.1).

Proposition 4.2 *If we apply the Walsh extension S_W to the linear HMP $O_p(W_n, z))$ (using the notation of Proposition 4.1) the resulting factor is*

$$F(z) = \begin{pmatrix} I & I & 2I & 0 \\ I & I & 0 & 2I \\ I & I & 0 & -2I \\ I & I & -2I & 0 \end{pmatrix} + z \begin{pmatrix} I & -I & 0 & 0 \\ -I & I & 0 & 0 \\ -I & I & 0 & 0 \\ I & -I & 0 & 0 \end{pmatrix} .$$

The minimal diagonal blocks after permutations have size 4.

The complexity is now much worse—five operations per item (six, actually, if we count only additions and implement the multiplier 2 as two additions). Interestingly, further extensions of size have similarly shaped factors while doubling the order leads to factors with complexity two per item as in Proposition 4.1.

Similar observations can be pursued for HMPs based on the S_P size extensions. These lead to transforms complementary to Walsh-Hadamard transforms with important applications in signal processing [1].

5. HMP — in-place implementation

It is obvious that if we apply a linear transform with an upper triangular matrix, then when we have calculated the first k elements of the result the first k elements of the input will not be needed any more and can be replaced by the output. That observation is the basis of the "in-place" implementation. Similarly, in the application of a matrix polynomial, in (2), we note that, once $\boldsymbol{y}_j$ is evaluated, $\boldsymbol{x}_j$ will not be needed to evaluate $\boldsymbol{y}_k$ for $k > j$.

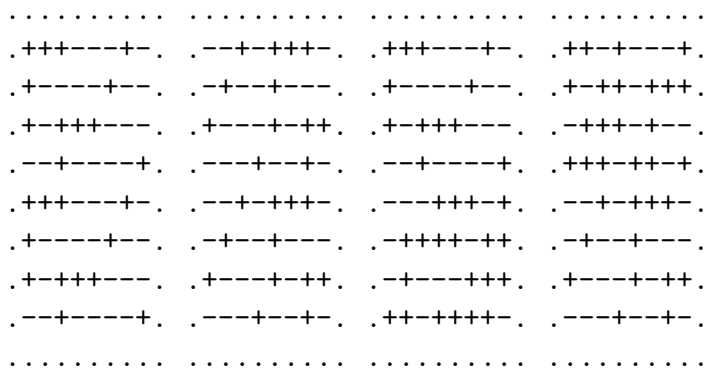
So memory requirements depend upon how we can implement the evaluation of $A_1\boldsymbol{x}_j$, that is $F_1\boldsymbol{x}_j$, as we are now discussing the application of a factor. As already pointed out in particular cases, we need to be able to permute rows and columns of F_1 into a block upper triangular form; we can then proceed as suggested above and the maximal size of the diagonal blocks gives the number of additional memory locations needed to calculate the transform in place.

It is important to realize that the rows and columns must be permuted in the same way, otherwise we would be storing the results corresponding to the diagonal blocks in the wrong places and overwriting what is still needed in using the next block.

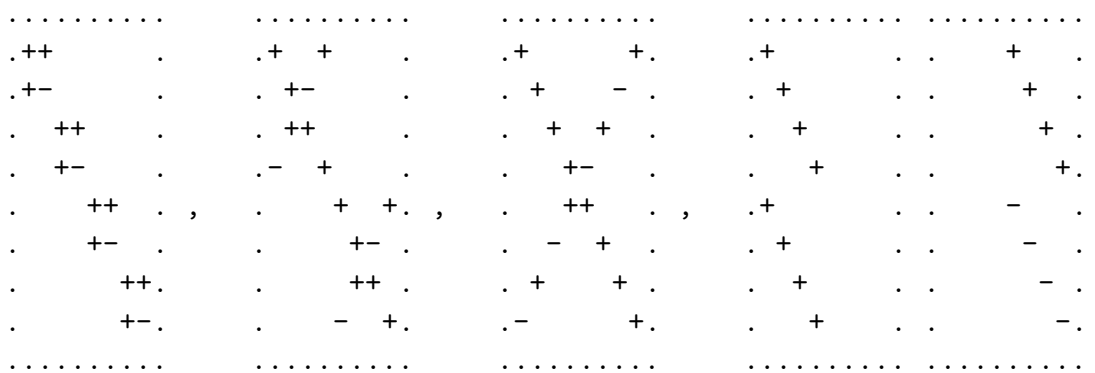
The problem of finding the block triangular form by a symmetric permutation is equivalent to that of determining the strongly connected components of a graph. It can be solved, for example, by Tarjan's algorithm [4].

6. An example

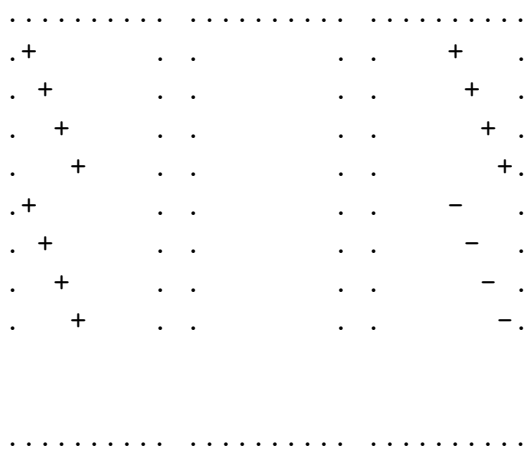
An 8×8 HMP of order 4 (that is, degree 3) constructed by two size extensions S_P followed by two order doubling O_p is visualized as:



where + and - denote 1 and -1, respectively, and we have surrounded the four coefficient matrices $A_1, \ldots, A_4$ by dots. Note that unlike in the Walsh extensions the pattern of ± 1 looks almost random in this PONS-like HMP. Nevertheless, the fast/in-place implementation is still achievable by the following five factors



and



Notice that application of each row of the original HMP involving 31 additions and subtractions is replaced by just 5 such operations. This was achieved by a factorization of a degree 3 matrix polynomial into the product of five matrix polynomials of degrees 0, 0, 0, 1 and 2, respectively.

One other observation important for implementation is the regular pattern in the structure of the factors which can be realized directly by the computer program and thus avoiding the need to store the original HMP or its factors.

7. Conclusion

In this paper we show that there is a large class of Hadamard matrices and Hadamard matrix polynomials which are constructed in such a way that using the new result of Theorem 2.3 can easily determine whether or not transforms can be implemented fast and in-place. Our approach suggests a simple algorithm which determines the polynomial matrix factors from which it is possible to see, in each case, how good the fast and in-place implementation will be. Theorem 2.3 resolves an important need because the savings available from fast and in-place transforms differ significantly from one case to another, as shown by two examples in Section 3. Indeed, the question what savings can be achieved by fast implementation is thus far from trivial. The possible savings result from two properties: the sparseness of the factors and the small integer sizes of their non-zero elements. An alternative factorization for orthogonal matrix polynomials mentioned in Section 2 ([9]) is not applicable here because the factors would neither be sparse nor have integer values.

Acknowledgements

The author thanks the anonymous referee for useful comments improving the presentation of the material.

This article is dedicated to my theses advisers and colleagues at the occasions of their birthdays. I have worked with them in the years 1956-1968 so it is not surprising that the topic has probably drifted far away from their interests. Nevertheless, I am grateful for their guidance and influence.

References

[1] Byrnes, J., Gertner, I., Ostheimer, G., and Ramalho, M.: Discrete one dimensional signal processing apparatus and method using energy spreading coding. U.S. patent number 5,913,186 (1999).

[2] Byrnes, J.S., Saffari, B., and Shapiro, H.: Energy spreading and data compression using the Prometheus orthonormal set. Proceedings of the IEEE Digital Signal Processing Workshop, Norway (1996), 9–12.

[3] Cooley, J. and Tukey, J.: An algorithm for the machine calculation of complex Fourier series. Math. Comp. **19** (1965), 297–301.

[4] Duff, I. and Reid, J.: An implementation of Tarjan's algorithm for the block triangularization of a matrix. ACM Transactions on Mathematical Software **4** (1978), 137–147.

[5] Fino, B. and Algazi, V.: Unified matrix treatment of the fast Walsh-Hadamard transform. IEEE Trans. on Computers **C-25** (1976), 1142–1146.

[6] Gohberg, L., Lancaster, P., and Rodman, L.: *Matrix polynomials.* Academic Press, New York, 1982.

[7] Kautsky, J. and Turcajová, R.: A matrix approach to discrete wavelets. In: C.K. Chui, L. Montefusco, and L. Puccio (Eds.), *Wavelets: Theory, Algorithms, and Applications*, *Wavelet Analysis and Its Applications*, vol. 5, pp. 117–136. Academic Press, 1994.

[8] Kautsky, J. and Turcajová, R.: Discrete biorthogonal wavelet transforms as block circulant matrices. Linear Algebra Appl. **223/224** (1995), 393–413.

[9] Kautsky, J. and Turcajová, R.: Pollen product factorization and construction of higher multiplicity wavelets. Linear Algebra Appl. **222** (1995), 241–260.

[10] Turcajova, R.: Construction of Hadamard Matrix Polynomials. URL: `http://www2.cs.cas.cz/semincm/abstracts/2004-06-17.Turcajova.html` (2004).

[11] Ultrafast efficient data compression. URL: `http://www.prometheus-us.com/Projects/UltrafastEfficientDataCompression.html` (1999).

[12] Vlasenko, V. and Rao, K.: Unified matrix treatment of discrete transforms. IEEE Trans. on Computers **C-28** (1979), 934–938.

[13] Yusuf, Z., Abbasi, S., and Alamoud, A.: A novel complete set of Walsh and inverse Walsh transforms for signal processing. *International Conference on Communication Systems and Network Technologies (CSNT)*, 2011, 504–509.

Conference Applications of Mathematics 2015, in honor of the birthday anniversaries of Ivo Babuška (90), Milan Práger (85), and Emil Vitásek (85)
J. Brandts, S. Korotov, M. Křížek, K. Segeth, J. Šístek, T. Vejchodský (Eds.)
Institute of Mathematics AS CR, Prague 2015

ON THE INTERPOLATION CONSTANTS OVER TRIANGULAR ELEMENTS

Kenta Kobayashi

Graduate School of Commerce and Management, Hitotsubashi University
Naka 2-1, Kunitachi, Tokyo 186-8601, Japan
kenta.k@r.hit-u.ac.jp

Abstract: We propose a simple method to obtain sharp upper bounds for the interpolation error constants over the given triangular elements. These constants are important for analysis of interpolation error and especially for the error analysis in the Finite Element Method. In our method, interpolation constants are bounded by the product of the solution of corresponding finite dimensional eigenvalue problems and constant which is slightly larger than one. Guaranteed upper bounds for these constants are obtained via the numerical verification method. Furthermore, we introduce remarkable formulas for the upper bounds of these constants.

Keywords: interpolation error constant, numerical verification method, Finite Element Method

MSC: 65D05, 65N15, 65D30

1. Introduction

The analysis of interpolation error is important in a lot of applications such as the approximate theory and the error estimation for the solution of Finite Element Method. In order to estimate the interpolation errors, we have to obtain the upper bounds of the constants which appear in some kinds of norm inequalities. These are called interpolation error constants.

Let T be given triangle in $\mathbb{R}^2$ and define function spaces $V^{1,1}(T), V^{1,2}(T), V^2(T)$ as follows:

$$V^{1,1}(T) = \left\{ \varphi \in H^1(T) \;\middle|\; \int_T \varphi \, dxdy = 0 \right\},$$
$$V^{1,2}(T) = \left\{ \varphi \in H^1(T) \;\middle|\; \int_{\gamma_k} \varphi \, ds = 0, \quad \forall k = 1,2,3 \right\},$$
$$V^2(T) = \left\{ \varphi \in H^2(T) \;\middle|\; \varphi(p_k) = 0, \quad \forall k = 1,2,3 \right\},$$

where p_1, p_2, p_3 and $\gamma_1, \gamma_2, \gamma_3$ are vertices and edges of T, respectively. Under these settings, it is known that the following interpolation error constants $C_1(T)$, $C_2(T)$, $C_3(T)$ and $C_4(T)$ exist:

$$C_1(T) = \sup_{u \in V^{1,1}(T)\backslash 0} \frac{\|u\|_{L^2(T)}}{\|\nabla u\|_{L^2(T)}}, \qquad C_2(T) = \sup_{u \in V^{1,2}(T)\backslash 0} \frac{\|u\|_{L^2(T)}}{\|\nabla u\|_{L^2(T)}},$$
$$C_3(T) = \sup_{u \in V^{2}(T)\backslash 0} \frac{\|u\|_{L^2(T)}}{|u|_{H^2(T)}}, \qquad C_4(T) = \sup_{u \in V^{2}(T)\backslash 0} \frac{\|\nabla u\|_{L^2(T)}}{|u|_{H^2(T)}}.$$

where $|\cdot|_{H^k(\Omega)}$ means H^k semi-norm defined later.

In this paper, we present a simple method to obtain explicit and sharp upper bounds for them. Furthermore, we obtained the following remarkable formulas for the upper bounds:

$$C_1(T) < K_1(T) = \sqrt{\frac{A^2+B^2+C^2}{28} - \frac{S^4}{A^2B^2C^2}},$$
$$C_2(T) < K_2(T) = \sqrt{\frac{A^2+B^2+C^2}{54} - \frac{S^4}{2A^2B^2C^2}},$$
$$C_3(T) < K_3(T) = \sqrt{\frac{A^2B^2+B^2C^2+C^2A^2}{83} - \frac{1}{24}\left(\frac{A^2B^2C^2}{A^2+B^2+C^2} + S^2\right)},$$
$$C_4(T) < K_4(T) = \sqrt{\frac{A^2B^2C^2}{16S^2} - \frac{A^2+B^2+C^2}{30} - \frac{S^2}{5}\left(\frac{1}{A^2}+\frac{1}{B^2}+\frac{1}{C^2}\right)},$$

where A, B, C are the edge lengths of triangle T and S is the area of T.

As we will show in Section 5, the upper bounds obtained by these formulas are sharp enough for the practical applications. Moreover, $K_1(T) \ldots K_4(T)$ are convenient for practical calculations since these formulas consists of just four arithmetic operations and the square root.

We have to note that, by our method, we can only prove these formulas for the "given" triangles. To prove the formulas for "any" triangle, we need some continuation techniques and the asymptotic analysis. More specifically, we first prove these formulas for finitely many specific triangles by slightly strict form, namely

$$C_j(T) < (1-\varepsilon)K_j(T)$$

for some small $\varepsilon > 0$ and then extend these results to general cases by the analytical evaluation and the asymptotic analysis. We indeed succeeded to prove it but we will show it in another paper because of the space limit.

2. Preceding works

In connection with the Finite Element Method, there is a plenty of works especially on the relation between $C_4(T)$ and the error estimates such as [4, 6, 3, 9, 10, 12, 19, 14, 24] for *a priori* error estimate and [4, 8, 14] for *a posteriori* error estimate.

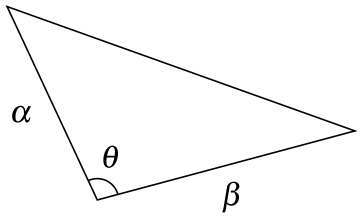


Figure 1: α, β and θ for triangle T.

On the explicit upper bound for $C_4(T)$, Arcangeli and Gout[2] obtained the following estimates:

$$C_4(T) \leq \frac{3d(T)^2}{\rho(T)}$$

where $d(T)$ is a diameter of T and $\rho(T)$ is a diameter of the inscribed circle of T. They also obtained the upper bound for $C_3(T)$ as follows:

$$C_3(T) \leq 3d(T)^2.$$

Meinguet and Descloux[17] improved their result and obtained

$$C_4(T) \leq \frac{1.21d(T)^2}{\rho(T)}.$$

Natterer [20] showed that $C_4(T)$ is bounded in terms of $C_4(T_{0,1})$ where $T_{0,1}$ is a isosceles right triangle whose edge lengths are $1, 1$ and $\sqrt{2}$. Specifically, they showed

$$C_4(T) \leq C_4(T_{0,1}) \cdot \frac{\alpha^2 + \beta^2 + \sqrt{\alpha^4 + 2\alpha^2\beta^2 \cos 2\theta + \beta^4}}{\sqrt{2(\alpha^2 + \beta^2 - \sqrt{\alpha^4 + 2\alpha^2\beta^2 \cos 2\theta + \beta^4})}}, \tag{1}$$

where α and β are the longest and second longest edge lengths and θ is an included angle (Fig. 1). In the same paper, they proved $C_4(T_{0,1}) \leq 0.81$. Nakao and Yamamoto [19] proved that

$$C_4(T_{0,1}) \leq 0.4939$$

by numerical verification method. Kikuchi and Liu [7] proved that $C_4(T_{0,1})$ is bounded by the maximum positive solution of transcendental equation for μ:

$$\frac{1}{\mu} + \tan\frac{1}{\mu} = 0$$

and showed

$$C_4(T_{0,1}) \leq 0.49293.$$

Moreover, Liu and Kikuchi [14] proved that

$$C_4(T) \leq C_4(T_{0,1}) \cdot \frac{1 + \cos\theta}{\sin\theta} \sqrt{\frac{\alpha^2 + \beta^2 + \sqrt{\alpha^4 + 2\alpha^2\beta^2 \cos 2\theta + \beta^4}}{2}}. \tag{2}$$

Note that the estimation (2) is consistent with the maximum angle condition [3] whereas the estimation (1) is not. In fact, if we fix β and θ and let $\alpha \to 0$, the right-hand side of (1) diverges to infinity whereas the right-hand side of (2) remains bounded.

$C_1(T)$ is known as the Poincaré-Friedrichs constant and Payne and Weinberger obtained

$$C_1(T) \leq \frac{d(T)}{\pi}.$$

This estimation is valid for any convex domain. For arbitrary triangle T, Laugesen and Siudeja [11] obtained

$$C_1(T) \leq \frac{d(T)}{j_{1,1}} \tag{3}$$

where $j_{1,1} = 3.83170597\ldots$ denotes the first positive root of the Bessel function J_1.

On the other hand, Kikuchi and Liu [7] proved that

$$C_1(T_{0,1}) = \frac{1}{\pi}$$

and

$$C_1(T) \leq C_1(T_{0,1})\sqrt{1 + |\cos\theta|}\ \max(\alpha, \beta). \tag{4}$$

There are only a few results for $C_2(T)$ itself. However, $C_2(T)$ is bounded by so called Babuška-Aziz constant whose existence is proved by Babuška and Aziz [3, Lemma 2.1]. From the upper bound for the Babuška-Aziz constant obtained by Liu and Kikuchi [14], we have

$$C_2(T) \leq 0.34856\sqrt{1 + |\cos\theta|}\ \max(\alpha, \beta).$$

For the most triangles, our formulas $K_1(T)\ldots K_4(T)$ give better upper bounds than the preceding results. The exception is that (3) or (4) provides slightly lower value than $K_1(T)$ for some triangles.

There are some results about computing lower bounds of eigenvalues of elliptic operators such as [1, 5, 13, 15, 16, 21, 23] which can be applied to compute upper bounds of $C_1(T)$ or $C_2(T)$. Compared to these results, our method is only applicable to the triangular domain but has the advantage that the sharp upper bounds can be obtained by a simple implementation.

3. Definitions and preliminaries

For given triangle T, let $p_1(T), p_2(T), p_3(T)$ be vertices of T and $\gamma_1(T), \gamma_2(T), \gamma_3(T)$ be edges $p_2(T)p_3(T)$, $p_3(T)p_1(T)$, $p_1(T)p_2(T)$, respectively. Let $n(T)$ be the outer normal unit vector on ∂T, $\nu(T)$ be the direction vector which takes counterclockwise direction through ∂T and $ds(T)$ be the line element on ∂T. We omit "(T)" if there is no possibility of confusion. We use Cartesian coordinates (x, y) and use the usual notation for L^2 norm and define H^k semi-norm $|\cdot|_{H^k(T)}$ by $|u|^2_{H^k(\Omega)} =$

$\sum_{j=0}^{k} \binom{k}{j} \left\| \frac{\partial^k u}{\partial x^j \partial y^{k-j}} \right\|^2_{L^2(\Omega)}$. $T_{a,b}$ denotes triangle whose vertices are $(0,0)$, $(1,0)$ and (a,b). We use subscripts to indicate partial derivatives.

Let Q_α and Q_β denote the following polynomial spaces:

$$Q_\alpha = \Big\{ a_1(x^2+y^2) + a_2 x + a_3 y + a_4 \;\Big|\; a_1, \ldots, a_4 \in \mathbb{R} \Big\},$$
$$Q_\beta = \Big\{ a_1 x^2 + a_2 xy + a_3 y^2 + a_4 x + a_5 y + a_6 \;\Big|\; a_1, \ldots, a_6 \in \mathbb{R} \Big\}.$$

Note that both Q_α and Q_β are invariant under constant shifts and rotations and thus they are independent of the choice of the coordinates. Let τ be the given triangle and we define two kinds of second order interpolation $\Pi_\tau^{(\alpha)}\varphi$ for $\varphi \in H^1(\tau)$ and $\Pi_\tau^{(\beta)}\varphi$ for $\varphi \in H^2(\tau)$ on triangle τ as follows:

$$\begin{cases} \Pi_\tau^{(\alpha)}\varphi \in Q_\alpha \\ \displaystyle\int_{\gamma_k} \Pi_\tau^{(\alpha)}\varphi \, ds = \int_{\gamma_k} \varphi \, ds, \qquad k = 1,2,3, \\ \displaystyle\iint_\tau \Pi_\tau^{(\alpha)}\varphi \, dxdy = \iint_\tau \varphi \, dxdy, \end{cases}$$

$$\begin{cases} \Pi_\tau^{(\beta)}\varphi \in Q_\beta \\ \Pi_\tau^{(\beta)}\varphi(p_k) = \varphi(p_k), \qquad k = 1,2,3, \\ \displaystyle\int_{\gamma_k} \nabla \Pi_\tau^{(\beta)}\varphi \cdot n \, ds = \int_{\gamma_k} \nabla\varphi \cdot n \, ds, \qquad k = 1,2,3. \end{cases}$$

In the rest of this section, we prepare some preliminary lemmas.

Lemma 1. *If $\varphi \in V^2(\tau)$ satisfies*

$$\int_{\gamma_k} \nabla\varphi \cdot n \, ds = 0, \quad k = 1,2,3,$$

then

$$\varphi_x, \ \varphi_y \in V^{1,2}(\tau)$$

holds.

Proof. From $\varphi(p_1) = \varphi(p_2) = \varphi(p_3) = 0$, we have

$$\int_{\gamma_k} \nabla\varphi \cdot \nu \, ds = 0, \qquad k = 1,2,3.$$

Then, together with the assumption,

$$\int_{\gamma_k} \nabla\varphi \cdot w \, ds = 0, \qquad k = 1,2,3,$$

holds for any fixed vector w, which proves the lemma. □

On the interpolations $\Pi_\tau^{(\alpha)}$ and $\Pi_\tau^{(\beta)}$, the following orthogonal properties hold:

Lemma 2. *For $\varphi \in H^1(\tau)$,*

$$\|\nabla \Pi_\tau^{(\alpha)} \varphi\|_{L^2(\tau)}^2 + \|\nabla(\varphi - \Pi_\tau^{(\alpha)} \varphi)\|_{L^2(\tau)}^2 = \|\nabla \varphi\|_{L^2(\tau)}^2.$$

Lemma 3. *For $\varphi \in H^2(\tau)$,*

$$|\Pi_\tau^{(\beta)} \varphi|_{H^2(\tau)}^2 + |\varphi - \Pi_\tau^{(\beta)} \varphi|_{H^2(\tau)}^2 = |\varphi|_{H^2(\tau)}^2.$$

Proof of Lemma 2. Since $\Pi_\tau^{(\alpha)} \varphi$ does not depend on the choice of the coordinates, we consider the x-axis to be aligned with the edge γ_3 and take $p_1 = (0,0)$, $p_2 = (h,0)$, $p_3 = (ah, bh)$ and

$$\Pi_\tau^{(\alpha)} \varphi = a_1(x^2 + y^2) + a_2 x + a_3 y + a_4.$$

Then, the divergence theorem yields

$$\begin{aligned}
&\|\nabla \varphi\|_{L^2(\tau)}^2 - \|\nabla \Pi_\tau^{(\alpha)} \varphi\|_{L^2(\tau)}^2 - \|\nabla(\varphi - \Pi_\tau^{(\alpha)} \varphi)\|_{L^2(\tau)}^2 \\
&\quad = 2 \iint_\tau \nabla(\varphi - \Pi_\tau^{(\alpha)} \varphi) \cdot \nabla \Pi_\tau^{(\alpha)} \varphi \, dxdy \\
&\quad = 2 \iint_\tau \operatorname{div}\Big((\varphi - \Pi_\tau^{(\alpha)} \varphi) \nabla \Pi_\tau^{(\alpha)} \varphi \Big) dxdy - 2 \iint_\tau (\varphi - \Pi_\tau^{(\alpha)} \varphi) \Delta \Pi_\tau^{(\alpha)} \varphi \, dxdy \\
&\quad = 2 \oint_{\partial\tau} (\varphi - \Pi_\tau^{(\alpha)} \varphi) \nabla \Pi_\tau^{(\alpha)} \varphi \cdot n \, ds - 8a_1 \iint_\tau (\varphi - \Pi_\tau^{(\alpha)} \varphi) \, dxdy \\
&\quad = 2 \oint_{\partial\tau} (\varphi - \Pi_\tau^{(\alpha)} \varphi) \begin{pmatrix} 2a_1 x + a_2 \\ 2a_1 y + a_3 \end{pmatrix} \cdot n \, ds \\
&\quad = 4a_1 \oint_{\partial\tau} (\varphi - \Pi_\tau^{(\alpha)} \varphi) \begin{pmatrix} x \\ y \end{pmatrix} \cdot n \, ds \\
&\quad = 4a_1 \int_{\gamma_1} (\varphi - \Pi_\tau^{(\alpha)} \varphi) \begin{pmatrix} x - h \\ y \end{pmatrix} \cdot n \, ds + 4a_1 \int_{\gamma_2} (\varphi - \Pi_\tau^{(\alpha)} \varphi) \begin{pmatrix} x - ah \\ y - bh \end{pmatrix} \cdot n \, ds \\
&\qquad\qquad + 4a_1 \int_{\gamma_3} (\varphi - \Pi_\tau^{(\alpha)} \varphi) \begin{pmatrix} x \\ y \end{pmatrix} \cdot n \, ds \\
&\quad = 4a_1 \int_{\gamma_1} \sqrt{(x-h)^2 + y^2} \, (\varphi - \Pi_\tau^{(\alpha)} \varphi) \, \nu \cdot n \, ds \\
&\qquad + 4a_1 \int_{\gamma_2} \sqrt{(x - ah)^2 + (y - bh)^2} \, (\varphi - \Pi_\tau^{(\alpha)} \varphi) \, \nu \cdot n \, ds \\
&\qquad\qquad + 4a_1 \int_{\gamma_3} \sqrt{x^2 + y^2} \, (\varphi - \Pi_\tau^{(\alpha)} \varphi) \, \nu \cdot n \, ds = 0
\end{aligned}$$

□

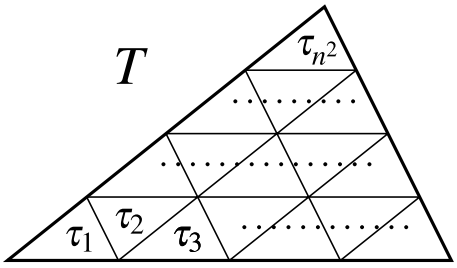


Figure 2: Divide T into n^2 congruent small triangles.

Proof of Lemma 3. Same as previous lemma, we take $p_1 = (0,0),\ p_2 = (h,0),\ p_3 = (ah, bh)$ and

$$\Pi_\tau^{(\beta)}\varphi = a_1x^2 + a_2xy + a_3y^2 + a_4x + a_5y + a_6.$$

Then, the divergence theorem yields

$$\begin{aligned}
&|\varphi|^2_{H^2(\tau)} - |\Pi_\tau^{(\beta)}\varphi|^2_{H^2(\tau)} - |\varphi - \Pi_\tau^{(\beta)}\varphi|^2_{H^2(\tau)} \\
&\quad= 2\iint_\tau \Big((\varphi - \Pi_\tau^{(\beta)}\varphi)_{xx}(\Pi_\tau^{(\beta)}\varphi)_{xx} + 2(\varphi - \Pi_\tau^{(\beta)}\varphi)_{xy}(\Pi_\tau^{(\beta)}\varphi)_{xy} \\
&\qquad\qquad + (\varphi - \Pi_\tau^{(\beta)}\varphi)_{yy}(\Pi_\tau^{(\beta)}\varphi)_{yy} \Big)\, dxdy \\
&\quad= 2\iint_\tau \operatorname{div} \begin{pmatrix} \nabla(\varphi - \Pi_\tau^{(\beta)}\varphi)\cdot\nabla(\Pi_\tau^{(\beta)}\varphi)_x \\ \nabla(\varphi - \Pi_\tau^{(\beta)}\varphi)\cdot\nabla(\Pi_\tau^{(\beta)}\varphi)_y \end{pmatrix} dxdy \\
&\quad= 2\oint_{\partial\tau} \begin{pmatrix} \nabla(\varphi - \Pi_\tau^{(\beta)}\varphi)\cdot\nabla(\Pi_\tau^{(\beta)}\varphi)_x \\ \nabla(\varphi - \Pi_\tau^{(\beta)}\varphi)\cdot\nabla(\Pi_\tau^{(\beta)}\varphi)_y \end{pmatrix} \cdot n\, ds \\
&\quad= 2\oint_{\partial\tau} \nabla(\varphi - \Pi_\tau^{(\beta)}\varphi)\cdot\nabla(\nabla\Pi_\tau^{(\beta)}\varphi\cdot n)\, ds \\
&\quad= 2\oint_{\partial\tau} \nabla(\varphi - \Pi_\tau^{(\beta)}\varphi)\cdot \begin{pmatrix} 2a_1 & a_2 \\ a_2 & 2a_3 \end{pmatrix} n\, ds.
\end{aligned}$$

Here, Lemma 1 yields

$$\int_{\gamma_k} (\varphi - \Pi_\tau^{(\beta)}\varphi)_x\, ds = \int_{\gamma_k} (\varphi - \Pi_\tau^{(\beta)}\varphi)_y\, ds = 0, \qquad k = 1,2,3,$$

which leads us to the conclusion. □

4. Our method to bound the constants

We divide triangle T into n^2 congruent small triangles $\tau_1, \ldots, \tau_{n^2}$ (Fig. 2). We assume that each τ_k is open set, namely, does not contain its boundary, and define

$$T' = \bigcup_{k=1}^{n^2} \tau_k.$$

Then we define $\Pi^{(\alpha)}u$ for $u \in H^1(T)$ and $\Pi^{(\beta)}u$ for $u \in H^2(T)$ as follows:

$$\Pi^{(\alpha)}u|_{\tau_k} = \Pi^{(\alpha)}_{\tau_k}u, \qquad \Pi^{(\beta)}u|_{\tau_k} = \Pi^{(\beta)}_{\tau_k}u.$$

Note that $\Pi^{(\alpha)}u$ and $\Pi^{(\beta)}u$ are not always continuous on T.

By solving finite dimensional generalized eigenvalue problems, we can obtain following constants:

$$C_1^{(n)}(T) = \sup_{u\in V^{1,1}(T)\setminus 0} \frac{\|\Pi^{(\alpha)}u\|_{L^2(T')}}{\|\nabla\Pi^{(\alpha)}u\|_{L^2(T')}}, \qquad C_2^{(n)}(T) = \sup_{u\in V^{1,2}(T)\setminus 0} \frac{\|\Pi^{(\alpha)}u\|_{L^2(T')}}{\|\nabla\Pi^{(\alpha)}u\|_{L^2(T')}},$$
$$C_3^{(n)}(T) = \sup_{u\in V^{2}(T)\setminus 0} \frac{\|\Pi^{(\beta)}u\|_{L^2(T')}}{|\Pi^{(\beta)}u|_{H^2(T')}}, \qquad C_4^{(n)}(T) = \sup_{u\in V^{2}(T)\setminus 0} \frac{\|\nabla\Pi^{(\beta)}u\|_{L^2(T')}}{|\Pi^{(\beta)}u|_{H^2(T')}}.$$

With respect to these constants, we have the following theorem:

Theorem 1.

$$C_1(T) \le \sqrt{\frac{n^2}{n^2-1}}\, C_1^{(n)}(T), \qquad C_2(T) \le \sqrt{\frac{n^2}{n^2-1}}\, C_2^{(n)}(T),$$
$$C_3(T) \le \sqrt{\frac{n^4}{n^4-1}}\, C_3^{(n)}(T), \qquad C_4(T) \le \sqrt{\frac{n^2}{n^2-1}}\, C_4^{(n)}(T),$$
$$C_4(T) \le \sqrt{C_4^{(n)}(T)^2 + \frac{C_2(T)^2}{n^2}},$$

Proof. We first note that the scaling properties $C_j(\tau_k) = C_j(T)/n$ for $j = 1, 2, 4$ and $C_3(\tau_k) = C_3(T)/n^2$ hold. This property can be easily shown by change of variables.

From Lemma 2, for $u \in V^{1,j}(T)$, $j = 1, 2$, we have

$$\begin{aligned}
\|u\|_{L^2(T)} &\le \|\Pi^{(\alpha)}u\|_{L^2(T')} + \|u - \Pi^{(\alpha)}u\|_{L^2(T')} \\
&= \|\Pi^{(\alpha)}u\|_{L^2(T')} + \sqrt{\sum_{k=1}^{n^2} \|u - \Pi^{(\alpha)}_{\tau_k}u\|^2_{L^2(\tau_k)}} \\
&\le C_j^{(n)}(T)\, \|\nabla\Pi^{(\alpha)}u\|_{L^2(T')} + \frac{C_j(T)}{n}\sqrt{\sum_{k=1}^{n^2} \|\nabla(u - \Pi^{(\alpha)}_{\tau_k}u)\|^2_{L^2(\tau_k)}} \\
&\le \sqrt{C_j^{(n)}(T)^2 + \frac{C_j(T)^2}{n^2}}\sqrt{\sum_{k=1}^{n^2} \left(\|\nabla\Pi^{(\alpha)}_{\tau_k}u\|^2_{L^2(\tau_k)} + \|\nabla(u - \Pi^{(\alpha)}_{\tau_k}u)|^2_{L^2(\tau_k)}\right)} \\
&= \sqrt{C_j^{(n)}(T)^2 + \frac{C_j(T)^2}{n^2}}\sqrt{\sum_{k=1}^{n^2} \|\nabla u\|^2_{L^2(\tau_k)}} \\
&= \sqrt{C_j^{(n)}(T)^2 + \frac{C_j(T)^2}{n^2}}\, \|\nabla u\|_{L^2(T)}.
\end{aligned}$$

Furthermore, from Lemma 3, for $u \in V^2(T)$,

$$\begin{aligned}
\|u\|_{L^2(T)} &\le \|\Pi^{(\beta)}u\|_{L^2(T')} + \|u - \Pi^{(\beta)}u\|_{L^2(T')} \\
&= \|\Pi^{(\beta)}u\|_{L^2(T')} + \sqrt{\sum_{k=1}^{n^2} \|u - \Pi^{(\beta)}_{\tau_k} u\|^2_{L^2(\tau_k)}} \\
&\le C_3^{(n)}(T)\ |\Pi^{(\beta)}u|_{H^2(T')} + \frac{C_3(T)}{n^2}\sqrt{\sum_{k=1}^{n^2} |u - \Pi^{(\beta)}_{\tau_k} u|^2_{H^2(\tau_k)}} \\
&\le \sqrt{C_3^{(n)}(T)^2 + \frac{C_3(T)^2}{n^4}}\sqrt{\sum_{k=1}^{n^2} \left(|\Pi^{(\beta)}_{\tau_k} u|^2_{H^2(\tau_k)} + |u - \Pi^{(\beta)}_{\tau_k} u|^2_{H^2(\tau_k)}\right)} \\
&= \sqrt{C_3^{(n)}(T)^2 + \frac{C_3(T)^2}{n^4}}\sqrt{\sum_{k=1}^{n^2} |u|^2_{H^2(\tau_k)}} \\
&= \sqrt{C_3^{(n)}(T)^2 + \frac{C_3(T)^2}{n^4}}\ |u|_{H^2(T)}
\end{aligned}$$

and

$$\begin{aligned}
\|\nabla u\|_{L^2(T)} &\le \|\nabla\Pi^{(\beta)}u\|_{L^2(T')} + \|\nabla(u - \Pi^{(\beta)}u)\|_{L^2(T')} \\
&= \|\nabla\Pi^{(\beta)}u\|_{L^2(T')} + \sqrt{\sum_{k=1}^{n^2} \|\nabla(u - \Pi^{(\beta)}_{\tau_k} u)\|^2_{L^2(\tau_k)}} \\
&\le C_4^{(n)}(T)\ |\Pi^{(\beta)}u|_{H^2(T')} + \frac{C_4(T)}{n}\sqrt{\sum_{k=1}^{n^2} |u - \Pi^{(\beta)}_{\tau_k} u|^2_{H^2(\tau_k)}} \\
&\le \sqrt{C_4^{(n)}(T)^2 + \frac{C_4(T)^2}{n^2}}\sqrt{\sum_{k=1}^{n^2} \left(|\Pi^{(\beta)}_{\tau_k} u|^2_{H^2(\tau_k)} + |u - \Pi^{(\beta)}_{\tau_k} u|^2_{H^2(\tau_k)}\right)} \\
&= \sqrt{C_4^{(n)}(T)^2 + \frac{C_4(T)^2}{n^2}}\sqrt{\sum_{k=1}^{n^2} |u|^2_{H^2(\tau_k)}} \\
&= \sqrt{C_4^{(n)}(T)^2 + \frac{C_4(T)^2}{n^2}}\ |u|_{H^2(T)}
\end{aligned}$$

hold. Using Lemma 1, we can also evaluate $\|\nabla(u - \Pi^{(\beta)}u)\|_{L^2(T')}$ in the first line of the previous expression by

$$\|\nabla(u-\Pi^{(\beta)}u)\|_{L^2(T')} = \sqrt{\sum_{k=1}^{n^2}\left(\|(u-\Pi^{(\beta)}_{\tau_k}u)_x\|^2_{L^2(\tau_k)} + \|(u-\Pi^{(\beta)}_{\tau_k}u)_y\|^2_{L^2(\tau_k)}\right)}$$
$$\leq \frac{C_2(T)}{n}\sqrt{\sum_{k=1}^{n^2}\left(\|\nabla(u-\Pi^{(\beta)}_{\tau_k}u)_x\|^2_{L^2(\tau_k)} + \|\nabla(u-\Pi^{(\beta)}_{\tau_k}u)_y\|^2_{L^2(\tau_k)}\right)}$$
$$= \frac{C_2(T)}{n}\sqrt{\sum_{k=1}^{n^2}|u-\Pi^{(\beta)}_{\tau_k}u|^2_{H^2(\tau_k)}}.$$

From above evaluations, we have the following:

$$C_1(T) \leq \sqrt{C_1^{(n)}(T)^2 + \frac{C_1(T)^2}{n^2}}, \qquad C_2(T) \leq \sqrt{C_2^{(n)}(T)^2 + \frac{C_2(T)^2}{n^2}},$$
$$C_3(T) \leq \sqrt{C_3^{(n)}(T)^2 + \frac{C_3(T)^2}{n^4}}, \qquad C_4(T) \leq \sqrt{C_4^{(n)}(T)^2 + \frac{C_4(T)^2}{n^2}},$$
$$C_4(T) \leq \sqrt{C_4^{(n)}(T)^2 + \frac{C_2(T)^2}{n^2}},$$

which leads us to the conclusion. □

This result shows that we can bound the constants $C_1(T)\dots C_4(T)$ by means of $C_1^{(n)}(T)\dots C_4^{(n)}(T)$. We can compute $C_1^{(n)}(T)\dots C_4^{(n)}(T)$ numerically and also obtain guaranteed results via the numerical verification method.

5. Numerical results

In this section, we show the values of the upper bounds for $C_1(T)\dots C_4(T)$ obtained by Theorem 1, that of $K_1(T)\dots K_4(T)$ in Section 1 and that of $C_1(T)\dots C_4(T)$ themselves. We can calculate $C_1^{(n)}(T)\dots C_4^{(n)}(T)$ via the numerical verification method with interval arithmetic using INTLAB, the MATLAB toolbox for the reliable computing [18, 22]. Let $\overline{C}_1^{(n)}(T)\dots\overline{C}_4^{(n)}(T)$ be the upper endpoints of the calculated intervals by INTLAB, then from Theorem 1, the upper bounds for $C_1(T)\dots C_4(T)$ are obtained as follows:

$$\overline{\overline{C}}_1^{(n)}(T) = \sqrt{\frac{n^2}{n^2-1}}\,\overline{C}_1^{(n)}(T), \qquad \overline{\overline{C}}_2^{(n)}(T) = \sqrt{\frac{n^2}{n^2-1}}\,\overline{C}_2^{(n)}(T),$$
$$\overline{\overline{C}}_3^{(n)}(T) = \sqrt{\frac{n^4}{n^4-1}}\,\overline{C}_3^{(n)}(T), \qquad \overline{\overline{C}}_4^{(n)}(T) = \sqrt{\frac{n^2}{n^2-1}}\,\overline{C}_4^{(n)}(T),$$
$$\overline{\overline{C}}_4^{\prime(n)}(T) = \sqrt{\overline{C}_4^{(n)}(T)^2 + \frac{\overline{\overline{C}}_2^{(n)}(T)^2}{n^2}}.$$

As for $C_1(T)\dots C_4(T)$ themselves, we cannot determine their values analytically. Therefore, we first compute the following values for $n \le 10$:

$$\widetilde{C}_1^{(n)}(T) = \sup_{u\in V^{1,1}(T)\cap\mathcal{P}_n\setminus 0} \frac{\|u\|_{L^2(T)}}{\|\nabla u\|_{L^2(T)}}, \qquad \widetilde{C}_2^{(n)}(T) = \sup_{u\in V^{1,2}(T)\cap\mathcal{P}_n\setminus 0} \frac{\|u\|_{L^2(T)}}{\|\nabla u\|_{L^2(T)}},$$
$$\widetilde{C}_3^{(n)}(T) = \sup_{u\in V^{2}(T)\cap\mathcal{P}_n\setminus 0} \frac{\|u\|_{L^2(T)}}{|u|_{H^2(T)}}, \qquad \widetilde{C}_4^{(n)}(T) = \sup_{u\in V^{2}(T)\cap\mathcal{P}_n\setminus 0} \frac{\|\nabla u\|_{L^2(T)}}{|u|_{H^2(T)}},$$

where $\mathcal{P}_n$ denote the space of polynomials with degree less than or equal to n, then apply the repeated Aitken extrapolation to obtain more accurate approximations $\widetilde{C}_1(T)\dots\widetilde{C}_4(T)$.

In the following tables, all numerical results are rounded up to seven decimal places. Note that $T_{a,b},\ 0 \le a \le 0.5,\ 0 < b \le 1$ provides all shapes of triangles and, due to the scaling property, the relative error between the upper bounds and the optimal values depends only on the shape of the triangle.

The numerical results show that the sharp and explicit upper bounds are obtained by our method and the formulas introduced in Section 1. We also checked that

$$\overline{\overline{C}}_j^{(20)}(T_{a,b}) < K_j(T_{a,b}), \qquad j = 1, 2, 3,$$
$$\overline{\overline{C}}_4^{\prime(20)}(T_{a,b}) < K_4(T_{a,b}),$$

holds for every triangles with $(a, b) = (k/100, l/100),\ 0 \le k \le 50,\ 1 \le l \le 100$.

T	Shape	$K_1(T)$	$\overline{\overline{C}}_1^{(10)}(T)$	$\overline{\overline{C}}_1^{(20)}(T)$	$\widetilde{C}_1(T)$
$T_{0,\,1}$	Isosceles right triangle	0.3340766	0.3212290	0.3190436	0.3183099
$T_{0,\,1/2}$		0.2771024	0.2740807	0.2723761	0.2718064
$T_{0,\,1/5}$		0.2681080	0.2648395	0.2632425	0.2627047
$T_{0,\,1/10}$		0.2674398	0.2635352	0.2619488	0.2614141
$T_{1/4,\,1}$		0.3030136	0.2911752	0.2893022	0.2886729
$T_{1/4,\,1/2}$		0.2459843	0.2436090	0.2420943	0.2415907
$T_{1/4,\,1/5}$		0.2434617	0.2329771	0.2312917	0.2307200
$T_{1/4,\,1/10}$		0.2420732	0.2310303	0.2292291	0.2285833
$T_{1/2,\sqrt{3}/2}$	Equilateral triangle	0.2683033	0.2408094	0.2392551	0.2387325
$T_{1/2,\,1/2}$	Isosceles right triangle	0.2362278	0.2271432	0.2255927	0.2250791
$T_{1/2,\,1/5}$		0.2350309	0.2150884	0.2129926	0.2122547
$T_{1/2,\,1/10}$		0.2327945	0.2124695	0.2100807	0.2091564

Table 1: Calculation results for $C_1(T)$.

6. Circumradius and $C_4(T)$

In Section 1, we claimed that the following estimate holds for the interpolation constant $C_4(T)$:

$$C_4(T) < K_4(T) = \sqrt{\frac{A^2B^2C^2}{16S^2} - \frac{A^2+B^2+C^2}{30} - \frac{S^2}{5}\left(\frac{1}{A^2}+\frac{1}{B^2}+\frac{1}{C^2}\right)},$$

T	Shape	$K_2(T)$	$\overline{\overline{C}}_2^{(10)}(T)$	$\overline{\overline{C}}_2^{(20)}(T)$	$\widetilde{C}_2(T)$
$T_{0,1}$	 Isosceles right triangle	0.2417625	0.2396039	0.2381772	0.2377024
$T_{0,1/2}$		0.2001158	0.1998408	0.1985657	0.1981418
$T_{0,1/5}$		0.1931751	0.1916921	0.1904436	0.1900288
$T_{0,1/10}$		0.1926085	0.1906412	0.1893972	0.1889838
$T_{1/4,1}$		0.2197865	0.2177021	0.2164124	0.2159829
$T_{1/4,1/2}$		0.1779313	0.1782025	0.1770818	0.1767091
$T_{1/4,1/5}$		0.1753980	0.1720157	0.1709011	0.1705287
$T_{1/4,1/10}$		0.1743207	0.1711858	0.1700506	0.1696660
$T_{1/2,\sqrt{3}/2}$	 Equilateral triangle	0.1948780	0.1906371	0.1895418	0.1891770
$T_{1/2,1/2}$	 Isosceles right triangle	0.1709519	0.1694255	0.1684167	0.1680810
$T_{1/2,1/5}$		0.1693067	0.1645693	0.1635627	0.1632276
$T_{1/2,1/10}$		0.1676363	0.1638830	0.1628606	0.1625187

Table 2: Calculation results for $C_2(T)$.

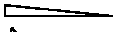
T	Shape	$K_3(T)$	$\overline{\overline{C}}_3^{(10)}(T)$	$\overline{\overline{C}}_3^{(20)}(T)$	$\widetilde{C}_3(T)$
$T_{0,1}$	 Isosceles right triangle	0.1702674	0.1684446	0.1675538	0.1672540
$T_{0,1/2}$		0.1184266	0.1180690	0.1175455	0.1173699
$T_{0,1/5}$		0.1107396	0.1096648	0.1092458	0.1091056
$T_{0,1/10}$		0.1099925	0.1087203	0.1083189	0.1081843
$T_{1/4,1}$		0.1487598	0.1464850	0.1458512	0.1456392
$T_{1/4,1/2}$		0.0950296	0.0946780	0.0942616	0.0941222
$T_{1/4,1/5}$		0.0855113	0.0849795	0.0844707	0.0842867
$T_{1/4,1/10}$		0.0843545	0.0837111	0.0831606	0.0829448
$T_{1/2,\sqrt{3}/2}$	 Equilateral triangle	0.1201799	0.1177043	0.1172419	0.1170872
$T_{1/2,1/2}$	 Isosceles right triangle	0.0851337	0.0842223	0.0837769	0.0836270
$T_{1/2,1/5}$		0.0732579	0.0727068	0.0719786	0.0716964
$T_{1/2,1/10}$		0.0715702	0.0710650	0.0702398	0.0698864

Table 3: Calculation results for $C_3(T)$.

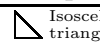
T	Shape	$K_4(T)$	$\overline{\overline{C}}_4^{(10)}(T)$	$\overline{\overline{C}}_4'^{(10)}(T)$	$\overline{\overline{C}}_4'^{(20)}(T)$	$\widetilde{C}_4(T)$
$T_{0,1}$	 Isosceles right triangle	0.4915961	0.4912760	0.4894003	0.4888906	0.4887225
$T_{0,1/2}$		0.3958115	0.3827571	0.3813624	0.3809004	0.3807482
$T_{0,1/5}$		0.3697886	0.3384254	0.3372742	0.3367584	0.3365883
$T_{0,1/10}$		0.3662945	0.3297106	0.3286114	0.3280661	0.3278854
$T_{1/4,1}$		0.4063828	0.3983769	0.3969774	0.3964682	0.3963006
$T_{1/4,1/2}$		0.3393941	0.3273684	0.3262146	0.3257826	0.3256403
$T_{1/4,1/5}$		0.5516444	0.5415574	0.5391173	0.5389133	0.5388452
$T_{1/4,1/10}$		0.9871946	0.9796800	0.9749196	0.9748225	0.9747889
$T_{1/2,\sqrt{3}/2}$	 Equilateral triangle	0.3476109	0.3200270	0.3189930	0.3185477	0.3184013
$T_{1/2,1/2}$	 Isosceles right triangle	0.3476109	0.3473846	0.3460583	0.3456979	0.3455790
$T_{1/2,1/5}$		0.6761400	0.6663349	0.6631990	0.6630533	0.6630043
$T_{1/2,1/10}$		1.2786662	1.2752049	1.2689187	1.2688525	1.2688286

Table 4: Calculation results for $C_4(T)$.

where A, B, C are the edge lengths of triangle T and S is the area of T. Since the circumradius of T is given by

$$R(T) = \frac{ABC}{4S},$$

we have the estimation

$$C_4(T) < R(T).$$

This fact is full of interesting suggestions for the error analysis in the Finite Element Method. See [9, 10] for the details.

7. Conclusion

We present a simple method to obtain sharp upper bounds for the interpolation error constants over the given triangular elements. These constants are important for analysis of interpolation error and especially for the error analysis in the Finite Element Method. Guaranteed upper bounds for these constants are obtained via the numerical verification method. Furthermore, we introduce remarkable formulas for the upper bounds of these constants. By the method explained in this paper, we can only prove these formulas for the given triangles. However, using some continuation techniques and asymptotic analysis, we are able to extend our results to the general cases. We will show the general proof in a forthcoming publication.

Acknowledgements

This work was supported by JSPS Grant-in-Aid for Scientific Research (C) Grant Number 25400198.

References

[1] Andreev, A. and Racheva, M.: Two-sided bounds of eigenvalues of second- and fourth-order elliptic operators. Appl. Math. **59** (2014), 371–390.

[2] Arcangeli, R. and Gout, J. L.: Sur l'évaluation de i'erreur d'interpolation de Lagrange dans un ouvert de $\mathbb{R}^n$. R.A.I.R.O. Analyse Numérique **10** (1976), 5–27.

[3] Babuška, I. and Aziz, A. K.: On the angle condition in the finite element method. SIAM J. Numer. Anal. **13** (1976), 214–226.

[4] Brenner, S. C. and Scott, L. R.: *The mathematical theory of Finite Element Methods*. Springer, 2002.

[5] Carstensen, C. and Gedicke, J.: Guaranteed lower bounds for eigenvalues. Math. Comp. **83(290)** (2014), 2605–2629.

[6] Ciarlet, P. G.: *The Finite Element Method for elliptic problems*. SIAM, 2002.

[7] Kikuchi, F. and Liu, X.: Estimation of interpolation error constants for the p_0 and p_1 triangular finite elements. Comput. Methods Appl. Mech. Engrg. **196** (2007), 3750–3758.

[8] Kikuchi, F. and Saito, H.: Remarks on a posteriori error estimation for finite element solutions. J. Comp. Appl. Math. **199** (2007), 329–336.

[9] Kobayashi, K. and Tsuchiya, T.: A Babuška-Aziz type proof of the circumradius condition. Japan J. Indust. Appl. Math. **31** (2014), 193–210.

[10] Kobayashi, K. and Tsuchiya, T.: On the circumradius condition for piecewise linear triangular elements. Japan J. Indust. Appl. Math. **32** (2015), 65–76.

[11] Laugesen, R. S. and Siudeja, B. A.: Minimizing Neumann fundamental tones of triangles: An optimal Poincaré inequality. J. Differential Equations **249** (2010), 118–135.

[12] Lehmann, R.: Computable error bounds in finite-element method. IMA J. Numer. Anal. **6** (1986), 265–271.

[13] Li, Q., Lin, Q., and Xie, H.: Nonconforming finite element approximations of the Steklov eigenvalue problem and its lower bound approximations. Appl. Math. **58** (2013), 129–151.

[14] Liu, X. and Kikuchi, F.: Analysis and estimation of error constants for p_0 and p_1 interpolations over triangular finite elements. J. Math. Sci. Univ. Tokyo **17** (2010), 27–78.

[15] Liu, X. and Oishi, S.: Guaranteed high-precision estimation for p_0 interpolation constants on triangular finite elements. Japan J. Indust. Appl. Math. **30** (2013), 635–652.

[16] Luo, F., Lin, Q., and Xie, H.: Computing the lower and upper bounds of Laplace eigenvalue problem: by combining conforming and non-conforming Finite Element Methods. Science China Mathematics **55** (2012), 1069–1082.

[17] Meinguet, J. and Descloux, J.: An operator-theoretical approach to error estimation. Numer. Math. **27** (1977), 307–326.

[18] Moore, R.E., Kearfott, R.B., and Cloud, M.J.: *Introduction to interval analysis.* Cambridge Univ. Press, 2009.

[19] Nakao, M. T. and Yamamoto, N.: A guaranteed bound of the optimal constant in the error estimates for linear triangular element. Comput. Suppl. **15** (2001), 163–173.

[20] Natterer, F.: Berechenbare Fehlerschranken für die Methode der Finite Elemente. Internat. Ser. Numer. Math. **28** (1975), 109–121.

[21] Repin, S.I.: Computable majorants of constants in the Poincaré and Friedrichs inequalities. J. Math. Sci. **186** (2012), 307–321.

[22] Rump, S.M.: Verification methods: Rigorous results using floating-point arithmetic. Acta Numer. **19** (2010), 287–449.

[23] Sebestova, I. and Vejchodsky, T.: Two-sided bounds for eigenvalues of differential operators with applications to Friedrichs', Poincaré, trace, and similar constants. SIAM J. Numer. Anal. **52** (2014), 308–329.

[24] Zlámal, M.: On the Finite Element Method. Numer. Math. **12** (1968), 394–409.

Conference Applications of Mathematics 2015, in honor of the birthday anniversaries of Ivo Babuška (90), Milan Práger (85), and Emil Vitásek (85)
J. Brandts, S. Korotov, M. Křížek, K. Segeth, J. Šístek, T. Vejchodský (Eds.)
Institute of Mathematics CAS, Prague 2015

WHY QUINTIC POLYNOMIAL EQUATIONS ARE NOT SOLVABLE IN RADICALS

Michal Křížek[1], Lawrence Somer[2]

[1] Institute of Mathematics, Academy of Sciences
Žitná 25, CZ – 115 67 Prague 1, Czech Republic
krizek@math.cas.cz
[2] Department of Mathematics, Catholic University of America
Washington, D.C. 20064, USA
somer@cua.edu

Abstract: We illustrate the main idea of Galois theory, by which roots of a polynomial equation of at least fifth degree with rational coefficients cannot general be expressed by radicals, i.e., by the operations $+$, $-$, $\cdot$, $:$, and $\sqrt[n]{\cdot}$. Therefore, higher order polynomial equations are usually solved by approximate methods. They can also be solved algebraically by means of ultraradicals.

Keywords: Galois theory, finite group, permutation, radical

MSC: 20D05, 13B05, 65H05

1. A brief historical survey

A classic problem in mathematics has been to solve polynomial equations with rational coefficients in terms of its coefficients by means of the operations $+$, $-$, $\cdot$, $:$, and $\sqrt[n]{\cdot}$ (this is the radical symbol and involves taking nth roots). For example, we can solve the quadratic equation

$$ax^2 + bx + c = 0, \quad a, b, c \in \mathbb{R},\ a \neq 0,$$

by the well-known quadratic formula

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}. \tag{1}$$

In the early to mid-1500s, solutions to the cubic and quartic equations by means of radicals were given by the Italian mathematicians Scipione del Ferro, Niccolo Tartaglia, Antonio Fiore, Gerolamo Cardano, and Lodovico Ferrari. In 1545, Cardano published an account of solutions of cubic and quartic equations by radicals in

Ars Magna [4]. By a suitable linear transfomation any cubic polynomial equation with real coefficients can be reduced to the form

$$x^3 + bx + c = 0,$$

for which Cardano proposed the following solution

$$x = \sqrt[3]{-\frac{c}{2} + \sqrt{\left(\frac{c}{2}\right)^2 + \left(\frac{b}{3}\right)^3}} - \sqrt[3]{\frac{c}{2} + \sqrt{\left(\frac{c}{2}\right)^2 + \left(\frac{b}{3}\right)^3}}.$$

For instance, the equation

$$x^3 + 9x - 26 = 0$$

implies that

$$x = \sqrt[3]{13 + \sqrt{13^2 + 3^3}} - \sqrt[3]{-13 + \sqrt{196}} = \sqrt[3]{27} - \sqrt[3]{1} = 2,$$

and thus this root can be separated:

$$x^3 + 9x - 26 = (x - 2)(x^2 + 2x + 13) = 0.$$

By (1), the remaining two roots are $x = 1 \pm \mathrm{i}2\sqrt{3}$.

Note that by a sophisticated transformation the solution of a quartic polynomial equation can be reduced to the solution of a cubic polynomial equation (see [13, p. 42]).

For centuries, it was an open question whether there existed a solution to the general quintic (fifth degree) equation by radicals. This question was settled in the negative by the Norwegian mathematician Niels Henrik Abel in 1824 (see [1, 2, 3]).

In this paper, we show that the equation

$$f(x) = 2x^5 - 10x + 5 = 0 \tag{2}$$

is not solvable by radicals [8].

We note that the derivative $f'(x) = 10x^4 - 10$ has exactly two real roots ± 1. Moreover, $f''(x) = 40x^3$ and the second derivative test of elementary calculus shows that f has one positive relative maximum at $x = -1$, one negative relative minimum at $x = 1$, and one point of inflection at $x = 0$. It is clear that the polynomial f has exactly three real zeros (cf. Fig. 1). Since its coefficients are real, we also see that f has exactly two imaginary zeros which are complex conjugates of each other.

2. Galois theory

We will show that equation (2) is not solvable by radicals by the use of Galois theory, named after the French mathematician Evariste Galois. In 1830, Galois wrote a groundbreaking paper [5] (see also [6]) that gave a criterion for determining whether any polynomial f of degree n with rational coefficients is solvable by radicals.

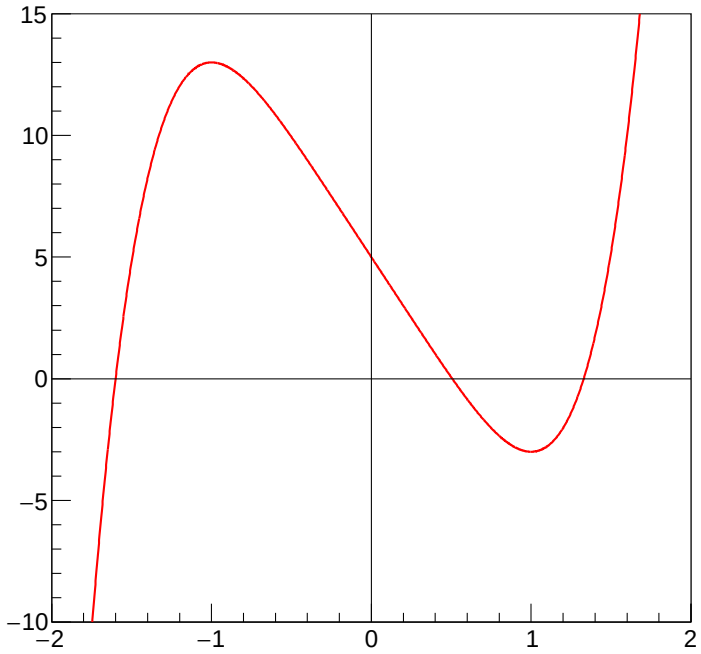


Figure 1: Graph of $y = f(x) = 2x^5 - 10x + 5$.

This criterion involves the Galois group G which is a group of permutations on the n roots of the polynomial f. Recall by the fundamental theorem of algebra that any polynomial of degree n has n roots over the complex numbers $\mathbb{C}$. Each element of the Galois group G transforms any valid polynomial equation with rational coefficients involving the roots of f into another valid equation involving these roots.

Let us take an example. Consider the polynomial equation

$$p(x) = x^4 - 4x - 5 = (x^2 + 1)(x^2 - 5) = 0. \tag{3}$$

There are four zeros: $x \in \{\pm \mathrm{i}, \pm\sqrt{5}\}$. It is clear that they form two natural pairs: i and $-\mathrm{i}$ go together and so do $\sqrt{5}$ and $-\sqrt{5}$. Indeed, it is impossible to distinguish i from $-\mathrm{i}$ and $\sqrt{5}$ from $-\sqrt{5}$ in the following sense. Write down any polynomial equation with rational coefficients that is satisfied by some selection from the four zeros. If we let

$$\alpha = \mathrm{i}, \quad \beta = -\mathrm{i}, \quad \gamma = \sqrt{5}, \quad \delta = -\sqrt{5},$$

then such equations include

$$\alpha^2 + 1 = 0, \quad \alpha + \beta = 0, \quad \delta^2 - 5 = 0, \quad \gamma + \delta = 0, \quad \alpha\gamma - \beta\delta = 0, \tag{4}$$

and so on. There are infinitely many valid equations of this kind. If we take any valid equation connecting α, β, γ, and δ and interchange α and β, we again get

a valid equation. The same is true if we interchange γ and δ. For example, the above equations lead by this process to

$$\beta^2 + 1 = 0, \quad \beta + \alpha = 0, \quad \gamma^2 - 5 = 0, \quad \delta + \gamma = 0, \quad \beta\gamma - \alpha\delta = 0,$$
$$\alpha\delta - \beta\gamma = 0, \quad \beta\delta - \alpha\gamma = 0,$$

and all of these are true. On the other hand, if we interchange α and γ, the second equation in (4) leads to the equation $\gamma + \beta = 0$, which is false.

The operations we are using are permutations of the zeros α, β, γ, and δ and thus are elements of S_4, which includes all $4! = 24$ possible permutations of the four symbols α, β, γ, and δ. In fact, in the usual permutation notation, the interchange of α and β is

$$R = \begin{pmatrix} \alpha & \beta & \gamma & \delta \\ \beta & \alpha & \gamma & \delta \end{pmatrix}$$

and that of γ and δ is

$$S = \begin{pmatrix} \alpha & \beta & \gamma & \delta \\ \alpha & \beta & \delta & \gamma \end{pmatrix}.$$

If these two permutations transform valid equations into valid equations, then so does the permutation obtained by performing them both in turn, which is

$$T = \begin{pmatrix} \alpha & \beta & \gamma & \delta \\ \beta & \alpha & \delta & \gamma \end{pmatrix}.$$

There is, of course, one other permutation with this property of preserving all valid equations, namely the identity permutation

$$I = \begin{pmatrix} \alpha & \beta & \gamma & \delta \\ \alpha & \beta & \gamma & \delta \end{pmatrix}.$$

One can check that only these four permutations in S_4 preserve valid equations, while the other twenty permutations in S_4 can turn a valid equation into a false equation. We can write permutations as products of disjoint cycles. Thus, using cycle notation, we can rewrite R, S, T, and I as

$$\begin{aligned} R &= (\alpha\beta)(\gamma)(\delta), \\ S &= (\gamma\delta)(\alpha)(\beta), \\ T &= (\alpha\beta)(\gamma\delta), \\ I &= (\alpha)(\beta)(\gamma)(\delta). \end{aligned}$$

Note that the permutations R, S, T, and I form a subgroup of S_4 under the operation of composition of permutations. Then we call

$$G = \{I, R, S, T\}$$

the Galois group of the equation (3).

3. Application of Galois theory to the quintic polynomials

We use the following facts from Galois theory (see [8, pp. 371–398] or [14]) to show that the equation (2) is not solvable in radicals. Note that the quintic equation (2) has 5 roots in $\mathbb{C}$ and thus its Galois group is a subgroup of S_5 with $5! = 120$ elements.

(A) *A quintic polynomial equation with rational coefficients is not solvable by radicals if its Galois group G is equal to S_5.*

(B) *If a polynomial with rational coefficients has degree n and is irreducible over the rationals, then the order of its Galois group G is divisible by n.*

(C) *By Cauchy's Theorem, if the order of a finite group is divisible by a prime p, then it has an element of order p.*

(D) *Let p be a prime. Then any element of order p in S_p is a p-cycle.*

(E) *By Eisenstein's Criterion, the polynomial*

$$f(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$$

with integer coefficients is irreducible over the rationals if there exists a prime p such that p does not divide a_n, p divides each of $a_{n-1}, a_{n-2}, \ldots, a_1, a_0$, and p^2 does not divide a_0.

(F) *Let f be a polynomial of degree n with rational coefficients. Suppose that exactly $n-2$ of the roots of f are real and the other two roots are imaginary. Let r_1 and r_2 be the two imaginary roots. Then r_1 and r_2 are complex conjugates of each other and the Galois group G of f contains the two-cycle $(r_1 r_2)(r_3)(r_4)\ldots(r_n)$. This mapping corresponds to complex conjugation which takes imaginary roots into their complex conjugate and leaves real roots fixed.*

(G) *Let f be a polynomial of prime degree p with rational coefficients. If the Galois group G of f contains both a p-cycle and a 2-cycle, then $G = S_p$.*

Theorem. *The polynomial equation* (2) *is not solvable by radicals.*

Proof. Let G be the Galois group of f. We will show that $G = S_5$. It will then follow by (A) that the equation $f(x) = 0$ is not solvable by radicals. By Fig. 1 and our earlier discussion, f has exactly three real roots and two imaginary roots r_1 and r_2 which are complex conjugates of each other. By Eisenstein's Criterion (E) with $p = 5$, we find that f is irreducible over the rationals. It follows by (B) that the order of G is divisible by 5. Since 5 is prime, we see by Cauchy's Theorem (C), that G has an element of order 5. Then by (D), we get that G contains a 5-cycle. By (F), G contains the 2-cycle $(r_1 r_2)(r_3)(r_4)\ldots(r_n)$. It now follows by (G) that the Galois group $G = S_5$. Hence, the equation (2) is not solvable by radicals. □

4. Conclusions

For a popular account of Galois theory, see [11]. It can be shown that for any $n \geq 5$ there exists a polynomial equation of degree n which is not solvable by radicals. This follows from Galois' Theorem which states: *The alternating group A_n is simple for $n \geq 5$* (see [10, p. 311]). Therefore, higher order polynonomial equations are usually solved by approximate methods (numerical, statistical, etc.). For example, the Lehmer-Schur method produces guaranteed error estimates, i.e., we can find arbitrarily small circles in the complex plane containing all roots of any polynomial (see [12]).

Note that the general quintic equation with rational coefficients can also be solved algebraically by other means than the use of radicals. Suppose that for any real number a we define the ultraradical $\sqrt[*]{a}$ to be the real zero of $x^5 + x - a$. It was shown by Erland Samuel Bring in 1796 and by George Birch Jerrard in 1852 (see [9]) that the quintic equation can be solved by the use of radicals and ultraradicals. In 1858, Charles Hermite [7] proved that the quintic equation can be solved in terms of elliptic modular functions.

Acknowledgement

This paper was supported by RVO 67985840.

References

[1] Abel, N. H.: *Mémoire sur les équations algébriques, on l'on démontré l'impossibilité de l'équation générale du cinquiéme degré*, (1824), Oeuvres Complètes de Niels Henrik Abel, vol. 1, Grøndahl, Christiana, 1881.

[2] Abel, N. H.: Beweis der Unmöglichkeit, algebraische Gleichungen von höheren Graden, als dem vierten, allgemein aufzuösen. J. Reine Angew. Math. **1** (1826), 65–84.

[3] Abel, N. H.: Démostration de l'impossibilité de la résolution des équations algébrique générales d'un degré supérieur du quatrième. Bulletin des scinces mathématiques, astronomiques, physiques et chimiques **6** (1826), 347–354.

[4] Cardano, G.: *Ars Magna of the rules of algebra*. T. R. Witmer, trans. and ed., Dover Publications, Mineola, New York, 1993, original edition 1545.

[5] Galois, E.: Mémoire sur les conditions de résolubilité des équations par radicaux. J. Math. Pures Appl. (9) (1830), 417–433.

[6] Galois, E.: Oeuvres mathématiques d'Évariste Galois, J. Math. Pures Appl. (9) **11** (1846), 381–444.

[7] Hermite, C.: Sur la résolution de l'équation du cinquième degré. Comptes Rendus de l'Academie des Sciences **46** (1858), 508–515.

[8] Hungerford, T. W.: *Abstract algebra. An introduction*, 2nd edition. Saunders College Publishing, Orlando, 1997.

[9] Jerrard, G. B.: *An essay on the resolution of equations*. Taylor and Francis, London, 1859.

[10] Křížek, M., Somer, L.: Architects of symmetry in finite nonabelian groups. Symmetry: Culture and Science **21** (2010), 333–344.

[11] Livio, M.: *The equation that couldn't be solved*. Simon and Schuster, New York, 2005.

[12] Ralston, A.: *A first course in numerical analysis*. McGraw-Hill, 1965.

[13] Rektorys, K., et al.: *Survey of applicable mathematics*, vol. I, 2nd edition. Kluwer, Dordrecht, 1994.

[14] Stewart, I.: *Galois theory*, 2nd edition. Chapman and Hall, London, 1989.

Conference Applications of Mathematics 2015, in honor of the birthday anniversaries of Ivo Babuška (90), Milan Práger (85), and Emil Vitásek (85)
J. Brandts, S. Korotov, M. Křížek, K. Segeth, J. Šístek, T. Vejchodský (Eds.)
Institute of Mathematics CAS, Prague 2015

A NOTE ON NECESSARY AND SUFFICIENT CONDITIONS FOR CONVERGENCE OF THE FINITE ELEMENT METHOD

Václav Kučera

Charles University in Prague, Faculty of Mathematics and Physics
Sokolovská 83, 186 75 Praha, Czech Republic
kucera@karlin.mff.cuni.cz

Abstract: In this short note, we present several ideas and observations concerning finite element convergence and the role of the maximum angle condition. Based on previous work, we formulate a hypothesis concerning a necessary condition for $O(h)$ convergence and show a simple relation to classical problems in measure theory and differential geometry which could lead to new insights in the area.

Keywords: finite element method, a priori error estimates, maximum angle condition

MSC: 65N30, 65N15, 53A05

1. Introduction

The finite element method (FEM) is among the most popular, if not the single most popular numerical method for the solution of partial differential equations. The theory and practice of the FEM has a long and rich history. One of the main questions is, of course, "when does it work". Specifically, for piecewise linear FEM, we ask when does the FEM have optimal $O(h)$ convergence in the $H^1(\Omega)$-norm. It was believed that the so-called *maximum angle condition* is a sufficient as well as necessary condition for $O(h)$ convergence. While the first is true, cf. [1, 2], the maximum angle condition is not *necessary* for $O(h)$ convergence, as was recently shown in [2] by a simple argument.

While the author believes that we are still far away from formulating a necessary and also sufficient condition for $O(h)$ convergence, in this short note we present some ideas and observations related to this question.

In Section 2.1, we investigate the refinement procedure from [2], where maximum angle condition satisfying triangulations are arbitrarily subdivided to obtain maximum angle violating triangulations with $O(h)$ convergence. We show that by such refinement, one cannot obtain triangulations containing *only* degenerate elements.

In Section 2.2, we review the only known counterexample of Babuška and Aziz [1], which together with the results of Section 2.1 leads to formulating a hypothesis on a necessary condition for $O(h)$ convergence. Namely, we hypothesize that elements satisfying the maximum angle condition must be "dense" in Ω.

We are unable to prove the presented hypothesis, in fact as far as the author knows, no nontrivial necessary condition is known in the literature. In Section 3, we present a simple connection between the question of FEM convergence and differential geometry, which could possibly give alternative insight into these and related questions of FEM convergence.

2. A hypothesis on a necessary condition for $O(h)$ convergence

We treat the following problem: Find $u : \Omega \subset \mathbb{R}^2 \to \mathbb{R}$ such that

$$-\Delta u = f, \quad u|_{\partial\Omega} = 0, \tag{1}$$

where Ω is a bounded polygonal domain with a Lipschitz boundary and $f \in L^2(\Omega)$. Defining $V = H_0^1(\Omega)$ and the associated bilinear form $a(u, v) = \int_\Omega \nabla u \cdot \nabla v \, \mathrm{d}x$, the corresponding weak form of (1) reads: Find $u \in V$ such that

$$a(u, v) = (f, v) \quad \forall v \in V.$$

The finite element method constructs a sequence of spaces $\{V_h\}_{h\in(0,h_0)}$ on conforming triangulations $\{\mathcal{T}_h\}_{h\in(0,h_0)}$ of Ω, where $V_h \subset V$ consists of globally continuous piecewise linear functions on $\mathcal{T}_h$. The FEM formulation then reads: Find $u_h \in V_h$ such that

$$a(u_h, v_h) = (f, v_h) \quad \forall v_h \in V_h.$$

Denoting h as the maximal diameter of all elements K in $\mathcal{T}_h$, the natural measure of convergence of u_h to u is estimation by powers of h. Specifically, in the energy norm of (1), we obtain at most $O(h)$ convergence if $u \in H^2(\Omega)$, i.e.

$$|u - u_h|_{H^1(\Omega)} \leq C(u)h \quad \forall u \in H^2(\Omega) \cap V, \tag{2}$$

where the constant $C(u)$ is typically written in the form $C|u|_{H^2(\Omega)}$. The question is, when can such a result be proved. Currently, the most general sufficient condition known for (2) is the maximum angle condition:

Definition 1. *A system of triangulations $\{\mathcal{T}_h\}_{h\in(0,h_0)}, h_0 > 0$ satisfies the* maximum angle condition*, if there exists $\alpha < \pi$ such that all angles in all triangles $K \in \mathcal{T}_h$ are less than α for all $h \in (0, h_0)$.*

Recently it was shown that the maximum angle condition is not necessary for $O(h)$ convergence. In [2], triangulations $\mathcal{T}_h$ satisfying Definition 1 are refined by subdividing each triangle $K \in \mathcal{T}_h$ arbitrarily, thus obtaining new triangulations $\widetilde{\mathcal{T}}_h$.

Since $\mathcal{T}_h$ satisfies Definition 1 and $\widetilde{\mathcal{T}}_h$ is a refinement of $\mathcal{T}_h$, then since $\mathcal{T}_h$ allows $O(h)$ convergence, so must $\widetilde{\mathcal{T}}_h$ by Céa's lemma.

The natural question arises, how far can one take these refinements. For example, taking $\mathcal{T}_h$ satisfying the maximum angle condition, can one construct $\widetilde{\mathcal{T}}_h$ such that the maximal angles of all triangles are arbitrarily close to π? We answer this question negatively in the following section.

2.1. Mesh subdivisions

In the following, we will need to distinguish between triangles $K \in \mathcal{T}_h$ "satisfying" and "violating" the maximum angle condition. Of course, this depends on the choice of α in Definition 1. For this purpose, we fix some α arbitrarily close to π. We will call K with maximum angle larger than α *degenerate* and *non-degenerate* otherwise. This terminology is clear: if $\{\mathcal{T}_h\}_{h\in(0,h_0)}$ violates Definition 1, than there exist triangles K in some $\mathcal{T}_h$ such that their maximum angle is arbitrarily close to π. Hence Definition 1 is violated for any choice of α. Therefore, in the end, the "maximum angle violating" property is independent of the specific choice of α.

Definition 2. *Let $\alpha \in (0, \pi)$ be close to π. A triangle $K \in \mathcal{T}_h$ is called* degenerate*, if the maximum angle in K is $> \alpha$. Otherwise, K is called* non-degenerate*.*

Now we show that a non-degenerate triangle K cannot be cut into degenerate triangles *only*. Hence the construction from [2] cannot give triangulations containing *only* degenerate triangles.

Lemma 1. *Let $\alpha \in (\frac{2}{3}\pi, \pi)$. Let K be a triangle with all angles less than α. Then there does not exist a finite conforming partition of K into triangles which all contain an angle greater than or equal to α.*

Proof. Assume on the contrary that such a partition $\mathcal{P}$ exists. Let t = number of triangles in $\mathcal{P}$, v_I = number of vertices of $\mathcal{P}$ contained in the interior of K and v_B = number of vertices of $\mathcal{P}$ lying on ∂K.

On one hand, the sum of all angles in $\mathcal{P}$ is πt. On the other hand, the same sum can be calculated by summing all angles surrounding all interior, boundary and corner vertices in $\mathcal{P}$. Thus we get

$$\pi t = 2\pi v_I + \pi v_B + \pi,$$

which simplifies to the Euler-type identity

$$t = 2v_I + v_B + 1. \tag{3}$$

Now, we calculate the number of angles in $\mathcal{P}$ that are greater than or equal to α. On one hand, we obtain t, since each triangle in $\mathcal{P}$ contains exactly one such angle. On the other hand, since $\alpha > \frac{2}{3}\pi$, each interior vertex of $\mathcal{P}$ can be the vertex

of (at most) two such angles. Similarly, each boundary vertex can be the vertex of (at most) one such angle and the corner vertices of T are all $< \alpha$. Thus

$$t \leq 2v_I + v_B. \tag{4}$$

Substituting (3) into (4), we get $1 \leq 0$, which is a contradiction. □

Lemma 1 can be interpreted informally in the following way: A non-degenerate triangle cannot be cut into degenerate triangles only, one always has at least one non-degenerate triangle in the resulting partition.

In [2], triangulations violating the maximum angle condition but possessing the $O(h)$ convergence property are constructed by taking a system of triangulations satisfying the maximum angle condition and subdividing each of its triangles arbitrarily. However, Lemma 1 states that each of these subdivided triangles K contains a triangle $\tilde{K}$ satisfying the same maximum angle condition as K. Therefore, using this procedure, one cannot produce large regions of degenerate triangles in the following sense.

Definition 3. *We say that the set of non-degenerate triangles is* dense *in* $\{\mathcal{T}_h\}_{h\in(0,h_0)}$, *if for all* $x \in \Omega$ *and all neighbourhoods* $\mathcal{U} \in \mathcal{O}(x)$ *there exists* $\tilde{h} \in (0, h_0)$ *such that for all* $h \in (0, \tilde{h})$ *there exists a non-degenerate* $K \in \mathcal{T}_h$ *such that* $K \subset \mathcal{U}$.

Now we will prove the main result, that using the procedure of [2], one can obtain only triangulations, where the set of non-degenerate triangles is dense in the sense of Definition 3. In particular, one cannot obtain triangulations with only degenerate triangles by subdividing triangulations satisfying the maximum angle condition.

Theorem 2. *Let* $\{\mathcal{T}_h\}_{h\in(0,h_0)}$ *satisfy the maximum angle condition. Let* $\{\widetilde{\mathcal{T}}_h\}_{h\in(0,h_0)}$ *be a set of conforming triangulations of* Ω *obtained from* $\{\mathcal{T}_h\}_{h\in(0,h_0)}$ *by subdividing each triangle in each* $\mathcal{T}_h$ *into a finite number of triangles. Then non-degenerate triangles are dense in* $\{\widetilde{\mathcal{T}}_h\}_{h\in(0,h_0)}$.

Proof. Choose $x \in \Omega$ and $\mathcal{U} \in \mathcal{O}(x)$. Then for sufficiently small $\tilde{h}$, for each $\mathcal{T}_h$, $h \in (0, \tilde{h})$ there exists $K \in \mathcal{T}_h$ such that $K \subset \mathcal{U}$. Since $\widetilde{\mathcal{T}}_h$ is obtained from $\mathcal{T}_h$ by subdividing each triangle, by Theorem 1 the partition of K must contain a non-degenerate triangle $\tilde{K}$. Since $K \subset \mathcal{U}$, then also $\tilde{K} \subset \mathcal{U}$, hence the set of non-degenerate triangles is dense in $\{\widetilde{\mathcal{T}}_h\}_{h\in(0,h_0)}$. □

2.2. The Babuška-Aziz counterexample

As far as the author is aware of, there exists only one counterexample to $O(h)$ convergence of the finite element method. This is the counterexample of Babuška and Aziz [1], further refined in [4]. The counterexample consists of a series of triangulations $\mathcal{T}_{m,n}$ of the unit square, where m and $2n$ denote the number of intervals into which the horizontal and vertical sides of the unit square are divided, cf. Figure 1. On these triangulations, the piecewise linear FEM is used to discretize Poisson's problem with the exact solution $\frac{1}{2}x(1-x)$. While the original paper [1] uses this

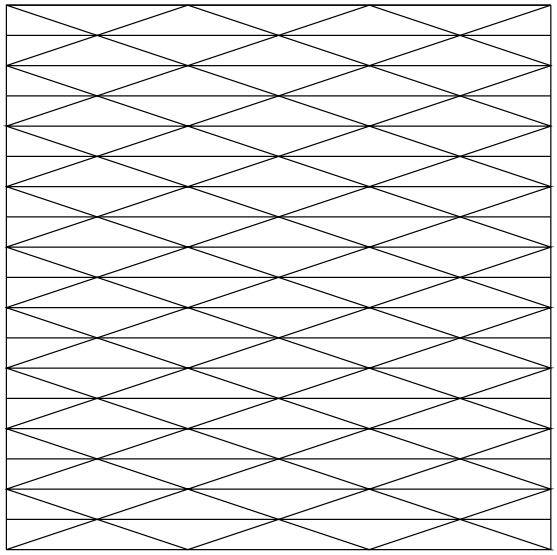


Figure 1: Babuška-Aziz triangulation $\mathcal{T}_{3,9}$ of the unit square

problem only to provide a counterexample to $O(h)$ convergence, in [4] a more detailed analysis is carried out and the error of the discrete solution $u_{m,n}$ is shown to satisfy

$$\|u - u_{m,n}\|_{H^1(\Omega)} \approx \min\{1, m/n^2\}. \tag{5}$$

In $\mathcal{T}_{m,n}$, the maximal edge length satisfies $h = 1/m$. Estimate (5) means that if the maximal condition is violated, i.e. $n \to \infty$ faster than m, then $O(h)$ convergence does not hold.

In the Babuška-Aziz counterexample, if all triangles (except the few triangles near the vertical boundaries) violate the maximal angle condition, then we lose $O(h)$ convergence. Therefore, in this counterexample, large open subsets of Ω containing only *degenerate* triangles destroy $O(h)$ convergence. In general, if another system of triangulations $\mathcal{T}_h$ coincided with $\mathcal{T}_{m,n}$ on a fixed open subset of Ω, then $\mathcal{T}_h$ would also not admit $O(h)$ convergence. Of course, $\mathcal{T}_{m,n}$ are highly structured, even periodic, and therefore represent only one possibility of triangulations containing only degenerate triangles. Theorem 2 states that such triangulations cannot be obtained by subdivision. Therefore, based on these considerations, we state the following hypothesis, which says that the result of Theorem 2 is a necessary condition for $O(h)$ convergence.

Hypothesis. *Let $\{\mathcal{T}_h\}_{h\in(0,h_0)}$ be a system of triangulations and $u_h \in V_h$ the corresponding discrete solutions. If $|u - u_h|_{H^1(\Omega)} \leq C(u)h$ for all $u \in H^2(\Omega)$, or some dense subset thereof, then triangles satisfying the maximum angle condition (non-degenerate triangles) are dense in $\{\mathcal{T}_h\}_{h\in(0,h_0)}$ in the sense of Definition* 3.

The strategy how to prove this hypothesis is as follows: show that triangulations similar to $\mathcal{T}_{m,n}$ violate $O(h)$ convergence. Here "similar" means that $\mathcal{T}_h$ should

contain only degenerate triangles (up to perhaps a few adjoining the boundary). However, for such general triangulation, we lack the simple structure of $\mathcal{T}_{m,n}$, which is an essential ingredient in the proofs presented in [1, 4].

3. Relation to differential geometry

The Babuška-Aziz counterexample is the only known counterexample to finite element convergence to date. Moreover, it is interesting that this counterexample coincides with a classical counterexample from the early stages of development of measure theory, the so-called *Schwarz lantern* [5]. The purpose of Schwarz's counterexample is to show that even for smooth surfaces, surface area cannot be defined as the limit of areas of approximating polyhedral surfaces. In fact, in Schwarz's counterexample, the surface areas of the approximating polyhedral surfaces tend to infinity, although the limit surface (in the Hausdorff metric) has finite surface area. This surprised the contemporary mathematical community, since this limit definition was standardly used, based on a flawed analogy with curve length.

Since two different areas of mathematics share the same counterexample, it is natural to ask whether there is some connection. Unsurprisingly, this is the case. The purpose of this section is to point out this connection and that this opens the door to obtain a different insight into the convergence of FEM via differential geometry and measure theory.

Definition 4. *Let $\Omega \subset \mathbb{R}^2$ and $v : \Omega \to \mathbb{R}$. Then the* graph *of v is defined as*

$$\operatorname{graph}(v) = \{(x, y, z) \in \mathbb{R}^3 : \ z = v(x, y) \text{ for } (x, y) \in \Omega\}.$$

In the following theorem, we show that FEM convergence implies convergence of surface areas of the corresponding graphs. By $A(v)$, we denote the surface area of $\operatorname{graph}(v)$ of a function $v : \Omega \to \mathbb{R}$, if it is well defined.

Theorem 3. *Let $u \in H^2(\Omega)$ and $u_h \in V_h$ for $h \in (0, h_0)$. If $u_h \to u$ in $H^1(\Omega)$ as $h \to 0$, then for a subsequence,* $\operatorname{graph}(u_{h_n}) \to \operatorname{graph}(u)$ *in the Hausdorff metric and $A(u_{h_n}) \to A(u)$.*

Proof. Since $H^2(\Omega), V_h \subset C(\bar{\Omega})$, then $u, u_h \in C(\bar{\Omega})$. Since $u_h \to u$ in $H^1(\Omega)$, also $u_h \to u$ in $L^2(\Omega)$. Therefore there exists a subsequence u_{h_n} converging to u pointwise almost everywhere in Ω. Since $u, u_{h_n} \subset C(\bar{\Omega})$, we have $u_{h_n} \to u$ uniformly in Ω. By the definition of the Hausdorff metric and uniform convergence, we immediately have $\operatorname{graph}(u_{h_n}) \to \operatorname{graph}(u)$ in the Hausdorff metric.

It remains to prove the convergence of surface areas. For a function $v \in H^1(\Omega)$, the area of $\operatorname{graph}(v)$ is given by

$$A(v) = \int_\Omega \sqrt{1 + \left(\tfrac{\partial v}{\partial x}\right)^2 + \left(\tfrac{\partial v}{\partial y}\right)^2} \mathrm{d}(x, y).$$

Therefore,

$$\big|A(u) - A(v)\big| \leq \int_\Omega \Big|\sqrt{1 + \left(\tfrac{\partial u}{\partial x}\right)^2 + \left(\tfrac{\partial u}{\partial y}\right)^2} - \sqrt{1 + \left(\tfrac{\partial v}{\partial x}\right)^2 + \left(\tfrac{\partial v}{\partial y}\right)^2}\Big| \mathrm{d}(x, y). \quad (6)$$

Using the easily verifiable inequality $|\sqrt{1+a^2+b^2} - \sqrt{1+c^2+d^2}| \leq |a-c| + |b-d|$ for all $a, b, c, d \in \mathbb{R}$, we obtain

$$\big|A(u) - A(v)\big| \leq \int_\Omega \Big(\big|\tfrac{\partial u}{\partial x} - \tfrac{\partial v}{\partial x}\big| + \big|\tfrac{\partial u}{\partial y} - \tfrac{\partial v}{\partial y}\big|\Big) \mathrm{d}(x, y) \leq \sqrt{2}|\Omega|^{1/2}|u - v|_{H^1(\Omega)} \quad (7)$$

by Hölder's inequality. Therefore, $u_{h_n} \to u$ in $H^1(\Omega)$ implies $A(u_{h_n}) \to A(u)$. □

Theorem 3 thus establishes a simple connection to the theory of approximation of surface area. Similar questions have been dealt with in the past decade in the field of discrete differential geometry. The question is, how do classical differential-geometric objects defined on polygonal (polyhedral) surfaces converge to those of the limit surface. In the case of Theorem 3, we are interested in results of the following type:

Theorem 4 ([3])**.** *If a sequence of polyhedral surfaces $\{M_n\}$ converges to a smooth surface M in the Hausdorff metric, then the following conditions are equivalent:*

convergence of area*,*

convergence of normal fields*,*

convergence of metric tensors*,*

convergence of Laplace-Beltrami operators*.*

Here convergence is always meant in the L^∞-sense for the corresponding term.

Theorems 3 and 4 yield a potential strategy for proving necessary conditions for FEM convergence. In the context of discrete differential geometry, the problem can be attacked from other points of view using different techniques than usual in the FEM community. We note that in [3], the theory used to prove Theorem 4 is used also to investigate convergence of other objects such as geodesics and mean curvature vectors. Unfortunately, it seems that there are no suitable more general results directly applicable to the convergence of FEM in existing discrete differential literature, although the analogy of the minimum angle condition is known in the community.

Acknowledgements

The research is supported by the Grant No. P201/13/00522S of the Czech Science Foundation. V. Kučera is a junior researcher at the University Center for Mathematical Modelling, Applied Analysis and Computational Mathematics (Math MAC).

References

[1] Babuška, I. and Aziz, A. K.: On the angle condition in the finite element method. SIAM J. Numer. Anal. **13 (2)** (1976), 214–226.

[2] Hannukainen, A., Korotov, S., and Křížek, M.: The maximum angle condition is not necessary for convergence of the finite element method. Numer. Math. **120** (2012), 79–88.

[3] Hildebrandt, K., Polthier, K., and Wardetzky, M.: On the convergence of metric and geometric properties of polyhedral surfaces. Geom. Dedicata **123** (2006), 89–112.

[4] Oswald, P.: Divergence of the FEM: Babuška-Aziz revisited. Appl. Math. **60 (5)** (2015).

[5] Schwarz, H. A.: Sur une définition erronée de l'aire d'une surface courbe. Ges. Math. Abhandl. II (1890), Springer-Verlag, 309–311.

Conference Applications of Mathematics 2015, in honor of the birthday anniversaries of Ivo Babuška (90), Milan Práger (85), and Emil Vitásek (85)
J. Brandts, S. Korotov, M. Křížek, K. Segeth, J. Šístek, T. Vejchodský (Eds.)
Institute of Mathematics CAS, Prague 2015

CONVERGENCE AND STABILITY OF HIGHER-ORDER FINITE ELEMENT SOLUTION OF REACTION-DIFFUSION EQUATION WITH TURING INSTABILITY

Pavel Kůs

Institute of Mathematics, Czech Academy of Sciences
Žitná 25, 1115 67 Praha 1, Czech Republic
kus@math.cas.cz

Abstract: In this contribution, higher-order finite element method is used for the solution of reaction-diffusion equation with Turing instability. Some aspects concerning convergence of the method for this particular problem are discussed. Our numerical tests confirm the convergence of the method, but for some very special choices of parameters, this convergence has very uncommon properties.

Keywords: convergence, finite element method, Turing instability, reaction-diffusion

MSC: 65N30, 35K57

1. Introduction

In this contribution we investigate convergence of higher-order finite element method for a reaction-diffusion problem exhibiting the Turing instability. The motivation of this work is to investigate convergence properties. There is not enough theoretical results regarding convergence theory for this particular application, so we try to observe some properties by performing numerical tests. We are interested in steady-state solutions only, which makes numerical experiments rather time demanding, since a lot of time steps have to be done before the steady-state solution is reached. It is known, that different initial conditions might lead to different steady states. Our interest is to investigate how the choice of finite element mesh and polynomial order influences the resulting steady state. In other words, we are interested in stability of the calculation with respect to choice of finite element approximation.

We are aware of the fact, that the selected method is not the most efficient for the given geometry and equation. There are different methods, which are able to exploit the square geometry such as FFT-based approach (used in [6]) or multigrid method, which is used to solve a similar problem in [4]. Our main interest is, however,

in testing the performance of higher-order FEM and we use this problem as a test example with certain unpleasant properties.

2. Turing instability

Reaction-diffusion equations are studied in various contexts. Our motivation is the study of systems exhibiting the Turing instability. In many applications, the equations describe an interaction of activator u and inhibitor v, see [5] for more motivation and explanations and [2] for more analysis and interesting applications. The investigated problem is the following:

$$\begin{aligned} \frac{\partial u}{\partial t} &= D\delta\triangle u + \alpha u + v - r_2 uv - \alpha r_3 uv^2, \\ \frac{\partial v}{\partial t} &= \delta\triangle v + \gamma u + \beta v + r_2 uv + \alpha r_3 uv^2. \end{aligned} \tag{1}$$

We will consider square domain $\Omega = [0, 200]^2$ and homogeneous Neumann boundary conditions for both u and v. As in the work [5], we will use the following coefficients, which are selected in such a way, that the Turing ("diffusion driven") instability leads to formation of patterns in the steady-state solution:

$$\alpha = \gamma = 0.899, \qquad \beta = -0.91, \qquad D = 0.45, \qquad r_2 = 2, \qquad r_3 = 3.5. \tag{2}$$

Parameters are fine-tuned in such a way, that Turing patterns develop, as can be seen in Fig. 1.

We will consider different values of scaling parameter δ, which can be viewed either as a ratio between the strength of diffusion and reaction or as a measure of the domain size. This choice will affect the appearance of the steady-state solution, as can be seen in Fig. 2.

Figure 1: Left: initial condition used for all following calculations. Middle and right: two intermediate time steps for the value $\delta = 24$. The corresponding steady-state solution is in Fig. 2 in the middle.

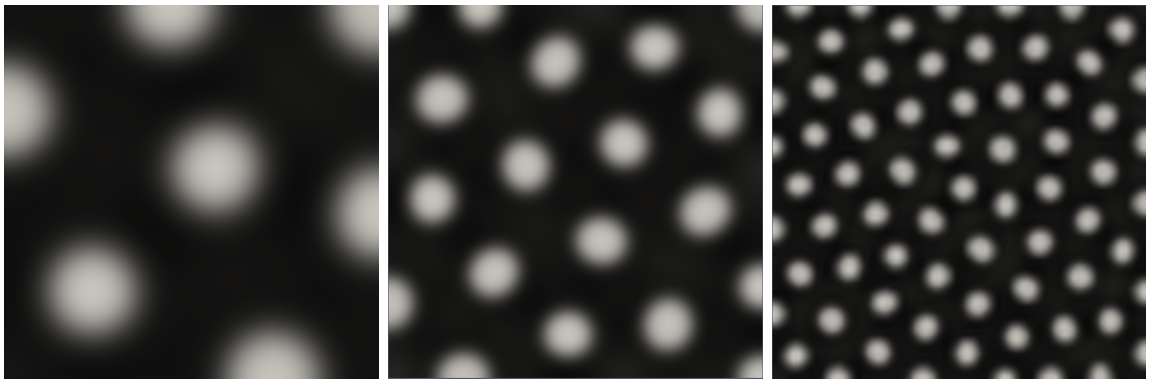


Figure 2: Steady state solutions for the value $\delta = 96$, 24 and 6, respectively.

3. Discretization

In this contribution, we will not explore more sophisticated time discretization schemes of higher order or adaptive choice of the time step length (as it is done in, e.g., [3]). We will use standard Crank-Nicolson method of second order with time levels t_n and fixed time step $dt = t_{n+1} - t_n$. Moreover, nonlinear terms of (1) will be treated explicitly. After the time semi-discretization, we obtain

$$\begin{aligned} \frac{u^{n+1} - u^n}{dt} &= \frac{1}{2}(D\delta\triangle u^{n+1} + \alpha u^{n+1} + v^{n+1} + D\delta\triangle u^n + \alpha u^n + v^n) \\ &\quad -r_2 u^n v^n - \alpha r_3 u^n (v^n)^2, \\ \frac{v^{n+1} - v^n}{dt} &= \frac{1}{2}(\delta\triangle v^{n+1} + \gamma u^{n+1} + \beta v^{n+1} + \delta\triangle v^n + \gamma u^n + \beta v^n) \\ &\quad +r_2 u^n v^n + \alpha r_3 u^n (v^n)^2, \end{aligned} \tag{3}$$

where u^n and v^n are solutions at time t_n. The space discretization of the resulting system is done in the usual finite element way. In this contribution we use only structured meshes with square elements with the same size and order in one mesh. We use different element sizes for different meshes and polynomial orders 1 to 4.

We are interested only in resulting steady-state solutions. We consider a solution to be steady-state, when the relative norm of difference of solutions in two consecutive time steps is below a prescribed tolerance 10^{-12}. As we have already stated, for a considered setting with fixed parameters, multiple steady state solutions can be found. The trivial solution $u = v = 0$ is always present. Apart from that, several nontrivial solutions can be found, depending on selected initial condition. These solutions might be qualitatively similar (exhibit the same pattern), but, in the sense of a mathematical function, they are completely different. For the considered setting, most initial conditions lead to steady-state solutions containing dots with similar sizes distributed in similar distances. The exact position and even amount of dots may,

however, be completely different (even when symmetries are taken into account), see, e.g., [6]. For the rest of this contribution, we will use the following initial condition:

$$u_I = \left(1 - \frac{|x - 125|}{25}\right)\left(1 - \frac{|y - 125|}{25}\right) \quad \text{for } (x, y) \in [100, 150]^2, \tag{4}$$

$u_I = 0$ otherwise. It is a piece-wise bilinear "hat", as can be seen in the left panel of Fig. 1. This function is contained in all finite element spaces used in this contribution, so there are no errors caused by inexact projection of the initial condition.

Practical implementation of the problem has been done using the deal.II library (see, e.g., [1]), which simplifies the use of higher-order basis functions. The temporal discretization has been done by hand inside the weak formulation (Rothe's method). A sparse direct linear solver has been used. Since the nonlinear part is discretized explicitly, calculations in all time steps have the same matrix, which can be factorized at the beginning. Thus, in each time step, only back substitution has to be performed. Even though, the calculations are very time-demanding, since many time steps have to be performed to obtain steady-state solution. We use constant time step $dt = 0.5$ and run the calculation until the relative norm of the difference of two consecutive solutions is below a prescribed tolerance 10^{-12}. The number of iterations depends on mesh, element order and parameter setting, but might exceed 50 000. Higher-order temporal discretization scheme and adaptive choice of time step could improve the situation, but it is not considered in this study. As can be seen from Fig. 3, steady state solutions for given value of δ converge as element size decreases and the number of degrees of freedom increases. This is the case for most (but not all) values of δ, as will be discussed later in the text.

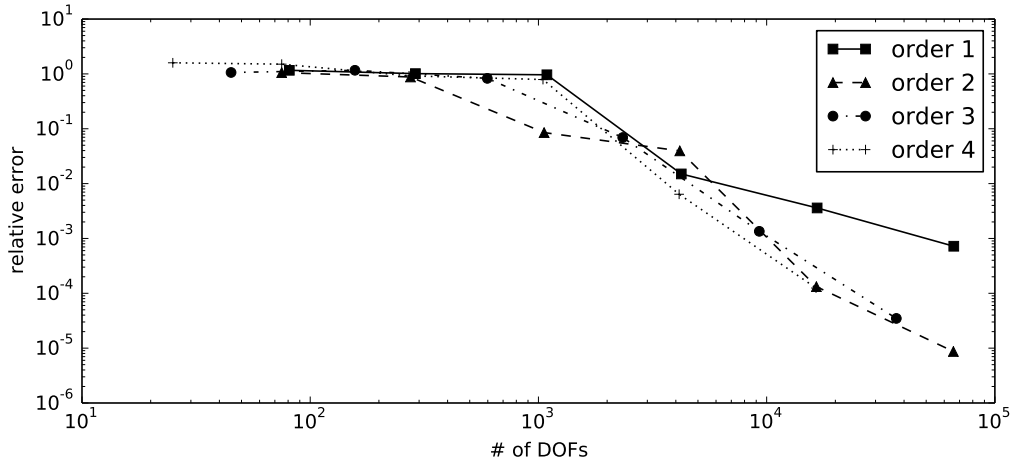


Figure 3: Convergence of FEM solutions for meshes with decreasing element sizes (and thus increasing number of degrees of freedom). The relative norm of difference from the solution on the finest mesh is shown. Results for the value $\delta = 24$.

4. Dependence on parameter δ

Interesting results can be obtained, when the value of the parameter δ is changed with small step. We performed series of calculations with gradually increasing δ, first and last such obtained steady-state solutions are depicted in left and right panel of Fig. 2. Most of the time, continuous dependence of the steady state solution on δ can be observed. Roughly speaking, we can observe that individual dots are getting smaller and their position is changing slowly. At some points, however, there is a bifurcation and a completely different solution is obtained with different number of dots in completely different positions. Many such bifurcations can be seen in Fig. 4, where each peak corresponds to big difference between two consecutive solutions with only slightly different value of δ. A detailed example of one such situation is given in Fig. 5. We remind that the same initial condition (4) is used for all calculations.

Let us now focus more closely on behavior of the solution close to the bifurcation points, where one solution changes to another one. We would like to investigate how this transition occurs, especially with respect to different meshes or polynomial orders. For simplicity, we will (in this contribution) focus only on transition between solutions shown in Fig. 5. For a given mesh, we use interval halving method to estimate the value of the bifurcation point. We do the following series of calculation. At the beginning, we set $\delta_B^1 = 78$, $\delta_B^2 = 79$. Than, at each step, we set $\delta_B = (\delta_B^1 + \delta_B^2)/2$ and find the corresponding steady state solution u_{δ_B}. Then we compare norms of differences of solutions and set $\delta_B^1 := \delta_B$ if $||u_{\delta_B^1} - u_{\delta_B}|| < ||u_{\delta_B^2} - u_{\delta_B}||$, or $\delta_B^2 := \delta_B$ otherwise. This process is repeated until $\delta_B^2 - \delta_B^1$ is sufficiently small (10^{-6} in our case). Throughout the whole process, solutions $u_{\delta_B^1}$ and $u_{\delta_B^2}$ are very different, since there is sharp jump between different types of solutions. Originally we thought, that

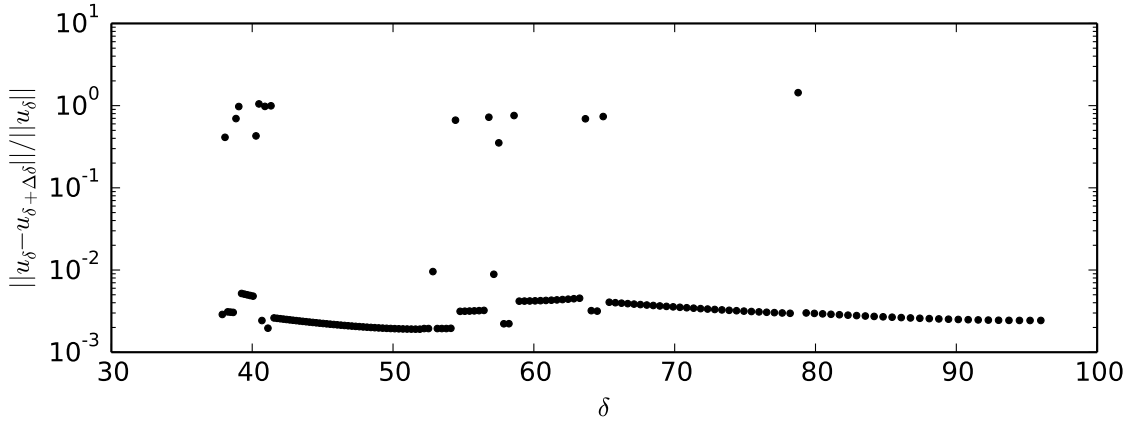


Figure 4: Relative difference of consecutive solutions in a series of calculations with increasing δ. The step $\Delta\delta$ by which δ is increased is approximately $\Delta\delta/\delta = 0.006$. The right-most peak corresponds to the situation from Fig. 5.

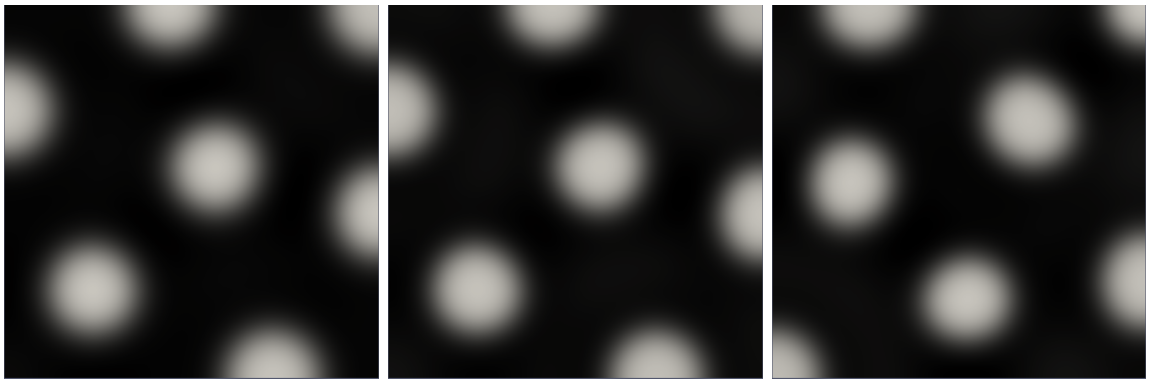


Figure 5: Steady states for value $\delta = 96$, 78.76 and 78.2, respectively. Even though the difference in parameter is much bigger between first and second value, corresponding solutions are hard to distinguish. The solution for slightly reduced δ is very different (right). This jump corresponds to the right-most peak in Fig. 4.

Figure 6: Steady states obtained on some meshes for $\delta = \delta_B$. These solutions form transition between two types of solutions observed in Fig. 5 and were not observed as peaks in Fig. 4, since the step of change of parameter δ was too large.

this value is the point of transition between solutions from Fig. 5. It turned out, however, that there are (at least) two more solutions obtained numerically (shown in Fig. 6). Values of δ_B obtained by this algorithm for a sequence of consecutively refined meshes can be seen in Fig. 7.

5. Convergence for fixed δ near bifurcation

In the previous section we described dependence of the solution on δ and discussed its behavior close to a bifurcation point. We have seen that this behavior depends on used FEM mesh and polynomial order. This brings us to a natural question, which is the main interest of this contribution. How will this behavior affect convergence of the method close to the bifurcation point δ_B?

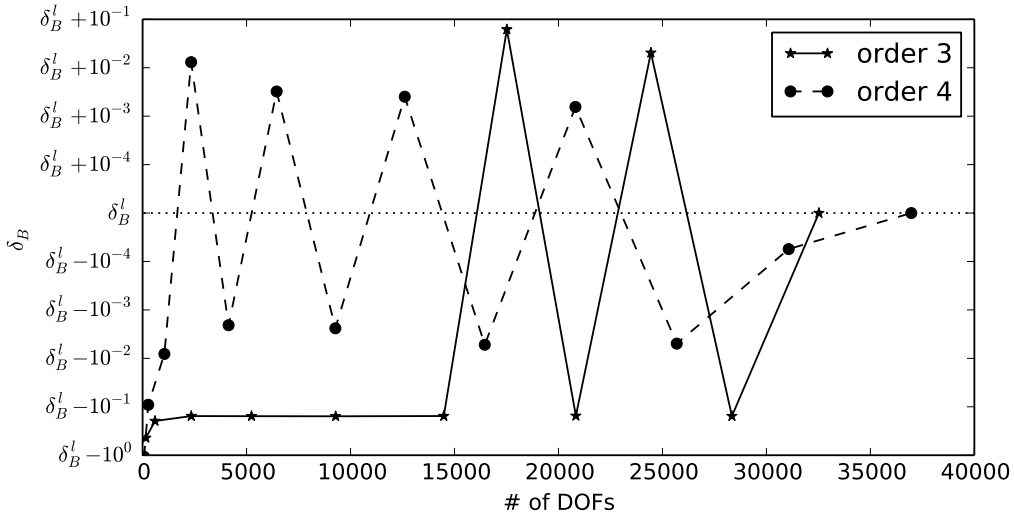


Figure 7: Bifurcation value δ_B for a sequence of consecutively refined meshes. Differences from the "limit" value δ_B^l obtained on the finest mesh are shown. Error in δ_B is less then 10^{-6}, which is sufficiently small compared to differences between values of δ_B for different meshes.

We performed series of calculations with fixed value $\delta = 78.500126$ (which is the "limit" value δ_B^l found by interval halving for finite elements of order 4, see Fig. 7) and with variable number of elements in the mesh. First of all, we found 2 more types of steady state solutions (shown in Fig. 6), distinguished from data presented in Fig. 4. Those new solutions form transition between solutions shown in Fig. 5, which are more "stable" in a sense that they are obtained for δ from a relatively large interval on all used meshes. This is not the case of "transition" solutions from Fig. 6. Another irregularity is the loss of symmetry with respect to line $y = x$, which is present in the initial condition and in all previously observed solutions. These two solutions, however, differ by symmetry with respect to the same axis. The question thus rises whether they really are solutions of the continuous system, or whether they are artificially created by discretization and roundoff errors.

The convergence process can be seen in Table 1. We can observe oscillations rather than smooth convergence. Each of the approximate solutions obtained on meshes with decreasing element size is approaching one of the four "types" of solutions, which are shown in Figs. 5 and 6. We denote them A, B, C and D, respectively. We do not claim that this means that the method does not converge as $h \to 0$. It may happen, however, that for h in the range given by capabilities of our computer, we are not able to determine, which of the completely different solutions that we are obtaining for different values of h is close to the exact solution for the given δ.

n :	16	20	24	28	32	36	40	44	48	52	56	60	64	68	72	76	80
order 3	A	B	A	B	A	B	C	B	C	B	C	B	C	B	D	B	C
order 4	A	C	D	C	C	D	D	D	C	C	D	C	D	D			

Table 1: Types of solutions, obtained by calculation with $\delta = 78.500126$. As A and B we denote solutions from Fig. 5, as C and D solutions from Fig. 6, respectively. Calculations for polynomial orders 3 and 4 are performed on meshes with n elements in each direction, n being the number in the first row. The mesh then consists of n^2 square elements of size $h = 200/n$.

6. Summary

We investigated a particular numerical scheme for the solution of reaction-diffusion problems. We have shown that the method usually converges and that it behaves according to expectations. However, for the parameter value close to a bifurcation point, we observe an oscillating sequence of approximate solutions. Moreover it is possible, that some of the approximate solutions, which are obtained for some meshes, are not approaching any solution of the continuous problem. Even though this behavior can be observed only for carefully selected parameters, we find it interesting, since it opposes the usual idea of convergence of the finite element method.

References

[1] Bangerth, W., Hartmann, R., and Kanschat, G.: deal.II – a general purpose object oriented finite element library. ACM Trans. Math. Softw. **33** (2007), 24/1–24/27.

[2] Barrio, R. A., Varea, C., Aragn, J. L., and Maini, P. K.: A two-dimensional numerical study of spatial pattern formation in interacting turing systems. Bulletin of Mathematical Biology **61** (1999), 483–505.

[3] Dolejší, V. and Kůs, P.: Adaptive backward difference formula–discontinuous Galerkin finite element method for the solution of conservation laws. Int. J. Numer. Meth. Engng. **73** (2008), 1739–1766.

[4] Landsberg, C. and Voigt, A.: A multigrid finite element method for reaction-diffusion systems on surfaces. Comput. Visual Sci. **13** (2010), 177–185.

[5] Liu, R. T., Liaw, S. S., and Maini, P. K.: Two-stage Turing model for generating pigment patterns on the leopard and the jaguar. Phys. Rev. E **74** (2006), 011 914:1–011 914:8.

[6] Rybář, V. and Vejchodský, T.: Variability of Turing patterns in reaction-diffusion systems. In: H. Bílková, M. Rozložník, and P. Tichý (Eds.), *Proceedings of the SNA'14*, pp. 87–90, 2014.

Conference Applications of Mathematics 2015, in honor of the birthday anniversaries of Ivo Babuška (90), Milan Práger (85), and Emil Vitásek (85)
J. Brandts, S. Korotov, M. Křížek, K. Segeth, J. Šístek, T. Vejchodský (Eds.)
Institute of Mathematics CAS, Prague 2015

USE OF A DIFFERENTIAL EVOLUTION ALGORITHM FOR THE OPTIMIZATION OF THE HEAT RADIATION INTENSITY

Jaroslav Mlýnek[1], Roman Knobloch[1], Radek Srb[2]

[1] Department of Mathematics, FP
jaroslav.mlynek@tul.cz, roman.knobloch@tul.cz
[2] Institute of Mechatronics and Computer Engineering
radek.srb@tul.cz

Technical University of Liberec, Studentská 2, Liberec, Czech Republic

Abstract: This article focuses on the heat radiation intensity optimization on the surface of an aluminium shell mould. The outer mould surface is heated by infrared heaters located above the mould and the inner mould surface is sprinkled with a special PVC powder. This is an economic way of producing artificial leathers in the automotive industry (e.g. the artificial leather on car dashboards). The article includes a description of a mathematical model that allows us to calculate the heat radiation intensity across the outer mould surface for every fixed location of the heaters. We also use this mathematical model for optimizing the locations of the heaters to generate uniform heat radiation intensity on the whole outer mould surface during the heating of the mould. In this way we obtain an even colour shade and material structure of the artificial leather. The problem of optimization is more complicated. Using gradient methods is not suitable because the minimized deviation function contains many local minima. A differential evolution algorithm is used during the process of optimization. The calculations were performed by a Matlab code written by the authors. The article contains a practical example including graphical outputs.

Keywords: heat radiation intensity, evolution optimization algorithm, mathematical model, experimental measurement, software implementation

MSC: 65K10, 78M50

1. Introduction

This article describes the calculation of radiation intensity on the whole mould surface for the fixed locations of infrared heaters above the mould and the process

of heat radiation intensity optimization on the mould surface. The problem of optimization is rather complex (the used moulds often have very complicated surfaces, during the process of optimization possible collisions between one heater and another as well as collisions between a heater and the mould surface must be avoided). The minimized deviation function has many local minima. Using gradient methods for finding the global minimum is therefore unsuitable. Thus, we used an evolution optimization algorithm. A differential evolution algorithm *DE/rand/1/bin* (see details in [6]) is used to find suitable locations of the heaters over the mould to optimize the heat radiation intensity on the whole outer mould surface. The manufacturer needs to implement the optimization procedure on the production line (after its verification in the Matlab system). Therefore, we need to know the optimization process in every detail and to be able to perform own modifications of the programmed optimization algorithm. We do not use existing commercially available software tools.

In practice, an aluminium mould is heated by a set of infrared heaters located above the outer mould surface. It is necessary to ensure the same heat radiation intensity (within a given tolerance) on the whole outer mould surface by finding a suitable locations for the heaters. In this way the same colour and material structure of the artificial leather are assured. Moulds which very often have complicated shapes and which weigh from 100 to 300 kilograms are used. The infrared heaters have a tubular form and their length is about 20 centimeters. Every heater is equipped with a mirror located above the radiation tube which reflects heat radiation in a set direction (see Figure 1).

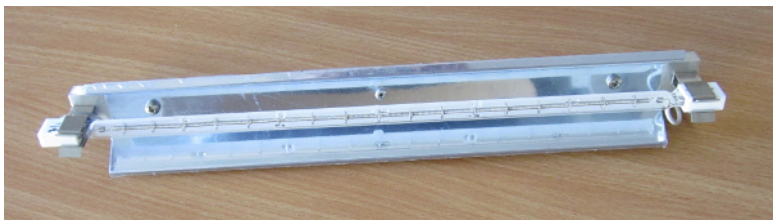


Figure 1: Infrared heater Ushio with heating power 2000 W.

2. Mathematical model of the heat radiation

In this chapter a mathematical model of the heat radiation produced by the infrared heaters on the outer mould surface is described. The heaters and the heated mould are represented in 3-dimensional Euclidean space E_3 using the Cartesian coordinate system (O, x_1, x_2, x_3) with basis vectors $e_1 = (1, 0, 0), e_2 = (0, 1, 0)$ and $e_3 = (0, 0, 1)$.

2.1. Representation of the heater

A heater is represented by a straight line segment of length d (see Figure 2). The location and orientation of a heater is defined by the following parameters:

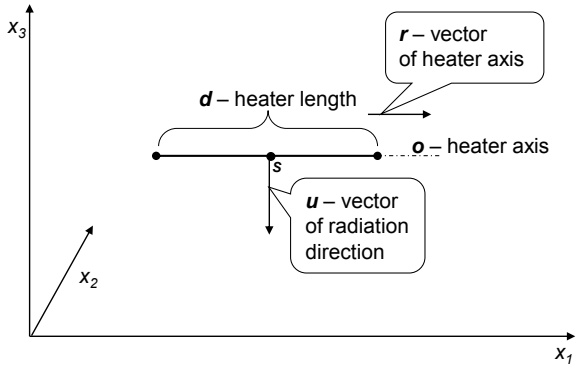


Figure 2: Schematic representation of the infrared heater.

(i) the coordinates of the heater centre $S = [s_1, s_2, s_3]$, (ii) the unit vector $u = (u_1, u_2, u_3)$ of the heat radiation direction, where component $u_3 < 0$ (i.e., the heater radiates "downward"), (iii) the vector of the heater axis $r = (r_1, r_2, r_3)$. Another way to determine the vector r is by using angle φ between the vertical projection of vector r onto the x_1x_2-plane and the positive part of axis x_1 (the vectors u and r are orthogonal, $0 \leq \varphi < \pi$). The location of each heater Z can be defined by the following 6 parameters

$$Z : (s_1, s_2, s_3, u_1, u_2, \varphi). \tag{1}$$

2.2. Representation of the mould

The outer mould surface P is described by elementary surfaces p_j, where $1 \leq j \leq N$. It holds that $P = \cup p_j$, where $1 \leq j \leq N$ and $int\ p_i \cap int\ p_j = \emptyset$ for $i \neq j$, $1 \leq i, j \leq N$. Each elementary surface p_j is described by the following parameters: (i) its centre of gravity $T_j = [t_1^j, t_2^j, t_3^j]$, (ii) the unit outer normal vector $v_j = (v_1^j, v_2^j, v_3^j)$ at the point T_j (we suppose v_j points "upwards"and therefore is defined through the first two components v_1^j and v_2^j), (iii) the area w_j of the elementary surface. Every elementary surface p_j thus can be defined by the following 6 parameters

$$p_j : (t_1^j, t_2^j, t_3^j, v_1^j, v_2^j, w_j). \tag{2}$$

2.3. Experimental measurement of the heater radiation intensity

We need to know the heat radiation intensity in the heater surroundings to calculate the total radiation intensity on the outer mould surface. The heater manufacturer does not provide the distribution function of the heat radiation intensity in the heater surroundings. We set up the experimental measurement of the heat radiation intensity as follows. The location of the heater is $Z : (0, 0, 0, 0, 0, 0)$ in accordance with relation (1), i.e., the centre S of the heater lies at the origin of the Cartesian coordinate system (O, x_1, x_2, x_3); the unit radiation vector has coordinates

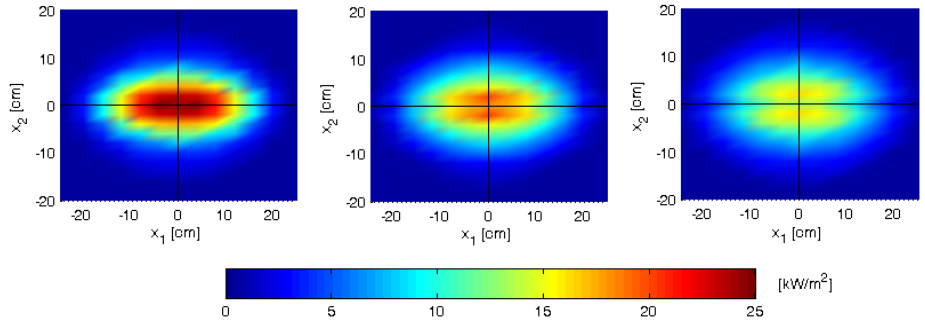


Figure 3: Heat radiation intensity in the planes at distances 9, 11 and 13 cm from the heater.

$u = (0, 0, -1)$ and the vector of the heater axis has coordinates $r = (1, 0, 0)$. We assume the heat radiation intensity across the elementary surface p_j is the same as at the centre of gravity T_j. The heat radiation intensity at T_j depends on the position of this point (determined by the first three parameter in the elementary surface p_j given by (2)) and on the direction of the outer normal vector v_j at point T_j (determined by the fourth and fifth parameters in the elementary surface p_j given by (2)). The heat radiation intensity I in the surroundings and below the heater was experimentally measured by a sensor at selected points $a = [a_1, a_2, a_3, a_4, a_5]$ (the first three parameters a_1, a_2, a_3 describe the position of the centre of gravity of a fictitious elementary surface and the fourth and fifth parameter describes the direction of the outer normal vector at the point $[a_1, a_2, a_3]$).

We use measured values $I(a)$ of heat radiation intensity at the selected points a and the linear interpolation function of five variables to calculate the heat radiation intensity $I(b)$ for the general point $b = [b_1, b_2, b_3, b_4, b_5]$ in the heater surroundings.

The measured heat radiation intensity, and its interpolated values in three parallel planes with x_1x_2-plane are shown in colour in Figure 3 in the case of 0^o deflection of the axis of the sensor (i.e., axis of the sensor is vertical). We use linear interpolation of a function of five variables. We assume that the point b holds $a_{j,i_j} \leq x_j^b \leq a_{j,i_j+1}$ for $1 \leq j \leq 5$. Let us denote $m_j = \frac{x_j^b - x_{j,i_j}}{a_{j,i_j+1} - a_{j,i_j}}$ for $1 \leq j \leq 5$. Then it holds for the interpolation value of radiation intensity $I(b)$ at the point b of heater Z

$$I(b) = I\left(x_1^b, x_2^b, x_3^b, x_4^b, x_5^b\right) = \sum_{k_1=i_1}^{i_1+1} \dots \sum_{k_5=i_5}^{i_5+1} I\left(a_{1,k_1}, a_{2,k_2}, a_{3,k_3}, a_{4,k_4}, a_{5,k_5}\right) \cdot \qquad (3)$$

$$\cdot \prod_{l=1}^{5} H\left(l, k_l - i_l\right).$$

The interpolation formula is described in detail in [1], p. 148, and [3].

2.4. General case of the heater location

In this subsection we explain a transformation of the general case of a heater location with reference to the special heater position solved in Subsection 2.3. For a heater in a general position, we briefly describe the transformation of the previous Cartesian coordinate system (O, e_1, e_2, e_3) into a positively oriented Cartesian system $(S, r, n, -u)$, where S is the centre of the heater, r is the heater axis vector, and u is the direction vector of the heat radiation. The vector n is determined by the vector product of the vectors $-u$ and r (see more detail in [2], [7]) and is defined by the following relation

$$n = (-u) \times r = \left(- \begin{vmatrix} u_2 & u_3 \\ r_2 & r_3 \end{vmatrix}, \begin{vmatrix} u_1 & u_3 \\ r_1 & r_3 \end{vmatrix}, - \begin{vmatrix} u_1 & u_2 \\ r_1 & r_2 \end{vmatrix} \right).$$

The vectors r, u and n are normalized to give the unit length. Then we can define an orthonormal transformation matrix

$$\mathbf{A} = \begin{pmatrix} r_1 & n_1 & -u_1 \\ r_2 & n_2 & -u_2 \\ r_3 & n_3 & -u_3 \end{pmatrix}.$$

Recall that for the elementary surface p_j, the respective triples T_j and v_j represent its centre of gravity and its outer normal vector in the Cartesian coordinate system (O, e_1, e_2, e_3). If S is the triple of parameters representing (in (O, e_1, e_2, e_3)) the centre of the heater that determines the coordinate system $(S, r, n, -u)$, then T_j and v_j are transformed as follows

$$\left(T_j^{'}\right)^T = \mathbf{A}^T \left(T_j - S\right)^T \quad \text{and} \quad \left(v_j^{'}\right)^T = \mathbf{A}^T v_j^T, \tag{4}$$

where $T_j^{'}$ and $v_j^{'}$ are the coordinates in $(S, r, n, -u)$. In this way, we transform the general case of the heater location to the measured case and we can calculate heat radiation intensity by using linear interpolation as described in the previous subsection (transformed point $T_j^{'}$ and vector $v_j^{'}$ correspond to point b in Subsection 2.3).

2.5. Calculation of the total heat radiation intensity

Now we describe the numerical computation procedure for the total heat radiation intensity on the mould surface. We denote by L_j the set of all heaters radiating on the jth elementary surface p_j $(1 \leq j \leq N)$ for the fixed locations of the heaters, and I_{jl} the heat radiation intensity of the lth heater on the p_j elementary surface. Then the total radiation intensity I_j on the elementary surface p_j is given by the following relation

$$I_j = \sum_{l \in L_j} I_{jl} \ . \tag{5}$$

The producer of artificial leathers recommends a constant value of the heat radiation intensity on the whole outer mould surface. Let us denote this constant value as I_{rec}. We can define function F, the deviation of the heat radiation intensity, by the relation

$$F = \frac{\sum_{j=1}^{N} |I_j - I_{rec}| w_j}{W} \tag{6}$$

and the deviation $\tilde{F}$ by the relation

$$\tilde{F} = \sqrt{\frac{1}{W} \cdot \sum_{j=1}^{N} (I_j - I_{rec})^2 w_j} \ , \tag{7}$$

where $W = \sum_{j=1}^{N} w_j$ and we highlight that w_j denotes the area of the elementary surface p_j. We need to find the locations of the heaters so that the value of deviation F (alternatively deviation $\tilde{F}$) is as small as possible.

3. Optimization of the heaters locations

Functions F and $\tilde{F}$ defined by (6) and (7) contain many local minima. Using gradient methods for finding global minima of the functions F and $\tilde{F}$ is not appropriate. If we used a gradient method, there would be a high probability that we would find only a local minimum of the function. Therefore, we use a differential evolution algorithm (more details in [6]) for finding an optimized minimum of function F (i.e., to optimize the locations of the heaters). The disadvantage of a differential evolution algorithm is its computational demandingness and slow convergence. The location of every heater is defined in accordance with the relation (1) by 6 parameters. Therefore $6M$ parameters are necessary to define the locations of all M heaters. One individual in the differential evolution algorithm represents one possible location of all $6M$ heaters. In the algorithm we successively construct populations of individuals. Every population includes NP individuals where every individual is a potential solution to our problem. We seek the individual $y_{min} \in C$ satisfying the condition

$$F(y_{min}) = min\{F(y); y \in C\}, \tag{8}$$

where $C \subset E_{6M}$ is the set we are searching for. Every element of C is formed by a set of $6M$ allowable parameters and this set defines just one constellation of the heaters above the mould. The identification of the individual y_{min} defined by (8) is not realistic in practice. But we are able to determine an optimized solution y_{opt}. The generated individuals are saved in the matrix $\mathbf{B}_{NP\times(6M+1)}$. Every row of this matrix represents one individual, y, and its evaluation, $F(y)$.

3.1. Differential evolution algorithm

Now we describe schematically the particular steps of the differential evolution algorithm named *DE/rand/1/bin* (for more details see [6] and [8]) which is applied to our problem.

We define a specimen which contains values ranges of each gene of the individual in the first step of the algorithm. Then we define an initial individual y_1 and randomly generate the initial generation of individuals. We create successively generations of individuals y and we are looking for an individual with the smallest value $F(y)$ (where function F is given by relation (6)) in the following steps of the algorithm. Note that four individuals y of a generation participate in the creation of individual y of the following generation. We describe the diagram of the algorithm.

Input: the initial individual y_1, population size NP, the number of used heaters M (dimension of the problem is $6M$), crossover probability CR, mutation factor f, the specified accuracy of the calculation ε.
Internal computation:
1. create an initial generation ($G = 0$) of NP individuals $y_i^G, 1 \leq i \leq NP$,
2.a) evaluate all the individuals y_i^G of the generation G (calculate $F(y_i^G)$ for every individual y_i^G), b) store the individuals y_i^G and their evaluations $F(y_i^G)$ into the matrix $\mathbf{B}$,
3. *repeat until* $min\{F(y_i^G); y_i^G \in \mathbf{B}\} < \varepsilon$
a) *for* $i := 1$ *step* 1 *to* NP *do*
(i) randomly select index $k_i \in \{1, 2, \ldots, 6M\}$,
(ii) randomly select indexes $r_1, r_2, r_3 \in \{1, 2, \ldots, NP\}$, where $r_t \neq i$ for $1 \leq t \leq 3$ and $r_1 \neq r_2, r_1 \neq r_3, r_2 \neq r_3$;
(iii) *for* $j := 1$ *step* 1 *to* $6M$ *do*
if $(rand(0,1) \leq CR$ *or* $j = k_i)$ *then*
$y_{i,j}^{trial} := y_{r_3,j}^G + f\left(y_{r_1,j}^G - y_{r_2,j}^G\right)$
else
$y_{i,j}^{trial} := y_{i,j}^G$
end if
end for (j)
(iv) *if* $F\left(y_i^{trial}\right) \leq F\left(y_i^G\right)$ *then* $y_i^{G+1} := y_i^{trial}$
else
$y_i^{G+1} := y_i^G$
end for(i),
b) store individuals y_i^{G+1} and their evaluations $F\left(y_i^{G+1}\right)$ $(1 \leq i \leq NP)$ of new generation $G + 1$ into the matrix $\mathbf{B}$, $G := G + 1$
end repeat.
Output:
the row of matrix $\mathbf{B}$ that contains corresponding value $min\{F(y_i^G); y_i^G \in \mathbf{B}\}$ represents the best found individual y_{opt}.

Note that function $rand(0,1)$ randomly chooses a number from the interval $\langle 0, 1\rangle$. The notation $y_{i,j}^G$ means the jth component of an individual y_i^G in Gth generation. The individual y_{opt} is the final optimized solution that contains information about the location of every heater in the form (1).

4. Practical example

Now we describe a practical example of the heating of an aluminium shell mould. The volume of the mould is $0.8 \times 0.4 \times 0.15\,\mathrm{m}^3$, mould thickness is 8 mm (see Figure 5), the number of elementary surfaces, $N = 2,064$; the heat radiation intensity recommended by the producer of artificial leathers, $I_{rec} = 47\,\mathrm{kW/m}^2$. We use 16 infrared heaters (i.e., $M = 16$) of the same type (producer Philips, power $1,600\,\mathrm{W}$, length 15 cm, width 4 cm). In the first step we calculate value $F(y_1)$ where the deviation of the heat radiation intensity F is defined by relation (6) and the initial individual y_1 corresponds to the following locations of the heaters. The centres of the heaters lie in the plane parallel to the x_1x_2-plane and at a distance of 10 cm from the centre of gravity T_j of the elementary surface p_j with the highest value $x_3^{T_j}$ $(1 \leq j \leq N)$. All the heaters have $r = (1, 0, 0)$ and $u = (0, 0, -1)$ (that is, all the heaters radiate downwards and they are parallel to the axis x_1). Then the deviation for this location of heaters is $F(y_1) = 20.74$.

We use the differential evolution algorithm described in subsection 3.1. to optimize the locations of the heaters. The parameters of the algorithm are as follows: population size $NP = 192$ (dimension of the problem is $6M = 96$), mutation factor $f = 0.98$ and crossover probability $CR = 0.60$. The heaters locations y_{tech} recommended by the producer technicians based on their experience in the production gives $F(y_{tech}) = 11.2204$. We obtain the optimized individual y_{opt} with value $F(y_{opt}) = 2.02$ after $4,000$ generations of the differential evolution algorithm. The dependence of the deviation $F(y_{opt})$ on the number of generations is shown in Figure 4. Furthermore, Figure 5 shows a graphical representation of heat radiation on the mould surface corresponding to individual $F(y_{opt})$ (where the levels of radiation intensity in $\mathrm{kW/m}^2$ correspond to the shades of grey colouring).

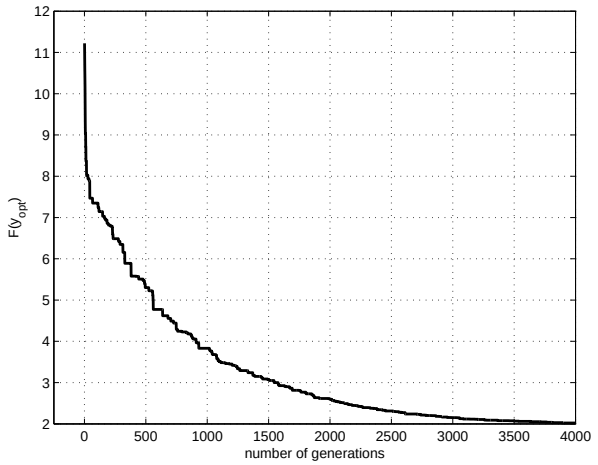


Figure 4: Dependence of $F(y_{opt})$ on the number of generations.

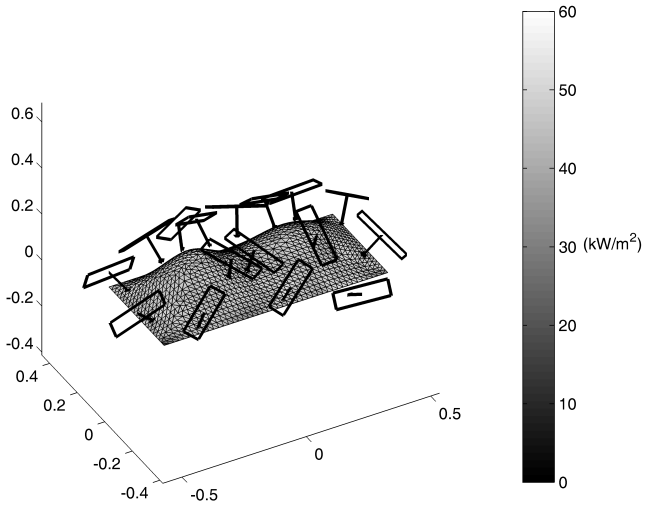


Figure 5: Heat radiation intensity (kW/m^2) on the mould surface and the locations of the heaters corresponding to the individual y_{opt}.

We made calculations on a PC computer with CPU: IntelCore i7-3770 CPU @3,4 GHz, RAM: 32 GB and GPU: GeForce GTX 460.

5. Conclusions

On the basis of practical calculations, we get a sufficiently exact solution for the optimized locations of heaters over the mould. We obtained more exact results using the differential evolution algorithm than using a genetic algorithm in numerical experiments (see [4], [5]). The temperature differences on the inner mould surface have to be maintained in the range of 3 °C during the mould heating process. The heat conductivity of the mould helps to unify different temperatures on the mould surface.

The locations of heaters determined on the basis of experience of technicians produces significantly worse results than the optimized locations. Generally, this approach is more time consuming (approximately two to three weeks depending on the mould size and the number of heaters). Furthermore, calculated optimization of the locations of heaters is more accurate and faster than optimization based on technicians experience.

The described method for manufacturing is an energy-efficient way of artificial leathers production. The given optimization process is advantageous for producer and induces virtually no additional cost.

Acknowledgements

This work has been supported by grants SGS-FP-TUL 21049 and SGS-FM-TUL.

References

[1] Antia, H. M.: *Numerical methods for scientists and engineers*. Birkhăuser Verlag, Berlin, 2002.

[2] Budinský, B.: *Analytical and differential geometry*. SNTL, Prague, 1983 (in Czech).

[3] Mlýnek, J. and Srb, R.: The process of aluminium mould warming in the car industry. Journal of Automation, Mobile Robotics and Intelligent Systems, Industrial Research Institute for Automation and Measurements PIAP, Warsaw **6 (2)** (2012), 47–51.

[4] Mlýnek, J. and Srb, R.: The process of an optimized heat radiation intensity calculation on a mould surface. In: K. G. Troitzsch (Ed.), *Proceedings of the 29th European Conference on Modelling and Simulation*, Koblenz, Germany (May 2012), 461–467, doi: 10.7148/2012-0461-0467.

[5] Mlýnek, J. and Srb, R.: Optimization of a heat radiation intensity on a mould surface in the car industry. In: R. Jablonski and T. Brezina (Eds.), *Proceedings of the 29th Conference Mechatronics 2011*, Faculty of Mechatronics, Warsaw University of Technology, Warsaw (September 2011), 531–540, doi: 10.1007/978-3-642-23244-2_64.

[6] Price, K. V., Storn, R. M., and Lampien, J. A.: *Differential evolution*. Springer-Verlag, Berlin, Heidelberg, 2005.

[7] Stocker, J. J.: *Differential geometry*. John Wiley&Sons, New York, 1989.

[8] Zhang, J., Sanderson, A. C.: *Adaptive differential evolution*. Springer-Verlag, Berlin, Heidelberg, 2009.

Conference Applications of Mathematics 2015, in honor of the birthday anniversaries of Ivo Babuška (90), Milan Práger (85), and Emil Vitásek (85)
J. Brandts, S. Korotov, M. Křížek, K. Segeth, J. Šístek, T. Vejchodský (Eds.)
Institute of Mathematics CAS, Prague 2015

DYNAMIC CONTACT PROBLEMS IN BONE NEOPLASM ANALYSES AND THE PRIMAL-DUAL ACTIVE SET (PDAS) METHOD

Jiří Nedoma

Institute of Computer Science, Czech Academy of Sciences
Pod Vodárenskou věží 2, 182 07 Prague 8, Czech Republic
nedoma@cs.cas.cz

Dedicated to Prof. Ivo Babuška, Dr. Milan Práger and Dr. Emil Vitásek on the occasion of their life jubilees.

Abstract: In the contribution growths of the neoplasms (benign and malignant tumors and cysts), located in a system of loaded bones, will be simulated. The main goal of the contribution is to present the useful methods and efficient algorithms for their solutions. Because the geometry of the system of loaded and possible fractured bones with enlarged neoplasms changes in time, the corresponding mathematical models of tumor's and cyst's evolutions lead to the coupled free boundary problems and the dynamic contact problems with or without friction. The discussed parts of these models will be based on the theory of dynamic contact problems without or with Tresca or Coulomb frictions in the visco-elastic rheology. The numerical solution of the problem with Coulomb friction is based on the semi-implicit scheme in time and the finite element method in space, where the Coulomb law of friction at every time level will be approximated by its value from the previous time level. The algorithm for the corresponding model of friction will be based on the discrete mortar formulation of the saddle point problem and the primal-dual active set algorithm. The algorithm for the Coulomb friction model will be based on the fixpoint algorithm, that will be an extension of the PDAS algorithm for the Tresca friction. In this algorithm the friction bound is iteratively modified using the normal component of the Lagrange multiplier. Thus the friction bound and the active and inactive sets are updated in every step of the iterative algorithm and at every time step corresponding to the semi-implicit scheme.

Keywords: dynamic contact problems, mathematical models of neoplasms - tumors and cysts, Coulomb and Tresca frictions, variational formulation, semi-implicit scheme, FEM, mortar aproximation, PDAS algorithm.

MSC: 65K10, 65C20, 65N15, 65N30, 74M15

1. Introduction

In biology and medical sciences mathematical models play an important role. The role of mathematical models are then to explain a set of biomedical experiments and analyses. During the last four decades, various neoplasms (cysts, benign and malign tumors) models have been developed, analyzed and discussed.

By **neoplasm** is meant a mass of tissue that forms when cells divide uncontrollably, that is, by an overproduction of cells. Neoplasms are benign tumors, malignant tumors or cancers and cysts. Cancers are of several types due to their origin, that is, due to the tissue from which they arise and the type of cells involved. A cancer of white blood cells is called leukemia, cancers arising in muscles and connective tissue are called sarcoma, and a cancer originated from epithelial cells is called carcinoma. A bone tumors are represented by abnormal growth of cells within the bone that are of (i) noncancerous types, and we speak about **benign bone tumors**, or (ii) cancerous types, and we speak about **malignant bone tumors**. In some cases the cancer cells invade into the blood or the lymphatic vessels and then are transported into another locations, where they create secondary tumors. This process is known as the **metastasis process**. Malign tumors rise relatively very quickly approximately 1mm/day. In all types of neoplasms a solid tumors can be detected when it reaches a size of several millimeters. Bone tumors are of primary types, originating within the bone tissues, or of secondary types, that result from the spread cancer cells from the primary tumors located in other tissues in the human body and we speak about **metastasis**. Growing tumors replace healthy tissue with abnormal benign or malignant tissues. Benign tumors are not life-threatening, expecting such benign tumors that are changed into malignant tumors. Benign bone tumors as well as cysts do not metastasize, that is, they do not spread to other tissues but remain situated in the bone or in the other tissue. Since bones are composed of hard mineralized tissues, they are more resistant to destruction than other soft tissues, but in some cases the loaded long bones, vertebra or jaw-bones with tumors and cysts can fracture. The classifications of neoplasms are published by the World Health Organization - WHO.

Cancers arise from one single tumor cell. The transformation from the normal cells into tumor cells are multistage processes, where the evolution of cells are regulated and controlled by genes constrained in their nucleus. A special feature in tumor growth is proliferation. Proliferating cells are causes of the tumor volume which varying in time. A tumor contains different populations of cells, such as (i) proliferating cells, i.e., cells that undergo abnormally fast mitosis; (ii) necrotic cells, i.e., cells that died due to a luck of nutrion; (iii) quiescent cells, i.e., cells that are alived but their rate of mitosis is balanced by the rate of natural death. By mitosis it is meant the process of cell division which results in the production of two daughter cells from an initial parent cell and that are identical with the parent cell.

Another type of neoplasms are **cysts** that are filled by fluid and that are formed either in bones or in soft tissues, respectively. **Cysts** are pathological cavity lined

by the own epithelium and in the cyst lumen filled by fluid or semi-fluid contents, that are not created by the accumulation of pus materials and generally are formed by a connective tissue walls. In this study we will limit ourselves to the odontogenic cysts only. **Odontogenic cysts** are cysts of the jaw-bone that are lined by an odontogenic epithelium (that is, avascular epithelial tissues). Odontogenic cysts are relatively slow growing and represent in early states of evolution no great problem and treat to human life. The main types are the radicular cysts, that grow relatively slowly and the keratocysts, that grow more rapidly.

2. Formulation of the problem

2.1. Formulation of the contact problem

Let the system of bones with neoplasms occupy a region $\Omega \in \mathbb{R}^N$, $N = 2, 3$, (Fig.1a,b,c), the geometry of which can be determined from the CT or MRI scans, respectively, and approximated by the visco-elasticity with short memory (Kelvin-Voigt type rheology).

Let $I = (0, t_p)$, $t_p > 0$, be a time interval. Let $\Omega \subset \mathbb{R}^N$, $N = 2, 3$, be a region occupied by a system of bodies (bones) of arbitrary shapes Ω^ι such that $\Omega = \cup_{\iota=1}^{r}(\Omega^\iota \cup \Gamma_{cv}^\iota)$. Let Ω^ι have Lipschitz boundaries $\partial\Omega^\iota$ and let us assume that $\partial\Omega = \Gamma_\tau \cup \Gamma_u \cup \Gamma_c$, where the disjoint parts Γ_τ, Γ_u, Γ_c are open subsets. Moreover, let $\Gamma_\tau = {}^1\Gamma_\tau \cup {}^2\Gamma_\tau$, $\Gamma_u = {}^1\Gamma_u \cup {}^2\Gamma_u$ and $\Gamma_c = \cup_{s,m}\Gamma_c^{sm}$, $\Gamma_c^{sm} = \partial\Omega^s \cap \partial\Omega^m$, $s \neq m$, $s, m \in \{1, \ldots, r\}$, Γ_c^{sm} represent the contact boundaries between the components of joints as well as between two opposite faces of cracks, $\Gamma_{cv} = \cup_s \Gamma_{cv}^s$, $\Gamma_{cv}^s \subset \partial\Omega_1^s \cap \partial\Omega_2^s$, represent virtual interfaces between regions Ω_1^s and Ω_2^s. It is evident that these boundaries are determined as results of the used neoplasm's growth models. Let $\Omega(t) = \Omega \times I$ denote the time-space domain and let $\Gamma_\tau(t) = \Gamma_\tau \times I$, $\Gamma_u(t) = \Gamma_u \times I$, $\Gamma_c(t) = \Gamma_c \times I$ denote the parts of its boundary $\partial\Omega(t) = \partial\Omega \times I$. In the study we will assume that the contact boundaries Γ_c^{sm} are between contact boundaries of joints (i.e., hip joints, knee joints, spine, temporomandibular joints, etc.) as well as contact boundaries between the opposite boundaries in the fractures of bones and/or of vertebra. In the case of e.g. vertebra fracture, the domain denoted as Ω^s will be divided into two parts denoted by Ω_1^s and Ω_2^s (see Figs 1a-c).

Furthermore, let $\mathbf{n}$ denote the outer normal vector of the boundary, $u_n = u_i n_i$, $\mathbf{u}_t = \mathbf{u} - u_n\mathbf{n}$, $\tau_n = \tau_{ij}n_j n_i$, $\boldsymbol{\tau}_t = \boldsymbol{\tau} - \tau_n\mathbf{n}$ be normal and tangential components of displacement and stress vectors $\mathbf{u} = (u_i)$, $\boldsymbol{\tau} = (\tau_i)$, $\tau_i = \tau_{ij}n_j$, $i, j = 1, \ldots, N$. Let $\mathbf{F}$, $\mathbf{P}$ be the body and surface forces, ρ the density. The respective time derivatives are denoted by "$\prime$". Let us denote by $\mathbf{u}' = (u_k')$ the velocity vector. To formulate the contact and friction conditions, let us introduce at each point of Γ_c^s the vectors $\mathbf{t}_i^s$, $i = N - 1$, spanning in the corresponding tangential plane. Let $\{\mathbf{n}^s, \mathbf{t}_i^s\}$, $i = 1, 2$, be an orthogonal basis in $\mathbb{R}^N$ for each point of Γ_c^s. To formulate the non-penetration condition we use a predefined relation between the points of the possible contact zones Γ_c. Therefore, we introduce a smooth mapping $\mathcal{R} : \Gamma_c^s \to \Gamma_c^m$ such that $\mathcal{R}(\Gamma_c^s) \subset \Gamma_c^m$, and assume that the mapping $\mathcal{R}$ is

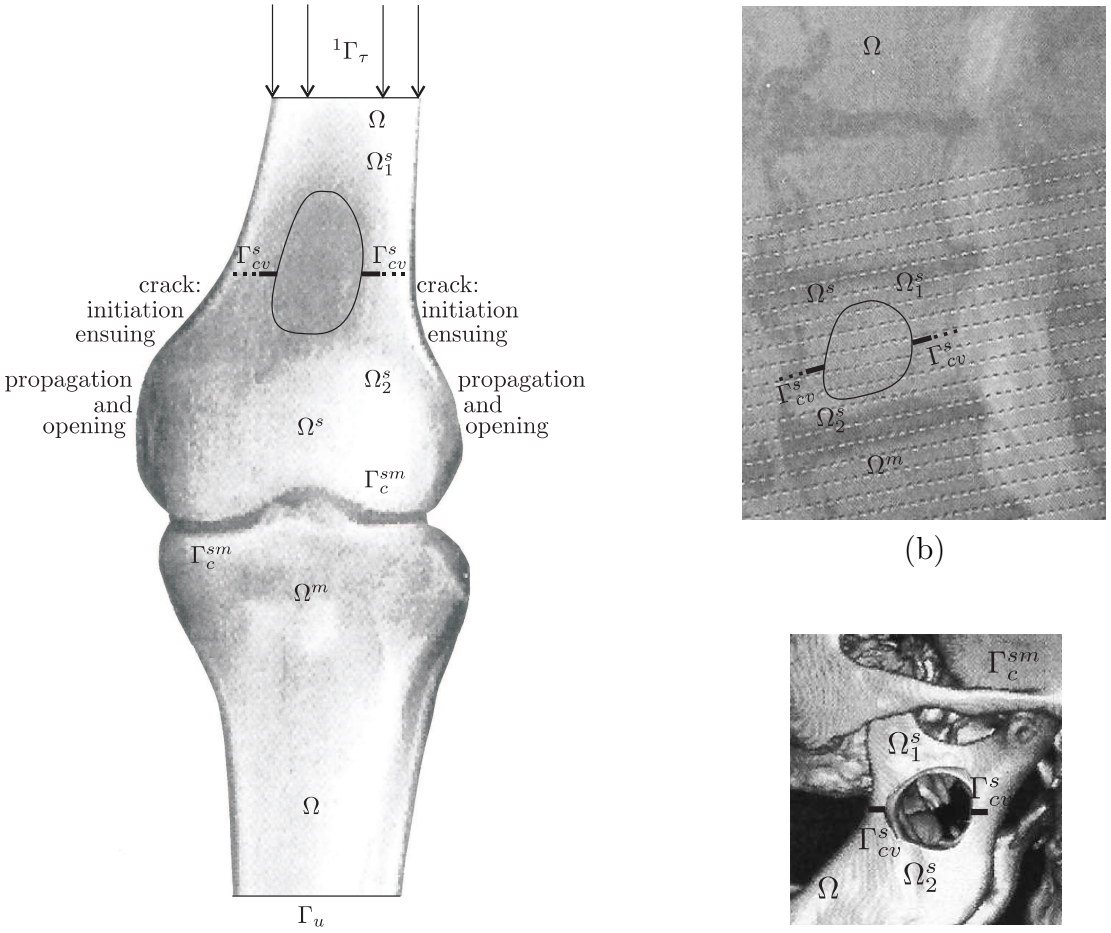


Figure 1: Mathematical models of the long bone and spine with tumors and the jaw-bone with cyst: crack initiation and ensuing crack propagation and crack opening are modelled on the basis of dynamic PDAS method for a crack problem with non-penetration: (a) the detail of knee joint with the tumor; (b) the detail of spine with the tumor; (c) the detail of jaw-bone with the cyst.

well defined and maps any $\mathbf{x} \in \Gamma_c^s$ to the intersection of the normal on Γ_c^s at $\mathbf{x}$ with Γ_c^m. Then $[\mathbf{u}]^{sm} := \mathbf{u}^s(\mathbf{x},t) - \mathbf{u}^m(\mathcal{R}(\mathbf{x},t))$, $[u_n]^{sm} := [\mathbf{u}]^{sm} \cdot \mathbf{n}^s$ is the jump in normal direction, $[\mathbf{u}_t]^{sm} = (\mathbf{u}^s(\mathbf{x},t) - \mathbf{u}^m(\mathcal{R}(\mathbf{x},t))) - [\mathbf{u}]^{sm} \cdot \mathbf{n}^s$ is the jump in the tangential direction and $\tau_n^s = (\mathbf{n}^s)^T \boldsymbol{\tau}^s(\mathbf{x},t)\mathbf{n}^s = (\mathbf{n}^s)^T \boldsymbol{\tau}^m(\mathcal{R}(\mathbf{x},t))\mathbf{n}^s$ is the boundary stress in normal direction on the possible contact part, and moreover, $(\mathbf{t}_i^s)^T \boldsymbol{\tau}^s(\mathbf{x},t)\mathbf{t}_i^s = (\mathbf{t}_i^s)^T \boldsymbol{\tau}^m(\mathcal{R}(\mathbf{x},t))\mathbf{t}_i^s$, $i = N-1$, is satisfied.

From the momentum conservation law the equation of motion is of the form

$$\rho \frac{\partial^2 u_i^\iota}{\partial t^2} = \frac{\partial \tau_{ij}^\iota}{\partial x_j} + F_i^\iota, \quad i,j = 1,\dots,N, \quad \iota = 1,\dots,r, \quad (\mathbf{x},t) \in \Omega^\iota(t) = \Omega^\iota \times I\,, \quad (1)$$

where

$$\begin{aligned}\tau_{ij}^{\iota} =&\tau_{ij}^{\iota}(\mathbf{u}^{\iota},\mathbf{u}'^{\iota}) = c_{ijkl}^{(0)\iota}(\mathbf{x})e_{kl}(\mathbf{u}^{\iota}) + c_{ijkl}^{(1)\iota}(\mathbf{x})e_{kl}(\mathbf{u}'^{\iota}) = \\ =&{}^{e}\tau_{ij}^{\iota}(\mathbf{u}^{\iota}) + {}^{\nu}\tau_{ij}^{\iota}(\mathbf{u}'^{\iota}), \quad i,j,k,l = 1,\dots,N\,,\ \iota = 1,\dots,r, \end{aligned} \tag{2}$$

where $c_{ijkl}^{(n)\iota}(\mathbf{x})$, $n = 0,1$, are anisotropic elastic and viscous coefficients and $e_{ij}(\mathbf{u})$ are components of the small strain tensor, N is the space dimension. For the tensors $c_{ijkl}^{(n)\iota}(\mathbf{x})$, $n = 0,1$, we assume that they satisfy the symmetric and Lipschitz conditions, that is,

$$\begin{aligned}&c_{ijkl}^{(n)\iota} \in L^{\infty}(\Omega^{\iota}), \quad n = 0,1,\ \iota = 1,\dots,r, \quad c_{ijkl}^{(n)\iota} = c_{jikl}^{(n)\iota} = c_{klij}^{(n)\iota} = c_{ijlk}^{(n)\iota}\,, \\ &c_{ijkl}^{(n)\iota}e_{ij}e_{kl} \geq c_0^{(n)\iota}e_{ij}e_{ij} \quad \forall e_{ij}, \quad e_{ij} = e_{ji} \quad \text{and a.e.} \quad \mathbf{x} \in \Omega^{\iota},\ c_0^{(n)\iota} > 0,\ \iota = 1,\dots,r, \\ &c_{ijkl}^{(n)\iota} = \lambda^{(n)\iota}\delta_{ij}\delta_{ij} + \mu^{(n)\iota}\left(\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}\right), n = 0,1, \text{for the isotropic bone materials,}\end{aligned} \tag{3}$$

where a repeated index implies summation from 1 to N.

On the contact boundaries between neighbouring bones and the neighbouring faces in the case of bone fractures the following non-penetration conditions and the Coulomb friction conditions

$$\left.\begin{aligned}&[u_n]^{sm} \leq d^{sm},\ \tau_n^s = \tau_n^m \equiv \tau_n^{sm} \leq 0, \\ &\left([u_n]^{sm} - d^{sm}\right)\tau_n^{sm} = 0, \\ &[\mathbf{u}_t']^{sm} = \mathbf{0} \ \Rightarrow\ |\boldsymbol{\tau}_t^{sm}| \leq \mathcal{F}_c^{sm}\,|\tau_n^{sm}(\mathbf{u})|\,, \\ &[\mathbf{u}_t']^{sm} \neq \mathbf{0} \ \Rightarrow\ \boldsymbol{\tau}_t^{sm} = -\mathcal{F}_c^{sm}\,|\tau_n^{sm}(\mathbf{u})|\,\tfrac{[\mathbf{u}_t']^{sm}}{|[\mathbf{u}_t']^{sm}|},\end{aligned}\right\} (\mathbf{x},t) \in \cup_{e,m}\Gamma_c^{sm} \times I, \tag{4}$$

are given and on the boundary $\partial\Omega(t)$ the following conditions

$$\tau_{ij}n_j = P_i, \quad i,j = 1,\dots,N, \quad (\mathbf{x},t) \in \Gamma_{\tau}(t) = \cup_{\iota=1}^{r}(\Gamma_{\tau} \cap \partial\Omega^{\iota}) \times I\,, \tag{5}$$

$$u_i = u_{2i}, \quad i = 1,\dots,N, \quad (\mathbf{x},t) \in\ \Gamma_u(t) = \cup_{\iota=1}^{r}(\Gamma_u \cap \partial\Omega^{\iota}) \times I\,, \tag{6}$$

are prescribed and the initial conditions

$$\mathbf{u}^{\iota}(\mathbf{x},0) = \mathbf{u}_0^{\iota}(\mathbf{x}), \quad \mathbf{u}'^{\iota}(\mathbf{x},0) = \mathbf{u}_1^{\iota}(\mathbf{x}), \quad \mathbf{x} \in \Omega^{\iota}\,, \tag{7}$$

are given, where $\boldsymbol{\tau}_t^{sm} \equiv \boldsymbol{\tau}_t^s = -\boldsymbol{\tau}_t^m$, $\mathcal{F}_c^{sm} = \mathcal{F}_c^{sm}(\mathbf{x},\mathbf{u}_t')$ is globally bounded, nonnegative, and satisfies the Carathéodory conditions [4, 18, 20] and $\mathbf{u}_0$, $\mathbf{u}_1$ are the given functions, $\mathbf{u}_2 \neq 0$ on ${}^1\Gamma_u$ or $= 0$ on ${}^2\Gamma_u$ has a time derivative $\mathbf{u}_2'$, and on $\cup\Gamma_c^{sm}$ due to the equilibrium of forces $\tau_{ij}(\mathbf{u}^s)n_j^s = -\tau_{ij}(\mathbf{u}^m)n_j^m$ and where $[\mathbf{v}]^{sm} = \mathbf{v}^s - \mathbf{v}^n$ is a jump (difference) of quantities $\mathbf{v}^s$ and $\mathbf{v}^m$ and d^{sm} is a gap, where

$$d^{sm}(\mathbf{x}) = \frac{\varphi^s(\mathbf{x}) - \varphi^m(\mathbf{x})}{\sqrt{1 + |\nabla\varphi^s(\mathbf{x})|^2}}\,,$$

where $\varphi^s, \varphi^m \in C^1$ are functions defined on an open subset Γ_c^{sm} of $\mathbb{R}^{N-1}$ parametrized the two contact boundaries, e.g. of joints in the first case, and the two opposite faces of the cracks in the second case. Thus the terms $d^{sm} \geq 0$ are the normalized gaps between the contact boundaries of Ω^s and Ω^m (e.g. of the joints or faces in the case of fractures) and between the two faces of the crack (i.e., Γ_c^s and Γ_c^m).

2.2. Formulation of coupled free boundary problems

Furthermore, we need to determine the evolution of neoplasms (tumors and cysts) in time, and then to determine the areas that are occupied by these tumors and cysts inside the system of bones, that create the investigated part of the human skeleton, and moreover, to determine their material compositions, all during the studied time period.

(A) The tumor growth case

The tumor's study and their growths are studied e.g. in [2, 3, 5] and in many others. Such models consist of a system of coupled partial differential equations and a mass conservation law. The problems then lead to solve the free boundary problems.

In the case of the tumor growth, we limit ourselves to avascular and vascular cases only. Let $u_c(\mathbf{x},t)$ denote the concentration of cells, and let $u_p(\mathbf{x},t)$, $u_q(\mathbf{x},t)$ and $u_D(\mathbf{x},t)$ denote the cell densities for proliferating, quiescent and dead cells, respectively, where $\mathbf{x}$ denotes a spatial coordinate and t time, $t \in I$, $\overline{I} \in [t_0, t_p]$, $t_0 \geq 0$, $t_p > 0$ (see [2, 3, 5]).

To determine the equation for the concentration $u_c(\mathbf{x},t)$, we must consider two cases — the avascular stage and the vascular stage. Then, for an avascular evolution of tumors we find

$$\varepsilon_0 \frac{\partial u_c}{\partial t} = D_c \nabla^2 u_c - \lambda u_c, \quad \varepsilon_0 = \frac{T_{\text{diffusion}}}{T_{\text{growth}}}, \tag{8}$$

where D_c is a diffusion coefficient, about which is assumed to be constant, λ is the nutrient consumption rate, ε_0 is the ratio of the nutrient diffusion time scale to the tumor growth time scale, $T_{\text{diffusion}} \sim 1$ minute, while $T_{\text{growth}} \sim 1$ day, so that ε_0 is small. For a vascular evolution of tumors the Eq. (8) must be replaced by

$$\varepsilon_0 \frac{\partial u_c}{\partial t} = D_c \nabla^2 u_c + \Gamma(u_{cB} - u_c) - \lambda u_c\,, \tag{9}$$

where u_{cB} is the nutrient concentration in the vasculature, Γ is the rate of the blood-tissue transfer, so that $\Gamma(u_{cB} - u_c)$ represents the nutrient concentration after the process of angiogenesis. Tumor angiogenesis refers to the ability of a tumor to stimulate new blood vessel formation.

In the case of vascularized tumors if we use the change of variables, that is, if we put

$$u_c - \frac{\Gamma u_{cB}}{\Gamma + \lambda} \to u_c, \quad \Gamma + \lambda \to \lambda\,, \tag{10}$$

then Eq. (9) is transformed to Eq. (8), that is, u_c in the avascular and vascular tumors are described by the same equation (8).

We assume that proliferating cells become quiescent at a rate $K_Q(u_c)$ that depends on the concentration $u_c(\mathbf{x},t)$ of a generic nourishment having an influence on

a tumor growth and that their death rate is $K_A(u_c)$, that also depends on $u_c(\mathbf{x},t)$. The quiescent cells become necrotic at a rate $K_D(u_c)$ that depends also on the concentration $u_c(\mathbf{x},t)$. The quiescent cells become proliferating at a rate $K_P(u_c)$ that also depends on the concentration of nutrient $u_c(\mathbf{x},t)$. The density of proliferating cells is increasing due to proliferation at a rate $K_B(u_c)$ that also depending on $u_c(\mathbf{x},t)$. Finally, the dead cells are removed from the tumor, as they decompose, at a constant rate K_R. Since cells proliferate and dead cells are removed from the tumor, there exists a continuous motion of cells within the tumor, which is represented by a velocity $\mathbf{v}$. Denoting by $\omega_n(t)$ a region occupied by a tumor at time t and $\partial\omega_n(t)$ its boundary, then the conservation of mass laws for the densities of the proliferating cells $u_p(\mathbf{x},t)$, the quiescent cells $u_q(\mathbf{x},t)$ and the dead cells $u_D(\mathbf{x},t)$ are as follows:

$$\frac{\partial u_p}{\partial t} + \operatorname{div}(u_p\mathbf{v}) = \left[K_B(u_c) - K_Q(u_c) - K_A(u_c)\right] u_p + K_P(u_c)u_q\,, \tag{11}$$

$$\frac{\partial u_q}{\partial t} + \operatorname{div}(u_q\mathbf{v}) = K_Q(u_c)u_p - \left[K_P(u_c) + K_D(u_c)\right] u_q\,, \tag{12}$$

$$\frac{\partial u_D}{\partial t} + \operatorname{div}(u_D\mathbf{v}) = K_A(u_c)u_p + K_D(u_c)u_q - K_R u_D\,. \tag{13}$$

Assuming that the tumor tissue is modelled by a porous medium and the moving cells by a fluid flow, then the velocity $\mathbf{v}$ of fluid flow is related to the fluid pressure σ by the Darcy law, thus

$$\mathbf{v} = -\beta\nabla\sigma, \quad \text{where} \quad \beta > 0\,. \tag{14}$$

Moreover, assuming that all cells are physically identical in volume and mass, therefore, their density is constant inside the tumor, that is,

$$u_p + u_q + u_D = N = \text{const.}$$

For simplicity, we can put $\beta = 1$ and $N = 1$.

Adding Eqs (11), (8) with (10), we find

$$\operatorname{div}\mathbf{v} = K_B(u_c)u_p - K_R u_D\,,$$

and substituting $u_D = 1 - u_p - u_q$, then we obtain the following problem describing the growth of the tumor:

Problem $(\mathcal{P}_T)$: Find u_c, u_p, u_q, σ satisfying the following system of equations

$$\varepsilon_0\frac{\partial u_c}{\partial t} = D_c\nabla^2 u_c - \lambda u_c \qquad \text{in} \quad \omega_n(t),\ t > 0\,, \tag{15}$$

$$\frac{\partial u_p}{\partial t} - \nabla\sigma\cdot\nabla u_p = f(u_c,u_p,u_q) \qquad \text{in} \quad \omega_n(t),\ t > 0\,, \tag{16}$$

$$\frac{\partial u_q}{\partial t} - \nabla\sigma\cdot\nabla u_q = g(u_c,u_p,u_q) \qquad \text{in} \quad \omega_n(t),\ t > 0, \tag{17}$$

$$\Delta\sigma = -h(u_c,u_p,u_q) \qquad \text{in} \quad \omega_n(t),\ t > 0\,, \tag{18}$$

where

$$
\begin{aligned}
f(u_c,u_p,u_q) &= [K_B(u_c) - K_Q(u_c) - K_A(u_c)]\, u_p + K_P(u_c)u_q - h(u_c,u_p,u_q)u_p,\\
g(u_c,u_p,u_q) &= K_Q(u_c)u_p - [K_p(u_c) + K_D(u_c)]\, u_q - h(u_c,u_p,u_q)u_q\,,\\
h(u_c,u_p,u_q) &= [K_B(u_c) + K_R]\, u_p + K_R u_q - K_R\,,
\end{aligned}
$$

with the boundary conditions on $\partial\omega_n(t)$

$$u_c = u_{c1} \qquad \text{on} \quad \partial\omega_n(t),\ t>0, \tag{19}$$

$$\sigma = \gamma\kappa,\ \frac{\partial\sigma}{\partial n} = -v_n \qquad \text{on} \quad \partial\omega_n(t),\ t>0, \tag{20}$$

and with the initial conditions

$$u_c(\mathbf{x},t_0) = u_{c0}(\mathbf{x}) \qquad \text{in} \quad \omega_n(t_0),\ u_{c0}(\mathbf{x}) \geq 0, \tag{21}$$

$$u_p(\mathbf{x},t_0) = u_{p0}(\mathbf{x}) \qquad \text{in} \quad \omega_n(t_0),\ u_{p0}(\mathbf{x}) \geq 0, \tag{22}$$

$$u_q(\mathbf{x},t_0) = u_{q0}(\mathbf{x}) \qquad \text{in} \quad \omega_n(t_0),\ u_{q0}(\mathbf{x}) \geq 0, \tag{23}$$

where $u_{p0}(\mathbf{x}) + u_{q0}(\mathbf{x}) \leq 1$, and where u_{c1} is a constant concentration of nutrients, v_n is the velocity of the free boundary, κ is the mean curvature, γ is the surface tension coefficient and u_{c0}, u_{p0}, u_{q0} are given functions.

Under the assumption that the initial data are smooth and the initial and boundary data are consistent with the Eq. (15) at $\partial\omega_n(t_0)$, we have the following result [3]:

Theorem 1 *Let the initial data be sufficiently smooth, the physical data be constant and the consistency conditions be satisfied, then there exists a unique smooth solution to Problem ($\mathcal{P}_T$) for $t \in \overline{I} = [0, t_p]$.*

(B) The case of the cystic growths

Our mathematical model of cystic growth is based on the diffusive mechanisms, cell birth and death, the idea of osmosis, the balance between osmotic and hydrostatic pressure forces within the cyst structure and its neighboring tissue. By the **osmosis** we understand the diffusive process of permeability between two different liquids which are mutually separated by a porous membrane.

Let us assume that the cyst occupies the region, we denote it by ω_c (e.g. it can be a sphere of radius R or of an arbitrary shape) with a thin epithelial rim of cells covering its surface. The lumen of the cyst is assumed to be filled by dead cellular material, consisting partly of osmotic material concentration C^+, with total mass S, generating an osmotic pressure P_0^+. Inside the cyst is observed the hydrostatic pressure, we denote it as P_h^+. The neighborhood of the cyst is created by a material, consisting of a fixed osmotic material of concentration C^-, generating an osmotic pressure P_0^-. The hydrostatic pressure here is P_h^-. According to the size of the cavity the thickness of the capsule and the epithelial layer can be neglected. The

growth of radicular cysts is of about a few millimeters per year, while in the keratocyst's case their growths are several times higher. The osmotic pressure difference $\Delta P_0 = P_0^+ - P_0^-$ relates to the difference in osmolality Δm, that is,

$$\Delta P_0 = \Delta m R_g T \sim 28.3 \; Nm^{-2} \,, \tag{24}$$

where Δm is the molar concentration of "osmotic active" molecular per litre (~ 0.011 Osml $\equiv 0.011$ mol), $R_g = 8.31$ J/mol.K is the ideal gas constant, T is the absolute temperature. For the hydrostatic pressure difference between the interior of the cyst and the neighborhood balances the osmotic pressure difference between the cyst interior and its neighborhood at the cyst rim, i.e.,

$$P_h^+ - P_h^- = P_0^+ - P_0^- = \alpha(C^+ - C^-), \quad \alpha = R_g T \,, \tag{25}$$

where the van Hoff equation was used, where α is the proportional coefficient R_g is the ideal gas constant, T is the temperature [21].

Since the cyst grows, cells migrate towards the interior of cavity, where they die and since the degraded material driving the osmosis does not penetrate the epithelial layer (i.e., membrane) it then start to be a part of osmotic material. The osmotic material is cummulated in the cavity of the cyst and only fluid can pass the semi-permeable epithelial membrane. Let "s" be the total amount of degraded material inside the cyst. Then the rate of change of mass of osmotic material in the core in time, i.e., of "$\dot{s} = \frac{ds}{dt}$", is proportional to the surface area of the covering epithelium, we denote it as S_c, then we have

$$\frac{ds}{dt} = \beta S_c \,, \tag{26}$$

where β is a supply rate of the osmotic material and it can change according to the type of cyst.

The hydrostatic pressure jump across the epithelial membrane balances the stresses in the semi-permeable membrane and the stresses on the cyst from the neighboring bone tissue. Thus

$$P_h^+ - P_h^- = f(\mathbf{r}, \dot{\mathbf{r}}) + f_b(\mathbf{r}, \dot{\mathbf{r}}) \,, \tag{27}$$

where f is the **physical stresses**, depending on the material properties of the cyst and the neighboring bone tissue, which in general is a function of a position vector $\mathbf{r}$ of the surface point, and $\dot{\mathbf{r}} = \frac{d\mathbf{r}}{dt}$ is the time derivative of $\mathbf{r}$, and f_b corresponds to the **biological stresses**. The natures of these stresses in situ are not known currently, therefore, the term $f_b(\mathbf{r}, \dot{\mathbf{r}})$ can be omitted, i.e., $f_b(\mathbf{r}, \dot{\mathbf{r}}) = 0$.

Ward et al. [23] expect that the material of surrounding tissue is mixture of elastic and non-elastic (viscous) materials and that it can be modelled by a linear viscoelastic fluid of Maxwell type with a stiffness E and a viscosity ν. The total strain is the sum of the elastic and viscous strains and the total strain rate is the

sum of its elastic and viscous strain rate,i.e., $\varepsilon = \varepsilon^e + \varepsilon^\nu, \quad \dot{\varepsilon} = \dot{\varepsilon}^e + \dot{\varepsilon}^\nu$, where $\dot{\varepsilon} = \frac{d\varepsilon}{dt}$. Since $\dot{\varepsilon}^e = \frac{\dot{f}}{E}$, and $\dot{\varepsilon}^\nu = \frac{f}{\nu}$, then we obtain

$$\dot{f} + \tau^{-1} f = E\dot{\varepsilon}\,, \tag{28}$$

where $\tau = \frac{\nu}{E}$ is the so-called **relaxation time**.

From (25) the osmotic pressure difference is equal to the hydrostatic pressure difference, i.e., $\frac{1}{\alpha}(P_0^+ - P_0^-) = \frac{1}{\alpha}(P_h^+ - P_h^-) = \frac{1}{\alpha} f(\mathbf{r}, \dot{\mathbf{r}})$, and therefore, the physical stresses $\frac{1}{\alpha} f(\mathbf{r}, \dot{\mathbf{r}}) = C^+ - C^-$. Hence, the concentration of degraded material

$$C^+ = C^- + \frac{1}{\alpha} f(\mathbf{r}, \dot{\mathbf{r}})\,, \tag{29}$$

that is, it is a linear function of the stresses, since C^- and α are assumed to be constant.

The concentration of material inside the cyst, given as its total mass "s" divided by the cavity volume v_c, is

$$C^+ = \frac{s}{v_c} = \frac{s}{|\omega_c|}\,, \tag{30}$$

where ω_c represents the region occupied by the cyst, i.e., $v_c = |\omega_c|$. When the cyst grows in a bony tissue, the bone is resorbed and the cyst grows as there it was no obstacle stopping it from expanding. Because $C^+ = \frac{s}{v_c(\mathbf{r})}$, then substituting $s = C^+ v_c(\mathbf{r})$ into (26), i.e., $\frac{ds}{dt} = \beta S_c$, and using (29), then after some modification, we obtain

$$\frac{\dot{v}_c}{\alpha} f(\mathbf{r}, \dot{\mathbf{r}})\dot{\mathbf{r}} + \dot{v}_c C^- \dot{\mathbf{r}} + \frac{v_c}{\alpha} \dot{f}(\mathbf{r}, \dot{\mathbf{r}}) = \beta S_c\,, \tag{31}$$

representing expression relating the cyst size, its shape and the physical stresses exerted by the stroma, where β is the core supply rate of osmotic material ([mol/m^2.s]) and is different for the radicular cysts and the keratocysts for which is several times higher than for radicular cysts.

Since we model the material which is a mixture of fluid, collagenous capsule, and crystalline structures, than it can be described as Maxwell's fluid. Due to (28) the stresses satisfy

$$\tau \dot{f}(\mathbf{r}, \dot{\mathbf{r}}) + f(\mathbf{r}, \dot{\mathbf{r}}) = \nu\dot{\varepsilon}\,, \tag{32}$$

as $\tau = \frac{\nu}{E}$. The problem will be complete, if the initial condition for $\mathbf{r}$ and f will be given. Thus, for $t = 0$

$$\mathbf{r}(0) = \mathbf{r}_0, \quad f(0) = f_0\,, \tag{33}$$

where $\mathbf{r}_0$ and f_0 are given.

Assuming that the cyst is of a spherical shape, then $v_c = \frac{4}{3}\pi R^3$ and $S_c = 4\pi R^2$, where R is a radius of the cyst. For more details see [23, 21, 19]. The problem can be solved by numerical methods for ODEs.

The biological materials of both types of tumors and both types of cysts are assumed to be near a fluids for which $\mu^{(0)} = 0$. When at $t \in I$ the shape of the cyst is known, it is possible to estimate a probable evolution of the cyst, and moreover, to determine a probable time of a cyst origin, similarly as in the previous case.

3. Stress-strain analysis of the loaded bone system with neoplasms

3.1. Mathematical model and its solution

The problem to be solved has the following classical formulation:

Problem $(\mathcal{P})$: Let $N = 2, 3$, $r \geq 2$. Find a displacement vector $\mathbf{u}^\iota : \overline{\Omega}^\iota \times I \to \mathbb{R}^N$ satisfying Eqs (1)–(3) and the contact conditions with the Coulomb friction (4), the boundary conditions (5)–(6) and initial conditions (7), where we assume that the geometry of ω_n and ω_c at $t = 0$ and the corresponding material coefficients were determined and that all anisotropic elastic and viscous coefficients satisfy the symmetric and Lipschitz conditions (3).

Since the problem with Coulomb friction formulated in displacements is up-to-date an open problem, therefore, for the existence analysis the contact conditions of nonpenetration (Signorini conditions) will be formulated in velocities, that is,

$$[u_n']^{sm} \leq d^{sm}, \quad \tau_n^s = \tau_n^m \equiv \tau_n^{sm} \leq 0, \quad ([u_n']^{sm} - d^{sm})\tau_n^{sm} = 0\,. \tag{34}$$

Let us introduce the spaces $L^{p,N}(\Omega)$, $p \in [1, +\infty)$, $L^\infty(\Omega)$, the Sobolev spaces $H^{1,N}(\Omega)$, $H_0^{1,N}(\Omega)$, $H^{\frac{1}{2},N}(\Gamma_c)$, $H_{00}^{\frac{1}{2},N}(\Gamma_c)$ by the usual way, and let $B(M)$ be the space of bounded functions endowed with the sup norm, and moreover, the spaces and sets

$$\begin{aligned}
&V_0 = \left\{\mathbf{v} | \mathbf{v} \in \sqcap_{\iota=1}^r H^{1,N}(\Omega^\iota), \mathbf{v} = 0 \text{ a.e. on } \Gamma_u\right\},\\
&V = \mathbf{u}_2 + V_0,\ \mathcal{V} = \mathbf{u}_2' + \mathcal{V}_0 = L^2(I; V),\quad K = \{\mathbf{v} \in V | [v_n]^{sm} \leq d^{sm} \text{ a.e. on } \Gamma_c^{sm}\}\,,\\
&\mathcal{K} = \left\{\mathbf{v} | \mathbf{v} \in L^2(I; \sqcap_{\iota=1}^s H^{1,N}(\Omega^\iota)), \mathbf{v} = \mathbf{u}_2' \text{ on } \Gamma_u(t),\ [v_n]^{sm} \leq 0 \text{ a.e. on } \Gamma_c^{sm}(t)\right\}\,.
\end{aligned}$$

Let $\rho^\iota \in C(\overline{\Omega}^\iota)$, $\rho^\iota \geq \rho_0^\iota > 0$, $c_{ijkl}^\iota \in L^\infty(\Omega^\iota)$, $\mathbf{F}^\iota, \mathbf{F}'^\iota \in L^2(I; L^{2,N}(\Omega^\iota))$, $\mathbf{P}, \mathbf{P}' \in L^2(I; L^{2,N}(\Gamma_\tau))$, $\mathbf{u}_0 \in K$, $\mathbf{u}_1 \in V$, $\mathbf{u}_2' \in L^2\left(I; \sqcap_{\iota=1}^r H^{1,N}(\Omega^\iota)\right)$, $d^{sm} \in H^{\frac{1}{2},N}(\Gamma_c^{sm})$, $d^{sm} \geq 0$ a.e. on Γ_c^{sm}, $\mathcal{F}_c^{sm} \in L^\infty(\Gamma_c^{sm})$, $\mathcal{F}_c^{sm} \geq 0$ a.e. on Γ_c^{sm}. In a special case if $\overline{\Gamma}_c^s = \cup_{\iota=1}^r \left(\partial\Omega^\iota \cap \Gamma_c^s\right) \setminus \Gamma_u^s$ then instead of the space $H^{\frac{1}{2},N}(\Gamma_c^{sm})$ we will use the space $H_{00}^{\frac{1}{2},N}(\Gamma_c^{sm})$.

The variational formulation of Problem $(\mathcal{P})$ will be obtained by the usual way. Thus,

Problem $(\mathcal{P})_v$**:** Find a displacement field $\mathbf{u} : \overline{I} \to V$ such that $\mathbf{u}(t) \in K$ for a.e. $t \in I$, and

$$\begin{aligned}
&(\mathbf{u}''(t), \mathbf{v} - \mathbf{u}(t)) + a^{(0)}(\mathbf{u}(t), \mathbf{v} - \mathbf{u}(t)) + a^{(1)}(\mathbf{u}'(t), \mathbf{v} - \mathbf{u}(t)) + j(\mathbf{v}) - j(\mathbf{u}(t)) \geq\\
&\geq (\mathbf{f}(t), \mathbf{v} - \mathbf{u}(t)) \quad \forall \mathbf{v} \in K,\ t \in I\,,
\end{aligned} \tag{35}$$

$$\mathbf{u}(\mathbf{x}, 0) = \mathbf{u}_0(\mathbf{x}), \mathbf{u}'(\mathbf{x}, 0) = \mathbf{u}_1(\mathbf{x})\,, \tag{36}$$

where the initial data $\mathbf{u}_0$, $\mathbf{u}_1$ are given functions as above, and where

$$(\mathbf{u}'', \mathbf{v}) = \sum_{\iota=1}^{r} (\mathbf{u}''^{\iota}, \mathbf{v}^{\iota}) = \int_{\Omega} \rho u_i'' v_i d\mathbf{x}\,,$$

$$a^{(n)}(\mathbf{u}^{\iota}, \mathbf{v}^{\iota}) = \sum_{\iota=1}^{r} a^{\iota}(\mathbf{u}^{\iota}, \mathbf{v}^{\iota}) = \int_{\Omega} c_{ijkl}^{(n)} e_{kl}(\mathbf{u}^{\iota}) e_{ij}(\mathbf{v}^{\iota}) d\mathbf{x}, \quad n = 0, 1\,,$$

$$(\mathbf{f}, \mathbf{v}) = \sum_{\iota=1}^{r} (\mathbf{f}^{\iota}, \mathbf{v}^{\iota}) = \int_{\Omega} \mathbf{F} \cdot \mathbf{v} d\mathbf{x} + \int_{\Gamma_\tau} \mathbf{P} \cdot \mathbf{v} ds\,,$$

$$j(\mathbf{v}) = \int_{\cup_{s,m} \Gamma_c^{sm}} \mathcal{F}_c^{sm} \left|\tau_n^{sm}(\mathbf{u}, \mathbf{u}')\right| \left([\mathbf{v}_t]^{sm}\right) ds,$$

and where the bilinear forms $a^{(n)}(\mathbf{u}, \mathbf{v})$, $n = 0, 1$, are symmetric in $\mathbf{u}$, $\mathbf{v}$ and satisfy $a^{(n)}(\mathbf{u}, \mathbf{u}) \geq c_0^{(n)} \|\mathbf{u}\|_{1,N}^2$, $c_0^{(n)} = \text{const} > 0$, $a^{(n)}(\mathbf{u}, \mathbf{v}) \leq c_1^{(n)} \|\mathbf{u}\|_{1,N} \|\mathbf{v}\|_{1,N}$, $c_1^{(n)} = \text{const} > 0$, $\mathbf{u}, \mathbf{v} \in V_0$, and moreover, where we assume that the initial data $\mathbf{u}_0$, $\mathbf{u}_1$ are given functions (e.g. they can be determined as solutions of static elastic contact problems).

To prove the existence of the solution of Problem $(\mathcal{P})_v$ the decomposition $\mathbf{v} - \mathbf{u} = \mathbf{v} - \mathbf{u} + \mathbf{u}' - \mathbf{u}' = \mathbf{w} - \mathbf{u}'$ will be used. The proof of the existence of the solution is based on the penalization and regularization techniques and is modification of that of [4].

3.2. Approximation of the problem by the Tresca model of friction

Let us assume that the Coulombian law of friction in every time level is approximated by its value g_c^{sm} from the previous time level, i.e., $g_c^{sm} \equiv \mathcal{F}_c^{sm} |\tau_n^{sm}(\mathbf{u}, \mathbf{u}')|)$ $(t - \Delta t)$. Thus g_c^{sm} is a non-negative function and has a meaning of a given friction limit (or a given friction bound, representing the magnitude of the limiting friction traction at which slip originates), and where $-g_c^{sm}$ has a meaning of a given frictional force, and Δt is a time element. Thus this problem is approximated by another problem in which in every time level we will solve the dynamic contact problem with the given friction, called the Tresca model of friction.

The corresponding variational problem is the following:

Problem $(\mathcal{P}_0)_v$: Find a displacement field $\mathbf{u} : \overline{I} \to V$ such that $\mathbf{u}(t) \in K$ for a.e. $t \in I$, and

$$(\mathbf{u}''(t), \mathbf{v} - \mathbf{u}(t)) + a^{(0)}(\mathbf{u}(t), \mathbf{v} - \mathbf{u}(t)) + a^{(1)}(\mathbf{u}'(t), \mathbf{v} - \mathbf{u}(t)) + j(\mathbf{v}) - j(\mathbf{u}(t)) \geq$$
$$\geq (\mathbf{f}(t), \mathbf{v} - \mathbf{u}(t)) \quad \forall \mathbf{v} \in K,\ t \in I\,, \tag{37}$$
$$\mathbf{u}(\mathbf{x}, 0) = \mathbf{u}_0(\mathbf{x}), \mathbf{u}'(\mathbf{x}, 0) = \mathbf{u}_1(\mathbf{x})\,, \tag{38}$$

where the initial data $\mathbf{u}_0$, $\mathbf{u}_1$ are given functions, and where $(\mathbf{u}'', \mathbf{v})$, $a^{(n)}(\mathbf{u}, \mathbf{v})$, $n = 0, 1$, $(\mathbf{f}, \mathbf{v})$ are defined above, and

$$j(\mathbf{v}) = \int_{\cup_{s,m} \Gamma_c^{sm}} g_c^{sm} [\mathbf{v}_t]^{sm} ds,$$

where the bilinear forms $a^{(n)}(\mathbf{u},\mathbf{v})$, $n=0,1$, are symmetric in $\mathbf{u}$, $\mathbf{v}$ and satisfy $a^{(n)}(\mathbf{u},\mathbf{u}) \geq c_0^{(n)}\|\mathbf{u}\|_{1,N}^2$, $c_0^{(n)} = \text{const} > 0$, $a^{(n)}(\mathbf{u},\mathbf{v}) \leq c_1^{(n)}\|\mathbf{u}\|_{1,N}\|\mathbf{v}\|_{1,N}$, $c_1^{(n)} = \text{const} > 0$, $\mathbf{u},\mathbf{v} \in V_0$.

The proof of the existence of the solution is based on the penalization and regularization techniques and is modification of that of [4], where the decomposition as above will be used.

3.3. Numerical solution

Let $\Omega = \cup_{\iota=1}^{r}(\Omega^\iota \cup \Gamma_{cv}^\iota)$ be approximated by $\Omega_h = \cup_{\iota=1}^{r}(\Omega_h^\iota \cup \Gamma_{cvh}^\iota)$ (a polygon in 2D and a polyhedron in 3D) with the boundary $\partial\Omega_h = \Gamma_{\tau h} \cup \Gamma_{uh} \cup \Gamma_{ch}$. Let $I=(0,t_p)$, $t_p>0$, let $m>0$ be an integer, then $\Delta t = t_p/m$, $t_i = i\Delta t$, $i=0,\ldots,m$. Let $\{\mathcal{T}_{h,\Omega_h}\}$ be a regular family of finite element partitions $\mathcal{T}_h$ of $\overline{\Omega}_h$ compatible to the boundary subsets $\overline{\Gamma}_{\tau h}$, $\overline{\Gamma}_{uh}$ and $\overline{\Gamma}_{ch}$. Let $V_h \subset V$ be the finite element space of linear elements corresponding to the partition $\mathcal{T}_h$, $K_h = V_h \cap K$ the set of continuous piecewise linear functions that vanish at the nodes of $\overline{\Gamma}_{uh}$ and whose normal components are nonpositive at the nodes on $\cup_{s,m}\Gamma_c^{sm}$; K_h is a nonempty, closed, convex subset of $V_h \subset V$. Let $\mathbf{u}_{0h} \in K_h$, $\mathbf{u}_{1h} \in V_h$ be approximations of $\mathbf{u}_0$ or $\mathbf{u}_1$. Let the end points $\overline{\Gamma}_{\tau h} \cup \overline{\Gamma}_{uh}$, $\overline{\Gamma}_{uh} \cup \overline{\Gamma}_{ch}$, $\overline{\Gamma}_{\tau h} \cup \overline{\Gamma}_{ch}$, coincide with the vertices of T_{hi}. Since the frictional term is assumed to be approximated by its value in the previous time level, the frictional term is approximated by a given friction limit. Then in every time level we have the following discrete problem:

Problem $(\mathcal{P})_h$: Find a displacement field $\mathbf{u}_h : \overline{I} \to V_h$ with $\mathbf{u}_h(0) = \mathbf{u}_{0h}$, $\mathbf{u}_h'(0) = \mathbf{u}_{1h}$, such that for a.e. $t \in I$, $\mathbf{u}_h(t) \in K_h$

$$
\begin{aligned}
&(\mathbf{u}_h''(t), \mathbf{v}_h - \mathbf{u}_h(t)) + a^{(0)}(\mathbf{u}_h(t), \mathbf{v}_h - \mathbf{u}_h(t)) + a^{(1)}(\mathbf{u}_h'(t), \mathbf{v}_h - \mathbf{u}_h(t)) + \\
&+ j(\mathbf{v}_h) - j(\mathbf{u}_h(t)) \geq (\mathbf{f}_h(t), \mathbf{v}_h - \mathbf{u}_h(t)) \quad \forall \mathbf{v}_h \in K_h, \quad \text{a.e. } t \in I\,,
\end{aligned} \tag{39}
$$

where

$$
\begin{aligned}
(\mathbf{u}_h'', \mathbf{v}_h) &= \sum_{\iota=1}^{r}(\mathbf{u}_h''^\iota, \mathbf{v}_h^\iota) = \int_{\Omega_h} \rho u_{hi}'' v_{hi} d\mathbf{x}\,, \\
a^{(n)}(\mathbf{u}_h, \mathbf{v}_h) &= \sum_{\iota=1}^{r} a^{(n)\iota}(\mathbf{u}_h^\iota, \mathbf{v}_h^\iota) = \int_{\Omega_h} c_{ijkl}^{(n)} e_{kl}(\mathbf{u}_h) e_{ij}(\mathbf{v}_h) d\mathbf{x}, \quad n=0,1, \\
(\mathbf{f}_h, \mathbf{v}_h) &= \sum_{\iota=1}^{r}(\mathbf{f}_h^\iota, \mathbf{v}_h^\iota) = \int_{\Omega_h} F_i v_{hi} d\mathbf{x} + \int_{\Gamma_{\tau_h}} P_i v_{hi} ds, \\
j(\mathbf{v}_h) &= \sum_{\iota=1}^{r} j^\iota(\mathbf{v}_h^\iota) = \int_{\cup_{s,m}\Gamma_{ch}^{sm}} g_{ch}^{sm} \left|[\mathbf{v}_{ht}]^{sm}\right| ds \equiv \langle g_{ch}^{sm}, \left|[\mathbf{v}_{ht}]^{sm}\right| \rangle_{\Gamma_{ch}^{sm}}\,.
\end{aligned}
$$

To prove the existence of discrete solution $\mathbf{u}_h$ the technique similar of that as in the continuous case, where the decomposition parallel as above, the penalty and regularization techniques are used.

3.4. Algorithm

The algorithm will be based on the semi-implicit scheme in time and the finite elements in space. Let $m > 0$ be an integer, then $\Delta t = t_p/m$, $t_i = i\Delta t$, $i = 0, 1, \ldots, m$. Approximating the derivatives by the differences, i.e., $\mathbf{u}_h'' = \frac{\mathbf{u}_h^{i+1} - 2\mathbf{u}_h^i + \mathbf{u}_h^{i-1}}{\Delta t^2}$, $\mathbf{u}_h' = \frac{\mathbf{u}_h^{i+1} - \mathbf{u}_h^i}{\Delta t}$, and setting $\mathbf{u}_h^i = \mathbf{u}_h(t_i)$, $\Delta\mathbf{u}_h^i = \mathbf{u}_h(t_i) - \mathbf{u}_h(t_{i-1})$, $\mathbf{u}_h^{i+1} \equiv \mathbf{u}_h$, $g_{ch}^{sm} = g_{ch}^{sm}(t_i) = \mathcal{F}_c^{sm}\left(\Delta t^{-1}[\Delta\mathbf{u}_{th}^i]^{sm}\right)\left|\tau_n^{sm}\left(\mathbf{u}_h^i, \frac{\Delta\mathbf{u}_h^i}{\Delta t}\right)\right|$, $(\mathbf{F}(t_{i+1}), \mathbf{v}_h) = \Delta t^2\,(\mathbf{f}_h(t_{i+1}), \mathbf{v}_h) + (2\mathbf{u}_h^i - \mathbf{u}_h^{i-1}, \mathbf{v}_h) + \Delta t a_h^{(1)}(\mathbf{u}_h^i, \mathbf{v}_h)$, $\mathbf{F}(t_{i+1}) \equiv \mathbf{f}_h$, then after some algebra in every time level $t = t_{i+1}$ we have to solve the following problem:

Problem $(\mathcal{P}_A)_h$**:** Find $\mathbf{u}_h \in K_h$, a.e. $t = t_{i+1} \in I$, such that

$$A(\mathbf{u}_h, \mathbf{v}_h - \mathbf{u}_h) + j(\mathbf{v}_h) - j(\mathbf{u}_h) \geq (\mathbf{f}_h, \mathbf{v}_h - \mathbf{u}_h), \quad \forall \mathbf{v}_h \in K_h,\ t = t_{i+1} \in I\,, \tag{40}$$

where

$$A(\mathbf{u}_h, \mathbf{v}_h) = (\mathbf{u}_h, \mathbf{v}_h) + \Delta t^2 a^{(0)}(\mathbf{u}_h, \mathbf{v}_h) + \Delta t a^{(1)}(\mathbf{u}_h, \mathbf{v}_h)\,,$$
$$j(\mathbf{v}_h) = \Delta t^2 \int_{\cup_{s,m}\Gamma_c^{sm}} g_{ch}^{sm}\,|[\mathbf{v}_{ht}]^{sm}|\,ds,$$

where g_{ch}^{sm} is the approximate given frictional limit. According to the above assumptions about the bilinear forms $a_h^{(n)}(\cdot,\cdot)$, $n = 0, 1$, and since $\rho \geq \rho_0 > 0$, then the bilinear form $A(\mathbf{u}_h, \mathbf{v}_h)$ is also symmetric in $\mathbf{u}_h$ and $\mathbf{v}_h$ and

$$A(\mathbf{u}_h, \mathbf{u}_h) \geq a_0 \|\mathbf{u}_h\|_{1,2}^2, \qquad a_0 = \text{const.} > 0,$$
$$|A(\mathbf{u}_h, \mathbf{v}_h)| \leq a_1 \|\mathbf{u}_h\|_{1,2}\|\mathbf{v}_h\|_{1,2}, \quad a_1 = \text{const.} > 0,\ \mathbf{u}_h, \mathbf{v}_h \in V_h,$$

hold.

The discretization error will be a function of the time step Δt and the mesh size h and thus the truncation error of the time and spatial discretization must tend to zero [1, 18, 20]. From the stability analysis for the critical time step size we have

$$\Delta t \leq \Delta t_{\text{crit}} = \gamma \frac{h^{(n)}}{\pi}\left(\frac{\rho^{(n)}}{E^{(n)}}\right), \tag{41}$$

where $h^{(n)}$ is the diameter of the corresponding (n)-th element, $h^{(n)} = c^{(n)}T_n$, T_n is the smallest period of the finite discretization with n degrees of freedom, $c^{(n)}$ is a dilatational wave velocity in the (n)-th material element, $\rho^{(n)}$ and $E^{(n)}$ are (average) values of the density and the Young modulus on the (n)-th element and γ is a reduction factor determined from the numerical experiments. Moreover, the algorithm is also consistent of order two, because the truncation error is of order Δt^2 in the displacements. Hence, the algorithm is convergent.

3.4.1. Mortar discretization

To give a saddle point formulation it is usually to introduce a Lagrange multiplier space $M = M_n \times M_t$, being the dual space of the trace space $W = \sqcap_s H^{\frac{1}{2},N}(\Gamma_c^s)$ (i.e., the trace space of V_0 restricted to $\cup_s \Gamma_c^s$) and its dual $W' = \sqcap_s H^{-\frac{1}{2},N}(\Gamma_c^s)$, assuming that Ω^ι, $\iota = 1, \ldots, r$, are domains with sufficiently smooth boundaries $\partial\Omega^\iota$, and the bilinear form $b(\cdot,\cdot)$ on the product space $V_0 \times M$. In the case if $\overline{\Gamma}_c^s = \cup_{\iota=1}^r (\partial\Omega^\iota \cap \Gamma_c^s) \setminus \Gamma_u^s$ we must use $H_{00}^{\frac{1}{2},N}(\Gamma_c^s)$ instead of $H^{\frac{1}{2},N}(\Gamma_c^s)$.

Let every polygonal domain Ω_h^ι, $\iota = 1, \ldots, r$, be covered by a triangulation $\mathcal{T}_{h,\Omega^\iota}$ in such a way that on the contact boundaries Γ_{ch}^{sm} the points of Γ_{ch}^s and Γ_{ch}^m are not identical, therefore, the mesh sizes $h_s \neq h_m$ and the global meshsize h is $h = \max_{\Omega_h}\{h_s, h_m\}$.

Let us introduce the discrete approximation of the Lagrange multiplier space $M_{hH} = M_{hn} \times M_{Ht}$, where

$$
\begin{aligned}
W_{hH}(\cup_s \Gamma_{ch}^s) =& W_{hn}(\cup_s \Gamma_{ch}^s) \times W_{Ht}(\cup_s \Gamma_{ch}^s) = \\
&= \left\{ \mathbf{v}_h^s \cdot \mathbf{n}^s|_{\cup_s \Gamma_{ch}^s}, \mathbf{v}_h \in V_h \right\} \times \left\{ \mathbf{v}_h^s \cdot \mathbf{t}^s|_{\cup_s \Gamma_{ch}^s}, \mathbf{v}_h \in V_h \right\}, \\
M_{hn} =& \Big\{ \mu_{hn} \in W_{hn}(\cup_s \Gamma_{ch}^s), \int_{\Gamma_c^s} \mu_{hn} \psi_h ds \geq 0, \\
&\forall \psi_h \in W_{hn},\ \psi_n \geq 0 \text{ a.e. on every } \Gamma_{ch}^s \Big\},
\end{aligned}
$$

$$
\begin{aligned}
M_{Ht} =& \Big\{ \mu_{Ht} \in W_{Ht}(\cup_s \Gamma_{ch}^s), \int_{\Gamma_{ch}^s} \boldsymbol{\mu}_{Ht} \boldsymbol{\psi}_H ds - \int_{\Gamma_{ch}^s} g_{ch}^{sm} |\boldsymbol{\psi}_H| ds \leq 0, \\
&\forall \boldsymbol{\psi}_H \in W_{Ht}(\cup_s \Gamma_{ch}^s) \Big\},
\end{aligned}
$$

Let

$$
\begin{aligned}
b(\boldsymbol{\mu}_{hH}, \mathbf{v}_h) = \langle \mu_{hn}, [\mathbf{v}_h \cdot \mathbf{n}]^s - d^{sm} \rangle_{\cup \Gamma_{ch}^s} + \int_{\cup_s \Gamma_{ch}^s} g_{ch}^{sm} \boldsymbol{\mu}_{Ht} \cdot [\mathbf{v}_{ht}]^s ds, \\
\boldsymbol{\mu}_{hH} \in M_{hH},\ \mathbf{v}_h \in V_{0h},
\end{aligned}
$$

where $[\mathbf{v}_h \cdot \mathbf{n}]^{sm} = v_{hn}^s(\mathbf{x},t) - v_{hn}^m(\mathcal{R}^{sm}(\mathbf{x},t))$, $[\mathbf{v}_{ht}]^{sm} = \mathbf{v}_{ht}^s(\mathbf{x},t) - \mathbf{v}_{ht}^m(\mathcal{R}^{sm}(\mathbf{x},t))$, where $\mathcal{R}^{sm}: \Gamma_{ch}^s(t) \mapsto \Gamma_{ch}^m(t)$, at $t \in I$, is a bijective map satisfying $\Gamma_{ch}^m(t) \subset \mathcal{R}^{sm}(\Gamma_{ch}^s(t))$, $t \in I$, and where $\langle \cdot, \cdot \rangle_{\Gamma_{ch}^s}$ denotes the duality pairing between W_{hH} and M_{hH}.

Then we have the following problem:

Problem $(\mathcal{P})_h$**:** In every time level find $(\boldsymbol{\lambda}_{hH}, \mathbf{u}_h) \in M_{hH} \times V_h$ satisfying

$$
\begin{aligned}
&A(\mathbf{u}_h, \mathbf{v}_h) + b(\boldsymbol{\lambda}_{hH}, \mathbf{v}_h) = (\mathbf{f}_h, \mathbf{v}_h) && \forall \mathbf{v}_h \in V_h = \sqcap_{\iota=1}^r V_h^\iota,\ t \in I, \\
&b(\boldsymbol{\mu}_{hH} - \boldsymbol{\lambda}_{hH}, \mathbf{v}_h) \leq \langle d^{sm}, \mu_{hn} - \lambda_{hn} \rangle_{\cup \Gamma_{ch}^s} && \forall \boldsymbol{\mu}_{hH} \in M_{hH},\ t \in I. \qquad (42)
\end{aligned}
$$

For the existence and uniqueness it is necessary to ensure that $\{\boldsymbol{\mu}_{hH} \in M_{hH}, b(\boldsymbol{\mu}_{hH}, \mathbf{v}_h) = 0, \forall \mathbf{v}_h \in V_h\} = \{\emptyset\}$.

Proposition 1 *Let $-\tau_n(\mathbf{u}) \in M_{hn}$. Then the problem (42) has a unique solution $(\boldsymbol{\lambda}_{hH}, \mathbf{u}_h) \in M_{hH} \times V_h$, a.e. $t \in I$. Moreover, we have*

$$\lambda_{hn}^s = -\tau_n^s(\mathbf{u}_h) \quad \textit{and} \quad g_c^s \boldsymbol{\lambda}_{Ht}^s = -\boldsymbol{\tau}_t^s(\mathbf{u}_h)\,,$$

where $\mathbf{u}_h$ is the solution of the discrete primal problem and $g_c^s \equiv g_{ch}^{sm}$.

3.4.2. Matrix formulation and the PDAS method

As usual in the mortar approach the contact boundary Γ_{ch}^{sm} has two sides, the "slave" side from the Ω_h^s and the "master" side from the Ω_h^m. The contact boundaries Γ_{ch}^{sm} are assumed to be a union of faces in the 3D case.

Let us assume that the space V is approximated by the discrete finite element space V_h of linear elements corresponding to the partition $\mathcal{T}_h$ and let $V_h = V_h^s \times V_h^m \subset V$ be introduced by such a way that the nodal basis functions on the mortar side will be biorthogonal with respect to the piecewise linear basis on the slave side.[1]

In the mortar approach, the Lagrange multiplier space is approximated by its $(N-1)$-dimensional mesh resulting from the N-dimensional triangulation on the slave side. In this case the discontinuous piecewise linear nodal basis functions for the dual Lagrange multiplier will be used. The discrete Lagrange multiplier space M_{hH} can be spanned as $M_{hH}^s = \mathrm{span}\{\psi_i \mathbf{e}_k,\, i = 1, \ldots, n_c,\, k = 1, \ldots, N\}$, $s \in \{1, \ldots, r\}$, where ψ_i is the i-th scalar dual basis function, $\mathbf{e}_k$ is the k-th unit vector, i.e., components of the unit Cartesian basis, n_c is the number of nodes on the slave side of $\overline{\Gamma}_{ch}^s$, i.e., the number of freedom of the space M_{hH} in each component.

Let W_{hH}^s be the vector valued trace space of V_{0h} restricted to $\cup_s \Gamma_{ch}^s$. Then for each $\mathbf{v}_h = \sum_i \gamma_i \varphi_i \in W_{hH}$ the discrete scalar product $\mathbf{v}_h \cdot \mathbf{n}_h^s = \sum_i (\gamma_i \cdot \mathbf{n}_i^s)\varphi_i$, where $\mathbf{n}_i^s$ denotes the outer normal of Ω^s at the node i. Similarly, for each $\boldsymbol{\mu}_h = \sum_i \boldsymbol{\alpha}_i \psi_i \in M_{hH}$ the discrete product $\boldsymbol{\mu}_h \cdot \mathbf{n}_h^s = \sum_i (\boldsymbol{\alpha}_i \cdot \mathbf{n}_i^s)\psi_i$. Let us define

$$M_{hH}^{s+} := \left\{ \boldsymbol{\mu}_{hH} \in M_{hH}^s \middle| \, \langle \boldsymbol{\mu}_{hH}, \mathbf{v}_h \rangle \geq 0, \mathbf{v}_h \in W_h^{s+} \right\},$$

where

$$W_h^{s+} := \{ \mathbf{v}_h \in W_{hH}(\cup \Gamma_{ch}^s) | \mathbf{v}_h \cdot \mathbf{n}_h^s \geq 0 \}$$

and

$$W_{hH} = W_{hH}(\cup_s \Gamma_{ch}^s) = W_{hn}(\cup_s \Gamma_{ch}^s) \times W_{Ht}(\cup_s \Gamma_{ch}^s) =$$

$$= \{ \mathbf{v}_h \cdot \mathbf{n}^s |_{\cup \Gamma_{ch}^s}, \mathbf{v}_h \in V_h \} \times \{ \mathbf{v}_h \cdot \mathbf{t}^s |_{\cup \Gamma_{ch}^s}, \mathbf{v}_h \in V_h \}$$

[1] Let $\{\psi_i\}_{i=1}^m$ be a suitable dual basis and $\{\varphi_j\}_{j=1}^m$ be the standard piecewise linear basis on the slave side, i.e., the basis of $W_{hH}(\Gamma_{sh}^{sm})$. We say that ψ_i and φ_j are biorthogonal if $\int_{\Gamma_{ch}^{sm}} \varphi_j \psi_i \, ds = \delta_{ij} \int_{\Gamma_{ch}^{sm}} \varphi_j \, ds$, $1 \leq i, j \leq m$, δ_{ij} is the Kronecker delta.

It can be shown ([20]) that M_{hH}^{s+} can be written as

$$M_{hH}^{s+} := \left\{ \boldsymbol{\mu}_{hH} = \sum_{i=1}^{m} \boldsymbol{\alpha}_i \psi_i | \boldsymbol{\alpha}_i \in \mathbb{R}^N, \boldsymbol{\alpha}_i = \alpha_i^n \mathbf{n}_i^s, \alpha_i^n \in \mathbb{R}, \alpha_i^n \geq 0,\ i \leq m \right\} .$$

Finally,

$$M_{hH}^{+} = \prod_s M_{hH}^{s+} ,$$
$$b(\boldsymbol{\mu}_{hH}, \mathbf{v}_h) = \langle \boldsymbol{\mu}_{hH}, [\mathbf{v}_h]^s \rangle_{\cup \Gamma_{ch}^s} .$$

For completeness, the discrete convex subset $K_h \subset V_h$ will be then defined as

$$K_h := \left\{ \mathbf{v}_h \in V_h | b(\boldsymbol{\mu}_{hH}, \mathbf{v}_h) \leq \int_{\cup \Gamma_{ch}^s} d_h^{sm} (\boldsymbol{\mu}_{hH} \cdot \mathbf{n}_h^s) ds,\ \boldsymbol{\mu}_{hH} \in M_{hH}^{+} \right\} ,$$

where d_h^{sm} is a suitable approximation of d^{sm} on W_{hH}.

Then the discrete mortar formulation of the saddle point problem for every time level is defined as follows:

Problem $(\mathcal{P}_{sp})_{dm}$**:** In every time level find $\mathbf{u}_h \in V_h$, $\boldsymbol{\lambda}_{hH} \in M_{hH}^{+}$, a.e. $t \in I$, $\boldsymbol{\lambda}_{hH} = (\lambda_{hn}, \boldsymbol{\lambda}_{Ht})$, satisfying

$$\begin{aligned} A(\mathbf{u}_h, \mathbf{v}_h) + b(\boldsymbol{\lambda}_{hH}, \mathbf{v}_h) &= (\mathbf{f}_h, \mathbf{v}_h) \quad \forall \mathbf{v}_h \in V_h,\ t \in I , \\ b(\boldsymbol{\mu}_{hH} - \boldsymbol{\lambda}_{hH}, \mathbf{v}_h) &\leq \langle d^{sm}, (\boldsymbol{\mu}_{hH} - \boldsymbol{\lambda}_{hH}) \cdot \mathbf{n}_h \rangle_{\cup \Gamma_{ch}^s} \quad \forall \boldsymbol{\mu}_{hH} \in M_{hH}^{+},\ t \in I, \end{aligned} \tag{43}$$

where

$$\begin{aligned} A(\mathbf{u}_h, \mathbf{v}_h) =& (\mathbf{u}_h, \mathbf{v}_h) + \Delta t^2 a^{(0)}(\mathbf{u}_h, \mathbf{v}_h) + \Delta t a^{(1)}(\mathbf{u}_h, \mathbf{v}_h) , \\ b(\boldsymbol{\mu}_{hH}, \mathbf{v}_h) =& \langle \boldsymbol{\mu}_{hH}, [\mathbf{v}_h]^s \rangle_{\cup \Gamma_{ch}^s} \quad \forall \mathbf{v}_h \in V_h,\ \boldsymbol{\mu}_{hH} \in M_{hH} . \end{aligned}$$

Let us decompose the set of all vertices of triangulation $\mathcal{T}_h = \cup_{i=1}^r \mathcal{T}_h^i$ into three disjoint parts $\mathcal{N}, \mathcal{M}, \mathcal{S}$, where $\mathcal{S}$ is a set of vertices on all $\mathcal{T}_h^s \cap \Gamma_{ch}^{sm}$, and $\mathcal{M}$ is a set of vertices on all $\mathcal{T}_h^m \cap \Gamma_{ch}^{sm}$, and $\mathcal{N}$ is a set of all the other one. The strong formulation of the non-penetration condition will be replaced by its weak discrete form

$$\int_{\cup \Gamma_{ch}^{sm}} [\mathbf{u}_h \cdot \mathbf{n}]^s \psi_p ds \leq \int_{\cup \Gamma_{ch}^{sm}} d_h^s \psi_p ds, \quad p \in \mathcal{S} , \tag{44}$$

that coupled the vertices on the slave side and the master side. Using a transformation of the basis of the space V_h in such a way that the weak non-penetration condition (44) in the new basis only deals with the vertices on the slave side. Moreover, the elimination of the Lagrange multiplicators $\boldsymbol{\Lambda}_{hH}$ can be easy made (see [11, 18, 20]). In this new basis the first equation of Problem $(\mathcal{P}_{sp})_{dm}$ for every $t \in I$ will be

expressed in the matrix form, that with respect to the sets $\mathcal{N}, \mathcal{M}, \mathcal{S}$, after using the modified basis bellow defined, after some modification ([18, 20]), we obtain the modified system

$$\hat{\mathbb{A}}_h \hat{\mathbf{U}}_h + \hat{\mathbb{B}}_h \hat{\mathbf{\Lambda}}_{hH} = \hat{\mathbb{F}}_h \,, \tag{45}$$

where $\hat{\mathbf{U}}_h$ is the displacement vector of nodal parameter with respect to the modified basis Φ, and where the modified stiffness matrix is of the form

$$\hat{\mathbb{A}}_h = Q\mathbb{A}_h Q^T = \begin{bmatrix} \mathbb{A}_{\mathcal{NN}} & \mathbb{A}_{\mathcal{NM}} + \mathbb{A}_{\mathcal{NS}}\hat{\mathbb{M}} & \mathbb{A}_{\mathcal{NS}} \\ \mathbb{A}_{\mathcal{MN}} + \hat{\mathbb{M}}^T\mathbb{A}_{\mathcal{SN}} & \mathbb{A}_{\mathcal{MM}} + \mathbb{A}_{\mathcal{MS}}\hat{\mathbb{M}} + \\ & +\hat{\mathbb{M}}^T\mathbb{A}_{\mathcal{SM}} + \hat{\mathbb{M}}^T\mathbb{A}_{\mathcal{SS}}\hat{\mathbb{M}} & \mathbb{A}_{\mathcal{MS}} + \hat{\mathbb{M}}^T\mathbb{A}_{\mathcal{SS}} \\ \mathbb{A}_{\mathcal{SN}} & \mathbb{A}_{\mathcal{SM}} + \mathbb{A}_{\mathcal{SS}}\hat{\mathbb{M}} & \mathbb{A}_{\mathcal{SS}} \end{bmatrix}$$

and the vector $\hat{\mathbb{F}}_h$ is of the form

$$\hat{\mathbb{F}}_h = Q\mathbb{F}_h = (\mathbb{F}_{\mathcal{N}},\, \mathbb{F}_{\mathcal{M}} + \hat{\mathbb{M}}^T\,\mathbb{F}_{\mathcal{S}},\, \mathbb{F}_{\mathcal{S}})^T \,,$$

$\hat{\mathbb{B}}_h = Q.(\mathbf{0}, -\mathbb{M}^T, \mathbf{0})^T = (\mathbb{O}, \mathbb{O}, \mathbb{D})^T$, and $\hat{\mathbb{M}}^T = \mathbb{D}^{-1}\mathbb{M}$, $\mathbb{M} = (\mathbb{M}[p,q])$, where $\mathbb{M}[p,q] = \int_{\cup\Gamma_{ch}^{sm}} \varphi_p\psi_q ds\, \mathbb{I}_3$, $p, \in \mathcal{S}, q \in \mathcal{M}$, and $\mathbb{D} = (\mathbb{D}[p,q])$, $\mathbb{D}[p,q] = \delta_{pq}\mathbb{I}_3 \cdot \int_{\cap\Gamma_{ch}^{sm}} \varphi_p\psi_q\, ds$, $p = q \in \mathcal{S}$, and where the used modified basis is of the form

$$\Phi = (\Phi_{\mathcal{N}}, \Phi_{\mathcal{M}}, \Phi_{\mathcal{S}}) = Q\varphi = \begin{bmatrix} \mathbb{I}_N & \mathbb{O} & \mathbb{O} \\ \mathbb{O} & \mathbb{I}_N & \hat{\mathbb{M}}^T \\ \mathbb{O} & \mathbb{O} & \mathbb{I}_N \end{bmatrix} \begin{bmatrix} \varphi_{\mathcal{N}} \\ \varphi_{\mathcal{M}} \\ \varphi_{\mathcal{S}} \end{bmatrix}.$$

If the displacement $\hat{\mathbb{U}}_h$ is known, then the Lagrange multiplier can be computed directly from (45) and then

$$\hat{\mathbf{\Lambda}}_{hH} = \mathbb{D}^{-1}(\hat{\mathbb{F}}_h - \hat{\mathbb{A}}_h\hat{\mathbb{U}}_h)_S. \tag{46}$$

The algebraic representation of the weak nonpenetration condition is associated with the transformed basis Φ is of the form ([10],[12])

$$\hat{\mathbb{U}}_{hn,p} \equiv (\mathbf{n}_p^s)^T\mathbb{D}[p,p]\hat{\mathbb{U}}_{hp} \leq d_p^{sm} \quad \forall p \in \mathcal{S}\,, \tag{47}$$

where $d_p^{sm} = \int_{\cup_s\Gamma_c^s} d_h^{sm}\psi_p ds$, $p \in \mathcal{S}$, and the coefficients at $\hat{\mathbb{U}}_{hq}$, $q \in \mathcal{M}$, are nullified.

Thus, in every time level, we will solve the following problem

$$\begin{aligned} &\hat{\mathbb{A}}_h\hat{\mathbb{U}}_h + \hat{\mathbb{B}}_h\mathbf{\Lambda}_{hH} = \hat{\mathbb{F}}_h, \\ &\hat{\mathbb{U}}_{hn,p} \leq d_p^{sm}, \mathbf{\Lambda}_{hn,p} \geq 0, (\hat{\mathbb{U}}_{hn,p} - d_p^{sm})\mathbf{\Lambda}_{hn,p} = 0, \quad \forall p \in \mathcal{S},\ t \in I, \end{aligned} \tag{48}$$

where the second line represents the Karush-Kuhn-Tucker conditions of a constrained optimization problem for inequality constraints, with the discrete Tresca friction conditions and with the discrete friction conditions

$$\begin{aligned} &\left|\mathbf{\Lambda}_{Ht,p}^s(p)\right| \leq g_p^s \ (= \mathcal{F}_c^{sm}\left|\mathbf{\Lambda}_{hn,p}^s\right|), \\ &\left|\mathbf{\Lambda}_{Ht,p}^s(p)\right| < g_p^s \ (= \mathcal{F}_c^{sm}\left|\mathbf{\Lambda}_{hn,p}^s\right|) \Rightarrow \mathbf{u}_{ht,p} = \mathbf{0}, \end{aligned}$$

$$\left|\mathbf{\Lambda}^s_{Ht,p}(p)\right| = g^s_p \ (= \mathcal{F}^{sm}_c \left|\mathbf{\Lambda}^s_{hn,p}\right|) \Rightarrow \exists \vartheta \geq 0$$
$$\text{such that} \quad \mathbf{\Lambda}^s_{Ht,p} = -\vartheta \mathbf{u}_{ht,p}, \quad \text{for all } p \in S, \tag{49}$$

where for the Tresca friction model $\mathcal{F}^{sm}_c|\mathbf{\Lambda}^s_{hn,p}| \equiv g^s_p \in H^{-\frac{1}{2}}(\Gamma^s_c)$, $g^s_p \geq 0$, $g^s_p = \int_{\Gamma^s_{ch}} g^s_{ch}\varphi_p ds$, and where

$$\mathbf{\Lambda}_{hn,p} = \mathbf{n}^{sT}_p \mathbb{D}[p,p]\mathbf{\Lambda}_{hH}(p), \quad \mathbf{\Lambda}_{hH}(p) \in \mathbb{R}^N,$$
$$\mathbf{\Lambda}_{Ht,p} = \mathbf{\Lambda}_{hH}(p) - (\mathbf{\Lambda}_{hH}(p)\cdot \mathbf{n}^s_p)\mathbf{n}^s_p = (\mathbf{\Lambda}_{hH}(p)\cdot \mathbf{t}^s_p)\mathbf{t}^s_p.$$

For $g^s_p = 0$ the condition (49) leads to homogeneous Neumann boundary conditions in tangential direction.

PDAS algorithm for the 3D case with friction of Tresca type. In the 3D model with the Tresca friction if the displacements $\mathbf{u}_h$ are known, the Lagrange multiplier $\mathbf{\Lambda}_{hH} = (\Lambda_{hn}, \mathbf{\Lambda}_{Ht})$ can be computed directly from (49a), that is,

$$\mathbf{\Lambda}_{hH} = \mathbb{D}^{-1}(\hat{\mathbb{F}}_h - \hat{\mathbb{A}}_h\hat{\mathbb{U}})_S, \tag{50}$$

where the subscript S denotes that we use only the entries of the vector corresponding to the nodes $p \in S$. For the normal and tangential components of the multiplier $\mathbf{\Lambda}_{hH}$ and of the relative decomposition $\mathbf{u}_h$ for a node point $p \in S$, we have

$$\hat{\mathbb{U}}_{hn,p} = \hat{\mathbb{U}}^T_p \mathbf{n}_p \in \mathbb{R}, \quad \hat{\mathbb{U}}_{Ht,p} = (\hat{\mathbb{U}}^T_p \mathbf{t}_{1p}, \quad \hat{\mathbb{U}}^T_p \mathbf{t}_{2p})^T \in \mathbb{R}^2,$$
$$\mathbf{\Lambda}^s_{hn,p} = (\mathbf{n}^s_p)^T \mathbb{D}[p,p]\mathbf{\Lambda}_{hH}(p) \in \mathbb{R}, \quad \mathbf{\Lambda}_{hH}(p) \in \mathbb{R}^3,$$
$$\mathbf{\Lambda}^s_{Ht,p} = \mathbf{\Lambda}_{hH}(p) - (\mathbf{\Lambda}_{hH}(p)\cdot \mathbf{n}^s_p)\mathbf{n}^s_p = \left(\mathbf{\Lambda}_{hH}(p)\cdot \mathbf{t}^s_p\right)\mathbf{t}^s_p =$$
$$= \left(\mathbf{\Lambda}^T_{hH}(p)\mathbb{D}[p,p]\mathbf{t}^s_{1,p}, \ \mathbf{\Lambda}^T_{hH}(p)\mathbb{D}[p,p]\mathbf{t}^s_{2,p}\right)^T \in \mathbb{R}^2.$$

Let $g^s_p > 0$, then the condition (49) is equivalent to $C_t\left(\hat{\mathbb{U}}_{t,p}, \mathbf{\Lambda}^s_{Ht,p}\right) = 0$ for all $p \in S$, where

$$C_t\left(\hat{\mathbb{U}}_{ht,p}, \mathbf{\Lambda}^s_{Ht,p}\right) = \max\left(g^s_{ch,p}, |\mathbf{\Lambda}^s_{Ht,p} + c_2\hat{\mathbb{U}}_{ht,p}|\right)\mathbf{\Lambda}^s_{Ht,p} - g^s_p\left(\mathbf{\Lambda}^s_{Ht,p} + c_2\hat{\mathbb{U}}_{ht,p}\right), \ c_2 > 0, \tag{51}$$

which will be a starting point of the algorithm, that will be based on a Newton-type algorithm for the solution of $C_t(\hat{\mathbb{U}}_{ht,p}, \mathbf{\Lambda}^s_{Ht,p}) = 0$. As it was shown in [7] the max-function and the Euclidean norm are semi-smooth, and therefore, a semi-smooth Newton method can be used. If the Euclidean norm $\left|\mathbf{\Lambda}^s_{Ht,p} + c_2\hat{\mathbb{U}}_{ht,p}\right| = 0$, then $\max\left(g^s_{ch,p}, |\mathbf{\Lambda}^s_{Ht,p} + c_2\hat{\mathbb{U}}_{ht,p}|\right) = g^s_{ch,p}$ and the Euclidean norm vanishes. Hence, the derivative of the Euclidean norm only occurs for points that are differentiable in the classical sense. The analysis of the generalized derivative of $C_t(\hat{\mathbb{U}}_{ht,p}, \mathbf{\Lambda}^s_{Ht,p})$ (see [12])

shows that the nodes of $\mathcal{S}$ are separated into the active set $\mathcal{A}_{Ht,k}$ and the inactive set $\mathcal{I}_{Ht,k}$, where

$$\mathcal{A}_{Ht,k} := \left\{ p \in S; |\mathbf{\Lambda}_{Ht,p}^{s,k-1} + c_2 \hat{\mathbb{U}}_{ht,p}^{k-1}| - g_{ch,p}^{s} > 0 \right\}, \tag{52}$$

$$\mathcal{I}_{Ht,k} := \left\{ p \in S; |\mathbf{\Lambda}_{Ht,p}^{s,k-1} + c_2 \hat{\mathbb{U}}_{ht,p}^{k-1}| - g_{ch,p}^{s} \leq 0 \right\}. \tag{53}$$

Since $\hat{\mathbb{B}}_h = (\mathbb{O}, \mathbb{O}, \mathbb{D})^T$, we decompose the matrix $\mathbb{D}$ into

$$\mathbb{D} = \begin{bmatrix} \mathbb{D}_{\mathcal{I}_k} & \mathbb{O} \\ \mathbb{O} & \mathbb{D}_{\mathcal{A}_k} \end{bmatrix}, \text{ since } \mathcal{S} = \mathcal{A}_k \cup \mathcal{I}_k.$$

This decomposition of nodes of S into the active $\mathcal{A}_{Ht,k}$ and inactive $\mathcal{I}_{Ht,k}$ sets is provided by the characteristic function in the generalized derivative of $C_t(\cdot,\cdot)$. The case if $g_{ch,p} = 0$ is in details analyzed in [12].

Summing all results for the frictionless contact problem and for the Tresca friction case, then Problem $(\mathcal{P})$ can be rewritten as

$$\begin{aligned} \hat{\mathbb{A}}_h \hat{\mathbb{U}}_h + \hat{\mathbb{B}}_h \mathbf{\Lambda}_{hH} &= \hat{\mathbb{F}}_h, \\ C_n \left(\hat{\mathbb{U}}_{hn,p}, \mathbf{\Lambda}_{hn,p} \right) &= 0, \\ C_t(\hat{\mathbb{U}}_{ht,p} \mathbf{\Lambda}_{Ht,p}) &= 0 \end{aligned} \tag{54}$$

for all vertices $p \in \mathcal{S}$ and $t \in I$.

The PDAS algorithm for the contact problem with friction in the Tresca sense is as follows:

Algorithm $(\mathcal{T})$:

STEP 1: Initiate the active sets $\mathcal{A}_{hn,1}$, $\mathcal{A}_{Ht,1}$ and the inactive sets $\mathcal{I}_{hn,1}$, $\mathcal{I}_{Ht,1}$ such that $\mathcal{S}_n = \mathcal{A}_{hn1} \cup \mathcal{I}_{hn1}$, $\mathcal{S}_t = \mathcal{A}_{Ht1} \cup \mathcal{I}_{Ht1}$, $\mathcal{A}_{hn1} \cap \mathcal{I}_{hn1} = \emptyset$, $\mathcal{A}_{Ht1} \cap \mathcal{I}_{Ht1} = \emptyset$ and introduce the initial value $(\hat{\mathbb{U}}^0, \mathbf{\Lambda}_{hH}^0)$, $c_1, c_2 \in (10^3, 10^4)$ and set $k = 1$, $c_1 > 0$, $c_2 > 0$, $m \in \mathbb{N}$.

STEP 2: Define the active and inactive sets

$$\begin{aligned} \mathcal{A}_{hn,k} &:= \left\{ p \in S; \mathbf{\Lambda}_{hn,p}^{s,k-1} + c_1 \left(\hat{\mathbb{U}}_{n,p}^{k-1,m} - d_p^{sm} \right) > 0 \right\}, \\ \mathcal{I}_{hn,k} &:= \left\{ p \in S; \mathbf{\Lambda}_{hn,p}^{s,k-1} + c_1 \left(\hat{\mathbb{U}}_{n,p}^{k-1,m} - d_p^{sm} \right) \leq 0 \right\}, \\ \mathcal{A}_{Ht,k} &:= \left\{ p \in S; \left| \mathbf{\Lambda}_{Ht,p}^{s,k-1} + c_2 \hat{\mathbb{U}}_{t,p}^{k-1,m} \right| - g_{ch,p}^{s} > 0 \right\}, \\ \mathcal{I}_{Ht,k} &:= \left\{ p \in S; \left| \mathbf{\Lambda}_{Ht,p}^{s,k-1} + c_2 \hat{\mathbb{U}}_{t,p}^{k-1,m} \right| - g_{ch,p}^{s} \geq 0 \right\}, \end{aligned}$$

STEP 3: For $i = 1, \ldots, m$, compute the generalized derivative in the sense of a semi-smooth Newton method, i.e.,

$$\hat{\mathbb{U}}_{hH}^{k,i} = G \left(\hat{\mathbb{U}}_{hH}^{k,i-1}, \mathcal{A}_{hn,k}, \mathcal{I}_{hn,k}, \mathcal{A}_{Ht,k}, \mathcal{I}_{Ht,k}, \hat{\mathbb{U}}_{hH}^{k-1,m}, \mathbf{\Lambda}_{hH}^{k-1} \right),$$

where by the symbol G we denote the generalized derivative in the sense of a semi-smooth Newton method.

STEP 4: **If** $\left|\hat{\mathbb{U}}_{hH}^{k,m} - \hat{\mathbb{U}}_{hH}^{k,0}\right| / \left|\hat{\mathbb{U}}_{hH}^{k,m}\right| < \varepsilon$ **then** STOP.

STEP 5: Compute the Lagrange multiplier due to (51), that is,

$$\mathbf{\Lambda}_{hH,k} = \mathbb{D}^{-1}\left(\hat{\mathbb{F}}_{hS} - \hat{\mathbb{A}}_{hS}\hat{\mathbb{U}}_{hH}^{k,m}\right) .$$

STEP 6: Set $\hat{\mathbb{U}}_{hH}^{k+1,0} = \hat{\mathbb{U}}_{hH}^{k,m}$, $k = k+1$ and **goto STEP 2.**

PDAS algorithm for the 3D case with Coulomb friction. The algorithms can be based on the fixpoint algorithm or on the full Newton method ([12]). We limit ourselves to the fixpoint algorithm only.

The Fixpoint Algorithm ($\mathcal{FP}$) is the extension of the above PDAS algorithm for the Tresca friction, in which the friction bound $g_{ch,p}^{s} = \mathcal{F}_c^{sm}|\mathbf{\Lambda}_{n,p}^{s}|$ is iteratively modified using the normal component of the Lagrange multiplier. Thus, we have the following algorithm, that the friction bound and the active and inactive sets are updated in every step.

Algorithm ($\mathcal{FP}$):

STEP 1: Initiate the initial value $\left(\hat{\mathbb{U}}^{0,0}, \mathbf{\Lambda}_{hH}^{0}\right)$, $c_1, c_2 \in (10^3, 10^4)$ and set $k = 1$, $k_0 \in \mathbb{N}$, $m \in \mathbb{N}$.

STEP 2: If $mod_{k_0}(k-1) = 0$, set $k_c = k-1$ and update the friction bound by $g_{ch,p}^{s,k_c} = \mathcal{F}_c^{sm} \max\{0, \mathbf{\Lambda}_{n,p}^{s,k_c}\}$, $p \in S$.

STEP 3: Define the active sets $\mathcal{A}_{hn,k}$, $\mathcal{A}_{Ht,k}$ and the inactive sets $\mathcal{I}_{hn,k}$, $\mathcal{I}_{Ht,k}$ by

$$\begin{aligned}
\mathcal{A}_{hn,k} &:= \left\{p \in S; \mathbf{\Lambda}_{hn,p}^{s,k-1} + c_1\left(\hat{\mathbb{U}}_{n,p}^{k-1,m} - d_p^{sm}\right) > 0\right\} , \\
\mathcal{I}_{hn,k} &:= \left\{p \in S; \mathbf{\Lambda}_{hn,p}^{s,k-1} + c_1\left(\hat{\mathbb{U}}_{n,p}^{k-1,m} - d_p^{sm}\right) \leq 0\right\} , \\
\mathcal{A}_{Ht,k} &:= \left\{p \in S; \left|\mathbf{\Lambda}_{Ht,p}^{s,k-1} + c_2\hat{\mathbb{U}}_{t,p}^{k-1,m}\right| - g_{ch,p}^{s,k_c} > 0\right\} , \\
\mathcal{I}_{Ht,k} &:= \left\{p \in S; \left|\mathbf{\Lambda}_{Ht,p}^{s,k-1} + c_2\hat{\mathbb{U}}_{t,p}^{k-1,m}\right| - g_{ch,p}^{s,k_c} \leq 0\right\} .
\end{aligned}$$

STEP 4: For $i = 1, \ldots, m$, compute the generalized derivative in the sense of a semi-smooth Newton method

$$\hat{\mathbb{U}}_{hH}^{k,i} = G\left(\hat{\mathbb{U}}_{hH}^{k,i-1}, \mathcal{A}_{hn,k}, \mathcal{I}_{hn,k}, \mathcal{A}_{Ht,k}, \mathcal{I}_{Ht,k}, \hat{\mathbb{U}}_{hH}^{k-1,m}, \mathbf{\Lambda}_{hH}^{k-1}\right) ,$$

where the symbol G has the same meaning as above.

STEP 5: Compute the Lagrange multiplier due to (51) as

$$\mathbf{\Lambda}_{hH}^{k} = \mathbb{D}^{-1}\left(\hat{\mathbf{F}}_{hS} - \hat{\mathbb{A}}_{hS}\hat{\mathbb{U}}_{hH}^{k,m}\right) .$$

STEP 6: If $\left\|\hat{\mathbb{U}}_{hH}^{k,m} - \hat{\mathbb{U}}_{hH}^{k_c,m}\right\| / \left\|\hat{\mathbb{U}}_{hH}^{k,m}\right\| < \varepsilon$ **then** STOP.

STEP 7: Set $\hat{\mathbb{U}}_{hH}^{k+1,0} = \hat{\mathbb{U}}_{hH}^{k,m}$ and $k = k+1$ and **goto** STEP 2.

If $m = \infty$, we obtain the exact version of the algorithm, in the previous case we speak about inexact algorithm. The algorithm is convergent for small coefficient of friction (see [6]).

3.5. Fracture of bones with neoplasms

With a persistent growth of the neoplasms, the possibility of fracture rises can be expected. Firstly, in locations with highest stresses the crack initiations can be occurred (Fig. 1a,b,c), and with continuous loading the cracks start to opening and propagate up-to the moment when the bone is fractured. In the real situations it is very difficult to determine the location of a crack, its initiation, its further opening and propagation and to determine the direction of its future propagation.

The geometry of the investigated system of bones with neoplasms is determined from the CT or MRI scan data. The locations of the acting contraction forces and their directions will be determined from the anatomy knowledges and their magnitudes (in N) will be determined from the cross-sectional area of the muscles (in mm^2), the averaged activation ratio, and a certain constant (in N/mm^2). On the bases of these CT or MRI data the finite element mesh will be generated. The contact boundaries will be approximated by such a way that the contact boundary is discretized from the both sides corresponding to the neighboring subdomains Ω^s and Ω^m, from the slave side and the master side, and then the unilateral contact conditions will be satisfied in all vertices of $\mathcal{T}_h^s \cap \Gamma_{ch}^{sm}$ from the slave side and in all vertices of $\mathcal{T}_h^m \cap \Gamma_{ch}^{sm}$ from the master side.

To determine the areas of possible fracture zones, we firstly determine the areas with maximal principle stresses, and therefore, the places where cracks are initiated. Thus we need to check, at each time step, when the crack is started to propagate and in which direction. In the first case the crack propagation criteria will be used, while in the second one the crack kinking criteria will be used. When a crack further propagate, the accuracy at the crack tip will be of great importance for determination of a possible fracture. Many numerical tools were developed to improve the accuracy at the crack tip. Since the stress field is singular in the vicinity of the crack tip, a concentric mesh around the crack tip can be coupled with singular elements, which can be used to model the stress field singularity. An other approach is based on the strain energy release rate, where a construction of ring elements in the neighborhood of the crack tip (Fig. 2a,b), is also used. Finally mesh refinement around the crack tip is necessary to keep a better precision in the vicinity of the crack. Since the

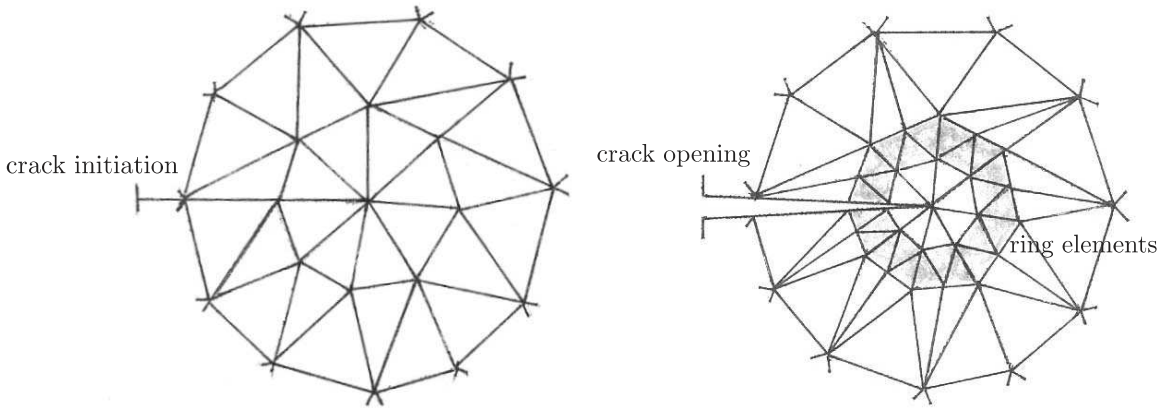


Figure 2: Location of the crack and the mesh around the crack tip: a) crack initiation; b) crack opening.

crack propagates, the crack tip moves along and the areas in the vicinity of crack are changed; thus, a new mesh is created and refined only in areas at the front of the propagated crack.

A location of a crack and its initiation and further opening are given in Fig. 2a,b. Many numerical algorithms have been applied to improve the accuracy at the crack tip and to determine a crack propagation direction. With a great advantage the automatic remeshing procedure at the crack tip, with a thickening of the mesh at the crack tip and using singular elements to model the singular stress-strain fields, can be used. To determine a crack propagation direction we compute eigenvalues and eigenvectors of the stress tensor in all determined mesh points nearest to the crack tip, i.e., we determine the principal stresses and their directions. The final direction of the crack propagation will be obtained as a weighted average of each direction with respect to the distance between the mesh point and the crack tip. Moreover, stress intensity factors, that is, strength singularity at the crack tip, can be used for determination of a crack propagation. Very useful algorithms are based on the dynamic contact problems with friction. Therefore, the PDAS algorithms discussed above can also be used for numerical studies of opening of cracks and fractures in loaded bones with neoplasms. Numerically, simpler versions of the free boundary problems can be firstly studied for the symmetric neoplasms.

4. Conclusion

At present about tumor's studies exist more than two millions research papers, predominantly of the oncological studies from the medical point of views, and only relatively small part of these papers are devoted to mathematical problems of oncology. Majority of these mathematical papers are devoted to studies on the response of a vascular tumor to chemotherapeudic treetments and effects of drug resistance, to studies on a tumor-induced angiogenesis, on a tumor-immune system dynamics and

minority of these papers are devoted to mathematical modelling of tumor's growth. These research works are connected with Profs A. Friedman, S. Cui, H. Byrne, L. Preziosi, M. A. Chaplain, S. J. Chapman, T. Roose, A. R. A. Andersson and many others. They analyzed the problems mathematically and under some assumptions on the physical parameters of the models, they prove the existence and the uniqueness of the solution of some free boundary problems. The studied models are predominantly assumed to be spherically symmetric.

The author, together with his co-workers, studied the problems concerning with the biomechanical problems of artificial replacements of human's joints, and moreover, e.g. a fractured lumbar spine, where the fracture passes practically horizontally through the vertebra, where the internal stabilized device was applied. Such a fracture is observed between the vertebra $Th12$ and $L3$, and is known as the Chance's fracture. The aim of this study was to obtain some knowledge about the situation and the behavior of fractured parts of the vertebra on their common contact boundary (because minor movements stimulate healing of the fracture), where the mathematical model was based on the contact problem in non-linear elasticity, where the non-linear elastic coefficients are strain dependent (see e.g. Nedoma et al. (2011) and the author's references presented here).

The PDAS method was firstly presented in the papers of Hintermüller et al. (2002), (2004), (2005) and in Wohlmuth and Krause (2003), Hüeber and Wohlmuth (2005), Hlaváček (2006) and many others, where the static contact problems with or without given friction were studied. Later Hüeber et al. (2008) applied the PDAS algorithm for $3D$ static contact problems with Coulomb friction, where they present two algorithms based on the fixpoint algorithm and on the full Newton method. Hüeber et al. (2005) studied the dynamic contact problem, where the Newmark algorithm with the PDAS algorithm was used. The author studied the quasi-static and dynamic problems with or without friction close of the nineties in connection with geodynamic problems, based on linear or non-linear elastic, thermo-(visco-)elastic and thermo-visco-plastic Bingham rheologies (Nedoma (1998a), (2005), (2006), (2010), (2012) and later in biomechanics (Nedoma (1998b), (2004), (2006), (2012) and Nedoma et al. (2011) and the author's references presented here. The PDAS algorithm presented in the paper is a continuations of results obtained in previous author's papers connected with the quasi-static and dynamic contact problems with or without friction in thermo-(visco-)elasticity. The presented PDAS algorithm as well as the PDAS algorithms of the previous mentioned papers are based on the author's idea and represent the own author's results. The novelty of these algorithms is that they practically pursue the techniques of proofs of dynamic problem with or without Coulomb (or Tresca) friction. From the medical point of view the aim of this paper is to give an optimal algorithm for application in connection with further oncological studies and in application concerned with a computer-aided orthopedic surgery. The presented method can be used also in geodynamic problems as well as in problems of technology.

Acknowledgements

This work was partly supported by the long term strategic development financing of the Institute of Computer Science ASCR v. v. i. (RVO:67985807). The author thanks to Ms. Hana Bílková for her help with typing of the paper and of preparing the figures.

References

[1] Belytschko, T., Liu, W. K., and Moran, B.: *Nonlinear finite elements for continua and structures*. Wiley, Chichester, 2000.

[2] Cui, S. and Friedman, A.: Analysis of a mathematical model of the growth of necrotic tumors. J. Math. Anal. Appl. **255** (2001), 636–677.

[3] Cui, S. and Friedman, A.: A free boundary problem for a singular system of differential equations: An application to a model of tumor growth. Trans. Amer. Math. Soc. **355 (9)** (2003), 3537–3590.

[4] Eck, C., Jarušek, J., and Krbec, J.: *Unilateral contact problems. Variational methods and existence theorems.* Chapman & Hall/CRC, Taylor & Francis Group, Boca Raton, London, New York, Singapore, 2005.

[5] Friedman, A.: Cancer models and their mathematical analysis. In: *Lect. Notes Math.*, vol. 1872, pp.223–246. Springer, 2006.

[6] Haslinger, J., Hlaváček, I., and Nečas, J.: Numerical methods for unilateral problems in solid mechanics. In: *Handbook of Numerical Analysis*, vol. ICV, pp. 313–486. Elsevier, Amsterdam, 1996.

[7] Hintermüller, M., Ito, K., and Kunish, K.: The primal-dual active set strategy as a semi-smooth Newton method. SIAM J. Optim. **13** (3) (2002), 865–888.

[8] Hintermüller, M., Kovtunenko, V.A., and Kunish, K.: The primal-dual active set method for a crack problem with non-penetration, IMA J. Appl. Math. **69 (1)** (2004) 1–26.

[9] Hlaváček, I.: Primárně duální metoda aktivních množin pro jednostranný kontakt pružných těles s daným třením, Tech.Rep. No 965, Institute of Computer Science AS CR v. v. i., Prague, 2006.

[10] Hlaváček, I.: The primal-dual active set method for unilateral contact of elastic bodies with given friction. Tech. Rep. No 965, Institute of Computer Science AS CR v. v. i., Prague, 2012 (in Czech).

[11] Hüeber, S. and Wohlmuth, B.: A primal-dual active set strategy for non-linear multibody contact problems. Comput. Meth. Appl. Mech. Eng. **194** (2005), 3147–3166.

[12] Hüeber, S., Stadler, G., and Wohlmuth, B.: A primal-dual active set algorithm for three-dimensional contact problems with Coulomb friction. SIAM J. Sci Comput. **30 (2)** (2008), 572–596.

[13] Nedoma, J.: *Numerical methods in applied geodynamics*. John Willey & Sons, Chichester, New York, 1998a.

[14] Nedoma, J.: Contact problems in biomechanics and iterative solution methods for constrained optimization. Theory. Tech. Rep. 756, Institute of Computer Science AS CR v. v. i., Prague, 1998b.

[15] Nedoma, J.: On the solution of contact problems with visco-plastic friction in the Bingham rheology: An application in Biomechanics. ICCSA 2004, *Lecture Notes in Computer Science* (3044), Springer, Berlin, Heidelber, 2004.

[16] Nedoma, J.: Mathematical models of the artificial total replacement of joints. I. Dynamical loading of TEP and TKR. Mathematical 2D and 3D Models. Tech. Rep. 950, Institute of Computer Science AS CR v. v. i., Prague, 2005 (in Czech).

[17] Nedoma, J.: On a solvability of contact problems with visco-plastic friction in the thermo-visco-plastic Bingham rheology. Future Gen. Comput. Syst. **22** (2006), 484–499.

[18] Nedoma, J.: Special problems in landslide modelling. Mathematical and computational methods. In: E. D. Werner and H. P. Friedman (Eds), *Landslides: Causes, Types and Effects*. Nova Sci. Publ., New York, NY, 2010.

[19] Nedoma, J.: Mathematical models of odontogenic cysts and of fractures of jawbones. An introductory study. TR-1166, ICS AS CR, Prague, 2012.

[20] Nedoma, J., et al.: *Mathematical and computational methods in biomechanics of human skeletal systems. An introduction.* John Wiley & Sons, Hoboken, NJ, 2011.

[21] Nedoma, J.: The Primal-Dual Active Set (PDAS) method for dynamic variational inequalities arising from the fractured bone neoplasm models. Tech. Rep. No 1167, Institute of Computer Science AS CR v. v. i., Prague, 2012.

[22] Tombs, M. P. and Peacocke, A. R.: *The osmotic pressure of biological macromolecules*. Clarendon Press, Oxford, UK, 1974.

[23] Ward, J. P., Magar, V., Franks, S. J., and Landini, G.: A mathematical model of the dynamics of odontogenic cyst growth. Analytical and Quantitative Cytology and Histology **26 (1)** (2004), 39–46.

[24] Wohlmuth, B.I. and Krause, R. Monotone multigrid methods on nonmathching grids for non-linear multibody contact problems. SIAM J. Sci. Comput. **25 (1)** (2003), 324–347.

Conference Applications of Mathematics 2015, in honor of the birthday anniversaries of Ivo Babuška (90), Milan Práger (85), and Emil Vitásek (85)
J. Brandts, S. Korotov, M. Křížek, K. Segeth, J. Šístek, T. Vejchodský (Eds.)
Institute of Mathematics AS CR, Prague 2015

NUMERICAL METHOD FOR THE MIXED VOLTERRA-FREDHOLM INTEGRAL EQUATIONS USING HYBRID LEGENDRE FUNCTIONS

S. Nemati[1], P. Lima[2], Y. Ordokhani[3]

[1] Department of Mathematics, Faculty of Mathematical Sciences,
University of Mazandaran
Babolsar, Iran
s.nemati@umz.ac.ir
[2] CEMAT-Departamento de Matematica, Instituto Superior Tecnico, UTL
Av. Rovisco Pais, 1049-001 Lisboa, Portugal
plima@math.ist.utl.pt
[3] Department of Mathematics, Alzahra University
Tehran, Iran
ordokhani@alzahra.ac.ir

Abstract: A new method is proposed for the numerical solution of linear mixed Volterra-Fredholm integral equations in one space variable. The proposed numerical algorithm combines the trapezoidal rule, for the integration in time, with piecewise polynomial approximation, for the space discretization. We extend the method to nonlinear mixed Volterra-Fredholm integral equations. Finally, the method is tested on a number of problems and numerical results are given.

Keywords: mixed Volterra-Fredholm integral equations, hybrid Legendre functions, piecewise polynomial approximation, trapezoidal method

MSC: 65R20, 41A30, 65D30

1. Introduction

In this paper, we are concerned with the numerical solution of the linear mixed Volterra-Fredholm integral equations of the form

$$u(x,t) = f(x,t) + \int_0^t \int_0^a K(x,t,y,z)u(y,z)dydz, \quad 0 \le x,y \le a,\ 0 \le z \le t \le T, \tag{1}$$

where $f(x,t)$ and $K(x,t,y,z)$ are given continuous real-valued functions defined on $[0,a]\times[0,T]$ and $\{(x,t,y,z):\ x,y\in[0,a],\ 0\le z\le t\le T\}$, respectively, and $u(x,t)$

is the unknown function to be determined. With this purpose, space discretisation is introduced, using a basis of hybrid Legendre functions, while time integration is performed using the trapezoidal rule. We will also consider an extension of the proposed method to nonlinear equations of the form

$$u(x,t) = f(x,t) + \int_0^t \int_0^a K(x,t,y,z)g(y,z,u(y,z))dydz, \\ 0 \le x, y \le a,\ 0 \le z \le t \le T, \quad (2)$$

where g is nonlinear in u.

Various problems in physics, mechanics and biology lead to nonlinear mixed type Volterra-Fredholm integral equations. In particular, such equations appear in modeling of the spatio-temporal development of an epidemic, theory of parabolic initial-boundary value problems, population dynamics, and Fourier problems [2, 4, 8].

In its general form, a mixed Volterra-Fredholm integral equation can be written as

$$u(\mathbf{x},t) = f(\mathbf{x},t) + \int_0^t \int_\Omega K(\mathbf{x},t,\mathbf{y},z,u(\mathbf{y},z))d\mathbf{y}dz, \quad (3)$$

where $u(\mathbf{x},t)$ is an unknown real-valued function defined on $D = \Omega \times [0,T]$ and Ω is a closed subset of $\mathbb{R}^n$, $n = 1,\ 2,\ 3$. The functions $f(\mathbf{x},t)$ and $K(\mathbf{x},t,\mathbf{y},z,u)$ are given functions defined on D and $S = \{(\mathbf{x},t,\mathbf{y},z,u) :\ \mathbf{x},\mathbf{y} \in \Omega,\ 0 \le z \le t \le T\}$, respectively [2].

Different numerical methods have been applied to approximate the solution of equation (3) (see for example [1, 3, 5]).

In this paper we use hybrid Legendre and block-pulse functions to solve equations of the forms (1) and (2). Hybrid Legendre functions have been applied extensively for solving differential and integral equations and systems, and proved to be a useful mathematical tool. In [6], a basis of shifted Legendre functions has been applied to the numerical solution of nonlinear two-dimensional Volterra integral equations.

In comparison with the methods used previously to solve equation (3), the advantage of the present method is the high convergence rate, specially with respect to the space variable, which allows to obtain accurate results using small matrices and with a low computational effort (see numerical examples in Section 5). Together with its simple implementation, this makes the present algorithm an efficient tool for the solution of this type of equations.

The organization of the rest of the paper is as follows: In Section 2 hybrid Legendre functions and their basic properties are described. In Section 3 we describe the numerical method used to solve equation (1). In Section 4, the method is extended to solve a class of nonlinear mixed Volterra-Fredholm integral equations. Numerical results are reported in Section 5 and conclusions are presented in Section 6.

2. Properties of hybrid Legendre functions

2.1. Definition and function approximation

Hybrid functions $b_{ij}(x)$, for $i = 1, 2, \ldots, k$, $j = 0, 1, \ldots, M$ and $h = a/k$ are defined on the interval $[0, a)$ as

$$b_{ij}(x) = \begin{cases} L_j(2x/h - 2i + 1), & (i-1)h \leq x < ih, \\ \\ 0, & \text{otherwise.} \end{cases}$$

Here, $L_j(x)$ denotes a Legendre polynomial of order j . Hybrid functions are orthogonal, since

$$\int_0^a b_{ij}(x) b_{mn}(x) = \begin{cases} h/(2j+1), & i = m \text{ and } j = n, \\ \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

Suppose that $V = L^2[0, a]$ and $\{b_{10}(x), b_{11}(x), \ldots, b_{kM}(x)\} \subset V$ is the set of hybrid Legendre functions and

$$B = \text{span}\{b_{10}(x), b_{11}(x), \ldots, b_{1M}(x), \ldots, b_{k0}(x), b_{k1}(x), \ldots, b_{kM}(x)\},$$

and $p(x)$ is an arbitrary element in V. Since B is a finite dimensional vector space, $p(x)$ has a unique best approximation $p_{k,M} \in B$, such that

$$\forall b \in B, \ \|p - p_{k,M}\|_2 \leq \|p - b\|_2.$$

Since $p_{k,M} \in B$, there exist unique coefficients $p_{10}, p_{11}, \ldots, p_{kM}$ such that

$$p(x) \simeq p_{k,M}(x) = \sum_{i=1}^{k} \sum_{j=0}^{M} p_{ij} b_{ij}(x) = P^T \psi(x), \tag{5}$$

where

$$P = [p_{10}, \ldots, p_{1M}, p_{20}, \ldots, p_{2M}, \ldots, p_{k0}, \ldots, p_{kM}]^T, \tag{6}$$

and

$$\psi(x) = [b_{10}(x), \ldots, b_{1M}(x), b_{20}(x), \ldots, b_{2M}(x), \ldots, b_{k0}(x), \ldots, b_{kM}(x)]^T. \tag{7}$$

The hybrid coefficients p_{ij}, $i = 1, 2, \ldots, k$, $j = 0, 1, \ldots, M$ are obtained as

$$p_{ij} = \frac{2j+1}{h} \int_{(i-1)h}^{ih} p(x) b_{ij}(x) dx.$$

We now briefly describe a technique that will be used to integrate hybrid Legendre functions.

2.2. Operational matrix of dual

The integration of the product of two hybrid vectors satisfies [7]:

$$\int_0^a \psi(x)\psi^T(x)dx = D, \tag{8}$$

where D is a $k(M+1)\times k(M+1)$ matrix of the form

$$D = \begin{pmatrix} d & O & O & \dots & O \\ O & d & O & \dots & O \\ O & O & d & \dots & O \\ \vdots & \vdots & \vdots & & \vdots \\ O & O & O & \dots & d \end{pmatrix},$$

in which O is the zero matrix of order $M+1$ and

$$d = h \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1/3 & 0 & \dots & 0 \\ 0 & 0 & 1/5 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \dots & 1/(2M+1) \end{pmatrix}.$$

3. Numerical method

In this section we apply a numerical method using hybrid Legendre functions to the numerical solution of mixed Volterra-Fredholm integral equations of the form (1). With this purpose, we consider the time step size τ as

$$\tau = \frac{T}{N}.$$

Then the mesh nodes are defined by

$$t_0 = 0, \quad t_n = t_{n-1} + \tau, \quad n = 1, 2, \dots, N.$$

Collocating equation (1) in t_n, $n = 0, 1, \dots, N$, yields:

$$u(x, t_n) = f(x, t_n) + \int_0^{t_n} \int_0^a K(x, t_n, y, z)u(y, z)dydz. \tag{9}$$

Considering the notations $u^n(x) = u(x, t_n)$ and $f^n(x) = f(x, t_n)$ in (9), we have

$$u^0(x) = f^0(x),$$

$$u^n(x) = f^n(x) + \int_0^{t_n} \int_0^a K(x, t_n, y, z)u(y, z)dydz, \quad n = 1, 2, \dots, N. \tag{10}$$

Using the trapezoidal rule to perform the integration on z in (10) we obtain the approximation

$$u^n(x) \simeq f^n(x) + \frac{t_n}{2n} \int_0^a \Big(K(x, t_n, y, t_0)u^0(y) + K(x, t_n, y, t_n)u^n(y) + \\ + 2\sum_{i=1}^{n-1} K(x, t_n, y, t_i)u^i(y) \Big) dy. \tag{11}$$

Introducing the notation $K^{n,i}(x, y) = K(x, t_n, y, t_i)$ in (11), yields:

$$u^n(x) = f^n(x) + \frac{t_n}{2n} \int_0^a \Big(K^{n,0}(x, y)u^0(y) + K^{n,n}(x, y)u^n(y) + 2\sum_{i=1}^{n-1} K^{n,i}(x, y)u^i(y) \Big) dy. \tag{12}$$

We approximate the functions in (12) using the method described in the previous section as

$$u^i(x) \simeq u^i_{k,M}(x) = U_i^T \psi(x) = \psi^T(x)U_i, \tag{13}$$

$$f^n(x) \simeq f^n_{k,M}(x) = F_n^T \psi(x) = \psi^T(x)F_n \tag{14}$$

$$K^{n,i}(x, y) \simeq K^{n,i}_{k,M}(x, y) = \psi^T(x)\kappa_{n,i}\psi(y), \tag{15}$$

where U_n, $n = 1,\ 2, \ldots, N$, in (13) is the unknown vector, of order $k(M + 1)$.
Substituting approximations (13)–(15) into equation (12) and using the operational matrix of dual, we obtain

$$U_n = F_n + \frac{t_n}{2n} \left[\kappa_{n,0} D U_0 + \kappa_{n,n} D U_n + 2\sum_{i=1}^{n-1} \kappa_{n,i} D U_i \right],$$

which can be rewritten as

$$(I - \frac{t_n}{2n}\kappa_{n,n} D)U_n = F_n + \frac{t_n}{2n} \left[\kappa_{n,0} D U_0 + 2\sum_{i=1}^{n-1} \kappa_{n,i} D U_i \right], \quad n = 1, \ldots, N. \tag{16}$$

Equations (16) form a system of $k(M + 1)$ linear equations in each step and can be solved easily using direct methods.
Therefore U_n, $n = 1,\ 2, \ldots, N$ can be computed via the recursive equation (16) using the initial value $U_0 = F_0$.

4. Numerical solution of nonlinear mixed Volterra-Fredholm integral equations

In this section we extend our numerical method to solve nonlinear mixed Volterra-Fredholm integral equations of the form (2).

Considering the same partition and notations as in Section 3 and collocating equation (2) in $t = t_n$ yields:

$$u^n(x) = f^n(x) + \int_0^{t_n} \int_0^a K(x, t_n, y, z) g(y, z, u(y, z)) dy dz. \tag{17}$$

Using the composite trapezoidal integration rule for the integral part of (17) leads to:

$$u^n(x) = f^n(x) + \frac{t_n}{2n} \int_0^a \Big(K^{n,0}(x, y) g(y, t_0, u^0(y)) + K^{n,n}(x, y) g(y, t_n, u^n(y)) + \\ + 2 \sum_{i=1}^{n-1} K^{n,i}(x, y) g(y, t_i, u^i(y)) \Big) dy. \tag{18}$$

Introducing the notation $g^i(y) = g(y, t_i, u^i(y))$ equation (18) can be written as

$$u^n(x) = f^n(x) + \frac{t_n}{2n} \int_0^a \Big(K^{n,0}(x, y) g^0(y) + K^{n,n}(x, y) g^n(y) + 2 \sum_{i=1}^{n-1} K^{n,i}(x, y) g^i(y) \Big) dy. \tag{19}$$

We approximate the functions $u^i(x)$, $f^n(x)$ and $K^{n,i}(x, y)$ in equation (19) using (13)–(15) and replace $g^i(y)$ with

$$g^i(y) \simeq g^i_{k,M}(x) = G_i^T \psi(x) = \psi^T(x) G_i, \tag{20}$$

where U_i and G_i are unknown vectors of dimension $k(M + 1)$. Then, substituting these approximations and using the operational matrix of dual in (19) yields:

$$U_n = F_n + \frac{t_n}{2n} \left[\kappa_{n,0} D G_0 + \kappa_{n,n} D G_n + 2 \sum_{i=1}^{n-1} \kappa_{n,i} D G_i \right], \tag{21}$$

which forms a system of $k(M + 1)$ linear algebraic equations in terms of $2k(M + 1)$ unknowns. In order to obtain a uniquely solvable system, we need $k(M+1)$ additional equations. For this purpose consider $k(M + 1)$ collocation points defined by

$$x_{i,j} = \frac{h}{2}(x_j + 2i - 1), \quad i = 1,\ 2, \ldots, k, \quad j = 0,\ 1, \ldots, M,$$

where x_j, $j = 0,\ 1, \ldots, M$ are the roots of Legendre polynomial of degree $M + 1$. Collocating the equation $g(x, t_n, U_n^T \psi(x)) = G_n^T \psi(x)$ in $x_{i,j}$, we obtain

$$g(x_{i,j}, t_n, U_n^T \psi(x_{i,j})) - G_n^T \psi(x_{i,j}) = 0, \quad \text{for } i = 1,\ 2, \ldots, k, \quad j = 0,\ 1, \ldots, M, \tag{22}$$

which is a system of $k(M+1)$ nonlinear equations in terms of the unknown elements of the vectors U_n and G_n. Finally, systems (21) and (22) together form a system of $2k(M + 1)$ equations and can be solved in terms of U_n and G_n using the Newton's iterative method. In the case $t = 0$, we have $U_0 = F_0$, and G_0 is obtained using the approximation of the function $g(x, 0, U_0^T \psi(x))$ (which is a known function) by the hybrid Legendre functions.

5. Numerical examples

In this section, the results of two numerical experiments are presented to validate accuracy, applicability and convergence of the proposed methods. In order to investigate the error of the method we introduce the following notations. The error norm is denoted by

$$e_n(x) = |u_n(x) - \tilde{u}_n(x)|,$$
$$E_{k,M,N}(t_n) = \|e_n(x)\|_2,$$

where $u_n(x)$ and $\tilde{u}_n(x)$ are the exact solution and the computed solution by the presented method at $t = t_n$ with selected k, M and N, respectively. For the convergence order, with respect to h, we use the estimate:

$$\rho_k(t_n) = \log_2 \left(E_{k,M,N}/E_{2k,M,N}\right);$$

and for the convergence order, with respect to τ, we write

$$\varrho_N(t_n) = \log_2 \left(E_{k,M,N}/E_{k,M,2N}\right).$$

When using different meshes in space (time), the stepsize h (resp. τ) of each subsequent mesh is twice smaller.

Example 1: Consider the following linear mixed Volterra-Fredholm integral equation as discussed in [5]

$$u(x,t) = f(x,t) + \int_0^t \int_0^2 K(x,t,y,z)u(y,z)dydz, \quad 0 \le t \le 1, \tag{23}$$

where

$$f(x,t) = \mathrm{e}^{-t}\left(\cos(x) + t\cos(x) + \frac{1}{2}t\cos(x-2)\sin(2)\right),$$
$$K(x,t,y,z) = -\cos(x-y)\mathrm{e}^{-(t-z)},$$

with the exact solution $u(x,t) = \mathrm{e}^{-t}\cos(x)$. After multiplying the exact solution by the kernel K we observe that the integrand function on the right-hand side of (23) does not depend on z. Therefore, the outer integral can be computed exactly and the final error of the numerical solution does not depend on τ. This is why in our tests we only check the convergence of the method, as $h \to 0$. We have applied the described numerical method with $M = 3$ and $M = 6$. In both cases, we have taken $N = 100$ and used three different meshes in space, with $k = 2,\ 4,\ 8$. The numerical results are given in Tables 1–2. They present 4-th order convergence in the case $M = 3$ and 7-th order convergence in the case $M = 6$.

Example 2: Consider the following nonlinear mixed Volterra-Fredholm integral equation, which arises in the mathematical modeling of the development of an epidemic [1, 3]:

$$u(x,t) = f(x,t) + \int_0^t \int_0^1 K(x,t,y,z)(1 - \mathrm{e}^{-u(y,z)})dydz \quad 0 \le t \le 1, \tag{24}$$

t	$E_{2,3,100}$	$E_{4,3,100}$	ρ_2	$E_{8,3,100}$	ρ_4
0.1	1.5659×10^{-4}	1.0029×10^{-5}	3.96	6.3032×10^{-7}	3.99
0.2	1.4168×10^{-4}	9.0747×10^{-6}	3.96	5.7034×10^{-7}	3.99
0.3	1.2820×10^{-4}	8.2111×10^{-6}	3.96	5.1606×10^{-7}	3.99
0.4	1.1600×10^{-4}	7.4297×10^{-6}	3.96	4.6695×10^{-7}	3.99
0.5	1.0496×10^{-4}	6.7227×10^{-6}	3.96	4.2252×10^{-7}	3.99
0.6	9.4976×10^{-5}	6.0829×10^{-6}	3.96	3.8231×10^{-7}	3.99
0.7	8.5938×10^{-5}	5.5040×10^{-6}	3.96	3.4593×10^{-7}	3.99
0.8	7.7760×10^{-5}	4.9803×10^{-6}	3.96	3.1301×10^{-7}	3.99
0.9	7.0360×10^{-5}	4.5063×10^{-6}	3.96	2.8322×10^{-7}	3.99
1.0	6.3664×10^{-5}	4.0775×10^{-6}	3.96	2.5627×10^{-7}	3.99

Table 1: Numerical results for Example 1

t	$E_{2,6,100}$	$E_{4,6,100}$	ρ_2	$E_{8,6,100}$	ρ_4
0.1	1.4847×10^{-8}	1.1527×10^{-10}	7.00	8.9936×10^{-13}	7.00
0.2	1.3434×10^{-8}	1.0430×10^{-10}	7.00	8.1378×10^{-13}	7.00
0.3	1.2156×10^{-8}	9.4374×10^{-11}	7.00	7.3634×10^{-13}	7.00
0.4	1.0999×10^{-8}	8.5393×10^{-11}	7.00	6.6626×10^{-13}	7.00
0.5	9.9528×10^{-9}	7.7267×10^{-11}	7.00	6.0286×10^{-13}	7.00
0.6	9.0056×10^{-9}	6.9914×10^{-11}	7.00	5.4549×10^{-13}	7.00
0.7	8.1486×10^{-9}	6.3261×10^{-11}	7.00	4.9358×10^{-13}	7.00
0.8	7.3732×10^{-9}	5.7241×10^{-11}	7.00	4.4661×10^{-13}	7.00
0.9	6.6715×10^{-9}	5.1794×10^{-11}	7.00	4.0411×10^{-13}	7.00
1.0	6.0366×10^{-9}	4.6865×10^{-11}	7.00	3.6565×10^{-13}	7.00

Table 2: Numerical results for Example 1

where

$$f(x,t) = -\ln\left(1 + \frac{xt}{1+t^2}\right) + \frac{xt^2}{8(1+t)(1+t^2)},$$

$$K(x,t,y,z) = \frac{x(1-y^2)}{(1+t)(1+z^2)}.$$

Its exact solution is $u(x,t) = -\ln(1 + xt/(1+t^2))$. The results of the numerical experiments with this example are displayed in Tables 3–4. In Table 3, τ is kept constant, with $N = 1000$, and $M = 2$ (quadratic polynomials). Note that with such value of N resulting from the time discretization is negligible when compared with the final error, so we can again investigate the dependence of the error on h. The error norms on three different meshes ($k = 2$,$k = 4$, and $k = 8$) show that the discretization error depends on h as $O(h^3)$. Finally, we have investigated the depen-

t	$E_{2,2,1000}$	$E_{4,2,1000}$	ρ_2	$E_{8,2,1000}$	ρ_4
0.1	6.6527×10^{-7}	8.3258×10^{-8}	2.99	1.0430×10^{-8}	2.99
0.2	4.3442×10^{-6}	5.4526×10^{-7}	2.99	6.8263×10^{-8}	2.99
0.3	1.1613×10^{-5}	1.4632×10^{-6}	2.98	1.8332×10^{-7}	2.99
0.4	2.1268×10^{-5}	2.6906×10^{-6}	2.98	3.3741×10^{-7}	2.99
0.5	3.1480×10^{-5}	3.9972×10^{-6}	2.97	5.0173×10^{-7}	2.99
0.6	4.0666×10^{-5}	5.1791×10^{-6}	2.97	6.5060×10^{-7}	2.99
0.7	4.7867×10^{-5}	6.1096×10^{-6}	2.97	7.6794×10^{-7}	2.99
0.8	5.2746×10^{-5}	6.7421×10^{-6}	2.96	8.4777×10^{-7}	2.99
0.9	5.5411×10^{-5}	7.0882×10^{-6}	2.96	8.9148×10^{-7}	2.99
1.0	5.6206×10^{-5}	7.1916×10^{-6}	2.96	9.0452×10^{-7}	2.99

Table 3: Numerical results for Example 2

t	$E_{16,2,20}$	$E_{16,2,40}$	ϱ_{20}	$E_{16,2,80}$	ϱ_{40}
0.1	1.6164×10^{-6}	4.0346×10^{-7}	2.00	1.0083×10^{-7}	2.00
0.2	5.5511×10^{-6}	1.3861×10^{-6}	2.00	3.4654×10^{-7}	1.99
0.3	1.0347×10^{-5}	2.5844×10^{-6}	2.00	6.4635×10^{-7}	1.99
0.4	1.4797×10^{-5}	3.6966×10^{-6}	2.00	9.2489×10^{-7}	1.99
0.5	1.8199×10^{-5}	4.5474×10^{-6}	2.00	1.1383×10^{-6}	1.99
0.6	2.0349×10^{-5}	5.0854×10^{-6}	2.00	1.2736×10^{-6}	1.99
0.7	2.1372×10^{-5}	5.3413×10^{-6}	2.00	1.3384×10^{-6}	1.99
0.8	2.1534×10^{-5}	5.3825×10^{-6}	2.00	1.3494×10^{-6}	1.99
0.9	2.1122×10^{-5}	5.2798×10^{-6}	2.00	1.3242×10^{-6}	1.99
1.0	2.0371×10^{-5}	5.0923×10^{-6}	2.00	1.2777×10^{-6}	1.99

Table 4: Numerical results for Example 2

dence of the error on τ. With this purpose, we have used three different stepsizes in time corresponding to $N = 20, N = 40$ and $N = 80$, keeping the stepsize h fixed ($k = 16$). For such stepsizes, the error resulting from the space discretization is much smaller than the component depending on τ. In this case, the results displayed in Table 4 show clearly that the error behaves as τ^2.

6. Conclusion

A new method is proposed for the numerical solution of linear and nonlinear mixed Volterra-Fredholm integral equations. The numerical scheme combines the trapezoidal rule, for integration in time, and piecewise polynomial approximation, for space discretisation. The hybrid Legendre functions and the operational matrix of dual are applied to reduce the problem to an algebraic system of nonlinear equations,

which is solved by the Newton method. The computational method was tested using a sample of numerical examples, including an equation arising in the modeling of spatio-temporal development of an epidemic (Example 2), which was formerly analysed by other authors [1, 3]. Our results for this example (see Tables 3–4) have the same degree of accuracy (6 digits) as the results presented in [3], obtained by means of a collocation scheme with Gaussian points, both in time and space. The numerical experiments suggest that the convergence order is $O(h^{M+1}) + O(\tau^2)$, which is in agreement with the known properties of methods based on piecewise polynomial collocation and trapezoidal rule. We leave as a future work the analysis of convergence.

References

[1] Banifatemi, E., Razzaghi, M., and Yousefi, S.: Two-dimensional Legendre wavelets method for the mixed Volterra-Fredholm integral equations. J. Vibr. Control **13** (2007), 1667–1675.

[2] Brunner, H.: *Collocation methods for Volterra integral and related functional equations*. Cambridge University Press, Cambridge, 2004.

[3] Brunner, H.: On the numerical solution of nonlinear Volterra-Fredholm integral equations by collocation methods. SIAM J. Numer. Anal. **27 (4)** (1990), 987–1000.

[4] Diekmann, O.: Thresholds and traveling for the geographical spread of infection. J. Math. Biol. **6** (1978), 109–130.

[5] Han, G. Q. and Zhang, L. Q.: Asymptotic error expansion for the trapezoidal Nystrom method of linear Volterra-Fredholm equations. J. Comput. Appl. Math. **51** (1994), 339–348.

[6] Nemati, S., Lima, P. M., and Ordokhani, Y.: Numerical solution of a class of two-dimensional Volterra integral equations using Legendre polynomials. J. Comput. Appl. Math. **242** (2013), 53–69.

[7] Razzaghi, M., and Marzban, H. R.: A hybrid analysis direct method in the calculus of variations. Int. J. Comput. Math. **75** (2000), 259–269.

[8] Thieme, H. R.: A model for the spatial spread of an epidemic. J. Math. Biol. **4** (1977) 337–351.

Conference Applications of Mathematics 2015, in honor of the birthday anniversaries of Ivo Babuška (90), Milan Práger (85), and Emil Vitásek (85)
J. Brandts, S. Korotov, M. Křížek, K. Segeth, J. Šístek, T. Vejchodský (Eds.)
Institute of Mathematics CAS, Prague 2015

A MULTI-SPACE ERROR ESTIMATION APPROACH FOR MESHFREE METHODS

Marcus Rüter[1], Jiun-Shyan Chen[2]

[1] Department of Civil and Environmental Engineering
University of California, Los Angeles, CA 90095, USA
marcus.ruter@ucla.edu
[2] Department of Structural Engineering
University of California, San Diego, La Jolla, CA 92093, USA
js-chen@ucsd.edu

Abstract: Error-controlled adaptive meshfree methods are presented for both global error measures, such as the energy norm, and goal-oriented error measures in terms of quantities of interest. The meshfree method chosen in this paper is the reproducing kernel particle method (RKPM), since it is based on a Galerkin scheme and therefore allows extensions of quality control approaches as already developed for the finite element method. Our approach of goal-oriented error estimation is based on the well-established technique using an auxiliary dual problem. To keep the formulation general and to add versatility, a multi-space approach is used, where the dual problem is solved numerically using a different approximation space than the one employed in the associated primal problem. This can be realized with meshfree methods at no additional cost. Possible merits of this multi-space approach are discussed and an illustrative numerical example is presented.

Keywords: reproducing kernel particle method (RKPM), meshfree methods, goal-oriented error estimation, dual problem

MSC: 65N15, 65N50, 74B05, 74R10

1. Introduction

In this paper, we confine our attention to error control of Galerkin-type meshfree methods, more specifically to error control of the reproducing kernel particle method (RKPM). From an error control point of view this has the advantage that error estimation techniques, as extensively developed for the finite element method, can generally be transferred to RKPM as they are both Galerkin methods.

Although meshfree methods offer several obvious advantages for *a posteriori* error control, the development of error estimators is surprisingly still in an early stage.

To date, the largest class of error estimators for meshfree methods constitutes of recovery-type error estimators. Mathematically more sound error estimators can be found in the class of residual-type error estimators. By construction, such error estimators have the virtue to offer error bounds. However, owing to several pessimistic inequalities that are usually used in their derivations, the error bounds of residual-type error estimators are typically not as sharp as the error approximations of recovery-type error estimators. Subclasses of this type of error estimation procedures constitute of explicit-type estimators, where the residual is used directly in the line of the pioneering works by Babuška & Rheinboldt [3, 4], and implicit-type estimators, where auxiliary local problems based on the residual are solved in the line of Bank & Weiser [6].

To the knowledge of the authors, in meshfree methods each of these subclasses currently consists of one representative. An explicit residual-type error estimator was developed by Duarte & Oden [8] for the meshfree method by the same authors, called *h-p* clouds. More recently, Vidal et al. [14] presented an implicit residual-type error estimator, where the auxiliary local problems are solved on patches of the integration cells so that no fluxes need to be taken into account. The authors are also the first ones who derived goal-oriented error estimators for meshfree methods.

In this paper, we follow Rüter & Chen [11] to add a new error estimator to the class of implicit residual-type error estimators for meshfree methods based on the finite element counterpart as introduced by Bank & Weiser [6] and further developed by Ainsworth & Oden [1] and others, see also Babuška & Strouboulis [5]. The error estimator presented in this paper is first derived for an energy-norm error control and is later extended to goal-oriented error estimation. It takes advantage of two key properties of meshfree methods. The first one is the high regularity of the meshfree solution which is reflected in a smooth stress field and thus in a smooth traction field. The second one is the independence of the particles from the integration cells which makes it possible to use different discretizations for the primal and for the dual problem, as used for the goal-oriented error estimator, at no additional cost. This multi-space approach thus bypasses the tedious transfer of discrete solutions from one mesh to the other as is required for mesh-based methods, such as the finite element method, as shown in Rüter et al. [12]. It is therefore tailored to meshfree methods and adds versatility and convenience to the goal-oriented error estimator proposed in this paper.

The paper is divided up as follows: in Section 2, the model problem of linear elasticity is presented. Furthermore, the meshfree method, RKPM, is introduced. Section 3 focuses on the derivation of a global implicit residual-type error estimator. In Section 4, the error estimator is extended to the case of goal-oriented error estimators where the error measure is given in terms of an arbitrary, user-defined quantity of interest. Thereby, emphasis is placed on the multi-space approach. The error estimator is then applied to a linear elastic fracture mechanics (LEFM) problem in Section 5. The paper concludes with Section 6, which summarizes the major findings achieved from theoretical and numerical points of view.

2. The model problem and its meshfree discretization

In this section, we briefly present the linear elasticity problem in its strong and weak forms. Furthermore, we show how a meshfree Galerkin method can be constructed based on reproducing kernel (RK) shape functions.

2.1. Strong and weak forms

We first introduce the elastic body which is given by the open, bounded domain $\Omega \subset \mathbb{R}^d$ with dimension $d \in \{1, 2, 3\}$. Its boundary $\Gamma = \partial\Omega$ consists of two disjoint parts $\Gamma_D \subset \Gamma$ and $\Gamma_N = \Gamma \setminus \bar{\Gamma}_D$, where, for simplicity, homogeneous Dirichlet and (generally inhomogeneous) Neumann boundary conditions are imposed, respectively.

The strong form of the elliptic and self-adjoint model problem of linear elasticity is to find the displacement field $\boldsymbol{u}$ such that the field equations

$$-\operatorname{div}\boldsymbol{\sigma}(\boldsymbol{u}) = \boldsymbol{f} \quad \text{in } \Omega \tag{1a}$$

$$\boldsymbol{\sigma} - \mathbb{C} : \boldsymbol{\varepsilon}(\boldsymbol{u}) = \boldsymbol{0} \quad \text{in } \Omega \tag{1b}$$

$$\boldsymbol{\varepsilon} - \boldsymbol{\nabla}^{\mathrm{sym}}\boldsymbol{u} = \boldsymbol{0} \quad \text{in } \Omega \tag{1c}$$

subjected to the boundary conditions

$$\boldsymbol{u} = \boldsymbol{0} \quad \text{on } \Gamma_D \tag{2a}$$

$$\boldsymbol{\sigma}(\boldsymbol{u}) \cdot \boldsymbol{n} = \bar{\boldsymbol{t}} \quad \text{on } \Gamma_N \tag{2b}$$

are fulfilled. In the above, $\boldsymbol{\sigma}$ denotes the stress tensor, $\boldsymbol{\varepsilon}$ is the strain tensor, and $\mathbb{C}$ is the elasticity tensor. Furthermore, on the right-hand sides of (1) and (2) we have prescribed body forces $\boldsymbol{f}$ and tractions $\bar{\boldsymbol{t}}$ that are assumed to be in the spaces $\boldsymbol{L}_2(\Omega)$ and $\boldsymbol{L}_2(\Gamma_N)$, respectively. Lastly, $\boldsymbol{n}$ denotes the unit outward normal.

In the classical weak formulation associated with (1) and (2) we seek for a solution $\boldsymbol{u}$ in the trial and test space $\mathcal{V}_0 = \{\boldsymbol{v} \in \boldsymbol{H}^1(\Omega)\,;\, \boldsymbol{v}|_{\Gamma_D} = \boldsymbol{0}\} \subset \mathcal{V} = \boldsymbol{H}^1(\Omega)$ such that

$$a(\boldsymbol{u}, \boldsymbol{v}) = F(\boldsymbol{v}) \quad \forall \boldsymbol{v} \in \mathcal{V}_0. \tag{3}$$

Here, a is a bilinear form defined on $\mathcal{V} \times \mathcal{V}$ as

$$a(\boldsymbol{u}, \boldsymbol{v}) = \int_\Omega \boldsymbol{\sigma}(\boldsymbol{u}) : \boldsymbol{\varepsilon}(\boldsymbol{v}) \, \mathrm{d}V \tag{4}$$

and F is a linear form defined on $\mathcal{V}$ as

$$F(\boldsymbol{v}) = \int_\Omega \boldsymbol{f} \cdot \boldsymbol{v} \, \mathrm{d}V + \int_{\Gamma_N} \bar{\boldsymbol{t}} \cdot \boldsymbol{v} \, \mathrm{d}A. \tag{5}$$

2.2. Reproducing kernel shape functions

In the associated meshfree Galerkin formulation of the weak form (3), we project (3) onto a suitable finite-dimensional subspace $\mathcal{V}_h \subset \mathcal{V}$ with

$$\mathcal{V}_h = \operatorname{span}\{\Psi_I\}_I^{n_P}, \tag{6}$$

where n_P is the number of particles $\boldsymbol{x}_I \in \bar{\Omega}$. A function $\boldsymbol{v}_h \in \mathcal{V}_h$ can then be expressed as

$$\boldsymbol{v}_h(\boldsymbol{x}) = \sum_{n_P} \Psi_I(\boldsymbol{x})\boldsymbol{v}_I \quad \forall \boldsymbol{x} \in \bar{\Omega}. \tag{7}$$

Here, $\boldsymbol{v}_I$ is a particle coefficient but, as opposed to the finite element method, in general not the value of $\boldsymbol{v}_h$ at particle $\boldsymbol{x}_I$, i.e. $\boldsymbol{v}_I \neq \boldsymbol{v}_h(\boldsymbol{x}_I)$, since, in general, the reproducing kernel (RK) shape functions Ψ_I do not possess the Kronecker-delta property (unlike the finite element shape functions), i.e. $\Psi_I(\boldsymbol{x}_J) \neq \delta_{IJ}$. Thus, Dirichlet boundary conditions cannot be satisfied by functions $\boldsymbol{v}_h \in \mathcal{V}_h$. As a consequence, it is obvious that, in general, $\mathcal{V}_h \not\subset \mathcal{V}_0$.

The meshfree RK shape function Ψ_I associated with a particle $\boldsymbol{x}_I$ takes the specific form

$$\Psi_I(\boldsymbol{x}) = \Phi(\boldsymbol{x} - \boldsymbol{x}_I)\boldsymbol{H}^T(\boldsymbol{0})\boldsymbol{M}^{-1}(\boldsymbol{x})\boldsymbol{H}(\boldsymbol{x} - \boldsymbol{x}_I) \tag{8}$$

with kernel function Φ, typically chosen as a cubic B-spline, vector of monomial basis functions $\boldsymbol{H}$, and the symmetric moment matrix

$$\boldsymbol{M}(\boldsymbol{x}) = \sum_{n_P} \Phi(\boldsymbol{x} - \boldsymbol{x}_I)\boldsymbol{H}(\boldsymbol{x} - \boldsymbol{x}_I)\boldsymbol{H}^T(\boldsymbol{x} - \boldsymbol{x}_I). \tag{9}$$

Note that since $\boldsymbol{M}$ needs to be invertible, the support of the kernel function Φ, i.e. $\operatorname{supp}\Phi$, needs to cover a sufficient amount of particles, see [7].

2.3. The reproducing kernel particle method

RKPM is a Galerkin method based on the RK shape functions as introduced in the previous section. As such, it is clear from the above that (homogeneous) Dirichlet boundary conditions on Γ_D are generally not fulfilled by the method. Without loss of generality, in this paper we make use of Nitsche's method to weakly impose the Dirichlet boundary conditions, see [10, 2], which leads to the following discrete problem associated with (3): find the RKPM solution $\boldsymbol{u}_h \in \mathcal{V}_h$ such that

$$a_h(\boldsymbol{u}_h, \boldsymbol{v}_h) = F(\boldsymbol{v}_h) \quad \forall \boldsymbol{v}_h \in \mathcal{V}_h. \tag{10}$$

Here, the discretization-dependent, symmetric bilinear form a_h is defined as

$$a_h(\boldsymbol{u}_h, \boldsymbol{v}_h) = a(\boldsymbol{u}_h, \boldsymbol{v}_h) - \int_{\Gamma_D} \boldsymbol{v}_h \cdot \boldsymbol{\sigma}(\boldsymbol{u}_h) \cdot \boldsymbol{n} \, \mathrm{d}A - \int_{\Gamma_D} \boldsymbol{u}_h \cdot \boldsymbol{\sigma}(\boldsymbol{v}_h) \cdot \boldsymbol{n} \, \mathrm{d}A + \frac{\beta}{h} \int_{\Gamma_D} \boldsymbol{u}_h \cdot \boldsymbol{v}_h \, \mathrm{d}A \tag{11}$$

with discretization parameter h and penalty parameter $\beta \in \mathbb{R}^+$ that takes the role of a stabilization parameter. Note that the right-hand side remains unchanged, since homogeneous boundary conditions are imposed on Γ_D.

3. Implicit energy norm residual-type error estimation

In what follows, we will derive an energy-norm error estimator of implicit residual type, which is based on a projected error residual equation to account for Nitsche's method and to get a symmetric form. The error estimator is established in terms of local forms of the projected error residual equation on each (Gauss) integration cell.

3.1. The error residual equation

The discretization error is defined in the usual way as $\boldsymbol{e} = \boldsymbol{u} - \boldsymbol{u}_h$. However, the error $\boldsymbol{e}$ as obtained by RKPM is an element of $\mathcal{V}_E = \{\boldsymbol{v} \in \boldsymbol{H}^1(\Omega)\,;\, \boldsymbol{v}|_{\Gamma_D} = \boldsymbol{e}|_{\Gamma_D}\} \subset \mathcal{V}$ rather than $\mathcal{V}_0$ as in mesh-based methods thanks to the Kronecker delta property of the mesh-based shape functions.

The starting point of our *a posteriori* error analysis will be the (extended) error residual equation

$$a(\boldsymbol{e}, \boldsymbol{v}) - \int_{\Gamma_D} \boldsymbol{v} \cdot \boldsymbol{\sigma}(\boldsymbol{e}) \cdot \boldsymbol{n} \, \mathrm{d}A = \hat{R}(\boldsymbol{v}) \quad \forall \boldsymbol{v} \in \mathcal{V}, \tag{12}$$

where the extended residual $\hat{R}$ is defined on $\mathcal{V}$ as

$$\hat{R}(\boldsymbol{v}) = F(\boldsymbol{v}) - a(\boldsymbol{u}_h, \boldsymbol{v}) + \int_{\Gamma_D} \boldsymbol{v} \cdot \boldsymbol{\sigma}(\boldsymbol{u}_h) \cdot \boldsymbol{n} \, \mathrm{d}A \tag{13a}$$

$$= \int_\Omega \boldsymbol{f} \cdot \boldsymbol{v} \, \mathrm{d}V + \int_{\Gamma_N} \bar{\boldsymbol{t}} \cdot \boldsymbol{v} \, \mathrm{d}A - \int_\Omega \boldsymbol{\sigma}(\boldsymbol{u}_h) : \boldsymbol{\varepsilon}(\boldsymbol{v}) \, \mathrm{d}V + \int_{\Gamma_D} \boldsymbol{t}_h \cdot \boldsymbol{v} \, \mathrm{d}A. \tag{13b}$$

Note that if $\boldsymbol{v}$ is chosen from $\mathcal{V}_0$, then (12) simplifies to the well-known equation $a(\boldsymbol{e}, \boldsymbol{v}) = R(\boldsymbol{v})$. Obviously, coercivity is an issue in (12). Therefore, and also to deal with the bilinear form a only, we propose to introduce a projection of the error in $\mathcal{V}_E$, denoted by $\hat{\boldsymbol{e}}$ and with the obvious property $\hat{\boldsymbol{e}}|_{\Gamma_D} = \boldsymbol{e}|_{\Gamma_D}$. This projection is defined via

$$a(\hat{\boldsymbol{e}}, \boldsymbol{v}) = a(\boldsymbol{e}, \boldsymbol{v}) - \int_{\Gamma_D} \boldsymbol{v} \cdot \boldsymbol{\sigma}(\boldsymbol{e}) \cdot \boldsymbol{n} \, \mathrm{d}A \quad \forall \boldsymbol{v} \in \mathcal{V} \tag{14}$$

and leads to the projected error residual equation

$$a(\hat{\boldsymbol{e}}, \boldsymbol{v}) = \hat{R}(\boldsymbol{v}) \quad \forall \boldsymbol{v} \in \mathcal{V} \tag{15}$$

with coercive bilinear form a. Using the discretization-dependent norm

$$\|\boldsymbol{v}\|^2_{\frac{1}{2},h} = \sum_{E \subset \Gamma_D} h_E^{-1} \|\boldsymbol{v}\|^2_{L_2(E)}, \tag{16}$$

with edge discretization parameter h_E and edge E, the Cauchy-Schwarz inequality, and an inverse estimate with positive constant C, it can be shown, see [11], that

$$|\!|\!| \boldsymbol{e} |\!|\!| \leq |\!|\!| \hat{\boldsymbol{e}} |\!|\!| + C \|\hat{\boldsymbol{e}}\|_{\frac{1}{2},h} \tag{17}$$

holds, where $|\!|\!| \cdot |\!|\!|$ is the energy norm. Note that if $\hat{\boldsymbol{e}} = \boldsymbol{e} = \boldsymbol{0}$ on Γ_D, then $|\!|\!| \hat{\boldsymbol{e}} |\!|\!| = |\!|\!| \boldsymbol{e} |\!|\!|$.

3.2. The energy-norm a posteriori error estimator

The general idea to derive an implicit residual-type *a posteriori* error estimator is to solve an approximation of the (projected) error residual equation (15). This

usually requires higher-order trial and test spaces because of the Galerkin orthogonality and would be computationally too expensive if computed globally. Therefore, the (projected) error residual equation (15) is solved in subdomains of the elastic body Ω, which can be chosen in a meshfree method as the integration cells.

The trial space for the local problems in each of the n_i integration cells Ω_i is thus defined as $\mathcal{V}_{E,i} = \{\boldsymbol{v}|_{\Omega_i} \,;\, \boldsymbol{v} \in \mathcal{V}_E\}$. Likewise, the local test space becomes $\mathcal{V}_i = \{\boldsymbol{v}|_{\Omega_i} \,;\, \boldsymbol{v} \in \mathcal{V}\}$. Consequently, the local bilinear form a_i on $\mathcal{V}_i \times \mathcal{V}_i$ is given by

$$a_i(\hat{\boldsymbol{e}}|_{\Omega_i}, \boldsymbol{v}) = \int_{\Omega_i} \boldsymbol{\sigma}(\hat{\boldsymbol{e}}|_{\Omega_i}) : \boldsymbol{\varepsilon}(\boldsymbol{v}) \, \mathrm{d}V. \tag{18}$$

Similarly, the local extended residual $\hat{R}_i$ on $\mathcal{V}_i$ reads

$$\hat{R}_i(\boldsymbol{v}) = \int_{\Omega_i} \boldsymbol{f}|_{\Omega_i} \cdot \boldsymbol{v} \, \mathrm{d}V + \sum_{l=1}^{n_l} \int_{E_l \subset \partial\Omega_i} \boldsymbol{t}_l \cdot \boldsymbol{v} \, \mathrm{d}A - \int_{\Omega_i} \boldsymbol{\sigma}(\boldsymbol{u}_h|_{\Omega_i}) : \boldsymbol{\varepsilon}(\boldsymbol{v}) \, \mathrm{d}V, \tag{19}$$

where n_l is the number of edges E_l of an integration cell and $\boldsymbol{t}_l = \boldsymbol{\sigma}(\boldsymbol{u}|_{\Omega_i}) \cdot \boldsymbol{n}$ are the exact tractions if $E_l \not\subset \Gamma_D$, otherwise $\boldsymbol{t}_l = \boldsymbol{\sigma}(\boldsymbol{u}_h|_{\Omega_i}) \cdot \boldsymbol{n}$ are the RKPM tractions.

With the above local forms (18) and (19) at hand, the projected discretization error restricted to an integration cell, i.e. $\hat{\boldsymbol{e}}|_{\Omega_i}$, satisfies the local form of the (projected) error residual equation (15) given as

$$a_i(\hat{\boldsymbol{e}}|_{\Omega_i}, \boldsymbol{v}) = \hat{R}_i(\boldsymbol{v}) \quad \forall \boldsymbol{v} \in \mathcal{V}_i. \tag{20}$$

Note that this a pure Neumann problem for each integration cell Ω_i. However, the extended residual $\hat{R}_i$ involves the generally unknown traction field $\boldsymbol{t}_l$. Therefore, we replace the exact tractions $\boldsymbol{t}_l$ in (19) by a computable traction field on each edge of the integration cell $E_l \subset \partial\Omega_i$, denoted by $\tilde{\boldsymbol{t}}_{l,h}$. The residual (19) then turns into

$$\tilde{R}_i(\boldsymbol{v}) = \int_{\Omega_i} \boldsymbol{f}|_{\Omega_i} \cdot \boldsymbol{v} \, \mathrm{d}V + \sum_{l=1}^{n_l} \int_{E_l \subset \partial\Omega_i} \tilde{\boldsymbol{t}}_{l,h} \cdot \boldsymbol{v} \, \mathrm{d}A - \int_{\Omega_i} \boldsymbol{\sigma}(\boldsymbol{u}_h|_{\Omega_i}) : \boldsymbol{\varepsilon}(\boldsymbol{v}) \, \mathrm{d}V. \tag{21}$$

From the requirement that the sum of the local residuals over all integration cells should match the global residual, it follows that $\tilde{\boldsymbol{t}}_{l,h}$ need to be compatible between the integration cells and that they need to fulfill the Neumann boundary conditions. The first requirement is already fulfilled by the RKPM tractions and the second requirement is obviously trivial to fulfill.

Replacing $\hat{R}_i$ by the computable residual $\tilde{R}_i$ in the local error residual equation (20) then yields the local problem

$$a_i(\boldsymbol{\psi}_i, \boldsymbol{v}) = \tilde{R}_i(\boldsymbol{v}) \quad \forall \boldsymbol{v} \in \mathcal{V}_i, \tag{22}$$

which we solve for a solution $\boldsymbol{\psi}_i \in \mathcal{V}_{E,i}$ that can be seen as an approximation of $\hat{\boldsymbol{e}}|_{\Omega_i}$ depending on the accuracy of the tractions $\tilde{\boldsymbol{t}}_{l,h}$. However, as mentioned above, (22) is

a pure Neumann problem. The solvability of (22) thus requires that the computable tractions $\tilde{\boldsymbol{t}}_{l,h}$ are also equilibrated, which results in additional computational efforts.

The summation of the local problems (22) over all n_i integration cells and subsequent application of the Cauchy-Schwarz inequality then yields the constant-free energy-norm estimator for the projected discretization error

$$\|\!|\hat{\boldsymbol{e}}|\!\| \leq \left(\sum_{n_i} a_i(\boldsymbol{\psi}_i, \boldsymbol{\psi}_i) \right)^{\frac{1}{2}}, \tag{23}$$

which bears resemblance to its finite element counterpart as originally introduced by Bank & Weiser [6] and Ainsworth & Oden [1].

From (17) we then immediately infer that the discretization error $\boldsymbol{e}$, measured in the energy norm, can be bounded from above as

$$\|\!|\boldsymbol{e}|\!\| \leq \left(\sum_{n_i} \eta_i \right)^{\frac{1}{2}} + C \left(\sum_{E_l \subset \Gamma_D} h_E^{-1} \eta_E \right)^{\frac{1}{2}}. \tag{24}$$

Here, $\eta_i = a_i(\boldsymbol{\psi}_i, \boldsymbol{\psi}_i)$ is the error estimator in the domain and $\eta_E = \|\hat{\boldsymbol{e}}\|^2_{L_2(E)}$ is the computable error on the Dirichlet boundary Γ_D.

Note that, in general, it is not required that the local problems (22) are solved with the same numerical method as used to approximate the model problem (3). Since the local problems are Neumann problems, the finite element method or RKPM can both be used with a sufficiently high polynomial order.

4. Goal-oriented a posteriori error estimation

For the practical engineer, further error measures than the energy norm are often of bigger interest, e.g. the error of the fracture criterion within the framework of linear elastic fracture mechanics (LEFM), see Stone & Babuška [13]. In the terminology of goal-oriented error estimation, the fracture criterion is an example of a quantity of interest and our aim is to control the error of such quantities of interest. The derivation of the goal-oriented error estimator follows the well-established duality strategy as originally elaborated by Eriksson et al. [9] and others.

4.1. The dual problem based on a multi-space approach

Let the quantity of interest be given by the linear or linearized functional Q defined on $\mathcal{V}_0$. Then the dual problem of (3) asks to find a solution $\overset{*}{\boldsymbol{u}} \in \mathcal{V}_0$ such that

$$a(\boldsymbol{v}, \overset{*}{\boldsymbol{u}}) = Q(\boldsymbol{v}) \quad \forall \boldsymbol{v} \in \mathcal{V}_0. \tag{25}$$

Note that since the linear elasticity problem is self adjoint, a is symmetric and thus (25) results from the primal problem (3) by simply replacing the right-hand side.

For the associated meshfree RKPM discretization of the above dual problem (25), we introduce the finite-dimensional subspace $\overset{*}{\mathcal{V}}_h \subset \mathcal{V}$. In the most general case, $\overset{*}{\mathcal{V}}_h$ is different from $\mathcal{V}_h$, which means that, compared to the discretization of the primal problem (3), we may use a different set of particles and even different RK shape functions for the Galerkin approximation of the dual solution. The homogeneous Dirichlet boundary conditions (2a) are again weakly imposed using Nitsche's method, which results in the discrete problem of finding a solution $\overset{*}{\boldsymbol{u}}_h \in \overset{*}{\mathcal{V}}_h$ such that

$$a_h(\boldsymbol{v}_h, \overset{*}{\boldsymbol{u}}_h) = Q(\boldsymbol{v}_h) \quad \forall \boldsymbol{v}_h \in \overset{*}{\mathcal{V}}_h \tag{26}$$

with associated discretization error $\overset{*}{\boldsymbol{e}} = \overset{*}{\boldsymbol{u}} - \overset{*}{\boldsymbol{u}}_h$.

4.2. The goal-oriented error estimator

To estimate the error of the quantity of interest Q, the general strategy is to set $\boldsymbol{v} = \boldsymbol{e}$ in the dual problem (25). Note, however, that $\boldsymbol{v}$ is required to be in the space $\mathcal{V}_0$ in (25), whereas $\boldsymbol{e}$ is an element of $\mathcal{V}_E$. Similar to the error residual equation in Section 3.1, we therefore need to extend the dual problem (25) to the case where $\boldsymbol{v}$ can be chosen from $\mathcal{V}$, which results in

$$\begin{aligned} Q(\boldsymbol{e}) = a(\boldsymbol{e}, \overset{*}{\boldsymbol{e}}) &- \int_{\Gamma_D} \boldsymbol{e} \cdot \boldsymbol{\sigma}(\overset{*}{\boldsymbol{e}}) \cdot \boldsymbol{n} \, \mathrm{d}A - \int_{\Gamma_D} \overset{*}{\boldsymbol{e}} \cdot \boldsymbol{\sigma}(\boldsymbol{e}) \cdot \boldsymbol{n} \, \mathrm{d}A \\ &+ \hat{R}(\overset{*}{\boldsymbol{u}}_h) - \int_{\Gamma_D} \boldsymbol{e} \cdot \boldsymbol{\sigma}(\overset{*}{\boldsymbol{u}}_h) \cdot \boldsymbol{n} \, \mathrm{d}A. \end{aligned} \tag{27}$$

Note that the last two terms in (27) are exactly computable.

As can be observed from (27), the (extended) residual of the primal problem $\hat{R}$ needs to be computed in terms of the RKPM solution of the dual problem $\overset{*}{\boldsymbol{u}}_h$. This mainly requires to integrate the dual RKPM solution $\overset{*}{\boldsymbol{u}}_h$ using the integration cells Ω_i of the primal problem. Since the particles are independent from the mesh, this computation is "a piece of cake", as opposed to a mesh-based method, where the transfer of the solution from one mesh to the other mesh can be tedious, see [12] for more details in this respect.

The first three terms on the right-hand side of (27) can now be estimated using the Cauchy-Schwarz inequality and an inverse estimate, see [11]. If the dual discretization is kept constant during adaptive refinements, then we arrive at the error estimate

$$|Q(\boldsymbol{e})| \leq C_1 |\!|\!| \hat{\boldsymbol{e}} |\!|\!| + C_2 \|\hat{\boldsymbol{e}}\|_{\frac{1}{2},h} + |R(\overset{*}{\boldsymbol{u}}_h) - \int_{\Gamma_D} \boldsymbol{e} \cdot \boldsymbol{\sigma}(\overset{*}{\boldsymbol{u}}_h) \cdot \boldsymbol{n} \, \mathrm{d}A|, \tag{28}$$

where the positive constants C_1 and C_2 include several constants. The first term in the above estimate can be estimated using the energy-norm error estimator (23).

If, in addition, the errors on the Dirichlet boundary vanish, this estimate can be simplified even further to

$$|Q(\boldsymbol{e})| \leq C_1 |\!|\!| \hat{\boldsymbol{e}} |\!|\!| + |R(\overset{*}{\boldsymbol{u}}_h)| \tag{29}$$

with constant $C_1 = |\!|\!| \overset{*}{\hat{\boldsymbol{e}}} |\!|\!|$.

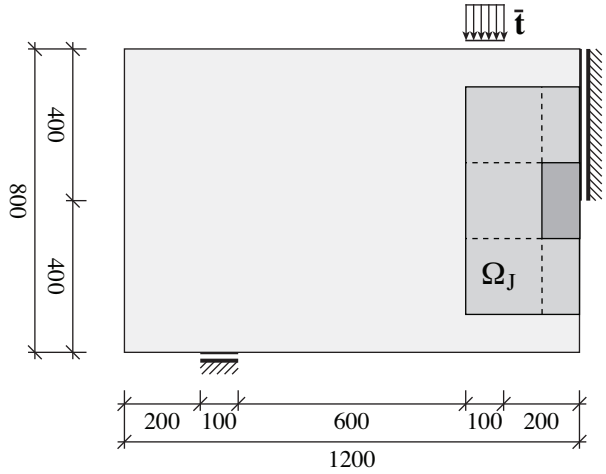


Figure 1: System and loading, measurements in mm

Figure 2: Primal error distribution

Figure 3: Dual error distribution

5. Numerical example: 4-point bending

In this section, we aim at investigating the goal-oriented error estimator for the case of the J-integral as an example for a nonlinear quantity of interest in LEFM. In a material force setting within Eshelbian mechanics, the J-integral takes the form

$$J(\boldsymbol{u}) = -\int_{\Omega_J} \boldsymbol{\Sigma}(\boldsymbol{u}) : \boldsymbol{H}(q\bar{\boldsymbol{x}})\, \mathrm{d}A. \tag{30}$$

Here, q is a C^0-function with $q = 1$ at the crack tip and $q = 0$ on $\Gamma_J = \partial\Omega_J \setminus \Gamma_c$. Moreover, $\boldsymbol{H}(\cdot) = \boldsymbol{\nabla}(\cdot)$ is the gradient tensor and $\boldsymbol{\Sigma} = W_s \boldsymbol{I} - \boldsymbol{H}^T \cdot \boldsymbol{\sigma}$ is the Newton-Eshelby stress tensor with the specific strain-energy function $W_s = 1/2\boldsymbol{\varepsilon} : \mathbb{C} : \boldsymbol{\varepsilon}$ and the identity tensor $\boldsymbol{I}$. The linearized quantity of interest Q is then defined as

$$Q(\boldsymbol{v}) = -\int_{\Omega_J} \boldsymbol{\Sigma}_{\mathrm{lin}}(\boldsymbol{u}_h) : \boldsymbol{H}(\boldsymbol{v})\, \mathrm{d}A, \tag{31}$$

where we introduced the linearized stress tensor

$$\boldsymbol{\Sigma}_{\mathrm{lin}}(\boldsymbol{u}_h) = \mathrm{div}(q\bar{\boldsymbol{x}})\boldsymbol{\sigma}(\boldsymbol{u}_h) - \boldsymbol{\sigma}(\boldsymbol{u}_h) \cdot \boldsymbol{H}^T(q\bar{\boldsymbol{x}}) - \boldsymbol{H}(q\bar{\boldsymbol{x}}) : \left[\boldsymbol{H}^T(\boldsymbol{u}_h) \cdot \mathbb{C}\right]. \tag{32}$$

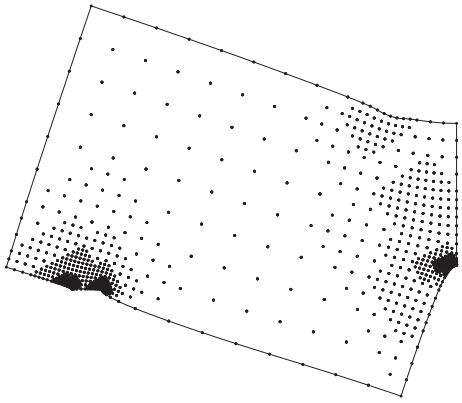


Figure 4: Adaptive primal refinement

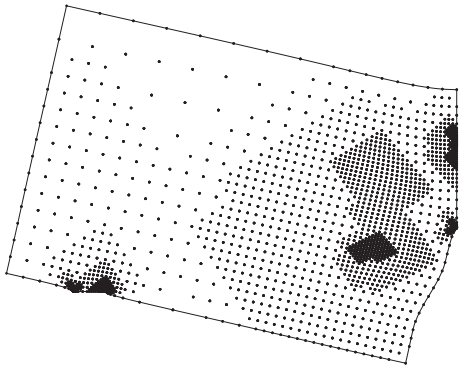


Figure 5: Adaptive dual refinement

The system in this example is a pre-cracked glass plate in plane-stress state subjected to 4-point bending, as illustrated in Figure 1. The material data is given in terms of Young's modulus $E = 64.000$ N/mm^2 and Poisson's ratio $\nu = 0.2$. The load in this example is $|\bar{\boldsymbol{t}}| = 1$ N/mm^2. The reference value $J(\boldsymbol{u}) = 0.016186862$ kJ/m^2 was computed using an adaptively refined Q_2 finite element mesh with 2.434.140 degrees of freedom (for the discretized half of the system).

As can be seen, the linearization depends on the solution of the primal problem. Therefore, one needs to solve the primal problem, use its solution to create the load of the dual problem, solve the dual problem and use the solution of both the primal and the dual problem to estimate the error of the J-integral. These steps can be easily carried out using the proposed multi-space RKPM approach.

In Figs. 2 and 3 the distribution of the exact error in each integration cell is visualized for both the primal and the dual problem (darker areas indicate larger errors). Accordingly, it can be expected that the error estimator finds the regions and refines the particles where large errors appear. This can be verified in terms of the 13-*th* primal refinement and 10-*th* dual refinement, see Figs. 4 and 5, respectively.

A comparison of various methods to deal with the dual problem is plotted in Figs. 6 and 7. It can be observed that the convergence rate is decreased when the dual discretization remains constant, which becomes clear from the error estimate (28). It can also be seen that when a fine dual discretization is used, the error is much smaller and thus the error tolerance could be reached with far less refinement steps compared to the other cases. A coarse and constant dual discretization, on the other hand, has the advantage of being computationally inexpensive. For a higher convergence rate, especially in the case of adaptive refinements, it is recommendable to refine both the primal and the dual discretization. The presented multi-space approach allows in this case to refine both problems independently and conveniently, which is not straightforwardly possible using a mesh-based method. Moreover, the refinement process itself is much easier using a meshfree method, because particles can be easily added and even removed from the discretization.

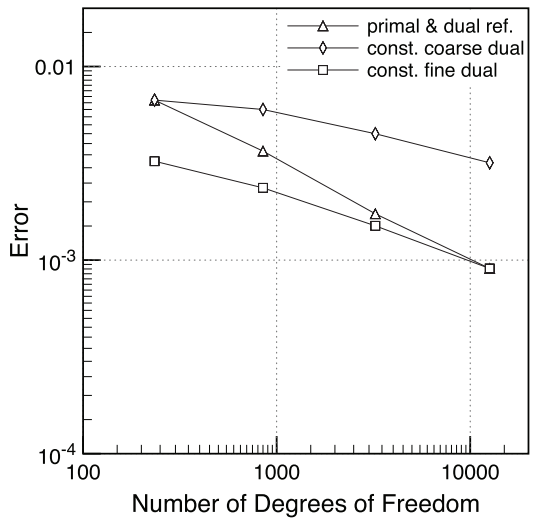


Figure 6: Estimated error $Q(\boldsymbol{e}) \approx J(\boldsymbol{u}) - J(\boldsymbol{u}_h)$, uniform refinements

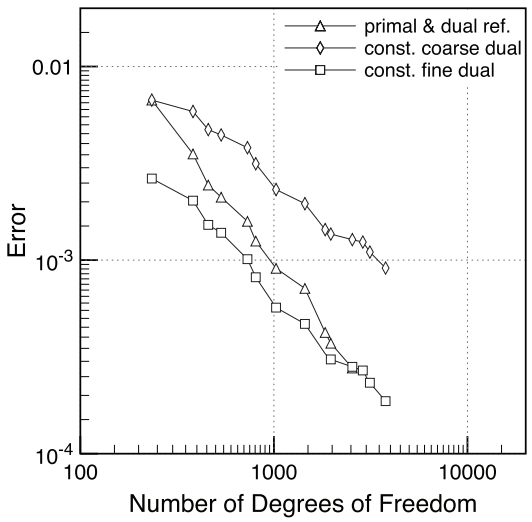


Figure 7: Estimated error $Q(\boldsymbol{e}) \approx J(\boldsymbol{u}) - J(\boldsymbol{u}_h)$, adaptive refinements

6. Conclusions

In this paper, both an implicit energy-norm and a goal-oriented *a posteriori* error estimator for RKPM approximations were derived and implemented. As it turned out, the main problem in the derivation of the error estimator is the violation of Dirichlet boundary conditions in meshfree methods. To cope with this problem, a projected error and thus a projected error residual equation were introduced. Since the RKPM particles are independent from the integration cells, a multi-space approach could easily be established allowing to use different discretizations for the primal and the dual problem. The error estimator was successfully applied to the J-integral as a crack propagation criterion in LEFM, offering the possibility to use only one (coarse or fine) dual solution within the refinement scheme and thus either obtaining less accurate but inexpensive (for the coarse solution) or more accurate but also more expensive (for the fine solution) error estimates.

Acknowledgements

The support of this work by DFG (German Research Foundation) under the grant no. RU 1213/2-1 is very much appreciated.

References

[1] Ainsworth, M. and Oden, J. T.: *A posteriori error estimation in finite element analysis*. John Wiley & Sons, New York, 2000.

[2] Babuška, I., Banerjee, U., and Osborn, J.: Survey of meshless and generalized finite element methods: A unified approach. Acta Numer. (2003), 1–125.

[3] Babuška, I. and Rheinboldt, W. C.: A-posteriori error estimates for the finite element method. Int. J. Numer. Meth. Engng **12** (1978), 1597–1615.

[4] Babuška, I. and Rheinboldt, W. C.: Error estimates for adaptive finite element computations. SIAM J. Numer. Anal. **15** (1978), 736–754.

[5] Babuška, I. and Strouboulis, T.: *The finite element method and its reliability.* Oxford University Press, Oxford, 2001.

[6] Bank, R. E. and Weiser, A.: Some a posteriori error estimators for elliptic partial differential equations. Math. Comp. **44** (1985), 283–301.

[7] Chen, J. S., Pan, C., Wu, C. T., and Liu, W. K.: Reproducing Kernel Particle Methods for large deformation analysis of non-linear structures. Comput. Methods Appl. Mech. Engrg. **139** (1996), 195–227.

[8] Duarte, C. A. and Oden, J. T.: An *h-p* adaptive method using clouds. Comput. Methods Appl. Mech. Engrg. **139** (1996), 237–262.

[9] Eriksson, K., Estep, D., Hansbo, P., and Johnson, C.: Introduction to adaptive methods for differential equations. Acta Numer. (1995), 106–158.

[10] Nitsche, J. A.: Über ein Variationsprinzip zur Lösung von Dirichlet-Problemen bei Verwendung von Teilräumen, die keinen Randbedingungen unterworfen sind. Abh. Math. Semin. Univ. Hambg. **36** (1971), 9–15.

[11] Rüter, M. and Chen, J. S.: A goal-oriented error estimator for meshfree methods based on a multi-space approach. (Submitted to Comput. Methods Appl. Mech. Engrg.) (2015).

[12] Rüter, M., Korotov, S., and Steenbock, C.: Goal-oriented error estimates based on different FE-spaces for the primal and the dual problem with applications to fracture mechanics. Comput. Mech. **39** (2007), 787–797.

[13] Stone, T. J. and Babuška, I.: A numerical method with a posteriori error estimation for determining the path taken by a propagating crack. Comput. Methods Appl. Mech. Engrg. **160** (1998), 245–271.

[14] Vidal, Y., Parés, N., Díez, P., and Huerta, A.: Bounds for quantities of interest and adaptivity in the element-free Galerkin method. Int. J. Numer. Meth. Engng. **76** (2008), 1782–1818.

Conference Applications of Mathematics 2015, in honor of the birthday anniversaries of Ivo Babuška (90), Milan Práger (85), and Emil Vitásek (85)
J. Brandts, S. Korotov, M. Křížek, K. Segeth, J. Šístek, T. Vejchodský (Eds.)
Institute of Mathematics AS CR, Prague 2015

ON THE NUMBER OF STATIONARY PATTERNS IN REACTION-DIFFUSION SYSTEMS

Vojtěch Rybář, Tomáš Vejchodský

Institute of Mathematics, Czech Academy of Sciences
Žitná 25, Prague 1, 115 67, Czech Republic
{rybar,vejchod}@math.cas.cz

Abstract: We study systems of two nonlinear reaction-diffusion partial differential equations undergoing diffusion driven instability. Such systems may have spatially inhomogeneous stationary solutions called Turing patterns. These solutions are typically non-unique and it is not clear how many of them exists. Since there are no analytical results available, we look for the number of distinct stationary solutions numerically. As a typical example, we investigate the reaction-diffusion system designed to model coat patterns in leopard and jaguar.

Keywords: diffusion driven instability, Turing patterns, classification of non-unique solutions

MSC: 35A02, 35K57, 35Q92

1. Introduction

Nonlinear systems of reaction-diffusion equations are universally recognized instruments for modelling various phenomena in chemistry, biology, and ecology. Their popular applications include, but are not limited to, symmetry breaking, biochemical reactions, tumour vascularization, predator-prey models or skin and coat patterns in animals.

Research of the diffusion driven instability initiated Alan Turing in 1952 by his seminal paper [13], where he presented a counter-intuitive property of systems of two reaction-diffusion equations of the form

$$\frac{\partial u}{\partial t} = D_1 \Delta u + f(u,v) \quad \text{in } (0,\infty)\times\Omega, \tag{1}$$

$$\frac{\partial v}{\partial t} = D_2 \Delta v + g(u,v) \quad \text{in } (0,\infty)\times\Omega, \tag{2}$$

where $u = u(t,x)$, $v = v(t,x)$ correspond to concentrations of two chemical species, the domain $\Omega \subset \mathbb{R}^2$ models a chemical reactor, D_1, D_2 are diffusion coefficients

and $f(u,v)$, $g(u,v)$ are nonlinear reaction terms representing chemical reactions. We assume existence of constants $u_s, v_s \in \mathbb{R}$ such that $f(u_s, v_s) = g(u_s, v_s) = 0$. Clearly, these constants form a stationary solution of system (1)–(2) known as the ground state. In addition, these constants can be seen as a solution to the system of ordinary differential equations (ODE) coming from (1)–(2) for $D_1 = D_2 = 0$. Turing demonstrated that if u_s, v_s is a linearly stable uniform stationary solution of this ODE system then u_s, v_s seen as a solution to system (1)–(2) can become unstable for $D_1 \neq 0$, $D_2 \neq 0$ and its small spatially inhomogeneous perturbations may evolve to an inhomogeneous steady state. Such steady state is known as Turing pattern. However, the diffusion driven instability can occur only for a limited set of values of D_1 and D_2 and possible other parameters of (1)–(2). These values can be identified by means of the well known linear analysis, see e.g. [7].

There is a number of models exhibiting the diffusion driven instability and Turing patterns. For example, Thomas model [11] of substrate inhibition, Schnakenberg model [10] describing a hypothetical trimolecular reaction, Gray and Scott model [3] for an autocatalytic reaction in a tank reactor, BMA model [1] for symmetry breaking in morphogenesis and LLM model [5] for pigment pattern generation on coats of leopards and jaguars.

The mentioned linear analysis describes well the initial evolution of small perturbations of the ground state, but it yields no information about their development if they grow sufficiently large. Existing analytical results about solutions farther away from the ground state are limited. Therefore, we study them numerically. We are then limited to an empirical study of a particular case only, but we aim to collect a large amount of data, process them by statistical methods and draw more general conclusions.

In this contribution we present results of a particular numerical study focused on the number of distinct Turing patters in the LLM model [5]. In this experiment we initiated the evolution with several thousand distinct initial conditions, solve them by a fast numerical scheme, and postprocess the obtained results by nontrivial methods to identify patterns that are identical up to natural symmetries of the problem. The following section introduces periodic boundary conditions and the corresponding symmetries of the problem. Section 3 briefly describes the spectral Fourier method. Section 4 introduces the LLM model, its particular setting, and explains how the computed results are postprocessed and analysed. Section 5 presents the obtained results, especially the numbers of distinct equivalence classes of stationary solutions. Finally, Section 6 draws conclusions and offers prospects for further research.

2. Boundary conditions and problem symmetries

Within this paper, we consider the domain Ω to be a square $\Omega = (0, L)^2$. The reaction-diffusion system (1)–(2) is usually equipped with no flux boundary conditions. However, we will consider periodic boundary conditions, because they are

natural for the spectral Fourier method, which we will use below. In particular, we consider these periodic boundary conditions:

$$u(0,y) = u(L,y) \quad \forall y \in (0,L) \quad \text{and} \quad u(x,0) = u(x,L) \quad \forall x \in (0,L), \tag{3}$$
$$v(0,y) = v(L,y) \quad \forall y \in (0,L) \quad \text{and} \quad v(x,0) = v(x,L) \quad \forall x \in (0,L), \tag{4}$$
$$\partial_x u(0,y) = \partial_x u(L,y) \quad \forall y \in (0,L) \quad \text{and} \quad \partial_y u(x,0) = \partial_y u(x,L) \quad \forall x \in (0,L), \tag{5}$$
$$\partial_x v(0,y) = \partial_x v(L,y) \quad \forall y \in (0,L) \quad \text{and} \quad \partial_y v(x,0) = \partial_y v(x,L) \quad \forall x \in (0,L). \tag{6}$$

Since the square domain and these periodic boundary conditions are invariant with respect to mirroring and rotations by $\pi/2$, we show that the stationary solutions of problem (1)–(2) with boundary conditions (3)–(6) possess the same symmetries. To be rigorous, we define the mirror image of a function u defined in Ω by

$$\bar{u}(x,y) = u(L-x,y). \tag{7}$$

Similarly, we define a function rotated by $\pi/2$ (counter-clockwise) as

$$\hat{u}(x,y) = u(L-y,x). \tag{8}$$

Further, the periodic boundary conditions enable to shift a stationary solution periodically in such a way that it remains a stationary solution. To be precise, we define a periodic shift of a function u by a vector $(r,s) \in (0,L)^2$ as

$$\tilde{u}(x,y) = \begin{cases} u(x+r, y+s) & \text{for } x \in (0, L-r), y \in (0, L-s), \\ u(x+r, y+s-L) & \text{for } x \in (0, L-r), y \in (L-s, L), \\ u(x+r-L, y+s) & \text{for } x \in (L-r, L), y \in (0, L-s), \\ u(x+r-L, y+s-L) & \text{for } x \in (L-r, L), y \in (L-s, L). \end{cases} \tag{9}$$

Lemma 1. *Let $u, v \in C^2(\Omega) \cap C^0(\overline{\Omega})$ form a stationary solution to problem (1)–(2) with boundary conditions (3)–(6). Then both pairs of functions $\bar{u}, \bar{v}$ and $\hat{u}, \hat{v}$ defined by (7) and (8), respectively, are stationary solutions to (1)–(2) with (3)–(6) as well. Moreover, if the shifted functions $\tilde{u}, \tilde{v}$ given by (9) with arbitrary $(r,s) \in (0,L)^2$ are both in $C^2(\Omega)$ then they again form a stationary solution to (1)–(2) with (3)–(6).*

Proof. It is easy to verify that all $\bar{u}, \bar{v}$ and $\hat{u}, \hat{v}$ are in $C^2(\Omega) \cap C^0(\overline{\Omega})$ and that they satisfy the periodic boundary conditions (3)–(6). Further, let $(x,y) \in \Omega$. Then it is easy to see that $\Delta\bar{u}(x,y) = \Delta u(L-x,y)$, $\Delta\hat{u}(x,y) = \Delta u(L-y,x)$, and similarly for $\bar{v}$ and $\hat{v}$. Since u, v satisfy equations (1)–(2) at both points $(L-x,y) \in \Omega$ and $(L-y,x) \in \Omega$, we conclude that both pairs $\bar{u}, \bar{v}$ and $\hat{u}, \hat{v}$ satisfy the same equations at (x,y).

Concerning the shifted functions $\tilde{u}, \tilde{v}$ we may proceed in the same way. The only difficulty is the fact that the periodic shift $\tilde{u}, \tilde{v}$ need not automatically be in $C^2(\Omega)$ and therefore the additional assumption is needed. □

3. Spectral Fourier collocation method

To solve problem (1)–(2) with periodic boundary conditions (3)–(6) efficiently, we employ the spectral Fourier collocation method [4, 12]. In order to briefly introduce the main idea of the method, we consider a 2π-periodic function $z(x)$ sampled on the spatial discretization grid $x_j = 2\pi j/N$ with $z_j = z(x_j)$, $j = 0, 1, 2, \ldots, N$. Note that we consider N to be even for simplicity. By periodicity of $z(x)$ we have $z_0 = z_N$. Using the discrete Fourier transform (DFT) and the inverse discrete Fourier transform (IDFT), both properly defined and discussed in [12], we can compute the derivatives $w_j = z'(x_j)$, $j = 1, \ldots, N$, by the following procedure:

1. Compute the DFT $\hat{z}_k = (2\pi/N) \sum_{j=1}^{N} \exp(-\mathrm{i}kx_j) z_j,\ k = -N/2+1, \ldots, N/2.$
2. Set $\hat{w}_k = \mathrm{i}k\hat{z}_k,\ k = -N/2+1, \ldots, N/2.$
3. Compute the IDFT $w_j = (1/(2\pi)) \sum_{k=-N/2+1}^{N/2} \exp(\mathrm{i}kx_j)\hat{w}_k,\ j = 1, \ldots, N.$

Similarly, the second derivative $w_j = z''(x_j)$ can be computed by the same procedure, but item 2 has to be replaced by $\hat{w}_k = -k^2\hat{z}_k,\ k = -N/2+1, \ldots, N/2.$

This idea can be easily applied in two dimensions as well. Let us consider partitions $x_m = 2\pi m/N$ and $y_n = 2\pi n/N$, $m, n = 1, 2, \ldots, N$, of $[0, 2\pi]$ corresponding to x and y directions. Set $u_{m,n} = u(x_m, y_n)$, $v_{m,n} = v(x_m, y_n)$, $f_{m,n} = f(u_{m,n}, v_{m,n})$, and $g_{m,n} = g(u_{m,n}, v_{m,n})$. The two-dimensional DFT of $u_{m,n}$ is defined as

$$\hat{u}_{k,\ell} = \frac{4\pi^2}{N^2} \sum_{m=1}^{N} \sum_{n=1}^{N} \exp(-\mathrm{i}(kx_m + \ell y_n)) u_{m,n}, \quad k, \ell = -N/2+1, \ldots, N/2, \tag{10}$$

and similarly for $\hat{v}_{k,\ell}$, $\hat{f}_{k,\ell}$, and $\hat{g}_{k,\ell}$. Correspondingly, the two-dimensional IDFT of $\hat{u}_{k,\ell}$ is

$$u_{m,n} = \frac{1}{4\pi^2} \sum_{k=-N/2+1}^{N/2} \sum_{\ell=-N/2+1}^{N/2} \exp(\mathrm{i}(kx_m + \ell y_n)) \hat{u}_{k,\ell}, \quad m, n = 1, 2, \ldots, N. \tag{11}$$

In order to use the DFT (10) for equations (1) and (2), we first transform variables to map $[0, L]$ into $[0, 2\pi]$ and then we obtain the following system of ordinary differential equations for the Fourier images $\hat{u}_{k,\ell}$ and $\hat{v}_{k,\ell}$:

$$\frac{\mathrm{d}\hat{u}_{k,\ell}}{\mathrm{d}t} = -D_1 \frac{4\pi^2}{L^2}(k^2 + \ell^2)\hat{u}_{k,\ell} + \hat{f}_{k,\ell} \quad k, \ell = -N/2+1, \ldots, N/2, \tag{12}$$

$$\frac{\mathrm{d}\hat{v}_{k,\ell}}{\mathrm{d}t} = -D_2 \frac{4\pi^2}{L^2}(k^2 + \ell^2)\hat{v}_{k,\ell} + \hat{g}_{k,\ell} \quad k, \ell = -N/2+1, \ldots, N/2. \tag{13}$$

Here, $\hat{f}_{k,\ell}$ and $\hat{g}_{k,\ell}$ are computed by the DFT (10) from $f_{m,n} = f(u_{m,n}, v_{m,n})$ and $g_{m,n} = g(u_{m,n}, v_{m,n})$, where the values of $u_{m,n}$ and $v_{m,n}$ have to be computed from the Fourier images $\hat{u}_{k,\ell}$ and $\hat{v}_{k,\ell}$ by the IDFT (11).

In order to solve the system of ordinary differential equations (12) efficiently, we utilize the fourth order Runge-Kutta method as in [4] and the fast Fourier transform.

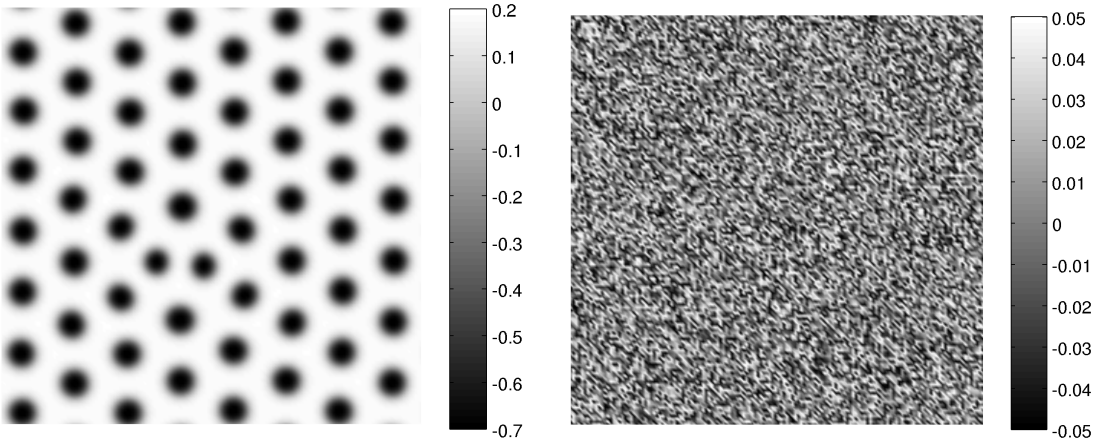


Figure 1: The left panel shows the component v of the stationary solution of problem (14)–(15) with periodic boundary conditions (3)–(6) as it evolved from a small random initial condition illustrated in the right panel.

4. Problem setting and postprocessing of results

For the numerical study presented below, we will consider the LLM model [5]. It consists of the following reaction-diffusion system:

$$\frac{\partial u}{\partial t} = D\delta\Delta u + \alpha u + v - r_2 uv - \alpha r_3 uv^2, \tag{14}$$

$$\frac{\partial v}{\partial t} = \delta\Delta v - \alpha u + \beta v + r_2 uv + \alpha r_3 uv^2 \tag{15}$$

with $D = 0.45$, $\delta = 6$, $\alpha = 0.899$, $\beta = -0.91$, $r_2 = 2$, and $r_3 = 3.5$. These parameter values yield stationary solutions that correspond to spotted patterns, see Figure 1 (left) for a typical pattern and Figure 1 (right) for the corresponding initial condition. Our main interest is to find how many different patterns can evolve from small random initial conditions. Note that the two components u and v are complementary to each other in the sense that local maxima of u correspond to local minima of v. Therefore, we concentrate on the component v only in what follows.

All patterns in this paper including Figure 1 are computed by the spectral Fourier collocation method with the following setting. The domain is a square $\Omega = (0, L)^2$ with $L = 200$. The discretization grid contains $N = 144$ points in every direction. Initial conditions are generated as a uniformly distributed random number within $(-0.05, 0.05)$ for every node of the grid. These initial conditions mimic small amplitude random fluctuations around the spatially homogeneous steady state. The time step of the fourth order Runge-Kutta method is chosen as $\Delta t = 1$ and the computation of the time evolution is terminated as soon as the relative difference of approximate solutions at two consecutive time steps is smaller than a prescribed tolerance. In particular, if $\|\cdot\|_{l^2}$ stands for the l^2-norm over the grid nodes and $v^{(k)}$

and $v^{(k+1)}$ denote the solution at times $t_k = k\Delta t$ and $t_{k+1} = (k+1)\Delta t$, then the computation is stopped if

$$\frac{\|v^{(k)} - v^{(k+1)}\|_{l^2}}{\|v^{(k)}\|_{l^2}} < 10^{-4}. \tag{16}$$

Using this setting, we generated a number of random initial conditions and computed the corresponding stationary solutions. However, based on the symmetries of the stationary solutions described in Lemma 1, each stationary solution represents a whole class of solutions equivalent up to one of the transformations (7)–(9). Therefore, finding the number of distinct stationary solutions, i.e. solutions that are not equivalent in the sense of transformations (7)–(9), is a nontrivial task. Especially challenging is the fact that any shift $(r, s) \in (0, L)^2$ yields a stationary solution and, hence, each class of solutions is uncountable.

5. Results

In total we computed stationary solutions for 5297 different random initial conditions. At first, we calculate the number of spots in each of these patterns. To compute this number we plot the v component of the given pattern as a bitmap image and utilize the Matlab Image Processing Toolbox [6]. In particular, we use the function `imfindcircles` which seeks circles in a given image and returns coordinates of their centres and radii. The number of spots is then simply equal to the number of returned centres. Having computed this number for all patterns, we then simply calculate how many patterns have a given number of spots. Figure 2 presents these data in the form of a histogram.

In this histogram we identify 15 possible numbers of spots. Number of spots varies between 50 and 65 with the intermediate numbers being naturally the most frequent. Surprising is the relatively wide range of these numbers. Researchers have usually a chance to observe a relatively small amount of patterns and then they tend to conclude that the number of spots is (almost) constant for the given parameter values and the size of the domain. However, our results show that it may vary considerably just due to the random variations in the initial condition. In this particular case, the variation in the number of spots is up to $\pm 15\,\%$.

Nevertheless, the main goal is to find the number of classes of solutions that are identical up to a combination of transformations (7)–(9). The computed number of spots serves as a first filter, because, clearly, if two patterns have a different number of spots, they cannot be equivalent. Thus, we split all patterns into 15 sets according to the number of spots and for each set we find classes of equivalent patterns as follows.

We keep a database of classes. Each class in this database is determined by one representative pattern. Initially the database is empty and the first pattern from the investigated set is chosen as the representative of the first class. Then for each pattern in the set, we find if it is equivalent to one of the stored representatives in the

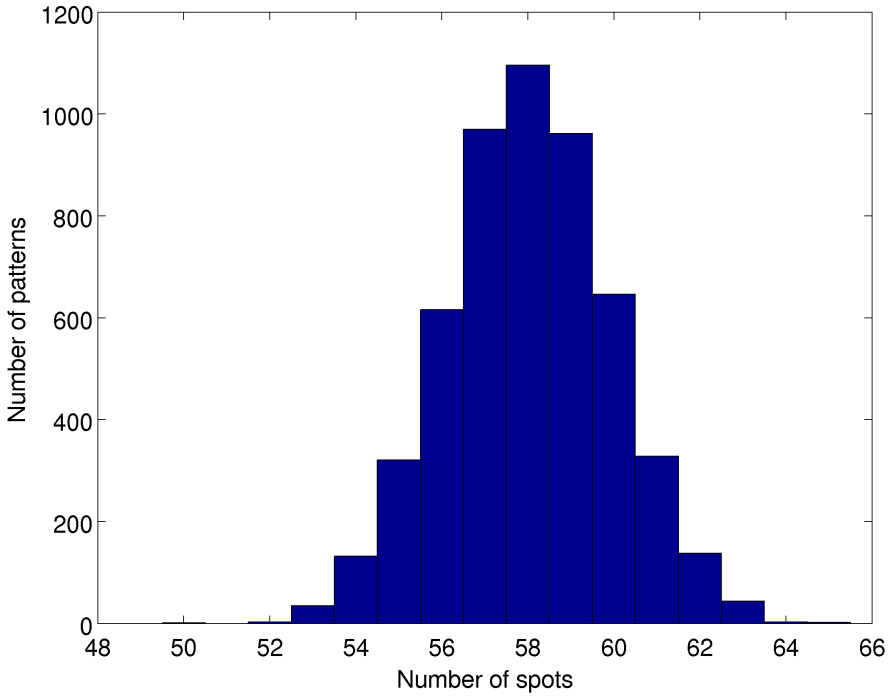


Figure 2: Distribution of the number of spots in stabilized patterns

database. If it is the case, we simply increase the counter of the number of patterns in the corresponding class. If not, then this pattern becomes a representative of a new class. As soon as we exhaust the entire investigated set of patterns, all classes of equivalent patterns are identified.

The important step in this algorithm is to decide whether a pattern is equivalent to a representative or not. In order to decide, we have to test all eight independent combinations of mirroring (7) and rotation (8) as well as all possible shifts (9). Fortunately, only the shifts that map a spot of the pattern to a spot of the representative are relevant and thus the total number of relevant shifts is finite and equals to the number of possible pairs of spots. If the number of spots in both the pattern and in the representative is n then the number of possible shifts is n^2.

The crucial operation here is the comparison of two patterns – the transformed pattern and the representative – and the decision whether they match or not. The point is that exact equality of two patterns cannot work here, because the patterns are polluted by various numerical errors. Therefore, we actually look for patterns matching up to a certain precision. We experimented with several measures and we obtained the best results by using the modified Hausdorff distance on the torus for the computed centres of spots. Note that the torus topology comes from the periodic boundary conditions (3)–(6).

The modified Hausdorff distance $\tilde{H}(A,B)$ in the torus is defined as

$$\tilde{H}(A,B) = \min\left\{\tilde{h}(A,B), \tilde{h}(B,A)\right\},$$

where $\tilde{h}(A,B)$ is the Hausdorff distance expressing the maximal distance between points from the set A to the set B, i.e.

$$\tilde{h}(A,B) = \max_{a\in A}\left\{\min_{b\in B}\tilde{d}(a,b)\right\},$$

where $\tilde{d}(a,b)$ is the torus distance of points $a=(a_1,a_2)$ and $b=(b_1,b_2)$:

$$\tilde{d}(a,b) = \left(\min\{|a_1-b_1|, L-|a_1-b_1|\}^2 + \min\{|a_2-b_2|, L-|a_2-b_2|\}^2\right)^{1/2}.$$

We use the modified Hausdorff distance rather then the Hausdorff distance itself, because it is symmetric and performs better [2]. In order to gain the required efficiency we use the Matlab mex-file implementation of the modified Hausdorff distance [9].

Using the modified Hausdorff distance on the torus, we compare two patterns based on the lists C_1 and C_2 of coordinates of centres of their spots. We simply compute $\tilde{H}(C_1,C_2)$ and test if it is below a chosen threshold. To be precise, we have to consider also the transformations (7)–(9). The two patterns to compare are determined by the lists C_1 and C_2 of their centres of spots. For the list C_1, we consider all its shifts (9) with $(r,s) = (b_1-a_1, b_2-a_2)$ for all $(a_1,a_2)\in C_1$ and all $(b_1,b_2)\in C_2$ together with eight possible combinations of mirroring (7) and rotations (8). Denoting the set of all these transformations by $\mathcal{R}$, we compute

$$\tilde{H}(C_1,C_2) = \min_{\rho\in\mathcal{R}}\tilde{H}(\rho(C_1),C_2), \tag{17}$$

where $\rho(C_1)$ stands for the transformation of the set C_1.

Practically, we have chosen the above mentioned threshold to be 0.75 and consider two patterns to match if $\tilde{H}(C_1,C_2) < 0.75$. However, we have to admit that the choice of this threshold is delicate, because it is difficult to distinguish whether two patterns differ due to numerical errors or whether they really correspond to different stationary solutions. For illustration, we present Figure 3 showing two sets of centres of spots with the distance $\tilde{H}(C_1,C_2) \approx 0.74941$.

The final results are summarized in Table 1, where we identified the number of distinct classes of stationary solutions. We observe that the number of these classes varies roughly between 10 and 20 % of the total number of patterns in each set with a higher number of spots. This ratio is naturally higher in the rare cases where the number of identified patterns is low.

An interesting observation is that there are considerably different numbers of distinct classes for pairs of sets with comparable total numbers patterns. See for example the cases of 56 and 60 spots, which both have slightly above 600 patterns, but the first case has only 67 distinct classes of solutions in contrast to 96 classes in the second case.

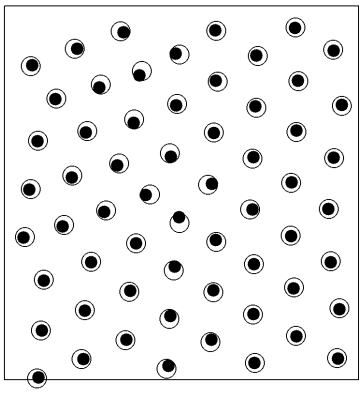


Figure 3: Centres of spots of two patterns with $\tilde{H}(C_1, C_2) \approx 0.74941$

Spots	50	52	53	54	55	56	57	58	59	60	61	62	63	64	65
Patterns	1	3	35	132	321	616	970	1096	962	646	328	138	44	3	2
Classes	1	3	6	22	34	67	70	123	115	96	34	27	20	3	2

Table 1: Number of classes for different number of spots in patterns

6. Conclusions

The results presented in Table 1 are especially interesting if they are compared with our previous results [8], where we performed a similar study, but for much smaller domain, namely, $\Omega = (0, 50)^2$. In that case, we obtained only two different numbers of spots and in total four distinct classes of solutions out of 6 000 computed patterns. In view of these results, the variety in both the number of spots and the number of classes of patterns we observe in Table 1 is rather surprising, even if we take into account the larger size of the domain Ω. These results strongly indicate that the number of distinct classes of patterns as well as the possible number of spots grow progressively with the size of the domain Ω.

Another observation is that the influence of boundary conditions on the final shape of the pattern is smaller if the domain is larger. Therefore, in the older study [8] we observe mainly the effects of boundary conditions, while in the current study we see mostly the natural variability of Turing patterns with the influence of boundary conditions being inferior.

Clearly, the presented results are burdened by uncertainties. For example, it is difficult to verify whether the numerical process of computing the stationary solution really converged. Even if the stopping criterion (16) is fulfilled, the pattern still might not be completely stationary. As a result, we can identify two patterns as distinct, but they both can eventually converge to the same stationary solution. This effect could hypothetically contribute to the observed high number of distinct classes of solutions.

Moreover, we obviously did not find all the possible classes of patterns. Particularly odd is the fact that we found classes with 50 and 52 spots, but none with 51 spots. Due to the nature of the problem, we believe that such patterns exist and that we just did not capture the rare stationary solutions. An experiment with a larger sample of computed patterns could help to estimate how many of these rare solutions we missed.

In the current numerical study we focused on the particular LLM model [5], but as a future project we plan to perform a similar study for other models. For example for the Thomas model [11]. Based on our experience, we expect that if we chose the problem parameters in such a way that the number of spots is similar to the current study, we obtain similar results.

Further, the obtained results can help us to understand additional subtle features of Turing patterns. For example, we can try to identify the natural period of the Turing patterns as a natural distance between two spots. This task can be accomplished by numerical methods and the results can help to find an analytical expression or estimate of the period and prove a corresponding theoretical result.

Acknowledgements

This work has been supported by grant SVV-2015-260226 of the Charles University Grant Agency and by RVO 67985840.

References

[1] Barrio, R., Maini, P., and Aragón, J.: Size-dependent symmetry breaking in models for morphogenesis. Physica D **2920** (2002), 1–12.

[2] Dubuisson, M. P. and Jain, A. K.: A modified Hausdorff distance for object matching. In: *Proc. of IAPR Int. Conf. on Pattern Recognition*, pp. A:566–568. Jerusalem, Israel, 1994.

[3] Gray, P. and Scott, S.: Autocatalytic reactions in the isothermal, continuous stirred tank reactor. Chem. Eng. Sci. **39** (1984), 1087–1097.

[4] Kassam, A. K.: Solving reaction-diffusion equations 10 times faster. Numerical Analysis Group Research Report **16** (Mathematical Institute, Oxford, 2003).

[5] Liu, R. T., Liaw, S. S., and Maini, P. K.: Two-stage Turing model for generating pigment patterns on the leopard and the jaguar. Physical Review **74** (2006), 011 914.

[6] MATLAB: *Image Processing Toolbox 2014b*. The MathWorks Inc., Natick, Massachusetts, United States, 2014.

[7] Murray, J. D.: *Mathematical biology. II. Spatial models and biomedical applications*. Springer-Verlag, New York, 2003.

[8] Rybář, V. and Vejchodský, T.: Variability of Turing patterns in reaction-diffusion systems. In: H. Bílková, M. Rozložník, and P. Tichý (Eds.), *Proceedings of the SNA'14*, pp. 87–90. Institute of Computer Science AS CR, Prague, 2014.

[9] Sasikanth, B.: Modified Hausdorff distance for 2D point sets. `https://www.mathworks.com/matlabcentral/fileexchange/30108`. Accessed 2015-05-25.

[10] Schnakenberg, J.: Simple chemical reaction systems with limit cycle behaviour. J. Theoret. Biol. **81** (1979), 389–400.

[11] Thomas, D.: Artificial enzyme membranes, transport, memory, and oscillatory phenomena. In: D. Thomas and J. P. Kernevez (Eds.), *Analysis and control of immobilized enzyme systems*, pp. 115–150. North Holland, Amsterdam, 1976.

[12] Trefethen, L. N.: *Spectral methods in MATLAB*. SIAM, Philadelphia, 2000.

[13] Turing, A. M.: The chemical basis of morphogenesis. Philosophical Transactions of the Royal Society B **237** (1952), 37–72.

Conference Applications of Mathematics 2015, in honor of the birthday anniversaries of Ivo Babuška (90), Milan Práger (85), and Emil Vitásek (85)
J. Brandts, S. Korotov, M. Křížek, K. Segeth, J. Šístek, T. Vejchodský (Eds.)
Institute of Mathematics CAS, Prague 2015

A NOTE ON TENSION SPLINE

Karel Segeth

Institute of Mathematics, Academy of Sciences
Žitná 25, CZ-115 67 Prague 1, Czech Republic
segeth@math.cas.cz

Abstract: Spline theory is mainly grounded on two approaches: the algebraic one (where splines are understood as piecewise smooth functions) and the variational one (where splines are obtained via minimization of quadratic functionals with constraints). We show that the general variational approach called smooth interpolation introduced by Talmi and Gilat covers not only the cubic spline but also the well known tension spline (called also spline in tension or spline with tension). We present the results of a 1D numerical example that show the advantages and drawbacks of the tension spline.

Keywords: smooth interpolation, tension spline, Fourier transform

MSC: 65D05, 65D07, 41A05

1. Introduction

In most practical cases, the minimum curvature (or cubic spline) method produces a visually pleasing smooth curve or surface. However, in some cases the method can create strong artificial oscillations in the curve derivative (surface gradient). A remedy suggested by Schweikert [6] is known as *tension spline*. The functional minimized includes the first derivative term in addition to the second derivative term.

Smooth approximation [10] is an approach to data interpolating or data fitting that employs the variational formulation of the problem in a normed space with constraints representing the approximation conditions. The cubic spline interpolation is also known to be the approximation of this kind.

For the 1D cubic spline, the objective is to minimize the L^2 norm of second derivative of the approximating function. A more sophisticated criterion is then to minimize, with some weights chosen, the integrals of the squared magnitude of some (or possibly all) derivatives of a sufficiently smooth approximating function. In the paper, we are concerned with the tension spline constructed by means of the smooth approximation theory (cf. also [4]), i.e. with the exact interpolation of the data at nodes and, at the same time, with the smoothness of the interpolating curve and its first derivative.

We are mostly interested in the case of a single independent variable in the paper. Assuming the approach of [8] and [10], we introduce the problem to be solved and the tools necessary to this aim in Sec. 3. We also present the general existence theorem for smooth interpolation proven in [8]. We are concerned with the use of basis system $\exp(\mathrm{i}kx)$ of exponential functions of pure imaginary argument for 1D smooth approximation problems in Secs. 4 and 5. We investigate some of its properties suitable for measuring the smoothness of the approximation and for generating the tension spline. We also show results of a 1D numerical experiment and discuss them to illustrate some properties of smooth approximation.

2. Problem of data approximation

Let us have a finite number N of (complex, in general) measured (sampled) values $f_1, f_2, \ldots, f_N \in C$ obtained at N nodes $X_1, X_2, \ldots, X_N \in R^n$. The nodes are assumed to be mutually distinct. We are usually interested also in the intermediate values corresponding to other points in some domain. Assume that $f_j = f(X_j)$ are measured values of some continuous function f while z is an approximating function to be constructed. The dimension n of the independent variable can be arbitrary. For the sake of simplicity we put $n = 1$ and assume that $X_1, X_2, \ldots, X_N \in \Omega$, where either $\Omega = [a, b]$ is a finite interval or $\Omega = (-\infty, \infty)$.

Data interpolation. The interpolating function z is constructed to fulfil the interpolation conditions

$$z(X_j) = f(X_j), \quad j = 1, \ldots, N. \tag{1}$$

Some additional conditions can be considered, e.g. the Hermite interpolation or minimization of some functionals applied to z.

The problem of data interpolation does not have a unique solution. The property (1) of the interpolating function is uniquely formulated by mathematical means but there are also requirements on the *subjective perception* of the behavior of the approximating curve or surface between nodes that can hardly be formalized [11].

3. Smooth approximation

We introduce an inner product space to formulate the additional constraints in the problem of smooth approximation [7, 8, 10]. Let $\widetilde{\mathcal{W}}$ be a linear vector space of complex valued functions g continuous together with their derivatives of all orders on the interval Ω. Let $\{B_l\}_{l=0}^{\infty}$ be a sequence of nonnegative numbers and L the smallest nonnegative integer such that $B_L > 0$ while $B_l = 0$ for $l < L$. For $g, h \in \widetilde{\mathcal{W}}$, put

$$(g, h)_L = \sum_{l=0}^{\infty} B_l \int_{\Omega} g^{(l)}(x)[h^{(l)}(x)]^* \,\mathrm{d}x, \tag{2}$$

$$|g|_L^2 = \sum_{l=0}^{\infty} B_l \int_{\Omega} |g^{(l)}(x)|^2 \,\mathrm{d}x, \tag{3}$$

where * denotes the complex conjugate.

If $L = 0$ (i.e. $B_0 > 0$), consider functions $g \in \widetilde{\mathcal{W}}$ such that the value of $|g|_0$ exists and is finite. Then $(g,h)_0 = (g,h)$ has the properties of *inner product* and the expression $|g|_0 = \|g\|$ is *norm* in a normed space $W_0 = \widetilde{\mathcal{W}}$.

Let $L > 0$. Consider again functions $g \in \widetilde{\mathcal{W}}$ such that the value of $|g|_L$ exists and is finite. Let $P_{L-1} \subset \widetilde{\mathcal{W}}$ be the subspace whose basis $\{\varphi_p\}$ consists of monomials

$$\varphi_p(x) = x^{p-1}, \quad p = 1, \ldots, L.$$

Then $(\varphi_p, \varphi_q)_L = 0$ and $|\varphi_p|_L = 0$ for $p, q = 1, \ldots, L$. Using (2) and (3), we construct the *quotient space* $\widetilde{\mathcal{W}}/P_{L-1}$ whose zero class is the subspace P_{L-1}. Finally, considering $(\cdot,\cdot)_L$ and $|\cdot|_L$ in every equivalence class, we see that they represent the inner product and norm in the normed space $W_L = \widetilde{\mathcal{W}}/P_{L-1}$.

W_L is the normed space where we minimize functionals and measure the smoothness of the interpolation. For an arbitrary $L \geq 0$, choose a *basis system* of functions $\{g_k\} \subset W_L$, $k = 1, 2, \ldots$, that is complete and orthogonal (in the inner product in W_L), i.e., $(g_k, g_m)_L = 0$ for $k \neq m$, $(g_k, g_k)_L = |g_k|_L^2 > 0$. If $L > 0$ then it is, moreover, $(\varphi_p, g_k)_L = 0$ for $p = 1, \ldots, L$, $k = 1, 2, \ldots$. The set $\{\varphi_p\}$ is empty for $L = 0$.

Smooth data interpolation. The *problem of smooth data interpolation* [10] consists in finding the coefficients A_k and a_p of the expression

$$z(x) = \sum_{k=1}^{\infty} A_k g_k(x) + \sum_{p=1}^{L} a_p \varphi_p(x) \tag{4}$$

such that

$$z(X_j) = f_j, \quad j = 1, \ldots, N, \tag{5}$$

and

$$\text{the quantity } \; |z|_L^2 \; \text{ attains its minimum.} \tag{6}$$

Introduce the *generating function*

$$R_L(x,y) = \sum_{k=1}^{\infty} \frac{g_k(x) g_k^*(y)}{|g_k|_L^2}. \tag{7}$$

We state in Theorem 1 that a finite linear combination of the values of the generating function R_L at particular nodes is used for the practical interpolation instead of the infinite linear combination in (4). Further put

$$R = [R_L(X_i, X_j)], \quad i, j = 1, \ldots, N,$$

where R is an $N \times N$ square Hermitian matrix, and if $L > 0$ introduce an $N \times L$ matrix

$$\Phi = [\varphi_p(X_j)], \quad j = 1, \ldots, N, \; p = 1, \ldots, L.$$

Theorem 1. *Let $X_i \neq X_j$ for all $i \neq j$. Assume that the generating function (7) converges for all $x, y \in \Omega$. If $L > 0$ let* rank $\Phi = L$. *Then the problem of smooth interpolation (4) to (6) has the unique solution*

$$z(x) = \sum_{j=1}^{N} \lambda_j R_L(x, X_j) + \sum_{p=1}^{L} a_p \varphi_p(x), \tag{8}$$

where the coefficients λ_j, $j = 1, \ldots, N$, and a_p, $p = 1, \ldots, L$, are the unique solution of a nonsingular system of $N + L$ linear algebraic equations.

Proof. The proof is given in [8]. □

The *problem of smooth curve fitting (data smoothing)*, where the interpolation condition (1) is not applied, is treated in more detail in [8], [10].

4. A particular choice of basis function system

Recall that we have put $n = 1$. Let the function f to be approximated be 2π-periodic in $[0, 2\pi]$. We choose exponential functions of pure imaginary argument for the periodic basis system $\{g_k\}$ in W_L. The following theorem shows important properties of the system.

Theorem 2. *Let there be an integer s, $s \geq L$, such that $B_l = 0$ for all $l > s$ in W_L. The system of periodic exponential functions of pure imaginary argument*

$$g_k(x) = \exp(-\mathrm{i}kx), \quad x \in [0, 2\pi], \ k = 0, \pm 1, \pm 2, \ldots, \tag{9}$$

is complete and orthogonal in W_L.

Proof. The proof is given in [9]. □

The range of k implies a minor change in the notation introduced above. For the basis system (9), notice that the generating function

$$R_L(x, y) = \sum_{k=-\infty}^{\infty} \frac{g_k(x) g_k^*(y)}{|g_k|_L^2} = \sum_{k=-\infty}^{\infty} \frac{\exp(-\mathrm{i}k(x-y))}{|g_k|_L^2} \tag{10}$$

is the Fourier series in $L^2(0, 2\pi)$ with the coefficients $|g_k|_L^{-2}$, where

$$|g_k|_L^2 = 2\pi \sum_{l=L}^{\infty} B_l k^{2l} \tag{11}$$

according to (3).

Let now the function f to be approximated be nonperiodic on $(-\infty, \infty)$ and $f^{(l)}(\pm\infty) = 0$ for all $l \geq 0$. Let us define the generating function $R_L(x, y)$ as the Fourier transform of the function $|g_k|_L^{-2}$ of continuous variable k,

$$R_L(x, y) = \int_{-\infty}^{\infty} \frac{\exp(-\mathrm{i}k(x - y))}{|g_k|_L^2} \, \mathrm{d}k, \tag{12}$$

if the integral exists. Using the effect of transition from the Fourier series (10) with the coefficients $|g_k|_L^{-2}$ to the Fourier transform (12) of the function $|g_k|_L^{-2}$ of continuous variable k (cf., e.g., [3]), we have transformed the basis functions, enriched their spectrum, and released the requirement of periodicity of f. Moreover, if the integral (12) does not exist, in many instances we can calculate $R_L(x, y)$ as the Fourier transform $\mathcal{F}$ of the generalized function $|g_k|_L^{-2}$ of k.

Tension spline. Choosing now a particular sequence $\{B_l\}$, we finish the definition of the inner product and norm (2), (3) in a particular space W_L and set, therefore, the minimization properties of the smooth interpolant. Let us thus put (cf. [4])

$$B_l = 0 \text{ for all } l \text{ with the exception of } B_1 = \alpha^2, \ \alpha > 0, \text{ and } B_2 = 1. \tag{13}$$

It means that we have $L = 1$ and minimize the L^2 norm of the first derivative (characterizing the oscillations) multiplied by α^2 plus the L^2 norm of the second derivative (characterizing the curvature) of the interpolant (4) in the form (8), i.e.

$$z(x) = \sum_{j=1}^{N} \lambda_j R_1(x, X_j) + a_1. \tag{14}$$

We get

$$|g_k|_1^2 = 2\pi(\alpha^2 k^2 + k^4)$$

from (11). Putting $r = |x - y|$, we arrive at

$$\begin{aligned} R_1(x, y) &= \mathcal{F}\left(\frac{1}{2\pi k^2(\alpha^2 + k^2)}\right) = \frac{1}{2\pi}\mathcal{F}\left(\frac{1}{\alpha^2 k^2} - \frac{1}{\alpha^2(k^2 + \alpha^2)}\right) \\ &= -\frac{1}{2\alpha^3}(\alpha|r| + \exp(-\alpha|r|)), \end{aligned} \tag{15}$$

where $\mathcal{F}$ denotes the Fourier transform of a generalized function (see [2], p. 375, formula 14 and p. 377, formula 29; and [1], formula 8.469.3). It is easy to find out that this version of smooth approximation is, in fact, equivalent to the tension spline interpolation [6] but introduced in a way different from [4].

There are further practical examples of smooth approximation where the integral (12) that defines the generating function can be calculated with the help of the Fourier transform.

Cubic spline. Choosing another particular sequence $\{B_l\}$, i.e. $B_l = 0$ for all l with the exception of $B_2 = 1$ (cf. [9], [10]), we have $L = 2$ and minimize the usual L^2 norm of the second derivative (curvature) of the interpolant (8) only, i.e.

$$z(x) = \sum_{j=1}^{N} \lambda_j R_2(x, X_j) + a_1 + a_2 x.$$

This sequence $\{B_l\}$ differs from (13) only by the condition $B_1 = 0$ that replaced the tension spline condition $B_1 = \alpha^2 > 0$. We get $|g_k|_2^2 = 2\pi k^4$ from (11) and arrive at (see [2])

$$R_2(x, y) = \mathcal{F}(1/(2\pi k^4)) = \tfrac{1}{12} r^3.$$

Apparently, this version of smooth approximation is, in fact, the *cubic spline interpolation* [5] considered to be a classical interpolation method and known for the above mentioned minimization property.

5. Computational comparison

We present results of a simple numerical experiment with the tension spline for $n = 1$. We employ the complete and orthogonal system (9) and the sequence (13) to introduce the space W_1. We use the interpolant (14), where R_1 is given by (15). The function to be interpolated is

$$f(x) = 5 + \frac{2}{1 + 16x^2}. \tag{16}$$

Apparently, it has "almost a pole" at $x = 0$. The tension spline interpolation of the function (16) has been constructed in several equidistant grids of N nodes on $[-1, 1]$ and for several values of α^2 including also $\alpha^2 = 0$, i.e. the cubic spline.

Some of the results of interpolation are in Fig. 1. We put $N = 9$ and compare tension splines with $\alpha^2 = 0$ (i.e. the cubic spline, solid line), $\alpha^2 = 1\ 000$ (dashed line), and $\alpha^2 = 10\ 000$ (dotted line). The interpolants are in the upper part of the figure, their first derivatives in the lower part along the x axis.

We see that the tension splines do not differ substantially from each other but their derivatives are very unlike. The derivative of the cubic spline is a smooth function while the derivative of the tension spline with $\alpha^2 = 10\ 000$ is similar to a piecewise constant function with smooth changes (not jumps) between the constant levels. This corresponds to the behavior of this tension spline if examined in a different scale: it resembles a piecewise linear curve but it is smooth, not sharp-cornered also at nodes, i.e. its derivative is continuous.

A proper choice of the parameter α^2 can provide a compromise interpolation solution with both tension spline and its derivative so smooth that they give a good, pleasing subjective impression.

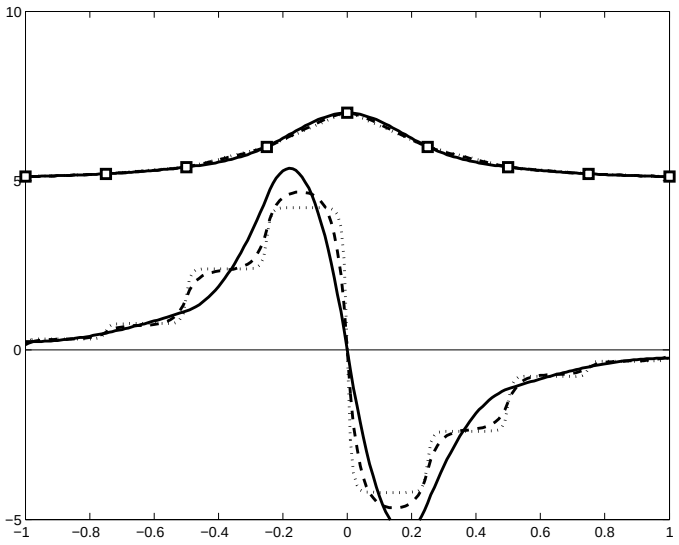


Figure 1: $N = 9$. The horizontal axis: independent variable, the vertical axis: interpolant (in the upper part of the figure) and its derivative (in the lower part). Cubic spline (tension $\alpha^2 = 0$): solid line, tension spline ($\alpha^2 = 1\ 000$): dashed line, tension spline ($\alpha^2 = 10\ 000$): dotted line.

6. Conclusion

The aim of this paper was to show that the generating function for the tension spline interpolation can be obtained by means of the Fourier transform of generalized functions. The Fourier transform can be successfully used to determine the generating function also in several other cases including $n = 2$ and $n = 3$. The example in Fig. 1 is a very simple illustration.

Acknowledgements

This work has been supported by project RVO 67985840 and by Czech Science Foundation grant P101/14-02067S.

References

[1] Gradshteyn, I. S. and Ryzhik, I. M.: *Table of integrals, series, and products.* Academic Press, Boston, 1994.

[2] Kreĭn, S. G. (Ed.): *Functional analysis* (Russian). Nauka, Moskva, 1964.

[3] Lanzos, C.: *Discourse on Fourier series.* Oliver & Boyd, Edinburgh, 1966.

[4] Mitáš, L. and Mitášová, H.: General variational approach to the interpolation problem. Comput. Math. Appl. **16** (1988), 983–992.

[5] Press, W. H., Flannery, B. P., Teukolsky, S. A., and Vetterling, W. T.: *Numerical recipes. The art of scientific computing.* Cambridge University Press, Cambridge, 1986.

[6] Schweikert, D. G.: An interpolation curve using a spline in tension. J. Math. and Phys. **45** (1966), 312–317.

[7] Segeth, K.: Smooth approximation of data with applications to interpolating and smoothing. In: *Programs and Algorithms of Numerical Mathematics* 16, pp. 181–186. Institute of Mathematics AS CR, Prague, 2013.

[8] Segeth, K.: Some computational aspects of smooth approximation. Computing **95** (2013), S695–S708.

[9] Segeth, K.: A periodic basis system of the smooth approximation space. Appl. Math. Comput. **267** (2015), 436–444.

[10] Talmi, A. and Gilat, G.: Method for smooth approximation of data. J. Comput. Phys. **23** (1977), 93–123.

[11] Ueberhuber, C. W.: *Numerical computation*, vol. 1. Springer, Berlin, 1997.

Conference Applications of Mathematics 2015, in honor of the birthday anniversaries of Ivo Babuška (90), Milan Práger (85), and Emil Vitásek (85)
J. Brandts, S. Korotov, M. Křížek, K. Segeth, J. Šístek, T. Vejchodský (Eds.)
Institute of Mathematics CAS, Prague 2015

GEOMETRIC DIAGRAM FOR REPRESENTING SHAPE QUALITY IN MESH REFINEMENT

José P. Suárez[1], Ángel Plaza[1,2] and Tania Moreno[3]

[1] MAGiC. Division of Mathematics, Graphics and Computation, IUMA, Information and Communication Systems, University of Las Palmas de Gran Canaria
Canary Islands, Spain
josepablo.suarez@ulpgc.es

[2] Department of Mathematics, University of Las Palmas de Gran Canaria
Canary Islands, Spain
angel.plaza@ulpgc.es

[3] Faculty of Mathematics and Informatics, University of Holguín, Avenida XX Aniversario, vía Guardalavaca, Holguín, Cuba
tmorenog@facinf.uho.edu.cu

Abstract: We review and discuss a method to normalize triangles by the longest-edge. A geometric diagram is described as a helpful tool for studying and interpreting the quality of triangle shapes during iterative mesh refinements. Modern CAE systems as those implementing the finite element method (FEM) require such tools for guiding the user about the quality of generated triangulations. In this paper we show that a similar method and corresponding geometric diagram in the three-dimensional case do not exist.

Keywords: mesh generation, longest-edge refinement, geometric diagram, mesh regularity

MSC: 65L50, 65M50

1. Introduction

Mesh generation and, in particular, the construction of 'quality' meshes is a major issue in many fields where computer modeling and engineering analysis are extensively used [4, 12]. Some mesh smoothing and mesh optimization strategies are described in [2, 3, 5, 6], and also different mesh quality metrics have been proposed in recent years [9].

For example in the finite element method (FEM), equilateral triangles are favored over obtuse or skinny triangles. Many of these methods employ forms of local and global triangle subdivision and seek to maintain well-shaped triangles. Here we consider several popular triangle subdivision schemes.

When refining a mesh, the aim is to use a given triangle subdivision scheme that may preserve the minimum angle condition in the sense that such smallest angle keeps as high as possible, [1, 15]. In the last years, many triangle partitions has been studied, including schemes using more than one point to configure the triangle halving [17]. The approach for partitioning single triangles is further used to develop algorithms that for a given input mesh and some refinement criteria through the elements, a new mesh is obtained with more element detail, see e.g. [8, 11].

Lastly the idea of using a two-dimensional diagram for representing triangle shape in mesh generation points to the works of [8, 13, 16]. While in [13] a more comprehensive development, theory and application of such diagram can be found, where the concept of normalization of triangles by the longest edge is the key for the computational construction of the diagram. It is worth to mention that this is not a novel idea. Already in 1939, Tuckey [18] observed the potential of using same normalization idea to construct a diagram in 2D. The main reason to introduce this diagram was the solution of triangles. In 1943 Tuckey presented a variation of the diagram for the same purpose, obviously a handle-method very far away of any computational automation. The idea behind the diagram is scaling any triangle so that the length of its longest edge is equal to one and the other sides x and y are in descending order; then the point with coordinates x and y will represent the triangle.

In this work we review some previous results on construction of a geometric diagram for assessing mesh quality, see [10, 13]. While diagram suits perfectly for the two-dimensional case its counterpart in the three-dimensional case does not. To demonstrate this, we provide a result showing that the normalization process of tetrahedra by the longest edges of tetrahedra is not possible, and thus the extension of a similar diagram to 3D case is not feasible either.

1.1. Mesh refinement schemes

By the longest-edge (LE) partition of a triangle t_0 we mean a subdivision scheme where the midpoint of the longest edge of a triangle t_0 is joint with the opposite vertex, see Figure 1 (a).

Another subdivision strategy of our interest is the 4TLE (Four Triangles Longest Edge) strategy [14], see Figure1(b). In this scheme, subdivision produces some subtriangles that are similar to certain previous triangles in the refinement tree generated. However, the other subtriangles are not in such similarity classes yet and we refer to them as new dissimilar triangles.

The other well-studied partition is the 7TLE (Seven Triangles Longest Edge) [10] where we position two equally spaced points per edge and, then the interior of the triangle is divided into seven sub-triangles in a manner compatible with the subdivision of the edges. Three of the new sub-triangles are similar to the original, two are similar to the new triangle also generated by the 4T-LE, and the other two triangles are, in general, better shaped [10].

It should be noted that, due to parallelism, the first three sub-triangles obtained are similar to the initial one t_0, whereas the second two sub-triangles are similar to

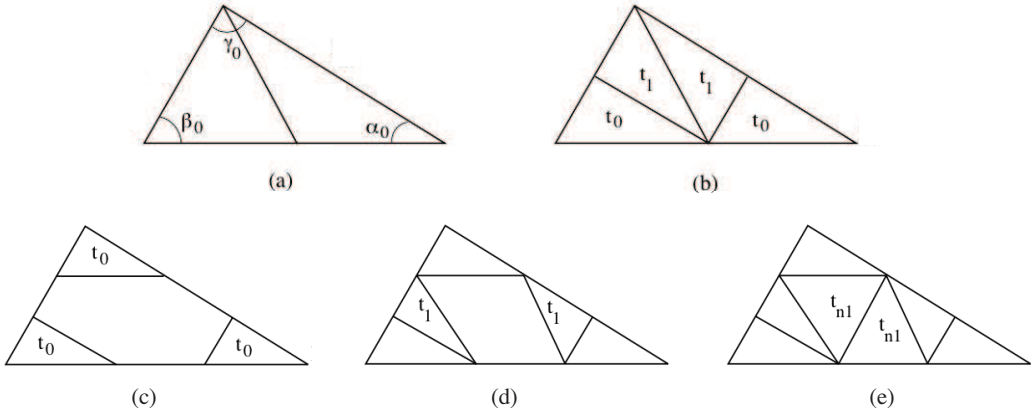


Figure 1: (a) LE (longest-edge) partition, (b) 4TLE (Four Triangles Longest Edge), (c)-(e) 7TLE (Seven Triangles Longest Edge).

the first-class Rivara triangle t_1. Finally, the last two triangles are not given with the 4T-LE partition and, consequently, will be called here, t_{n1}. Note also that the area of sub-triangles t_0 and t_1 is 1/9 of the area of the initial triangle, whereas the area of each sub-triangle t_{n1} is 2/9 of the area of the initial triangle.

2. Geometric diagram

Now we present the main ideas for the construction of the geometric diagram in the context of 2D mesh refinement, for details, see [10, 13].

We first consider the normalized triangles which share the longest edge defined by the points $(0, 0)$ and $(1, 0)$. The apex of the triangle (x, y) is a point chosen uniformly in the region defined by the unit circle centered at $(1, 0)$, and by the vertical line $x = \frac{1}{2}$ (see Figure 2).

The geometric diagram is constructed as follows: (1) For a given triangle (or sub-triangle) the longest edge is scaled to have the unit length. This forms the base of the diagram. (2) It follows that the set of all triangles is bounded by this horizontal segment (longest edge) and by two bounding exterior circular arcs of unit radius, as shown in Figure 3.

In the diagram of Figure 3 (left) we demarcate shaded regions to classify triangles based on ranges of the *largest angle* γ within circular arcs as shown; e.g. the lowermost subregion corresponds to obtuse triangles with large angles near π and the uppermost subregion (exterior to the semicircle of radius $\frac{1}{2}$ centered on the unit base) corresponds to acute-angled triangles. The equilateral triangle corresponds to the apex with coordinates $\left(\frac{1}{2}, \frac{\sqrt{3}}{2}\right)$. As the vertex of a triangle moves from this point along either boundary arc, the maximum angle increases from $\frac{\pi}{3}$ to approach a right angle at the degenerate 'needle triangle' limit near $(0, 0)$ or $(1, 0)$. Similarly, in the

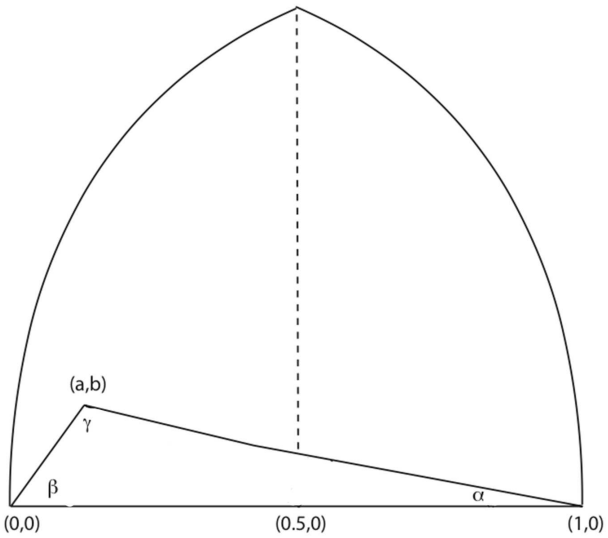


Figure 2: A random triangle normalized by its longest edge inside the diagram region.

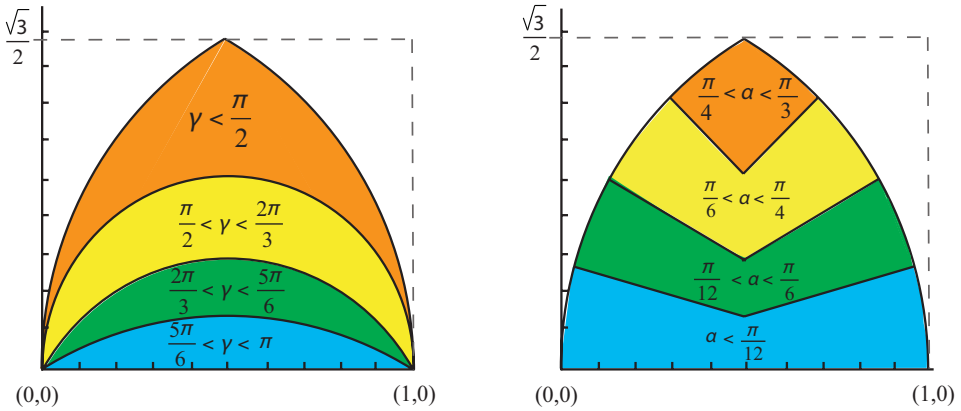


Figure 3: Diagram for triangles showing different regions corresponding to variations values of the largest angle γ and the smallest angle α.

diagram of Figure 3 (right) we demarcate by segments of straight lines emanating from $(0,0)$ and $(1,0)$, shaded subregions that bound the *smallest angle* α of a triangle. Color shading makes the respectively subregions easier to identify. The topmost subregion between the exterior circular arcs and the lines for smallest angle $\alpha = \frac{\pi}{4}$ corresponds to triangles with $\frac{\pi}{4} < \alpha < \frac{\pi}{3}$. The v-shaped subregion below this is for the case $\frac{\pi}{6} < \alpha < \frac{\pi}{4}$ and so on, with the lowest shaded region for $\alpha < \frac{\pi}{12}$. From the shaded regions in these two diagrams, it is clear that slender triangles with large

obtuse angles and small acute angles will be located close to the center part of the base and triangles close to equilateral shape will be near the apex of the diagram. It follows that one can use this diagram to investigate the evolution of triangle shapes under subdivision as we show further.

It is obvious that the right triangle (i.e. $\gamma = \pi/2$) is a separator of the familiar acute and obtuse triangle classes. The locus of points corresponding to this separator is easily identified from elementary geometry as the semicircle with unit diagonal base in our mapping diagram. Points above this semicircle $\left|z - \frac{1}{2}\right| = \frac{1}{2}$, where $z = (x, y)$ is the apex of a triangle, correspond to acute triangles and points below correspond to obtuse triangles. We will not show this semicircle in the color class diagrams following, but this property should be kept in mind for other reasons. Further we focus on the number of dissimilar triangles that are generated by various triangle subdivision schemes.

2.1. Generating the diagram by the Monte-Carlo experiment

Once we have depicted the basics involved in the diagram, we focus on computing those regions that are specifically related to the shapes appearing in a given triangle partition, for example 4TLE, 7TLE, etc.

Let us begin by describing a Monte-Carlo computational experiment that can be used to visually distinguish the classes of triangles by the number of dissimilar triangles generated by the 4TLE partition. We proceed as follows: (**1**) Select a point within the mapping domain comprised by the horizontal segment and by the two bounding exterior circular arcs. This point (x, y) defines the apex of a target triangle. (**2**) For this selected triangle, 4TLE refinement is successively applied as long as a new dissimilar triangle appears. This means that we recursively apply 4TLE and stop when the shapes of new generated triangles are the same as those already generated in previous refinement steps. (**3**) The number of such refinements to reach termination defines the number of dissimilar triangles associated with the initial triangle and this numerical value is assigned to the initial point (x, y) chosen. (**4**) This process is progressively applied to a large sample of triangles (points) uniformly distributed over the mapping domain. (**5**) Finally, we graph the respective values of dissimilar triangles in a corresponding color map, see Figure 4 (left).

2.2. Structure of regions for the 4TLE refinement

For clarity, the number of dissimilar triangles has been added inside several colored regions in Figure 4. Thus, the numerical value 2 corresponds to two dissimilar triangles is associated with the region above the pair of arcs that intersect on the vertical line of symmetry near the point $(0.5, 0.3)$. The region below this corresponds to 3 dissimilar triangles, and so on as the base is approached. Viewed another way, obtuse needle-like triangles near the base will require many refinements before new dissimilar triangles no longer appear. Later, we will explore this point and plot associated trajectories corresponding to migration of new triangles.

From Figure 4 (right), we deduce that the separator for classes 2 and 3 in case

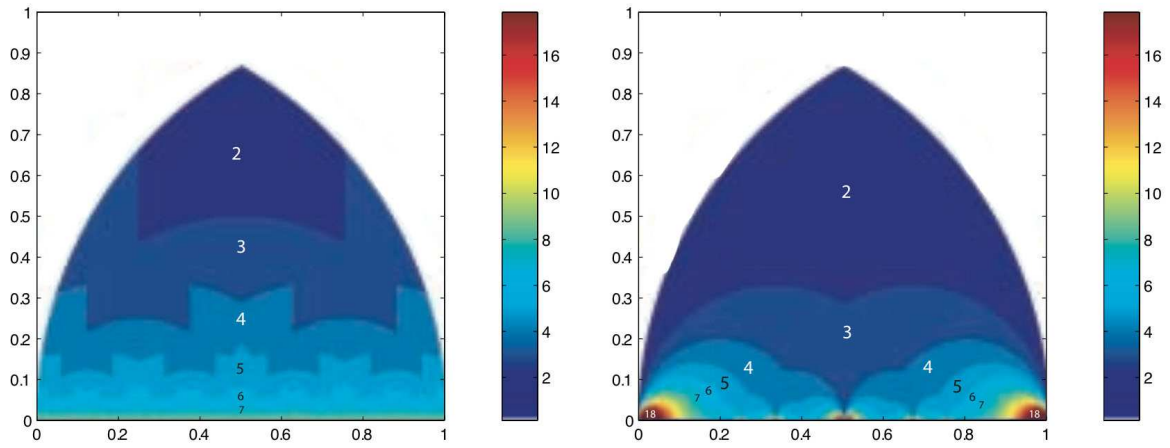


Figure 4: Subregions for dissimilar triangle classes generated by the Monte-Carlo simulation for the 4TLE (left) and the 7TLE (right) partitions, respectively.

of 7TLE is given by the segments of two circles of radius $\frac{1}{3}$ centered respectively at $x = \frac{1}{3}$ and $x = \frac{2}{3}$. That is, the curves $\left|z - \frac{1}{3}\right| = \frac{1}{3}$ and $\left|z - \frac{2}{3}\right| = \frac{1}{3}$. The curves for the subsequent separator between 3 and 4 have slightly more complicated shapes. Moreover, this shape is again evident on two smaller scales at the level of the next separator between 4 and 5. The pattern appears to continue to repeat in a fractal-like manner as the higher value separators are identified. A more formal mathematical approach, based on mapping in the complex plane, follows and utilizes the concept of antecedent triangle for the 4TLE partition.

One may use two functions, named f_L and f_R showing a 'trace back' from the equilateral triangle t_0, with apex $\left(\frac{1}{2}, \frac{\sqrt{3}}{2}\right)$, see Figure 5 (a). From equilateral triangle t_0 situated on the intersection of the exterior boundary curves, the only antecedent that generates it after subdivision is triangle t_1. Note that t_1 is an obtuse triangle located exactly where the pair of boundary curves intersect on the vertical line of symmetry at the point $y = \frac{\sqrt{3}}{6}$. Continuing the traceback, this obtuse triangle t_1 is the result of the subdivision of two antecedent triangles as marked, with the right antecedent denoted t_2 and the left one being symmetrically located in the left part of the diagram as expected. Again, each of these t_2 triangles is located at the intersection of two boundary curves. The next pair of antecedents of the right half from left to right are t_3' and t_3 respectively. As before, t_3' and t_3 are located at the intersections of respective pairs of boundary curves that demarcate a change in similarity class and the process continues downward in the diagram with the antecedents approaching the degenerate case of planar obtuse triangles on the horizontal line. Another traceback example is given in Figure 5 (b), starting with apex $(0.67, 0.43)$ (and obviously has a similar path reflected in $x = \frac{1}{2}$).

The class separators determined experimentally by the Monte-Carlo experiment may also be generated mathematically as a recursive composition of left and right

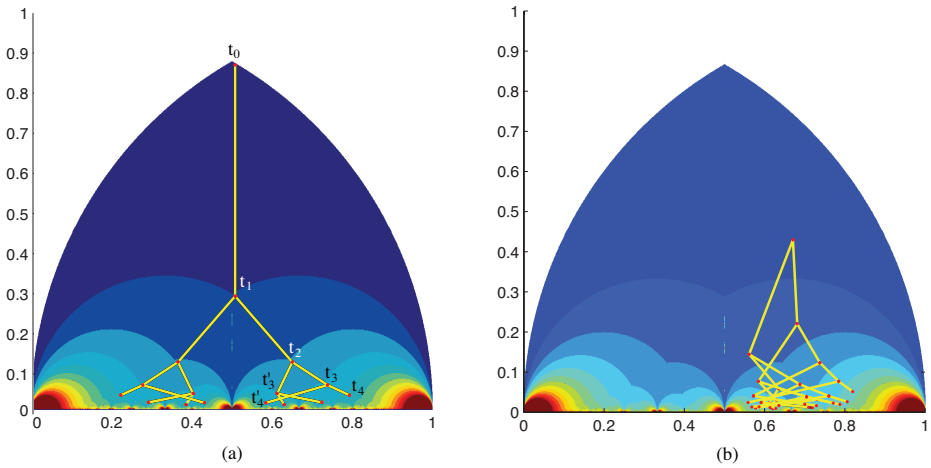


Figure 5: Traceback curves showing antecedents in successive regions: (a) traceback from equilateral triangle apex with antecedents on border curves and (b) traceback from acute triangles interior to region.

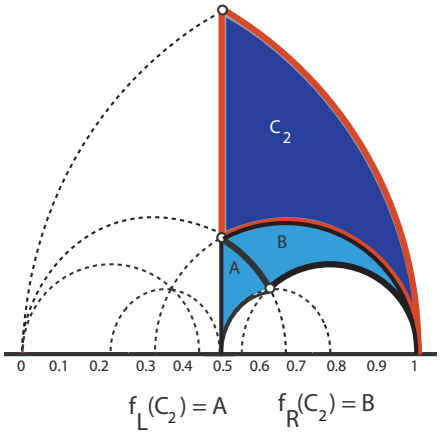


Figure 6: Boundary curves for the separator of order 2 and 3 in the 4TLE partition.

mapsf_L and f_R, for details see [13]. Then it follows that a set of boundary curves for those separators is easily obtained, see Figure 6.

2.3. Structure of regions for the 7TLE refinement

We recursively apply 7T-LE and stop when the shapes of new generated triangles are the same as those already generated in previous refinement steps. The number of such refinements to reach termination defines the number of dissimilar triangles associated with the initial triangles associated with the initial triangle and this numerical value is as- signed to the initial point (x, y) chosen. This process is

progressively applied to a large sample of triangles (points) uniformly distributed over the domain. Finally, we graph the respective values of dissimilar triangles in a corresponding color map to obtain the result in Figure 4 (right).

Note that for the 7TLE partition, a lower bound for the maximum of the smallest angles for triangles $t_n 1$ is $\alpha = 30°$ corresponding to the apex with $x = 1$, or the apex with $x = 3$.

Unfortunately, deriving classes separator as done respectively for the 4TLE is still an open problem. The main reason is that the 7TLE refinement produces seven new triangles that complicates the calculation of functions f_L and f_R.

Note that the diagram is useful to tackle the so-called self-improvement property of partitions for 4TLE, 7TLE and others triangle schemes. It also is feasible to use the diagram for assessing which algorithm is more convenient for mesh refinement. For example, using combinations of partitions, i.e. using one type for some targeted triangle cases and the other type for the others may yield improved algorithms. For example, in [10] it has been proposed a combined scheme for improvement of the mesh, by combining longest-edge based with other self-similar partitions, depending on the number of points inserted per edge.

3. Toward a geometric diagram in the three-dimensional space

Concerning extension of the previous geometric diagram to the three-dimensional case, we give some helpful ideas resulting to a negative conclusion: there does not exist a similar geometric diagram for representing mesh quality during mesh refinements in 3D.

We say that two triangles (two tetrahedra in three dimensions) are in the same similarity class if there exists a similarity transformation that transforms one of these triangles (tetrahedra in three dimensions) into the other.

Theorem 1. *Let us have a given segment $\overline{AB}$ of length 1 in the Euclidean plane. Let $\mathcal{T}$ be the class of all triangles T such that $\overline{AB}$ is the longest edge of T. Then there exists a subclass $\mathcal{D}$ of $\mathcal{T}$ satisfying the following conditions.*

1. *For each similarity class of triangles $\mathcal{S}$, there are one and only one element of $\mathcal{S}$ in $\mathcal{D}$.*

2. $$diam(\bigcup_{T\in\mathcal{D}} T) = 1$$

As a consequence of Theorem 1 we have that for the two-dimensional case, there exists a diagram for normalizing triangles with respect to the longest edge, as showed in Section 2. The utility of this diagram lead us to think about the possibility of make an analogous diagram for the representation of tetrahedra in the three-dimensional Euclidean space. However, we will see some difficulties that appears in the attempt to construct a natural generalization of this diagram in the three-dimensional case.

Theorem 2. *Let us have a given segment* $\overline{AB}$ *of length* 1 *in the three-dimensional Eucidean space. Let* $\mathcal{D}$ *be the geometric place of all tetrahedra* $\mathcal{T}$ *such that* $\overline{AB}$ *is one longest edge of* $\mathcal{T}$*. Then it follows that the diameter (maximum of longest edge) in the set* $\mathcal{D}$ *is greater than* 1*.*

Skecth of the Proof: Note that $\mathcal{D}$ include a regular tetrahedron ABP_1P_2 and a tetrahedron ABQ_1Q_2 such that $|\overline{Q_1A}| = |\overline{Q_1B}| = |\overline{Q_2A}| = |\overline{Q_2B}| = \frac{\sqrt{2}}{2}$ and the dihedral angle between faces ABQ_1 y ABQ_2 equals 150°. We will see that there exist i_0, j_0, $i_0 \in \{1, 2\}$, $j_0 \in \{1, 2\}$ such that $|\overline{P_{i_0}Q_{j_0}}| > 1$.

Let M be the midpoint of $\overline{AB}$. Note that $\angle P_1MP_2 = \arccos(\frac{1}{3}) = 70.5288°$, $\angle Q_1MQ_2 = 150°$ and the points M, P_1, P_2, Q_1, Q_2 are coplanar. Then there exist i_0, j_0, $i_0 \in \{1, 2\}$, $j_0 \in \{1, 2\}$ such that $\triangle P_{i_0}MQ_{j_0}$ is an obtuse triangle in M.

Moreover, $|\overline{P_iM}| = \frac{\sqrt{3}}{2}$, $i \in \{1, 2\}$ and $|\overline{Q_jM}| = \frac{1}{2}$, $j \in \{1, 2\}$ from where in the obtuse triangle $\triangle P_{i_0}MQ_{j_0}$ we have

$$|\overline{P_{i_0}Q_{j_0}}| > \sqrt{|\overline{P_{i_0}M}|^2 + |\overline{Q_{j_0}M}|^2} = \frac{3}{4} + \frac{1}{4} = 1.$$

We then conclude that

$$\operatorname{diam}(\bigcup_{T \in \mathcal{D}} T) > 1$$

□

4. Conclusions

In this work we have recovered the idea of the geometric diagram for assessing quality in triangle mesh refinement. In FEM, error indicators give the trends of the behavior of the solution through an iterative process. Classically, the angle condition has been set as the standard form for the good quality of a mesh, where acute, right, and in general, those triangle shapes very close the regular triangles behave better, [15, 1]. In this work we provide a visual tool, called the geometric diagram, for inspecting shape evolution in the refinement of meshes. Fortunately, the tool is clearly of utility as has been shown in the study of two triangle partition schemes, 4TLE and 7TLE. In addition, we provide a new result consisting in confirming that an extension to a similar geometric diagram for the three-dimensional case is not possible, partially due to the impossibility to normalize by the longest edge of tetrahedra. This last result however, open to a new scene where an idea of different geometric diagram by normalizing e.g. edges or faces of tetrahedra may be feasible.

References

[1] Babuška, I. and Aziz, A. K.: On the angle condition in the finite element method. SIAM J. Numer. Anal. **13** (1976), 214–226.

[2] Branets, L. and Carey, G. F.: Smoothing and adaptive redistribution for grids with irregular valence and hanging nodes. In: *Proceedings of the 13th International Meshing Roundtable*, pp. 333–344, 2004.

[3] Branets, L. and Carey, G. F.: A local cell quality metric and variational grid smoothing algorithm. Engrg. Comput. **21** (2005), 19–28.

[4] Carey, G. F.: *Computational grids: generation, refinement and solution strategies*. Taylor and Francis, 1997.

[5] Freitag, L. A. and Plassman, P.: Local optimization-based simplicial mesh untangling and improvement. Int. J. Num. Meth. Engrg. **49** (2000), 109–125.

[6] Garimella, R. V., Shashkov, M. J., and Knupp, P. M.: Triangular and quadrilateral surface mesh quality optimization using local parametrization. Comp. Meth. Appl. Mech. and Engrg. **193** (2004), 913–928.

[7] Hannukainen, A., Korotov, S., and Křížek, M.: On numerical regularity of the face-to-face longest-edge bisection algorithm for tetrahedral partitions. Sci. Comput. Program. **90** (2014), 34–41.

[8] Hannukainen, A., Korotov, S., and Křížek, M.: On global and local mesh refinements by a generalized conforming bisection algorithm. J. Comput. Appl. Math. **235** (2010), 419–436.

[9] Knupp, P. M.: Algebraic mesh quality metrics. SIAM J. Sci. Comput. **23** (2001), 193–218.

[10] Márquez, A., Moreno-González, A., Plaza, Á., and Suárez, J. P.: The seven-triangle longest-side partition of triangles and mesh quality improvement. Finite Elem. Anal. Des. **44** (2008), 748–758.

[11] Padrón, M. A., Suárez, J. P., and Plaza, A.: Refinement based on longest-edge and self-similar four-triangle partitions. Math. Comput. Simulation **75** (2007), 251–262.

[12] Knupp, P. M. and Steinberg S.: *The fundamentals of grid generation*. CRC Press, Boca Raton, FL, 1994.

[13] Plaza, Á., Suárez, J. P., and Carey, G. F.: A geometric diagram and hybrid scheme for triangle subdivision. Comput. Aided Geom. Des., **24** (2007), 19–27.

[14] Plaza, Á., Suárez, J. P., Padrón, M. A., Falcón, S., and Amieiro, D.: Mesh quality improvement and other properties in the four-triangles longest-edge partition. Comput. Aided Geom. Design **21** (2004), 353–369.

[15] Korotov, S., Křížek, M., and Kropáč, A.: Strong regularity of a family of face-to-face partitions generated by the longest-edge bisection algorithm. Comput. Math. Math. Phys. **48** (2008), 1687–1698.

[16] Rivara, M. C.: LEPP-bisection algorithms, applications and mathematical properties. Appl. Num. Math. **59** (2009), 2218–2235.

[17] Suárez, J. P., Moreno, T., Abad, P., and Plaza, Á.: Properties of the longest-edge n-section refinement scheme for triangular meshes. Appl. Math. Letters **25** (2012), 2037–2039.

[18] Tuckey, C. O.: A diagram for the study and solution of triangles. The Mathematical Gazette, **23** (1939), 150–154.

[19] Tuckey. C. O.: A diagram for the solution of triangles. The Mathematical Gazette, **27** (1943), 1–3.

Conference Applications of Mathematics 2015, in honor of the birthday anniversaries of Ivo Babuška (90), Milan Práger (85), and Emil Vitásek (85)
J. Brandts, S. Korotov, M. Křížek, K. Segeth, J. Šístek, T. Vejchodský (Eds.)
Institute of Mathematics CAS, Prague 2015

SOME REMARKS ON MIXED APPROXIMATION PROBLEM

Irena Sýkorová

University of Economics
Ekonomická 957, 148 00 Prague 4, Czech Republic
sykorova@vse.cz

Dedicated to my father Milan Práger on his 85th birthday

Abstract: Several years ago, we discussed the problem of approximation polynomials with Milan Práger. This paper is a natural continuation of the work we collaborated on. An important part of numerical analysis is the problem of finding an approximation of a given function. This problem can be solved in many ways. The aim of this paper is to show how interpolation can be combined with the Chebyshev approximation.

Keywords: interpolation, approximation, Chebyshev approximation, Remez algorithm

MSC: 65D05, 41A05, 41A50

1. Introduction

Numerical analysis often requires approximating a given real-valued function f, continuous on a closed interval $[a, b]$, by another function g that is more suitable for computing and that only slightly differs from the given function. The function g is in most cases a polynomial. In [2] and [5], there are described three basic ways of approximation of the given function f.

1. *Interpolation approximation* is such a replacement of the given function f by a new function g which satisfies the following condition:

$$f(x_j) = g(x_j) \tag{1}$$

at the given points $a \leq x_0 < x_1 < \cdots < x_n \leq b$. Sometimes we also additionally require the coincidence of the derivatives $f^{(i)}(x_j) = g^{(i)}(x_j)$ for $i = 1, 2, \ldots, r$.

2. *The Chebyshev approximation* consists in the minimization of the maximum norm. The desired polynomial g satisfies

$$E_n(f) = ||f - g|| = \max_{x\in[a,b]} |f(x) - g(x)| \leq \max_{x\in[a,b]} |f(x) - h(x)|, \tag{2}$$

where h is an arbitrary polynomial of degree at most n.

3. *The least squares method* finds a function g which fits with the given function f in such a way that the sum of squares of the differences $f(x_j)-g(x_j)$, sometimes multiplied by a suitable weight function w,

$$\sum_{j=0}^{n} w(x_j)[f(x_j) - g(x_j)]^2$$

is minimal. In the usual case, $g(x) = \sum_{k=0}^{m} c_k g_k(x)$ and the sum to be minimized is

$$\sum_{j=0}^{n} w(x_j) \left[f(x_j) - \sum_{k=0}^{m} c_k g_k(x_j) \right]^2 .$$

2. Mixed approximation polynomial

Now we try to construct a mixed polynomial which is a combination of interpolation (1) and the Chebyshev approximation (2). The *mixed approximation polynomial* is a polynomial s of degree at most n which approximates the function f in the Chebyshev sense, i.e.

$$\|f - s\| = \max_{x\in[a,b]} |f(x) - s(x)| \tag{3}$$

is minimal and at the endpoints of the interval it fulfils, in addition, the interpolation conditions

$$s(a) = f(a) \qquad \text{and} \qquad s(b) = f(b). \tag{4}$$

Such a polynomial s has similar properties as the Chebyshev approximation.

If $f \in C[a, b]$ and s is a polynomial of degree at most n such that $s(a) = f(a)$ and $s(b) = f(b)$ then the following holds.

a) Suppose there exist a constant c and n points $x_1 < x_2 < \cdots < x_n$ in the interval (a, b) such that

$$\operatorname{sign}[(-1)^i(f(x_i) - s(x_i))] = c \qquad \text{for} \quad i = 1, \ldots, n.$$

Then

$$E_n(f) \geq \min_{i=1,\ldots,n} |f(x_i) - s(x_i)|.$$

b) The polynomial s is the best approximation of the function f in the sense of Chebyshev if and only if there exist at least n points $x_1 < x_2 < \cdots < x_n$ in the interval (a, b) with the property

$$f(x_i) - s(x_i) = \alpha(-1)^i\|f - s\| \qquad \text{for} \quad i = 1, \ldots, n,$$

where $\alpha = 1$ or $\alpha = -1$ for all i simultaneously. The set of the points $\{x_i\}_{i=1}^n$ is called the Chebyshev alternant.

c) The polynomial s is unique.

The proof is given in [3], but it is only a slight modification of corresponding proofs in the standard case.

3. Construction of mixed approximation polynomial

Now we show a construction of the polynomial s which is the best approximation of a given function $f \in C[a,b]$ in the maximum norm and furthermore satisfies $s(a) = f(a)$, $s(b) = f(b)$. We use the Remez algorithm which sequentially improves the polynomial using the alternant property, see [1], [5].

The initial approximation of the alternant $x_1^{(0)}, \ldots, x_n^{(0)}$ can be arbitrary, but the points must be mutually different. From the kth approximation of the alternant $x_1^{(k)}, \ldots, x_n^{(k)}$ we will construct next approximation $x_1^{(k+1)}, \ldots, x_n^{(k+1)}$. Having the kth approximation, we can construct a polynomial

$$s^{(k)}(x) = \sum_{j=0}^{n} c_j^{(k)} x^j$$

of the degree at most n such that it holds

$$f(x_i^{(k)}) - s^{(k)}(x_i^{(k)}) = (-1)^i E^{(k)} \qquad \text{for} \quad i = 1, \ldots, n,$$

where $E^{(k)}$ is some constant which we have to determine. The conditions $s(a) = f(a)$ and $s(b) = f(b)$ are fulfilled at the endpoints of the interval $[a,b]$. We have a system of $(n+2)$ linear equations for the coefficients $c_0^{(k)}$, $\ldots$, $c_n^{(k)}$ and the constant $E^{(k)}$.

$$\begin{aligned} \sum_{j=0}^{n} c_j^{(k)} \left(x_i^{(k)}\right)^j + (-1)^i E^{(k)} &= f\left(x_i^{(k)}\right), \\ \sum_{j=0}^{n} c_j^k a &= f(a), \qquad (5) \\ \sum_{j=0}^{n} c_j^k b &= f(b). \end{aligned}$$

The determinant of this system (5) is nonzero, so there exists a unique solution, see [4].

Now we denote

$$R^{(k)}(x) = f(x) - s^{(k)}(x) \qquad (6)$$

and choose arbitrarily the number q such that $q \in (0,1)$.

Let the points $x_1^{(k)}, \ldots, x_n^{(k)}$ be given. Then we are looking for a new set of points $x_1^{(k+1)}, \ldots, x_n^{(k+1)}$ so that $x_i^{(k+1)} \in [x_{i-1}^{(k)}, x_{i+1}^{(k)}]$ for $i = 1, \ldots, n$, and the following conditions are fulfilled:

$$\max_{1 \le i \le n} \left| R^{(k)}\left(x_i^{(k+1)}\right) \right| \geq |E^{(k)}| + q(\|R^{(k)}\| - |E^{(k)}|), \qquad (7)$$

$$R^{(k)}\left(x_i^{(k+1)}\right) R^{(k)}\left(x_{i+1}^{(k+1)}\right) \leq 0, \qquad i = 1, \ldots, n-1, \qquad (8)$$

$$\left| R^{(k)}\left(x_i^{(k+1)}\right) \right| \geq |E^k|, \qquad i = 1, \ldots, n. \qquad (9)$$

This choice is not unique. We show that we can construct the $(k+1)$st approximation of the alternant such that the required properties are satisfied. We choose a point $y^{(k+1)}$ such that it holds

$$\left|R^{(k)}\left(y^{(k+1)}\right)\right| \geq |E^{(k)}| + q(\|R^{(k)}\| - |E^{(k)}|).$$

Since the right-hand side is at most equal to $\|R^{(k)}\|$, then the choice which satisfies the condition (7) is possible.

When we have the kth approximation of the alternant, we can define the $(k+1)$st approximation in the following way. One point is the point $y^{(k+1)}$ and the other points are suitable points from the previous kth approximation.

If $E^k = 0$ the $(k+1)$st approximation will be a set containing the point $y^{(k+1)}$ and arbitrary $n-1$ points of the kth approximation, i.e. an arbitrary point of the kth approximation can be replaced by the point $y^{(k+1)}$. By ordering of the points of the $(k+1)$st approximation, these points will be put in the corresponding intervals. The conditions (8) and (9) are also fulfilled, because $E^k = 0$ and then $R^{(k)}\left(x_i^{(k)}\right) = 0$ for every i, too.

If $E^k \neq 0$, there exists $R^{(k)}\left(x_i^{(k)}\right) \neq 0$ for every i. Now we describe three cases which can occur:

1. $y^{(k+1)} \in [a, x_1^{(k)}]$,
2. $y^{(k+1)} \in [x_n^{(k)}, b]$,
3. $y^{(k+1)} \in [x_1^{(k)}, x_n^{(k)}]$.

In case 1, we put $x_1^{(k+1)} = y^{(k+1)}$ and for $i = 2, \cdots, n$ we define

$$\begin{aligned} x_i^{(k+1)} &= x_i^{(k)} \quad \text{if} \quad R^{(k)}\left(x_1^{(k)}\right) R^{(k)}\left(x_1^{(k+1)}\right) > 0 \quad \text{or} \\ x_i^{(k+1)} &= x_{i-1}^{(k)} \quad \text{if} \quad R^{(k)}\left(x_1^{(k)}\right) R^{(k)}\left(x_1^{(k+1)}\right) < 0. \end{aligned}$$

We drop the point $x_1^{(k)}$ or $x_n^{(k)}$ from the previous approximation.

In case 2, we put $x_n^{(k+1)} = y^{(k+1)}$ and for $i = 1, \cdots, n-1$ we define

$$\begin{aligned} x_i^{(k+1)} &= x_i^{(k)} \quad \text{if} \quad R^{(k)}\left(x_n^{(k)}\right) R^{(k)}\left(x_n^{(k+1)}\right) > 0 \quad \text{or} \\ x_i^{(k+1)} &= x_{i+1}^{(k)} \quad \text{if} \quad R^{(k)}\left(x_n^{(k)}\right) R^{(k)}\left(x_n^{(k+1)}\right) < 0. \end{aligned}$$

We drop the point $x_n^{(k)}$ or $x_1^{(k)}$ from the previous approximation.

In case 3, we denote by i_0 the subscript such that $y^{(k+1)} \in [x_{i_0}^{(k)}, x_{i_0+1}^{(k)}]$. Then we put $x_{i_0}^{(k+1)} = y^{(k+1)}$ and $x_i^{(k+1)} = x_i^{(k)}$ for $i \neq i_0$ if $R^{(k)}\left(y^{(k+1)}\right) R^{(k)}\left(x_{i_0}^{(k)}\right) > 0$. We drop the point $x_{i_0}^{(k)}$ from the previous approximation. Or we put $x_{i_0+1}^{(k+1)} = y^{(k+1)}$ and $x_i^{(k+1)} = x_i^{(k)}$ for $i \neq i_0 + 1$ if $R^{(k)}\left(y^{(k+1)}\right) R^{(k)}\left(x_{i_0+1}^{(k)}\right) > 0$. We drop the point $x_{i_0+1}^{(k)}$ from the previous approximation.

In this choice, the points of the $(k+1)$st approximation are in the corresponding intervals and all conditions are fulfilled. A convergence of this process for a sufficiently smooth function is proved in [4].

4. Examples

We complete the previous theory with several simple numerical experiments. We present results for the function $f_1(x) = \mathrm{e}^x$ on the interval $[0, 1]$. The speed of the convergence in two special cases are summarized in the following Tables 1 and 2. The iterations of the Remez algorithm at the points of the alternant in each table are given. The point which is changed in the corresponding iteration is printed in boldface digits. The points of the uniform partition of the interval are chosen for the first iteration. The accuracy of computing the points of the alternant is 0.01. We solved our task for $n = 3$, $n = 4$, $n = 5$, $n = 6$, but we present only the tables for $n = 3$ and $n = 6$.

Under each table, the mixed approximation polynomial corresponding to the last row is written. Its coefficients are shown to three decimal places.

Table 3 shows the dependence of the approximation error on the degree of the polynomial used. Approximation error is indicated by means of the absolute value of the extremes of the difference $f(x) - s(x)$ between the given function and the

No. of it.	x_1	x_2	x_3
1	0.25	0.50	0.75
2	0.25	0.50	**0.89**
3	**0.15**	0.50	0.89
4	**0.13**	0.50	0.89
5	0.13	**0.52**	0.89

$$s(x) = 0.280x^3 + 0.424x^2 + 1.014x + 1$$

Table 1: Convergence of the alternant, $n = 3$.

No. of it.	x_1	x_2	x_3	x_4	x_5	x_6
1	0.14	0.29	0.43	0.57	0.71	0.86
2	0.14	0.29	0.43	0.57	0.71	**0.96**
3	**0.06**	0.29	0.43	0.57	0.71	0.96
4	0.06	0.29	0.43	0.57	**0.83**	0.96
5	0.06	**0.20**	0.43	0.57	0.83	0.96
6	0.06	0.20	0.43	**0.63**	0.83	0.96
7	**0.05**	0.20	0.43	0.63	0.83	0.96
8	0.05	0.20	**0.40**	0.63	0.83	0.96

$$s(x) = 0.002x^6 + 0.007x^5 + 0.043x^4 + 0.166x^3 + 0.500x^2 + x + 1$$

Table 2: Convergence of the alternant, $n = 6$.

Pol. deg.	max	min	difference
3	0.7489 E−3	0.7471 E−3	0.0018 E−3
4	0.3519 E−4	0.3472 E−4	0.0047 E−4
5	0.1419 E−5	0.1370 E−5	0.0049 E−5
6	0.4972 E−7	0.4717 E−7	0.0255 E−7

Table 3: Approximation convergence in dependence of the polynomial degree.

mixed approximation polynomial. In the table, the maximum and minimum of the absolute value of the extremes and their difference are given. The absolute values of the maximum and the minimum for the theoretical approximation should be identical and the difference should be 0. We did not get the polynomial of the best approximation. Further iterations did not bring any new information in the frame of the chosen accuracy.

The examples were calculated by the mathematical software MATLAB. The examples demonstrate a very fast convergence of the Remez algorithm and very fast increase of the accuracy with increase of the degree of the polynomial.

References

[1] Práger, M.: *Numerická matematika I.* SPN, Praha, 1981.

[2] Ralston, A.: *A first course in numerical analysis.* Oxford University Press, New York, 2001.

[3] Sýkorová, I.: On a mixed approximation problem. Mundus Symbolicus **7** (1999), 63–68.

[4] Sýkorová, I.: An iterative method for the computation of the mixed approximation polynomial. Mundus Symbolicus **8** (2000), 61–69.

[5] Vitásek, E.: *Numerické metody.* SNTL, Praha, 1987.

Conference Applications of Mathematics 2015, in honor of the birthday anniversaries of Ivo Babuška (90), Milan Práger (85), and Emil Vitásek (85)
J. Brandts, S. Korotov, M. Křížek, K. Segeth, J. Šístek, T. Vejchodský (Eds.)
Institute of Mathematics AS CR, Prague 2015

ON THE QUALITY OF LOCAL FLUX RECONSTRUCTIONS FOR GUARANTEED ERROR BOUNDS

Tomáš Vejchodský

Institute of Mathematics, Czech Academy of Sciences
Žitná 25, Praha 1, Czech Republic
vejchod@math.cas.cz

Abstract: In this contribution we consider elliptic problems of a reaction-diffusion type discretized by the finite element method and study the quality of guaranteed upper bounds of the error. In particular, we concentrate on complementary error bounds whose values are determined by suitable flux reconstructions. We present numerical experiments comparing the performance of the local flux reconstruction of Ainsworth and Vejchodský [2] and the reconstruction of Braess and Schöberl [5]. We evaluate the efficiency of these flux reconstructions by their comparison with the optimal flux reconstruction computed as a global minimization problem.

Keywords: a posteriori error estimates, complementarity, index of effectivity, elliptic problem, reaction-diffusion, singular perturbation

MSC: 65N15, 65N30

1. Introduction

The popularity of the complementary error bounds has grown during recent years due to their favourable properties. They provide guaranteed upper bounds on the error, they are locally efficient, robust, and there are fast algorithms for their computation. The main idea of these error bounds goes back the 'method of hypercircle' [15, 19] and it was developed in [3, 8, 9, 10, 12, 23] and other papers. During recent years this idea attracted a lot of attention [13, 16, 20, 21] and it was used even for partial differential eigenvalue problems [18]. Error bounds of this type for reaction-diffusion problems were recently presented in [6, 11, 17, 22, 24] and elsewhere.

However, the papers [1, 2] are the only one (to our knowledge) where a locally computable and robust upper bound on the error of the finite element solution is presented. The main goal of this contribution is to compare the accuracy of this error bound with the bound proposed in [5] for the Poisson problem, see also [7] and [4, Algorithm 9.3].

We consider the following linear elliptic problem of the reaction-diffusion type with mixed boundary conditions:

$$-\Delta u + \kappa^2 u = f \quad \text{in } \Omega; \qquad u = 0 \quad \text{on } \Gamma_{\mathrm{D}}; \qquad \partial u/\partial \boldsymbol{n} = g_{\mathrm{N}} \quad \text{on } \Gamma_{\mathrm{N}}. \tag{1}$$

Here, $\Omega \subset \mathbb{R}^2$ is a domain, $\boldsymbol{n}$ stands for the unit outward-facing normal vector to the boundary $\partial\Omega$, the portions Γ_{D} and Γ_{N} of the boundary $\partial\Omega$ are open, disjoint, and satisfy $\overline{\Gamma}_{\mathrm{D}} \cup \overline{\Gamma}_{\mathrm{N}} = \partial\Omega$. We assume the reaction coefficient $\kappa \geq 0$ to be piecewise constant. In order to guarantee unique solvability of (1), we assume that $\kappa > 0$ in a subdomain of Ω of a positive measure or that Γ_{D} has a positive measure.

In order to discretize this problem by the standard lowest-order finite elements, we approximate the domain Ω by a polygon Ω_h, Γ_{D} by $\Gamma_{\mathrm{D},h} \subset \partial\Omega_h$, and Γ_{N} by $\Gamma_{\mathrm{N},h} \subset \partial\Omega_h$. The weak solution in the domain Ω_h is defined as $\tilde{u} \in V = \{v \in H^1(\Omega_h) : v = 0 \text{ on } \Gamma_{\mathrm{D},h}\}$ such that

$$\mathcal{B}(\tilde{u}, v) = \mathcal{F}(v) \quad \forall v \in V, \tag{2}$$

where

$$\mathcal{B}(\tilde{u}, v) = \int_{\Omega_h} (\boldsymbol{\nabla}\tilde{u} \cdot \boldsymbol{\nabla} v + \kappa^2 \tilde{u} v)\,\mathrm{d}\boldsymbol{x} \quad \text{and} \quad \mathcal{F}(v) = \int_{\Omega_h} f v\,\mathrm{d}\boldsymbol{x} + \int_{\Gamma_{\mathrm{N},h}} g_{\mathrm{N}} v\,\mathrm{d}\boldsymbol{s}$$

for $\tilde{u}, v \in H^1(\Omega_h)$. We remark that $H^1(\Omega_h)$ stands for the usual Sobolev space $W^{1,2}(\Omega_h)$.

The approximation of the general domain Ω by a polygon Ω_h introduces a boundary approximation error $u - \tilde{u}$. In this paper we strictly distinguish between the solution in Ω and the solution in Ω_h in order to emphasize the fact that the error estimators discussed below do not include the boundary approximation error. Moreover, the numerical examples in Section 5 are posed in a circular disc and, therefore, there is a nonzero boundary approximation error. Such examples enable us to discuss the relative size of the boundary approximation error with respect to the other components of the total error estimated by the computed error bounds.

We discretize problem (2) by the lowest-order finite element method. Therefore, we consider a triangulation $\mathcal{T}_h$ of Ω_h consisting of triangular elements. The union of all triangles in $\mathcal{T}_h$ is $\overline{\Omega}_h$, the interiors of triangles in $\mathcal{T}_h$ are pairwise disjoint, and every edge of each triangle lies either on $\partial\Omega_h$ or it is completely shared by exactly two neighbouring triangles. The discretization parameter is defined as $h = \max_{K \in \mathcal{T}_h} h_K$, where $h_K = \operatorname{diam} K$. We also assume that the triangulation $\mathcal{T}_h$ is compatible with the piecewise constant coefficient κ and denote the value of κ on an element $K \in \mathcal{T}_h$ by κ_K. Using this triangulation, we define the usual finite element space

$$V_h = \{u_h \in V : u_h|_K \in P^1(K)\ \forall K \in \mathcal{T}_h\},$$

where $P^1(K)$ stands for the space of linear functions on the triangle $K \in \mathcal{T}_h$. Finally, the finite element formulation of problem (2) reads: find $u_h \in V_h$ such that

$$\mathcal{B}(u_h, v_h) = \mathcal{F}(v_h) \quad \forall v_h \in V_h. \tag{3}$$

Let us note that problem (1) can be diffusion dominated or singularly perturbed depending on the size of the reaction coefficient κ. The behaviour of the finite element method depends on the size of the discretization parameter h with respect to κ. If κh is small then possible boundary layers, which may occur if κ is large, are well resolved and the finite element solution is accurate. However, if κh is large, then boundary layers are not well captured by the mesh, the finite element solution exhibits spurious oscillations and its error is relatively large. This error behaviour has to be reflected by the error bounds. Therefore, we often distinguish the cases of small and large κh and observe differences in the accuracy.

2. Complementary error bounds

In this section we present two types of complementary error bounds. These bounds are similar, but they slightly differ in definition, assumptions, and applicability. Surprisingly, they differ considerably in performance. One bound provides accurate results for problems with large κh, while the other one for small κh.

First, let us introduce certain notation. Let $|\!|\!| v |\!|\!|^2 = \mathcal{B}(v,v)$ be the energy norm and $\|v\|_K$ be the $L^2(K)$ norm of v. Let $\Pi_K f \in P^1(K)$ be $L^2(K)$-orthogonal projection of f onto $P^1(K)$, $K \in \mathcal{T}_h$. Similarly, if γ is an edge of a triangle $K \in \mathcal{T}_h$, then $\Pi_\gamma g_{\mathrm{N}}$ is the $L^2(\gamma)$-orthogonal projection of $g_{\mathrm{N}} \in L^2(\gamma)$ onto $P^1(\gamma)$. We also define oscillation terms

$$\mathrm{osc}_K(f) = \min\left\{\frac{h_K}{\pi}, \frac{1}{\kappa_K}\right\} \|f - \Pi_K f\|_K\,, \quad \mathrm{osc}_\gamma(g_{\mathrm{N}}) = \min\{C_{\mathrm{T}}, \overline{C}_{\mathrm{T}}\}\, \|g_{\mathrm{N}} - \Pi_\gamma g_{\mathrm{N}}\|_\gamma\,,$$

where $K \in \mathcal{T}_h$ and $\gamma \subset \Gamma_{\mathrm{N},h} \cap \partial K$ is an edge. Constants C_{T} and $\overline{C}_{\mathrm{T}}$ are defined in [2] as

$$\begin{aligned} C_{\mathrm{T}}^2 &= \frac{|\gamma|}{d|K|}\frac{1}{\kappa_K}\sqrt{(2h_K)^2 + (d/\kappa_K)^2}, \\ \overline{C}_{\mathrm{T}}^2 &= \frac{|\gamma|}{d|K|}\min\{h_K/\pi, \kappa_K^{-1}\}\left(2h_K + d\min\{h_K/\pi, \kappa_K^{-1}\}\right), \end{aligned}$$

where $d = 2$ is the dimension.

To handle the Neumann boundary conditions, we seek the flux reconstruction in

$$\boldsymbol{W} = \{\boldsymbol{\tau} \in \boldsymbol{H}(\mathrm{div}, \Omega_h) : \boldsymbol{\tau}\cdot\boldsymbol{n} = \Pi_\gamma g_{\mathrm{N}} \text{ on all edges } \gamma \subset \Gamma_{\mathrm{N},h} \cap \partial K \text{ of all } K \in \mathcal{T}_h\}.$$

Having a flux reconstruction $\boldsymbol{\tau} \in \boldsymbol{W}$, we can consider the error bound $\eta(\boldsymbol{\tau})$ in the general form

$$\eta^2(\boldsymbol{\tau}) = \sum_{K\in\mathcal{T}_h}\left[\eta_K(\boldsymbol{\tau}) + \mathrm{osc}_K(f) + \sum_{\gamma\subset\Gamma_{\mathrm{N},h}\cap\partial K}\mathrm{osc}_\gamma(g_{\mathrm{N}})\right]^2, \tag{4}$$

where $\eta_K(\boldsymbol{\tau})$ is an error indicator computed from the values of $\boldsymbol{\tau}$ restricted to K only. We introduce two error indicators. If $\boldsymbol{\tau}$ on an element K satisfies the equilibration condition

$$\int_K (\Pi_K f - \kappa_K^2 u_h + \operatorname{div} \boldsymbol{\tau}|_K)\,\mathrm{d}\boldsymbol{x} = 0, \tag{5}$$

then we set

$$\eta_K^{\mathrm{a}}(\boldsymbol{\tau}) = \|\boldsymbol{\tau} - \boldsymbol{\nabla} u_h\|_K + \frac{h_K}{\pi} \left\|\Pi_K f - \kappa_K^2 u_h + \operatorname{div} \boldsymbol{\tau}\right\|_K, \tag{6}$$

otherwise $\eta_K^{\mathrm{a}}(\boldsymbol{\tau})$ is undefined. If $\kappa_K > 0$, then we put

$$\eta_K^{\mathrm{b}}(\boldsymbol{\tau}) = \left(\|\boldsymbol{\tau} - \boldsymbol{\nabla} u_h\|_K^2 + \kappa_K^{-2} \left\|\Pi_K f - \kappa_K^2 u_h + \operatorname{div} \boldsymbol{\tau}\right\|_K^2\right)^{1/2}, \tag{7}$$

otherwise $\eta_K^{\mathrm{b}}(\boldsymbol{\tau})$ is undefined. The following theorem proves that both these error indicators provide guaranteed upper bounds on the energy norm of the error.

Theorem 1. *Let $\tilde{u} \in V$ be the weak solution (2). Let both $u_h \in V$ and $\boldsymbol{\tau} \in \boldsymbol{W}$ be arbitrary. Then*

$$|\!|\!| \tilde{u} - u_h |\!|\!| \le \eta^{\mathrm{ab}}(\boldsymbol{\tau}), \tag{8}$$

where $\eta^{\mathrm{ab}}(\boldsymbol{\tau})$ is given by (4) with

$$\eta_K(\boldsymbol{\tau}) = \begin{cases} \min\{\eta_K^{\mathrm{a}}(\boldsymbol{\tau}), \eta_K^{\mathrm{b}}(\boldsymbol{\tau})\} & \text{if (5) holds in } K \in \mathcal{T}_h \text{ and } \kappa_K > 0, \\ \eta_K^{\mathrm{a}}(\boldsymbol{\tau}) & \text{if (5) holds in } K \in \mathcal{T}_h \text{ and } \kappa_K = 0, \\ \eta_K^{\mathrm{b}}(\boldsymbol{\tau}) & \text{if (5) does not hold in } K \in \mathcal{T}_h \text{ and } \kappa_K > 0, \end{cases}$$

Proof. Proofs of different variants of this theorem can be found in many places in the literature. The main idea traces back to the method of hypercircle and [19]. For the reader's convenience we briefly present the main steps of the proof and refer to [2] for details.

Let $v \in V$ be arbitrary. Using the weak formulation (2) for $\tilde{u}$, splitting the integrals in definitions of $\mathcal{B}$ and $\mathcal{F}$ into sums over all elements in $\mathcal{T}_h$, and applying the divergence theorem for $\boldsymbol{\tau} \in \boldsymbol{W}$, we obtain the identity

$$\mathcal{B}(\tilde{u} - u_h, v) = \sum_{K \in \mathcal{T}_h} \left[\int_K (\boldsymbol{\tau} - \boldsymbol{\nabla} u_h) \cdot \boldsymbol{\nabla} v\,\mathrm{d}\boldsymbol{x} + \int_K (\Pi_K f - \kappa_K^2 u_h + \operatorname{div} \boldsymbol{\tau}) v\,\mathrm{d}\boldsymbol{x} \right.$$
$$\left. + \int_K (f - \Pi_K f) v\,\mathrm{d}\boldsymbol{x} + \sum_{\gamma \subset \Gamma_{\mathrm{N},h} \cap \partial K} \int_\gamma (g_{\mathrm{N}} - \Pi_\gamma g_{\mathrm{N}}) v\,\mathrm{d}\boldsymbol{s}\right]. \tag{9}$$

For brevity, let us denote $\boldsymbol{g} = \boldsymbol{\tau} - \boldsymbol{\nabla} u_h$ and $r_K = \Pi_K f - \kappa_K^2 u_h + \operatorname{div} \boldsymbol{\tau}$. If the equilibration condition (5) is satisfied in K then we obtain

$$\int_K r_K v\,\mathrm{d}\boldsymbol{x} \le \int_K r_K (v - \bar{v}_K)\,\mathrm{d}\boldsymbol{x} \le \|r_K\|_K \|v - \bar{v}_K\|_K \le \frac{h_K}{\pi} \|r_K\|_K \|\boldsymbol{\nabla} v\|_K,$$

where $\bar{v}_K = |K|^{-1} \int_K v \,\mathrm{d}\boldsymbol{x}$ and we use the Poincaré inequality [14]. Using this estimate, we easily bound the first two integrals on the right-hand side of (9) as

$$\int_K \boldsymbol{g} \cdot \boldsymbol{\nabla} v \,\mathrm{d}\boldsymbol{x} + \int_K r_K v \,\mathrm{d}\boldsymbol{x} \leq \eta_K^{\mathrm{a}}(\boldsymbol{\tau}) \left\|\boldsymbol{\nabla} v\right\|_K \leq \eta_K^{\mathrm{a}}(\boldsymbol{\tau}) |\!|\!| v |\!|\!|_K, \tag{10}$$

where $|\!|\!| v |\!|\!|_K^2 = \left\|\boldsymbol{\nabla} v\right\|_K^2 + \kappa_K^2 \left\| v \right\|_K^2$ stands for the local energy norm. Alternatively, if $\kappa_K > 0$, we can bound these two integrals as

$$\int_K \boldsymbol{g} \cdot \boldsymbol{\nabla} v \,\mathrm{d}\boldsymbol{x} + \int_K r_K v \,\mathrm{d}\boldsymbol{x} \leq \left\|\boldsymbol{g}\right\|_K \left\|\boldsymbol{\nabla} v\right\|_K + \left\|r_K\right\|_K \left\| v \right\|_K \leq \eta_K^{\mathrm{b}}(\boldsymbol{\tau}) |\!|\!| v |\!|\!|_K. \tag{11}$$

To finish the proof we use (10) and (11) in (9), estimate the last two integrals on the right-hand side of (9) by the corresponding oscillation terms, see [2], substitute $v = \tilde{u} - u_h$, and apply the Cauchy-Schwarz inequality. □

Note that neither η_K^{a} nor η_K^{b} provide an error bound if (5) does not hold and $\kappa_K = 0$. Further note that the error indicators η_K^{a} and η_K^{b} given in (6) and (7) coincide if $\boldsymbol{\tau} \in \boldsymbol{W}$ is chosen in such a way that $\Pi_K f - \kappa_K^2 u_h + \operatorname{div} \boldsymbol{\tau} = 0$ for all $K \in \mathcal{T}_h$. In this case $\eta_K^{\mathrm{a}}(\boldsymbol{\tau}) = \eta_K^{\mathrm{b}}(\boldsymbol{\tau}) = \left\|\boldsymbol{\tau} - \boldsymbol{\nabla} u_h\right\|_K$ provides the upper bound (8) even for $\kappa_K = 0$, see [2].

However, in general the indicators (6) and (7) differ. Considering the optimal flux reconstruction, the indicator η_K^{a} typically yields smaller values than η_K^{b} for small $\kappa_K h_K$ including $\kappa_K = 0$. However, if $\kappa_K h_K$ is large then η_K^{b} provides tight and robust upper bound and η_K^{a} overestimates the error unacceptably. Moreover, formulas (6) and (7) for η_K^{a} and η_K^{b} have different structures, which can be unpleasant from the practical point of view. Therefore, we unify both these indicators into a single one, which is comparatively accurate as $\min\{\eta_K^{\mathrm{a}}(\boldsymbol{\tau}), \eta_K^{\mathrm{b}}(\boldsymbol{\tau})\}$, but always (slightly) greater or equal.

Lemma 2. *Let $K \in \mathcal{T}_h$ and let η_K^{a} and η_K^{b} be given by* (6) *and* (7)*. Further, let $\boldsymbol{\tau} \in \boldsymbol{W}$ and let $\boldsymbol{\tau}|_K$ satisfy the equilibration condition* (5)*. Finally, let $\kappa_K > 0$. Then*

$$\min\{\eta_K^{\mathrm{a}}(\boldsymbol{\tau}), \eta_K^{\mathrm{b}}(\boldsymbol{\tau})\} \leq \eta_K^{\mathrm{c}}(\boldsymbol{\tau}), \tag{12}$$

where

$$\eta_K^{\mathrm{c}}(\boldsymbol{\tau}) = \left\|\boldsymbol{\tau} - \boldsymbol{\nabla} u_h\right\|_K + \min\left\{\frac{h_K}{\pi}, \frac{1}{\kappa_K}\right\} \left\|\Pi_K f - \kappa_K^2 u_h + \operatorname{div} \boldsymbol{\tau}\right\|_K. \tag{13}$$

Moreover,

$$\eta_K^{\mathrm{c}}(\boldsymbol{\tau}) \leq \sqrt{2} \min\{\eta_K^{\mathrm{a}}(\boldsymbol{\tau}), \eta_K^{\mathrm{b}}(\boldsymbol{\tau})\}. \tag{14}$$

Proof. Inequality (12) follows easily from the simple estimate

$$\eta_K^{\mathrm{b}}(\boldsymbol{\tau}) \leq \left\|\boldsymbol{\tau} - \boldsymbol{\nabla} u_h\right\|_K + \kappa_K^{-1} \left\|\Pi_K f - \kappa_K^2 u_h + \operatorname{div} \boldsymbol{\tau}\right\|_K.$$

Similarly, inequality (14) follows from the estimate

$$\left\|\boldsymbol{\tau} - \boldsymbol{\nabla} u_h\right\|_K + \kappa_K^{-1} \left\|\Pi_K f - \kappa_K^2 u_h + \operatorname{div} \boldsymbol{\tau}\right\|_K \leq \sqrt{2} \eta_K^{\mathrm{b}}(\boldsymbol{\tau}).$$

□

Lemma 2 implies that we can replace $\min\{\eta_K^{\mathrm{a}}(\boldsymbol{\tau}), \eta_K^{\mathrm{b}}(\boldsymbol{\tau})\}$ by a simpler indicator $\eta_K^{\mathrm{c}}(\boldsymbol{\tau})$ in (8) and the upper bound property still holds. On the other hand, indicator $\eta_K^{\mathrm{c}}(\boldsymbol{\tau})$ is not as tight upper bound as $\min\{\eta_K^{\mathrm{a}}(\boldsymbol{\tau}), \eta_K^{\mathrm{b}}(\boldsymbol{\tau})\}$. It can overestimate it, but at most by a factor of $\sqrt{2}$.

3. Local flux reconstructions

All error indicators η^{a}, η^{b}, and η^{c} provide an upper bound on the energy norm of the error for a wide class of fluxes $\boldsymbol{\tau} \in \boldsymbol{W}$. However, an arbitrary choice of $\boldsymbol{\tau} \in \boldsymbol{W}$ would yield a large overestimation of the error. Therefore, the goal is to construct flux $\boldsymbol{\tau} \in \boldsymbol{W}$ that yields a tight bound. Tight bounds are provided by reconstructions of Ainsworth and Vejchodský [2] and Braess and Schöberl [5] and in this paper we compare their accuracy.

The reconstruction of Ainsworth and Vejchodský [2] is based on a fast algorithm to compute boundary fluxes on element edges. These boundary fluxes are computed by solving small so-called 'topology' systems of linear algebraic equations on patches of elements sharing a common vertex. Subsequently, the flux $\boldsymbol{\tau} \in \boldsymbol{W}$ is reconstructed element-by-element using explicit formulae that differ for small and large values of κh. The resulting error bound is locally efficient and robust with respect to both the mesh size h and the reaction coefficient κ [2] over the entire range of values of κh. For future reference, we denote this flux by $\boldsymbol{\tau}_h^{\mathrm{AV}}$.

The reconstruction of Braess and Schöberl [5] is based on a solution of local problems on patches of elements around vertices of the triangulation. These local problems are formulated as mixed finite element problems and correspond to the minimization of the error bound localized to the patch with an equilibration constraint. Although this flux reconstruction was originally designed for pure diffusion problems, its generalization to the reaction-diffusion case is straightforward. However, this straightforward generalization does not yield good results for large values of κh as we will see below. For future reference, we denote this flux by $\boldsymbol{\tau}_h^{\mathrm{BS}}$.

Both of these flux reconstructions have similar features. For example, in both cases the flux reconstruction is local and based on patches of elements sharing a common vertex. If $\kappa_K h_K$ is small, namely at most of order 1, then the flux $\boldsymbol{\tau}_h^{\mathrm{AV}}$ lies in the Brezzi-Douglas-Marini space $\mathbf{BDM}^2(\mathcal{T}_h) = \{\boldsymbol{w}_h \in \boldsymbol{H}(\mathrm{div}, \Omega_h) : \boldsymbol{w}_h|_K \in [P^2(K)]^2 \ \forall K \in \mathcal{T}_h\}$, while the flux $\boldsymbol{\tau}_h^{\mathrm{BS}}$ lies in the Raviart-Thomas-Nédélec space $\mathbf{RTN}^1(\mathcal{T}_h) = \{\boldsymbol{w}_h \in \boldsymbol{H}(\mathrm{div}, \Omega_h) : \boldsymbol{w}_h|_K \in [P^1(K)]^2 \oplus \boldsymbol{x}P^1(K) \ \forall K \in \mathcal{T}_h\}$. Spaces $\mathbf{BDM}^2(\mathcal{T}_h)$ and $\mathbf{RTN}^1(\mathcal{T}_h)$ are quite similar. They both contain piecewise quadratic vector fields and $\mathbf{RTN}^1(\mathcal{T}_h) \subset \mathbf{BDM}^2(\mathcal{T}_h)$. In addition, these flux reconstructions are exactly equilibrated, i.e. $\Pi_K f - \kappa_K^2 u_h + \mathrm{div}\, \boldsymbol{\tau} = 0$ in all elements $K \in \mathcal{T}_h$ for both $\boldsymbol{\tau} = \boldsymbol{\tau}_h^{\mathrm{AV}}$ and $\boldsymbol{\tau} = \boldsymbol{\tau}_h^{\mathrm{BS}}$, provided $\kappa_K h_K$ is small. This means that in this case all three error indicators η_K^{a}, η_K^{b}, and η_K^{c} are actually equal for both $\boldsymbol{\tau}_h^{\mathrm{AV}}$ and $\boldsymbol{\tau}_h^{\mathrm{BS}}$. The situation is slightly different if $\kappa_K h_K$ is large, because then the reconstruction $\boldsymbol{\tau}_h^{\mathrm{AV}}$ no longer satisfies the exact equilibration condition and it does not lie in $\mathbf{BDM}^2(\mathcal{T}_h)$ any more. Instead, it lies in $\mathbf{BDM}^2(\mathcal{T}_h^*)$, where $\mathcal{T}_h^*$ is a certain special refinement of $\mathcal{T}_h$, and the employed error bound is η_K^{b}.

These similarities motivate our interest in the comparison of these two approaches. We compare them numerically on a couple of examples and find what reconstruction provides more accurate results. The second question is, what is the absolute accuracy of these local reconstructions and what is their potential for improvement. To answer this, we find the optimal flux reconstruction in the space $\mathbf{RTN}^1(\mathcal{T}_h)$. The optimal flux is obtained by a global minimization of the error bound under the weakest equilibration constraints.

4. Global flux reconstructions

In this section we present a procedure yielding the optimal flux reconstruction in a certain finite dimensional affine subspace $\boldsymbol{W}_h \subset \boldsymbol{W}$. The idea is to minimize the error bound (8) over $\boldsymbol{W}_h$. Since this error bound consists of a sum of error indicators and oscillation terms which are independent of $\boldsymbol{\tau}$ we minimize the sum of indicators only. The three error indicators we defined above correspond to the following three minimization problems:

$$\boldsymbol{\tau}_h^{\mathrm{a}} = \operatorname*{arg\,min}_{\boldsymbol{\tau}_h \in \widetilde{\boldsymbol{W}}_h} \sum_{K \in \mathcal{T}_h} \left[\eta_K^{\mathrm{a}}(\boldsymbol{\tau}_h)\right]^2, \tag{15}$$

$$\boldsymbol{\tau}_h^{\mathrm{b}} = \operatorname*{arg\,min}_{\boldsymbol{\tau}_h \in \boldsymbol{W}_h} \sum_{K \in \mathcal{T}_h} \left[\eta_K^{\mathrm{b}}(\boldsymbol{\tau}_h)\right]^2, \tag{16}$$

$$\boldsymbol{\tau}_h^{\mathrm{c}} = \operatorname*{arg\,min}_{\boldsymbol{\tau}_h \in \widetilde{\boldsymbol{W}}_h} \sum_{K \in \mathcal{T}_h} \left[\eta_K^{\mathrm{c}}(\boldsymbol{\tau}_h)\right]^2, \tag{17}$$

where $\widetilde{\boldsymbol{W}}_h = \{\boldsymbol{\tau}_h \in \boldsymbol{W}_h : \text{condition (5) holds for all } K \in \mathcal{T}_h\}$ is a subset of $\boldsymbol{W}_h$. Recalling the definitions (6) and (13) of η_K^{a} and η_K^{c}, we notice that the structure of problems (15) and (17) is the same. They are both constrained minimization problems and the only difference of indicators $\eta_K^{\mathrm{a}} \in \widetilde{\boldsymbol{W}}_h$ and $\eta_K^{\mathrm{c}} \in \widetilde{\boldsymbol{W}}_h$ is the constant multiple of the second term. On the other hand, minimization problem (16) is unconstrained and the indicator $\eta_K^{\mathrm{b}} \in \boldsymbol{W}_h$ has a different structure.

Clearly, problem (16) is a quadratic minimization problem, but problems (15) and (17) are not quadratic. Since minimization of quadratic functionals is straightforward, we transform problems (15) and (17) such that they correspond to the minimization of a functional quadratic in $\boldsymbol{\tau}_h$. For example, in case (15), we use inequality

$$\left[A_K(\boldsymbol{\tau}_h) + B_K(\boldsymbol{\tau}_h)\right]^2 \le \left(1 + \frac{1}{\xi_K}\right) A_K^2(\boldsymbol{\tau}_h) + (1 + \xi_K)\, B_K^2(\boldsymbol{\tau}_h),$$

where $A_K(\boldsymbol{\tau}_h) = \|\boldsymbol{\tau}_h - \boldsymbol{\nabla} u_h\|_K$, $B_K(\boldsymbol{\tau}_h) = (h_K/\pi)\,\|\Pi_K f - \kappa_K^2 u_h + \operatorname{div} \boldsymbol{\tau}_h\|_K$, and $\xi_K > 0$ is arbitrary. This inequality holds as equality if $\xi_K = A_K(\boldsymbol{\tau}_h)/B_K(\boldsymbol{\tau}_h)$. Thus, instead of minimizing the left-hand side of this inequality over $\boldsymbol{\tau}_h$, we equivalently minimize the right hand side over both $\xi_K > 0$ and $\boldsymbol{\tau}_h$. Note that the right-hand side is already quadratic in $\boldsymbol{\tau}_h$, but the nonlinear nature of the minimization problem

cannot be avoided and manifests itself in the nonlinear minimization with respect to ξ_K.

Using this approach, we reformulate all problems (15)–(17) to the minimization of the functional

$$J(\alpha_K, \beta_K, \boldsymbol{\tau}_h) = \sum_{K\in\mathcal{T}_h} \alpha_K \left\|\boldsymbol{\tau}_h - \boldsymbol{\nabla} u_h\right\|_K^2 + \beta_K \left\|\Pi_K f - \kappa_K^2 u_h + \operatorname{div} \boldsymbol{\tau}_h\right\|_K^2, \tag{18}$$

where α_K and β_K are suitable constants defined for all elements $K \in \mathcal{T}_h$. For convenience, we use the notation $P^0(\mathcal{T}_h) = \{\xi \in L^1(\Omega_h) : \xi|_K = \xi_K \text{ is a constant } \forall K \in \mathcal{T}_h\}$. Problems (15)–(17), respectively, are then equivalent to:

$$(\boldsymbol{\tau}_h^{\mathrm{a}}, \xi^{\mathrm{a}}) = \operatorname*{arg\,min}_{\boldsymbol{\tau}_h\in\widetilde{\boldsymbol{W}}_h,\ \xi\in P^0(\mathcal{T}_h)} J\left(1+\xi_K^{-1}, (1+\xi_K)h_K^2/\pi^2, \boldsymbol{\tau}_h\right), \tag{19}$$

$$\boldsymbol{\tau}_h^{\mathrm{b}} = \operatorname*{arg\,min}_{\boldsymbol{\tau}_h\in\boldsymbol{W}_h} J\left(1, \kappa_K^{-2}, \boldsymbol{\tau}_h\right), \tag{20}$$

$$(\boldsymbol{\tau}_h^{\mathrm{c}}, \xi^{\mathrm{c}}) = \operatorname*{arg\,min}_{\boldsymbol{\tau}_h\in\widetilde{\boldsymbol{W}}_h,\ \xi\in P^0(\mathcal{T}_h)} J\left(1+\xi_K^{-1}, (1+\xi_K)\min\{h_K^2/\pi^2, \kappa_K^{-2}\}, \boldsymbol{\tau}_h\right). \tag{21}$$

Note that in practice we solve problem (19) iteratively. We start with the natural choice $\xi_K = 1$ for all $K \in \mathcal{T}_h$, fix it, and solve the quadratic minimization problem for $\boldsymbol{\tau}_h \in \widetilde{\boldsymbol{W}}_h$. Then we update ξ_K to $\xi_K = A_K(\boldsymbol{\tau}_h)/B_K(\boldsymbol{\tau}_h)$ for all $K \in \mathcal{T}_h$ and repeat the procedure until we find an (approximate) fixed point for ξ_K. The case of problem (21) is completely analogous.

Thus, for fixed ξ_K, both problems (19) and (21) are quadratic minimization problems for the functional (18) with suitable and fixed choices of constants α_K and β_K. Namely, $\alpha_K = 1 + \xi_K^{-1}$ and $\beta_K = (1 + \xi_K)h_K^2/\pi^2$ for problem (19) and $\alpha_K = 1+\xi_K^{-1}$ and $\beta_K = (1+\xi_K)\min\{h_K^2/\pi^2, \kappa_K^{-2}\}$ for problem (21). The constraint for these minimizations is the equilibration (5), see the definition of $\widetilde{\boldsymbol{W}}_h$. The solution of this constrained minimization problem can be obtained by solving the corresponding Euler-Lagrange equations: find $\boldsymbol{\tau}_h \in \boldsymbol{W}_h$ and $d_h \in P^0(\mathcal{T}_h)$ such that

$$\mathcal{B}^*(\boldsymbol{\tau}_h, \boldsymbol{w}_h) + \mathcal{Q}^*(d_h, \boldsymbol{w}_h) = \mathcal{F}^*(\boldsymbol{w}_h) \quad \forall \boldsymbol{w}_h \in \boldsymbol{W}_h, \tag{22}$$

$$-\mathcal{Q}^*(q_h, \boldsymbol{\tau}_h) = \mathcal{G}^*(q_h) \quad \forall q_h \in P^0(\mathcal{T}_h), \tag{23}$$

where

$$\mathcal{B}^*(\boldsymbol{\tau}_h, \boldsymbol{w}_h) = \sum_{K\in\mathcal{T}_h} \int_K \left(\alpha_K \boldsymbol{\tau}_h \cdot \boldsymbol{w}_h + \beta_K \operatorname{div} \boldsymbol{\tau}_h \operatorname{div} \boldsymbol{w}_h\right) \mathrm{d}\boldsymbol{x},$$

$$\mathcal{Q}^*(d_h, \boldsymbol{w}_h) = \sum_{K\in\mathcal{T}_h} \int_K d_h \operatorname{div} \boldsymbol{w}_h \,\mathrm{d}\boldsymbol{x},$$

$$\mathcal{F}^*(\boldsymbol{w}_h) = \sum_{K\in\mathcal{T}_h} \int_K \left(\alpha_K \boldsymbol{\nabla} u_h \cdot \boldsymbol{w}_h - \beta_K(\Pi_K f - \kappa_K^2 u_h)\operatorname{div} \boldsymbol{w}_h\right) \mathrm{d}\boldsymbol{x},$$

$$\mathcal{G}^*(q_h) = \sum_{K\in\mathcal{T}_h} \int_K (\Pi_K f - \kappa_K^2 u_h) q_h \,\mathrm{d}\boldsymbol{x}.$$

Note that equality (23) corresponds to the equilibration constraint (5) and that d_h is the Lagrange multiplier. Consequently, if $\boldsymbol{\tau}_h \in \boldsymbol{W}_h$ solves (22)–(23) then it lies actually in $\widetilde{\boldsymbol{W}_h}$.

The case of the minimization problem (20) is even simpler. It is a quadratic minimization with no constraints. Therefore, its solution $\boldsymbol{\tau}_h \in \boldsymbol{W}_h$ is given by the corresponding Euler-Lagrange equations

$$\mathcal{B}^*(\boldsymbol{\tau}_h, \boldsymbol{w}_h) = \mathcal{F}^*(\boldsymbol{w}_h) \quad \forall \boldsymbol{w}_h \in \boldsymbol{W}_h, \tag{24}$$

where the constants α_K and β_K are 1 and κ_K^{-2}, respectively.

5. Numerical results

In this section, we consider two examples of reaction-diffusion problems. We solve them on a series of uniformly refined meshes and compute several error bounds of the form (8). In particular, we compute three error bounds η^{a}, η^{b}, and η^{c}, which are obtained from (4) by using indicators η_K^{a}, η_K^{b}, and η_K^{c} in place of η_K. In addition, we compute five different flux reconstructions. Namely, the local reconstructions $\boldsymbol{\tau}_h^{\mathrm{AV}}$ and $\boldsymbol{\tau}_h^{\mathrm{BS}}$ described in Section 3, and three global reconstructions $\boldsymbol{\tau}_h^{\mathrm{a}}$, $\boldsymbol{\tau}_h^{\mathrm{b}}$, and $\boldsymbol{\tau}_h^{\mathrm{c}}$ in $\boldsymbol{W}_h = \mathbf{RTN}^1(\mathcal{T}_h) \cap \boldsymbol{W}$, see Section 4. Recall that reconstructions $\boldsymbol{\tau}_h^{\mathrm{AV}}$ and $\boldsymbol{\tau}_h^{\mathrm{BS}}$ are fully equilibrated and thus $\eta^{\mathrm{a}}(\boldsymbol{\tau}_h^{\mathrm{AV}}) = \eta^{\mathrm{b}}(\boldsymbol{\tau}_h^{\mathrm{AV}}) = \eta^{\mathrm{c}}(\boldsymbol{\tau}_h^{\mathrm{AV}})$ and $\eta^{\mathrm{a}}(\boldsymbol{\tau}_h^{\mathrm{BS}}) = \eta^{\mathrm{b}}(\boldsymbol{\tau}_h^{\mathrm{BS}}) = \eta^{\mathrm{c}}(\boldsymbol{\tau}_h^{\mathrm{BS}})$. For simplicity, we denote these two numbers by $\eta(\boldsymbol{\tau}_h^{\mathrm{AV}})$ and $\eta(\boldsymbol{\tau}_h^{\mathrm{BS}})$, respectively. Further, we use Lemma 2 to improve the error bound obtained by $\eta_K^{\mathrm{c}}(\boldsymbol{\tau}_h^{\mathrm{c}})$. Once, we have computed $\boldsymbol{\tau}_h^{\mathrm{c}}$, which is an expensive calculation, we can virtually for free evaluate the error bound

$$\eta^{\min}(\boldsymbol{\tau}_h^{\mathrm{c}}) = \min\{\eta^{\mathrm{a}}(\boldsymbol{\tau}_h^{\mathrm{c}}), \eta^{\mathrm{b}}(\boldsymbol{\tau}_h^{\mathrm{c}})\},$$

which is guaranteed to be less than or equal to $\eta_K^{\mathrm{c}}(\boldsymbol{\tau}_h^{\mathrm{c}})$ and Theorem 1 implies that it is still an upper bound on the error. In order to compare the accuracy of these error bounds we use the index of effectivity $I_{\mathrm{eff}} = \eta(\boldsymbol{\tau}_h)/|||u - u_h|||$, where u is the exact solution of problem (1) defined in Ω.

Example 1. Let us consider problem (1) in the domain of the shape of three quarters of a circular disk. Namely $\Omega = \{(r,\theta) : 0 \le r < R \text{ and } \pi/2 < \theta < 2\pi\}$, where (r,θ) are the usual polar coordinates. We set $\Gamma_{\mathrm{D}} = \partial\Omega$, $\Gamma_{\mathrm{N}} = \emptyset$, and $f(r,\theta) = \left(32R^{-4/3}/9 + \kappa^2 r^{2/3} - \kappa^2 R^{-4/3} r^2\right) \sin(2\theta/3 - \pi/3)$. The exact solution to this problem $u(r,\theta) = (r^{2/3} - R^{-4/3} r^2) \sin(2\theta/3 - \pi/3)$ has a singularity at the reentrant corner and we will use it to compute the energy norm $|||u - u_h|||$ of the error. For simplicity, we consider $R = 1$ and solve the problem for various constant values of κ.

The coarsest mesh we use is shown in Figure 1 (left). We then uniformly refine this mesh several times and compute the indices of effectivity for the above described error bounds on this sequence of meshes. Figure 2 (left) presents these results for

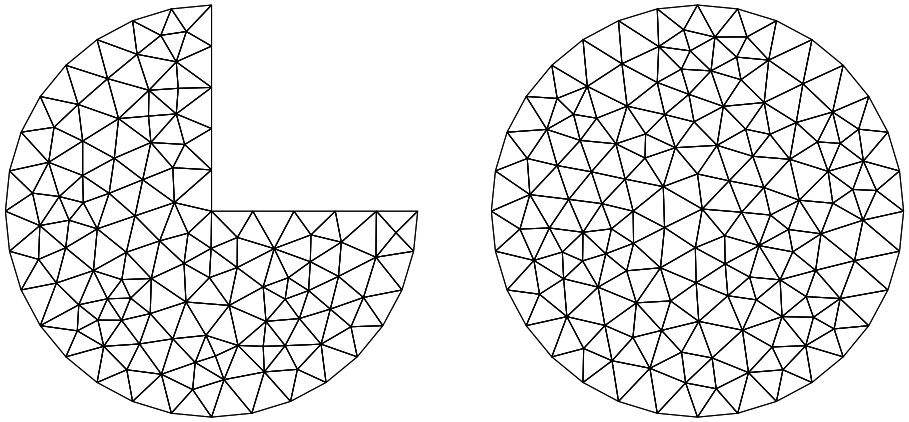


Figure 1: Domains and the coarsest meshes for Examples 1 (left) and 2 (right).

$\kappa = 100$. These results confirm that all these error bounds behave robustly with respect to the mesh size, although the case $\kappa = 100$ seems to be difficult for error bounds of this type and several of them yield indices of effectivity up to 5.

Figure 2 (right) shows how these indices of effectivity vary with κ, provided the mesh is fixed. We have chosen two times refined initial mesh. We observe that not all the error bounds provide robust bounds over this range of κ. Estimators $\eta(\boldsymbol{\tau}_h^{\mathrm{BS}})$ and $\eta^{\mathrm{a}}(\boldsymbol{\tau}_h^{\mathrm{a}})$ overestimate the error hugely if κ is large ($\kappa \geq 100$ in this case). This is not too surprising, because these two error bounds are not designed to be robust in the singularly perturbed case. On the other hand, for small values of κ (below 100) all error bounds provide very accurate results with indices of effectivity below 1.2. Only the bound $\eta(\boldsymbol{\tau}_h^{\mathrm{AV}})$ yields indices of effectivity around 1.7.

Example 2. Let Ω be a unit disk, $f = 1$, and homogeneous Dirichlet boundary conditions be prescribed on the boundary of Ω. Then, the exact solution of problem (1) is $u = (1 - r^2)/4$ for $\kappa = 0$ and $u = \kappa^{-2}(1 - I_0(\kappa r)/I_0(\kappa))$ for $\kappa > 0$. Here, $r^2 = x^2 + y^2$ and I_0 stands for the modified Bessel function of the first kind.

As in Example 1, we solve this problem on a series of uniformly refined meshes, where the coarsest mesh is presented in Figure 1 (right). Figure 3 presents the results in the same manner as Figure 2. Conclusions are basically the same as for Example 1. A difference is that in this example all error bounds provide consistently better results than in Example 1. The reason probably is that the exact solution in this example has no singularity and that the right-hand side f is constant and thus, there are no quadrature errors and the oscillation term vanishes.

If κ is small (below 100) then all error bounds yield almost exact results. An exception is $\eta(\boldsymbol{\tau}_h^{\mathrm{AV}})$ which overestimates the error by about 7 % with a worse accuracy already for $\kappa = 10$. On the other hand if κ is large (above 100) then the local bound $\eta(\boldsymbol{\tau}_h^{\mathrm{BS}})$ and the global bound $\eta^{\mathrm{a}}(\boldsymbol{\tau}_h^{\mathrm{a}})$ overestimate the error hugely. However, all the other error bounds provide almost exact results. The intermediate range of values of κ around 100 seems to be problematic for all considered error bounds, because they all exhibit the least accurate values there.

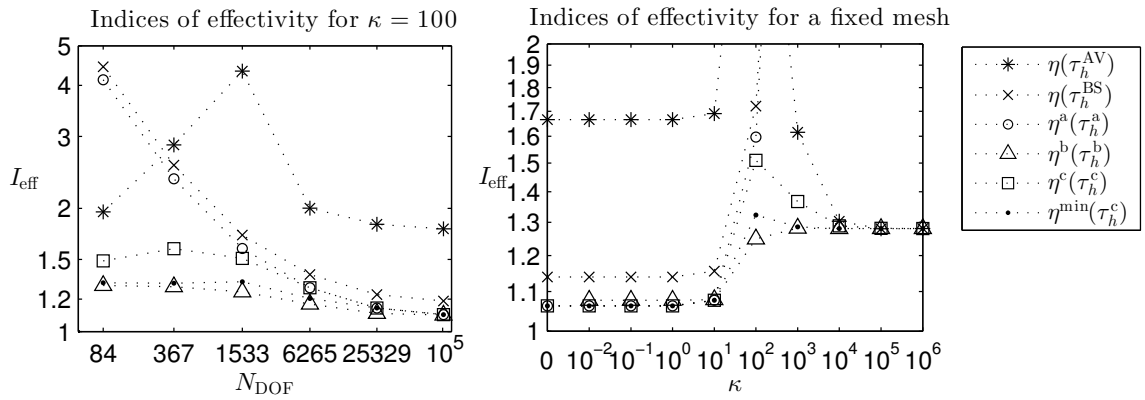


Figure 2: Results of Examples 1. The left panel shows the variations of the index of effectivity for various error bounds on a sequence of uniformly refined meshes for $\kappa = 100$. The mesh sizes h for these meshes are approximately $0.24, 0.12, 0.060, 0.030, 0.015, 0.0075$, respectively. The right panel presents their variation with respect to κ on the mesh with $N_{\text{DOF}} = 1533$ ($h \approx 0.060$), i.e. the two times refined the initial mesh.

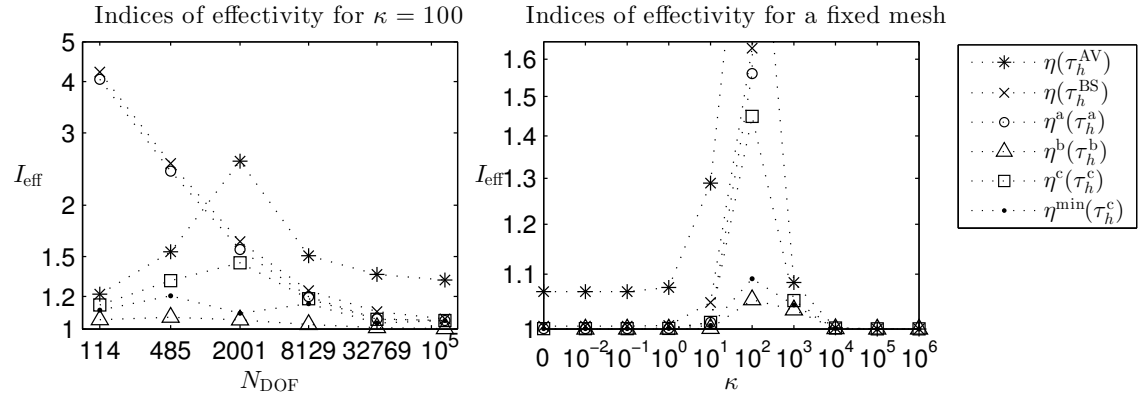


Figure 3: Results of Examples 2. The left panel shows the variations of the index of effectivity for various error bounds on a sequence of uniformly refined meshes for $\kappa = 100$. The mesh sizes h for these meshes are approximately $0.24, 0.12, 0.062, 0.031, 0.016, 0.0078$, respectively. The right panel presents their variation with respect to κ on the mesh with $N_{\text{DOF}} = 2001$ ($h \approx 0.062$), i.e. the two times refined the initial mesh.

6. Conclusions

We have compared the accuracy of two local flux reconstructions and assessed their accuracy with respect to optimal reconstructions computed as global minimization problems. We observe that both the locally computed error bounds provide good accuracy if κh is smaller than approximately $1/2$. However, the bound $\eta(\boldsymbol{\tau}_h^{\text{BS}})$ provides results close to the optimal values computed by the global minimization and

performs considerably better than $\eta(\boldsymbol{\tau}_h^{\mathrm{AV}})$. On the other hand, if κh is larger than approximately 50 then the bound $\eta(\boldsymbol{\tau}_h^{\mathrm{BS}})$ overestimates the true error unacceptably. The reason is the unnatural form of the error bound and too restrictive equilibration of $\boldsymbol{\tau}_h^{\mathrm{BS}}$. Similarly, even the globally computed bound $\eta^{\mathrm{a}}(\boldsymbol{\tau}_h^{\mathrm{a}})$ overestimates the error unacceptably. Nevertheless, all the other error bounds provide accurate results. Namely, the local flux reconstruction $\boldsymbol{\tau}_h^{\mathrm{AV}}$ yields practically as accurate results as the globally computed optimal reconstructions. The intermediate values of κh seem to be problematic for the accuracy of error bounds of the considered type. Although all error bounds provide acceptable results for these values of κh, their accuracy is worse and sometimes considerably worse than their accuracy for other values of κh. (Bound $\eta^{\mathrm{b}}(\boldsymbol{\tau}_h^{\mathrm{b}})$ in Example 1 being the only exception in the provided examples.)

In general, comparing the two locally computed error bounds, we may conclude that $\eta(\boldsymbol{\tau}_h^{\mathrm{BS}})$ is very accurate and yields close to optimal results for κh small. However, if κh is large then than $\eta(\boldsymbol{\tau}_h^{\mathrm{BS}})$ fails. The second locally constructed error bound $\eta(\boldsymbol{\tau}_h^{\mathrm{AV}})$ is less accurate for small values of κh, but still provides acceptable values. For large values of κh it gives nearly optimal results.

The secondary conclusion we can draw from the performed experiments, concerns the globally computed reconstructions and the three possible forms of the error bounds. The form η^{a} cannot be recommended in general, because it provides accurate results for small values of κh only. The form η^{b} provides accurate results over the whole range of values of $\kappa h > 0$. Even more, it provides the best results except for cases with small κh, where it is only slightly worse than η^{a}. The disadvantage of η^{b} is the fact that it is undefined in the important case $\kappa = 0$. Therefore, we can recommend to use η^{c} as a robust solution. The bound η^{c} and especially its improved variant $\eta^{\min}$ provides results that are close to the best in all cases.

The obtained results suggest several directions for future investigations. First, there is a potential for further improvements of the bound $\eta(\boldsymbol{\tau}_h^{\mathrm{AV}})$, which is not as accurate as it could be for small and intermediate values of κh. Second, the bound $\eta(\boldsymbol{\tau}_h^{\mathrm{AV}})$ is almost optimal for large values of κh, but this flux reconstruction is constructed on the refined mesh $\mathcal{T}_h^*$. However, the performance of the global flux reconstructions $\boldsymbol{\tau}_h^{\mathrm{b}}$ and $\boldsymbol{\tau}_h^{\mathrm{c}}$ clearly shows that a robust reconstruction is possible even on the original mesh $\mathcal{T}_h$. Therefore, we may try to simplify the construction of $\boldsymbol{\tau}_h^{\mathrm{AV}}$ for large values of κh and define it on $\mathcal{T}_h$ only while keeping its robust and accurate performance. Third, the bound $\eta(\boldsymbol{\tau}_h^{\mathrm{BS}})$ can be improved and redefined in such a way that it is robust and accurate even in the singularly perturbed case.

Finally, let us point out that the presented error bounds estimate the error $\tilde{u} - u_h$, which includes the discretization error, quadrature errors, round-off errors, and the error of the solver of linear algebraic equations. However, these error bounds ignore the domain approximation error $u - \tilde{u}$. Therefore, they could theoretically underestimate the total error $u - u_h$ in the case of large domain approximation error. Both the presented examples exhibit nonzero domain approximation error, but the used meshes seem to approximate the exact domain Ω well, because we do not observe any indices of effectivity below zero.

Acknowledgements

This work has been supported by grant No. P101/14-02067S of the Czech Science Foundation and by RVO 67985840.

References

[1] Ainsworth, M. and Vejchodský, T.: Fully computable robust a posteriori error bounds for singularly perturbed reaction–diffusion problems. Numer. Math. **119** (2011), 219–243.

[2] Ainsworth, M. and Vejchodský, T.: Robust error bounds for finite element approximation of reaction-diffusion problems with non-constant reaction coefficient in arbitrary space dimension. Comput. Methods Appl. Mech. Engrg. **281** (2014), 184–199.

[3] Aubin, J. P. and Burchard, H. G.: Some aspects of the method of the hypercircle applied to elliptic variational problems. In: *Numerical Solution of Partial Differential Equations, II (SYNSPADE 1970) (Proc. Sympos., Univ. of Maryland, College Park, Md., 1970)*, pp. 1–67. Academic Press, New York, 1971.

[4] Braess, D.: *Finite elements: Theory, fast solvers, and applications in elasticity theory*. Cambridge University Press, Cambridge, 2007, 3rd edn.

[5] Braess, D. and Schöberl, J.: Equilibrated residual error estimator for edge elements. Math. Comp. **77** (2008), 651–672.

[6] Cheddadi, I., Fučík, R., Prieto, M.I., and Vohralík, M.: Guaranteed and robust a posteriori error estimates for singularly perturbed reaction–diffusion problems. M2AN Math. Model. Numer. Anal. **43** (2009), 867–888.

[7] Destuynder, P. and Métivet, B.: Explicit error bounds in a conforming finite element method. Math. Comp. **68** (1999), 1379–1396.

[8] Haslinger, J. and Hlaváček, I.: Convergence of a finite element method based on the dual variational formulation. Apl. Mat. **21** (1976), 43–65.

[9] Kelly, D. W.: The self-equilibration of residuals and complementary a posteriori error estimates in the finite element method. Internat. J. Numer. Methods Engrg. **20** (1984), 1491–1506.

[10] Křížek, M.: Conforming equilibrium finite element methods for some elliptic plane problems. RAIRO Anal. Numér. **17** (1983), 35–65.

[11] Kunert, G.: A posterior H^1 error estimation for a singularly perturbed reaction diffusion problem on anisotropic meshes. IMA J. Numer. Anal. **25** (2005), 408–428.

[12] Ladevèze, P. and Leguillon, D.: Error estimate procedure in the finite element method and applications. SIAM J. Numer. Anal. **20** (1983), 485–509.

[13] Parés, N., Santos, H., and Díez, P.: Guaranteed energy error bounds for the Poisson equation using a flux-free approach: solving the local problems in subdomains. Internat. J. Numer. Methods Engrg. **79** (2009), 1203–1244.

[14] Payne, L. E. and Weinberger, H. F.: An optimal Poincaré inequality for convex domains. Arch. Rational Mech. Anal. **5** (1960), 286–292 (1960).

[15] Prager, W. and Synge, J. L.: Approximations in elasticity based on the concept of function space. Quart. Appl. Math. **5** (1947), 241–269.

[16] Repin, S.: *A posteriori estimates for partial differential equations*, *Radon Series on Computational and Applied Mathematics*, vol. 4. de Gruyter, Berlin, 2008.

[17] Repin, S. and Sauter, S.: Functional a posteriori estimates for the reaction-diffusion problem. C. R. Math. Acad. Sci. Paris **343** (2006), 349–354.

[18] Šebestová, I. and Vejchodský, T.: Two-sided bounds for eigenvalues of differential operators with applications to Friedrichs, Poincaré, trace, and similar constants. SIAM J. Numer. Anal. **52** (2014), 308–329.

[19] Synge, J. L.: *The hypercircle in mathematical physics: a method for the approximate solution of boundary value problems*. Cambridge University Press, New York, 1957.

[20] Vejchodský, T.: Guaranteed and locally computable a posteriori error estimate. IMA J. Numer. Anal. **26** (2006), 525–540.

[21] Vejchodský, T.: Complementarity based a posteriori error estimates and their properties. Math. Comput. Simulation **82** (2012), 2033–2046.

[22] Verfürth, R.: A note on constant-free a posteriori error estimates. SIAM J. Numer. Anal. **47** (2009), 3180–3194.

[23] de Veubeke, B.F.: Displacement and equilibrium models in the finite element method. In: O. Zienkiewicz and G. Hollister (Eds.), *Stress Analysis*, pp. 145–197. Wiley, London, 1965.

[24] Zhang, B., Chen, S., and Zhao, J.: Guaranteed a posteriori error estimates for nonconforming finite element approximations to a singularly perturbed reaction–diffusion problem. Appl. Numer. Math. **94** (2015), 1–15.

Conference Applications of Mathematics 2015, in honor of the birthday anniversaries of Ivo Babuška (90), Milan Práger (85), and Emil Vitásek (85)
J. Brandts, S. Korotov, M. Křížek, K. Segeth, J. Šístek, T. Vejchodský (Eds.)
Institute of Mathematics CAS, Prague 2015

VISCOSITY SOLUTIONS TO A NEW PHASE-FIELD MODEL FOR MARTENSITIC PHASE TRANSFORMATIONS

Peicheng Zhu

Department of Mathematics, Shanghai University
Shangda Road 99, Shanghai 200444
P. R. China
pczhu@shu.edu.cn

Abstract: We investigate a new phase-field model which describes martensitic phase transitions, driven by material forces, in solid materials, e.g., shape memory alloys. This model is a nonlinear degenerate parabolic equation of second order, its principal part is not in divergence form in multi-dimensional case. We prove the existence of viscosity solutions to an initial-boundary value problem for this model.

Keywords: phase-field model, martensitic phase transitions, viscosity solution, initial-boundary value problem

MSC: 35D40, 35K65

1. Introduction

The result presented in this talk is mainly from a recent work [8] by the author and his coworker.

Martensitic transformations are displacive, diffusionless and are responsible for the formation of some microstructures, like martensite which is a key microstructure of some materials and thus determines properties of those materials, for example, shape memory effect [16, 17] of shape memory alloys. Martensite can grow at temperatures close to absolute zero and at speeds in excess of 1000ms^{-1}. Thus it is very difficult to obtain, by observing this process directly, useful information to understand its mechanism, instead mathematical modeling is a powerful tool, for instance, phase-field method has been proved extremely powerful to both theoretical and numerical analysis of phenomena in materials science (see e.g., [9, 10, 20]).

To understand this type of rapidly changing processes, the author and his coworker proposed in [2, 4] a new phase-field model, which consists of a linear elasticity system and a nonlinear degenerate parabolic equation of second order. In this talk we neglect the elasticity effect of solids and formulate a little simpler model. To formulate

an initial-boundary value problem for this model, we first introduce some notations. Let Ω be an open bounded domain in $\mathbb{R}^3$ with smooth boundary $\partial\Omega$. It represents the points of a material body. Define $Q_t := (0,t)\times\Omega$. Then the model reads

$$S_t = -c\left(\hat{\psi}'(S) - \nu\Delta_x S\right)|\nabla_x S| \tag{1}$$

which is satisfied in Q_T with $T>0$. Here, S is an order parameter taking the values between 0 and 1, and $S\approx 0$ and $S\approx 1$ indicate that the material is in phases γ and γ', respectively. ∇_x and Δ_x are, respectively, the gradient and Laplace operators, and S_t denotes the partial derivative of S with respect to t, and

$$|\nabla_x S| = \Big(\sum_{i=1}^{3}|\partial_{x_i}S|^2\Big)^{\frac{1}{2}}.$$

$\hat{\psi}'(S)$ is the derivative of the function $\hat{\psi}(S)$ which is taken as a double-well potential so that $\hat{\psi}(S)$ has at least two local minima, say $S=0$ and $S=1$, and a maximum in-between. It holds that $\hat{\psi}\,'(0)=\hat{\psi}\,'(1)=0$. c,ν are positive constants.

To derive the model, we choose a free energy $\Psi(t)=\int_\Omega \psi(S,\nabla_x S)dx$ with the density

$$\psi(S,\nabla_x S) = \hat{\psi}(S) + \frac{\nu}{2}|\nabla_x S|^2.$$

Straightforward computations show that if equation (1) is satisfied, then the validness of the second law of thermodynamics is guaranteed (cf. Alber and Zhu [2, 3]).

We add, respectively, the following Dirichlet boundary and initial conditions

$$S|_{[0,T]\times\partial\Omega} = 0, \tag{2}$$
$$S|_{\{t=0\}\times\bar{\Omega}} = S_0. \tag{3}$$

Thus we complete the formulation of the initial-boundary value problem.

Let us now compare this model with the Allen-Cahn model which has been widely accepted as a model for phase separation driven by mean curvature, and comprises of

$$S_t = -c\,(\hat{\psi}'(S) - \nu\Delta_x S). \tag{4}$$

This differs from (1) by the gradient term $|\nabla_x S|$. We conclude that:

(i) Equation (1) is degenerate, non-uniformly parabolic with non-smooth coefficients; while (4) is uniformly parabolic with smooth coefficients.

(ii) Our model implies that after a part of a material changes to, say, phase 1 over an open sub-region, then we have $\nabla_x S=0$ which together with Eq. (1) implies $S_t=0$, thus S keeps the same value which means the material is kept in phase 1 over that sub-region. This is confirmed by observation. However in the Allen-Cahn model there is no such property, namely, the material is still changing after it achieves its

equilibrium over an open sub-domain because even if $\nabla_x S = 0$ over an open sub-region one cannot obtain from (4) that $S_t = 0$ over that sub-domain. Thus, the Allen-Cahn model is suitable for phase separation.

We shall prove the existence of viscosity solutions to problem (1)–(3). The principle part, i.e. $c\nu|\nabla_x S|\Delta_x S$, of this model is not in divergence form, and the order parameter equation is degenerate. Thus to investigate the validity of problem (1)–(3), we employ the notion of viscosity solution. Introduce Hamiltonian H by

$$H(S, q, r) = -c\,(\hat{\psi}'(S) - \nu\, r)|q|, \quad q \in \mathbb{R}^3,\ r \in \mathbb{R}. \tag{5}$$

Definition 1.1 *A function S which belongs to the space $C(\bar{Q}_T)$, is called a viscosity solution to problem* (1)–(3) *if S satisfies both i) and ii) below:*

i) S is a sub-viscosity solution to (1)–(3)*, i.e. for any function $\phi(t, x)$ in $C^{2,1}(\bar{Q}_T)$, if $S - \phi$ attains its local maximum at (τ, y), then*

$$\phi_t(\tau, y) \le H(S(\tau, y), \nabla_x \phi(\tau, y), \Delta_x \phi(\tau, y)), \tag{6}$$

and there holds that $S(t, x) \le 0$ for all $(t, x) \in [0, T] \times \partial\Omega$, and that $S(0, x) \le S_0(x)$ for all $x \in \Omega$;

ii) S is a super-viscosity solution to (1)–(3)*, i.e. for any function $\phi(t, x)$ in $C^{2,1}(\bar{Q}_T)$, if $S - \phi$ achieves its local minimum at (τ, y), then*

$$\phi_t(\tau, y) \ge H(S(\tau, y), \nabla_x \phi(\tau, y), \Delta_x \phi(\tau, y)), \tag{7}$$

and there holds that $S(t, x) \ge 0$ for all $(t, x) \in [0, T] \times \partial\Omega$, and that $S(0, x) \ge S_0(x)$ for all $x \in \Omega$.

Now we may state the main result.

Theorem 1.1 *Let T be a given positive constant. Suppose that $\partial\Omega \in C^{2+\beta}$ for some real positive number $\beta \in (0, 1)$, and that $S_0 \in W_0^{1,\infty}(\Omega)$ satisfies $0 \le S_0(x) \le 1$ for almost every $x \in \bar{\Omega}$. Furthermore, we assume that the potential $\hat{\psi}$ is C^2-continuous.*

Then there exists a viscosity solution S to problem (1) – (3) *in the sense of Definition 1.1, such that $0 \le S(t, x) \le 1$ for almost every $(t, x) \in \bar{Q}_T$.*

$$S \in C(\bar{Q}_T) \cap L^\infty(0, T; W_0^{1,\infty}(\Omega)),\ S_t \in L^2(Q_T). \tag{8}$$

The main difficulties in the proof of Theorem 1.1 are as follows: First, the equation of S is nonlinear, and its principal part cannot be rewritten in the divergence form, moreover, we shall find that *a priori* estimates of the highest derivative of approximate solutions depend on a term which is a function of the gradient of the order parameter, and plays a role of weight. This term is not uniformly bounded from below with respect to a small parameter, thus it leads to that standard lemmas of compactness do not apply to our problem. So we apply the concept of viscosity

solutions. Second, equation (1) is non-uniform, degenerate and its coefficients are not smooth.

Our strategies for overcoming these difficulties are in order. We make a suitable smooth approximation of the non-smooth term, then the equation becomes a uniformly parabolic one with smooth coefficients. We first derive the energy estimates, and also the uniform L^∞-bound of the gradient of S with the help of a technique from the book by Ladyzenskaya et al. [15]. The main idea behind the technique is to show that the measure of the set $\mathcal{A}_K(t) = \{x \in \Omega \mid z(t,x) > K\}$ is zero for sufficiently large K, where z is a nonlinear function in $\nabla_x v$ and v is defined by $S = \phi(v)$ with ϕ being a smooth nonlinear function. However we find it is not able to do this in one step, instead we must divide $\mathcal{A}_K(t) = \{x \in \Omega \mid z(t,x) > K\}$ into $\cup_{i=1}^{\infty} \mathcal{A}_{K,i}(t) = \{x \in \Omega \mid K + i - 1 < z(t,x) \le K + i\}$ and prove the measure of each subset is zero when K is sufficiently large. After modifying that technique in this way, we can make use of the good term to each subset and establish the L^∞-bound of the gradient of S. Then we employ these estimates to obtain the compactness of the approximate solutions.

We recall some literature related closely to our results. For the viscosity solutions, we refer to Crandall and Lions [12], Crandall, Ishii and Lions [11]. For the model investigated in this talk, the study from various aspects has been carried out, see Alber and Zhu [2, 3, 4, 5, 6, 7], Kawashima and Zhu [14], Ou and Zhu [18], Zhu [21, 22, 23]. Acharya et al. in [1], and Hildebrand et al. in [13] study a model which is closely related to ours.

Notations. Let m, n be nonnegative integers, and $p, q \ge 1$. α denotes a real number in $(0,1)$. Let $L^p(\Omega)$, $W^{m,p}(\Omega)$ are standard Lebesgue and Sobolev spaces, and $H^m(\Omega) = W^{m,2}(\Omega)$. We denote by $C^{m+\alpha}(\overline{\Omega})$ the space of $m-$times differentiable functions on $\overline{\Omega}$, whose $m-$th derivative is Hölder continuous with exponent α. The space $C^{\alpha,\alpha/2}(\overline{Q}_T)$ consists of all functions on $\overline{Q}_T$, which are Hölder continuous in the parabolic distance $\mathrm{d}((t,x),(s,y)) := \sqrt{|t-s| + |x-y|^2}$. $C^{m,n}(\overline{Q}_T)$ and $C^{m+\alpha,n+\alpha/2}(\overline{Q}_T)$ are the spaces of functions, whose x-derivatives up to order m and t-derivatives up to order n belong to $C(\overline{Q}_T)$ or to $C^{\alpha,\alpha/2}(\overline{Q}_T)$, respectively.

2. Existence of solutions

2.1. Approximate solutions

To construct approximate solutions, we formulate an approximate problem to the original problem (1)–(3). To this end, for $\kappa > 0$, we smooth the term $|\nabla_x S|$ as follows

$$|\nabla_x S|_\kappa = \sqrt{|\nabla_x S|^2 + \kappa^2},$$

and choose a sequence $S_0^\kappa \in C_0^\infty(\Omega)$ such that

$$\|S_0^\kappa - S_0\|_{H^1(\Omega)} \to 0$$

as $\kappa \to 0$ since $C_0^\infty(\Omega)$ is dense in $H_0^1(\Omega)$.

Then we can approximate the initial-boundary value problem (1)–(3) by the following problem

$$S_t = c\nu|\nabla_x S|_\kappa \Delta_x S - c\hat{\psi}'(S)(|\nabla_x S|_\kappa - \kappa), \tag{9}$$

and the boundary and initial conditions become

$$S|_{[0,T]\times\partial\Omega} = 0, \tag{10}$$
$$S|_{\{0\}\times\bar{\Omega}} = S_0^\kappa. \tag{11}$$

For the sake of simplicity, we use the following notations. Define

$$a_{ij} = a_{ij}(\nabla_x S) = c\nu|\nabla_x S|_\kappa \delta_{ij}, \text{ and} \tag{12}$$
$$a = a(S, \nabla_x S) = c\hat{\psi}'(S)(|\nabla_x S|_\kappa - \kappa) \tag{13}$$

where δ_{ij} is the Kronecker delta, $i, j = 1, 2, 3$. Straightforward computations show that

$$\frac{c\sqrt{2}\,\nu}{2}(\kappa + |p|)\xi^2 \le a_{ij}\xi_i\xi_j \le c\nu(\kappa + |p|)\xi^2, \tag{14}$$
$$\left|\frac{\partial a_{ij}}{\partial p_k}\right| \le c\nu, \tag{15}$$
$$|a(S,p)| \le \mu_1(|S|)P(|p|)(\kappa + |p|)^3, \tag{16}$$
$$-\frac{\partial a(S,p)}{\partial S} \le \mu_2(|S|)P(|p|)(\kappa + |p|)^3, \tag{17}$$
$$\left|\frac{\partial a(S,p)}{\partial p_k}\right| \le \mu_3(|S|)P(|p|)(\kappa + |p|)^2. \tag{18}$$

where $P(|p|) = (\kappa + |p|)^{-2}$.

Recalling an existence theorem from [15, p. 558], we check that all conditions of this theorem are satisfied for any given $\kappa > 0$, thus we can formulate the following theorem.

Theorem 2.1 *Let $T > 0$. Assume that $\partial\Omega \in C^{2+\beta}$ with some $\beta \in (0, 1)$. For any given κ, the coefficient functions $a_{ij}(p)$ and $a(S, p)$ are continuously differentiable with respect to their arguments S, p, and* (14) – (18) *are satisfied. Suppose that the following compatibility conditions are satisfied*

$$S_0|_{\partial\Omega} = 0, \tag{19}$$
$$\nu|\nabla_x S_0(x)|_\kappa \Delta_x S_0(x) - \hat{\psi}'(S_0(x))(|\nabla_x S_0(x)|_\kappa - \kappa) = 0 \tag{20}$$

for all $x \in \partial\Omega$.

Then there exists a solution $S \in C^{2+\alpha,1+\alpha/2}(\bar{Q}_T)$ of problem (9) – (11). *This solution has derivatives $S_{tx_i} \in L^2(Q_T)$, $i = 1, 2, 3$.*

2.2. A priori estimates

We list in this subsection *a priori* estimates which are uniform in $\kappa \in (0,1]$, for the approximate solutions. For simplicity we denote $\|f\| = \|f\|_{L^2(\Omega)}$. C is a universal constant which is independent of κ and may vary from line to line.

This subsection is devoted to uniform bound of S and to the energy estimates.

Lemma 2.1 *There hold for almost every* $t \in [0,T]$

$$\|S^\kappa\|_{L^2(0,T;W^{1,\infty}(\Omega))} \leq C, \tag{21}$$

$$\int_0^t \int_\Omega (|\nabla_x S^\kappa|_\kappa |\Delta_x S^\kappa|^2 + |S_t^\kappa|^2) \mathrm{d}\tau \mathrm{d}x \leq C. \tag{22}$$

2.3. Weak solutions to the phase-field model

In this subsection we shall make use of the *a priori* estimates to investigate the limits of the approximate solutions by using the following lemma of compactness.

Lemma 2.2 (Aubin-Lions) *Let* B_0, B, B_1 *be Banach spaces satisfying that* B_0, B_1 *are reflexive and*

$$B_0 \subset\subset B \subset B_1.$$

Here, by $\subset\subset$ *we denote the compact imbedding. Define*

$$W = \left\{ f \mid f \in L^{p_0}(0,T;B_0), f' = \frac{df}{dt} \in L^{p_1}(0,T;B_1) \right\}$$

with T *being a given positive number and* $1 < p_0, p_1 < +\infty$.

Then the embedding of W *into* $L^{p_0}(0,T;B)$ *is compact.*

Proof of Theorem 1.1. We choose

$$B_0 = W^{1,\infty}(\Omega), B = C(\bar{\Omega}), B_1 = L^2(\Omega),$$

and $p_0 = p, p_1 = 2$ (where p is an arbitrary positive number greater than 1), then we infer from Lemma 2.2 that S^κ is a compact sequence in $C(\bar{Q}_T)$. Then by the standard argument for passing to limits in the theory of viscosity solutions, we complete the proof of Theorem 1.1.

Acknowledgement

This work is partly supported by Starting-up Grant for 1000 Plan Scholars from Shanghai University, P. R. China.

References

[1] Acharya, A., Matthies, K., and Zimmer, J.: Traveling wave solutions for a quasi-linear model of field dislocation mechanics. J. Mech. Phys. Solids **58** (2010), 2043–2053.

[2] Alber, H.-D. and Zhu, P.: Solutions to a model with nonuniformly parabolic terms for phase evolution driven by configurational forces. SIAM J. Appl. Math. **66 (2)** (2006), 680–699.

[3] Alber, H.-D. and Zhu, P.: Evolution of phase boundaries by configurational forces. Arch. Rational Mech. Anal. **185** (2007), 235–286.

[4] Alber, H.-D. and Zhu, P.: Solutions to a model for interface motion by interface diffusion. Proc. Royal Soc. Edinburgh. **138A** (2008), 923–955.

[5] Alber, H.-D. and Zhu, P.: Interface motion by interface diffusion driven by bulk energy: justification of a diffusive interface model. Continuum Mech. Thermodyn. **23 (2)**, (2011), 139–176.

[6] Alber, H.-D. and Zhu, P.: Solutions to a model with Neumann boundary conditions for phase transitions driven by configurational forces. Nonlinear Anal. Real World Appl. **12 (3)** (2011), 1797–1809.

[7] Alber, H.-D. and Zhu, P.: Comparison of a rapidly converging phase field model for interfaces in solids with the Allen-Cahn model. J. Elasticity **111** (2012), 153–221.

[8] Alber, H.-D. and Zhu, P.: Viscosity solutions to a new model for solid-solid phase transitions driven by material forces. Manuscript, 2015.

[9] Bhadeshia, H.: Mathematical models in materials science. Materials Sci. Tech. **24 (2)** (2008), 128–136.

[10] Chen, L.: Phase-field models for microstructure evolution. Annu. Rev. Mater. Res. **32** (2002), 113–140.

[11] Crandall, M., Ishii, H., and Lions, P.: User's guide to viscosity solutions of second order elliptic partial differential equations. Bull. AMS. **27** (1992), 1–67.

[12] Crandall, M. and Lions, P.: Viscosity solutions of Hamilton-Jacobi equations. Trans. Amer. Math. Soc. **277** (1983), 1–42.

[13] Hildebrand, F. and Miehe, C.: A regularized sharp-interface model for phase transformation accounting for prescribed sharp-interface kinetics. Proc. Appl. Math. Mech. **10** (2010), 673–676.

[14] Kawashima, S. and Zhu, P.: Traveling waves for models of phase transitions of solids driven by configurational forces. Discr. Conti. Dyna. Systems B. **15 (1)** (2011), 309–323.

[15] Ladyzenskaya, O., Solonnikov, V., and Uralceva, N.: *Linear and quasilinear equations of parabolic type*. Translations of Math. Monographs **23**, AMS, Providence, 1968.

[16] Levitas, V., Idesman, A., and Preston, D.: Microscale simulation of martensitic microstructure evolution. Phys. Rev. Letters **93** (2004), 105701-1–105701-4.

[17] Otsuka, K. and Wayman, C.: *Shape memory materials*. Cambridge Univ. Press, 1998.

[18] Ou, Y. and Zhu, P.: Spherically symmetric solutions to a model for phase transitions driven by configurational forces. J. Math. Phys. **52** (2011), 093708 pp. 21.

[19] Qin, R. and Bhadeshia, H.: Phase field method. Materials Sci. Tech. **26** (2010), 803–811.

[20] Steinbach, I.: Phase-field models in materials science. Modelling Simul. Mater. Sci. Eng. **17** (2009), 073001-1–073001-31.

[21] Zhu, P.: Solvability via viscosity solutions for a model of phase transitions driven by configurational forces. J. Diff. Eqn. **251** (2011), 2833–2852.

[22] Zhu, P.: *Solid-solid phase transitions driven by configurational forces: A phase-field model and its validity*. Lambert Academy Publishing (LAP), Germany, 2011.

[23] Zhu, P.: Regularity of solutions to a model for solid-solid phase transitions driven by configurational forces. J. Math. Anal. Appl. **389** (2012), 1159–1172.

LIST OF AUTHORS

LIST OF PARTICIPANTS

Monika Balázsová
Charles University in Prague, Czech Republic, b.moncsi@gmail.com

Larisa Beilina
Chalmers University of Technology and Gothenburg University, Sweden, larisa.beilina@chalmers.se

Michal Beneš
Czech Technical University, Prague, Czech Republic, benes@mat.fsv.cvut.cz

Hana Bílková
Czech Academy of Sciences, Prague, Czech Republic, hanka@cs.cas.cz

Radim Blaheta
Czech Academy of Sciences, Ostrava, Czech Republic, radim.blaheta@ugn.cas.cz

Jan Brandts
University of Amsterdam, the Netherlands, janbrandts@gmail.com

Roberto Castelli
VU University Amsterdam, the Netherlands, r.castelli@vu.nl

Marta Čertíková
Czech Technical University, Prague, Czech Republic, Marta.Certikova@fs.cvut.cz

Jan Chleboun
Czech Technical University, Prague, Czech Republic, chleboun@mat.fsv.cvut.cz

Yana Di
Chinese Academy of Sciences, Beijing, China, yndi@lsec.cc.ac.cn

Vít Dolejší
Charles University, Prague, Czech Republic, dolejsi@karlin.mff.cuni.cz

Jurjen Duintjer Tebbens
Czech Academy of Sciences, Prague, Czech Republic, duintjertebbens@cs.cas.cz

István Faragó
Eötvös Loránd University, Budapest, Hungary, faragois@cs.elte.hu

Miloslav Feistauer
Charles University, Prague, Czech Republic, feist@karlin.mff.cuni.cz

Drahoslava Janovská
University of Chemistry and Technology, Prague, Czech Republic, Drahoslava.Janovska@vscht.cz

Vladimír Janovský
Charles University, Prague, Czech Republic, janovsky@karlin.mff.cuni.cz

Xia Ji
Chinese Academy of Sciences, Beijing, China, jixia@lsec.cc.ac.cn

Lucie Kárná
Czech Technical University, Prague, Czech Republic, karna@fd.cvut.cz

Inderpreet Kaur
Basque Center for Applied Mathematics, Bilbao, Spain, ikaur@bcamath.org

Jaroslav Kautský
Flinders University, Adelaide, Australia, jardakau@internode.on.net

Ielizaveta Kholmetska
Czech Technical University, Prague, Czech Republic,
ielizaveta.kholmetska@fsv.cvut.cz

Štěpán Klapka
AŽD Praha s. r. o., Czech Republic, klapka.stepan@azd.cz

Kenta Kobayashi
Hitotsubashi University, Tokyo, Japan, kenta.k@r.hit-u.ac.jp

Radek Kolman
Czech Academy of Sciences, Prague, Czech Republic, kolman@it.cas.cz

Sergey Korotov
Basque Center for Applied Mathematics, Bilbao, Spain, korotov@bcamath.org

Karel Kozel
Czech Technical University, Prague, Czech Republic, Karel.Kozel@fs.cvut.cz

Michal Křížek
Czech Academy of Sciences, Prague, Czech Republic, krizek@math.cas.cz

Václav Kučera
Charles University, Prague, Czech Republic, kucera@karlin.mff.cuni.cz

Pavel Kůs
Czech Academy of Sciences, Prague, Czech Republic, kus@math.cas.cz

Torsten Linß
FernUniversität in Hagen, Germany, torsten.linss@fernuni-hagen.de

Ivo Marek
Czech Technical University, Prague, Czech Republic, marekivo@mat.fsv.cvut.cz

Jaroslav Mlýnek
Technical University of Liberec, Czech Republic, jaroslav.mlynek@tul.cz

Šárka Nečasová
Czech Academy of Sciences, Prague, Czech Republic, matus@math.cas.cz

Jiří Nedoma
Czech Academy of Sciences, Prague, Czech Republic, nedoma@cs.cas.cz

Ivana Pultarová
Czech Technical University, Prague, Czech Republic, ivana@mat.fsv.cvut.cz

Milan Práger
Czech Academy of Sciences, Prague, Czech Republic, prager@math.cas.cz

Jiří Rákosník
Czech Academy of Sciences, Prague, Czech Republic, rakosnik@math.cas.cz

Hans-Goerg Roos
Technical University of Dresden, Germany, hans-goerg.roos@tu-dresden.de

Miroslav Rozložník
Czech Academy of Sciences, Prague, Czech Republic, miro@cs.cas.cz

Karel Segeth
Czech Academy of Sciences, Prague, Czech Republic, segeth@math.cas.cz

Jan Šembera
Technical University of Liberec, Czech Republic, jan.sembera@tul.cz

Erdoğan Şen
Namik Kemal University, Tekirdag, Turkey, erdogan.math@gmail.com

Jakub Šístek
Czech Academy of Sciences, Prague, Czech Republic, sistek@math.cas.cz

Lawrence Somer
The Catholic University of America, Washington, D.C., U.S.A., somer@cua.edu

Zdeněk Strakoš
Charles University, Prague, Czech Republic, strakos@karlin.mff.cuni.cz

Petr Sváček
Czech Technical University, Prague, Czech Republic, Petr.Svacek@fs.cvut.cz

Irena Sýkorová
University of Economics, Prague, Czech Republic, irena.sykorova@vse.cz

Stanislav Sysala
Czech Academy of Sciences, Ostrava, Czech Republic,
stanislav.sysala@ugn.cas.cz

Takuya Tsuchiya
Ehime University, Matsuyama, Japan, tsuchiya@math.sci.ehime-u.ac.jp

Tomáš Vejchodský
Czech Academy of Sciences, Prague, Czech Republic, vejchod@math.cas.cz

Emil Vitásek
Czech Academy of Sciences, Prague, Czech Republic, vitas@math.cas.cz

John Whiteman
Brunel University, Uxbridge, United Kingdom, john.whiteman@brunel.ac.uk

Hehu Xie
Chinese Academy of Sciences, Beijing, China, hhxie@lsec.cc.ac.cn

Jan Zeman
Czech Technical University, Prague, Czech Republic, zemanj@cml.fsv.cvut.cz

Shuhua Zhang
Tianjin University of Finance and Economics, China, shuhua55@126.com

Zhimin Zhang
Beijing Computational Science Research Center, China and Wayne State University, U.S.A., zmzhang@csrc.ac.cn

PROGRAM OF THE CONFERENCE

Wednesday, November 18

13.00 – 14.00 Registration

14.00 – 15.00 Opening
Jiří Rákosník, Director of the Institute of Mathematics
Presentation of the Medal of the Czech Mathematical Society to Milan Práger and Emil Vitásek
Karel Segeth: Professor Ivo Babuška
Ivo Babuška: Courant element: before and after (video record)
Michal Křížek: Asteroid no. 36060. My wonderful numerical analysis teachers – Milan Práger and Emil Vitásek

15.00 – 15.30 Jan Chleboun
On uncertain data in the modeling of magnetostrictive energy harvesting

15.30 – 16.00 Coffee Break

16.00 – 16.30 John Whiteman
Towards a proof-of-concept for acoustic localisation of coronary artery stenoses

16.30 – 17.00 István Faragó
Qualitative properties in discrete space-time models of epidemic propagation

17.00 – 17.30 Sergey Korotov
Conforming post-refinements of adjacent 3D meshes

Thursday, November 19

9.00 – 9.30 Miloslav Feistauer
Discontinuous Galerkin method for the solution of dynamic elasticity problems and applications to fluid-structure interaction

9.30 – 10.00 Vít Dolejší
hp-adaptive discontinuous Galerkin method for PDEs

10.00 – 10.30 Zhimin Zhang
Some recent development in superconvergence theory

10.30 – 11.00 Coffee Break

11.00 – 11.30 Drahoslava Janovská
Filippov systems with DAE

11.30 – 12.00 Vladimír Janovský
A numerical analysis of a lumped parameter friction model

12.00 – 14.00 Lunch Break

14.00 – 14.20 Kenta Kobayashi
On the interpolation constants over triangular elements

14.20 – 14.40 Monika Balázsová
Stability analysis of the space-time discontinuous Galerkin method in the ALE framework

14.40 – 15.00 Michal Beneš
Multi-time-step domain decomposition methods for parabolic problems

15.00 – 15.20 Larisa Beilina
Iteratively regularized adaptive finite element method in the reconstruction of coefficients in Maxwell's equations

15.20 – 15.40 Coffee Break

15.40 – 16.00 Petr Sváček
On application of extended finite element method for two phase flows with treatment of surface tension and contact angles

16.00 – 16.20 Pavel Kůs
Convergence and stability of higher-order finite element solution of diffusion-reaction equation with Turing instability

16.20 – 16.40 Erdoğan Şen
The regularized trace formula for differential operator equation with unbounded operator coefficient

16.40 – 17.00 Xia Ji
C^0 IPG for transmission eigenvalue problems

18.00 – 23.00 Conference Dinner, U Seminaristy Restaurant, Spálená St. 45

Friday, November 20

9.00 – 9.30 Zdeněk Strakoš
Preconditioning and the conjugate gradient method in the context of solving PDEs

9.30 – 10.00 Radim Blaheta
Poroelasticity: LBB, locking phenomena, preconditioning

10.00 – 10.30 Hehu Xie
A full multigrid method for eigenvalue problems

10.30 – 11.00 Coffee Break

11.00 – 11.30 Takuya Tsuchiya
Error estimates for Lagrange interpolations on triangles

11.30 – 12.00 Torsten Linss
Maximum-norm a posteriori error estimates for parabolic problems

12.00 – 14.00 Lunch Break

14.00 – 14.20 Jan Zeman
Guaranteed a-posteriori error bounds in homogenization via Fourier-Galerkin methods

14.20 – 14.40 Roberto Castelli
Analytical enclosure of fundamental matrix solution with applications

14.40 – 15.00 Lucie Kárná
How message doubling improve error detection in BSC model

15.00 – 15.20 Irena Sýkorová
Some remarks on function approximation problem

15.20 – 15.40 Coffee Break

15.40 – 16.00 Gianni Pagnini
Wildland fire propagation modelling: A novel approach reconciling models based on moving interface methods and on reaction-diffusion equations

16.00 – 16.20 Yana Di
Numerical simulations on adsorption of the surfactant

16.20 – 16.40 Shuhua Zhang
Modeling and computation of transboundary industrial pollution with emission permits trading by stochastic differential game

16.40 – 17.00 Jaroslav Mlýnek
Optimization of heat radiation intensity and use of evolutionary algorithm

17.00 – 17.20 Jiří Nedoma
Dynamic contact problems in bone neoplasm analyses and the primal-dual active set (PDAS) method

Saturday, November 21

10.00 – 12.00 A walk through the Old Town