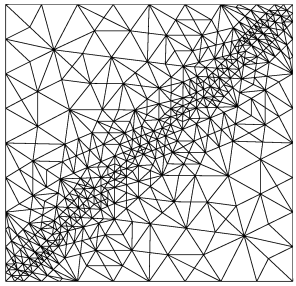
Proceedings of the International Conference

Applications of Mathematics 2013

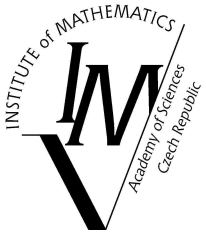
Prague, May 15–17, 2013

In honor of the 70th birthday of Karel Segeth

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



Institute of Mathematics
Academy of Sciences of the Czech Republic
Prague 2013



ISBN 978-80-85823-61-5
Institute of Mathematics
Academy of Sciences of the Czech Republic
Prague 2013

Professor Karel Segeth is seventy

Karel Segeth was born on May 10, 1943 in Prague. His father taught biology and geography in secondary school and his mother was a pediatrician. While in elementary and secondary school Karel regularly took part at the Mathematical Olympiad and got several diplomas. In 1964, he finished his studies at the Faculty of Mathematics and Physics of Charles University in Prague and started to work as research assistant in the Mathematical Institute of the Czechoslovak Academy of Sciences. He spent three months of 1966 in academical institutions in Novosibirsk, Moscow, and Kiev. During the period 1969–1970 he worked at the University of Maryland in College Park, where he developed numerical software for Prof. Ivo Babuška. In 1969 he received the academic title RNDr. from the Faculty of Mathematics and Physics of Charles University and three years later he defended his doctoral thesis *On universally optimal quadrature formulae involving values of derivatives of integrand* at the Mathematical Institute of the Academy and got the scientific degree Candidate of Sciences (equivalent to PhD). His advisor was Ivo Babuška. In 1996 Karel Segeth passed his habilitation at the Faculty of Mathematics and Physics of Charles University and received the title Doc. (equivalent to Associate Professor). In 2004 he became Full Professor in Applied Mathematics at the University of West Bohemia in Pilsen.

The scientific activities of Prof. Segeth are very broad. Among computational methods for numerical solution of partial differential equations, he deals with problems in geophysics, archaeology, and also in medicine (e.g. diffusion in layered structures of the human brain). No wonder that he publishes his results in an extensive spectrum of scientific journals, such as Numerische Mathematik, Geophysics, Applications of Mathematics, Biophysical Journal, Czechoslovak Mathematical Journal, Tectonophysics, International Journal for Numerical Methods in Fluids, Computers & Geosciences, and Mathematics and Computers in Simulation.

The main research interest of Prof. Segeth is the solution of problems of mathematical physics and numerical modeling of physical phenomena (e.g. semiconductor devices, electric and magnetic fields). At present, Prof. Segeth examines mainly *a posteriori* bounds for the discretization error in numerical solutions of differential equations. Their analytical solution in explicit form is usually not known. Therefore, some approximate methods need to be used. Once the approximate solution is computed, the discretization error can be estimated *a posteriori* by means of sophisticated mathematical methods. Prof. Segeth focuses on the finite element and finite volume methods for numerical solution of boundary value problems for partial differential equations of elliptic type and also the method of lines for solving initial-boundary value problems for nonlinear evolution equations of parabolic type. This has a close connection to his interest in numerical solution of large systems of algebraic equations by the methods of cyclic reduction and conjugate gradients, fast Fourier transform, the multigrid method, etc. Prof. Segeth showed the practical importance of *a posteriori* error estimates of the discretization error, which can be effectively used in the finite element method for adaptive mesh refinements.

Further areas of interest of Prof. Segeth are mathematical methods for solving real-life problems in geophysics and archaeology. At present he deals with numerical simulation of solid particles in slowly flowing viscous liquids. For many years he cooperated with Professor Irwin Scollar from the Laboratorium für Feldarcheologie, Rheinisches Landesmuseum in Bonn. By means of the Fourier analysis of aerial photographs or terrain data (gravitational or electromagnetic) they developed methods for discovering new archaeological deposits (see Pokroky Mat. Fyz. Astronom. 2011, pp. 213–227) and mineral resources. The Fourier transform is also at the basis of one of his other favourite topics, namely the implementation of so-called fast algorithms (see e.g. his paper in Pokroky Mat. Fyz. Astronom. 2008, pp. 199–210). He has been contributing to this journal for many years. He published many articles in it and prepared several interesting translations.

He wrote his first monograph *Mathematical Modeling in Electromagnetic Prospecting Methods*, Charles University, Prague, 1982, 133 pp., together with Václav Bezvoda. Segeth's rich experience with the method of lines and numerical algebra are included in another monograph *Higher-Order Finite Element Methods* (coauthors P. Šolín and I. Doležel), Chapman & Hall/CRC, London, 2004, 403 pp., which has got many citations. He also contributed to Rektorys' *Survey of Applicable Mathematics*, Prometheus, Prague 1995, whose English version appeared in the prestigious publishing house Kluwer in 1994. Prof. Segeth is coeditor of 16 conference proceedings *Programs and Algorithms of Numerical Mathematics*, that he coorganizes with his colleagues from the Institute of Mathematics. His excellent knowledge of languages helped him to translate several important monographs on numerical linear algebra and continuum mechanics. Together with Petr Přikryl he translated the monograph by Jindřich Nečas and Ivan Hlaváček from Czech into English, as well as a book by Miroslav Fiedler. They also translated the famous treatise of A. A. Samarskij and J. S. Nikolajev from Russian into Czech, and another monograph by G. I. Marchuk.

Prof. Segeth has a rich pedagogical experience due to the many decades that he worked at several Czech universities, such as the Faculty of Mathematics and Physics and Faculty of Sciences of Charles University in Prague, Faculty of Mechanical Engineering of Czech Technical University in Prague, Faculty of Applied Sciences of the University of West Bohemia in Pilsen and Technical University of Liberec. He lectured numerical methods for solving large sparse systems, numerical software, programming in FORTRAN, numerical modeling of problems in electrical engineering, but also basic courses in mathematics. He is the author or coauthor of eight lecture notes. He was advisor of ten diploma students and of PhD students M. Pospíšek, P. Vaněk, V. V. Vlček, and M. Zítka. He was invited to give lectures at several world-wide known universities: Wayne State University in Detroit, the University of Texas at Austin, A & M University of Texas, the University of Maryland, the State University of New York, Keio University of Yokohama, Flinders University in Adelaide as well as at many European universities.

Due to his brilliant organization capabilities he was the Secretary of the Scientific Collegium for Mathematics of the Czechoslovak Academy of Sciences from 1982 to 1992, which was headed at that time by Prof. Miloš Zlámal. In 1994 Karel Segeth succeeded Dr. Milan Práger as the Head of the Department of Constructive Methods of Mathematical Analysis of the Mathematical Institute, and at the same time he was elected as the Head of the Scientific Council of the Mathematical Institute. After that he was the Director of the Mathematical Institute for two periods (1996–2000 and 2000–2004). From 2004 to 2009 he was the Head of the Department of Mathematics and Didactics of Mathematics of the Technical University of Liberec. During the period 2004–2008 he also headed the Department of Applied Mathematics there. He was a member of five scientific councils of university faculties at Prague, Olomouc, Liberec, and Pilsen. At present he is still a member of the Scientific Council of the University of West Bohemia in Pilsen. Together with the Union of Czech Mathematicians and Physicists and the Czech Society for Mechanics, and with great enthusiasm, he has started to organize the *Babuška Prize* for the best student work in the field of Computer Science in 1994.

Since 1996 Prof. Segeth is a member of the Union of Czech Mathematicians and Physicists. In 2003 and 2004 he received two memorial medals from the Faculty of Mathematics and Physics of Charles University. He became the Deserving Member of the Union of Czech Mathematicians and Physicists in 2006.

To commemorate the 70th birthday of Prof. Karel Segeth we organized the International Conference *Applications of Mathematics 2013* at the Institute of Mathematics in Žitná 25, Prague 1, from 15 to 17 May 2013 (see `www.math.cas.cz/~am2013`).

The Scientific Committee consisted of

Ivo Babuška (University of Texas at Austin, USA)
Jan Brandts (University of Amsterdam, Netherlands)
Antti Hannukainen (Aalto University, Finland)
Sergey Korotov (Basque Center for Applied Mathematics, Spain)
Qun Lin (Academy of Mathematics and Systems Science, China)
Liping Liu (Lakehead University, Canada)
Milan Práger (Academy of Sciences, Czech Republic)
Lawrence Somer (Catholic University of America, USA)
Emil Vitásek (Academy of Sciences, Czech Republic)
Shuhua Zhang (Tianjin University of Finance and Economics, China)
Zhimin Zhang (Wayne State University, USA)

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

Karel Segeth is married with Dr. Jitka Segethová, a granddaughter of mathematician Prof. Josef Holubář. She taught mostly numerical methods at the Faculty of Mathematics and Physics of Charles University in Prague. Karel and Jitka have two daughters, Jitka and Jana, and two grandchildren. We wish Prof. Karel Segeth and his family enduring happiness and good health.

Jan Brandts and Michal Křížek

The Organizing Committee is gratefull to all authors for their contributions and to Grant MTM2011-24766 of MICINN (Spain).

LIST OF PUBLICATIONS OF KAREL SEGETH

with references to Zentralblatt and Mathematical Reviews

B. Books and chapters in monographs

[B1] V. Bezvoda, K. Segeth: *Mathematical modeling in electromagnetic prospecting methods.* Praha, Univerzita Karlova 1982, 133 pp.

[B2] J. Segethová, K. Segeth: *Numerical methods in linear algebra.* Survey of Applicable Mathematics. 2nd revised ed. Dordrecht, Kluwer Academic Publishers 1994, Vol. 2, 594–647.

[B3] J. Segethová, K. Segeth: *Numerické metody lineární algebry.* Přehled užité matematiky. 6. přepracované vydání. Praha, Prometheus 1995, díl 2, 552–602.

[B4] P. Šolín, K. Segeth, I. Doležel: *Higher-order finite element methods.* Studies in Advanced Mathematics. Boca Raton, Chapman & Hall/CRC 2004, 403 pp. + CD ROM. Zbl 1032.65132, MR2000261

[B5] K. Segeth, J. Chleboun: *Ivo Babuška.* Sto českých vědců v exilu. Praha, Academia 2011, 215–218.

[B6] K. Segeth: *On the advantages and drawbacks of a posteriori error estimation for fourth-order elliptic problems.* Numerical Methods for differential equations, optimization, and technological problems. Dordrecht, Springer 2013, 145–158.

J. Research papers published in journals

[J1] K. Segeth: *Sravnenie točnosti nekotorych formulirovok krajevych uslovij pri ispol'zovanii metoda setok.* Apl. Mat. **10** (1965), 302–307. MR0184444

[J2] K. Segeth: *On quadrature formulae involving values of derivatives.* Z. Angew. Math. Mech. **48** (1968), T104–T105. Zbl 0184.38102

[J3] K. Segeth: *On universally optimal quadrature formulae involving values of derivatives of integrand.* Czechoslovak Math. J. **19** (1969), 605–675. Zbl 0188.13203, MR0260177

[J4] V. Bezvoda, K. Segeth: *A two-layer ground in the field of an infinitely long cable.* Geophys. Prospect. **18** (1970), 343–351.

[J5] K. Segeth: *Universal approximation by hill functions.* Czechoslovak Math. J. **22** (1972), 612–640. Zbl 0247.41011, MR0310502

[J6] K. Segeth: *Numerické experimenty s univerzálními kopečkovými funkcemi.* Acta Polytech. Práce ČVUT Ser. IV (1973), 189–193.

[J7] K. Segeth: *Universal approximation by systems of hill functions.* Apl. Mat. **19** (1974), 403–436. Zbl 0305.41011, MR0388812

[J8] K. Segeth: *A remark on a class of universal hill functions.* Acta Univ. Carolin. - Math. Phys. **15** (1974), 155–156. Zbl 0314.41008, MR0390598

[J9] V. Bezvoda, K. Segeth, E. Stehlík: *Dvojdimenzionální směrová analýza geologických dat.* Acta Polytech. Práce ČVUT Ser. IV (1976), 113–118.

[J10] V. Bezvoda, K. Segeth: *A contribution to the theory of electromagnetic induction of a line source.* Stud. Geophys. Geod. **20** (1976), 366–377.

[J11] V. Kolář, P. Přikryl, K. Segeth, J. Šťastna: *Vliv normálových napětí na Tomsův jev II.* Vodohospodárský časopis **26** (1978), 34–48.

[J12] V. Bezvoda, E. Jelínková, K. Segeth: *Modern methods of the separation of regional and residual portions of potential fields.* Acta Univ. Carolin. - Geologica **27** (1980), 135–150.

[J13] V. Bezvoda, K. Segeth: *Directional and frequency filtering of geophysical data measured in a rectangular net.* Gerlands Beiträge Geophys. **90** (1981), 133–146.

[J14] K. Segeth: *Roundoff errors in the fast computation of discrete convolutions.* Apl. Mat. **26** (1981), 241–262. Zbl 0474.65025, MR0623505

[J15] V. Červ, K. Segeth: *A comparison of the accuracy of the finite-difference solution to boundary-value problems for the Helmholtz equation obtained by direct and iterative methods.* Apl. Mat. **27** (1982), 375–390. Zbl 0511.65074, MR0674982

[J16] V. Bezvoda, K. Segeth: *On the resolving power of the VLF method.* Pure Appl. Geophys. **120** (1982), 348–364.

[J17] K. Segeth: *On the choice of iteration parameters in the Stone incomplete factorization.* Apl. Mat. **28** (1983), 295–306. Zbl 0532.65020, MR0710177

[J18] V. Bezvoda, K. Segeth, Č. Tomek: *Combination and comparison of various filtering techniques in processing gravity data.* Ann. Geophys. **1** (1983), 229–234.

[J19] V. Hašek, P. Matula, K. Segeth, J. Vignatiová: *Použití výpočetní techniky při strojovém zpracování geofyzikálních dat v archeologii.* Sb. prací Filoz. fak. Brněnské Univ. **E29** (1984), 195–200.

[J20] V. Bezvoda, K. Segeth: *The electromagnetic response of an inhomogeneous layered earth - a general one-dimensional approach.* Geophysics **50** (1985), 434–442.

[J21] J. Jelínek, K. Segeth, T. R. Overton: *Three-dimensional reconstruction from projections.* Apl. Mat. **30** (1985), 92–109. Zbl 0576.65128, MR0778981

[J22] V. Bezvoda, K. Segeth: *Počítačové zpracování obrazů.* Řízení v kultuře **10** (1985), 67–73.

[J23] I. Scollar, B. Weidner, K. Segeth: *Display of archaeological magnetic data.* Geophysics **51** (1986), 623–633.

[J24] V. Bezvoda, R. Farzan, K. Segeth, G. Takó: *On numerical evaluation of integrals involving Bessel functions.* Apl. Mat. **31** (1986), 396–410. Zbl 0614.65012, MR0863034

[J25] M. Bálek, V. Hašek, Z. Měřínský, K. Segeth: *Metodický přínos kombinace letecké prospekce a geofyzikálních metod při archeologickém výzkumu na Moravě.* Archeologické rozhledy **38** (1986), 550–574, 598–600.

[J26] V. Bezvoda, E. Jelínková, K. Segeth: *Evaluation of geochemical data acquired from regular grids.* Math. Geology **18** (1986), 823–843.

[J27] V. Bezvoda, K. Segeth: *An application of fast algorithms to numerical electromagnetic modeling.* Geophys. Prospect. **35** (1987), 312–322.

[J28] V. Bezvoda, J. Ježek, K. Segeth: *A comment on "A computer program to perform transformations of gravimetric and aeromagnetic surveys".* Computers Geosci. **14** (1988), 123–124.

[J29] M. Bálek, V. Hašek, V. Ondruš, K. Segeth: *Aerial survey and geophysical methods in archaeological investigations of neolithic circular objects in Moravia.* Przeglgad Archeologiczny **36** (1989), 5–16.

[J30] V. Bezvoda, J. Ježek, K. Segeth: *FREDPACK - a program package for linear filtering in frequency domain.* Computers Geosci. **16** (1990), 1123–1154.

[J31] V. Bezvoda, J. Hrabě, K. Segeth: *Linear filters for solving the direct problem of potential fields.* Geophysics **57** (1992), 1348–1351.

[J32] V. Bezvoda, K. Segeth: *Programs available for two-dimensional numerical modeling of the electromagnetic field.* Acta Univ. Carolin. - Math. Phys. **33** (1992), 39–52.

[J33] V. Bezvoda, J. Hrabě, K. Segeth: *Discussion on "A FORTRAN-77 computer program for three-dimensional analysis of gravity anomalies with variable density contrast".* Computers Geosci. **18** (1992), 1287.

[J34] K. Segeth: *Grid adjustment based on a posteriori error estimators.* Appl. Math. **38** (1993), 488–504. Zbl 0797.65068, MR1241452

[J35] V. Hašek, H. Petrová, K. Segeth: *Graphic representation methods in archaeological prospection in Moravia.* Sb. prací Filoz. Fak. Brněnské Univ. **E38** (1993), 111–117.

[J36] K. Segeth: *A posteriori error estimates for parabolic differential systems solved by the finite element method of lines.* Appl. Math. **39** (1994), 415–443. Zbl 0822.65068, MR1298731

[J37] M. Křížek, K. Segeth: *Co přináší základní výzkum v numerické matematice?* Vesmír **74** (1995), 206–207.

[J38] K. Segeth: *Grid adjustment for parabolic systems based on a posteriori error estimates.* J. Comput. Appl. Math. **63** (1995), 349–355. Zbl 0939.65107, MR1365575

[J39] J. Ježek, K. Schulmann, K. Segeth: *Fabric evolution of rigid inclusions during mixed coaxial and simple shear flows.* Tectonophysics **257** (1996), 203–221.

[J40] K. Segeth: *A posteriori error estimates for parabolic equations applied to the space grid adjustment.* Z. Angew. Math. Mech. **76** (1996), Suppl. 1, 531–532. Zbl 0900.65268

[J41] K. Kronrádová, K. Segeth, O. Kronrád, E. Kindler: *Mathematical and simulation model for education and manpower planning.* Z. Angew. Math. Mech. **77** (1997), Suppl. 2, 601–602. Zbl 0900.90331

[J42] K. Segeth: *A posteriori error estimates in the finite element method of lines.* Z. Angew. Math. Mech. **77** (1997), Suppl. 2, 671–672. Zbl 0900.65279

[J43] J. Ježek, S. Saic, K. Segeth, K. Schulmann: *Three-dimensional hydrodynamical modelling of viscous flow around a rotating ellipsoidal inclusion.* Computers Geosci. **25** (1999), 547–558.

[J44] K. Segeth: *A posteriori error estimation with the finite element method of lines for a nonlinear parabolic equation in one space dimension.* Numer. Math. **83** (1999), 455–475. Zbl 0936.65113, MR1715561

[J45] P. Přikryl, R. Černý, V. Havlík, K. Segeth, P. Stupka, J. Toman: *Deposition of waste water into deep mines.* Environmetrics **10** (1999), 457–466.

[J46] K. Segeth: *A posteriori error estimates and grid adjustment for a nonlinear parabolic equation.* Math. Comput. Simulation **50** (1999), 331–338. MR1717661

[J47] J. Ježek, S. Saic, K. Segeth: *Numerical modellng of the movement of a rigid particle in viscous fluid.* Appl. Math. **44** (1999), 469–479. Zbl 1060.76537, MR1727983

[J48] P. Šolín, K. Segeth: *Performance of various ODE solvers on FV-semi-discretized nonstationary compressible Euler equations.* Acta Tech. CSAV **47** (2002), 47–66. MR1898098

[J49] P. Šolín, K. Segeth: *Description of the multi-dimensional finite volume solver EULER.* Appl. Math. **47** (2002), 169–185. Zbl 1090.65532, MR1894668

[J50] P. Šolín, K. Segeth: *Examples of non-uniqueness of almost-unidirectional gas flow.* Math. Comput. Simulation **61** (2003), 229–237. Zbl 1215.76067, MR1983671

[J51] P. Šolín, K. Segeth: *Application of the method of lines to unsteady compressible Euler equations.* Internat. J. Numer. Methods Fluids **41** (2003), 519–535. Zbl 1078.76590, MR1951794

[J52] P. Šolín, K. Segeth: *Non-uniqueness of almost unidirectional inviscid compressible flow.* Appl. Math. **49** (2004), 247–268. Zbl 1099.76053, MR2059429

[J53] J. Hrabě, S. Hrabětová, K. Segeth: *A model of effective diffusion and tortuosity in the extracellular space of the brain.* Biophys. J. **87** (2004), 1606–1617.

[J54] P. Šolín, K. Segeth: *A new sequence of hierarchic prismatic elements satisfying de Rham diagram on hybrid meshes.* J. Numer. Math. **13** (2005), 295–318. Zbl 1089.78022, MR2189550

[J55] P. Šolín, K. Segeth: *Hierarchic higher-order hermite elements on hybrid triangular/quadrilateral meshes.* Math. Comput. Simulation **76** (2007), 198–204. Zbl 1135.65393, MR2392478

[J56] K. Segeth: *A review of some a posteriori error estimates for adaptive finite element methods.* Math. Comput. Simulation **80** (2010), 1589–1600. Zbl 1196.65173, MR2647253

[J57] K. Segeth: *Fourierova analýza dvojrozměrných terénních dat.* Pokroky mat. fyz. astronom. **56** (2011), 213–227.

[J58] K. Segeth: *A comparison of a posteriori error estimates for biharmonic problems solved by the FEM.* J. Comput. Appl. Math. **236** (2012), no. 18, 4788–4797. Zbl 1250.65133, MR2946409

P. Research papers published in reviewed proceedings

[P1] K. Segeth: *On universally optimal quadrature formulae involving values of derivatives of integrand.* Basic Problems of Numerical Mathematics 2. Communications. (Proceedings of Conference, Liblice 1967.) Praha, Matematický ústav ČSAV 1967, 10 pp.

[P2] V. Bezvoda, K. Segeth: *Řešení Helmholtzovy rovnice metodou konečných prvků.* Sborník konference o aplikacích matematiky. (Olomouc 1973.) Olomouc, Přírodovědecká fakulta UP 1973, 40–42.

[P3] V. Bezvoda, K. Segeth: *Výpočet harmonického pole nekonečného kabelu.* Teorie a počítače v geofyzice. (Sborník 4. pracovního semináře, Loučná n. Desnou 1974.) Brno, Geofyzika, n.p., 1974, 287–296.

[P4] V. Bezvoda, K. Segeth, E. Stehlík: *Lineární filtrace jednorozměrných a dvojrozměrných dat.* Sborník 6. celostátní konference geofyziků. (Plzeň 1975.) Brno, Geofyzika, n.p., 1975, díl 3, 373–389.

[P5] V. Bezvoda, K. Segeth: *Program pro lineární filtraci dvojrozměrných dat.* Problémy současné gravimetrie. (Sborník referátů celostátního semináře, Liblice 1976.) Praha, Geofyzikální ústav ČSAV 1976, 93–110.

[P6] K. Segeth: *Teorie aproximací v metodě konečných prvků.* Sborník přednášek letní školy o numerickém řešení eliptických rovnic metodou konečných prvků. (Praha 1974.) Praha, Univerzita Karlova 1976, 57–77.

[P7] K. Segeth: *Aproximace v metodě konečných prvků.* Software a algoritmy numerické matematiky. (Sborník referátů letní školy, Zadov 1975.) Praha, JČSMF 1976, 130–141.

[P8] K. Segeth: *Řešení okrajových úloh pro eliptické diferenciální rovnice metodou konečných prvků.* Aplikovaná matematika v inženýrské praxi. (Sborník přednášek postgraduálního kursu, Praha 1976.) Praha, pobočka ČSVTS na FEL ČVUT 1976, 33 pp.

[P9] V. Bezvoda, K. Segeth: *Řešení jedné okrajové úlohy se singulárními daty metodou konečných prvků.* Sborník 3. semináře o metodě konečných prvků a variačních metodách. (Plzeň 1977.) Plzeň, Škoda, n.p., 1977, díl 1, 1-9.

[P10] K. Segeth: *Universal approximation in the finite element method.* Theory of Nonlinear Operators. Constructive Aspects. (Proceedings of International Summer School, Berlin 1975.) Berlin, Akademie-Verlag 1977, 389–393. Zbl 0392.65030, MR0468247

[P11] K. Segeth: *Řešení okrajových úloh pro eliptické diferenciální rovnice metodou konečných prvků.* Aplikovaná matematika v inženýrské praxi. (Sborník přednášek postgraduálního kursu, Praha 1977.) Praha, pobočka ČSVTS na FEL ČVUT 1977, 30 pp.

[P12] M. Práger, K. Segeth: *Rychlé algoritmy pro řešení úloh matematické fyziky.* Software a algoritmy numerické matematiky 2. (Sborník referátů letní školy, Trojanovice 1977.) Praha, JČSMF 1978, 41–54.

[P13] V. Bezvoda, K. Segeth: *Mathematical modeling of electromagnetic fields.* The Use of Finite Element Method and Finite Difference Method in Geophysics. (Proceedings of Summer School, Liblice 1977.) Praha, Geofyzikální ústav ČSAV 1978, 329–332.

[P14] K. Segeth: *Approximation in the finite element method I.* The Use of Finite Element Method and Finite Difference Method in Geophysics. (Proceedings of Summer School, Liblice 1977.) Praha, Geofyzikální ústav ČSAV 1978, 61–79. MR0541735

[P15] K. Segeth: *Finite element method. An introductory course.* The Use of Finite Element Method and Finite Difference Method in Geophysics. (Proceedings of Summer School, Liblice 1977.) Praha, Geofyzikální ústav ČSAV 1978, 95–118. MR0541737

[P16] V. Bezvoda, J. Matouš, K. Segeth: *Lineární filtrace profilových měření metodou VDV v okolí Příbrami.* Hornická Příbram ve vědě a technice. (Sborník přednášek symposia, Příbram 1979.) Sekce užitá geofyzika. Příbram, Geofyzika, n.p., 1979, 187–204.

[P17] V. Bezvoda, V. Hašek, K. Segeth: *Objektivní metody zpracování geofyzikálních dat v archeologii.* Aplikace geofyzikálních metod v archeologii a moderní metody terénního výzkumu a dokumentace. (Sborník referátů 1. celostátní konference, Petrov n. Desnou 1979.) Brno, Geofyzika, n.p., 1979, 37–40.

[P18] V. Bezvoda, J. Dvořák, K. Segeth: *Poznatky z použití metody lineární filtrace při interpretaci geofyzikálních dat.* Sborník referátů 7. celostátní konference geofyziků. (Gottwaldov 1980.) Sekce S6 (Komplexní geofyzikální interpretace a syntézy). Brno, Geofyzika, n.p., 1980, 29–34.

[P19] K. Segeth: *Rychlé algoritmy pro řešení eliptických parciálních diferenciálních rovnic.* Software a algoritmy numerické matematiky 3. (Sborník referátů letní školy, Nové Město na Moravě 1979.) Praha, JČSMF 1980, 133–145.

[P20] K. Segeth: *Rychlé metody pro řešení řídkých soustav lineárních algebraických rovnic.* Sborník 4. semináře o metodě konečných prvků a variačních metodách. (Plzeň 1981.) Plzeň, Škoda, k.p., 1981, díl 2, 285–287.

[P21] V. Bezvoda, E. Jelínková, K. Segeth: *Výpočet přímé úlohy gravimetrie pomocí Fourierovy transformace.* Problémy současné gravimetrie. (Sborník referátů celostátního semináře, Zvíkovské Podhradí 1980.) Brno, Geofyzika, n.p., 1981, díl 2, 157–163.

[P22] V. Bezvoda, J. Dvořák, K. Segeth: *Použití matematického modelování při interpretaci anomálií VDV na Příbramsku.* Hornická Příbram ve vědě a technice. (Sborník přednášek symposia, Příbram 1982.) Sekce užitá geofyzika. Příbram, Geofyzika, n.p., 1982, 107–121.

[P23] K. Segeth: *Soubor programů pro řešení eliptických okrajových úloh rychlými metodami.* Programy a algoritmy numerické matematiky. (Sborník semináře, Alšovice 1983.) Praha, Matematický ústav ČSAV 1983, 19 pp.

[P24] K. Segeth: *Numerical experiments with the Stone incomplete triangular decomposition.* Mathematical Models in Physics and Chemistry and Numerical Methods of Their Realization. (Proceedings of Seminar, Visegrád 1982.) Teubner-Texte zur Mathematik 61. Leipzig, Teubner 1984, 226–236. Zbl 0548.65015, MR0790545

[P25] K. Segeth: *Stone incomplete factorization for the iterative solution of linear algebraic systems.* Software a algoritmy numerické matematiky 5. (Sborník referátů letní školy, Dlouhá Ves 1983.) Praha, JČSMF 1984, část 1, 163–181.

[P26] K. Segeth: *The iterative use of fast algorithms for the solution of elliptic partial differential equations.* Software a algoritmy numerické matematiky 5. (Sborník referátů letní školy, Dlouhá Ves 1983.) Praha, JČSMF 1984, část 2, 291–311.

[P27] K. Segeth: *Iterační použití rychlých algoritmů pro řešení soustav lineárních algebraických rovnic.* Programy a algoritmy numerické matematiky 2. (Sborník semináře, Alšovice 1984.) Praha, Matematický ústav ČSAV 1984, část 2, 16 pp.

[P28] K. Segeth: *Iterační řešení soustav lineárních algebraických rovnic pomocí přímých rychlých metod.* Matematické metody řešení fyzikálních problémů. (Sborník semináře, Cikháj 1985.) Brno, Přírodovědecká fakulta UJEP 1986, 64–72.

[P29] K. Segeth: *MUGTAPE 84 - programy pro metodu více sítí.* Programy a algoritmy numerické matematiky 3. (Sborník kursu, Alšovice 1986.) Praha, Matematický ústav ČSAV 1986, 7 pp.

[P30] K. Segeth: *Conjugate direction methods for the solution of sparse linear algebraic systems.* Software a algoritmy numerické matematiky 6. (Sborník referátů letní školy, Doksy 1985.) Praha, JČSMF 1986, 49–54.

[P31] V. Bezvoda, K. Segeth: *An application of cyclic reduction to the trial and error modeling of electromagnetic field.* Mathematical Methods for Solving Inverse Problems of Geophysical Fields. (Proceedings of Seminar, Smolenice 1986.) Bratislava, Geophysical Institute of the SAV 1987, 97–104.

[P32] P. Šilhán, M. Pospíšek, K. Segeth: *Univerzální programové prostředky pro řešení okrajových úloh a jejich užití při návrhu integrovaných obvodů.* Návrh obvodů počítačem. (Sborník přednášek semináře, Praha 1987.) Praha, Tesla VÚST A.S. Popova 1987, 240–249.

[P33] V. Bezvoda, J. Ježek, S. Saic, K. Segeth: *Employment of the PERICOLOR system in gravimetry.* Problémy současné gravimetrie. (Sborník referátů celostátního semináře, Liblice 1986.) Praha, Geofyzikální ústav ČSAV 1987, 175–183.

[P34] P. Šilhán, M. Pospíšek, K. Segeth: *Program pro dvourozměrnou analýzu funkce obecného elementu IO.* Návrh obvodů počítačem. (Sborník přednášek semináře, Praha 1988.) Praha, Tesla VÚST A.S. Popova 1988, 302–305.

[P35] V. Hašek, Z. Měřínský, J. Odstrčil, K. Segeth, R. Záhora: *Geophysical data processing systems in Czechoslovak archaeology.* Computer and Quantitative Methods in Archaeology. (Proceedings of Conference, Birmingham 1988.) Oxford, B.A.R. 1988, 195–200.

[P36] V. Bezvoda, J. Ježek, S. Saic, K. Segeth, Č. Tomek: *Employment of image processing systems in applied geophysics.* Proceedings of the 33rd International Geophysical Symposium. (Praha 1988.) Praha, Geofyzika, s.p., 1988, Vol. B(II), 171–179.

[P37] K. Segeth: *MLAT solutions to the nonlinear Poisson equation in semiconductor device models.* Proceedings of the 2nd International Symposium on Numerical Analysis. (Praha 1987.) Teubner Texte zur Mathematik 107. Leipzig, Teubner 1988, 273–276. Zbl 0675.65128, MR1171719

[P38] K. Segeth: *Víceúrovňové adaptivní techniky pro řešení okrajových úloh a jejich programová realizace.* Matematické metódy riešenia fyzikálnych problémov. (Zborník prednášok 2. celoštátneho seminára, Stará Turá 1988.) Bratislava, Matematicko-fyzikálna fakulta UK 1988, 61–68.

[P39] K. Segeth: *Dostupné programové vybavení pro víceúrovňové adaptivní techniky.* Software a algoritmy numerické matematiky 7. (Sborník referátů letní školy, Písek 1987.) Praha, JČSMF 1988, 57–72.

[P40] K. Segeth: *Soubor programů PLTMG pro adaptivní diskretizaci na více sítích.* Programy a algoritmy numerické matematiky 4. (Sborník kursu, Alšovice 1988.) Praha, Matematický ústav ČSAV 1988, 82–89.

[P41] V. Bezvoda, K. Segeth: *Efficient methods for solving 2-D magnetotelluric problems.* Inverse Modeling in Exploration Geophysics. (Proceedings of Seminar, Berlin (West) 1988.) Braunschweig, Vieweg 1989, 247–257.

[P42] V. Hašek, Z. Měřínský, K. Segeth: *New trends in processing and interpretation of geophysical data in Czechoslovak archaeology.* Communication in Archaeology. (Proceedings of the 2nd World Archaeological Congress, Barquisimeto 1990.) Winchester, IBM UK Scientific Centre 1990. Vol. 1, Data Visualization, 27–34, 117–118.

[P43] V. Bezvoda, J. Ježek, K. Segeth: *Číslicová filtrace dvojrozměrných dat.* Numerické metódy a ich aplikácie. (Zborník prednášok seminára, Modra 1990.) Bratislava, JSMF 1990, 147–151.

[P44] K. Segeth: *FREDPACK PC - soubor programů pro zpracování dvojrozměrných dat ve Fourierově oblasti.* Programy a algoritmy numerické matematiky 5. (Sborník kursu, Alšovice 1990.) Praha, Matematický ústav ČSAV 1990, 187–193.

[P45] V. Hašek, Z. Měřínský, K. Segeth: *New trends in processing and interpretation of geophysical data in Czechoslovak archaeology.* Science and Archaeology Vol. 32. Stoke-on-Trent, Research Centre for Computer Archaeology 1990, 39–42.

[P46] M. Pospíšek, K. Segeth, P. Šilhán: *Numerical modeling of semiconductor devices.* Colloquia Mathematica Societatis János Bolyai 59. Numerical Methods, Miskolc 1990. (Proceedings of Conference, Miskolc 1990.) Budapest, János Bolyai Mathematical Society 1991, 319–328. Zbl 0744.65097, MR1161241

[P47] K. Segeth: *Algorithms for linear filtering of two-dimensional data.* Software and Algorithms of Numerical Mathematics 9. (Proceedings of Summer School, Horní Údolí 1991.) Praha, JČMF 1991, 5-17.

[P48] V. Bezvoda, J. Hrabě, K. Segeth: *Fast algorithm for solving potential field problems.* Theory and Practice of Geophysical Data Inversion. (Proceedings of the 8th International Mathematical Geophysics Seminar on Model Optimization in Exploration Geophysics, Berlin (West) 1990.) Braunschweig, Vieweg 1991, 99–106.

[P49] V. Bezvoda, J. Hrabě, J. Ježek, K. Segeth: *3-D inverse gravity problem solved in the frequency domain.* Advances in Gravimetry. (Proceedings of Seminar, Smolenice 1990.) Bratislava, Geofyzikálny ústav SAV 1991, 105–108.

[P50] M. Práger, K. Segeth: *Aposteriorní odhady chyby řešení eliptických a parabolických úloh a adaptace sítě.* Programy a algoritmy numerické matematiky 6. (Sborník kursu, Bratříkov 1992.) Praha, Matematický ústav ČSAV 1992, 121–138.

[P51] V. Hašek, H. Petrová, K. Segeth: *Graphic representation methods in archaeological prospection in Czechoslovakia.* Computing the Past. (Proceedings of the Conference on Computer Applications and Quantitative Methods in Archaeology, Aarhus 1992.) Aarhus, Aarhus University Press 1992, 63–66.

[P52] K. Segeth: *A posteriori estimates and grid adjustment.* Numerical Mathematics in Theory and Practice. (Proceedings of Seminar, Plzeň 1993.) Plzeň, Západočeská univerzita 1993, 64–75.

[P53] K. Segeth: *A grading function algorithm for space grid adjustment in the method of lines.* Software and Algorithms of Numerical Mathematics 10. (Proceedings of Summer School, Cheb 1993.) Plzeň, University of West Bohemia 1993, 139–152.

[P54] K. Segeth: *ODEPACK - software pro řešení počátečních úloh pro soustavy obyčejných diferenciálních rovnic.* Programy a algoritmy numerické matematiky 7. (Sborník semináře, Bratříkov 1994.) Praha, Matematický ústav AV ČR 1994, 159–172.

[P55] K. Segeth: *Numerical experience with grid adjustment based on a posteriori error estimators.* The Mathematics of Finite Elements and Applications. (Proceedings of Conference, Uxbridge 1993.) Chichester, Wiley 1994, 391.

[P56] K. Segeth: *Numerical experiments with space grid adjustment in the finite element method of lines.* Numerické metódy a ich aplikácie. (Zborník prednášok seminára, Bratislava 1995.) Bratislava, JSMF 1995, 64–69.

[P57] K. Segeth: *Space grid adjustment for parabolic systems.* Proceedings of the 1st International Conference on Difference Equations. (San Antonio, 1994.) Luxembourg, Gordon and Breach Publishers 1995, 469–481. Zbl 0860.65088, MR1678699

[P58] K. Segeth: *Linear filtering of two-dimensional data in the frequency domain.* Proceedings of the European Symposium on Computing in Archaeology. (Saint Germain en Laye 1991.) Paris, CNRS 1995, 420–426.

[P59] K. Segeth: *Are adaptive space grids advantageous for the finite element method of lines?* Software and Algorithms of Numerical Mathematics 11. (Proceedings of Summer School, Železná Ruda 1995.) Plzeň, University of West Bohemia 1996, 226–240.

[P60] K. Segeth: *Je bilanční metoda lepší než metoda konečných prvků?* Programy a algoritmy numerické matematiky 8. (Sborník semináře, Janov n. Nisou 1996.) Praha, Matematický ústav AV ČR 1996, 184–193.

[P61] K. Segeth: *Grid adjustment employing a grading function.* Proceedings of the Prague Mathematical Conference. (Praha 1996.) Praha, Icaris 1996, 295–300. Zbl 0963.65526, MR1703496

[P62] K. Segeth: *Základy technologie řídkých matic.* Sborník MOSD 96. (Pernink 1996.) Plzeň, Západočeská univerzita 1996, 19–30.

[P63] P. Přikryl, R. Černý, V. Havlík, K. Segeth, P. Stupka, J. Toman: *Deposition of waste water into deep mines.* Proceedings of the 8th International Conference on Quantitative Methods for the Environmental Sciences. (Innsbruck 1997.) Innsbruck, Universität Innsbruck 1997, 64–65.

[P64] K. Segeth: *Newtonova metoda.* Sborník MOSD 97. (Pernink 1997.) Plzeň, Západočeská univerzita 1997, 31–40.

[P65] K. Segeth: *A posteriori error estimates for a nonlinear parabolic equation.* Equadiff 9 CD ROM. (Proceedings of Conference, Brno 1997.) Brno, Masaryk University 1998, Equadiff 9 Papers, 255–262.

[P66] J. Ježek, S. Saic, K. Segeth: *Numerické modelování pohybu rigidní částice ve viskózní kapalině.* Programy a algoritmy numerické matematiky 9. (Sborník semináře, Kořenov 1998.) Praha, Matematický ústav AV ČR 1998, 85–95.

[P67] J. Ježek, S. Saic, K. Segeth: *Numerical modelling of rigid objects in a ductile matrix.* Proceedings of the 4th Annual Conference of the International Association for Mathematical Geology. (Ischia 1998.) Houston, IAMG 1998, 821–826.

[P68] R. Černý, P. Jelínek, K. Segeth, P. Přikryl: *Matematické modelování transportních jevů v nádrži.* Sborník konference Orlice'98. (Králíky 1998.) Žamberk, Sdružení obcí a měst Orlice 1998, 62–69.

[P69] K. Segeth: *Spolehlivost numerických výpočtů.* Sborník MOSD 98. (Pernink 1998.) Plzeň, Západočeská univerzita 1998, 15–22.

[P70] K. Segeth: *ODE solvers as a tool in the adaptive method of lines.* Software and Algorithms of Numerical Mathematics 13. (Proceedings of Summer School, Nečtiny 1999.) Plzeň, University of West Bohemia 1999, 239–269.

[P71] K. Segeth: *A damped Newton iterative algorithm.* Lecture Notes of IMAMM 99. (Proceedings of Summer School, Nečtiny 1999.) Plzeň, University of West Bohemia 1999, 189–195.

[P72] K. Segeth: *Adaptive finite element method of lines for nonlinear parabolic equations.* Numerical Mathematics and Advanced Applications 3. (Proceedings of ENUMATH Conference, Jyväskylä 1999.) Singapore, World Scientific 2000, 707–714. Zbl 0969.65086

[P73] K. Segeth, P. Šolín: *Aplikace metody přímek na stlačitelné proudění (výpočetní aspekty).* Programy a algoritmy numerické matematiky 10. (Sborník semináře, Lázně Libverda 2000.) Praha, Matematický ústav AV ČR 2000, 162–173.

[P74] K. Segeth: *Matematické a numerické modelování polovodičové součástky.* Sborník MOSD 01. (Pernink 2001.) Plzeň, Západočeská univerzita 2001, 69–76.

[P75] P. Šolín, K. Segeth: *Some remarks on the method of lines applied to non-stationary compressible Euler equations.* ACOMEN 2002 CD ROM. (Proceedings of Conference, Liege 2002.) Liege, Université de Liege and Universiteit Ghent 2002, 10 pp.

[P76] P. Šolín, K. Segeth: *Non-uniqueness of solution to quasi-1D compressible Euler equations.* Equadiff 10 CD ROM. (Proceedings of Conference, Praha 2001.) Brno, Masaryk University Publishing House 2002, Equadiff 10 Papers, 379–389.

[P77] P. Přikryl, K. Segeth, R. Černý: *Computational modeling of pulsed laser induced phase change processes in II-VI semiconductors.* Heat Transfer 7. (Proceedings of the International Conference Advanced Computational Methods in Heat Transfer, Halkidiki 2002.) Southampton, WIT Press 2002, 205–214.

[P78] K. Segeth, P. Šolín, M. Kočiřík: *Some algorithmic aspects of higher-order finite element schemes in multidimensions.* Software and Algorithms of Numerical Mathematics 14. (Proceedings of Summer School, Kvilda 2001.) Plzeň, University of West Bohemia 2002, 199–221.

[P79] K. Segeth, P. Šolín: *Adaptive higher-order finite element solution of PDE's.* Programy a algoritmy numerické matematiky 11. (Sborník semináře, Dolní Maxov 2002.) Praha, Matematický ústav AV ČR 2002, 232–248.

[P80] K. Segeth: *Rothe method and method of lines. A brief discussion.* Mathematical and Computer Modelling in Science and Engineering. (Proceedings of International Conference, Prague 2003.) Prague, Czech Technical University 2003, 316–320.

[P81] K. Segeth: *On some fast algorithms.* Numerical Analysis. (Proceedings of Seminar, Ostrava 2003.) Ostrava, VŠB - Technical University 2003, 53-54.

[P82] M. Křížek, K. Segeth: *Metoda sdružených gradientů.* Sborník příspěvků 3. konference o matematice a fyzice na vysokých školách technických. (Brno 2003.) Brno, Vojenská akademie 2003, 9–13.

[P83] P. Šolín, K. Segeth, I. Doležel, M. Zítka: *Design of scalar and vector-valued hierarchic finite elements in 2D and 3D.* ADMOS 2003 CD ROM. (Proceedings of Conference, Göteborg 2003.) Barcelona, CIMNE 2003, 21 pp.

[P84] K. Segeth, P. Šolín: *Application of the method of lines to flow problems.* SIMONA 2003. (Proceedings of Workshop, Liberec 2003.) Liberec, Technical University 2003, 1–15.

[P85] M. Zítka, P. Šolín, K. Segeth: *PARSYS_2D - a higher-order FE solver for systems of nonlinear elliptic and parabolic equations.* ECCOMAS 2004 CD ROM Vol. 2. (Proceedings of Congress, Jyväskylä 2004.) Jyväskylä, University of Jyväskylä 2004, 15 pp.

[P86] K. Segeth, P. Šolín, M. Zítka: *Higher-order methods of lines and error estimates for 2D nonlinear parabolic problems.* Software and Algorithms of Numerical Mathematics 15. (Proceedings of Summer School, Hejnice 2003.) Plzeň, University of West Bohemia 2004, 109–117.

[P87] M. Zítka, K. Segeth, P. Šolín: *Higher-order FEM for a system of nonlinear parabolic PDE's in 2D with a-posteriori error estimates.* Numerical Mathematics and Advanced Applications 5. (Proceedings of ENUMATH Conference, Prague 2003.) Berlin, Springer 2004, 854–863. Zbl 1056.65091, MR2121431

[P88] P. Šolín, K. Segeth: *Three ways of interpolation on finite elements.* Programs and Algorithms of Numerical Mathematics 12. (Proceedings of Seminar, Dolní Maxov 2004.) Prague, Mathematical Institute of the AS CR 2004, 230–241.

[P89] P. Šolín, K. Segeth, I. Doležel: *Numerical quadrature for higherorder finite element methods.* Lecture Notes of IMAMM 03. (Proceedings of Summer School, Rožnov p. Radhošťem 2003.) Ostrava, VŠB - Technická univerzita 2005, 121–130.

[P90] K. Segeth, P. Šolín, I. Doležel: *Higher-order numerical quadrature in 2D and 3D.* Mezinárodní konference Prezentace matematiky ICPM'04. (Sborník konference, Liberec 2004.) Liberec, Technická univerzita v Liberci 2005, 203–210.

[P91] K. Segeth, P. Šolín, M. Zítka: *Singularities in electro- and magnetostatics, and their efficient resolution by hp-FEM.* Proceedings of the Seminar of Applied Mathematics. (Prague 2005.) Prague, Czech Technical University 2005, 155–172.

[P92] P. Šolín, K. Segeth: *Towards optimal shape functions for hierarchical Hermite elements.* Software and Algorithms of Numerical Mathematics 16. (Proceedings of Summer School, Srní 2005.) Plzeň, University of West Bohemia 2006, 236–244.

[P93] K. Segeth: *Conjugate gradient method.* Mezinárodní konference Prezentace matematiky ICPM'05. (Sborník konference, Liberec 2005.) Liberec, Technická univerzita v Liberci 2006, 335–341.

[P94] K. Segeth, P. Šolín: *Finite element approximation of boundary layers.* Simulation, Modelling, and Numerical Analysis. (Proceedings of the 3rd International Workshop SIMONA, Liberec 2006.) Liberec, Technical University of Liberec 2006, 131–140.

[P95] K. Segeth: *Shape functions for hierarchic Hermite elements.* International Conference Presentation of Mathematics '06. (Proceedings of Conference, Liberec 2006.) Liberec, Technical University of Liberec 2006, 95–102.

[P96] K. Segeth: *Jurij Vega and his connections to the Royal Bohemian Learned Society.* Baron Jurij Vega and His Times. (Proceedings of Conference Jurij Vega and His Time, Ljubljana 2004.) Ljubljana, DMFA 2006, 205–215.

[P97] K. Segeth, P. Šolín: *On some a posteriori error estimation results for the method of lines.* Programs and Algorithms of Numerical Mathematics 13. (Proceedings of Conference, Prague 2006.) Prague, Mathematical Institute of the AS CR 2006, 229–234.

[P98] K. Segeth: *Comparison of some FEM approximations of boundary layers.* International Conference Presentation of Mathematics '07. (Proceedings of Conference, Liberec 2007.) Liberec, Technical University of Liberec 2007, 103–110.

[P99] K. Segeth, P. Šolín: *A posteriori error estimates in the h-adaptive FEM of lines.* Seminar on Numerical Analysis SNA'08. (Proceedings of Seminar, Liberec 2008.) Liberec, Technical University of Liberec 2008, 106–109.

[P100] K. Segeth: *A posteriori error estimates for adaptive finite element methods.* International Conference Presentation of Mathematics '08. (Proceedings of Conference, Liberec 2008.) Liberec, Technical University of Liberec 2008, 73–80.

[P101] P. Šolín, K. Segeth, I. Doležel: *Space-time adaptive hp-FEM: Methodology overview.* Programs and Algorithms of Numerical Mathematics 14. (Proceedings of Seminar, Dolní Maxov 2008.) Prague, Institute of Mathematics of the AS CR 2008, 185–200. Zbl pre05802259, MR2522075

[P102] K. Segeth: *A comparison of some analytical and computational a posteriori error estimates in the FEM.* Simulace, modelování a nejrůznější aplikace. (Sborník semináře SIMONA 2009, Liberec 2009.) Liberec, Technická univerzita v Liberci 2009, 126–138.

[P103] K. Segeth: *Computational and analytical a posteriori error estimates for finite element methods.* International Conference Presentation of Mathematics '09. (Proceedings of Conference, Liberec 2009.) Liberec, Technical University of Liberec 2010, 93–100.

[P104] K. Segeth: *A comparison of some a posteriori error estimates for fourth order problems.* Programs and Algorithms of Numerical Mathematics 15. (Proceedings of Seminar, Dolní Maxov 2010.) Prague, Institute of Mathematics of the AS CR 2010, 164–170.

[P104] K. Segeth: *On a posteriori error estimates for biharmonic problems.* Seminar on Numerical Analysis SNA'11. (Proceedings of Seminar, Rožnov p. R. 2011.) Ostrava, Institute of Geonics of the AS CR 2011, 100–103.

[P105] K. Segeth: *On the advantages and drawbacks of a posteriori error estimation for fourth order elliptic problems.* Computational Analysis and Optimization CAO2011. (Proceedings of ECCOMAS Thematic Conference, Jyväskylä 2011.) Jyväskylä, University of Jyväskylä 2011, 160–164.

[P106] J. Ježek, K. Segeth: *Numerical experiments with smooth approximation.* Seminar on Numerical Analysis SNA'12. (Proceedings of Seminar, Liberec 2012.) Liberec, Technical University of Liberec 2012, 93–96.

[P107] K. Segeth: *Smooth approximation and its application to some 1D problems.* Applications of Mathematics 2012. (Proceedings of Conference, Prague 2012.) Prague, Institute of Mathematics of the AS CR 2012, 243–252.

[P108] K. Segeth: *Smooth approximation of data with applications to interpolating and smoothing.* Programs and Algorithms of Numerical Mathematics 16 (Proceedings of Seminar, Dolní Maxov 2012), Prague, Institute of Mathematics of the AS CR 2012, 181–186.

U. University texts

[U1] V. Bezvoda, B. Melichar, K. Segeth: *Matematické stroje.* Praha, katedra užité geofyziky přírodovědecké fakulty UK 1968, 102 pp.

[U2] V. Bezvoda, K. Segeth: *Programování a kybernetika pro posluchače geologie a geografie I.* Praha, SPN 1981, 187 pp.

[U3] V. Bezvoda, K. Segeth: *Programování a kybernetika pro posluchače geologie a geografie II.* Praha, SPN 1985, 268 pp.

[U4] V. Bezvoda, J. Ježek, S. Saic, K. Segeth: *Dvojrozměrná diskrétní Fourierova transformace a její použití I. Teorie a obecné užití.* Praha, SPN 1988, 181 pp.

[U5] K. Segeth: *Numerický software I.* Praha, Karolinum 1998, 99 pp.

[U6] M. Křížek, K. Segeth: *Numerické modelování problémů elektrotechniky.* Praha, Karolinum 2001, 198 pp.

[U7] K. Segeth: *Numerické metody algebry.* Praha, Česká technika-nakladatelství ČVUT 2011, 117 pp.

V. Research reports

[V1] K. Segeth: *Problems of universal approximation by hill functions.* Tech. Note BN-619. College Park, MD, Institute for Fluid Dynamics and Applied Mathematics, University of Maryland 1970, 24 pp.

[V2] I. Babuška, J. Segethová, K. Segeth: *Numerical experiments with finite element method I.* Tech. Note BN-669. College Park, MD, Institute for Fluid Dynamics and Applied Mathematics, University of Maryland 1970, 17 pp.

[V3] K. Segeth: *Programy pro řešení parciálních diferenciálních rovnic.* Informační řada B, č. 18. Praha, sektor výpočetních systémů ÚTIA ČSAV 1979, 3 pp.

[V4] V. Červ, K. Segeth: *Numerical experiments for a comparison of the accuracy of the finite-difference solution to elliptic boundary-value problems obtained by direct and iterative methods.* Praha, Matematický ústav ČSAV 1980, 20 pp.

[V5] M. Práger, P. Přikryl, K. Segeth: *Numerický software pro řešení diferenciálních rovnic.* Praha, Matematický ústav ČSAV 1980, 27 pp.

[V6] P. Přikryl, K. Segeth, E. Vitásek: *Matematické modelování fyzikálních procesů se změnou fáze.* Praha, Matematický ústav ČSAV 1980, 33 pp.

[V7] V. Bezvoda, J. Ježek, S. Saic, K. Segeth: *FOURFIVE 83. Soubor programů pro Fourierovu transformaci, filtraci a vizualizaci dvojrozměrných dat.* Praha, Přírodovědecká fakulta UK 1983, 19 pp.

[V8] K. Segeth: *Rychlé metody pro řešení soustav lineárních algebraických rovnic. Doprovodný text k přednáškám 1. semestru kursu numerické matematiky.* Praha, pobočka ČSVTS při SVÚSS 1984, 74 pp.

[V9] J. Havel, M. Vošvrda, J. Michálek, I. Bajla, K. Segeth, M. Vajteršic: *Zpracování vícerozměrných signálů. Prognostické téma.* Praha, ÚTIA ČSAV 1985, 41 pp.

[V10] V. Bezvoda, J. Ježek, S. Saic, K. Segeth: *FOURFIVE 85. Soubor programů pro Fourierovu transformaci, filtraci a vizualizaci dvojrozměrných dat.* Praha, Přírodovědecká fakulta UK 1985, 24 pp.

[V11] M. Práger, M. Křížek, P. Přikryl, K. Segeth: *Zpráva o části dílčího úkolu SPZV III-9-1/2 řešené v MÚ ČSAV v letech 1981-1985.* Praha, Matematický ústav ČSAV 1985, 21 pp.

[V12] K. Segeth: *Programy pro metodu více sítí - MUGTAPE 84. Část 1, 2, 3, 4.* Dílčí výzkumná zpráva 1880 02 752/2 až 5. Praha, Tesla VÚST A. S. Popova 1987, 15, 18, 26, 24 pp.

[V13] K. Segeth: *STONE. Podprogram pro řešení řídké soustavy lineárních algebraických rovnic speciálního tvaru Stoneovou iterační metodou.* Praha, Matematický ústav ČSAV 1987, 11 pp.

[V14] J. Segethová, K. Segeth: *Soudobé efektivní numerické metody řešení lineárních algebraických rovnic. (Text přednášek Kursu numerické matematiky I.)* Praha, Dům techniky ČSVTS 1988, 104 pp.

[V15] V. Červ, J. Pek, K. Segeth: *Modelling of magnetotelluric field in 2D anisotropic media.* Praha, Mathematical Institute of the Academy of Sciences 1996, 10 pp.

[V16] K. Kronrádová, J. Hořejš, L. Kalinová, E. Kindler, I. Kronrád, O. Kronrád, I. Marek, E. Nováková, K. Segeth, K. Zimmermann, M. Žemlička: *Modelování sítě škol a jeho regionální a sociální aspekty.* Zpráva o řešení projektu RS 96108. Praha, Matematicko-fyzikální fakulta UK 1996, 107 pp.

[V17] R. Černý, V. Havlík, J. Toman, P. Přikryl, K. Segeth: *Likvidace nadbilančních vod z odkaliště Bytíz uložením do podzemí ložiska Příbram.* 1. etapa. Výzkumná zpráva. Praha, Stavební fakulta ČVUT 1997, 45 pp.

R. Preprints

[R1] K. Segeth: *Universal approximation by systems of hill functions.* Praha, Matematický ústav ČSAV 1974, 55 pp.

[R2] V. Bezvoda, K. Segeth: *A contribution to the theory of electromagnetic induction of a line source.* Praha, Matematický ústav ČSAV 1975, 23 pp.

[R3] K. Segeth: *Roundoff errors in the fast computation of discrete convolutions.* Praha, Matematický ústav ČSAV 1979, 34 pp.

[R4] V. Bezvoda, K. Segeth: *Přednášky z programování v jazyce Fortran.* Praha, FV SSM přírodovědecké fakulty UK 1980, 94 pp.

[R5] K. Segeth: *On the choice of iteration parameters in the Stone incomplete factorization.* Praha, Matematický ústav ČSAV 1983, 21 pp.

[R6] K. Segeth: *The iterative use of fast algorithms for the solution of elliptic partial differential equations.* Praha, Matematický ústav ČSAV 1983, 22 pp.

[R7] K. Segeth: *A posteriori error estimates for parabolic differential systems solved by the finite element method of lines.* Praha, Mathematical Institute of the Academy of Sciences 1993, 29 pp.

[R8] K. Segeth: *A posteriori error estimation with the finite element method of lines for a nonlinear parabolic equation in one space dimension.* Praha, Mathematical Institute of the Academy of Sciences 1998, 23 pp.

D. Dissertations

[D1] K. Segeth: *Formulace okrajových podmínek při užití metody sítí.* Diplomní práce. Praha, Matematicko-fyzikální fakulta UK 1964, 87 pp.

[D2] K. Segeth: *On universally optimal quadrature formulae involving values of derivatives of integrand.* Kandidátská disertace. Praha, Matematický ústav ČSAV, 1968, 80 pp.

[D3] K. Segeth: *A posteriori error estimates and grid adjustment in the finite element method of lines.* Habilitační práce. Praha, Matematicko-fyzikální fakulta UK 1995, 66 pp.

T. Translations

[T1] G. E. Forsythe: *Co je uspokojivý program pro řešení kvadratické rovnice?* Program výchovy informatiků na vysoké škole. Co dělat než přijde informatik. (Učební texty postgraduálního kursu POKAM.) Praha, Matematicko-fyzikální fakulta UK 1968. (Jointly with P. Přikryl.)

[T2] J. Nečas, I. Hlaváček: *Mathematical theory of elastic and elasto-plastic bodies: an introduction.* Amsterdam - Oxford - New York, Elsevier 1981.

[T3] A. A. Samarskij, J. S. Nikolajev: *Numerické řešení velkých řídkých soustav.* Praha, Academia 1984. (Jointly with P. Přikryl.)

[T4] M. Fiedler: *Special matrices and their applications in numerical mathematics.* Dordrecht, M. Nijhoff 1986. (Jointly with P. Přikryl.)

[T5] G. I. Marčuk: *Metody numerické matematiky.* Praha, Academia 1987. (Jointly with P. Přikryl.) MR0931536

[T6] D. G. Saari, Z. Xia: *Do nekonečna v konečném čase.* Pokroky mat. fyz. astronom. 42 (1997), 90–102. (Jointly with M. Křížek.)

[T7] C. Pomerance: *Vyprávění o dvou sítech.* Pokroky mat. fyz. astronom. 43 (1998), 9–29. (Jointly with J. Chleboun and M. Křížek.)

[T8] R. D. Mauldin: *Zobecnění Velké Fermatovy věty: Bealova domněnka a problém o cenu.* Pokroky mat. fyz. astronom. 43 (1998), 104–107. (Jointly with M. Křížek.)

[T9] T. C. Hales: *Dělové koule a včelí plásty.* Pokroky mat. fyz. astronom. 46 (2001), 101–118. (Jointly with J. Chleboun and M. Křížek.)

[T10] J. Stillwell: *Příběh stodvacetistěnu v* $\mathbb{R}^4$. Pokroky mat. fyz. astronom. 46 (2001), 265–280. (Jointly with M. Křížek and I. Saxl.)

Contents

Conference Applications of Mathematics 2013
in honor of the 70th birthday of Karel Segeth.
Jan Brandts, Sergey Korotov, Michal Křížek,
Jakub Šístek, and Tomáš Vejchodský (Eds.),
Institute of Mathematics AS CR, Prague 2013

COUNTING TRIANGLES THAT SHARE THEIR VERTICES WITH THE UNIT N-CUBE

Jan Brandts, Apo Cihangir

Korteweg-de Vries Institute, University of Amsterdam
Science Park 904, 1098 XH Amsterdam, The Netherlands
janbrandts@gmail.com

Abstract

This paper is about 0/1-triangles, which are the simplest nontrivial examples of 0/1-polytopes: convex hulls of a subset of vertices of the unit n-cube I^n. We consider the subclasses of right 0/1-triangles, and acute 0/1-triangles, which only have acute angles. They can be explicitly counted and enumerated, also modulo the symmetries of I^n.

1. Introduction

A 0/1-polytope [3] is the convex hull of a subset of the set of vertices $\mathbb{B}^n$ of the unit n-cube I^n. Since I^n has 2^n vertices, the number of subsets of $\mathbb{B}^n$ equals $2^{(2^n)}$, a number that grows so quickly that the practical study of 0/1-polytopes is a complicated matter. Therefore, it is convenient to consider two 0/1-polytopes as equivalent if there exists an n-cube symmetry that maps one onto the other. The group H_n of symmetries of I^n is called the hyperoctahedral group. It is generated by the reflections in the n hyperplanes that orthogonally intersect the coordinate axes at their midpoints, and the transposition of labels of coordinate axes. The number of elements of H_n, its order, is $n!2^n$. An orbit of a 0/1-polytope under the action of H_n, or in other words, the set of images of the polytope under each of the cube's symmetries, can therefore contain at most $n!2^n$ elements. Under the proposed equivalence, the number of equivalence classes of 0/1-polytopes can, in principle, be counted using Pólya's Enumeration Theorem. This requires the explicit computation of the so-called cycle index of H_n. In [2], it is described how to compute this cycle index, but the procedure is nontrivial and does not lead to a general formula in n. Also, it does not distinguish between 0/1-polytopes whose dimension equals n, and the ones that are less-dimensional. This explains why only for $n \leq 6$ it is known how many equivalence classes of n-dimensional 0/1-polytopes exist.

1.1. Goal and outline of this paper

In this paper, we will fully characterize the 0/1-polytopes that are the convex hull of three different vertices of I^n, the 0/1-triangles. Next to individual vertices

and line segments, these are the simplest 0/1-polytopes. We will count the number of 0/1-triangles in I^n, and also the number of elements in the disjoint subsets of right and acute 0/1-triangles. This will be done in Section 2. In Section 3 we will count the number of 0/1-equivalence classes of such triangles. We will also enumerate them, by which we mean that we list from each equivalence class a unique member.

2. Counting and enumerating all 0/1-triangles in I^n

Let $n \geq 2$. A 0/1-triangle is the convex hull of three distinct vertices of the unit n-cube I^n. We will write Δ_n for the set of 0/1-triangles in I^n. A first observation is that no three vertices of I^n lie on the same line, and thus that each $\mathcal{T} \in \Delta_n$ is nondegenerate. A second observation is that each $\mathcal{T} \in \Delta_n$ is nonobtuse, by which we mean that all its angles are less than or equal to 90°. This is true because the inner product between two vectors $u, v \in \mathbb{B}^n$, the set of 0/1-vectors of length n representing the vertices of I^n, is nonnegative. Thus, any angle between two edges that meet at the origin is nonobtuse. By symmetry, this also holds for angles located at other vertices of I^n. This leads to the following proposition.

Proposition 2.1 *The number $|\Delta_n|$ of elements of the set Δ_n of 0/1-triangles in I^n equals*

$$|\Delta_n| = \binom{2^n}{3} = \frac{1}{6} 2^n \left(2^n - 1\right) \left(2^n - 2\right). \tag{1}$$

and each $\mathcal{T} \in \Delta_n$ is nondegenerate, and moreover nonobtuse.

We can divide the triangles in Δ_n into two subsets, the subset R_n of right triangles, and the subset A_n of acute triangles, which are the triangles that have three acute angles,

$$\Delta_n = A_n \cup R_n \quad \text{and} \quad A_n \cap R_n = \emptyset,$$

with as immediate consequence that

$$|\Delta_n| = |A_n| + |R_n|. \tag{2}$$

It is possible to count the number $|R_n|$ of right triangles, and thus to count $|A_n|$ as well.

Theorem 2.2 *The number $|R_n|$ of right 0/1-triangles in I^n equals*

$$|R_n| = 2^{n-1} \left(3^n - 2^{n+1} + 1\right). \tag{3}$$

Proof. We will first count the right triangles $\mathcal{T} \in R_n$ that have their right angle at the origin. The other two vertices $u, v \in \mathbb{B}^n$ of such a $\mathcal{T}$ are nonzero and orthogonal. If u has $k < n$ zero entries, there are $2^k - 1$ different $v \neq 0$ such that $u \perp v$. The number of $u \in \mathbb{B}^n$ with k zero entries is $\binom{n}{k}$, leading to a total of

$$\frac{1}{2} \sum_{k=1}^{n-1} \binom{n}{k} (2^k - 1)$$

right triangles with right angle at the origin, where the factor of a half is due to the fact that the roles of u and v can be interchanged. As a consequence,

$$|R_n| = 2^n \cdot \frac{1}{2} \sum_{k=1}^{n-1} \binom{n}{k} (2^k - 1), \tag{4}$$

because the right angle can be located at any of the 2^n vertices of I^n. Using the binomial formula

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k},$$

with $x = 2$ and $y = 1$, and also with $x = y = 1$, the expression in (4) can be easily simplified until (3) remains. This proves the theorem. □

Corollary 2.3 *The number $|A_n|$ of acute 0/1-triangles in I^n equals*

$$|A_n| = \frac{1}{6} 2^n \left(4^n - 3^{n+1} + 3 \cdot 2^n - 1\right). \tag{5}$$

Proof. Substitute expressions (1) and (3) into (2) and rearrange some terms. □

The following table gives the values of $|\Delta_n|, |R_n|$ and $|A_n|$ for small values of n. The asymptotic behavior of $|R_n| = \mathcal{O}(n^6)$ and $|A_n| = \mathcal{O}(n^8)$ is clearly visible.

n	$\lvert R_n\rvert$	$\lvert A_n\rvert$	$\lvert\Delta_n\rvert$
2	4	0	4
3	48	8	56
4	400	160	560
5	2880	2080	4960
6	19264	22400	41664
7	123648	217728	341376
8	774400	1989120	2763520
9	4776960	17461760	22238720
10	29185024	149248000	178433024

Neither $|R_n|$ nor $|A_n|$ is mentioned in the Online Encyclopedia of Integer Sequences (OEIS). But the scaled sequence $|R_n|/2^{n-1}$ can be found under label A028243 and has annotation *essentially Stirling numbers of second kind*, whereas $|A_n|/2^n$ has label A000453, *Stirling numbers of the second kind, $S(n, 4)$.*

3. Counting and enumerating modulo cube symmetries

In the previous section we described how to generate and count right and acute 0/1-triangles. We did not take into account 0/1-equivalence, as described in Section 1. This will be done here. We will count the number of 0/1-equivalence classes of right and acute 0/1-triangles, and explicitly give one representative for each equivalence class.

3.1. Matrix representation and 0/1-equivalence

Apart from the empty set, we will represent a 0/1-polytope $\mathcal{P} \subset I^n$ by a 0/1-matrix P of size $n \times p$ whose columns are the p coordinate vectors in $\mathbb{B}^n$ of its vertices. Since we do not allow multiple vertices, this can be done in exactly $p!$ different ways. If we assign to each $n \times p$ 0/1-matrix U an integer vector

$$\nu(U) = \mathcal{S}(v_n^\top U), \quad \text{where} \quad v_n^\top = \left(1, 2, 4, \ldots, 2^{n-1}\right), \tag{6}$$

and where $\mathcal{S}$ sorts the integer vector in its argument in increasing order, we see that each matrix representation P of $\mathcal{P}$ has the same vector value $\nu(P)$. Moreover, if the 0/1 polytopes $\mathcal{P}_1$ and $\mathcal{P}_2$ are distinct subsets of I^n, then their vertex sets are distinct [3], and hence for given matrix representations P_1 of $\mathcal{P}_1$ and P_2 of $\mathcal{P}_2$ we have that $\nu(P_1) \neq \nu(P_2)$. Therefore, with a slight abuse of notation, we will also consider ν as an injective map on the set of all nonempty 0/1-polytopes into the set consisting of all vectors up to length $2^n - 1$.

Let P_1 be a matrix representing a 0/1-polytope $\mathcal{P}_1$. Then $\mathcal{P}_2$ is a 0/1-polytope that is 0/1-equivalent to $\mathcal{P}_1$ if and only if $\mathcal{P}_2$ has a matrix representation P_2 that can be transformed into P_1 by permuting and negating some rows of P_2. A row negation is to replace the zeros by ones, and the ones by zeros within a row. The negation of row j corresponds to the reflection of I^n into the hyperplane with equation $2x_j = 1$, whereas the exchange of rows i and j corresponds to the relabeling of coordinate axes i and j.

Definition 3.1 The minimal representative within the 0/1-equivalence class $\mathcal{E}(\mathcal{P})$ of a given 0/1-polytope $\mathcal{P}$ is the unique element $\mathcal{P}^* \in \mathcal{E}(\mathcal{P})$ for which $\nu(\mathcal{P}^*)$ is lexicographically smaller than $\nu(\mathcal{P})$ for all $\mathcal{P} \in \mathcal{E}(\mathcal{P}), \mathcal{P} \neq \mathcal{P}^*$. The minimal matrix representation P^* of the equivalence class $\mathcal{E}(\mathcal{P})$ is the matrix representation P^* for $\mathcal{P}^*$ for which $v^\top P^*$ is increasing.

In the following section we will see some examples of matrix representations and of geometrical invariants under cube symmetries.

3.2. Congruence versus 0/1-equivalence

If two 0/1-polytopes $\mathcal{P}_1$ and $\mathcal{P}_2$ are 0/1-equivalent, $\mathcal{P}_1$ can be transformed into $\mathcal{P}_2$ by a cube symmetry, which is a congruence. Conversely, it is well known that congruent 0/1-polytopes need not be 0/1-equivalent. An example, adapted from [3], is given by the two full-dimensional 5-simplices in I^5 represented by the matrices

$$P_1 = \begin{bmatrix} 0 & 1 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \quad \text{and} \quad P_2 = \begin{bmatrix} 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

If for $j \in \{1, 2\}$ we write R_j for P_j with its first column removed, then $R_1^\top R_1 = R_2^\top R_2$ and since R_2 is invertible, $(R_1 R_2^{-1})^\top (R_1 R_2^{-1}) = I$. Thus $Q = R_1 R_2^{-1}$ is orthogonal,

and we conclude that $R_1 = QR_2$ and hence $P_1 = QP_2$, proving the congruence. To disprove 0/1-equivalence, consider the effect of cube symmetries on the vector of row sums of a matrix. Row permutations do not alter the values, only permute them, whereas a row negation replaces a row sum s by $p - s$, where p is the number of columns. Since P_2 has two row sums equal to 1, whereas P_1 has only one row sum equal to one and no row sum equal to $6 - 1 = 5$, we see that $\mathcal{P}_1 \notin \mathcal{E}(\mathcal{P}_2)$. Geometrically speaking, P_2 has two exterior facets, which are facets that lie in a facet of I^5, and P_1 has only one. Obviously, cube symmetries preserve such exterior facets.

In spite of the above, it is known that equivalence does indeed hold for all full dimensional 0/1-polytopes of dimension $n \leq 4$. The full dimensionality cannot be omitted, as is shown by the following counter example,

$$P_1 = \begin{bmatrix} 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \quad \text{and} \quad P_2 = \begin{bmatrix} 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Both matrices represent a regular tetrahedron in I^4, but the tetrahedron at the left has a zero row, and hence lies in a cube facet (basically in I^3), whereas the right one has neither a zero row nor a row with ones. Thus, they are not 0/1-equivalent. This also shows that for 0/1-tetrahedra in I^n, congruence and 0/1-equivalence are not the same. For 0/1-triangles however, they are.

Theorem 3.2 *If 0/1-triangles $\mathcal{T}_1$ and $\mathcal{T}_2$ are congruent, then they are 0/1-equivalent.*

Proof. Let $\mathcal{T}_1, \mathcal{T}_2 \in \Delta_n$ be congruent. Then their edge lengths and angles are equal. Therefore, it is possible to apply a cube symmetry S_1 to $\mathcal{T}_1$ such that the origin is a vertex of $S_1(\mathcal{T}_1)$, while its remaining vertices are $v_1, w_1 \in \mathbb{B}^n$, and then to apply a cube symmetry S_2 to $\mathcal{T}_2$ such that the origin is a vertex of $S_2(\mathcal{T}_2)$ and its remaining vertices are $v_2, w_2 \in \mathbb{B}^n$, such that

$$\|v_1\| = \|v_2\| = \sqrt{p}, \quad \|w_1\| = \|w_2\| = \sqrt{q}, \quad \text{and} \quad v_1^\top w_1 = v_2^\top w_2 = r, \tag{7}$$

for certain integers p, q, r. Due to (7), the $3 \times n$ matrices $P_1 = (0|v_1|w_1)$ representing $\mathcal{T}_1$ and $P_2 = (0|v_2|w_2)$ representing $\mathcal{T}_2$, both have r rows equal to $(0, 1, 1)$, and consequently, $p - r$ rows equal to $(0, 1, 0)$ and $q - r$ rows equal to $(0, 0, 1)$. And since P_1 and P_2 have the same rows, $\mathcal{T}_1$ and $\mathcal{T}_2$ are 0/1-equivalent. □

3.3. The minimal matrix representation for each 0/1-equivalence class

We will now formulate necessary and sufficient conditions under which a matrix is a minimal matrix representation of an equivalence class $\mathcal{E}(\mathcal{T})$. The necessity of the block form of the matrix in 8 was already described in [1] in a more general context.

Theorem 3.3 *An $n \times 3$ matrix P^* is a minimal matrix representation of an equivalence class $\mathcal{E}(\mathcal{T})$ of 0/1-triangles in I^n if and only if*

$$P^* = \left[\begin{array}{ccc} 0 & 1 & 1 \\ \vdots & \vdots & \vdots \\ 0 & 1 & 1 \\ \hline 0 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 0 & 1 & 0 \\ \hline 0 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 0 & 0 & 1 \\ \hline 0 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 0 & 0 & 0 \end{array}\right], \quad (P^*)^\top P^* = \begin{bmatrix} 0 & 0 & 0 \\ 0 & p & r \\ 0 & r & q \end{bmatrix}, \quad 1 \le p \le q \le p-r+q-r \le n-r \le n. \tag{8}$$

Note that p, q and $p - r + q - r$ are the squares of the lengths of the edges of the triangle.

Proof. Suppose that P^* is a minimal matrix representation. Then the block form given in (8) is necessary for the following reasons. Firstly, any vertex of any triangle can be mapped onto the origin by a cube symmetry, hence the first column of P^* needs to be zero. Secondly, if there is a zero entry in the second column in row i and an entry equal to one in row j with $j > i$, interchanging rows i and j would decrease the second value in $v_n^\top P^*$ while the first value of $v_n^t P^*$ remains zero, contradicting the minimality. Further, if in the third column there is a zero entry in row i with $i < p$ and an entry equal to one in row j with $i < j \le p$, then interchanging rows i and j would decrease the third value in $v_n^t P$ while the first value of $v_n^t P$ remains zero, and the second also remains the same because in the second column, two entries equal to one are swapped, contradicting the minimality. Finally, if in the third column there is a zero entry in row i with $i \ge p$ and an entry equal to one in row j with $i < j$, then interchanging rows i and j would decrease the third value in $v_n^\top P$ while the first value of $v_n^\top P$ remains zero, and the second also remains the same because in the second column, two entries equal to zero are swapped. This shows the necessity of the block form in (8).

Additionally, the set of inequalities $1 \le p \le q \le p-r+q-r \le n-r$ is necessary for the following reasons. Firstly, the second column of P needs to be nonzero, hence $1 \le p$. Secondly, $p \le q$ or otherwise swapping the block with rows (0 1 0) with the block with rows (0 0 1) followed by swapping the second and third column, would result in a zero first column, and a second column with $q < p$ entries equal to one, and this would reduce the second value of $v_n^\top P$ while the first remains zero. Thirdly, $q \le p - r + q - r$, or equivalently $r \le p - r$ or otherwise negating all rows that have

a one in the second column, followed by interchanging the first and second column, followed by restoring the block form by interchanging the block with rows (0 1 1) with the block with rows (0 1 0), would result in a matrix with zero first and second column and also the third and fourth block of rows unchanged. However, there would be $p-r$ ones at the top of the third column instead of r, and if $r > p-r$, this would reduce the third value of $v_n^\top P^*$ while the first and second remain unchanged, contradicting the minimality. The next inequality, equivalent to $p+q-r \leq n$, is necessary because $p+q-r$ is the number of nonzero rows of P^*, which must, of course, be bounded by n. Finally, the rightmost inequality is necessary because the other ones do not yet guarantee that r is nonnegative.

Now we prove that the given conditions in (8) are sufficient. Firstly, since the first entry of $v_n^\top P^*$ equals zero, this value cannot be reduced. Secondly, since the triangle has no edge with length less than $\sqrt{p}$, also the second entry of $v_n^\top P^*$ cannot be reduced. The third column of P^* represents one of the two remaining edges of the triangle. The third entry of $v_n^\top P^*$ is minimal for the edge whose inner product with the second column of p is maximal, because this minimizes the number of rows equal to (0 0 1). This follows from the requirement $r \leq p-r$. □

As a consequence, we can directly characterize the equivalence classes of right triangles in I^n.

Corollary 3.4 *An $n \times 3$ matrix P^* is a minimal matrix representation of an equivalence class $\mathcal{E}(\mathcal{T})$ of right 0/1-triangles in I^n if and only if (8) holds with $r = 0$.*

Proof. If (8) holds with $r = 0$, the matrix P^* in (8) obviously represents a right triangle, and due to Theorem 3.2, this representation is minimal. Conversely, suppose that P^* is a minimal representation of a right triangle. Then (8) holds due to Theorem 3.2. We will prove that additionally, $r = 0$. Writing $P^* = (0\ u\ v)$ with $u, v \in \mathbb{B}^n$, we that either $u \perp v$ or $u - v \perp u$ or $u - v \perp v$. The second of these options, $u - v \perp u$, implies that P^* has no rows equal to (0 1 0), or in other words, that $p = r$. But due to the inequality $q \leq p - r + q - r$ from (8), this implies that $r = 0$. Consequently, also $p = r = 0$, contradicting $p \geq 1$. The third option $u - v \perp v$ similarly implies that P^* has no rows equal to (0 0 1), hence $q = r$, hence the inequality $p \leq p - r + q - r$ from (8) implies that $r = 0$. Therefore $q = r = 0$, contradicting $q \geq 1$. The only option left is $u \perp v$, which indeed implies $r = 0$. □

Corollary 3.5 *An $n \times 3$ matrix P^* is a minimal matrix representation of an equivalence class $\mathcal{E}(\mathcal{T})$ of acute 0/1-triangles in I^n if and only if (8) holds with $r > 0$.*

Proof. Follows immediately from Theorem 3.2 and Corollary 3.4. □

3.4. Counting the 0/1-equivalence classes of right and acute 0/1-triangles

In order to count the number of equivalence classes of 0/1-triangles in I^n, by Theorem 3.3 we only have to count the number of triples (p, q, r) such that

$$1 \leq p \leq q \leq p + q - 2r \leq n - r \leq n. \tag{9}$$

We will do this by fixing a value for r and counting the tupels (p, q) that satisfy the resulting equation. The following lemmas will be of use.

Lemma 3.6 *Let $m \geq 1$ be an integer. The number of integer tupels (a, b) satisfying*

$$1 \leq a \leq b \leq m - a \tag{10}$$

equals

$$\left\lfloor \frac{m}{2} \right\rfloor \left\lceil \frac{m}{2} \right\rceil, \tag{11}$$

where $\lfloor \cdot \rfloor$ is the floor-operator and $\lceil \cdot \rceil$ the ceil-operator.

Proof. Only for values of a with $1 \leq a \leq \lfloor m/2 \rfloor$, we have that $a \leq m - a$. The number of integers between such an a and $m - a$ equals $m + 1 - 2a$. This leads to a total of

$$\sum_{a=1}^{\lfloor m/2 \rfloor} m + 1 - 2a = \left\lfloor \frac{m}{2} \right\rfloor (m+1) - 2 \cdot \frac{1}{2} \left\lfloor \frac{m}{2} \right\rfloor \left(\left\lfloor \frac{m}{2} \right\rfloor + 1 \right) \tag{12}$$

tupels (a, b) that satisfy (10). Using the relation

$$m = \left\lceil \frac{m}{2} \right\rceil + \left\lfloor \frac{m}{2} \right\rfloor, \tag{13}$$

together with Lemma 3.10, this leads, after some simplifications, to the statement. □

Corollary 3.7 *The number of $0/1$-equivalence classes of right triangles in I^n equals*

$$\left\lfloor \frac{n}{2} \right\rfloor \left\lceil \frac{n}{2} \right\rceil. \tag{14}$$

Proof. According to Corollary 3.4, we need to count to number of tupels (p, q) satisfying

$$1 \leq p \leq q \leq p + q \leq n. \tag{15}$$

Since the inequality $q \leq p + q$ is always valid, it can be removed. Thus, we only need to count the number of tupels (p, q) such that $1 \leq p \leq q \leq n - p$, which was done in Lemma 3.6. □

In the next lemma we will count equivalence classes of triangles for fixed values of $r \geq 1$. It will turn out that if $3r > n$, there are no solutions. Moreover, substituting $r = 0$ in (16) below does not yield the result of Corollary 3.7. After its proof it is explained why not.

Lemma 3.8 *For given $r \geq 1$ with $3r \leq n$, the number of tupels (p, q) satisfying (9) equals*

$$\left\lfloor \frac{n - 3r + 2}{2} \right\rfloor \left\lceil \frac{n - 3r + 2}{2} \right\rceil. \tag{16}$$

Proof. Let $r \geq 1$ be fixed. If $p < 2r$, there are no integers q that satisfy the third inequality $q \leq p + q - 2r$ in (9). If $p \geq 2r$, this inequality holds for all q and can thus be removed. Thus, we only need to count the tupels (p, q) for which

$$2r \leq p \leq q \leq n + r - p. \tag{17}$$

For such tupels to exist, we need that $2r \leq n+r-p$, but since $p \geq 2r$ this translates into $2r \leq n + r - 2r$. This explains the requirement $3r \leq n$ in the statement of this lemma. To count the tupels, subtract $2r - 1$ from each term in (17), and define $a = p - (2r - 1), b = q - (2r - 1)$, and $m = n - 3r + 2$, then

$$1 \leq a \leq b \leq n + r - a - 2(2r - 1) = n - 3r + 2 - a = m - a. \tag{18}$$

Applying Lemma 3.6 gives the number of tupels (a, b) satisfying these inequalities in terms of m, and substituting back $m = n - 3r + 2$ proves the statement. □

Remark 3.9 Choosing $r = 0$ in (16) does not give (14). This is because setting $r = 0$ in (17) does not imply $1 \leq p$, as is required in Theorem 3.3, whereas for $r \geq 1$, it does.

We will now count the number of equivalence classes of acute triangles. First another lemma.

Lemma 3.10 *For nonnegative integers k we have that $(k \bmod 2)^2 = k \bmod 2$, and hence*

$$\left\lfloor \frac{k}{2} \right\rfloor \left\lceil \frac{k}{2} \right\rceil = \left(\frac{k - k \bmod 2}{2} \right) \left(\frac{k + k \bmod 2}{2} \right) = \frac{1}{4}(k^2 - k \bmod 2). \tag{19}$$

Moreover,

$$\sum_{k=1}^{n} k \bmod 2 = \left\lfloor \frac{n+1}{2} \right\rfloor, \quad \text{and} \quad \left\lfloor \frac{n - \lfloor \frac{n}{3} \rfloor}{2} \right\rfloor = \left\lfloor \frac{n+1}{3} \right\rfloor. \tag{20}$$

Proof. Elementary, and thus left to the reader. □

Theorem 3.11 *The number of $0/1$-equivalence classes of acute triangles in I^n equals*

$$\left\lfloor \frac{2n^3 + 3n^2 - 6n + 9}{72} \right\rfloor. \tag{21}$$

Proof. We need to sum the expression in (16) over all $r \geq 1$ satisfying $3r \leq n$. Now, since $(n - 3r + 2) \bmod 2 = (n - r) \bmod 2$, we find using Lemma 3.10 that

$$\sum_{r=1}^{\lfloor \frac{n}{3} \rfloor} \left\lfloor \frac{n - 3r + 2}{2} \right\rfloor \left\lceil \frac{n - 3r + 2}{2} \right\rceil = \frac{1}{4} \sum_{r=1}^{\lfloor \frac{n}{3} \rfloor} (n - 3r + 2)^2 - \frac{1}{4} \sum_{r=1}^{\lfloor \frac{n}{3} \rfloor} (n - r) \bmod 2. \tag{22}$$

The first sum in the right-hand side of (22) can be evaluated using standard expressions for sums of squares as

$$\sum_{r=1}^{\lfloor \frac{n}{3} \rfloor} (n-3r+2)^2 = \left\lfloor \frac{n}{3} \right\rfloor (n+2) \left(n-1-3 \left\lfloor \frac{n}{3} \right\rfloor \right) + \frac{3}{2} \left\lfloor \frac{n}{3} \right\rfloor \left(\left\lfloor \frac{n}{3} \right\rfloor + 1 \right) \left(2 \left\lfloor \frac{n}{3} \right\rfloor + 1 \right). \tag{23}$$

Using Lemma 3.10 again, the second sum in the right-hand side of (22) evaluates to

$$\sum_{r=1}^{\lfloor \frac{n}{3} \rfloor} (n-r) \bmod 2 = \sum_{r=1}^{n-1} r \bmod 2 - \sum_{r=1}^{n-\lfloor \frac{n}{3} \rfloor - 1} r \bmod 2 = \left\lfloor \frac{n}{2} \right\rfloor - \left\lfloor \frac{n+1}{3} \right\rfloor. \tag{24}$$

Combining (22), (23) and (24), the number of equivalence classes of acute 0/1-triangles equals

$$\frac{1}{4} \left(\left\lfloor \frac{n}{3} \right\rfloor (n+2) \left(n-1-3 \left\lfloor \frac{n}{3} \right\rfloor \right) + \frac{3}{2} \left\lfloor \frac{n}{3} \right\rfloor \left(\left\lfloor \frac{n}{3} \right\rfloor + 1 \right) \left(2 \left\lfloor \frac{n}{3} \right\rfloor + 1 \right) - \left\lfloor \frac{n}{2} \right\rfloor + \left\lfloor \frac{n+1}{3} \right\rfloor \right). \tag{25}$$

To verify that this expression equals (21) is a tedious task, but can be done as follows. First, we substitute $n = 6k + \ell$ with $\ell \in \{0, \ldots, 5\}$ into (21), which after simplifications results in

$$6k^3 + \frac{3}{2}(2\ell+1)k^2 + \frac{1}{2}\left(\ell^2+\ell-1\right)k + \left\lfloor \frac{1}{36}\ell^3 + \frac{1}{24}\ell^2 - \frac{1}{12}\ell + \frac{1}{8} \right\rfloor, \tag{26}$$

where we have used that $2\ell+1$ and $\ell^2+\ell-1 = \ell(\ell+1)-1$ are both odd, which implies that the sum of the first three terms in (26) is indeed an integer for all k and ℓ.

Next, substitute $n = 6k + \ell$ with $\ell \in \{0, 1, 2\}$ in (25), and note that it simplifies to

$$6k^3 + \frac{3}{2}(2\ell+1)k^2 + \frac{1}{2}\left(\ell^2+\ell-1\right)k, \tag{27}$$

which equals the expression in (26) because for $\ell \in \{0, 1, 2\}$ the floor results in zero. Finally, set $n = 6k + \ell$ with $\ell \in \{3, 4, 5\}$ in (25). After simplification there remains

$$6k^3 + \frac{3}{2}(2\ell+1)k^2 + \frac{1}{2}\left(\ell^2+\ell-1\right)k + \frac{1}{4}\left(\ell^2 - 2\ell + 1 - \left\lfloor \frac{\ell}{2} \right\rfloor + \left\lfloor \frac{\ell+1}{3} \right\rfloor \right). \tag{28}$$

Comparing (26) with (28), it can be easily verified that for $\ell \in \{3, 4, 5\}$,

$$\frac{1}{4}\left(\ell^2 - 2\ell + 1 - \left\lfloor \frac{\ell}{2} \right\rfloor + \left\lfloor \frac{\ell+1}{3} \right\rfloor \right) = \left\lfloor \frac{1}{36}\ell^3 + \frac{1}{24}\ell^2 - \frac{1}{12}\ell + \frac{1}{8} \right\rfloor. \tag{29}$$

And this proves the theorem. □

Below are listed the numbers r_n and a_n of 0/1-equivalence classes of right and acute 0/1-triangles and their sum d_n for small values of n.

n	2	3	4	5	6	7	8	9	10
r_n	1	2	4	6	9	12	16	20	25
a_n	0	1	2	4	7	11	16	23	31
d_n	1	3	6	10	16	23	32	43	56

In the OEIS, the sequence r_n has label A002620, sequence a_n has label A181120, and d_n has label A034198. Only the latter has as description "number of distinct triangles on vertices of n-dimensional cube", the other two are not associated with counting triangles in I^n.

On a personal note

I met Karel Segeth for the first time in the beginning of October 1997, when I was 29 years old. Karel was director of the Mathematical Institute of the Academy of Sciences of the Czech Republic, and I had just arrived to take up a one year visiting position at his Institute. He invited me to his director's office for a cup of tea, and to welcome me. I recall being impressed and a bit nervous, and listened to what Professor Segeth had to say, in his typical (although at that time, of course, I did not know this) calm and amiable tone of voice. He seemed to be the type of person taking his responsibilities seriously; the greater was my surprise when he good-humouredly laughed at my humble wish to take up a Czech language course now that I had arrived in Prague, and actually rather cheekily added: "Pardon me, but I'm afraid you will never learn to speak Czech!". Notwithstanding cheekiness, he immediately organized for me to be enroled in a Czech language course provided by the Academy of Sciences, and until this day I still get goose bumps when I recall the teacher, a strict lady who asked me questions when, and only when, I had completely lost track of things. It was the beginning of my personal quest to prove Karel wrong, a quest that still goes on today, and which, of course, I can never complete. It was also the beginning of a wonderful year in Prague. When I left the institute, Karel spoke the words "Please, come again!".

And so I did. In the almost fifteen years since my first stay in Prague, I have visited the Institute many times a year. Instead of -or maybe better, next to- being an impressive director, Karel became a fellow mathematician, a trustworthy source of Czech culture and history, a fixed point in the audience of my mathematical presentations, and a good companion in not always politically correct jokes and a celebrational glass of spirit. And each time when I left, he spoke the words "Please, come again!". What choice do I have, than to follow his advice?

I wish Karel all the best, and hope to see him regularly at the Institute; at the seminar, the corridor, at Michal's office, the printer room, and to enjoy his typical humor and wisdom for many years to come. Happy seventieth birthday!

Jan Brandts

References

[1] Aichholzer, O.: Extremal properties of 0/1-polytopes of dimension 5. In [3], pp. 111–130, 2000.

[2] Chen, W. Y. C.: Induced cycle structures of the hyperoctahedral group. SIAM J. Discrete Math. **6**(3) (1993), 353–362.

[3] Kalai, G., Ziegler, G. M. (Eds): *Polytopes – Combinatorics and computation.* DMV Seminar, Band 29, Birkhäuser Verlag, Basel, Boston, Berlin, 2000.

Conference Applications of Mathematics 2013
in honor of the 70th birthday of Karel Segeth.
Jan Brandts, Sergey Korotov, Michal Křížek,
Jakub Šístek, and Tomáš Vejchodský (Eds.),
Institute of Mathematics AS CR, Prague 2013

ANALYTICAL SOLUTION OF ROTATIONALLY SYMMETRIC STOKES FLOW NEAR CORNERS

Pavel Burda[1,2], Jaroslav Novotný[3], Jakub Šístek[4]

[1] Department of Applied Mathematics, Czech Technical University
Karlovo náměstí 13, CZ-121 35 Praha 2, Czech Republic
pavel.burda@fs.cvut.cz
[2] VŠB - Technical University of Ostrava
17. listopadu 15/2172, CZ-708 33 Ostrava-Poruba, Czech Republic
[3] Institute of Thermomechanics, Academy of Sciences of the Czech Republic
Dolejškova 5, CZ-182 00 Praha 8, Czech Republic
novotny@it.cas.cz
[4] Institute of Mathematics, Academy of Sciences of the Czech Republic
Žitná 25, CZ-115 67 Praha 1, Czech Republic
sistek@math.cas.cz

Abstract

We present analytical solution of the Stokes problem in rotationally symmetric domains. This is then used to find the asymptotic behaviour of the solution in the vicinity of corners, also for Navier-Stokes equations. We apply this to construct very precise numerical finite element solution.

1. Introduction

In this paper we analyze the singularities arising in rotationally symmetric tubes with nonsmooth walls, e.g. with forward and/or backward steps, or jumps in diameter. The goal of the paper is to contribute to the asymptotic behaviour of the Stokes flow near 'corners' of the rotationally symmetric tubes. We follow up the methodology used in the paper [5] for the Stokes flow in 2D domains.

We start with the general stream function-vorticity formulation. Then by means of the cylindrical coordinates together with rotational symmetry we derive equations for vorticity and stream function in z, ρ geometry (z axial, ρ radial coordinate) as e.g. in a domain in Fig. 1, or Fig. 2.

Then we perform the transformation to polar coordinates r, ϑ, where the point P in Fig. 1 is the pole. So we get the equations for both stream function and vorticity in polar coordinates r, ϑ. Continuing as in [5] we derive the analytical solution for the singularity near the corner P.

Let us note that the asymptotic behaviour applies also to Navier-Stokes equations. The results will be applied to the flow in a tube with forward and/or backward steps.

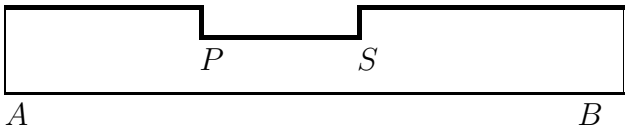


Figure 1: Example of the solution domain in cylindrical z, ρ geometry.

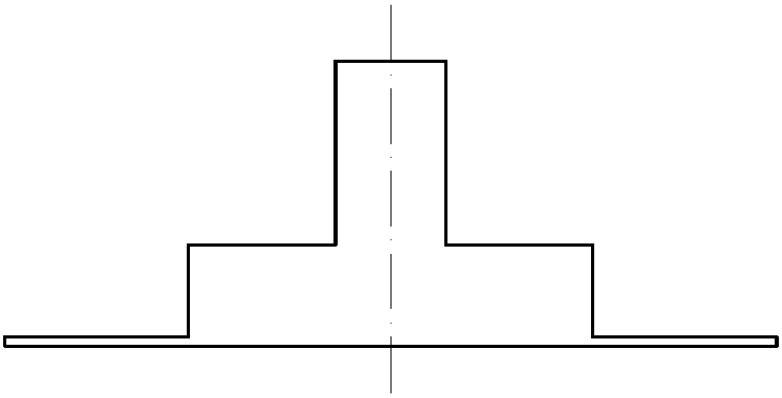


Figure 2: The hydrostatic cell (rotationally symmetric).

2. Stream function-vorticity formulation of the Stokes flow in cylindrical geometry

We start with the general 3D stationary Stokes flow in stream function-vorticity formulation, see e.g. Peyret, Taylor [8],

$$\boldsymbol{\Omega} = \nabla \times \mathbf{V}, \tag{1}$$

$$-\nu \nabla^2 \boldsymbol{\Omega} = \frac{1}{\rho} \nabla \times \mathbf{f}, \tag{2}$$

$$\mathbf{V} = \nabla \times \boldsymbol{\Psi}, \tag{3}$$

$$\nabla^2 \boldsymbol{\Psi} + \boldsymbol{\Omega} = \mathbf{0}, \tag{4}$$

where $\mathbf{V}$ is the vector of velocity, $\boldsymbol{\Omega}$ is the vector of vorticity, $\boldsymbol{\Psi}$ is the stream function vector, ν the kinematic viscosity, ρ is the density, and $\mathbf{f}$ is the external force.

In the paper we study the Stokes flow in the rotationally symmetric tubes, like e.g. the hydrostatic cell, see Fig. 2, or tube on Fig. 1 with line AB as the axis of symmetry.

We first transform the equations (1)–(4) to cylindrical coordinates z, ρ, φ and use the rotational symmetry $\left(\frac{\partial}{\partial \varphi}(\cdot) = 0\right)$. In what follows we use the following formulas (see Batchelor [1])

$$\nabla^2 \mathbf{F} = \left(\nabla^2 F_z, \nabla^2 F_\rho - \frac{F_\rho}{\rho^2}, \nabla^2 F_\varphi - \frac{F_\varphi}{\rho^2} \right), \tag{5}$$

$$\nabla \times \mathbf{F} = \left(\frac{1}{\rho} \frac{\partial}{\partial \rho} \left(\rho F_\varphi \right), -\frac{\partial F_\varphi}{\partial z}, \frac{\partial F_\rho}{\partial z} - \frac{\partial F_z}{\partial \rho} \right), \tag{6}$$

$$\nabla^2 g = \frac{\partial^2 g}{\partial z^2} + \frac{1}{\rho} \frac{\partial}{\partial \rho} \left(\rho \frac{\partial g}{\partial \rho} \right). \tag{7}$$

Denoting

$$\mathbf{V} = \left(V_z, V_\rho, 0 \right),$$

we get, by (1) and (6),

$$\mathbf{\Omega} = \left(0, 0, \frac{\partial V_\rho}{\partial z} - \frac{\partial V_z}{\partial \rho} \right).$$

Now we denote the scalar vorticity

$$\omega = \frac{\partial V_\rho}{\partial z} - \frac{\partial V_z}{\partial \rho}.$$

Similarly we denote the scalar stream function $\psi = \psi_\varphi$ and, by (3) and (6), we get

$$\mathbf{V} = \left(\frac{1}{\rho} \frac{\partial}{\partial \rho} \left(\rho \psi \right), -\frac{\partial \psi}{\partial z}, 0 \right),$$

so that the velocity components are

$$V_z = \frac{\partial \psi}{\partial \rho} + \frac{1}{\rho} \psi, \tag{8}$$

$$V_\rho = -\frac{\partial \psi}{\partial z}.$$

By (4) and (5),

$$\omega = -\left(\nabla^2 \psi - \frac{\psi}{\rho^2} \right),$$

so that, by (7)

$$\boxed{-\omega = \frac{\partial^2 \psi}{\partial z^2} + \frac{\partial^2 \psi}{\partial \rho^2} + \frac{1}{\rho} \frac{\partial \psi}{\partial \rho} - \frac{\psi}{\rho^2}.} \tag{9}$$

In the paper we assume the external forces $\mathbf{f} = 0$, so that, by (2) and (5),

$$\nu \left(\nabla^2 \omega - \frac{\omega}{\rho^2} \right) = 0,$$

which, together with (7) gives

$$\boxed{\nu \left(\frac{\partial^2 \omega}{\partial z^2} + \frac{\partial^2 \omega}{\partial \rho^2} + \frac{1}{\rho} \frac{\partial \omega}{\partial \rho} - \frac{\omega}{\rho^2} \right) = 0.} \tag{10}$$

The equations (9) and (10) together with (8) describe the flow in a rotationally symmetric domain. If we add appropriate boundary conditions, like e.g. prescribed velocity at the inflow, zero velocity on the wall, symmetry condition on the axis of symmetry, and 'do nothing' condition at the outflow, then the flow is uniquely determined.

In the paper we are interested in the behaviour of the solution near the singular points, like e.g. the points P, Q in Fig. 1. This will be the subject of the next section.

3. Stream function and vorticity near the singular points

In order to investigate the behaviour of the flow in the vicinity of the singular point P, we transform the equations (9) and (10) to polar coordinates r, ϑ with pole in the point $P = [z_0, \rho_0]$. Without loss of generality we take $z_0 = 0$. So the transformation is

$$\begin{aligned} z &= r\cos\vartheta, \\ \rho &= \rho_0 + r\sin\vartheta. \end{aligned} \tag{11}$$

The stream function $\psi(z,\rho)$ after transformation will be denoted, for a moment, as $\psi^*(r,\vartheta)$ i.e.

$$\psi^*(r,\vartheta) = \psi(z,\rho) = \psi(r\cos\vartheta, \rho_0 + r\sin\vartheta).$$

Then, using the chain rule, equation (9) gives the equality

$$\frac{\partial^2\psi^*}{\partial r^2} + \frac{1}{r}\frac{\partial\psi^*}{\partial r} + \frac{1}{r^2}\frac{\partial^2\psi^*}{\partial\vartheta^2} + \frac{1}{\rho}\left(\frac{\partial\psi^*}{\partial r} + \frac{1}{r}\frac{\partial\psi^*}{\partial\vartheta}\right) - \frac{\psi^*}{\rho^2} = -\omega^*. \tag{12}$$

As we are interested in the behaviour of the solution in a small neighborhood of the point P, we assume

$$r \ll \rho. \tag{13}$$

Then we may neglect the terms with the coefficients $\frac{1}{\rho}$ and $\frac{1}{\rho^2}$ in (12) and we get the equation for the stream function in polar coordinates (stars deleted)

$$\boxed{\frac{\partial^2\psi}{\partial r^2} + \frac{1}{r}\frac{\partial\psi}{\partial r} + \frac{1}{r^2}\frac{\partial^2\psi}{\partial\vartheta^2} = -\omega.} \tag{14}$$

The same transformation (11) is done for the vorticity ω in (10)

$$\omega^*(r,\vartheta) = \omega(z,\rho) = \omega(r\cos\vartheta, \rho_0 + r\sin\vartheta).$$

Again, using the chain rule, the equation (10) gives the equality (positive constant ν is omitted)

$$\frac{\partial^2\omega^*}{\partial r^2} + \frac{1}{r}\frac{\partial\omega^*}{\partial r} + \frac{1}{r^2}\frac{\partial^2\omega^*}{\partial\vartheta^2} + \frac{1}{\rho}\left(\frac{\partial\omega^*}{\partial r} + \frac{1}{r}\frac{\partial\omega^*}{\partial\vartheta}\right) - \frac{\omega^*}{\rho^2} = 0.$$

Due to assumption (13) we get the equation for the vorticity in polar coordinates (stars deleted)

$$\boxed{\frac{\partial^2\omega}{\partial r^2}+\frac{1}{r}\frac{\partial\omega}{\partial r}+\frac{1}{r^2}\frac{\partial^2\omega}{\partial\vartheta^2}=0.} \tag{15}$$

Let us note that the velocity components u_r, u_ϑ are related to the stream function as follows

$$u_r=\frac{1}{r}\frac{\partial\psi}{\partial\vartheta}, \qquad u_\vartheta=-\frac{\partial\psi}{\partial r}. \tag{16}$$

If we compare equations (14), resp. (15) that apply to rotationally symmetric flow, with the equations (4), resp. (5) in [5] that apply to plain flow, we see that they are identical. In other words, we proved the following assertion.

Assertion 1. Under the assumption (13), the asymptotic behaviour of the Stokes flow near the corners of the rotationally symmetric tube is the same as that of the Stokes flow in a 2D channel.

4. Analytical solution for singularities

In the paper [5] we used the separation of variables

$$\psi(r,\vartheta)=P(r)\,F(\vartheta), \tag{17}$$

$$\omega(r,\vartheta)=R(r)\,G(\vartheta) \tag{18}$$

in order to find the singular part of the solution of the equations (14) and (15) in the neighborhood of the point P (see Fig. 1). There it was done for the channel flow. Now, due to the identical equations, cf. Assertion 1, we may proceed in the same way also in the case of rotationally symmetric flow.

Namely, we consider fluid flow in the rotationally symmetric region with boundary corner of nonconvex internal angle α, cf. Fig. 1. We assume a rigid boundary and nonslip boundary conditions, so that the boundary conditions for the stream function are

$$\psi(r,0)=0, \qquad \psi(r,\alpha)=0, \tag{19}$$

$$\frac{\partial\psi}{\partial\vartheta}(r,0)=0, \qquad \frac{\partial\psi}{\partial\vartheta}(r,\alpha)=0. \tag{20}$$

As proved in [5], the singular part of the stream function ψ is

$$\psi(r,\vartheta)=r^{\gamma+1}F(\vartheta), \tag{21}$$

where γ is the solution of the algebraic equation

$$\gamma^2\sin^2\alpha-\alpha\sin^2\gamma=0. \tag{22}$$

In the case of domains in Figs. 1 and 2 the angle is

$$\alpha=\frac{3}{2}\pi. \tag{23}$$

So that, by (22),

$$\gamma = 0.5444837, \tag{24}$$

and we get for the stream function the asymptotic behaviour near the angle $\frac{3\pi}{2}$:

$$\psi(r,\vartheta) = r^{1.54448} \cdot F(\vartheta), \tag{25}$$

where the function F does not depend on r. Consequently, for the velocity components, by (16) we have the asymptotics

$$\begin{aligned} u_r &= r^{\gamma}\ F_1(\vartheta) = r^{0.54448}\ F_1(\vartheta), \\ u_\vartheta &= r^{\gamma}\ F_2(\vartheta) = r^{0.54448}\ F_2(\vartheta), \end{aligned} \tag{26}$$

where the functions $F_1(\vartheta), F_2(\vartheta)$ are independent of r.

For pressure, we derived in [5] the asymptotic behaviour

$$p \approx r^{\gamma-1}\Phi_p(\vartheta) \approx r^{-0.45552}\Phi_p(\vartheta), \tag{27}$$

where the function $\Phi_p(\vartheta)$ is independent of r.

Let us note that for 2D channel flow, the same asymptotics were also found by a different technique in Kondratiev [6] and in Ladeveze and Peyret [7]. For rotationally symmetric flow the technique based on Kondratiev was used in [2]. Further, we note that the asymptotics (26) and (27) apply also to the Navier-Stokes equations, see e.g. [2].

5. Application to finite element solution of Navier-Stokes equations

In [4] and [5] we described the way how to make use of the asymptotics of the solution near the singular points. Together with the a priori error estimates we suggested and applied the algorithm for designing the finite element mesh in the neighbourhood of the singular point. Due to the Assertion 1, the results obtained in [4] may be applied to axisymmetric flows, using the 2D domain as a cross section of the axisymmetric tube.

For evaluating the achieved accuracy of the approximate solution, we use the a posteriori error estimator, see e.g. [3].

6. Numerical results

We study flow in the rotationally symmetric domain of the hydrostatic cell from Fig. 2. Similar results were obtained for a two-dimensional flow problem in [4]. Figure 3 shows the shape of the singularity in pressure solution near the bottom corner of the hydrostatic cell. In Fig. 4, we compare the asymptotic behaviour of pressure near the bottom corner of the cell obtained by formula (27) with the solution in the horizontal cut obtained by FEM. Let us note that the finite element mesh was not designed by our algorithm here, and we used a simple local refinement offered by the program GMSH. A more precise FEM solution would need a finer mesh.

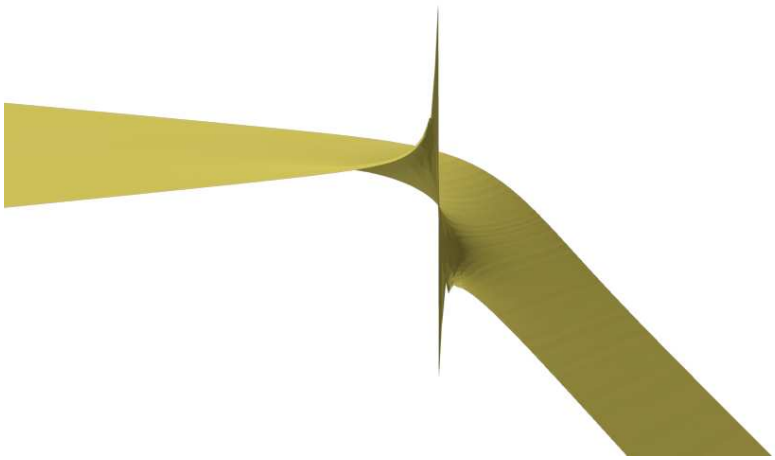


Figure 3: Singularity of pressure in hydrostatic cell.

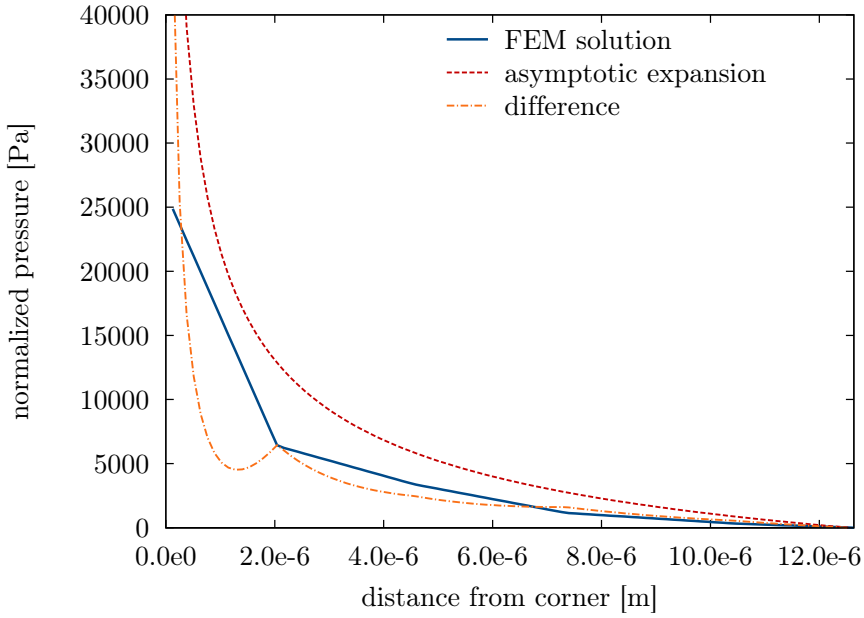


Figure 4: Pressure near the bottom corner: asymptotic versus FEM solution.

7. Conclusion

In the paper, we are interested in Stokes problem with singularities caused by nonconvex corners in rotationally symmetric domains. We proved that the asymptotic behaviour of the Stokes flow near the corners of the rotationally symmetric tube is the same as that of the Stokes flow in a 2D channel. For the Stokes flow we find analytically the principal part of the asymptotics of the solution in the vicinity of corners. This result may be used on one hand to construct the finite element mesh

adjusted to singularity. This mesh is then used to find a very precise solution to Stokes but also Navier-Stokes equations. On the other hand, the analytical solution of the Stokes flow near corners may be used to test other methods.

Acknowledgements

This work was supported by the IT4Innovations Centre of Excellence project, reg. no. CZ.1.05/1.1.00/02.0070, and by the Academy of Sciences of the Czech Republic through RVO:61388998 and RVO:67985840.

References

[1] Batchelor, G. K.: *An introduction to fluid dynamics*. Cambridge University Press, 1967.

[2] Burda, P.: On the FEM for the Navier-Stokes equations in domains with corner singularities. In: M. Křížek, et al. (Eds.), *Finite element methods, Supeconvergence, Post-Processing and A Posteriori Estimates*, pp. 41–52. Marcel Dekker, New York, 1998.

[3] Burda, P., Novotný, J., Sousedík, B.: A posteriori error estimates applied to flow in a channel with corners. Math. Comput. Simulation **6**1 (2003), 375–383.

[4] Burda, P., Novotný, J., Šístek J.: Precise FEM solution of a corner singularity using an adjusted mesh. Internat. J. Numer. Meth. Fluids **4**7 (2005), 1285–1292.

[5] Burda, P., Novotný, J., Šístek J.: Analytical solution of Stokes flow near corners and applications to numerical solution of Navier-Stokes equations with high precision. In: J. Brandts, et al. (Eds.), Proc. Conf. Applications of Mathematics 2012, Prague, May 2-5, 2012, Inst. Math. Acad. Sci., pp. 43–54.

[6] Kondratiev, V. A.: Asimptotika rešenija uravnienija Nav'je-Stoksa v okrestnosti uglovoj točki granicy. Prikl. Mat. i Mech. **1** (1967) 119–123.

[7] Ladevéze, J., Peyret, R.: Calcul numérique d'une solution avec singularité des équations de Navier-Stokes: écoulement dans un canal avec variation brusque de section. J. de Mécanique **1**3 (1974), 367–396.

[8] Peyret, R., Taylor, T. D.: *Computational methods for fluid flow*. Springer, Berlin, 1983.

Conference Applications of Mathematics 2013
in honor of the 70th birthday of Karel Segeth.
Jan Brandts, Sergey Korotov, Michal Křížek,
Jakub Šístek, and Tomáš Vejchodský (Eds.),
Institute of Mathematics AS CR, Prague 2013

A METHOD TO RIGOROUSLY ENCLOSE EIGENPAIRS OF COMPLEX INTERVAL MATRICES

Roberto Castelli[1], Jean-Philippe Lessard[2]

[1] BCAM – Basque Center for Applied Mathematics
Alameda de Mazarredo 14, 48009 Bilbao, Spain
rcastelli@bcamath.org

[2] Université Laval, Département de Math. et de Stat., Québec, QC, G1V 0A6, Canada
and BCAM - Basque Center for Applied Mathematics
Alameda de Mazarredo 14, 48009 Bilbao, Spain
jean-philippe.lessard@mat.ulaval.ca

Abstract

In this paper, a rigorous computational method to enclose eigenpairs of complex interval matrices is proposed. Each eigenpair $x = (\lambda, \mathrm{v})$ is found by solving a nonlinear equation of the form $f(x) = 0$ via a contraction argument. The set-up of the method relies on the notion of *radii polynomials*, which provide an efficient mean of determining a domain on which the contraction mapping theorem is applicable.

1. Introduction

Computing eigenvalues and eigenvectors of matrices is a central problem in many fields of applied sciences involving mathematical modelling. When applied to real-life phenomena, models need to consider the occurrence of diverse errors in the data, due for instance to inaccuracy of measurements or noise effects. Such uncertainty in the data can be represented by intervals. In the context of studying a matrix with uncertain entries, interval matrices can be considered. Our goal here is to develop a rigorous computational method to enclose eigenpairs of complex interval matrices.

Before proceeding further, note that bounds for eigendecompositions of standard (non interval) matrices are abundant, ranging from classical perturbation theory like Bauer-Fike residual and condition number based theorems [4], via Kato-Temple bounds [5], to Rayleigh-Ritz bounds [6, 7, 8, 9], to bounds coming from Newton-Kantorovich type arguments [1, 2], to pseudospectral bounds, and so on. Many such results can, for instance, lead to bounds on the nearest eigenpair to a given approximation. Also, while the problem of computing rigorous bounds for the eigenvalue set of interval matrices is well studied, see [10, 11] and the references therein, a not so large literature has been produced regarding the simultaneous enclosure of the eigenvalues and eigenvectors of interval matrices. In this direction we refer to [1, 12],

where different techniques have been developed to enclose simple eigenvalues and corresponding eigenvectors, while for double or nearly double eigenvalues a method has been introduced in [13]. For the rigorous enclosure of multiple or nearly multiple eigenvalues of complex matrices, a contribution has been made by S. Rump in [3, 14].

In this paper, we propose the new idea of enclosing rigorously the eigenpairs of complex interval matrices by using the notion of the *radii polynomials*, which provide a computationally efficient way of determining a domain on which the contraction mapping theorem is applicable. The radii polynomials approach, which is similar to the approaches of Newton-Kantorovich and the Krawczyk operator, aims at demonstrating existence and local uniqueness of solutions of nonlinear equations. The Newton-Kantorovich approach fixes *a priori* the radius r of a ball $B(r)$ around a numerically computed eigenpair and attempt to demonstrate the existence of a contraction on $B(r)$. Similarly, the Krawczyk operator approach consists of applying directly the operator to interval vectors (in the form of small neighbourhoods around a numerical approximation) and then attempt to verify *a posteriori* the hypotheses of a contraction mapping argument [15, 16]. On the other hand, the radii polynomials are *a priori* conditions that are derived analytically, and once they are theoretically constructed, they are used to *solve* for the sets (also in the form of small neighbourhoods of a numerical solution) on which a Newton-like operator is a contraction. The radii polynomials were originally introduced in [17] to compute equilibria of PDEs with the goal of minimizing the extra computational cost required to prove existence of solutions of infinite dimensional PDEs [18].

The paper is organized as follows. In Section 2, the method is introduced to enclose rigorously eigenpairs of non interval matrices and in Section 3 it is generalized to the case of interval matrices. In Section 4, we present applications and compare our method to the method of [14] and to a method based on the Krawczyk operator.

2. The computational method

We fix some notation. We denote by $\mathbb{IC}^{n\times n}$ the set of complex matrices with interval entries, $A \in \mathbb{C}^{n\times n}$ an $n \times n$ complex matrix and $\boldsymbol{A} \in \mathbb{IC}^{n\times n}$ an $n \times n$ interval complex matrix, meaning that any entry of $\boldsymbol{A}$ is a complex interval of the form

$$\boldsymbol{A}_{i,j} = [Re(\hat{A}_{i,j}) \pm rad^{(1)}_{i,j}] + \imath[Im(\hat{A}_{i,j}) \pm rad^{(2)}_{i,j}], \quad rad^{(1)}_{i,j}, rad^{(2)}_{i,j} \in \mathbb{R}_+ \ ,$$

where $\hat{A} \in \mathbb{C}^{n\times n}$ is called the center of $\boldsymbol{A}$ while $rad^{(1)}_{i,j}, rad^{(2)}_{i,j}$ are called the radii of the real and imaginary part of $\boldsymbol{A}_{i,j}$, resp. We denote $A \in \boldsymbol{A}$, if $A_{i,j} \in \boldsymbol{A}_{i,j}$ for any $1 \le i, j \le n$. Bold face letters will always denote interval quantities. Moreover,

- $|\cdot|$ is the complex absolute value and, in case of matrices $M \in \mathbb{C}^{n\times m}$, it acts component-wise, that is $|M|_{i,j} = |M_{i,j}|$;
- given two real matrices M, N, we write $M \preceq N$ if and only if $M_{i,j} \le N_{i,j}$ for all i, j. The same notation holds for $\prec$, $\succ$ and $\succeq$;

- I_n denotes the $n \times n$ dimensional identity matrix, $\mathbb{1}_n$ is the column vector of length n with all the entries equal to 1;
- given any matrix $M \in \mathbb{C}^{n\times m}$, the object $(M)_{\hat{k}}$ stands for the $n\times(m-1)$ matrix obtained by deleting the k-th column of M.

Given $\boldsymbol{A} \in \mathbb{IC}^{n\times n}$, we aim to enclose eigenpairs of any $A \in \boldsymbol{A}$. To simplify the exposition, we first present the method in the context of non interval matrices $A \in \mathbb{C}^{n\times n}$. Minor modifications are needed for the extension to the interval case.

Suppose that an approximate eigenpair of A has been computed, that is $(\bar{\lambda}, \bar{\mathrm{v}})$ such that $A\bar{\mathrm{v}} \approx \bar{\lambda}\bar{\mathrm{v}}$ and let $f(x)$ be the function $f : \mathbb{C}^n \to \mathbb{C}^n$ that maps a point $x = (\lambda, \mathrm{v}_1, \mathrm{v}_2, \ldots, \mathrm{v}_{k-1}, \mathrm{v}_{k+1}, \ldots, \mathrm{v}_n)$ to

$$f(x) = A \begin{bmatrix} \mathrm{v}_1 \\ \vdots \\ \bar{\mathrm{v}}_k \\ \vdots \\ \mathrm{v}_n \end{bmatrix} - \lambda \begin{bmatrix} \mathrm{v}_1 \\ \vdots \\ \bar{\mathrm{v}}_k \\ \vdots \\ \mathrm{v}_n \end{bmatrix}, \tag{1}$$

where $\bar{\mathrm{v}}_k$ is the largest component in absolute value of $\bar{\mathrm{v}}$. Fixing $\mathrm{v}_k = \bar{\mathrm{v}}_k$ ensures that the solution is isolated. Note that the more standard approach of fixing $\|\mathrm{v}\| = 1$ will fail to provide isolation if v is complex. Indeed, given an eigenpair $(\lambda, \mathrm{v}) \in \mathbb{C}^{n+1}$ where v is complex, then for any θ, $(\lambda, \mathrm{e}^{\mathrm{i}\theta}\mathrm{v})$ is also an eigenpair and $\|\mathrm{e}^{\mathrm{i}\theta}\mathrm{v}\| = 1$.

By definition, a solution x of $f(x) = 0$ corresponds to an eigenpair (λ, v) of A with the eigenvalue λ given by the first component of x and the eigenvector $\mathrm{v} = (\mathrm{v}_1, \ldots, \mathrm{v}_{k-1}, \bar{\mathrm{v}}_k, \mathrm{v}_{k+1}, \ldots, \mathrm{v}_n)$. We then aim at proving existence of zeros of $f(x)$ together with rigorous bounds. Denoting $\bar{x} = (\bar{\lambda}, \bar{\mathrm{v}}_1, \bar{\mathrm{v}}_2, \ldots, \bar{\mathrm{v}}_{k-1}, \bar{\mathrm{v}}_{k+1}, \ldots, \bar{\mathrm{v}}_n)$ and $Df(\bar{x})$ the Jacobian matrix of f at $\bar{x}$, one has that

$$Df(\bar{x}) = \left(- \left[\begin{array}{c} \bar{\mathrm{v}}_1 \\ \vdots \\ \bar{\mathrm{v}}_k \\ \vdots \\ \bar{\mathrm{v}}_n \end{array} \right] \middle| (A - \bar{\lambda} I_n)_{\hat{k}} \right). \tag{2}$$

We find zeros of f by introducing a fixed point operator T. Endow $\mathbb{C}^n$ with the norm $\|x\|_\infty = \max_{i=1,\ldots,n}\{|x_i|\}$. Consider $R \approx Df(\bar{x})^{-1}$ an invertible matrix. Define

$$T : \mathbb{C}^n \to \mathbb{C}^n : x \mapsto T(x) = x - Rf(x) \tag{3}$$

so that fixed points of T are in bijection with zeros of f. In practice, the matrix R is computed numerically in MATLAB. Note that getting a good approximate inverse is fundamental for our method to provide sharp bounds. Indeed, as one shall see shortly, the better the approximate inverse R is, the smaller the bound $Z^{(1)}$ in (4) will be. Since fixed points of T correspond to zeros of $f(x)$, the idea is to construct a small set $B \subset \mathbb{C}^n$ such that $T : B \to B$ is a contraction, and then to apply the contraction mapping theorem to conclude about the existence of a unique fixed point

of T in B. Note that $\bar{x}$ is an approximate zero of f and the operator T has been defined as a Newton-like operator around the point $\bar{x}$, thus it is advantageous to test the contractibility of T on neighbourhoods of $\bar{x}$ in $\mathbb{C}^n$. More precisely, denote by $B(r) = \{x \in \mathbb{C}^n, \|x\|_\infty \leq r\}$ the closed ball of radius r around the origin and let $B_{\bar{x}}(r) = \bar{x} + B(r)$ be the ball with the same radius and centered at $\bar{x}$. Treating r as a variable, we choose the balls $B_{\bar{x}}(r)$ as the candidate sets where to check if T is a contraction. The next result provides a recipe to determine the radius r.

Theorem 2.1. *Consider $\bar{x} \in \mathbb{C}^n$ and R a real $n \times n$ invertible matrix. Consider the nonlinear problem (1) and bounds $Y, Z^{(1)}, Z^{(2)} \in \mathbb{R}^n$ such that*

$$|Rf(\bar{x})| \preceq Y, \quad |I_n - R \cdot Df(\bar{x})|\mathbb{1}_n \preceq Z^{(1)}, \quad 2|R|(\mathbb{1}_n)_{\hat{k}} \preceq Z^{(2)}. \tag{4}$$

Define the radii polynomials $p_1(r), p_2(r), \ldots, p_n(r)$ by

$$p_i(r) = Z_i^{(2)} r^2 + (Z_i^{(1)} - 1)r + Y_i, \tag{5}$$

and define $\mathcal{I} = \cap_{i=1}^n \{r > 0 : p_i(r) < 0\}$. If $\mathcal{I} \neq \emptyset$, then for any $r \in \mathcal{I}$, there exists a unique $\hat{x} \in B_{\bar{x}}(r)$ such that $f(\hat{x}) = 0$.

Proof. Consider $r \in \mathcal{I} \neq \emptyset$. Recalling (3), consider $T(x) = x - Rf(x)$. Then

$$\begin{aligned}
\sup_{b,c \in B(r)} |DT(\bar{x}+b)c| &= \sup_{b,c \in B(r)} |(I_n - R \cdot Df(\bar{x}))c + R(Df(\bar{x}) - Df(\bar{x}+b))c| \\
&\preceq \sup_{b,c \in B(r)} |(I_n - R \cdot Df(\bar{x}))c| + |R(Df(\bar{x}) - Df(\bar{x}+b))c| \\
&\preceq Z^{(1)} r + Z^{(2)} r^2.
\end{aligned}$$

In the last inequality, we used that for any $b = (b_\lambda, b_1, \ldots, b_{k-1}, b_k, \ldots, b_n) \in B(r)$

$$(Df(\bar{x}) - Df(\bar{x}+b)) = \left(\begin{bmatrix} b_1 \\ \vdots \\ b_{k-1} \\ 0 \\ b_{k+1} \\ \vdots \\ b_n \end{bmatrix} \middle| (b_\lambda I_n)_{\hat{k}} \right).$$

Note that the k-th row of the above matrix is null. Since $|b_i| \leq r$, we have that $|(Df(\bar{x}) - Df(\bar{x}+b))c| \preceq 2r^2(\mathbb{1}_n)_{\hat{k}}$ and therefore, $\sup_{b,c \in B(r)} |R[(Df(\bar{x}) - Df(\bar{x}+b))c]| \preceq 2r^2|R|(\mathbb{1}_n)_{\hat{k}} = Z^{(2)} r^2$.

Letting $Z(r) := Z^{(1)} r + Z^{(2)} r^2$, we get that $\sup_{b,c \in B(r)} |DT(\bar{x}+b)c| \preceq Z(r)$. The Mean Value Theorem applied component-wise to T implies that for any $x, y \in B_{\bar{x}}(r)$ and for any $i = 1, \ldots, n$, $T_i(x) - T_i(y) = DT_i(z)(x-y)$, for some $z \in \{tx + (1-t)y : t \in [0,1]\} \subset B_{\bar{x}}(r)$. Then,

$$|T_i(x) - T_i(y)| = \left| DT_i(z) \frac{r(x-y)}{\|x-y\|_\infty} \right| \frac{1}{r} \|x-y\|_\infty \leq \frac{Z_i(r)}{r} \|x-y\|_\infty \leq Z_i(r). \tag{6}$$

Let $x \in B_{\bar{x}}(r)$ and $y = \bar{x}$ in (6), and using that $T(\bar{x}) - \bar{x} = -Rf(\bar{x})$, one has that

$$|T_i(x) - \bar{x}_i| \leq |T_i(x) - T_i(\bar{x})| + |T_i(\bar{x}) - \bar{x}_i| \leq Z_i(r) + Y_i = Z_i^{(2)} r^2 + Z_i^{(1)} r + Y_i < r$$

by the hypothesis that $p_i(r) < 0$, which follows from the fact that $r \in \mathcal{I}$. That shows that $T(B_{\bar{x}}(r)) \subseteq B_{\bar{x}}(r)$. From (6), it follows that

$$\|T(x) - T(y)\|_\infty = \max_i \{|T_i(x) - T_i(y)|\} \leq \frac{\|Z(r)\|_\infty}{r} \|x - y\|_\infty.$$

Since $Z_i(r) \leq Z_i(r) + Y_i < r$ for any $i = 1, \ldots, n$, it follows that $\|Z(r)\|_\infty < r$. Hence, T is a contraction with contraction constant $\frac{\|Z(r)\|_\infty}{r} < 1$. From the contraction mapping theorem, there exists a unique $\hat{x} \in B_{\bar{x}}(r)$ such that $T(\hat{x}) = \hat{x}$. By invertibility of R, there exists a unique $\hat{x} \in B_{\bar{x}}(r)$ such that $f(\hat{x}) = 0$. □

In summary, given an approximate eigenpair $(\bar{\lambda}, \bar{\mathrm{v}})$, the method consists of computing rigorously the bounds $Y, Z^{(1)}, Z^{(2)}$ given in (4), and then to check whether there exists an interval $\mathcal{I}$ where all the polynomials $p_i(r)$ are negative. If $\mathcal{I} \neq \emptyset$ we select $r = \inf \mathcal{I}$ and we conclude that $f = 0$ has a unique solution within the ball $B_{\bar{x}}(r)$. In practice, we get the existence of an eigenpair (λ, v) of A, with $|\lambda - \bar{\lambda}| \leq r$, $|\mathrm{v}_j - \bar{\mathrm{v}}_j| \leq r$, for $j \neq k$ and $\mathrm{v}_k = \bar{\mathrm{v}}_k$. To prove the existence of another eigenpair of A, it is necessary to provide a different numerical approximate solution $(\bar{\lambda}, \bar{\mathrm{v}})$, different from the previous one, and to repeat the computation.

3. Extension to the interval case

Besides few modifications necessary to deal with interval quantities, the procedure to compute rigorously bounds for the eigenpairs of an interval matrix $\boldsymbol{A} \in \mathbb{IC}^{n \times n}$ is basically the same as for the scalar case. However, a fundamental difference is that all the computations are done using interval arithmetic [10], in which any of the basic operations $\circ \in \{+, -, \cdot, /\}$ is extended to the interval case in order to satisfy the general assumption

$$\forall P \in \boldsymbol{P} \quad \forall Q \in \boldsymbol{Q}, \quad P \circ Q \in \boldsymbol{P} \circ \boldsymbol{Q} \; . \tag{7}$$

Given an interval complex valued matrix $\boldsymbol{A}$, we now address the problem wether or not we can rigorously enclose the eigenpairs of any $A \in \boldsymbol{A}$. Recall that $\hat{A}$ is the center of the interval matrix $\boldsymbol{A}$. We first compute $(\bar{\lambda}, \bar{\mathrm{v}})$ an approximate eigenpair of $\hat{A}$ and, as before, define $\bar{x} = (\bar{\lambda}, \bar{\mathrm{v}}_1, \bar{\mathrm{v}}_2, \ldots, \bar{\mathrm{v}}_{k-1}, \bar{\mathrm{v}}_{k+1}, \ldots, \bar{\mathrm{v}}_n)$, where the missing component $\bar{\mathrm{v}}_k$ is chosen so that $|\bar{\mathrm{v}}_k| = \max_j \{|\bar{\mathrm{v}}_j|\}$. Then, replacing the scalar matrix A in (1) by the interval matrix $\boldsymbol{A}$, the function $f(x)$ and the Jacobian matrix $Df(\bar{x})$ defined in (1) and (2) are replaced respectively by $\boldsymbol{f} : \mathbb{C}^n \to \mathbb{IC}^n$ and by an interval matrix $\boldsymbol{Df}(\bar{x})$ that represents a linear operator from $\mathbb{C}^n$ to $\mathbb{IC}^n$. We choose R to be a numerical inverse of $\widehat{Df}(\bar{x})$, the center of $\boldsymbol{Df}(\bar{x})$, and we proceed to the definition of the operator $\boldsymbol{T}(x) = x - R\boldsymbol{f}(x)$ and to the bounds $Y, Z^{(1)}, Z^{(2)}$, as done before

with the boldface quantities in place of the previous one. Clearly some quantities on the left hand side of relations (4) are now intervals, thus we define component-wise $Y, Z^{(1)}, Z^{(2)}$ as the supremum over the intervals involved, yielding uniform bounds

$$|R\boldsymbol{f}(\bar{x})| \preceq Y, \quad |I_n - R \cdot \boldsymbol{Df}(\bar{x})|\mathbb{1}_n \preceq Z^{(1)}, \quad 2|R|(\mathbb{1}_n)_{\hat{k}} \preceq Z^{(2)}.$$

As in Theorem 2.1, define the radii polynomials by $p_i(r) = Z_i^{(2)} r^2 + (Z_i^{(1)} - 1)r + Y_i$, for $i = 1, \ldots, n$. If $r \in \mathcal{I} = \cap_{i=1}^n \{r > 0 : p_i(r) < 0\}$, then for all $A \in \boldsymbol{A}$, there exists a unique $(\lambda, \mathrm{v}) \in B_{\bar{x}}(r)$ such that $|\lambda - \bar{\lambda}| \leq r$, $|\mathrm{v}_j - \bar{\mathrm{v}}_j| \leq r$, $\mathrm{v}_k = \bar{\mathrm{v}}_k$, and $A\mathrm{v} = \lambda\mathrm{v}$. In other words, r is a uniform bound in $\boldsymbol{A}$ for the existence of an eigenpair of any $A \in \boldsymbol{A}$. Indeed, having fixed $(\bar{\lambda}, \bar{\mathrm{v}})$ and $R \approx \left(\widehat{Df}(\bar{x})\right)^{-1}$, for any $A \in \boldsymbol{A}$ define $f_A(x)$ and $Df_A(\bar{x})$ as in (1) and (2), and the fixed point operator $T_A(x) = x - Rf_A(x)$. The fundamental inclusion (7) implies that $f_A(x) \in \boldsymbol{f}(x)$, $Df_A(\bar{x}) \in \boldsymbol{Df}(\bar{x})$ and $T_A(x) \in \boldsymbol{T}(x)$, for any $A \in \boldsymbol{A}$ and $x \in \mathbb{C}^n$. Thus, as A varies in $\boldsymbol{A}$, the bounds (4), with T_A in place of T, are satisfied for the same $Y, Z^{(1)}, Z^{(2)}, r$ proving the existence of a fixed point in $B_{\bar{x}}(r)$ for any T_A and consequently an eigenpair for any $A \in \boldsymbol{A}$.

4. Results

In this section we report some computational results. All the computations have been done in MATLAB supported by the package INTLAB [19] where the interval arithmetic routines have been implemented. The approximate eigenpairs $(\bar{\lambda}, \bar{\mathrm{v}})$ of $\hat{A}$ have been computed running the standard `eig.m` function in MATLAB. In order to avoid rounding error and to obtain rigorous results, we emphasize that the computational algorithm treats any matrix as an interval matrix. Thus, even if one wishes to deal with a scalar matrix A, the method first constructs a (narrow) interval matrix around A and perform all the computation with interval arithmetics.

Example 1. Consider the interval matrix $\boldsymbol{A}$ centered at

$$\hat{A} = \begin{bmatrix} -10.55360193 & 5.33379647 & -5.24740415 \\ 0.31403414 & 2.33062549 & -3.32865541 \\ -7.49045333 & 5.01386821 & -5.44369022 \end{bmatrix}$$

with radius $rad = 9.66146973 \cdot 10^{-7}$, meaning that each entry $\boldsymbol{A}(i,j)$ consists of the interval $[\hat{A}(i,j) - rad, \hat{A}(i,j) + rad]$. Using the method of Section 2, it results that any $A \in \boldsymbol{A}$ admits three eigenpairs $(\lambda_i, \mathrm{v}_i)$, $i = 1, 2, 3$ each one lying in the ball of radius r_i around the approximate values $(\bar{\lambda}_i, \bar{\mathrm{v}}_i)$ given in Table 1.

We remark that in the general situation the genuine solution (λ, v) of the eigenproblem is proved to exist in a complex neighborhood of the approximate solution $(\bar{\lambda}, \bar{\mathrm{v}})$. Therefore, even if one or both $\bar{\lambda}$ and $\bar{\mathrm{v}}$ are real vectors, the same cannot be concluded for λ or v. However, if the matrix A and the approximate solution $\bar{\lambda}$ and $\bar{\mathrm{v}}$ are real and the computation is successful, then the genuine solution so obtained by solving the radii polynomials is also real. Indeed, suppose the contrary, that is the

	$i=1$	$i=2$	$i=3$
r_i	$2.7747640834393 \cdot 10^{-6}$	$3.5677963538014 \cdot 10^{-5}$	$3.6494066386385 \cdot 10^{-5}$
$\bar{\lambda}_i$	-13.9620493680589	$-9.3632556453596 \cdot 10^{-14}$	0.2953827013923
$\bar{\mathrm{v}}_i$	$\begin{bmatrix} -0.77788012985175 \\ -0.11136179959087 \\ -0.61846669528252 \end{bmatrix}$	$\begin{bmatrix} 0.12133203779007 \\ 0.80769996168880 \\ 0.57697427021802 \end{bmatrix}$	$\begin{bmatrix} 0.15662675418092 \\ 0.83598562894630 \\ 0.52592403260356 \end{bmatrix}$

Table 1: Rigorous enclosure of the eigenpairs of $\boldsymbol{A}$.

exact solution λ and v are complex. Since A is real, the complex conjugate couple $(\mathcal{C}(\lambda), \mathcal{C}(\mathrm{v})))$ is also a solution of the eigenproblem, $A\mathcal{C}(\mathrm{v}) = \mathcal{C}(\lambda)\mathcal{C}(\mathrm{v})$. But both the solutions (λ, v) and $(\mathcal{C}(\lambda), \mathcal{C}(\mathrm{v})))$ belong to the same ball in $\mathbb{C}^n$ around $\bar{x}$ and this violates the uniqueness result stated in Theorem 2.1. The same argument extends in the case of interval matrices.

Example 2: matrices with interval entries of large radius. In this example, we rigorously enclose all eigenpairs of an interval matrix $\boldsymbol{A}$ constructed as follows: consider the complex number $\lambda_0 = 0$ and $\lambda_j = \mathrm{e}^{\mathrm{i}\frac{2\pi}{5}j}$, $j = 1, \ldots, 5$ and define D as the diagonal matrix with entries λ_i, $i = 0, \ldots, 5$. Let $\hat{A} = XDX^{-1}$, for a random matrix X with values in the complex square $[-1, 1] + \mathrm{i}[-1, 1]$ and finally let $\boldsymbol{A}$ be the interval complex matrix centered at $\hat{A}$ with component-wise radius rad both in the real and imaginary part. For different values of rad we compute the enclosure of the eigenvalues of $\boldsymbol{A}$. Consider $\bar{\lambda}_i$ the approximate eigenvalues of $\hat{A}$ given by $\bar{\lambda}_0 = 0$, $\bar{\lambda}_1 = 0.30901 + 0.95105\mathrm{i}$, $\bar{\lambda}_2 = -0.80901 + 0.58778\mathrm{i}$, $\bar{\lambda}_3 = -0.80901 - 0.58778\mathrm{i}$, $\bar{\lambda}_4 = 0.30901 - 0.95105\mathrm{i}$ and $\bar{\lambda}_5 = 1$. Denote by r_i, $i = 0, \ldots 5$ the radius of the ball in the complex plane centered at $\bar{\lambda}_i$ inside which, for any $A \in \boldsymbol{A}$, a unique eigenvalues of A has been proved to exist. The results are presented in Table 2. See also Figure 1 for the enclosure of the six eigenvalues of any $A \in \boldsymbol{A}$ for $rad = 1.3 \cdot 10^{-3}$.

We see in Table 2 that for values of $rad \approx 10^{-3}$ the method starts to fail. A natural question is whether it is possible to predict up for which values of rad the method will be successful. We underline that the technique we propose is a verification method, therefore among other conditions, the success or the failure is strictly related to the accuracy of the approximate solution that could change from one computation to the other. Hence it is not possible to determine a priori the maximum value for rad. However we can to get an idea of what happens when rad increases and, based on that, we can guess which is the maximal admissible value of rad. A necessary condition for the method to be successful is that all the radii polynomials defined in (5) cross the r-axis. That occurs if

$$(Z_i^{(1)} - 1)^2 - 4Y_i Z_i^{(2)} > 0 \tag{8}$$

rad	r_0	r_1	r_2	r_3	r_4	r_5
$1 \cdot 10^{-5}$	$1.2 \cdot 10^{-4}$	$1.0 \cdot 10^{-4}$	$9.2 \cdot 10^{-5}$	$1.4 \cdot 10^{-4}$	$1.3 \cdot 10^{-4}$	$1.0 \cdot 10^{-4}$
$1 \cdot 10^{-4}$	0.0013	0.0011	0.0001	0.0015	0.0014	0.0012
$1 \cdot 10^{-3}$	0.0149	0.0115	0.0103	0.0184	0.0168	0.0122
$1.5 \cdot 10^{-3}$	0.0254	0.0186	0.0163	–	0.0307	0.0197
$2.0 \cdot 10^{-3}$	–	0.0272	0.0234	–	–	0.0287
$2.5 \cdot 10^{-3}$	–	0.0390	0.0322	–	–	0.0407
$3.0 \cdot 10^{-3}$	–	–	0.0450	–	–	–
$3.5 \cdot 10^{-3}$	–	–	–	–	–	–

Table 2: Enclosures of the eigenpairs of the complex interval matrix $\boldsymbol{A}$, as rad grows.

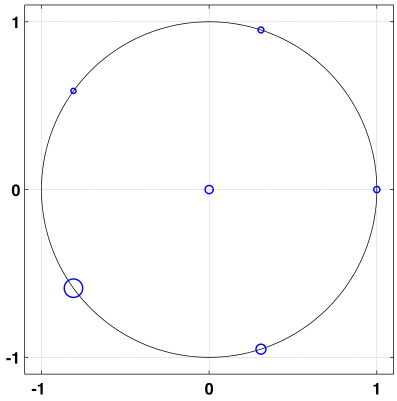


Figure 1: Balls in $\mathbb{C}$ enclosing the six eigenvalues of any $A \in \boldsymbol{A}$ for $rad = 1.3 \cdot 10^{-3}$.

for all $i = 1, \ldots, n$. Roughly speaking, if the value of rad increases, while the numerical solution is kept fixed, the norms of the components of the vectors Y and $Z^{(1)}$ increase. Thus there is a value of rad large enough such that some of the inequalities (8) are not satisfied anymore. To be more precise, assuming that $\bar{x}$ is a good numerical approximate solution, we can estimate $|f(\bar{x})|_\infty \approx |\bar{x}|_1 rad$, where $|x|_1 = \sum_i |x_i|$. Then we can write $|Y|_\infty \approx \|R\|_\infty |\bar{x}|_1 rad$. Concerning $Z^{(1)}$, we see from (2) that the radius of $\boldsymbol{Df}$ is the same as the radius of $\boldsymbol{A}$. Then, assuming that the matrix R has been properly computed so that $I - R \cdot \widehat{Df}(\bar{x}) \approx 0$, we estimate $|Z^{(1)}|_\infty \approx n\|R\|_\infty rad$. Finally $|Z^{(2)}|_\infty \approx 2\|R\|_\infty$. By substituting into (8), we obtain

$$rad_{max} = \frac{n + 4\|R\|_\infty |\bar{x}|_1 - 2\sqrt{4\|R\|_\infty^2 |\bar{x}|_1^2 + 2n\|R\|_\infty |\bar{x}|_1}}{n^2 \|R\|_\infty}.$$

For $\boldsymbol{A}$ considered above, rad_{max} are computed and presented in Table 3. Note there that the prediction of rad_{max} is more precise when the dimension n is larger.

Table 4 displays the results for the enclosure of two eigenpairs of $\boldsymbol{A}$ around $\hat{A} = XDX^{-1}$, where X is a random matrix and $D = diag(1, 2, \ldots, n)$ both for $n = 50$ and $n = 100$ and the value of rad_{max}.

#	0	1	2	3	4	5
rad_{max}	0.0016	0.0024	0.0028	0.0012	0.0014	0.0022

.

Table 3: rad_{max} as a function of $\bar{\lambda}_i$, for $i = 0, 1, \ldots, 5$.

$n = 50$

$\bar{\lambda}_{20} = 20$, $rad_{max} = 5.698 \cdot 10^{-6}$		$\bar{\lambda}_{47} = 47$, $rad_{max} = 1.701 \cdot 10^{-5}$	
rad	r	rad	r
$1 \cdot 10^{-7}$	$3.806 \cdot 10^{-5}$	$1 \cdot 10^{-7}$	$2.163 \cdot 10^{-5}$
$2 \cdot 10^{-6}$	$8.399 \cdot 10^{-4}$	$1 \cdot 10^{-5}$	$2.621 \cdot 10^{-3}$
$5 \cdot 10^{-6}$	$1.956 \cdot 10^{-3}$	$1.7 \cdot 10^{-5}$	$6.134 \cdot 10^{-3}$
$6 \cdot 10^{-6}$	-	$1.8 \cdot 10^{-5}$	-

$n = 100$

$\bar{\lambda}_5 = 5$, $rad_{max} = 2.753 \cdot 10^{-6}$		$\bar{\lambda}_{95} = 95$, $rad_{max} = 7.9715 \cdot 10^{-7}$	
rad	r	rad	r
$1 \cdot 10^{-7}$	$6.474 \cdot 10^{-5}$	$1 \cdot 10^{-7}$	$1.239 \cdot 10^{-4}$
$1 \cdot 10^{-6}$	$7.154 \cdot 10^{-4}$	$6 \cdot 10^{-7}$	$9.630 \cdot 10^{-4}$
$2 \cdot 10^{-6}$	$1.693 \cdot 10^{-3}$	$8 \cdot 10^{-7}$	$1.857 \cdot 10^{-3}$
$3 \cdot 10^{-6}$	-	$9 \cdot 10^{-7}$	-

Table 4: Test the theoretically derived rad_{max} to some rad used in computations.

Example 3: comparison. We now compare our method, denoted by radiipol, with two different algorithms developed by S. Rump. The first one, denoted by verifyeig, has been introduced in [14] with the primary goal of computing enclosures of multiple of nearly multiple eigenvalues (and related eigenvectors) of interval matrices. The second one, denoted by verifynlss, is based on a Krawczyk operator [15, 16] and is a general routine to rigorously compute well separated zeros of nonlinear functions. In fact, in the code `verifyeig.m` (available in the library INTLAB [19]), where the method verifyeig has been implemented, the author suggests to use verifynlss to compute simple and well separated eigenpairs. This method is implemented in the code `verifynlss.m` in the library INTLAB [19]. Table 5 provides the average of the radius of the balls enclosing the exact eigenpairs for each method.

For both experiments the test matrices $\boldsymbol{A}$ have been constructed as in the previous section: given N we define $D \in \mathbb{C}^{N+1,N+1}$ as a diagonal matrix with entries given by N equispaced values on the unit circle in the complex plane and 0, i.e. $diag(D) = [0, e^{i\frac{2\pi}{N}j}], j = 1, \ldots, N$. Then let $\hat{A} = XDX^{-1}$, where X is a complex random matrix with entries in the complex square $[-1, 1] + i[-1, 1]$ and finally define $\boldsymbol{A}$ as the interval complex matrix centered in $\hat{A}$ and of radius rad.

	N=5			N=10			
rad	10^{-20}	10^{-10}	10^{-4}	10^{-10}	10^{-5}	10^{-4}	10^{-3}
radiipol	$9.14{\cdot}10^{-15}$	$2.76{\cdot}10^{-9}$	0.0019	$3.25{\cdot}10^{-9}$	$4.61{\cdot}10^{-4}$	–	–
verifyeig	$4.69\cdot10^{-15}$	$2.07{\cdot}10^{-9}$	0.0016	$2.08{\cdot}10^{-9}$	$3.02{\cdot}10^{-4}$	0.0049	–
verifynlss	$6.26\cdot10^{-9}$	–	–	–	–	–	–

	N=50			N=100			N=150
rad	10^{-10}	10^{-8}	10^{-5}	10^{-10}	10^{-8}	10^{-7}	10^{-10}
radiipol	$2.69{\cdot}10^{-7}$	$4.94{\cdot}10^{-5}$	–	$9.02{\cdot}10^{-7}$	–	–	$1.31{\cdot}10^{-6}$
verifyeig	$5.59{\cdot}10^{-8}$	$9.45{\cdot}10^{-6}$	–	$1.31{\cdot}10^{-7}$	$2.07{\cdot}10^{-5}$	–	$1.64{\cdot}10^{-7}$
verifynlss	–	–	–	–	–	–	–

Table 5: Each number is the average of the radius of the disks enclosing the eigenvalues for each method. Comparison of the accuracy of the three methods as the dimension N and the radius rad of the test matrix $\boldsymbol{A}$ change. The entry $-$ means that the method fails in the enclosure of at least one of the eigenpair.

The results presented in Table 5 confirm that the new approach radiipol is satisfactory from the point of view of the accuracy of the results. Indeed, while the algorithm verifynlss fails quite soon as N and rad increase (it fails for $rad = 0$ and for all $N \geq 15$), the new algorithm is successful also for large entries of $\boldsymbol{A}$.

Acknowledgements

This work was supported by grant No. MTM2011–24766 of the MICINN, Spain.

References

[1] Yamamoto, T.: Error bounds for computed eigenvalues and eigenvectors. Numer. Math. **34**(2) (1980), 189–199.

[2] Yamamoto, T.: Error bounds for computed eigenvalues and eigenvectors. II. Numer. Math. **40**(2) (1982), 201–206.

[3] Rump, S. M. and Zemke, J.-P. M.: On eigenvector bounds. BIT **43**(4) (2003), 823–837.

[4] Bauer, F. L. and Fike, C. T.: Norms and exclusion theorems. Numer. Math. **2** (1960), 137–141.

[5] Sleijpen, G. L. G., van den Eshof, J., and Smit, P.: Optimal a priori error bounds for the Rayleigh-Ritz method. Math. Comp. **72**(242) (2003), 677–684.

[6] Beattie, C.: Harmonic Ritz and Lehmann bounds. Electron. Trans. Numer. Anal. **7** (1998), 18–39. Large scale eigenvalue problems (Argonne, IL, 1997).

[7] Parlett, B. N.: *The symmetric eigenvalue problem, volume 20 of Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1998. Corrected reprint of the 1980 original.

[8] Stewart, G. W.: *Matrix algorithms. Vol. II.* Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2001. Eigensystems.

[9] Stewart, G. W. and Sun, J. G.: *Matrix perturbation theory*. Computer Science and Scientific Computing, Academic Press Inc., Boston, MA, 1990.

[10] Alefeld, G. and Mayer, G.: Interval analysis: theory and applications. J. Comput. Appl. Math. **121**(1-2) (2000), 421–464. Numerical analysis in the 20th century, Vol. I, Approximation theory.

[11] Hladík, M., Daney, D., and Tsigaridas, E.: Bounds on real eigenvalues and singular values of interval matrices. SIAM J. Matrix Anal. Appl. **31**(4) (2009/10), 2116–2129.

[12] Mayer, G.: Result verification for eigenvectors and eigenvalues. In: *Topics in validated computations (Oldenburg, 1993)*, vol. 5 of *Stud. Comput. Math.*, pp. 209–276. North-Holland, Amsterdam, 1994.

[13] Alefeld, G. and Spreuer, H.: Iterative improvement of componentwise error bounds for invariant subspaces belonging to a double or nearly double eigenvalue. Computing **36**(4) (1986), 321–334.

[14] Rump, S. M.: Computational error bounds for multiple or nearly multiple eigenvalues. Linear Algebra Appl. **324**(1-3) (2001), 209–226. Special issue on linear algebra in self-validating methods.

[15] Krawczyk, R.: Newton-Algorithmen zur Bestimmung von Nullstellen mit Fehlerschranken. Computing (Arch. Elektron. Rechnen) **4** (1969), 187–201.

[16] Moore, R. E.: A test for existence of solutions to nonlinear systems. SIAM J. Numer. Anal. **14**(4) (1977), 611–615.

[17] Day, S., Lessard, J.-P., and Mischaikow, K.: Validated continuation for equilibria of PDEs. SIAM J. Numer. Anal. **45**(4) (2007), 1398–1424 (electronic).

[18] Gameiro, M. and Lessard, J.-P.: Rigorous computation of smooth branches of equilibria for the three dimensional Cahn-Hilliard equation. Numer. Math. **117**(4) (2011), 753–778.

[19] Rump, S. M.: INTLAB – INTerval LABoratory. In: T. Csendes (Ed.), *Developments in Reliable Computing*, pp. 77–104. Kluwer Academic Publishers, Dordrecht, 1999. `http://www.ti3.tu-harburg.de/rump/`.

Conference Applications of Mathematics 2013
in honor of the 70th birthday of Karel Segeth.
Jan Brandts, Sergey Korotov, Michal Křížek,
Jakub Šístek, and Tomáš Vejchodský (Eds.),
Institute of Mathematics AS CR, Prague 2013

HP-ANISOTROPIC MESH ADAPTATION TECHNIQUE BASED ON INTERPOLATION ERROR ESTIMATES

Vít Dolejší

Charles University Prague, Faculty of Mathematics and Physics
Sokolovská 83, 186 75, Prague, Czech Republic
dolejsi@karlin.mff.cuni.cz

Abstract

We present a completely new *hp*-anisotropic mesh adaptation technique for the numerical solution of partial differential equations with the aid of a discontinuous piecewise polynomial approximation. This approach generates general anisotropic triangular grids and the corresponding degrees of polynomial approximation based on the minimization of the interpolation error. We develop the theoretical background of this approach and present a numerical example demonstrating the efficiency of this anisotropic strategy in comparison with an isotropic one.

1. Introduction

Adaptive methods exhibit an efficient tool for the numerical solution of partial differential equations (PDEs). Our aim is to develop an adaptive technique which is able to generate general *hp*-anisotropic grids which can be employed in the framework of discontinuous Galerkin method based on a discontinuous piecewise polynomial approximation. The shape of an anisotropic element is extended in one dominant direction.

The *hp*-adaptive method allows the adaptation in the element size h as well as in the polynomial degree p. Several strategies of *hp*-adaptation have been proposed over the years, see, e.g., [14] or [11] for a survey. Based on many theoretical works, e.g., monographs [15] or papers [1, 5, 17] we expect that an error converges at an exponential rate in the number of degrees of freedom. However, most of *hp*-adaptive methods deal with h-isotropic refinement when the element marked for h-refinement is split (isotropically) into several (usually four in 2D) daughter elements. Some exception is, e.g., [13] where quadrilateral elements can be split onto two daughter elements by a line in a either vertical or horizontal direction.

Our goal is to generate anisotropic grids similarly to those ones developed, e.g., in [4, 6, 9, 12, 16], for the first order finite volume and finite element methods. In these works, the Hessian matrices (matrices of second order derivatives) are employed for the definition of a Riemann metric. Then the highly anisotropic triangular

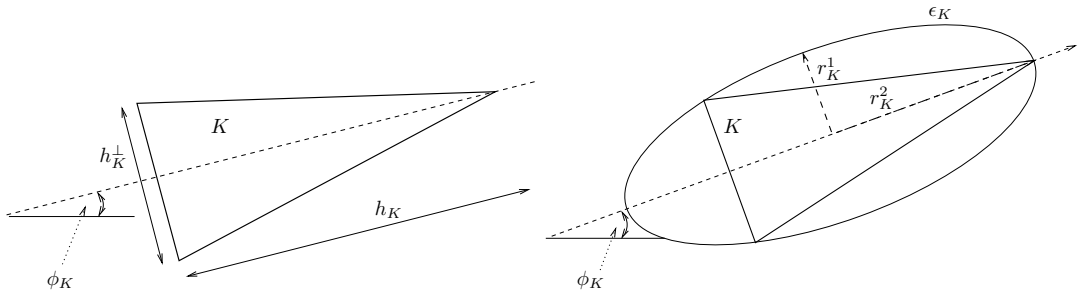


Figure 1: An anisotropic element K characterized by h_K, $h_K^\perp$ and ϕ_K (left), and an anisotropic element K characterized by r_K^1, r_K^2 and ϕ_K with the corresponding ellipse (right).

grids, which are quasi-uniform in this metric, are constructed. However, the Hessian matrices correspond to the interpolation error for a piecewise linear approximation. In [2, 3], the Riemann metric (defining the anisotropic mesh) is developed for a high degree of polynomial approximation. This approach is based on a particular definition of the magnitude, orientation, and anisotropic ratio for the higher order derivative of a function u to characterize its anisotropic behaviour. Being inspired by these papers, we develop here a new strategy which is able to generate anisotropic triangular grids and the corresponding degree of polynomial approximation for each element of the mesh. This approach is based on the approximation of the interpolation error in the L^∞-norm by the leading terms of the Taylor expansion. The aim is to keep the interpolation error under a given tolerance and to minimize the number of degree of freedom.

2. An anisotropic element

In this section, we describe an anisotropy of triangles in a plane domain. Let $K \subset \mathbb{R}^2$ be an acute isosceles triangle, see Figure 1, left. By h_K we denote its size in the direction of its axis, $h_K^\perp$ denotes its size in the direction perpendicular of its axis and $\phi_K \in [0, \pi)$ denotes the angle between its axis and the axis x_1, see Figure 1, left. The triple $(h_K, h_K^\perp, \phi_K)$ defines the *anisotropy* of element K.

We can define the anisotropy in an alternative way. Let $\lambda_K^1 > 0$, $\lambda_K^2 > 0$, and $\phi_K \in [0, \pi)$. We define the matrix M_K by

$$M_K := R^{\mathrm{T}}(\phi_K) \begin{pmatrix} \lambda_K^1 & 0 \\ 0 & \lambda_K^2 \end{pmatrix} R(\phi_K) = \begin{pmatrix} a_K & b_K \\ b_K & c_K \end{pmatrix}, \tag{1}$$

where $R(\phi_K)$ is the the rotation matrix

$$R(\phi_K) := \begin{pmatrix} \cos\phi_K & -\sin\phi_K \\ \sin\phi_K & \cos\phi_K \end{pmatrix} \tag{2}$$

and $R^{\mathrm{T}}(\phi_K)$ is its transpose matrix. Obviously, M_K is a symmetric positive definite matrix having eigenvalues λ_K^1, λ_K^2. The equation

$$x^{\mathrm{T}} M_K x = a_K x_1^2 + 2b_K x_1 x_2 + c_K x_2^2 \leq 1, \quad x = (x_1, x_2) \in \mathbb{R}^2, \tag{3}$$

defines an ellipse ϵ_K with the centre at origin, the semi-axes lengths

$$r_K^1 = 1/\sqrt{\lambda_K^1}, \quad r_K^2 = 1/\sqrt{\lambda_K^2} \tag{4}$$

and the angle between the axis x_1 and the major axis of ϵ_K is ϕ_K, see Figure 1, right.

Let K denotes an acute isosceles triangle which is inscribed into ellipse ϵ_K and which has the maximal possible area, see Figure 1, right. We say that K *is generated by* M_k. Hence, the anisotropy of this triangle K can be defined by the triple $(\lambda_K^1, \lambda_K^2, \phi_K)$ or the triple (r_K^1, r_K^2, ϕ_K). With the aid of techniques [7], we can derive direct relations between triples $(h_K, h_K^\perp, \phi_K)$ and $(\lambda_K^1, \lambda_K^2, \phi_K)$ (or (r_K^1, r_K^2, ϕ_K)). Namely, $h_K = \frac{3}{2} r_K^2$ and $h_K^\perp = 2\sqrt{3} r_K^1$.

Let $\boldsymbol{e}_i,\ i = 1, 2, 3$ denote the edges of the triangle K inscribed into ϵ_K and having the maximal area. The edges $\boldsymbol{e}_i,\ i = 1, 2, 3$ are considered as vectors from $\mathbb{R}^2$ given by their endpoints. In [6] we proved that

$$\|\boldsymbol{e}_i\|_{M_K} = \sqrt{3}, \quad i = 1, 2, 3, \tag{5}$$

where $\|\boldsymbol{e}_i\|_{M_K} := \left(\boldsymbol{e}_i^{\mathrm{T}} M_K \boldsymbol{e}_i\right)^{1/2}$ is the size of $\boldsymbol{e}_i$ in the Riemann metric generated by M_K, compare with Definition 3.1 bellow. Hence, K is the *equilateral triangle* in the *metric* generated by M_K.

3. *hp*-anisotropic meshes

Let the computational domain $\Omega \subset \mathbb{R}^2$ be bounded with a polygonal boundary $\partial\Omega$. Let $\mathscr{T}_h$ $(h > 0)$ be a partition of the closure $\overline{\Omega}$ of the domain Ω into a finite number of closed triangles K with mutually disjoint interiors. We call $\mathscr{T}_h = \{K\}_{K \in \mathscr{T}_h}$ a *triangulation* of Ω and assume that $\mathscr{T}_h$ is conforming.

Moreover, to each $K \in \mathscr{T}_h$, we assign a positive integer p_K (=local polynomial degree of polynomial approximation on K). Then we define the set $\mathbf{p} := \{p_K; K \in \mathscr{T}_h\}$ and the pair

$$\mathscr{T}_{h\mathbf{p}} := \{\mathscr{T}_h, \mathbf{p}\} \tag{6}$$

is called the *hp-mesh*.

For the given *hp*-mesh $\mathscr{T}_{h\mathbf{p}}$, we construct the space of piecewise polynomial discontinuous functions by

$$S_{h\mathbf{p}} = \{v \in L^2(\Omega);\ v|_K \in P^{p_K}(K)\ \forall K \in \mathscr{T}_h\}, \tag{7}$$

where $P^{p_K}(K)$ is the space of polynomials or degree $\leq p_K$ on $K \in \mathscr{T}_h$. The dimension of $S_{h\mathbf{p}}$ can be expressed (for two-dimensional domain) by

$$N_{h\mathbf{p}} := \sum_{K \in \mathscr{T}_h} (p_K + 1)(p_K + 2)/2. \tag{8}$$

We call this quantity the *size* of the *hp*-mesh $\mathscr{T}_{h\mathbf{p}}$.

Finally, by $\mathscr{F}_h$ we denote the set of edges of $\mathscr{T}_h$. Here the edges $\boldsymbol{e} \in \mathscr{F}_h$ are considered as vectors from $\mathbb{R}^2$ given by its endpoints. The orientation of the edges is arbitrary.

Similarly as in [4, 6, 9, 12, 16], we define the anisotropic triangular grid as a quasi-uniform grid in a Riemann metric.

Definition 3.1. *Let $\mathbf{M} : \Omega \to \mathbb{R}^{2\times 2}$ be a continuous mapping such that for each $x \in \Omega$, the matrix $\mathbf{M}(x)$ is symmetric and positive definite. Moreover, let $\boldsymbol{v}_0, \boldsymbol{v}_1 \in \mathbb{R}^2$ such that $\boldsymbol{v}_0 \in \Omega$ and $\boldsymbol{v}_0 + \boldsymbol{v}_1 \in \Omega$. The mapping $\boldsymbol{v} : [0,1] \to \mathbb{R}^2$, $\boldsymbol{v}(t) = \boldsymbol{v}_0 + t\boldsymbol{v}_1$, $t \in [0,1]$ defines a straight edge in Ω. Furthermore, we set*

$$\|\boldsymbol{v}\|_{\mathbf{M}} := \int_0^1 \left(\boldsymbol{v}'(t)^{\mathrm{T}}\mathbf{M}(\boldsymbol{v}_0 + t\boldsymbol{v}_1)\boldsymbol{v}'(t)\right)^{1/2} \mathrm{d}t = \int_0^1 \left(\boldsymbol{v}_1^{\mathrm{T}}\mathbf{M}(\boldsymbol{v}_0 + t\boldsymbol{v}_1)\boldsymbol{v}_1\right)^{1/2} \mathrm{d}t. \quad (9)$$

We call $\mathbf{M}$ the Riemann metric *on Ω and $\|\boldsymbol{v}\|_{\mathbf{M}}$ defines the* size *of edge $\boldsymbol{v}$ in the* Riemann metric $\mathbf{M}$.

Remark 3.2. *Let us note that if $\mathbf{M}$ is constant along $\boldsymbol{v}$ then* (9) *reduces to $\|\boldsymbol{v}\|_{\mathbf{M}} = (\boldsymbol{v}_1^{\mathrm{T}}\mathbf{M}\boldsymbol{v}_1)^{1/2}$. Moreover, if $\mathbf{M}(x) = \mathbb{I}\ \forall x \in \boldsymbol{v}$ ($\mathbb{I}$= the identity matrix) then the size of $\boldsymbol{v}$ in the Riemann metric $\mathbf{M}$ is equal to its length in the Euclidean metric.*

In virtue of (5), we define a triangulation corresponding to the metric $\mathbf{M}$.

Definition 3.3. *Let $\omega > 0$ be a given constant. Let $\mathbf{M}$ be the Riemann metric defined on Ω, $\mathscr{T}_h$ be a triangulation of Ω and $\mathscr{F}_h$ the corresponding set of edges. We say that the triangulation $\mathscr{T}_h$ is* generated by metric $\mathbf{M}$ *if*

$$\|\boldsymbol{e}\|_{\mathbf{M}} = \omega \quad \forall \boldsymbol{e} \in \mathscr{F}_h. \quad (10)$$

Remark 3.4. *For the given metric $\mathbf{M}$, there does not exist (except special cases) any triangulation generated by $\mathbf{M}$ in virtue of Definition 3.3. However, we can construct a triangulation which satisfies* (10) *approximately by the* least square technique*, see [6, 9]. Therefore, we replace* (10) *by $\|\boldsymbol{e}\|_{\mathbf{M}} \approx \omega\ \forall \boldsymbol{e} \in \mathscr{F}_h$ in the sense of the least square method. Moreover, let us note that for practical reasons, it is sufficient to evaluate the metric $\mathbf{M}$ only in a finite number of nodes $x \in \Omega$.*

Finally, let $\mathcal{P} : \Omega \to [0, \infty)$ be a given function. We define

$$p_K := \mathrm{int}\left[\frac{1}{|K|}\int_K \mathcal{P}(x)\,\mathrm{d}x\right], \quad K \in \mathscr{T}_h, \quad (11)$$

where $\mathrm{int}[a] := \lfloor a + 1/2 \rfloor$ denotes the integer part of the number $a + 1/2$, $a \geq 0$. We call $\mathcal{P}$ the *polynomial degree distribution function*.

We conclude that for the given Riemann metric $\mathbf{M}$ and for the given polynomial degree distribution function $\mathcal{P}$, there exists a *hp*-mesh $\mathscr{T}_{h\mathbf{p}} = \{\mathscr{T}_h, \mathbf{p}\}$, where $\mathscr{T}_h$ is given by Definition 3.3 in the sense of Remark 3.4 and $\mathbf{p}$ by (11). Our aim is to define the metric $\mathbf{M}$ and the polynomial degree distribution function $\mathcal{P}$ such that the corresponding *hp*-mesh is optimal in the sense specified later.

4. Interpolation error

For simplicity, we deal with the space of functions $V := C^\infty(\Omega)$. Let $\bar{x} = (x_1, x_2) \in \Omega$ be arbitrary but fixed. Let $p > 0$ be an integer, we define the interpolation operator $\Pi_{hp} : V \to P^p(\bar{\Omega})$ such that

$$\frac{\partial^k}{\partial x_1^l \partial x_2^{k-l}} \Pi_{hp} u(\bar{x}) = \frac{\partial^k}{\partial x_1^l \partial x_2^{k-l}} u(\bar{x}) \qquad \begin{matrix} \forall l = 0, \dots, k\,, \\ \forall k = 0, \dots, p\,. \end{matrix} \tag{12}$$

Therefore, $\Pi_{hp} u$ is the polynomial function of degree p on Ω which has the same value and the same values of all partial derivatives up to order p at $\bar{x}$ as the function u.

Using the Taylor expansion at $\bar{x} = (\bar{x}_1, \bar{x}_2)$, we have

$$u(x) = \sum_{k=0}^{p+1} \frac{1}{k!} \left(\sum_{l=0}^{k} \binom{k}{l} \frac{\partial^k u(\bar{x})}{\partial x_1^l \partial x_2^{k-l}} (x_1 - \bar{x}_1)^l (x_2 - \bar{x}_2)^{k-l} \right) + O(|x - \bar{x}|^{p+2}), \tag{13}$$

where $\binom{k}{l} = \frac{k!}{l\,!(k-l)!}$. From (12) and (13) we obtain

$$u(x) - \Pi_{hp} u(x) = E_{\mathrm{I}}^p(x) + O(|x - \bar{x}|^{p+2}), \tag{14}$$

where

$$E_{\mathrm{I}}^p(x) := \frac{1}{(p+1)!} \sum_{l=0}^{p+1} \left[\binom{p+1}{l} \frac{\partial^{p+1} u(\bar{x})}{\partial x_1^l \partial x_2^{p+1-l}} (x_1 - \bar{x}_1)^l (x_2 - \bar{x}_2)^{p+1-l} \right] \tag{15}$$

is the *interpolation error function* of degree $p = 0, 1, \dots$.

At this point, we consider the following task: Let $u \in V$, $\bar{x} \in \Omega$, $\omega > 0$ and $p > 0$ be given, we seek a triangle K' with barycentre at $\bar{x}$ such that

(C1) $E_{\mathrm{I}}^p(x) \le \omega$ for all $x \in K'$,

(C2) the area (two-dimensional Lebesgue measure) of K' is maximal.

The condition (C2) follows from the observation that a mesh having the maximal possible triangles has a small number of degree of freedom.

Let $B_1 := \{\xi;\ \xi = (\xi_1, \xi_2) \in \mathbb{R}^2,\ \xi_1^2 + \xi_2^2 = 1\}$ denote the unit sphere (in the Euclidean metric) in $\mathbb{R}^2$. We define the k^{th}-(scaled) *directional derivative* of $u \in V$ in $x \in \Omega$ and in the direction ξ by

$$\mathrm{d}^k u(x; \xi) := \frac{1}{k!} \sum_{l=0}^{k} \binom{k}{l} \frac{\partial^k u(x)}{\partial x_1^l \partial x_2^{k-l}} \xi_1^l \xi_2^{k-l}, \quad x \in \Omega,\ \xi = (\xi_1, \xi_2) \in B_1. \tag{16}$$

Therefore, from (15) and (16), we have

$$E_{\mathrm{I}}^p(x) = \mathrm{d}^{p+1} u \left(\bar{x}; \frac{x - \bar{x}}{|x - \bar{x}|} \right) |x - \bar{x}|^{p+1}, \quad p = 0, 1, \dots,\ x \in \Omega. \tag{17}$$

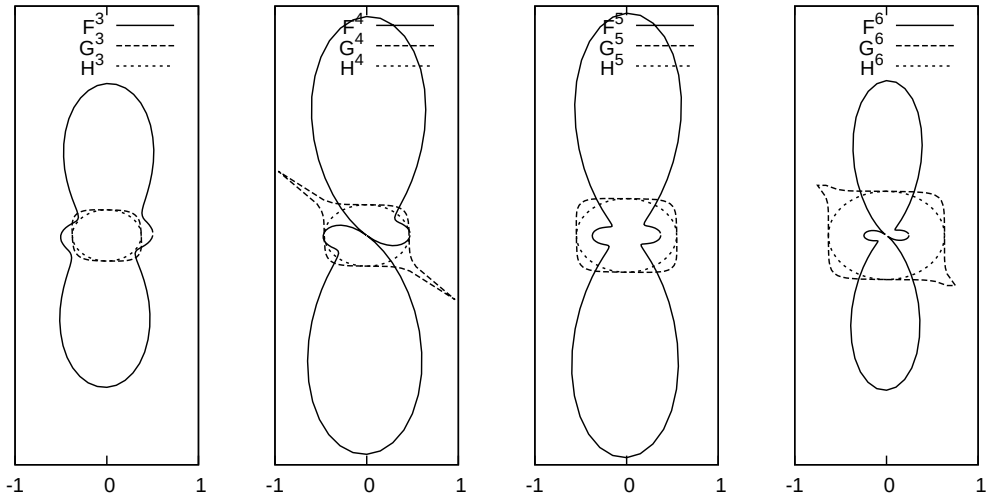


Figure 2: The curve F^p, the domain G^p and the ellipse H^p for $p = 3, 4, 5, 6$, $\bar{x} = (0, 0)$ and the function u given by (20).

Let $u \in V$, $\bar{x} \in \Omega$, $\omega > 0$ and $p > 0$ be given. We define the sets

$$F^p := \left\{ x \in \mathbb{R}^2;\ x = \bar{x} + \xi \left| \mathrm{d}^{p+1} u(\bar{x}; \xi) \right|,\ \xi \in B_1 \right\}, \tag{18}$$

$$G^p := \left\{ x \in \mathbb{R}^2;\ x = \bar{x} + t\xi \left(\frac{\omega}{|\mathrm{d}^{p+1} u(\bar{x}; \xi)|} \right)^{\frac{1}{p+1}},\ t \in [0; 1],\ \xi \in B_1 \right\}, \tag{19}$$

where $p = 1, 2, \ldots$. If $x \in F^p$ then the directional derivative $\mathrm{d}^{p+1} u(\bar{x}, \cdot)$ in the direction $(x - \bar{x})/|x - \bar{x}|$ is equal to $|x - \bar{x}|$. Moreover, in virtue of (17) and (19), G^p is the set such that $E^p_{\mathrm{I}}(x) \leq \omega\ \forall x \in G^p$. The set F^p is one-dimensional continuous curve in $\mathbb{R}^2$ whereas G^p is two dimensional sub-domain of $\mathbb{R}^2$ (it may be unbounded if $\mathrm{d}^{p+1} u(\bar{x}; \xi) = 0$ for some ξ). Figure 2 shows the curve F^p and the domain G^p for $p = 3, 4, 5, 6$, $\bar{x} = (0, 0)$ and the function

$$u(x_1, x_2) = 10x_1^{10} + 2x_1^{10}x_2^6 + x_1^9 x_2 + 2x_1^8 x_2^3 - x_1^7 x_2^5 + 8x_1^4 x_2^6 + 2x_2^{10}. \tag{20}$$

From (19) we find that if K is a triangle with the barycentre $\bar{x}$ such that $K \subset G^p$ for some p then $E^p_{\mathrm{I}}(x) \leq \omega$ for all $x \in K$. In order to minimize the number of degree of freedom of $S_{h_{\mathbf{p}}}$, the aim is to have triangle K such that $K \subset G^p$ and K has the maximal possible area.

5. Definition of the metric

In the following, with the aid of the results from Section 4, we define the Riemann metric $\mathbf{M}$ and the polynomial degree distribution function $\mathcal{P}$ introduced in Section 3.

Let $\bar{x} \in \Omega$, $u \in V$ and $p \geq 1$. Let $\xi_p^{\max} \in B_1$ be the direction which maximizes $|\mathrm{d}^p u(\bar{x};\xi)|$ and $\xi_p^{\perp}$ the direction orthogonal, i.e,

$$\xi_p^{\max} := \arg\max_{\xi \in B_1} |\mathrm{d}^p u(\bar{x};\xi)|, \quad \xi_p^{\perp} \in B_1, \quad \xi_p^{\max} \cdot \xi_p^{\perp} = 0. \tag{21}$$

Then we define quantities

$$h_p^{\max} := \left(\frac{\omega}{\left|\mathrm{d}^{p+1} u(\bar{x};\xi_p^{\max})\right|} \right)^{1/(p+1)}, \quad h_p^{\min} := \left(\frac{\omega}{\left|\mathrm{d}^{p+1} u(\bar{x};\xi_p^{\perp})\right|} \right)^{1/(p+1)}. \tag{22}$$

Let us note that $h_p^{\max} \leq h_p^{\min}$. Moreover, let $\phi_p \in [0, 2\pi)$ be such that $\xi_p^{\max} = (\cos\phi_p, \sin\phi_p) \in B_1$. Hence, the triple

$$\{h_p^{\min}, h_p^{\max}, \phi_p\} \tag{23}$$

defines the ellipse H^p which touches G^p at the nearest point to $\bar{x}$, see Figure 2. Moreover, we have observed experimentally that H^p is almost included in G^p.

Therefore, in virtue of (1), (4) and Definition 3.1, we define the metric $\mathbf{M}$ at $\bar{x}$ by $\mathbf{M}(\bar{x}) := M_p$, where

$$M_p := R^{\mathrm{T}}(\phi_p) \begin{pmatrix} 1/(h_p^{\max})^2 & 0 \\ 0 & 1/(h_p^{\min})^2 \end{pmatrix} R(\phi_p), \quad K \in \mathscr{T}_h,\ p \geq 1, \tag{24}$$

and $R(\phi_p)$ is given by (2).

Finally, we have to define the polynomial degree distribution function $\mathcal{P}(x)$ at $\bar{x} \in \Omega$. For each integer $p \geq 1$ we have matrix $\mathbf{M}(\bar{x}) := M_p$. We seek some criterion choosing giving the optimal degree of polynomial approximation p. The aim is to minimize $N_{h\mathbf{p}}$ (=size of the hp-mesh). The area of the element K generated by M_p is proportional to the area of the ellipse defined by relation $\xi^{\mathrm{T}} M_p \xi = 1$, $\xi \in B_1$, namely $|K| = (2\sqrt{3}/2) h_p^{\max} h_p^{\min}$. If $|K|$ is an average volume of triangles from $\mathscr{T}_h$ then we need approximately $\lfloor |\Omega|/|K| \rfloor$ triangles. If p is the degree of polynomial approximation, the total number of freedom for one element is $(p+1)(p+2)/2$ and the value $N_{h\mathbf{p}}$ can be estimated (up to a constant)

$$N_{h\mathbf{p}} \approx \frac{(p+1)(p+2)}{2} \frac{|\Omega|}{|K|}. \tag{25}$$

Then we deduce that in order to minimize $N_{h\mathbf{p}}$, we need to choose the degree of polynomial approximation p such that

$$\mathcal{P}(\bar{x}) = \arg\min_{p=1,2,\ldots} \frac{(p+1)(p+2)}{h_p^{\max} h_p^{\min}}. \tag{26}$$

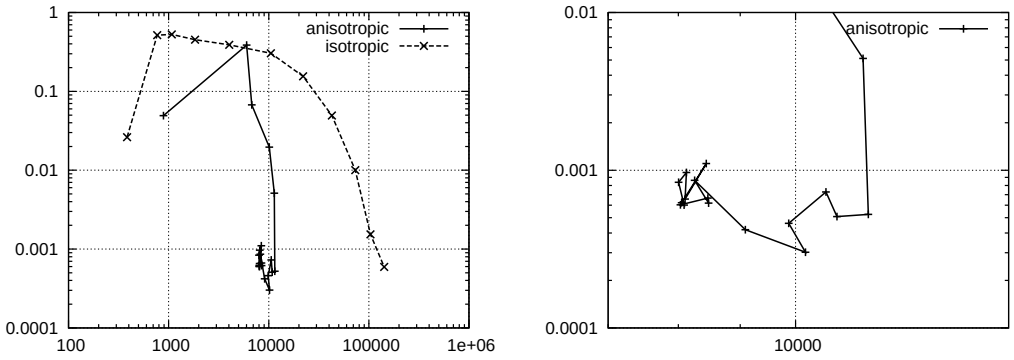


Figure 3: Comparison of the isotropic and the anisotropic *hp*-adaptation, the dependence of the error in the X-norm with respect to the degree of freedom $N_{h\mathbf{p}}$, the total view (left) and the detail (right).

6. Numerical implementation

In Sections 2–5, we developed the method which defines the metric $\mathbf{M}(x)$ and the polynomial degree distribution function $\mathcal{P}(x)$ for $x \in \Omega$. Hence, in virtue of the conclusion of Section 3, we have defined the *hp*-mesh for a given function $u \in V_h$.

The aim is to employ this strategy for the numerical solution of partial differential equations. Since the exact solution u is unknown, the natural approach is to apply the previous *hp*-anisotropic mesh adaptation method to some smoothing of the approximate solution $u_{h\mathbf{p}} \in S_{h\mathbf{p}}$. We obtain iteratively better and better *hp*-grids and the corresponding approximate solutions. Moreover, for practical computation, it is not necessary to evaluate $\mathbf{M}(x)$ and $\mathcal{P}(x)$ for all $x \in \Omega$. It is enough to compute $\mathbf{M}(x_K)$ and $\mathcal{P}(x_K)$ for all elements K of the given mesh (x_K is the barycentre of K), similarly as in [6, 9].

We demonstrate the potential of the proposed *hp*-anisotropic mesh adaptation method by a comparison with the isotropic *hp*-adaptation method presented in [8]. We consider the scalar linear convection-diffusion equation (similarly as in [10])

$$-\varepsilon\triangle u - \frac{\partial u}{\partial x_1} - \frac{\partial u}{\partial x_2} = g \quad \text{in } \Omega := (0,1)^2, \tag{27}$$

where $\varepsilon > 0$ is a constant diffusion coefficient. We prescribe a Dirichlet boundary condition on $\partial\Omega$ and the source term g such that the exact solution has the form $u(x_1, x_2) = \left(c_1 + c_2(1-x_1) + \mathrm{e}^{-x_1/\varepsilon}\right)\left(c_1 + c_2(1-x_2) + \mathrm{e}^{-x_2/\varepsilon}\right)$ with $c_1 = -\mathrm{e}^{-1/\varepsilon}$, $c_2 = -1 - c_1$. The solution contains two boundary layers along $x_1 = 0$ and $x_2 = 0$, whose width is proportional to ε. Here we consider $\varepsilon = 10^{-3}$.

We solve (27) with the aid of discontinuous Galerkin method with an interior penalty. Figure 3 shows the convergence of the computational error in the norm $\|\cdot\|_X^2 := \|\cdot\|_{L^2(\Omega)}^2 + \varepsilon|\cdot|_{H^1(\Omega)}^2$ with respect to the number of degree of freedom. We observe that the *hp*-anisotropic mesh adaptation is more efficient. Moreover, the

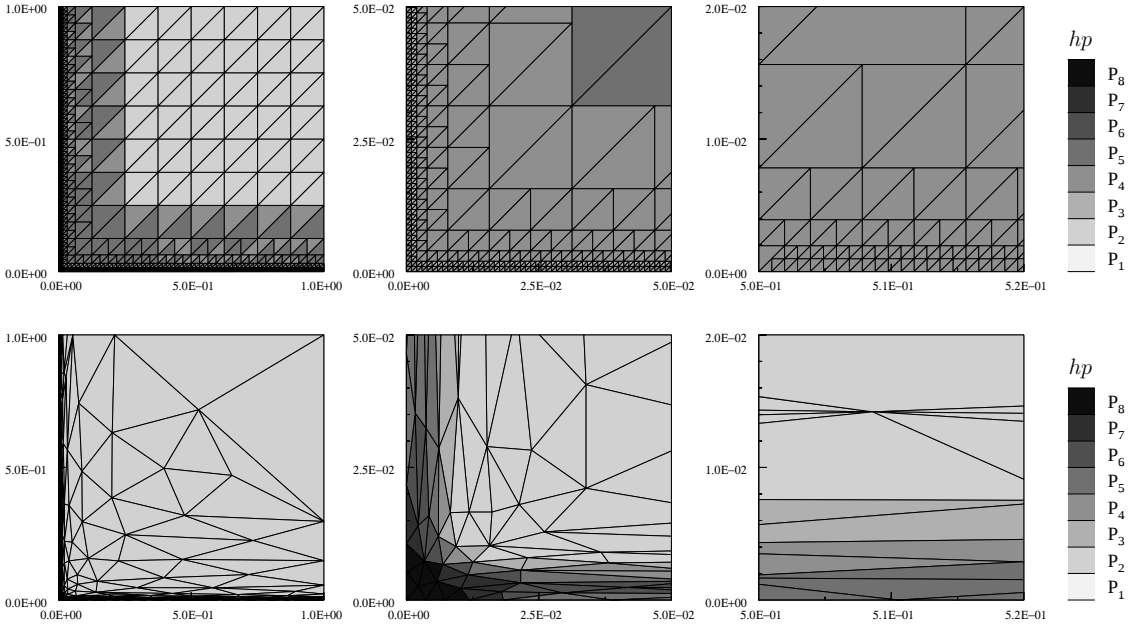


Figure 4: Example (E1): the final *hp*-meshes obtained be the isotropic (top) and the anisotropic (bottom) *hp*-adaptation, the total view (left), the detail around the corner (centre) and the detail of the boundary layer (right).

proposed technique is able to reduce the number of degree of freedom and to keep the level of the computational error during the optimization of the *hp*-mesh. Figure 4 shows the final grids obtained by the isotropic and the anisotropic technique.

Acknowledgements

This work was supported by grant No. 13-00522S of the Czech Science Foundation.

References

[1] Babuška, I. and Suri, M.: The $p-$ and $hp-$FEM a survey. SIAM Review **36** (1994), 578–632.

[2] Cao, W.: Anisotropic measures of third order derivatives and the quadratic interpolation error on triangular elements. SIAM J. Sci. Comput. **29** (2007), 756–781.

[3] Cao, W.: An interpolation error estimate in R^2 based on the anisotropic measures of higher order derivatives. Math. Comp. **77** (2008), 265–286.

[4] Castro-Díaz, M. J., Borouchaki, H., George, P. L., Hecht, F., and Mohammadi, B.: Anisotropic Adaptive Mesch Generation in Two Dimensions for CFD. In: J.A. Désidéri, C. Hirsch, P. Le Tallec, M. Pandolfi, and J. Périaux (Eds.), *Computational Fluid Dynamics '96*. Wiley, Chichester, Paris, 1996 pp. 181–186.

[5] Demkowicz, L., Rachowicz, W., and Devloo, P.: A fully automatic *hp*-adaptivity. J. Sci. Comput. **17** (2002), 117–142.

[6] Dolejší, V.: Anisotropic mesh adaptation for finite volume and finite element methods on triangular meshes. Comput. Vis. Sci. **1** (1998), 165–178.

[7] Dolejší, V.: *Adaptive higher order methods for compressible flow*, chap. Anisotropic mesh adaptation method. Charles University Prague, Faculty of Mathematics and Physics, 2003. Habilitation thesis.

[8] Dolejší, V.: *hp*-DGFEM for nonlinear convection-diffusion problems. Math. Comput. Simul. (submitted). Preprint No. MATH-knm-2012/2, Charles University Prague, www.karlin.mff.cuni.cz/ms-preprints.

[9] Dolejší, V. and Felcman, J.: Anisotropic mesh adaptation and its application for scalar diffusion equations. Numer. Methods Partial Differential Equations **20** (2004), 576–608.

[10] Dolejší, V. and Roos, H. G.: BDF-FEM for parabolic singularly perturbed problems with exponential layers on layer-adapted meshes in space. Neural Parallel Sci. Comput. **18** (2010), 221–235.

[11] Eibner, T. and Melenk, J. M.: An adaptive strategy for *hp*-FEM based on testing for analyticity. Comput. Mech. **39** (2007), 575–595.

[12] Fortin, M., Vallet, M. G., Dompierre, J., Bourgault, Y., and Habashi, W. G.: Anisotropic Mesh Adaptation: Theory, Validation and Applications. In: J. A. Désidéri, C. Hirsch, P. Le Tallec, M. Pandolfi, and J. Périaux (Eds.), *Computational Fluid Dynamics '96*. Wiley, Chichester, Paris, 1996 pp. 174–180.

[13] Giani, S. and Houston, P.: Anisotropic *hp*-adaptive discontinuous Galerkin finite element methods for compressible fluid flows. International Journal of Numerical Analysis and Modeling **9** (2012), 928–949.

[14] Houston, P. and Süli, E.: A note on the design of *hp*–adaptive finite element methods for elliptic partial differential equations. Comput. Methods Appl. Mech. Engrg. **194** (2005), 229–243.

[15] Schwab, C.: *p- and hp-Finite Element Methods*. Clarendon Press, Oxford, 1998.

[16] Simpson, R. B.: Anisotropic mesh transformations and optimal error control. Appl. Numer. Math. **14** (1994), 183–198.

[17] Šolín, P. and Demkowicz, L.: Goal-oriented *hp*-adaptivity for elliptic problems. Comput. Methods Appl. Mech. Engrg. **193** (2004), 449–468.

Conference Applications of Mathematics 2013
in honor of the 70th birthday of Karel Segeth.
Jan Brandts, Sergey Korotov, Michal Křížek,
Jakub Šístek, and Tomáš Vejchodský (Eds.),
Institute of Mathematics AS CR, Prague 2013

CONVERGENCE AND STABILITY CONSTANT OF THE THETA-METHOD

István Faragó
Eötvös Loránd University, Institute of Mathematics and
HAS-ELTE Numerical Analysis and Large Networks Research Group
Pázmány P. s. 1/c, 1117 Budapest, Hungary
faragois@cs.elte.hu

Abstract

The Euler methods are the most popular, simplest and widely used methods for the solution of the Cauchy problem for the first order ODE. The simplest and usual generalization of these methods are the so called theta-methods (notated also as θ-methods), which are, in fact, the convex linear combination of the two basic variants of the Euler methods, namely of the explicit Euler method (EEM) and of the implicit Euler method (IEM). This family of the methods is well-known and it is introduced almost in any arbitrary textbook of the numerical analysis, and their consistency is given. However, in its qualitative investigation the convergence is proven for the EEM, only, almost everywhere. At the same time, for the rest of the methods it is usually missed (e.g., [1, 2, 7, 8]). While the consistency is investigated, the stability (and hence, the convergence) property is usually shown as a consequence of some more general theory. In this communication we will present an easy and elementary prove for the convergence of the general methods for the scalar ODE problem. This proof is direct and it is available for the non-specialists, too.

1. Motivation and basic of the theta-method

Many different problems (physical, chemical, etc.) can be described by the initial-value problem for first order ordinary differential equation (ODE) of the form

$$\frac{du}{dt} = f(t,u), \quad t \in (0,T), \tag{1}$$

$$u(0) = u_0. \tag{2}$$

We note that, using the semidiscretization, the time-dependent partial differential equations also lead to the problem (1)–(2). Hence, the solution of such problem plays a crucial role in mathematical modelling. (For simplicity, in sequel we consider only the scalar problem, i.e., when $f : \mathbb{R}^2 \to \mathbb{R}$.) We know that under the global Lipshitz condition, i.e., in case

$$|f(t,s_1) - f(t,s_2)| \le L|s_1 - s_2| \quad \text{for all } (t,s_1),(t,s_2) \in \mathrm{dom}(f) \tag{3}$$

with the Lipschitz constant $L > 0$, the problem (1)–(2) has unique solution on the entire domain $\mathrm{dom}(f)$.

Since we have no hope of solving the vast majority of differential equations in explicit, analytic form, the design of suitable numerical algorithms for accurately approximating solutions is essential. The ubiquity of differential equations throughout mathematics and its applications has driven the tremendous research effort devoted to numerical solution schemes, some dating back to the beginnings of the calculus. Therefore, we apply some numerical method. Hence, the numerical integration of the problem (1)–(2) – under the condition (3) – is one of the most typical tasks in the numerical modelling of real-life problems.

Our aim is to define some numerical solution at some fixed point $t^\star \in (0, T)$ to the Cauchy problem (1)–(2). Therefore, we construct the sequence of the uniform meshes with the mesh-size $h = t^\star/N$ of the form

$$\omega_h = \{t_n = n \cdot h, n = 0, 1, \ldots, N\},$$

and our aim is to define at the mesh-point $t^\star = t_N$ a suitable approximation y_N on each fixed mesh.

This requires to give the rule how to define the mesh-function $y_h : \omega_h \to \mathbb{R}$. The most popular, simplest and widely used method are the so-called single step (one-step) schemes, particularly, the theta-method, which is frequently notated as θ-method. Using the notation $y_h(t_n) = y_n$, the θ-method is defined as

$$\begin{aligned} &y_n = y_{n-1} + h\left(\theta f(t_n, y_n) + (1-\theta) f(t_{n-1}, y_{n-1})\right), \quad n = 1, \ldots, N, \\ &y_0 = u_0. \end{aligned} \tag{4}$$

Here $\theta \in [0, 1]$ is a fixed parameter, and, it is for $\theta = 0$ explicit, otherwise implicit method. The θ-method is considered here as basic method since it represents the most simple Runge-Kutta method (and also linear multistep method). For stiff systems the cases $\theta = 0.5$ trapezoidal rule and $\theta = 1$ implicit (backward) Euler are of practical interest, for non-stiff systems we can also consider $\theta = 0$ explicit (forward) Euler.

In mathematics and computational science, these methods are most basic method for numerical integration of ordinary differential equations and they are the simplest Runge-Kutta methods.

Let us define the local truncation error for the θ-method, under the assumption that f (and hence, the solution $u(t)$) is sufficiently smooth.

As it is known, the local truncation error $l_n(h)$ for the θ-method can be defined as

$$l_n(h) = u(t_n) - u(t_{n-1}) - h\theta f(t_n, u(t_n)) - h(1-\theta) f(t_{n-1}, u(t_{n-1})), \tag{5}$$

where $u(t)$ stands for the solution of the problem (1)–(2). Therefore, we have the relation

$$l_n(h) = u(t_n) - u(t_{n-1}) - h\theta u'(t_n) - h(1-\theta) u'(t_{n-1}). \tag{6}$$

Hence, by expanding $u(t_n) = u(t_{n-1} + h)$ and $u'(t_n) = u'(t_{n-1} + h)$ into the Taylor series around the point $t = t_{n-1}$, we get for the local approximation error the relation

$$l_n(h) = (1/2 - \theta) h^2 u''(t_{n-1}) + (1/6 - \theta/2) h^3 u'''(t_{n-1}) + \mathcal{O}(h^4). \tag{7}$$

The order of a numerical method is defined by the local truncation error: when $l_n(h) = \mathcal{O}(h^{p+1})$ then the method is called consistent of order p. This means that for both Euler methods ($\theta = 0$ and $\theta = 1$) the order of consistency is equal to one, while for the trapezoidal rule ($\theta = 0.5$) the order of consistency is equal to two.

However, as it is well-known, the consistency itself does not guarantee the convergence of a numerical method, the stability is also required.

Roughly speaking, the consistency is the characterization of the local (truncation) error of the method, which is the error committed by one step of the method. (That is, it is the difference between the result given by the method, assuming that no error was made in earlier steps and hence having the exact solution.) On the other hand, the stability guarantees that the numerical method produces a bounded solution whenever the solution of the exact differential equation is bounded, in other words, the local truncation errors are damped out. The convergence means that the numerical solution approximates the solution of the original problem, i.e., a numerical method is said to be convergent if the numerical solution converges to the exact solution as the step size of mesh h tends to zero.

Although the consistency analysis of the θ-method is introduced almost in any arbitrary textbook of the numerical analysis, typically the stability (and hence, the convergence) is shown directly for the explicit method, only.

Our aim is to give an easy and elementary prove for the convergence of the general θ-method, i.e., we consider the implicit methods. The proof is direct and it is available for the non-specialists, too. Moreover, we give the expression for the stability constant of the θ-method.

This paper extends the results of the paper [4] in two directions: we prove the convergence of any implicit θ-method, and we also give sharp estimate for the stability constant, improving the result obtained in paper [4].

The paper is organized as follows. In Section 2, for sake of completeness, we formulate the basic results for the explicit Euler method, proving its convergence and stability constant. Section 3 contains the simple and compact proof of the convergence of the θ-method, and we define the order of its convergence, too. Finally, we finish the paper with giving some remarks and conclusions.

2. Convergence and the stability constant of the explicit Euler method

In this section we use a sequence of meshes ω_h and we define the numerical solution at some fixed point $t^\star \in (0, T)$ to the Cauchy problem (1)–(2) for the θ-method with $\theta = 0$, i.e., by using the scheme

$$\begin{aligned} y_n &= y_{n-1} + hf(t_{n-1}, y_{n-1}), \quad n = 1, 2, \ldots, N, \\ y_0 &= u_0 \end{aligned} \tag{8}$$

with $Nh = t^\star$.

The following statement will be used several times within the paper.

Lemma 2.1 *Let $a \geq 1$, $b \geq 0$, and s_n be such numbers that the inequalities*

$$|s_n| \leq a|s_{n-1}| + b, \quad n = 1, 2, \ldots \tag{9}$$

hold. Then the estimate

$$|s_n| \leq a^n \left(|s_0| + n\frac{b}{a} \right), \quad n = 0, 1, 2, \ldots \tag{10}$$

is valid.

Proof. By using induction, we can readily verify the statement. Indeed, for $n = 0$ (10) is clearly valid. Now, under the assumption that (10) holds for $n - 1$, from (9) we have

$$\begin{aligned} |s_n| &\leq a \left[a^{n-1} \left(|s_0| + (n-1)\frac{b}{a} \right) \right] + b \\ &= a^n \left(|s_0| + n\frac{b}{a} \right) \underbrace{-a^{n-1}b + b}_{\leq 0} \leq a^n \left(|s_0| + n\frac{b}{a} \right), \end{aligned} \tag{11}$$

which yields the statement. □

For the EEM the local truncation error at the mesh-point $t = t_n$ can be written as

$$l_n(h) = u(t_n) - u(t_{n-1}) - hu'(t_{n-1}) = \frac{h^2}{2} u''(\vartheta_n^{EEM}), \tag{12}$$

where $\vartheta_n^{EEM} \in (t_{n-1}, t_n)$ is a given value. Hence, setting $M_2 = \max_{[0,t^\star]} |u''|$, we get

$$l_n(h) \leq l(h) := M_2 \frac{h^2}{2}. \tag{13}$$

Let us consider the EEM defined by the one-step recursion (8). Due to (12), we have

$$u(t_n) = u(t_{n-1}) + hf(t_{n-1}, u(t_{n-1})) + l_n(h). \tag{14}$$

Hence, for the global error $e_n = u(t_n) - y_n$ at the mesh-point $t = t_n$ we get the recursion in the form

$$e_n = e_{n-1} + h\left(f(t_{n-1}, u(t_{n-1})) - f(t_{n-1}, y_{n-1}) \right) + l_n(h). \tag{15}$$

Hence, using the Lipschitz property (3) and (13), we obtain

$$|e_n| \leq |e_{n-1}| + hL|e_{n-1}| + l(h) = (1 + Lh)|e_{n-1}| + l(h), \tag{16}$$

for any $n = 1, 2, \ldots, N$. Then, by choosing $a = 1 + Lh$ and $b = l(h)$, and using the inequality $1 + x \leq \exp(x)$ for $x \geq 0$, Lemma 2.1 implies the estimate

$$|e_n| \leq [\exp(hL)]^n \left[|e_0| + \frac{nl(h)}{1 + Lh} \right] \leq [\exp(hL)]^n \left[|e_0| + nl(h) \right]. \tag{17}$$

Since $nh = t_n \le t^\star$, the following relations obviously hold for any $n = 1, 2, \ldots, N$:

$$[\exp(hL)]^n = \exp(Lhn) = \exp(Lt_n) \le \exp(Lt^\star),$$
$$nl(h) = nM_2\frac{h^2}{2} = \frac{M_2 t_n}{2}h \le \frac{M_2 t^\star}{2}h.$$

Because $e_0 = 0$, the relation (17) results in the estimate

$$|e_n| \le C_{EEM} \cdot h, \tag{18}$$

for all $n = 1, 2, \ldots, N$ with $C_{EEM} = \exp(Lt^\star)\dfrac{M_2 t^\star}{2}$. Putting $n = N$ into (18), we get

$$|e_N| \le C_{EEM} \cdot h. \tag{19}$$

This proves the first order convergence of the EEM with the stability constant C_{EEM}.

3. Convergence of the implicit theta methods

The convergence of the implicit θ-method (i.e., for $\theta \in (0, 1]$) cannot be proven directly as it was done previously. The main reason is that from the corresponding error recursion the inequality (9) cannot be obtained directly, due to the implicitness with respect to e_n. The usual way of proving the convergence of the θ-method is to show the zero-stability, by using its first characteristic polynomial. (The proof is complicated, and it can be found in [6, 10].)

In the sequel, using Lemma 2.1, we give an elementary proof of the convergence.

To this aim, we first give a uniform estimate for the local approximation error, which, by (6), has the form

$$\begin{aligned} l_n(h) &= u(t_n) - u(t_{n-1}) - (1-\theta)hu'(t_{n-1}) - \theta u'(t_n) \\ &= \theta\left(u(t_n) - u(t_{n-1}) - hu'(t_n)\right) + (1-\theta)\left(u(t_n) - u(t_{n-1}) - hu'(t_{n-1})\right). \end{aligned} \tag{20}$$

The Taylor polynomial with Lagrange remainder gives

$$\begin{aligned} u(t_{n-1}) &= u(t_n) - hu'(t_n) + \frac{h^2}{2}u''(t_n) - \frac{h^3}{6}u'''(\vartheta_n^1), \\ u(t_n) &= u(t_{n-1}) + hu'(t_{n-1}) + \frac{h^2}{2}u''(t_{n-1}) + \frac{h^3}{6}u'''(\vartheta_n^2). \end{aligned} \tag{21}$$

Using the relation $u''(t_n) = u''(t_{n-1}) + hu'''(\vartheta_n^3)$ (where $\vartheta_n^i \in (t_{n-1}, t_n)$ for $i = 1, 2, 3$), substitution (21) into (20) results in the equality

$$l_n(h) = \frac{h^2}{2}(1-2\theta)u''(t_{n-1}) + \frac{h^3}{6}\left(-3\theta u'''(\vartheta_n^3) + \theta u'''(\vartheta_n^1) + (1-\theta)u'''(\vartheta_n^2)\right). \tag{22}$$

Hence, using the notation $M_3 = \max_{[0,t^\star]}|u'''|$, we obtain

$$|l_n(h)| \le l(h) = C_2^\theta h^2 + C_3^\theta h^3, \tag{23}$$

where

$$C_2^\theta = \frac{|1-2\theta|}{2} M_2, \quad C_3^\theta = \frac{1+3\theta}{6} M_3. \tag{24}$$

We consider the θ-method, which means that the values y_n at the mesh-points ω_h are defined by the one-step recursion (4). Rearranging the local truncation error for θ-method of the form (5), and using the formula (4), for global error e_n we get the recursion

$$\begin{aligned} e_n = e_{n-1} &+ h\theta \left(f(t_n, u(t_n)) - f(t_n, y_n) \right) \\ &+ h(1-\theta) \left(f(t_{n-1}, u(t_{n-1})) - f(t_{n-1}, y_{n-1}) \right) + l_n(h), \quad n = 1, \ldots, N, \end{aligned} \tag{25}$$

with $e_0 = 0$. This equality, by using the Lipschitz continuity, implies the relation

$$|e_n| \le |e_{n-1}| + \theta L h |e_n| + (1-\theta) L h |e_{n-1}| + |l_n(h)|, \quad n = 1, \ldots, N. \tag{26}$$

Using the uniform estimate (23), (26) yields that with the choice

$$a = \frac{1 + (1-\theta)Lh}{1 - \theta L h}, \quad b = \frac{l(h)}{1 - \theta L h} \tag{27}$$

the recursion

$$|e_n| \le a|e_{n-1}| + b, \quad n = 1, 2, \ldots, N\,, \tag{28}$$

holds for the values

$$0 < h < \frac{1}{\theta L}. \tag{29}$$

Due to the obvious relations

$$a = 1 + \frac{Lh}{1 - \theta L h} \ge 1, \quad b \ge 0, \tag{30}$$

Lemma 2.1 is applicable to the recursion (28), which results in the validity of the estimate

$$|e_n| \le a^n \left(|e_0| + n\frac{b}{a} \right) = a^n \left(|e_0| + t_n \frac{l(h)}{h} \frac{1}{1 + (1-\theta)Lh} \right) \le a^n \left(t_n \frac{l(h)}{h} \right), \tag{31}$$

for any $n = 0, 1, 2, \ldots, N$ and h, satisfying (29).

We give an estimate for a^n. According to (30), we have

$$a = 1 + \frac{Lh}{1 - \theta L h} = 1 + \frac{1}{\theta} \cdot \frac{\theta L h}{1 - \theta L h}. \tag{32}$$

Let $\varepsilon > 0$ be arbitrary fixed number. Then for any $x \in (0, \varepsilon/(1+\varepsilon))$ the inequality $x^2/(1-x) \le \varepsilon x$ holds. Therefore, owning to the identity

$$\frac{x}{1-x} = x + \frac{x^2}{1-x},$$

we have the estimate

$$\frac{x}{1-x} \le (1+\varepsilon)x, \quad \text{for any } x \in \left(0, \frac{\varepsilon}{1+\varepsilon}\right). \tag{33}$$

Applying (33) to the second term on right-hand side (32), we obtain

$$a < 1 + \frac{1}{\theta} \cdot (1+\varepsilon)\theta L h = 1 + (1+\varepsilon)Lh \quad \text{for any } h \in (0, h_0), \tag{34}$$

where

$$h_0 = h_0(\varepsilon) = \frac{\varepsilon}{(1+\varepsilon)\theta L}. \tag{35}$$

Hence, using again the estimation $1 + s < \exp(s)$ for $s > 0$, we get

$$a^n < \exp\left(L(1+\varepsilon)t_n\right), \quad h \in (0, h_0). \tag{36}$$

Since for $\varepsilon > 0$ the inequality $\varepsilon/(1+\varepsilon) < 1$ holds, therefore under the condition $h \in (0, h_0)$ the requirement (29) is satisfied, too. Hence, based on relations (31), (23) and (36), we can formulate our results in the following statements.

Theorem 3.1 *Let $\varepsilon > 0$ be any fixed number and ω_h a mesh with mesh-size $h < h_0$, where h_0 is given in* (35)*. Then for the global error e_n of the θ-method with $\theta \in (0, 1]$ the estimate*

$$|e_n| \le t_n \left(C_2^\theta h + C_3^\theta h^2\right) \exp\left(L(1+\varepsilon)t_n\right) \tag{37}$$

holds for any $n = 1, 2, \ldots, N$, with the constants C_2^θ and C_3^θ defined in (24)*.*

Let us apply Theorem 3.1 for the value $n = N$. Then we have the following statement.

Corollary 3.2 *Under the assumptions and notations of the Theorem 3.1, for the global error e_N the estimate*

$$|e_N| \le t^\star \left(C_2^\theta h + C_3^\theta h^2\right) \exp\left(L(1+\varepsilon)t^\star\right) \tag{38}$$

holds.

The formula (38) gives an estimate for the global error at the mesh-point $t^\star = t_N = Nh$ of the θ-method with $\theta \in (0, 1]$ for any fixed $h \in (0, h_0)$. Moreover, ε depends on h_0, and, due to (35), ε also tends to zero as $h_0 \to 0$. Therefore, letting $h_0 \to 0$ on both sides of (38), we get the following statement.

Theorem 3.3 *The θ-method with any fixed $\theta \in (0, 1]$ is convergent at any fixed point $t^\star \in (0, T)$. Moreover, it is of the first order for $\theta \ne 0.5$, and of the second order for $\theta = 0.5$, with the stability constants $C_2^\theta t^\star \exp(Lt^\star)$ and $C_3^\theta t^\star \exp(Lt^\star)$, respectively.*

Since for the explicit Euler method we have $\theta = 0$ and $C_2^0 = C_{EEM}$ (c.f. formulas (15) and (24)), we can summarize our results in the following statement.

Theorem 3.4 *For the Cauchy problem* (1)–(2) *under the Lipschitz condition* (3) *the θ-method with any fixed $\theta \in [0,1]$ is convergent at any fixed point $t^\star \in (0,T)$. The rate of convergence of the method is equal to two for $\theta = 0.5$, otherwise it is of the first order. The stability constant C^θ of the method is defined as*

$$C^\theta = \begin{cases} \dfrac{1+3\theta}{6} M_3 t^\star \exp(Lt^\star) & \text{for } \theta = 0.5, \\ \dfrac{|1-2\theta|}{2} M_2 t^\star \exp(Lt^\star) & \text{for } \theta \neq 0.5, \end{cases} \tag{39}$$

respectively.

4. Concluding remarks

Finally, we give some comments.

♢ The convergence on the interval $[0, t^\star]$ yields the relation

$$\lim_{h \to 0} \max_{n=1,2,\ldots,N} |e_n| = 0.$$

As one can easily see, based on the relations (15) (for the EEM) and (37) (for the θ-method) the global error $|e_n|$ at any mesh-point can be bounded by the expression $C_{EEM} \cdot h$ (for the EEM) and by term, standing on the right-hand side of (38) (for the IEM). This means that both methods are convergent on the interval $[0, t^\star]$ with the same order.

♢ In our paper we did not consider roundoff error, which is always present in computer calculations. At the present time there is no universally accepted method to analyze roundoff error after a large number of time steps. The three main methods for analyzing roundoff accumulation are the analytical method, the probabilistic method and the interval arithmetic method, each of which has both advantages and disadvantages.

♢ In the implicit θ-method in each step we must solve a -usually non-linear- equations, namely, the root of the equation. This can be done by using some iterative method such as direct (function) iteration, Newton method and modified Newton method.

♢ In this paper we have been concerned with the stability and accuracy properties of the Euler methods in the asymptotic limit of $h \to 0$ and $N \to \infty$ while $N \cdot h$ is fixed. However, it is of practical significance to investigate the performance of methods in the case of fixed $h > 0$ and $n \to \infty$. Specifically, we would like to ensure that when applied to an initial value problem whose solution decays

to zero as $t \to \infty$, the Euler methods exhibit a similar behavior, for fixed $h > 0$ and $t_n \to \infty$. This problem is investigated on the famous Dahlquist scalar test equation, and it requires the so called A-stability property [3]. As it is known (e.g. in [8]), the θ-method is A-stable ("absolute stable") for the values $\theta \in [0.5, 1]$, otherwise the θ-method is bounded only under some strict condition for h. The latter makes these methods (including the EEM, too) unusable for several classes of the problem, like stiff problems.

♢ Why consider the θ-method, i.e., analyze the method with any θ in $[0, 1]$, not just 0, 0.5 and 1? We can list several reasons.

- The concept of order is based on assumption that error is concentrated on the leading order of Taylor series expansion (on real computers, h is small, but finite). As formula (7) shows, the case $\theta = 1/3$ gets rid of $\mathcal{O}(h^3)$ while retaining $\mathcal{O}(h^2)$. Hence, for different types of $f(t, u)$ one can tune θ to control whether $\mathcal{O}(h^3)$ and higher order terms or $\mathcal{O}(h^2)$ and higher order terms contribute to the overall error when h is finite.
- It may be possible to choose a θ that generates a close-to-optimal or smaller error. E.g., in [9] it is shown that the optimality criterion

$$\min_{\theta} \max_{-\infty<z<0} |\exp(z) - R(z)|$$

  leads to the value $\theta \approx 0.878$.
- θ-method is an example of a general approach to designing algorithms in which geometric intuition is replaced by Taylor series expansion. Invariably the implicit function theorem is also used in the design and analysis of this scheme.
- The implicit Euler method (the case $\theta = 1$) is very practical: it is a simple yet robust method for solving stiff ODE's.
- In some applications, a value such as $\theta = 0.55$ is used as trade-off between extended stability and second order accuracy.

♢ The qualitative analysis of the θ-method is investigated in several works, mainly, by its use to the numerical solution of some semidiscretized linear parabolic problems, (e.g. [5, 11]).

Acknowledgements

This work has been supported by the Hungarian Research Grant OTKA K 67819.

References

[1] Ascher, U. M. and Petzold, L. R.: *Computer methods for ordinary differential equations and differential-algebraic equations*. SIAM, Philadelphia, 1998.

[2] Bachvalov, N. S.: *Numerical methods*. Nauka, Moscow, 1975. (in Russian)

[3] Dahlquist, G.: Convergence and stability in the numerical integration of ordinary differential equations. Math. Scand. **4** (1956), 33–53.

[4] Faragó, I.: Note on the convergence of the implicit Euler method. (submitted)

[5] Faragó, I., Horváth, R.: Continuous and discrete parabolic operators and their qualitative properties. IMA J. Numer. Anal. **29** (2009), 606–631.

[6] Isaacson, E. and Keller, H. B.: *Analysis of numerical methods.* Wiley, New York, 1966.

[7] LeVeque, R.: *Finite difference mehods for ordinary and partial differential equations.* SIAM, Philadelphia, 2007.

[8] Lambert, J. D.: *Numerical methods for ordinary differential systems: The initial value problem.* John Wiley and Sons, Chicester, 1991.

[9] Liniger, W.: Global accuracy and A-stability of one- and two-step integration formulae for stiff ordinary differential equations. Lecture Notes in Mathematics **109** (1969), 188–193.

[10] Suli, E.: *Numerical solution of ordinary differential equations.* Oxford, 2010.

[11] Szabó, T.: On the discretization time-step in the finite element theta-method of the discrete heat equation. Lect. Notes Comp. Sci. **5434** (2009), 564–571.

Conference Applications of Mathematics 2013
in honor of the 70th birthday of Karel Segeth.
Jan Brandts, Sergey Korotov, Michal Křížek,
Jakub Šístek, and Tomáš Vejchodský (Eds.),
Institute of Mathematics AS CR, Prague 2013

SOLUTIONS OF HYPERSINGULAR INTEGRAL EQUATIONS OVER CIRCULAR DOMAINS BY A SPECTRAL METHOD

Leandro Farina[1,2], Juliana S. Ziebell[3]

[1]Instituto de Matemática, Universidade Federal do Rio Grande do Sul
Porto Alegre, RS, 91509-900, Brazil
farina@mat.ufrgs.br
[2]BCAM - Basque Center for Applied Mathematics,
Mazarredo 14, E48009 Bilbao, Basque Country, Spain
lfarina@bcamath.org
[3]Instituto de Matemática, Estatística e Física,
Universidade Federal do Rio Grande, Rio Grande, RS, 96201-900, Brazil
ju_sziebell@yahoo.com.br

Abstract

The problem of a solving a class of hypersingular integral equations over the boundary of a nonplanar disc is considered. The solution is obtained by an expansion in basis functions that are orthogonal over the unit disc. A Fourier series in the azimuthal angle, with the Fourier coefficients expanded in terms of Gegenbauer polynomials is employed. These integral equations appear in the study of the interaction of water waves with submerged thin plates.

1. Introduction

The aim of the present study is to consider the semi-analytical solution of a class of two-dimensional hypersingular integral equations. These equations can arise in the study of the interaction of water waves with submerged plates and the method of solution can be classified in the general area of spectral methods.

When the physical problem is two-dimensional and thus, the hypersingular integral equations is onedimensional, an efficient method can be applied for solution based on expansions in terms of Chebyshev polynomials. These problems can be related to scattering by flat [17] and curved [18] submerged plates, and by surface-piercing plates [18], and the trapping of water waves by submerged plates [19]. They used an expansion-collocation method to solve the one-dimensional hypersingular integral equations, in which the unknown is expanded using Chebyshev polynomials of the second kind. This method is very effective, and its convergence has been proved by Golberg [5, 6] and by Ervin & Stephan [1], in various function spaces. Ervin & Stephan [1] obtained the rate of convergence in appropriate Sobolev spaces. See also Frenkel [4] and Kaya & Erdogan [8].

The three-dimensional scattering by a thin disc, in deep water was investigated by Farina and Martin [2] and by Ziebell and Farina [20]. The authors solved the governing twodimensional hypersingular integral equation numerically using a spectral method using as basis functions, Gegenbauer polynomials in the radial variable. Physcally, when the plate is very close to the free surface, resonant frequencies can occur and this phenomenon has been inverstigated by Farina [3].

In this work, we ilustrate the spectral method by choosing the of problem a submerged disc is perturbed out of it original plane, so the disc could be denominated wrinkled or rough. This type of problem has been solved approximately, for circular caps and rough discs by Ziebell and Farina [20].

A similar problem in acoustics has been studied by Jansson [7], where the scattering of an acoustic wave from a thin circular disc was investigated by an integral equation method where the disc is modelled as part of an infinite interface between two half-spaces; this interface is then perturbed. However, this approach causes the behaviour of the solution near the edge of the disc to produce singularities at the edge of the disc.

Before presenting the integral equation that we will focus on, let us present in the next section, the physical and differential problem that originates it.

2. Formulation

A Cartesian coordinate system is chosen, in which z is directed vertically downwards into the fluid. We take the mean free surface lying at $z = 0$. We assume the presence of a submerged body into the fluid with a smooth, closed and bounded surface S. We suppose that the motions of the fluid are of small-amplitude, time-harmonic, that the fluid is incompressible and inviscid, and that the motion is irrotational. We denote ϕ as the potential flow and $[\phi]$ as the discontinuity in ϕ across S. Thus, the time dependent velocity potential is $\mathrm{Re}\{\phi(x, z, t)\}\mathrm{e}^{-\mathrm{i}\omega t}$, where ω is the angular frequency.

The conditions to be satisfied by ϕ are Laplace's equation

$$\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}\right)\phi = 0 \quad \text{in the fluid}$$

along with the free-surface condition

$$K\phi + \frac{\partial \phi}{\partial z} = 0 \ \text{ on } \ z = 0,$$

where $K = \omega^2/g$; g being the acceleration due to gravity.

On the surface of the body, the normal velocity is prescribed by

$$\frac{\partial \phi}{\partial n} = V \ \text{ in } \ S, \tag{1}$$

where V is a given function and $\frac{\partial}{\partial n}$ denotes normal differentiation.

Additionally, ϕ must satisfy a radiation condition:

$$r^{1/2}\left(\frac{\partial\phi}{\partial r} - \mathrm{i}K\phi\right) \to 0 \text{ when } r = (x^2+y^2)^{1/2} \to \infty.$$

The points P, Q denote points in the fluid and the points p, q denote points on the submerged body.

The free surface Green function for this problem is given by

$$G(P,Q) \equiv G(\xi,\eta,\zeta;x,y,z) = G_0(R, z-\zeta) + G_1(R, z+\zeta), \tag{2}$$

where $R = ((x-\xi)^2+(y-\eta)^2)^{1/2}$, $G_0(R,z-\zeta) = (R^2+(z-\zeta)^2)^{-1/2}$ and

$$G_1(R,z+\zeta) = \int\!\!\!\!\!\!\diagup_0^\infty \mathrm{e}^{-k(z+\zeta)} J_0(kR)\frac{k+K}{k-K}\,dk. \tag{3}$$

Here J_0 is the Bessel function of order zero. The path integral defining G_1 above runs below the singularity K. G satisfies the free surface condition, the Laplace equation, and have a weak singularity at $P = Q$.

For any harmonic function ϕ, satisfying $\phi = O(r^{-1})$ as $r \to \infty$, we have from Green's second identity, the following integral representation.

$$\phi(P) = \frac{1}{4\pi}\int_S \left(\phi(q)\frac{\partial}{\partial n_q}G(P,q) - G(P,q)\frac{\partial\phi}{\partial n_q}\right) dS_q, \tag{4}$$

where $\frac{\partial}{\partial n_q}$ denotes normal differentiation at q on S.

Now, for a thin body with surface Ω, denote the two sides of Ω by Ω^+ and Ω_- and define the discontinuity in ϕ across Ω by

$$[\phi] = \lim_{Q\to q^+}\phi(Q) - \lim_{Q\to q^-}\phi(Q),$$

where $q \in \Omega$, $q^- \in \Omega^-$, $q^+ \in \Omega^+$ and Q is a point in the fluid. Thus, equation (4) reduces to

$$\phi(P) = \frac{1}{4\pi}\int_\Omega [\phi(q)]\frac{\partial}{\partial n_q}G(P,q)\,dS, \tag{5}$$

where $n_q = n_q^+$ denotes now the normal unit vector at q on Ω^+ . Applying boundary condition (1) to (5) gives

$$\frac{1}{4\pi}\int\!\!\!\!\!\!\diagup_\Omega [\phi(q)]\frac{\partial^2}{\partial n_q \partial n_q}\,G(p,q)\,dS_q = V(p), \quad p \in \Omega, \tag{6}$$

where the integral must be interpreted in the Hadamard finite-part sense. Equation (6) is the governing hypersingular integral equation for $[\phi]$; this is to be solved subject to the edge condition

$$[\phi] = 0 \quad \text{in} \quad \partial\Omega.$$

Now let

$$\Omega : z = F(x,y) + \frac{b}{2},\ (x,y) \in D,$$

where D is the unit disc in the xy-plane and $\frac{b}{2}$ is the depth to which the body is submerged. Let $p, q \in \Omega$ such that $p = (\xi, \eta, \zeta),\ q = (x, y, z)$. The normal vector to Ω is then given by

$$\mathbf{N} = \left(-\frac{\partial F}{\partial x}, -\frac{\partial F}{\partial y}, 1 \right)$$

and a unit normal vector is therefore, expressed by $\mathbf{n} = \mathbf{N}/|\mathbf{N}|$. Using the notation

$$w(x,y) = [\phi(q)], \tag{7}$$

it can be shown by a direct calculation that formula (5) becomes

$$\phi(\xi,\eta,\zeta) = \frac{1}{4\pi} \int_D w(x,y) \frac{\mathbf{N} \cdot \mathbf{R_F}}{R_F^3}\, dS + \frac{1}{4\pi} \int_D w(x,y) (\nabla G_1 \cdot \mathbf{N})\, dS, \tag{8}$$

where $\mathbf{R_F} = (\xi - x, y - \eta, \zeta - F(x,y)),\ R_F = |\mathbf{R_F}|$ and $dS = dx\, dy$.

Our goal now is to clarify and understand the governing equation (6). In order to do this, consider the following definitions and notations.

$$F_1 = \frac{\partial F}{\partial x},\ F_2 = \frac{\partial F}{\partial y} \tag{9}$$

with F_1^0 and F_2^0 being the corresponding functions at (ξ, η). Let also $\Lambda = \frac{F(x,y) - F(\xi,\eta)}{R}$ and $\bar{\Lambda} = \frac{F(x,y) + F(\xi,\eta)}{R}$ and define the angle Θ by $x - \xi = R\cos\Theta$ and $y - \eta = R\sin\Theta$.

Projecting onto D, we can rewrite (6) as

$$\frac{1}{4\pi} \not\!\!\int_D H\ w(q) dA + \frac{1}{4\pi} \int_D W\ w(q) dA = V(p), \quad p \in D, \tag{10}$$

where (see [12])

$$\begin{aligned} H(\xi,\eta;x,y) \ &= \ \frac{1}{R^3} \left(\frac{1 + F_1 {F_1}^0 + F_2 {F_2}^0}{(1+\Lambda^2)^{\frac{3}{2}}} \right. \\ &\quad \left. - 3 \frac{(F_1 \cos\Theta + F_2 \sin\Theta - 1)({F_1}^0 \cos\Theta + {F_2}^0 \sin\Theta - 1)}{(1+\Lambda^2)^{\frac{5}{2}}} \right) \end{aligned} \tag{11}$$

and

$$W = \left. \frac{\partial^2 G_1}{\partial n_q \partial n_p} \right|_D = \not\!\!\int_0^\infty \mathrm{e}^{-k\bar{\Lambda}R} \mathrm{e}^{-kb}\, \mathcal{K}\, \frac{k+K}{k-K}\, dk, \tag{12}$$

where

$$\begin{aligned}\mathcal{K} &= F_1F_1^0\frac{k}{2R}(2\sin^2\Theta J_1(kR)+kR\cos^2\Theta(J_0(kR)-J_2(kR)))\\ &+F_2F_2^0\frac{k}{2R}(2\cos^2\Theta J_1(kR)+kR\sin^2\Theta(J_0(kR)-J_2(kR)))\\ &+(F_2F_1^0+F_1F_2^0)\frac{k}{2R}\cos\Theta\sin\Theta\,(kR(J_0(kR)-J_2(kR))-2J_1(kR))\\ &+(F_1^0-F_1)k^2\cos\Theta J_1(kR)\\ &+(F_2^0-F_2)k^2\sin\Theta J_1(kR)\\ &+k^2J_0(kR). \end{aligned} \tag{13}$$

Equation (10) is the governing equation for the problem of any submerged non planar circular disc Ω in water of infinite depth. Its solution gives the jump in the velocity potential ϕ across Ω. With this information, one can evaluate ϕ at any point P in the fluid by using (8). Equation (10) could be solved numerically, although not by the semi-analytical expansion-collocation method proposed by Farina e Martin [2] for the solution of hypersingular integral equations on a disc. Alternatively, an approximation to the solution could be obtained by a boundary perturbation method. We present such a method next. This method follows the one proposed by Martin [12] for treating the problem of a wrinkled disc in an unbounded fluid.

3. Perturbation method

We now assume that

$$V(p)=n_3,\quad n_3=\frac{1}{\sqrt{F_1^2+F_2^2+1}}, \tag{14}$$

where n_3 is the vertical component of the unit normal vector to the disc. This simplifies the following analysis and corresponds physically a situation where the disc performs heave (vertical) oscillations. Thus the problem stated in Section 2 becomes a radiation problem.

In order to consider a perturbation of the flat disc, we introduce the function f such that

$$F(x,y)=\epsilon f(x,y), \tag{15}$$

where ϵ is a small parameter and f is independent of ϵ. In [12] it is shown that

$$H=\frac{1}{R^3}\{1+\epsilon^2K_2+O(\epsilon^4)\},$$

where

$$K_2=f_1f_1^0+f_2f_2^0-\frac{3}{2}\lambda^2-3(f_1\cos\Theta+f_2\sin\Theta-\lambda)(f_1^0\cos\Theta+f_2^0\sin\Theta),$$

$\lambda = (f(x,y) - f(x\,\mathrm{i},\eta))/R$ and f_j, f_j^0 are defined similarly to F_j, F_j^0; see the comments after (9).

In order to get a similar expression for W, substitute (15) in (12), giving

$$W = W_0 + \epsilon W_1 + \epsilon^2 W_2, \tag{16}$$

where

$$W_0 = \int_0^\infty \mathrm{e}^{-k(\epsilon(f(x,y)+f(\xi,\eta))+b)} k^2 J_0(kR)\, \frac{k+K}{k-K}\, dk, \tag{17}$$

$$\begin{aligned} W_1 &= [(f_1^0 - f_1)\cos\Theta + (f_2^0 - f_2)\sin\,\Theta] \\ &\quad \times \int_0^\infty \frac{k+K}{k-K}\, \mathrm{e}^{-k(\epsilon(f(x,y)+f(\xi,\eta))+b)} k^2 J_1(kR)\, dk, \end{aligned} \tag{18}$$

and

$$\begin{aligned} W_2 &= \left[\frac{\sin^2\,\Theta}{R} f_1 f_1^0 + \frac{\cos^2\Theta}{R} f_2 f_2^0 - (f_2 f_1^0 + f_1 f_2^0)\frac{\sin(2\Theta)}{2R}\right] \\ &\quad \times \int_0^\infty \mathrm{e}^{-k(\epsilon(f(x,y)+f(\xi,\eta))+b)} k J_1(kR)\, \frac{k+K}{k-K}\, dk \\ &\quad + \left[\cos^2\Theta f_1 f_1^0 + \sin^2\,\Theta f_2 f_2^0 - (f_2 f_1^0 + f_1 f_2^0)\,\frac{\sin(2\Theta)}{2}\right] \\ &\quad \times \frac{1}{2}\int_0^\infty \mathrm{e}^{-k(\epsilon(f(x,y)+f(\xi,\eta))+b)} k^2 (J_0(kR) - J_2(kR))\, \frac{k+K}{k-K}\, dk. \end{aligned} \tag{19}$$

Expanding $\mathrm{e}^{-k(\epsilon(f(x,y)+f(\xi,\eta))+b)}$ in Taylor's series, we obtain

$$\begin{aligned} W_0 &= W_{00} + \epsilon W_{01} + \epsilon^2 W_{02}, \\ W_1 &= W_{10} + \epsilon W_{11} + \epsilon^2 W_{12}, \\ W_2 &= W_{20} + \epsilon W_{21} + \epsilon^2 W_{22}, \end{aligned}$$

where

$$W_{00} = \int_0^\infty \frac{k+K}{k-K} \mathrm{e}^{-kb} k^2 J_0(kR)\, dk \tag{20}$$

and the functions $W_{01}, \ldots, W_{22}$ are given in appendix A.

Substituting (15) in (14) and expanding in Taylor series, we get

$$n_3 = 1 + \frac{1}{2}\left(f_1^2 + f_2^2\right)\epsilon^2 + O(\epsilon^3). \tag{21}$$

Similarly, for w, assume

$$w = w_0 + \epsilon w_1 + \epsilon^2 w_2 + ... \tag{22}$$

Now, substituting (16) and (22) in (10), with V given by (21), we obtain

$$\begin{aligned}
\frac{1}{4\pi}\not\int_D w_0\frac{dA}{R^3}+\frac{1}{4\pi}\int_D W_{00}\,w_0\,dA &= 1, && (23)\\
\frac{1}{4\pi}\not\int_D w_1\frac{dA}{R^3}+\frac{1}{4\pi}\int_D W_{00}\,w_1\,dA &= -\frac{1}{4\pi}\int_D (W_{10}+W_{01})w_0\,dA, && (24)\\
\frac{1}{4\pi}\not\int_D w_2\frac{dA}{R^3}+\frac{1}{4\pi}\int_D W_{00}\,w_2\,dA &= -\frac{1}{4\pi}\not\int_D K_2 w_0\frac{dA}{R^3}\\
&\quad -\frac{1}{4\pi}\int_D (W_{02}+W_{11}+W_{20})w_0\,dA\\
&\quad -\frac{1}{4\pi}\int_D (W_{01}+W_{10})w_1\,dA\\
&\quad +\frac{1}{2}\left(f_1^2+f_2^2\right). && (25)
\end{aligned}$$

Note that equation (23) appears in [11, eq. 4.1] and in [2, eq. 17]. Thus, the first order equation of the present perturbation method recovers the governing equation for the *plane* disc: this corresponds to the problem of a horizontal and plane circular disc performing heave oscillations.

By defining the integral operators

$$\begin{aligned}
H_{ij}w &= \int_D W_{ij}w\;dA \qquad \forall\; i,j\in\{0,1,2\},\\
\mathcal{H}w &= \not\int_D w\;\frac{dA}{R^3},\\
\mathcal{K}_2w &= \not\int_D K_2 w\;\frac{dA}{R^3},
\end{aligned}$$

we can write equations (23)-(25) in a more compact form as

$$(\mathcal{H}+H_{00})\,w_0 = 1, \tag{26}$$

$$(\mathcal{H}+H_{00})\,w_1 = -(H_{10}+H_{01})w_0, \tag{27}$$

$$(\mathcal{H}+H_{00})\,w_2 = -(\mathcal{K}_2+H_{02}+H_{11}+H_{20})w_0-(H_{01}+H_{10})w_1+\frac{1}{2}(f_1^2+f_2^2) \tag{28}$$

Equations (26)–(28) form a sequence of integral equations that approach the solution of the governing equation (10). Note that the simple structure of these equations offers an alternative to the solution of the problem: in order to solve it, one has just to invert the integral operator $H_{00}+\mathcal{H}$. Note further that the function f is only present in the right-hand side of the equations. This means that all the information about the specific geometry of the plate is in these terms of the equations. Thus, it is possible to *pre-solve* the problem for any perturbation of the disc by inverting the operator mentioned above. This can be done efficiently by the numerical method presented in Section 4.

4. Alternative expressions and numerical method

In this section we show how to compute a solution of the problem formulated in the section above.

4.1. Alternative expressions for W

The integrands of the integral equations (26)–(28) involve the regular part of the free surface Green function, that is, G_1, and its derivatives. The numerical implementation of these functions are not trivial. Specifically, these integrands present path integrals that involve Bessel functions. Nevertheless, we can express these integrals in terms of Bessel functions and Struve functions which are suitable for more efficient numerical calculation. According to [13] (see also [15, 16]), we have

$$\begin{aligned} G_1 &= \int\!\!\!\!\!\!\diagup_0^\infty \frac{k+K}{k-K}\mathrm{e}^{-k(z+\zeta)}J_0(kR)\,dk \\ &= K\left[(X^2+Y^2)^{-1/2} - \pi\mathrm{e}^{-Y}(H_0(X)+Y_0(X)) - 2\int_0^Y \mathrm{e}^{t-Y}(X^2+t^2)^{-1/2}dt\right] \\ &\quad -2\pi\mathrm{i}K\mathrm{e}^{-Y}J_0(X), \end{aligned} \tag{29}$$

where $X = KR$, $Y = K(z+\zeta)$, H_0 is the Struve function of order 0 and J_0 and Y_0 denote the Bessel functions of the first and second kind, respectively. Expression (29) is suitable for numerical calculation; this has been used is several computer codes for water wave analysis. See for instance [10].

Using (29), it can be shown that the integrands $W_{00}, \ldots, W_{22}$, originally written as (39–45) in appendix A, admit similar representations. For example,

$$\begin{aligned} W_{00} &= 2K^2(R^2+b^2)^{-1/2} + (2Kb-1)(R^2+b^2)^{-3/2} + 3b^2(R^2+b^2)^{-5/2} \\ &\quad -\pi K^3\mathrm{e}^{-Kb}(H_0(KR)+Y_0(KR)) - 2K^3\mathrm{e}^{-Kb}\int_0^{Kb}\mathrm{e}^t((KR)^2+t^2)^{-1/2}\,dt \\ &\quad +2\pi\mathrm{i}K^3\mathrm{e}^{-Kb}J_0(KR) \end{aligned} \tag{30}$$

is an alternative expression for W_{00}, which allows more efficient numerical computation than (39) does. The expression (30) does not involve path integrals whose calculation are computationally expensive. Furthermore, the Struve and Bessel functions present in this alternative term are efficiently computed by approximating orthogonal polynomials; see [14]. Integrals such as the one in (30) can be efficiently computed; see [13] and [15]. Similar alternative expressions for the W_{01}, W_{02}, W_{10} and W_{20} are shown in appendix B.

4.2. Expansion-collocation method

4.2.1. Review of the one-dimensional theory

In two-dimensions, many wave problems involving thin plates can be reduced to an equation of the form

$$\int\!\!\!\!\!\!\diagup_{-1}^1 \left\{\frac{1}{(x-t)^2} + H(x,t)\right\} v(t)\,dt = f(x) \quad \text{for } -1 < x < 1, \tag{31}$$

supplemented by two boundary conditions, which we take to be $v(-1) = v(1) = 0$. Here, v is the unknown function, f is prescribed and the kernel H is known. Assuming that f is sufficiently smooth, the solution v has square-root zeros at the end-points. This suggests that we write

$$v(x) = \sqrt{1 - x^2}\, u(x).$$

Then, we expand u using a set of orthogonal polynomials; a good choice is to use Chebyshev polynomials of the second kind, U_n, defined by

$$U_n(\cos\theta) = \frac{\sin(n+1)\theta}{\sin\theta}, \qquad n = 0, 1, 2, \ldots.$$

This is a good choice because of the formula

$$\frac{1}{\pi} \barint_{-1}^{1} \frac{\sqrt{1 - t^2}\, U_n(t)}{(x - t)^2}\, dt = -(n+1)U_n(x). \tag{32}$$

Thus, we approximate u by

$$\sum_{n=0}^{N} a_n U_n(x),$$

substitute into (31) and evaluate the hypersingular integral analytically, using (32). To find the $(N+1)$ coefficients a_n, we collocate at $(N+1)$ points; good choices are the zeros of T_{N+1} or U_{N+1}, where T_n is a Chebyshev polynomial of the first kind.

4.2.2. The two-dimensional theory

The governing equations (26)–(28), obtained by the perturbation method in section 3, can be written in the same form, which is

$$(\mathcal{H} + H_{00})u = g, \tag{33}$$

where g is a known function, which can involve solutions of lower order problems. As a particular case, the plane disc equation (26) has an axisymmetric solution and can be solved by reducing it to a non singular one dimensional Fredholm integral equation of the second kind [11, eq. 7.6]. A simple numerical method can be used for this equation; for instance a Nyström method combined with the Gauss-Legendre quadrature rule, as employed by Martin and Farina [11]. However, as the solutions of equations (27) and (28) are not axisymmetric, we need a more general method of solution. We employ the expansion-collocation method used by Farina and Martin [2] for solving an equation of the form of (33). In fact, this method does not require that $V = 1$. This forcing could be any function of two variables; for instance, this could represent an incident wave and in this way, the problem would be a scattering one. In order to describe the expansion-collocation method, introduce cylindrical polar coordinates (r, θ, z), so that $x = r\cos\theta$ and $y = r\sin\theta$. Then, the disc is given by

$$D = \{(r, \theta, z) : 0 \le r < 1, -\pi \le \theta < \pi, z = b/2\}. \tag{34}$$

If we write $\xi = s\cos\alpha$, $\eta = s\sin\alpha$, we have

$$R^3 = [r^2 + s^2 - 2rs\cos(\theta - \alpha)]^{3/2}.$$

Hence we can write (33) as

$$\frac{1}{4\pi} \neq\!\!\!\!\!\int_D u(s,\alpha) \left\{ \frac{1}{R^3} + W_{00}(r,\theta;s,\alpha;b,K) \right\} s\,ds\,d\alpha = g(r,\theta), \quad (r,\theta) \in D, \tag{35}$$

We shall expand u using the basis functions B_k^m, defined by

$$B_k^m(r,\theta) = P_{m+2k+1}^m(\sqrt{1-r^2})\,\mathrm{e}^{\mathrm{i}m\theta}, \qquad k,m = 0,1,\ldots,$$

where P_n^m is an associated Legendre function. The radial part of these basis functions can also be expressed in terms of Gegenbauer polynomials.

The functions $\{B_k^m\}$ are orthogonal over the unit disc with respect to the weight $(1-r^2)^{-1/2}$.

The next formula, due to Krenk [9] is essential in the construction of the method:

$$\frac{1}{4\pi} \neq\!\!\!\!\!\int_S \frac{1}{R^3} B_k^m(s,\alpha)\,s\,ds\,d\alpha = C_k^m \frac{B_k^m(r,\theta)}{\sqrt{1-r^2}}, \tag{36}$$

where

$$C_k^m = -\frac{\pi}{4}\frac{(2k+1)!}{(2m+2k+1)!}[P_{m+2k+1}^{m+1}(0)]^2$$

Equation (36) allows us to evaluate the hypersingular integrals analytically[1]. To exploit (36), we expand $[\phi]$ in terms of the functions B_k^m. For brevity, we write

$$[\phi] = w \approx \sum_{k,m}^{N} a_k^m\, B_k^m := \sum_{k=0}^{N_1}\sum_{m=0}^{N_2} a_k^m\, B_k^m. \tag{37}$$

Substituting (37) in the integral equation (35) and then evaluating the hypersingular integrals analytically using (36), we obtain

$$\sum_{k,m}^{N} a_k^m \left\{ C_k^m \frac{B_k^m(r,\theta)}{\sqrt{1-r^2}} + \frac{1}{4\pi}\int_S B_k^m(s,\alpha) W_{00}(r,\theta;s,\alpha;d,K)\,s\,ds\,d\alpha \right\} = g(r,\theta), \quad (r,\theta) \in D. \tag{38}$$

[1]Another consequence of formula (36) is that the functions $B_k^m(r,\theta)/\sqrt{1-r^2}$ can bee seen as eigenfunctions of the integral operator $\bar{\mathcal{H}}$ defined by

$$\bar{\mathcal{H}}v(r,\theta) = \neq\!\!\!\!\!\int_D \frac{1}{R^3} v(s,\alpha)\,\sqrt{1-s^2}\,s\,ds\,d\alpha.$$

It remains to determine the unknown coefficients a_k^m. We use a collocation method, in which evaluation of (38) at $(N_1+1)(N_2+1)$ points on the disc gives a linear system for the coefficients a_k^m. For a discussion on the choice of the collocation points on a disc and other numerical issues on the collocation-expansion method, including its analogue for two-dimensional water wave problems, see [2]. Numerical results showing the effectiviness of the method were presented by Ziebell and Farina [20] for spherical caps and rough discs.

5. Discussion

We have presented a spectral method for solving a class of hypersingular equations over a nonplanar circular disc. The motivation of the problem comes from a interaction of water waves with a submerged thin non-planar surface. By using a boundary perturbation method, we formulate the problem in terms of sequence of hypersingular integral equations, $(\mathcal{H} + H_{00})w_n = g_n$, over a flat disc. This approach allows the application of a efficient semi-analytical method where the solution is expanded in terms of Gegenbauer polynomials. This is the analogue of a spectral method used for the solutions of one-dimensional hypersingular integral equations in terms of Chebyshev polynomials.

Acknowledgements

The first author has done part of the research on this article while a member of the EU project FP7-295217 – HPC-GA. The second author acknowledges financial support from CAPES and CNPq.

Appendices

A. Expansion terms for W

$$W_{00} = \fint_0^\infty \frac{k+K}{k-K} \mathrm{e}^{-kb} k^2 J_0(kR)\, dk, \tag{39}$$

$$W_{01} = -(f(x,y) + f(\xi,\eta)) \fint_0^\infty \frac{k+K}{k-K} \mathrm{e}^{-kb} k^3 J_0(kR)\, dk, \tag{40}$$

$$W_{02} = \frac{1}{2}(f(x,y) + f(\xi,\eta))^2 \fint_0^\infty \frac{k+K}{k-K} \mathrm{e}^{-kb} k^4 J_0(kR)\, dk, \tag{41}$$

$$W_{10} = -[(f_1 - f_1^0)\cos\Theta + (f_2 - f_2^0)\sin\Theta] \fint_0^\infty \frac{k+K}{k-K} k^2 \mathrm{e}^{-kb} J_1(kR)\, dk, \tag{42}$$

$$W_{11} = [(f_1 - f_1^0)\cos\Theta + (f_2 - f_2^0)\sin\Theta](f(x,y) + f(\xi,\eta)) \fint_0^\infty \frac{k+K}{k-K} k^3 \mathrm{e}^{-kb} J_1(kR)\, dk, \tag{43}$$

$$W_{12} = -\frac{1}{2}[(f_1 - f_1^0)\cos\Theta + (f_2 - f_2^0)\sin\Theta](f(x,y) + f(\xi,\eta))^2 \oint_0^\infty \frac{k+K}{k-K} k^4 \mathrm{e}^{-kb} J_1(kR)\, dk, \tag{44}$$

$$\begin{aligned} W_{20} = &\left[\frac{\sin^2\Theta}{R} f_1 f_1^0 + \frac{\cos^2\Theta}{R} f_2 f_2^0 - (f_2 f_1^0 + f_1 f_2^0)\frac{\sin(2\Theta)}{2R}\right] \oint_0^\infty \mathrm{e}^{-kb} k J_1(kR)\, \frac{k+K}{k-K}\, dk \\ &+ \left[\cos^2\Theta f_1 f_1^0 \sin^2\Theta f_2 f_2^0 - (f_2 f_1^0 + f_1 f_2^0)\, \frac{\sin(2\Theta)}{2}\right] \\ &\times \frac{1}{2}\oint_0^\infty \mathrm{e}^{-kb} k^2 (J_0(kR) - J_2(kR))\, \frac{k+K}{k-K}\, dk, \end{aligned} \tag{45}$$

$$\begin{aligned} W_{21} = &(f(x,y) + f(\xi,\eta)) \left\{ \left[-\frac{\sin^2\Theta}{R} f_1 f_1^0 - \frac{\cos^2\Theta}{R} f_2 f_2^0 + (f_2 f_1^0 + f_1 f_2^0)\, \frac{\sin(2\Theta)}{2R} \right] \right. \\ &\times \oint_0^\infty \mathrm{e}^{-kb} k^2 J_1(kR)\, \frac{k+K}{k-K}\, dk \\ &+ \left[-\cos^2\Theta f_1 f_1^0 - \sin^2\Theta f_2 f_2^0 + (f_2 f_1^0 + f_1 f_2^0)\, \frac{\sin(2\Theta)}{2} \right] \\ &\left. \times \frac{1}{2}\oint_0^\infty \mathrm{e}^{-kb} k^3 (J_0(kR) - J_2(kR)) \frac{k+K}{k-K}\, dk \right\}, \end{aligned} \tag{46}$$

$$\begin{aligned} W_{22} = &\frac{1}{2}(f(x,y) + f(\xi,\eta))^2 \left\{ \left[\frac{\sin^2\Theta}{R} f_1 f_1^0 + \frac{\cos^2\Theta}{R} f_2 f_2^0 - (f_2 f_1^0 + f_1 f_2^0)\, \frac{\sin(2\Theta)}{2R} \right] \right. \\ &\times \oint_0^\infty \mathrm{e}^{-kb} k^3 J_1(kR)\, \frac{k+K}{k-K}\, dk \\ &+ \left[\cos^2\Theta f_1 f_1^0 \sin^2\Theta f_2 f_2^0 - (f_2 f_1^0 + f_1 f_2^0)\, \frac{\sin(2\Theta)}{2} \right] \\ &\left. \times \frac{1}{2}\oint_0^\infty \mathrm{e}^{-kb} k^4 (J_0(kR) - J_2(kR))\, \frac{k+K}{k-K}\, dk \right\}. \end{aligned} \tag{47}$$

B. Alternatives expressions

$$\begin{aligned} W_{01} = &(f(x,y) + f(\xi,\eta)) \Big[-2bK^3 (R^2 + b^2)^{-1/2} \\ &+ (2K - 2K^2 b)(R^2 + b^2)^{-3/2} + (9b - 6Kb^2)(R^2 + b^2)^{-5/2} \\ &- 15b^3 (R^2 + b^2)^{-7/2} + \pi K^4 \mathrm{e}^{-Kb}(H_0(KR) + Y_0(KR)) \\ &+ 2\, K^3 \mathrm{e}^{-Kb} \int_0^{Kb} \mathrm{e}^t ((KR)^2 + t^2)^{-1/2}\, dt + 2\pi \mathrm{i} K^4 \mathrm{e}^{-Kb} J_0(KR) \Big], \end{aligned} \tag{48}$$

$$W_{02} = \frac{1}{2}(f(x,y)+f(\xi,\eta))^2\Big[-2K^2(R^2+b^2)^{-3/2}+(6K^2b^2+3Kb^2-18b+5)$$
$$\times(R^2+b^2)^{-5/2}+(-75K^2b^4+30Kb^3-15b^2)(R^2+b^2)^{-7/2}$$
$$+105K^2b^6(R^2+b^2)^{-9/2}-\pi K^5\mathrm{e}^{-Kb}(H_0(KR)+Y_0(KR))$$
$$-2K^3\mathrm{e}^{-Kb}\int_0^{Kb}\mathrm{e}^t((KR)^2+t^2)^{-1/2}\,dt-2\pi\mathrm{i}K^5\mathrm{e}^{-Kb}J_0(X)\Big], \tag{49}$$

$$W_{10} = [(f_1-f_1^0)\cos\Theta+(f_2-f_2^0)\sin\,\Theta](f(x,y+f(\xi,\eta))\Big[-2KR(R^2+b^2)^{-3/2}$$
$$-3R(R^2+b^2)^{-5/2}+\pi K^3\mathrm{e}^{-Kb}\left(H_1(KR)+Y_1(KR)-\frac{2}{\pi}\right)$$
$$+2K^4R\,\mathrm{e}^{-Kb}\int_0^{Kb}\mathrm{e}^{-Kb}((KR)^2+t^2)^{-3/2}\,dt$$
$$-2\pi\mathrm{i}K^4\mathrm{e}^{-Kb}J_1(KR)\Big], \tag{50}$$

and

$$W_{20} = \left[\frac{-\sin^2\,\Theta}{R}f_1f_1^0-\frac{\cos^2\Theta}{R}f_2f_2^0+(f_2f_1^0+f_1f_2^0)\frac{\sin\,(2\Theta)}{2R}\right]\Big[-R(R^2+b^2)^{-3/2}$$
$$-\pi K^2\mathrm{e}^{-Kb}(H_0'(X)+Y_0'(X))+2K^3\mathrm{e}^{-Kb}R\int_0^{Kb}\mathrm{e}^t((KR)^2+t^2)^{-3/2}\,dt$$
$$+2\pi\mathrm{i}K^2\mathrm{e}^{-Kb}J_1(KR)\Big]+\left[-\cos^2\Theta f_1f_1^0-\sin^2\Theta f_2f_2^0+(f_2f_1^0+f_1f_2^0)\,\frac{\sin\,(2\Theta)}{2}\right]$$
$$\left[3R(R^2+b^2)^{-5/2}-\frac{1}{2}\pi K^3\mathrm{e}^{-Kb}\left(H_2(KR)+Y_2(KR)-H_0(KR)-Y_0(KR)\right.\right.$$
$$\left.-\frac{KR}{2\sqrt{\pi}\,\Gamma(5/2)}\right)+2K^3\mathrm{e}^{-Kb}\int_0^{Kb}\mathrm{e}^t((KR)^2+t^2)^{-3/2}\,dt$$
$$-6K^3\mathrm{e}^{-Kb}\int_0^{Kb}\mathrm{e}^t((KR)^2+t^2)^{-5/2}\,dt-\pi\mathrm{i}K^3\mathrm{e}^{-Kb}(J_2(KR)-J_0(KR))\Big], \tag{51}$$

where Γ denotes the Gamma function.

References

[1] Ervin, V. J., Stephan, E. P.: Collocation with Chebyshev polynomials for a hypersingular integral equation on an interval. J. Comp. & Appl. Math. **43** (1992), 221–229.

[2] Farina, L., Martin, P. A.: Scattering of water waves by a submerged disc using a hypersingular integral equation. Applied Ocean Research **20** (1998), 121–134.

[3] Farina, L.: Water wave radiation by a heaving submerged horizontal disk very near the free surface. Physics of Fluids **22** (2010), 057102.

[4] Frenkel, A.: A Tschebyshev expansion of singular integrodifferential equations with a $\partial^2 \ln|s-t|/\partial s\,\partial t$ kernel. J. Comp. Phys. **51** (1983), 335–342.

[5] Golberg, M. A.: The convergence of several algorithms for solving integral equations with finite-part integrals. J. Integ. Equations. **5** (1983), 329–340.

[6] Golberg, M. A.: The convergence of several algorithms for solving integral equations with finite-part integrals. II. J. Integ. Equations. **9** (1985), 267–275.

[7] Jansson, P.-Å.: Acoustic scattering from a rough dircular disk. J. Acoust. Soc. Am. **99**(2) (1996), 672–681.

[8] Kaya, A. C., Erdogan, F.: On the solution of integral equations with strongly singular kernels. Q. Appl. Math. **45** (1987), 105–122.

[9] Krenk, S.: A circular crack under asymmetric loads and some related integral equations. J. Appl. Mech. **46** (1979), 821–826.

[10] Lee, C.-H., Newman, J. N.: Computations of wave effects using the panel method. In: S. Chakrabarti, (Ed.), *Numerical modeling in fluid-structure interaction*, WIT Press, Southampton, 2005.

[11] Martin, P. A., Farina, L.: Radiation of water waves by a heaving submerged horizontal disc. J. Fluid Mech. **337** (1997), 365–379.

[12] Martin, P. A.: On potential flow past wrinkled discs. Proc. Royal Soc. of London, Series A. **454** (1998), 2209–2221.

[13] Newman, J. N.: Double-precision evaluation of the oscillatory source potential. Journal of Ship Research **28** (1984), 151–154.

[14] Newman, J. N.: Approximations for the bessel and struve functions. Mathematics of Computation **43** (1984), 551–556.

[15] Newman, J. N.: Algorithms for the free-surface green function. J. Eng. Math. **19** (1985), 57–67.

[16] Newman, J. N.: Approximation of free-surface Green funtions. In: P. A. Martin and G. R. Wickham, (Eds.), *Wave Asymptotics*, pp. 107–135. Cambridge University Press, 1992.

[17] Parsons, N. F., Martin, P. A.: Scattering of water waves by submerged plates using hypersingular integral equations. Appl. Ocean Research **14** (1992), 313–321.

[18] Parsons, N. F., Martin, P. A.: Scattering of water waves by submerged curved plates and by surface-piercing flat plates. Appl. Ocean Res. **16** (1994), 129–139.

[19] Parsons, N. F., Martin, P. A.: Trapping of water waves by submerged plates using hypersingular integral equations. J. Fluid Mech. **284** (1995), 359–375.

[20] Ziebell, J. S., Farina, L.: Water wave radiation by a submerged rough disc. Wave Motion **49** (2012), 34–49.

Conference Applications of Mathematics 2013
in honor of the 70th birthday of Karel Segeth.
Jan Brandts, Sergey Korotov, Michal Křížek,
Jakub Šístek, and Tomáš Vejchodský (Eds.),
Institute of Mathematics AS CR, Prague 2013

A SHORT PHILOSOPHICAL NOTE ON THE ORIGIN OF SMOOTHED AGGREGATIONS

Pavla Fraňková, Milan Hanuš, Hana Kopincová, Roman Kužel,
Petr Vaněk, Zbyněk Vastl

University of West Bohemia
Univerzitni 22, 306 14 Pilsen, Czech Republic
frankova@kma.zcu.cz, mhanus@kma.zcu.cz, kopincov@kma.zcu.cz, rkuzel@kma.zcu.cz,
ptrvnk@kma.zcu.cz, zvastl@kma.zcu.cz

Abstract

We derive the smoothed aggregation two-level method from the variational objective to minimize the *final error* after finishing the entire iteration. This contrasts to a standard variational two-level method, where the coarse-grid correction vector is chosen to minimize the error after coarse-grid correction procedure, which represents merely an intermediate stage of computing. Thus, we enforce *the global minimization of the error*. The method with smoothed prolongator is thus interpreted as a qualitatively different, and more optimal, algorithm than the standard multigrid.

1. Introduction

The smoothed aggregation method [13, 14, 15, 12] proved to be a very efficient tool for solving various types of elliptic problems and their singular perturbations. In this short note, we turn to the very roots of smoothed aggregation method and derive its two-level variant on a systematic basis.

The multilevel method consists in combination of a coarse-grid correction and smoothing. The coarse-grid correction of a standard two-level method is derived using the A-orthogonal projection of an error to the range of the prolongator. In other words, the coarse-grid correction vector is chosen to minimize the error *after coarse-grid correction procedure*. This means, the standard two-level method minimizes the error in an intermediate stage of the iteration, while we are, naturally, interested in minimizing *the final error after accomplishing the entire iteration*. In other words, we strive to minimize the error after coarse-grid correction and subsequent smoothing. The two-level smoothed aggregation method is obtained by solving this minimization problem. This, in the opinion of the authors, explains its remarkable robustness.

We derive the two-level smoothed aggregation method from the variational objective to minimize the error after coarse-grid correction and subsequent post-smoothing. Then, by a trivial argument, we extend our result to the two-level method with pre-smoothing, coarse-grid correction and post-smoothing.

The minimization of error after coarse-grid correction and subsequent smoothing leads to a method with smoothed prolongator. We can say that by smoothing the prolongator, we adapt the coarse-space (the range of the prolongator) to the post-smoother so that the resulting iteration is as efficient as possible. Our short explanation applies to any two-level method with smoothed prolongator. The particular case we have in mind is, however, a method with smoothed *tentative* prolongator given by generalized unknowns aggregations [15]. The discrete basis functions of the coarse-space (the columns of the prolongator) given by unknowns aggregations have no overlap; the natural overlap of discrete basis functions (like it is in the case of finite element basis functions) is created by smoothing and, for additive point-wise smoothers, leads to sparse coarse-level matrix.

Our argument is basically trivial. It, however, shows a fundamental property of the method with smoothed prolongator, that is essential. This argument is known to the authors for a long time, but has never been published.

We conclude our paper by a numerical test. Namely, we demonstrate experimentally that smoothed aggregation method with powerful smoother and small coarse-space solves efficiently highly anisotropic problems without the need to perform semi-coarsening (the coarsening that follows only strong connections).

2. Two-level method

We solve a system of linear algebraic equations

$$A\mathbf{x} = \mathbf{f}, \tag{1}$$

where A is a symmetric positive definite matrix of order n and $\mathbf{f} \in \mathbb{R}^n$. We assume that an injective linear *prolongator* $p : \mathbb{R}^m \to \mathbb{R}^n,\ m < n$ is given.

The two-level method consists in the combination of a *coarse-grid correction* and *smoothing*. The smoothing means using point-wise iterative methods at the beginning and at the end of the iteration. The coarse-grid correction is derived by correcting an error $\mathbf{e}$ by a coarse-level vector $\mathbf{v}$ so that the resulting error $\mathbf{e} - p\mathbf{v}$ is minimal in A-norm. In other words, we solve the minimization problem

$$\text{find } \mathbf{v} \in \mathbb{R}^m \text{ so that } \|\mathbf{e} - p\mathbf{v}\|_A \text{ is minimal.} \tag{2}$$

It is well-known that such vector $p\mathbf{v}$ is an A-orthogonal projection of the error $\mathbf{e}$ onto Range(p), with the projection operator given by

$$Q = p(p^T Ap)^{-1} p^T A.$$

Thus, the error propagation opeartor of the coarse-grid correction is given by $I - Q = I - p(p^T Ap)^{-1} p^T A$ and the error propagation operator of the two-level method by

$$E_{TGM} = S_{post}[I - p(p^T Ap)^{-1} p^T A]S_{pre}, \tag{3}$$

where S_{pre} and S_{post} are error propagation operators of pre- and post- smoothing iterations, respectively.

Clearly, for the error $\mathbf{e}(\mathbf{x}) \equiv \mathbf{x} - A^{-1}\mathbf{f}$ we have $A\mathbf{e}(\mathbf{x}) = A\mathbf{x} - \mathbf{f}$. Hence, the coarse-grid correction can be algorithmized as

$$\mathbf{x} \leftarrow \mathbf{x} - p(p^T Ap)^{-1} p^T (A\mathbf{x} - \mathbf{f})$$

and the variational two-level algorithm with post-smoothing step proceeds as follows:

Algorithm 1

1. *Pre-smooth:* $\mathbf{x} \leftarrow \mathcal{S}_{pre}(\mathbf{x}, \mathbf{f})$,
2. *evaluate the residual:* $\mathbf{d} = A\mathbf{x} - \mathbf{f}$,
3. *restrict the residual:* $\mathbf{d}_2 = p^T\mathbf{d}$,
4. *solve a coarse-level problem* $A_2\mathbf{v} = \mathbf{d}_2, \quad A_2 = p^T Ap$,
5. *correct the approximation* $\mathbf{x} = \mathbf{x} - p\mathbf{v}$,
6. *post-smooth* $\mathbf{x} = \mathcal{S}_{post}(\mathbf{x}, \mathbf{f})$.

Here, $\mathcal{S}_{pre}(.,.)$ and $\mathcal{S}_{post}(.,.)$, respectively, represent one or more iterations of point-wise iterative methods for solving (1).

The coarse-grid correction vector $\mathbf{v}$ is chosen to minimize the error after Step 5 of Algorithm 1. Thus, we conclude that in the case of a standard variational multigrid, the coarse-grid correction procedure minimizes the error in an intermediate stage of the iteration, while we are in fact interested in minimizing the final error after accomplishing the entire iteration. This means to minimize the error after coarse-grid correction with subsequent smoothing.

3. The smoothed aggregation two-level method

In the smoothed aggregation method, we construct the coarse-grid correction to minimize the error *after coarse-grid correction with subsequent smoothing*, which means the final error on the exit of the iteration procedure. The minimization of the error after pre-smoothing, coarse-grid correction and post-smoothing then follows immediately by a trivial argument.

Let S be the error propagation operator of the post-smoother $\mathcal{S}(.,.) = \mathcal{S}_{post}(.,.)$. Throughout this section we assume that S is sparse. This is due to the fact that the above minimization problem leads to smoothed prolongator $P = Sp$ and we need a sparse coarse-level matrix $A_2 = P^T AP$. The additive point-wise smoothing methods have, in general, sparse error propagation operator; this is the case of Jacobi method or Richardson's iteration.

For a multilevel method with post-smoothing only, the error after coarse-grid correction and subsequent smoothing is given by

$$S(\mathbf{e} - p\mathbf{v}), \tag{4}$$

where $\mathbf{v}$ is a correction vector and $\mathbf{e}$ the error on the entry of the iteration procedure. We choose $\mathbf{v}$ so that the error in (4) is minimal in A-norm, that is, we solve the minimization problem

$$\text{find } \mathbf{v} \in \mathbb{R}^m \text{ such that } \|S(\mathbf{e} - p\mathbf{v})\|_A \text{ is minimal.} \tag{5}$$

Since $\|S(\mathbf{e} - p\mathbf{v})\|_A = \|\mathbf{e} - p\mathbf{v}\|_{S^TAS}$, the minimum is attained for $\mathbf{v}$ satisfying

$$\langle S^TAS(\mathbf{e} - p\mathbf{v}), p\mathbf{w}\rangle = 0 \ \forall \mathbf{w} \in \mathbb{R}^m.$$

We have $\langle S^TAS(\mathbf{e} - p\mathbf{v}), p\mathbf{w}\rangle = \langle p^TS^TAS(\mathbf{e} - p\mathbf{v}), \mathbf{w}\rangle$, hence the above identity is equivalent to $p^TS^TASp\mathbf{v} = p^TS^TAS\mathbf{e}$ and setting $P = Sp$, it becomes

$$P^TAP\mathbf{v} = P^TAS\mathbf{e}. \tag{6}$$

Here, $\mathbf{e}$ is the error on the entry of the iteration procedure. Assume for now that P is injective. Then by (6), we have $\mathbf{v} = (P^TAP)^{-1}P^TAS\mathbf{e}$ and the error after coarse-grid correction and subsequent smoothing is given by

$$S(\mathbf{e} - p\mathbf{v}) = S\left[\mathbf{e} - p(P^TAP)^{-1}P^TAS\mathbf{e}\right] = \left[I - P(P^TAP)^{-1}P^TA\right]S\mathbf{e}. \tag{7}$$

By comparing the operator

$$E = \left[I - P(P^TAP)^{-1}P^TA\right]S \tag{8}$$

on the right-hand side of (7) with (3), we identify E as the error propagation operator of the variational multigrid with smoothed prolongator $P = Sp$ and pre-smoothing step given by $\mathbf{x} \leftarrow \mathcal{S}(\mathbf{x}, \mathbf{f})$. The algorithm is as follows:

Algorithm 2

1. *Pre-smooth:* $\mathbf{x} \leftarrow \mathcal{S}(\mathbf{x}, \mathbf{f})$,
2. *evaluate the residual:* $\mathbf{d} = A\mathbf{x} - \mathbf{f}$,
3. *restrict the residual:* $\mathbf{d}_2 = P^T\mathbf{d}$,
4. *solve the coarse-level problem:* $A_2\mathbf{v} = \mathbf{d}_2, \quad A_2 = P^TAP$,
5. *correct the approximation:* $\mathbf{x} \leftarrow \mathbf{x} - P\mathbf{v}$.

Remark 3.1 Note that in the process of the deriving the algorithm in (7), our post-smoother have become a pre-smoother. Nothing was lost in that process; the algorithm minimizes the final error and takes into account the pre-smoother.

Remark 3.2 The smoothed prolongator $P = Sp$ is potentially non-injective, hence the coarse-level matrix $A_2 = P^T AP$ is potentially singular. In this case, we need to replace the inverse of $P^T AP$ in (7) by a pseudo-inverse.

We summarize our considerations in the form of a theorem.

Theorem 3.3 *The error propagation operator E in (8) (the error propagation operator of Algorithm 2) satisfies the identity*

$$\|E\mathbf{e}\|_A = \inf_{\mathbf{v}\in\mathbb{R}^m} \|S(\mathbf{e} - p\mathbf{v})\|_A$$

for all $\mathbf{e} \in \mathbb{R}^n$.

Proof. The proof follows directly from the fact that Algorithm 2 was derived from variational objective (5). □

Remark 3.4 One may also start with the variational objective to minimize the final error after performing the pre-smoothing, the coarse-grid correction and the post-smoothing. Such extension is trivial, the pre-smoother has no influence on the coarse-grid correction operator $I - P(P^T AP)^{-1}P^T A$ and influences only its argument. Indeed, asuming the error propagation operator of the pre-smoother is S^* (the A-adjoint operator), the final error is given by $S(S^*\mathbf{e} - p\mathbf{v})$ and we solve the minimization problem

$$\text{for } \mathbf{e} \in \mathbb{R}^n \text{ find } \mathbf{v} \in \mathbb{R}^m \ : \ \|S(S^*\mathbf{e} - p\mathbf{v})\|_A \text{ is minimal.} \tag{9}$$

Fundamentally, this is the same minimization problem as (5); to derive the corresponding algorithm, it is simply sufficient to follow our manipulations from (5) to (7) with $\mathbf{e} \leftarrow S^*\mathbf{e}$. This way, we end up with a two-level method that has the error propagation operator

$$E = \left[I - P(P^T AP)^{-1}P^T A\right] SS^*, \tag{10}$$

(see (3)) that is, with the algorithm

Algorithm 3

1. *Pre-smooth:* $\mathbf{x} \leftarrow \mathcal{S}_t(\mathbf{x}, \mathbf{f})$*, where $\mathcal{S}_t$ is an iterative method with error propagation operator S^*,*
2. *pre-smooth:* $\mathbf{x} \leftarrow \mathcal{S}(\mathbf{x}, \mathbf{f})$*, where $\mathcal{S}$ is an iterative method with error propagation operator S,*
3. *evaluate the residual:* $\mathbf{d} = A\mathbf{x} - \mathbf{f}$*,*
4. *restrict the residual:* $\mathbf{d}_2 = P^T\mathbf{d}$*,*
5. *solve the coarse-level problem:* $A_2\mathbf{v} = \mathbf{d}_2, \quad A_2 = P^T AP$*,*
6. *correct the approximation:* $\mathbf{x} \leftarrow \mathbf{x} - P\mathbf{v}$*.*

We summarize the content of Remark 3.4 as a theorem.

Theorem 3.5 *The error propagation operator (10) of Algorithm 3 satisfies the identity*

$$\|E\mathbf{e}\|_A = \inf_{\mathbf{v}\in\mathbb{R}^m} \|S(S^*\mathbf{e} - p\mathbf{v})\|_A$$

for all $\mathbf{e} \in \mathbb{R}^n$.

Proof. The proof follows directly from the fact that Algorithm 3 was derived from variational objective (9). □

Remark 3.6 Our manipulations hold equally for a general pre-smoother with error propagation operator $M \neq S^*$, simply by replacing $S^* \leftarrow M$. The error propagation operator M has no influence on the coarse-space and thus it does not have to be sparse.

4. Numerical example

To demonstrate the robustness of smoothed aggregation method, we consider the algorithm of [6] which is a modification of the method proposed and analyzed in [8] and [10]. Its relationship to Algorithm 2 is obvious. This method uses the smoothing iterative method $\mathcal{S}(\cdot,\cdot)$ which is a sequence of Richardson's iterations with carefully chosen iteration parameters. The error propagation operator S of the smoother $\mathcal{S}(\cdot,\cdot)$ is therefore a polynomial in the matrix A.

In this method, we use massive smoother S and a small coarse-space resulting in sparse coarse-level matrix.

Let $\bar{\lambda} \geq \varrho(A)$ and d be the desired degree of the smoothing polynomial S. We set

$$\alpha_i = \left[\frac{\bar{\lambda}}{2}\left(1 - \cos\frac{2i\pi}{2d+1}\right)\right]^{-1}, \quad i = 1,\ldots,d, \tag{11}$$

$$S = (I - \alpha_1 A)\ldots(I - \alpha_d A) \tag{12}$$

and

$$P = Sp.$$

Here, p is a *tentative prolongator* given by generalized unknowns aggregation. The simplest aggregation method is described in this section.

The smoother S is chosen to minimize $\varrho(S^2A)$. The reason for this comes from the fact that the convergence of the method of [6] is guided by the constant C in the weak approximation condition

$$\forall \mathbf{e} \in \mathbb{R}^n \; \exists \mathbf{v} \in \mathbb{R}^m \;:\; \|\mathbf{e} - p\mathbf{v}\| \leq \frac{C}{\sqrt{\varrho(S^2A)}}\|\mathbf{e}\|_A. \tag{13}$$

The smaller $\varrho(S^2A)$, the easier it becomes to satisfy (13) with a reasonable (sufficiently small) constant. It holds that ([6])

$$\bar{\lambda}_{S^2A} \equiv \frac{\bar{\lambda}}{(1+2d)^2} \geq \varrho(S^2A). \tag{14}$$

The aggregates $\{\mathcal{A}_j\}$ are sets of fine-level degrees of freedom that form a disjoint covering of the set of all fine-level degrees of freedom. For example, we can choose aggregates to form a decomposition of the set of degrees of freedom induced by a geometrically reasonable partitioning of the computational domain. For standard discretizations of scalar elliptic problems, the tentative prolongator matrix p is the $n \times m$ matrix ($m =$ the number of aggregates)

$$p_{ij} = \begin{cases} 1 & \text{if } i \in \mathcal{A}_j, \\ 0 & \text{otherwise,} \end{cases} \tag{15}$$

that is, the j-th column is created by restricting a vector of ones onto the j-th aggregate, with zeroes elsewhere. Thus, the aggregation method can be viewed as a piecewise constant coarsening in a discrete sense. The generalized aggregation method, suitable for non-scalar elliptic problems (like that of linear elasticity), is described in [15].

Algorithm 4 *Given the degree d of the smoothing polynomial $S = pol(A)$, the smoothed prolongator $P = Sp$ where p is the tentative prolongator and the prolongator smoother S is given by (12), the upper bound $\bar{\lambda} \geq \varrho(A)$ and a parameter $\omega \in (0, 1)$, one iteration of the two-level algorithm*

$$\mathbf{x} \leftarrow TG(\mathbf{x}, \mathbf{f})$$

proceeds as follows:

1. *perform*
$$\mathbf{x} \leftarrow \mathbf{x} - \frac{\omega}{\bar{\lambda}_{S^2A}} S^2(A\mathbf{x} - \mathbf{f}),$$
where $\bar{\lambda}_{S^2A}$ is given by (14) and S by (12),
2. *perform the iteration with symmetric error propagation operator S given by (12), that is,*
for $i = 1, \ldots, d$ do
$$\mathbf{x} \leftarrow (I - \alpha_i A)\,\mathbf{x} + \alpha_i \mathbf{f},$$
3. *evaluate the residual $\mathbf{d} = A\mathbf{x} - \mathbf{f}$,*
4. *restrict the residual $\mathbf{d}_2 = P^T\mathbf{d}$,*
5. *solve the coarse-level problem $A_2\mathbf{v} = \mathbf{d}_2$, $A_2 = P^TAP$,*
6. *correct the approximation $\mathbf{x} \leftarrow \mathbf{x} - P\mathbf{v}$,*
7. *for $i = 1, \ldots, d$ do*
$$\mathbf{x} \leftarrow (I - \alpha_i A)\,\mathbf{x} + \alpha_i \mathbf{f},$$
8. *perform*
$$\mathbf{x} \leftarrow \mathbf{x} - \frac{\omega}{\bar{\lambda}_{S^2A}} S^2(A\mathbf{x} - \mathbf{f}).$$

512 000 dofs, coarse space 512 dofs, $\deg(S) = 7$, $H/h = 9$.		
ε	rate of conv. q_N	no. iter. N
1000	0.321	19
100	0.241	15
10	0.137	11
1	0.131	11
0.1	0.221	14
0.01	0.317	19
0.001	0.300	18

Table 1: 3D anisotropic problem

Thus, Algorithm 4 is a symmetrized version of Algorithm 2 with added smoothing in steps 1 and 8.

It is generally believed that in order to solve efficiently an anisotropic problem, one has to perform coarsening only by following *strong connections*. This technique is called *semi-coarsening*. In our case, we form aggregates by coarsening by a factor of 10 in all 3 spatial directions, which means, we do not perform semi-coarsening. Despite of this fact, our method gives satisfactory results regardless of the anisotropy coefficient ε. In this experiment, the symmetric Algorithm 4 is used as a conjugate gradient method preconditioner.

Test problem

- Problem:

$$-\left(\frac{\partial^2}{\partial x^2} + \varepsilon\frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}\right) u = f \text{ on } \Omega = (0,1)^3,\ u = 0 \text{ on } \partial\Omega. \tag{16}$$

- Mesh: $82 \times 82 \times 82$ regular square mesh, 512 000 unconstrained degrees of freedom.
- Aggregates: cubic groups of $10 \times 10 \times 10$ unconstrained vertices.
- Coarse-space size: 512 degrees of freedom.
- Degree of smoothing polynomial: 7.
- Stopping criterion: relative residual $< 10^{-9}$.

The results are summed up in Table 1. Note that here, the estimate of the rate of convergence after N iterations is defined as

$$q_N = \left(\|A\mathbf{x}^N - \mathbf{f}\|/\|A\mathbf{x}^0 - \mathbf{f}\|\right)^{\frac{1}{N}}.$$

Here, $\mathbf{x}^i$ denotes the i-th iteration.

Acknowledgements

This work was sponsored by the TAČR (Technologická Agentura České Republiky) grant TA01020352, ITI (Institut Teoretické Informatiky) grant 1R0545, Department of the Navy Grant N62909-11-1-7032.

References

[1] Bramble, J. H., Pasciak, J. E., Wang, J., and Xu, J.: Convergence estimates for multigrid algorithms without regularity assumptions. Math. Comp. **57** (1991).

[2] Xu, J. and Zikatanov, L. T.: The method of alternating projections and the method of subspace corrections in Hilbert space. J. Amer. Math. Soc. **15** (2002).

[3] Vassilevski, P. S.: *Multilevel Block Factorization Preconditioners*. Matrix-Based Analysis and Algorithms for Solving Finite Element Equations, Springer, New York, 2008.

[4] Brandt, A.: Algebraic multigrid theory: the symmetric case. Appl. Math. Comput. **19** (1986).

[5] Ciarlet, P. G.: *The finite element method for elliptic problems.* Series "Studies in Mathematics and its Applications", North-Holland, Amsterdam, 1978.

[6] Brousek, J., Fraňková, P., Kopincová, H., Kužel, R., Tezaur, R., Vaněk, P., and Vastl, Z.: *An overview of multilevel methods with aggressive coarsening and massive polynomial smoothing*, in preparation.

[7] Křížková, J. and Vaněk, P.: Two-level preconditioner with small coarse grid appropriate for unstructured meshes. Numer. Linear Algebra Appl. **3**(4) (1996).

[8] Guillard, H., Kužel, R., Vaněk, P., and Vastl, Z.: An alternative to domain decomposition methods based on polynomial smoothing. Numer. Linear Algebra Appl., submitted.

[9] Vaněk, P.: Smoothed prolongation multigrid with rapid coarsening and massive smoothing. To appear in Appl. Math.

[10] Vaněk, P., Brezina, M., and Tezaur, R.: Two-grid method for linear elasticity on unstructured meshes. SIAM J. Sci Comput. **21** (1999).

[11] Brezina, M., Heberton, C., Mandel, J., and Vaněk, P.: An iterative method with convergence rate chosen a priori. UCD/CCR Report no. 140, 1999.

[12] Vaněk, P., Mandel, J., and Brezina, M.: Algebraic multigrid by smoothed aggregation for second and fourth order elliptic problems. Computing **56** (1996).

[13] Vaněk, P.: Acceleration of convergence of a two-level algorithm by smoothing transfer operator. Appl. Math. **37** (1992).

[14] Vaněk, P.: Fast multigrid solver. Appl. Math. **40**(1) (1995).

[15] Vaněk, P., Brezina, M., and Mandel, J.: Convergence of algebraic multigrid based on smoothed aggregations. Numer. Math. **88**(3) (2001).

[16] Vaněk, P., Brezina, M: Nearly optimal convergence result for multigrid with aggressive coarsening and polynomial smoothing. Appl. Math, to appear.

[17] Brezina, M., Vaněk, P., Vassilevski, P. S.: An improved convergence analysis of the smoothed aggregation algebraic multigrid. Numer. Linear Algebra Appl., to appear.

Conference Applications of Mathematics 2013
in honor of the 70th birthday of Karel Segeth.
Jan Brandts, Sergey Korotov, Michal Křížek,
Jakub Šístek, and Tomáš Vejchodský (Eds.),
Institute of Mathematics AS CR, Prague 2013

INTERPLAY OF SIMPLE STOCHASTIC GAMES AS MODELS FOR THE ECONOMY

Ubaldo Garibaldi[1], Tijana Radivojević[2], Enrico Scalas[2,3]

[1] IMEM-CNR, Physics Department, Genoa University
Via Dodecaneso 33, 16146 Genoa, Italy
garibaldi@fisica.unige.it
[2] BCAM - Basque Center for Applied Mathematics
Alameda de Mazarredo 14, 48009 Bilbao, Basque Country, Spain
tradivojevic@bcamath.org
[3] Dipartimento di Scienze e Innovazione Tecnologica, Università del Piemonte Orientale
"Amedeo Avogadro"
Viale Michel 11, 15121 Alessandria, Italy
enrico.scalas@mfn.unipmn.it

Abstract

Using the interplay among three simple exchange games, one may give a satisfactory representation of a conservative economic system where total wealth and number of agents do not change in time. With these games it is possible to investigate the emergence of statistical equilibrium in a simple pure-exchange environment. The exchange dynamics is composed of three mechanisms: a decentralized interaction, which mimics the pair-wise exchange of wealth between two economic agents, a failure mechanism, which takes into account occasional failures of agents and includes wealth redistribution favoring richer agents, and a centralized mechanism, which describes the result of a redistributive effort. According to the interplay between these three mechanisms, their relative strength, as well as the details of redistribution, different outcomes are possible.

> ...But Mr. Lebeziatnikov who keeps up with modern ideas explained the other day that compassion is forbidden nowadays by science itself, and that that's what is done now in England, where there is political economy...
>
> *Crime and Punishment, Chapter 1, Fyodor Dostoevsky*

1. Introductory considerations

In economics, distributional problems emerge in contexts of economic growth and allocation of resources, among others. Distributions of relevant economic variables are important for policy making purposes, however, there is a general policy problem that was emphasized by Federico Caffè [1]:

> ...when, in economic reasoning, the social wealth distribution is assumed 'given', this means that the existing distribution is accepted, without evaluating whether it is good or bad, acceptable or unacceptable... this must be explicitly done further clarifying that the conclusions are conditioned on the acceptability of the distributional set-up.

In the past, we have studied simple exchange mechanisms (games) based on exact probabilistic dynamics and leading to statistical equilibrium distributions [3, 4, 5, 7]. Our aim is to give a satisfactory representation of an economy and investigate the statistical equilibrium by means of interplay between these mechanisms. One of them describes centralized activities in terms of taxation and redistribution of wealth and it produces exponential tails. In order to include a process that gives power-law tails we involve a mechanism consisting of occasional failures of agents together with redistribution of their wealth. This mechanism is discussed in [3], Chapter 10, and it leads to the Yule distribution, which was originally proposed by Yule to account for the data on biological species [10]. It was the idea of Simon to use it in order to describe a class of distributions that appears in a wide range of empirical data, including economic phenomena [8]. The last but crucial mechanism represents pairwise exchange of wealth between agents and its interplay ensures that the system does not break down due to the failures of agents.

Since it is difficult to study the interplay of the three games analytically, we want to develop a statistical procedure based on statistical inference from the data to obtain relevant distributional properties of economic variables. In particular, we focus on the wealth distribution. However, here the emphasis will be on modelling and studying the aggregate wealth distribution in a conservative system where the number of agents and the total wealth do not change in time. For a simple trading rule in an active market, where number of agents and money is not conserved, Kusmartsev found that the wealth distribution has a general Bose-Einstein form, whose parameters depend on wealth exchange parameter, i.e. activity of agents [6]. Regarding conservation of wealth in the system, as pointed out in [9], ordinary agents in an economy can only exchange money with each other, so there is a local conservation of wealth. However, a government or a central bank can cause a change of wealth, but as long as it does not cause hyperinflation, the system can be close to statistical equilibrium, with slowly changing parameters.

1.1. Descriptions for the state of the system

The basic random variables for the description of the games are introduced in [3]. A general framework for agent-based models consists of the allocation of n objects among g agents (categories). Categories represent economic agents and objects may represent money or wealth. In this paper objects will be called coins.

The most complete description of the states is in terms of individual (coin) configurations $\mathbf{X}$. An *individual description* $\mathbf{X} = (X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$, where $x_i \in \{1, \ldots, g\}$, is a list telling us, for each coin to which agent it belongs.

The total number of configurations for n coins distributed among g agents is g^n. A *statistical description* $\mathbf{Y} = (Y_1 = n_1, Y_2 = n_2, \ldots, Y_g = n_g)$ is a list giving us the number of coins for each agent, with the constraint $\sum_1^g n_i = n$. The total number of these agent descriptions for g agents sharing n coins is $\binom{n+g-1}{n}$. A *partition description* $\mathbf{Z} = (Z_0 = z_0, Z_1 = z_1, \ldots, Z_n = z_n)$ is the number of agents with zero coins, one coin, etc., with the constraints for $\mathbf{Z}$ $\sum_0^n z_i = g$, $\sum_0^n i z_i = n$. This is the less complete description, commonly referred to as *wealth distribution* and it will be mostly used throughout this paper. Figure 1 shows a state of a system and illustrates the meaning of the various descriptions.

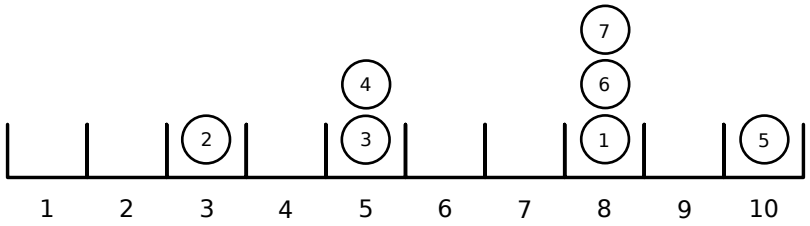


Figure 1: A state of a system with ten agents and seven coins. The individual description of the state is $x_1 = 8,\ x_2 = 3,\ x_3 = 5, \ldots, x_7 = 8$. The statistical description is $y_1 = y_2 = y_4 = y_6 = y_7 = y_9 = 0,\ y_3 = y_{10} = 1,\ y_5 = 2,\ y_8 = 3$ and the wealth distribution in this case is $z_0 = 6,\ z_1 = 2,\ z_2 = 1,\ z_3 = 1,\ z_4 = z_5 = z_6 = z_7 = 0$.

2. Simple exchange games

2.1. Random coin exchange (Bennati-Drăgulescu-Yakovenko)

Bennati-Drăgulescu-Yakovenko (BDY) model was introduced in Bennati's work (1988, 1993) and rediscovered in [2]. Later it was studied in [7]. It is a discrete model, where number of agents and wealth measured by coins are conserved. The BDY game is played as follows. In the system of g agents sharing n coins, at each time step, two agents are randomly selected. The selection is such that each pair of agents has equal probability to be chosen. One of the agents (randomly chosen) becomes the loser and gives one coin to the other, who becomes the winner. Indebtedness is not possible, i.e. if the loser has zero coins, the move is not taken into account and a new pair of players is selected. In order to avoid null moves, the game can be formulated in the following way – a loser is chosen randomly from the agents who have at least one coin and the winner is chosen among all agents, randomly as well. In case the loser and the winner coincide, there will be no change in the state of the system.

The appropriate description of the system is the statistical description, in terms of agents. Let assume that at a given time t, the agents are described by the state $\mathbf{Y}_t = (n_1, ..., n_g) := \mathbf{n}$ and at the next step by the $\mathbf{Y}_{t+1} = (n_1, ..., n_i - 1, ..., n_j + 1, .., n_g) := \mathbf{n}_i^j$, that corresponds to a loss of the agent i and a win of the agent j.

The transition between these states follows a homogenous Markov dynamics with transition probability:

$$\mathbb{P}(\mathbf{n}_i^j|\mathbf{n}) = \frac{1-\delta_{n_i,0}}{g-z_0(\mathbf{n})}\frac{1}{g}, \tag{1}$$

where $\delta_{n_i,0}$ is the usual Kronecker's delta equal to 1 for $n_i = 0$ and zero otherwise. The first part is the probability of selecting a loser (agent i) from all agents with at least one coin and the second part $(1/g)$ is the probability that the agent j is the winner. The sequence $\mathbf{Y}_0, \mathbf{Y}_1, \ldots$ is a finite Markov chain with irreducible set of states and no periodicity. Therefore, an invariant probability distribution exists and it coincides with the equilibrium distribution. Its form is the following:

$$\pi(\mathbf{n}) = C \cdot (g - z_0(\mathbf{n})), \quad C = \left[\sum_{k=1}^{g} \binom{g}{k}\binom{n-1}{n-k}\right]^{-1} \tag{2}$$

and it can be derived by means of detailed balance given that the chain is reversible. The exact solution of this problem is not as simple as it appears in [2]. One can see that the invariant probability distribution for this model depends on number of agents with zero coins (z_0); more precisely it is proportional to the number of agents with at least one coin in their pocket, and hence, it is not uniform.

At the beginning of each simulation in this paper n/g coins are given to the each agent, i.e. the initial wealth distribution is a Dirac delta, $\delta(n/g - i)$ for $i = 0, 1, \ldots,$. Results of simulations present expected wealth distributions, given in terms of partition vector $\mathbf{Z} = (Z_0, Z_1, \ldots, Z_n)$, namely time means of relative frequencies of agents with i coins, that approximate $\mathbb{E}(Z_i)/g$.

Figure 2 shows the expected wealth distribution in the system with dynamics given by the BDY game. Time means of relative frequencies of agents with $0, 1, \ldots, 500$ coins obtained from simulations are compared with theoretical expected wealth distribution given by exact formulas and with exponential distribution, which is the distribution in the limit of large density and large number of agents ($n \gg g \gg 1$). A detailed derivation of the expected wealth distribution can be found in [7] and in general, it is not exponential – it becomes exponential only in the appropriate limit. Therefore, conclusions from [2] on exponential wealth distribution are not fully correct if one considers equation (2).

2.2. Taxation and redistribution

The second exchange game mimics taxation and redistribution in a simplified way. The taxation-redistribution model was introduced in [4]. There are still n coins to be allocated among g agents and n and g are conserved in time.

The simplest form of this game consists in taking a coin from one agent and redistributing it to another agent. Taxation is represented by a step where a coin is taken from an agent and temporarily removed from the population, so the state of the system is

$$\mathbf{n}_i := (n_1, ..., n_i - 1, ..., n_g).$$

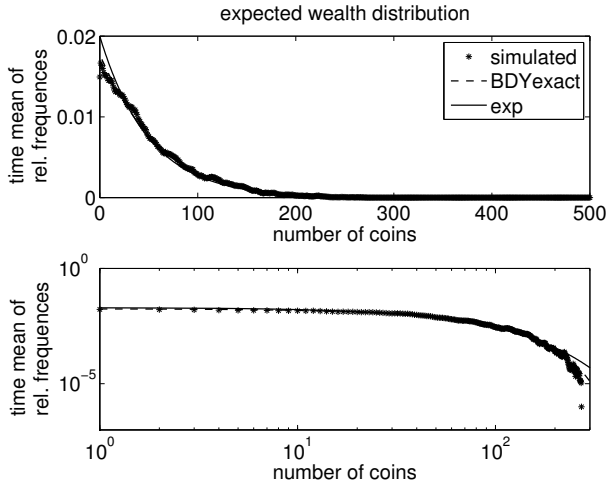


Figure 2: Time mean of relative frequencies of $g = 10$ agents sharing $n = 500$ coins obtained from simulation of the BDY game (stars) compared with theoretical values for expected wealth distribution (dashed line) and with distribution in the limit of large systems, Exp(g/n) (line), shown in linear (up) and logarithmic scale (down). The values of the random variables $\mathbf{Z}_t$ were sampled and averaged over 10^5 of Monte Carlo steps, after an equilibration of 10^4 steps.

By redistribution, one means a step where a coin is given back to an agent, i.e.

$$\mathbf{n}^j := (n_1, ..., n_j + 1, ..., n_g).$$

If the system is in the state $\mathbf{Y}_t = (n_1, ..., n_g) := \mathbf{n}$, at the next step of this game, possible values of $\mathbf{Y}_{t+1}$ will be $\mathbf{Y}_{t+1} = (n_1, ..., n_i - 1, ..., n_j + 1, .., n_g) := \mathbf{n}_i^j$, corresponding to a loss of the ith agent due to taxation and a gain of the jth agent due to redistribution. The transition probability between these states is:

$$\mathbb{P}(\mathbf{n}_i^j|\mathbf{n}) = \frac{n_i}{n}\frac{\alpha_j + n_j - \delta_{i,j}}{\theta + n - 1}, \tag{3}$$

where $(\alpha_1, ..., \alpha_g)$ are weights for redistribution and $\theta = \sum_j \alpha_j$. Taxation consists of random selection of a coin, in opposition to the BDY move where the selection refers to agents. If a coin is randomly selected out of n coins, the probability of selecting a coin belonging to the agent i is n_i/n, where n_i is the number of coins of the agent i. Hence, the agents are taxed proportionally to their wealth. Then, this coin is redistributed to the agents following the rule that the jth agent will receive a coin with probability proportional to $\alpha_j + n_j$, where n_j is the number of coins of the agent j and α_j is a suitable weight. The redistribution policy is determined by the values of α_j. Positive values make rich agents richer, and the effect is larger the smaller is α_j, $\alpha_j \to \infty$ determines a redistribution where all agents are equivalent, so the redistribution mechanism becomes random, whereas negative values of α_j tend to favor poor agents.

Equation (3) defines the transition probability matrix of an irreducible Markov chain which is also aperiodic. Hence, there exists an invariant probability distribution, which coincides with the equilibrium distribution. In this case it is the Pólya distribution:

$$\pi(\mathbf{n}) = \frac{n!}{\theta^{[n]}} \prod_{i=1}^{g} \frac{\alpha_i^{[n_i]}}{n_i!} \tag{4}$$

and it can be derived by means of detailed balance given that the chain is reversible. $x^{[n]}$ is the Pochhammer symbol representing the rising (or upper) factorial defined by $x^{[n]} = x(x+1)...(x+n-1)$.

Let us suppose that $\alpha_j = \alpha$ for all j. Depending on the choice of α, one can obtain different equilibrium situations. Marginalizing equation (4) for a single agent (all the agents follow the same probability distribution) in the continuous limit of large systems, for α positive, the taxation and redistribution model is approximately described by the gamma distribution (see [3], Chapter 5), whose form factor is just the initial redistribution weight. If α is negative, then the limiting distribution is the hypergeometric distribution and in the case $\alpha \to \infty$ it is the Poisson distribution. Note that in the case of equidistributed agents, one has that $\mathbb{E}(Z_i) = g\mathbb{P}(Y_1 = i)$; in other words, the knowledge of the marginal occupation distribution immediately gives the expected wealth distribution.

Instead of taxation and redistribution of only one coin at each time step, we will consider *block taxation* in which $m \leq n$ coins are randomly taken from agents and then redistributed according to the same probability, i.e. proportionally to the actual wealth of agents and the chosen weight for redistribution. Block taxation can be written in the following way

$$\mathbf{n}' = \mathbf{n} - \mathbf{m} + \mathbf{m}', \tag{5}$$

where $\mathbf{n} = (n_1, \ldots, n_g)$ is the initial agent description, $\mathbf{m} = (m_1, \ldots, m_g)$ is the taxation vector and $\mathbf{m}' = (m'_1, \ldots, m'_g)$ is the redistribution vector, with the constraints $\sum_{i=1}^{g} m_i = m$ and $\sum_{j=1}^{g} m'_j = m$. This leads to the same equilibrium distribution, given by the equation (4).

Figure 3 presents the expected wealth distribution in the system with dynamics governed by the taxation and redistribution game. It shows time mean of relative frequencies of agents with $0, 1, \ldots, 500$ coins obtained from simulations, compared with theoretical values for the expected wealth distribution and with gamma distribution which is the distribution in the continuous limit of large systems.

2.3. Zipf-Simon-Yule

The third important game is the Zipf-Simon-Yule one, initially described in [5]. In order to consider a system that is conservative, in terms of total wealth and number of agents, as in previous two games, we will modify the given Zipf-Simon-Yule model. This game includes a failure probability, which is independent of agents' wealth. An

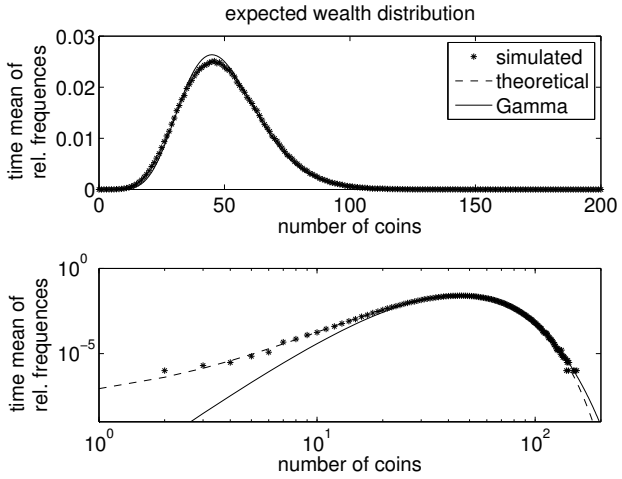


Figure 3: Expected wealth distribution in a system of $g=10$ agents and $n=500$ coins for the 250-block taxation and redistribution game with redistribution weight $\alpha = 10$. Time mean of relative frequencies of agents with i coins obtained from simulation (stars) are compared with theoretical expected wealth distribution (dashed line) and with distribution in the limit of large systems, Gamma$(\alpha, n/\alpha g)$ (line), in linear (up) and logarithmic scale (down). Values from simulation were sampled and averaged over 10^5 Monte Carlo steps, after an equilibration of 10^4 steps.

agent with coins is randomly selected and all his coins are removed. Therefore, the probability of failure for the ith agent is:

$$\mathbb{P}(\mathbf{n}_{(i)}|\mathbf{n}) = \frac{1}{g - z_0(\mathbf{n})}\mathbf{1}_{\{n_i>0\}}, \tag{6}$$

where $\mathbf{n}_{(i)} = (n_1, \ldots, 0, \ldots, n_g)$ is an agent description vector with a zero element on the ith position. The wealth of the failed agent is then redistributed to the agents with probability proportional to their actual wealth, but the last coin is given back to the failed agent. This move can be regarded as a sort of "compassionate capitalism". This is a trick to avoid absorbing states; without this move a failed agent should be cancelled out forever, and after g moves the process would stop. If we assume that the ith agent had m coins that are removed, then the probabilities for redistributing each of those m coins are the following:

$$\begin{cases} \mathbb{P}(X_1 = j|\mathbf{n}_{(i)}) = \frac{n_j}{n-m}\,, \\ \vdots \\ \mathbb{P}(X_{s+1} = j|\mathbf{n}_{(i)}, j_1, \ldots, j_s) = \frac{n_j'}{n-m+s}\,, \\ \vdots \\ \mathbb{P}(X_m = j|\mathbf{n}_{(i)}, j_1, \ldots, j_{m-1}) = \delta_{j,i}\,. \end{cases} \tag{7}$$

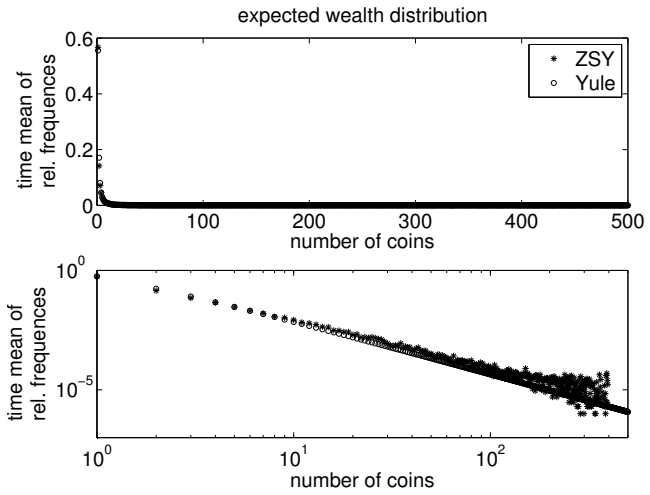


Figure 4: Expected wealth distribution in a system of $g=100$ agents sharing $n=500$ coins, shown in linear (up) and logarithmic scale (down). Stars represent the time mean of relative frequencies of agents with i coins obtained from simulation of the ZSY game, sampled and averaged over 10^4 Monte Carlo steps, after an equilibration of 10^4 steps. Circles represent the Yule distribution with parameter $\rho \approx 1$.

The sequence $\mathbf{Y}_0, \mathbf{Y}_1, \ldots, \mathbf{Y}_n, \ldots$ is a finite Markov chain with an irreducible set of states and no periodicity. Therefore, the invariant and equilibrium distribution exists, but it is not easy to find, because the chain is irreversible, and one cannot apply detailed balance.

The described mechanism alone produces power-law tails. Figure 4 presents the expected wealth distribution obtained from simulation of this game and fitted with a Yule distribution, which is the discrete counterpart of the Pareto distribution and has the following form

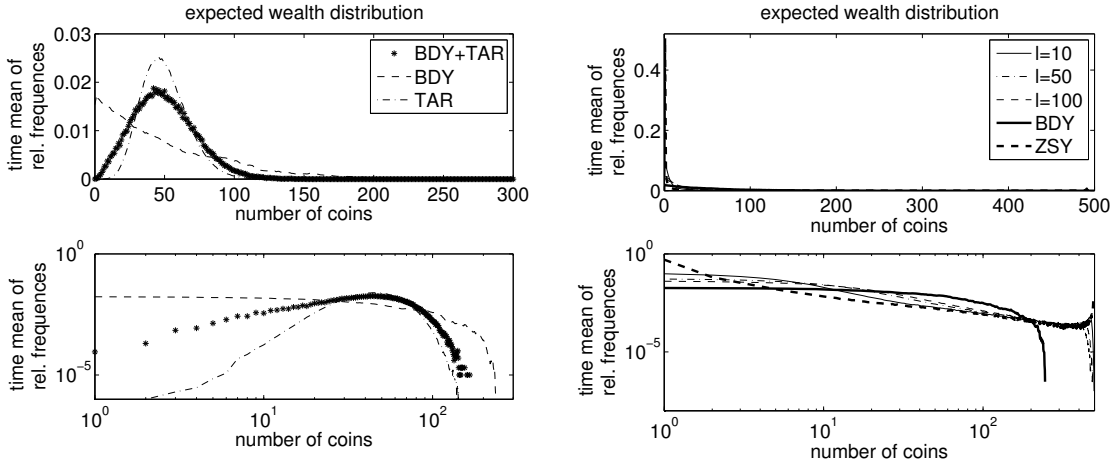
$$\mathbb{P}(U = i) = \rho B(i, \rho + 1) \quad \text{for } i \in \mathbb{N}, \rho \in \mathbb{R}_+ ,$$

where U denotes a random variable and B is the Beta function.

3. "Super-moves"

In the case of the first two models (BDY and TAR), computer simulations are not really necessary because the equilibrium distributions can be analytically derived. On the contrary, they are necessary in the case of the Zipf-Simon-Yule model. However they are necessary if one thinks of a system where two or three games are played sequentially.

In order to know what is the shape of the expected wealth distribution in the equilibrium state in the long-term limit of an economy, we perform Markov chain Monte Carlo simulations for various combinations of three games described in previous section.



(a) (b)

Figure 5: Expected wealth distribution in a system of 10 agents sharing 500 coins for the mixture of the two games. (a) One step consists of $l = 50$ steps of BDY game and one step of 250-block TAR game with redistribution weight $\alpha = 10$. (b) One step consists of one step of ZSY game and $l = 10$, $l = 50$, $l = 100$ steps of BDY game. Time mean of relative frequencies of agents with i coins obtained from simulations of two games are compared with distribution from pure BDY and TAR games (a) and BDY and ZSY games (b). Values from simulation were sampled and averaged over 10^4 Monte Carlo steps, after an equilibration of 10^4 steps.

Let us first consider an alternation of the BDY and TAR game. Suppose that the BDY game is played l times and then one step of the TAR game is performed. The first game shifts the initial expected wealth distribution towards an exponential, and such a distribution becomes the initial one for the second game. Then, the second game shifts the distribution towards a gamma and these steps can be iterated many times. One can guess that the equilibrium distribution of the joint process will be a mixture of the two pure ones, with weights proportional to the frequency of the two processes. This conjecture is qualitatively corroborated in figure 5 for alternating the BDY and TAR games (a) and also for alternating ZSY and BDY games (b), where the resulting expected wealth distribution is compared with distributions from pure games.

In order to mimic what happens in a real economy, we will use the following combination of steps. On each "day" we run a move of BDY, at the end of each "month=l days" we have a ZSY failure, at the end of each "year=k months" we run a TAR, i.e. taxation and redistribution of the coins. After the failure of an agent, modeled by the ZSY game, we have l iterations of coin exchanges between agents, a process giving to the failed agent the opportunity to recover some wealth; in this way the "compassionate capitalism" mechanism can be avoided. These random coin exchanges give the background noise of the economy. Once per year taxation and

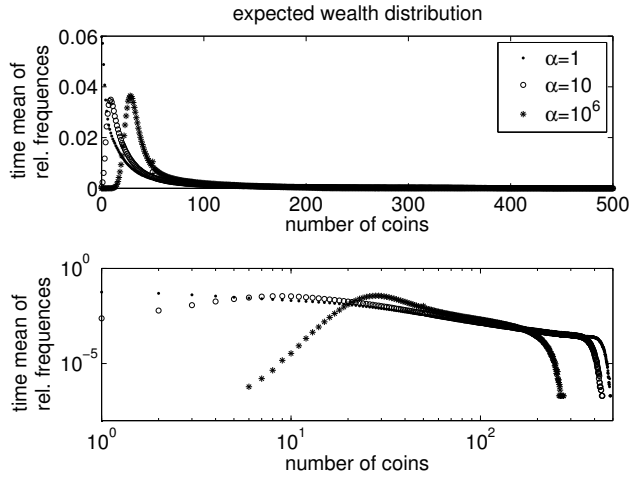


Figure 6: Expected wealth distribution in the system for the mixture of the three games described by the equation (8) with $l = 100$ and $k = 10$. Time mean of relative frequencies of 10 agents sharing 500 coins sampled and averaged over 10^3 realizations of 500 Monte Carlo steps and different redistribution parameter $\alpha = 1$ (dots), $\alpha = 10$ (circles) and $\alpha = 10^6$ (stars).

redistribution of the coins appear as centralized mechanisms depending on weights which characterize the redistribution policy. In summary, the "super-move" for the mixture of the three games at time step t (the last "day" of each "year") can be presented as

$$\mathbf{P}_t = (\mathbb{P}(Y_{1,t} = \cdot), \ldots, \mathbb{P}(Y_{g,t} = \cdot)) = \mathbf{P}_{t-1} \cdot ((BDY^l \cdot ZSY)^k \cdot TAR),\ t = 1, 2, ... \quad (8)$$

and $(BDY^l \cdot ZSY)^k \cdot TAR$ is the stochastic matrix for an irreducible, aperiodic Markov chain representing the described alternation of the three games.

Figure 6 shows the resulting expected wealth distribution in the system with dynamics given by the equation (8). Details of the simulations can be found in the caption.

4. Summary and outlook

In this paper we proposed a representation of a conservative economic system, where total wealth and number of agents do not change, using the interplay among three games, previously described separately. The exchange dynamics in the system is composed of pair-wise interactions between economic agents, a mechanism for occasional failures of agents including redistribution of their wealth and a centralized mechanism, which presents redistribution policy. Depending on the relative strength of these mechanisms, the nature of the interplay between them, the specification of redistribution, various outcomes are possible.

The presented model is general enough to be applied to the description of both aggregate wealth distribution, and to the distribution of firm sizes. It can be extended

in several directions. One can take into account a heterogeneous population of agents and investigate the presence of asymmetry in the initial endowments on the long run dynamics of the model. This case becomes relevant when one wants to describe the aggregate effect of a policy switch between different redistributive regimes. Another extension could include saving propensity and analyze the resulting distributional properties. Even random failures can be easily taken into account. Future work will investigate the correspondence of the model with real data. There is only one probabilistic parameter, namely α, the weight of the redistribution policy, to be estimated. Other parameters, such as l and k are to be considered fully phenomenological.

Acknowledgements

This work was supported by the Italian grant PRIN 2009, 2009H8WPX5_002, *Finitary and non-finitary probabilistic methods in economics*.

References

[1] Caffè, F.: *Lezioni di politica economica*. Bollati-Boringhieri, 1978.

[2] Drăgulescu, A. and Yakovenko, V. M.: Statistical mechanics of money. European Phys. J. B **17** (2000), 723–729.

[3] Garibaldi, U. and Scalas, E.: *Finitary probabilistic methods in econophysics*. Cambridge University Press, Cambridge UK, 2010.

[4] Garibaldi, U., Scalas, E. and Viarengo, P.: Statistical equilibrium in simple exchange games II: the redistribution game. European Phys. J. B **60** (2007), 241–246.

[5] Garibaldi, U., Costantini, D., Donadio, S., and Viarengo, P.: Herding and clustering in economics: the Yule-Zipf-Simon model. Comput. Economics **27**(1) (2006), 115–134.

[6] Kürten, K. E. and Kusmartsev, F. V.: Bose-Einstein distribution of money in a free-market economy II. Europhys. Lett. **93** (2011), 28003.

[7] Scalas, E., Garibaldi, U., and Donadio, S.: Statistical equilibrium in simple exchange games I – methods of solution and application to the Bennati-Dragulescu-Yakovenko (BDY) game. European Phys. J. B **53** (2006), 267–272.

[8] Simon, H. A.: On a class of skew distribution functions. Biometrika **42** (1955), 425–440.

[9] Yakovenko V. M.: Econophysics, statistical mechanics approach to. In: R. A. Meyers (Ed.), *Encyclopedia of Complexity and System Science*, Springer, 2009.

[10] Yule, G. U.: A mathematical theory of evolution, based on the conclusions of Dr J. C.Willis. F.R.S. Phil. Trans. B, 213, 21, 1924.

Conference Applications of Mathematics 2013
in honor of the 70th birthday of Karel Segeth.
Jan Brandts, Sergey Korotov, Michal Křížek,
Jakub Šístek, and Tomáš Vejchodský (Eds.),
Institute of Mathematics AS CR, Prague 2013

NUMERICAL APPROXIMATION OF DENSITY DEPENDENT DIFFUSION IN AGE-STRUCTURED POPULATION DYNAMICS

Luca Gerardo-Giorda

BCAM – Basque Center for Applied Mathematics
Bilbao, Spain
lgerardo@bcamath.org

Abstract

We study a numerical method for the diffusion of an age-structured population in a spatial environment. We extend the method proposed in [2] for linear diffusion problem, to the nonlinear case, where the diffusion coefficients depend on the total population. We integrate separately the age and time variables by finite differences and we discretize the space variable by finite elements. We provide stability and convergence results and we illustrate our approach with some numerical result.

1. Introduction

The mathematical problem describing the spatial dispersal of an age-structured population in a region Ω consists in a reaction-diffusion equation for the population density, together with a given initial condition, an integral condition at age $a = 0$, giving the newborns rate, and boundary conditions on $\partial\Omega$ depending on specific features of the population and of the environment. An almost complete review of the results concerning existence, uniqueness and asymptotic behaviour of the solution of age-structured diffusion models can be found in the book by A. Okubo and S. A. Levin ([10], Sec.10.8).

The earliest age-structured models did not include a spatial distribution of the population density (see e.g. [5]). Under the hypothesis of space homogeneity, the problem reduces to a pure first order hyperbolic partial differential equation, which was naturally solved by integration along characteristics in age and time (see for instance [6, 7, 9]). This integration method entails the use of the same discretization step in age and time. However, the presence of different time scales in the dynamics (which is typically the case when space is involved) suggests the use of different steps in the discretization of time and age. This was the approach followed by A. de Roos in [3], and B. Ayati *et al.* in [1], where an approximation space in age is built by discontinuous piecewise polynomials moving along characteristic lines. In [2] a new approach was introduced for the linear diffusion case, where the age and time variables are decoupled and discretized separately by finite differences, while

the space variable is discretized by finite elements. The problem is advanced in time by semi-implicit scheme, while a parabolic problem in age and space is solved within the single time step.

In plenty of application of practical interest, the diffusion coefficient depends on the total population itself, and the associated problem is nonlinear (see, e.g. [8]). In this paper we present the extension of the method introduced in [2] to the case of nonlinear diffusion coefficients.

The paper is organized as follows. In Section 2 we describe the nonlinear model we are dealing with. In Sections 3 through 5 we present the finite dimensional approximation, and in Section 6 we outline the algorithmic aspects of the procedure. In Section 7 we state the stability and convergence analysis of the method, and in Section 8 we present some numerical results to illustrate our method.

2. Setting of the problem

We consider an age-structured population diffusing in a bounded spatial domain $\Omega \subset \mathbf{R}^d$, $d = 1, 2, 3$, with boundary $\partial\Omega \in C^2$. We denote by $\rho(t, a, x)$ the density per unit space and age of the population at time t, where $a \in [0, a_\dagger]$ and $x \in \Omega$. The population at time t in a given location $x \in \Omega$, and the total population at time t are thus given by

$$p(t,x) = \int_0^{a_\dagger} \rho(t,a,x)\, da, \qquad\qquad P(t) = \int_\Omega p(t,x)\, dx. \tag{1}$$

We assume the diffusion process to be density- and age-driven, namely the diffusion coefficient in (t, x) depends on the population $p(t, x)$ at the corresponding location in space and time, and on the age of the individuals.

Given a final time $T > 0$, the population density $\rho(t, a, x) \in C(0, T; L^2(0, a_\dagger; H^1(\Omega)))$ satisfies the nonlinear model problem

$$\begin{aligned}
&\rho_t + \rho_a - \text{div}\,(k(p(t,x),a)\,\nabla\rho) = f(t,x) - \mu(a)\,\rho \quad \text{in } (0,T)\times(0,a_\dagger)\times\Omega\,,\\
&\rho(0,a,x) = \rho_0(a,x) \qquad\qquad \text{in } (0,a_\dagger)\times\Omega\,,\\
&\rho(t,0,x) = \int_0^{a_\dagger} \beta(a)\,\rho(t,a,x)\; da \qquad\qquad \text{in } (0,T)\times\Omega\,,\\
&k(p(t,x))\,\mathbf{n}\cdot\nabla p = 0 \qquad\qquad \text{on } (0,T)\times(0,a_\dagger)\times\partial\Omega\,,
\end{aligned} \tag{2}$$

where $p(t, x)$ is given in (1), the operators div$(\cdot)$ and $\nabla(\cdot)$ are the standard divergence and gradient operators in Ω, and $\vec{n}$ is the unit vector normal to $\partial\Omega$ pointing outwards.

The coefficients $\mu(a)$ and $\beta(a)$ represent the age-specific mortality and the age-specific fertility, respectively, which are supposed to be non-negative functions of age only. In (2), ρ_0 is the given non-negative initial age distribution, while the integral condition is the so-called renewal condition, providing the newborns rate. Finally, we

consider an isolated environment by choosing a zero-flux boundary condition, which reflects the absence of both immigration and emigration, but other boundary conditions can be considered as well (for instance, an homogeneous Dirichlet boundary condition would model an hostile habitat at the boundary of Ω). We refer to [10] for issues concerning existence and uniqueness of a nonnegative solution of (2).

We impose standard conditions on the diffusion coefficient to ensure ellipticity of the associated bilinear form.

$$k \in L^\infty(\mathbb{R}^+ \times (0, a_\dagger)), \quad 0 < k_0 \leq k(p, a) \leq k_+, \tag{3}$$

and we assume that the age-specific fertility $\beta(\cdot)$ is measurable and essentialy bounded, namely there exists a constant β_+ such that

$$0 \leq \beta(a) \leq \beta_+. \tag{4}$$

Finally, we assume the age-specific mortality $\mu(\cdot)$ to be a measurable function, satisfying

$$\int_0^{a_\dagger} \mu(\sigma) d\sigma = +\infty, \tag{5}$$

in order to guarantee that the probability for an individual to survive at age a, which is defined as

$$\pi(a) = \exp\left(-\int_0^a \mu(\sigma) d\sigma\right), \tag{6}$$

vanishes at the maximum age $a_\dagger$. The numerical issues arising from the unbounded coefficient $\mu(a)$ can be avoided by performing a standard change of variable.

We let $\rho(t, a, x) = \pi(a)u(t, a, x)$, and we reduce ourselves to the problem of finding $u(t, a, x) \in C(0, T; L^2(0, a_\dagger; H^1(\Omega)))$ such that

$$\begin{aligned}
u_t + u_a - \operatorname{div}\left(k(p(t,x), a)\, \nabla u\right) &= f(t,x) && \text{in } (0,T) \times (0, a_\dagger) \times \Omega\,, \\
p(t,x) &= \int_0^{a_\dagger} \pi(a) u(t,a,x)\, da && \text{in } (0,T) \times \Omega\,, \\
u(0,a,x) &= u_0(a,x) && \text{in } (0, a_\dagger) \times \Omega \\
u(t,0,x) &= \int_0^{a_\dagger} m(a) u(t,a,x)\ da && \text{in } (0,T) \times \Omega\,, \\
k(a,x)\, \mathbf{n} \cdot \nabla u &= 0 && \text{on } (0,T) \times (0, a_\dagger) \times \partial\Omega\,,
\end{aligned} \tag{7}$$

where now $u_0(a, x) = \frac{\rho_0(a,x)}{\pi(a)}$, while $m(a) = \frac{\beta(a)}{\pi(a)}$ is the so called maternity function. Notice that $m \in L^\infty(0, a)$ as for all $a \in (0, a)$ we have $m(a) \leq \beta_+$.

We focus here on the numerical treatment of the problem and we assume throughout the paper existence and uniqueness of smooth, nonnegative solutions [10].

3. Time discretization

Let $t^n = n\Delta t$ $(n = 0, 1, \ldots, N_t)$ be a partition of the interval $(0, T)$ into N_t subintervals (for simplicity we consider an uniform discretization, adaptivity in time being

beyond the scope of this paper). We denote with $u^n(a,x)$ and $p^n(x)$ the approximations of $u(t^n,a,x)$ and $p(t^n,x)$, respectively, and we advance in time equation (7) by means of a semi-implicit scheme, where both the initial condition in age and the diffusion coefficient are computed at the previous time step. Moving from t^{n-1} to t^n we solve the following parabolic problem in age and space.

Find $u^n \in L^2(0,a_\dagger;H^1(\Omega))$ such that for all $v \in H^1(\Omega)$

$$\begin{aligned} &\frac{d}{da}\,\langle u^n, v\rangle + A(p^{n-1};a;\,u^n,v) + \frac{1}{\Delta t}(u^n,v) = (f,v) + \frac{1}{\Delta t}(u^{n-1},v) \\ &u^n(0,x) = \int_0^{a_\dagger} m(a)\,u^{n-1}(a,x)\;da, \qquad\qquad p^n(x) = \int_0^{a_\dagger} \pi(a)\,u^n(a,x)\;da, \end{aligned} \tag{8}$$

where $\langle\cdot,\cdot\rangle$ denotes the duality pairing between $H^1(\Omega)$ and $H^{-1}(\Omega)$, and where $A(p^{n-1};a;\cdot,\cdot)$ is the bilinear form given by

$$A(p^{n-1};a;\,w,v) = \int_\Omega k(p^{n-1}(x),a)\nabla w\cdot\nabla v\,dx.$$

By standard coercivity arguments one can prove existence and uniqueness for the solution of (8).

Remark 3.1 *The coercivity and the continuity of the bilinear form* $A(p^{n-1};a;\cdot,\cdot) + \frac{1}{\Delta t}(\cdot,\cdot)$ *are straightforward. Moreover the fact that the maternity function* $m \in L^\infty(0,a_\dagger)$ *guarantees that* $u^n(0,x) \in L^2(\Omega)$ *as long as* $u^{n-1} \in L^2([0,a_\dagger]\times\Omega)$.

4. Space discretization

We discretize in space equation (8) by means of finite elements (see [11] for an introduction to finite element methods). Let then $\Omega = \bigcup_{j=1}^N K_j$, where each $K_j = T_{K_j}(E)$ is an element of the triangulation, E is the reference simplex and T_{K_j} is an invertible affine map. The associated finite element space is then

$$V_h = \left\{\; \varphi_h \in C^0(\Omega)\,|\,\varphi_{h|K_j}\circ T_{K_j} \in \mathbb{P}_k(E)\right\},$$

where $\mathbb{P}_k(E)$ is the space of polynomials of degree at most k on E. A semi-discrete problem in space is then obtained by applying a Galerkin procedure to (8) and choosing a finite element basis for V_h. Letting $\{\varphi_j\}_{j=1,..,N_h}$ be the nodal basis of the finite element space V_h, the semi-discrete solution $u_h^n(a,x)$ is given by

$$u_h^n(a,x) = \sum_{j=1}^{N_h} u_j^n(a)\,\varphi_j(x).$$

By denoting with $\mathbf{u}_h^n(a) = (u_1^n(a), \ldots, u_{N_h}^n(a))^T$, since the finite element basis functions depend only on space, we can rewrite problem (8) as

$$\begin{aligned} &M\frac{d\mathbf{u}_h^n}{da} + \mathcal{A}^{(n-1)}(a)\,\mathbf{u}_h^n + \frac{1}{\Delta t}M\mathbf{u}_h^n = \mathbf{f}^n + \frac{1}{\Delta t}M\mathbf{u}_h^{n-1}, \\ &\mathbf{u}_h^n(0) = \int_0^{a_\dagger} m(a)\,\mathbf{u}_h^{n-1}(a)\;da, \qquad\qquad \mathbf{p}_h^n = \int_0^{a_\dagger} \pi(a)\,\mathbf{u}_h^n(a)\;da, \end{aligned} \tag{9}$$

where M is the mass matrix ($M_{ij} = \int_\Omega \varphi_j\varphi_i\,dx$) and $\mathcal{A}^{(n-1)}$ is the stiffness matrix associated to the bilinear form $A(\mathbf{p}_h^{n-1}; a; \cdot, \cdot)$, ($\left[\mathcal{A}^{(n-1)}(a)\right]_{ij} = A(\mathbf{p}_h^{n-1}; a; \varphi_j, \varphi_i)$).

5. Age discretization

We advance in age the differential problem in (9) by means of the θ-method (see [11]). Let then $a^m = m\Delta a$ ($m = 0, 1, \ldots, N_a$) be a partition of the age interval $[0, a_\dagger]$ into N_a subintervals of uniform amplitude. For $j = 1, .., N_h$, we let $u_j^{n,m}$ denote the approximation of $u_j^n(a^m)$, and the approximation to $u(t^n, a^m, x)$ is then given by

$$u_h^{n,m}(x) = \sum_{j=1}^{N_h} u_j^{n,m}\,\varphi_j(x).$$

We denote by $\mathbf{u}_h^{n,m} = (u_1^{n,m}, \ldots, u_{N_h}^{n,m})^T$ the unknown vector at time t^n and age a^m, and we advance from age level a^m to a^{m+1} by the θ-method, which reads, for $0 \le \theta \le 1$,

$$\begin{aligned} &M\frac{\mathbf{u}_h^{n,m} - \mathbf{u}_h^{n,m-1}}{\Delta a} + \theta\left(\mathcal{A}_m^{(n-1)}\,\mathbf{u}_h^{n,m} + \frac{1}{\Delta t}M\mathbf{u}_h^{n,m}\right) + \\ &\quad (1-\theta)\left(\mathcal{A}_{m-1}^{(n-1)}\,\mathbf{u}_h^{n,m-1} + \frac{1}{\Delta t}M\mathbf{u}_h^{n,m-1}\right) = \\ &\qquad \theta\left(\mathbf{f}^{n,m} + \frac{1}{\Delta t}M\mathbf{u}_h^{n-1,m}\right) + (1-\theta)\left(\mathbf{f}^{n,m-1} + \frac{1}{\Delta t}M\mathbf{u}_h^{n-1,m-1}\right), \end{aligned} \tag{10}$$

where $\mathcal{A}_m^{(n-1)} = \mathcal{A}^{(n-1)}(a^m)$. If $\theta = 0$ we have the Forward Euler method (fully explicit in age), if $\theta = 1$ we have the Backward Euler method (fully implicit in age), while $\theta = 1/2$ corresponds to the Crank-Nicholson method [11].

Finally, the values of $\mathbf{u}_h^{n,0}$ and $\mathbf{p}_h^n$ will be computed by replacing the integrals in (9) with suitable quadrature rules. In the numerical result section, we us in both cases a second order Simpson quadrature rule over two adjacent intervals.

6. Stability and convergence

Denoting by $\mathbf{U}_h^n = (\mathbf{u}_h^{n,0}, \mathbf{u}_h^{n,1}, \ldots, \mathbf{u}_h^{n,N_a})$ the approximate solution at time $t = t^n$, we define the discrete $L^1(0, a_\dagger; L^2(\Omega))$ norm as

$$\|\mathbf{U}_h^n\|_{\mathcal{L}^1(0,a_\dagger;L^2(\Omega))} = \sum_{m=0}^{N_a} \Delta a\,\|\mathbf{u}_h^{n,m}\|_0\,,$$

where $\|\cdot\|_0$ is the standard $L^2(\Omega)$ norm. Under some mild assumption on the exact solution, the following stability and convergence results for the proposed scheme (with $\theta = 1$) hold.

Proposition 6.1 (Stability) *For any $n = 1, \ldots, N_t$, the following estimate holds:*

$$\|\mathbf{U}_h^n\|_{\mathcal{L}^1(0,a_\dagger;L^2(\Omega))} \leq \left(1 + e^{a_\dagger \beta_+^2 T}\right) \|\mathbf{U}_h^0\|_{\mathcal{L}^1(0,a_\dagger;L^2(\Omega))},$$

where β_+ is the one in (4). □

Proposition 6.2 (Convergence) *Let $\mathcal{T}_h$ be a regular family of triangulations on Ω. Assume that the solution $u(t,a,x)$ of the continuous problem is such that, for all $t \in (0,T)$, $\dfrac{\partial u}{\partial a}(t,\cdot,\cdot)$, $\dfrac{\partial u}{\partial t}(t,\cdot,\cdot) \in L^1(0,a_\dagger;H^1(\Omega))$, and $\dfrac{\partial^2 u}{\partial a^2}(t,\cdot,\cdot)$, $\dfrac{\partial^2 u}{\partial t^2}(t,\cdot,\cdot) \in L^1(0,a_\dagger;L^2(\Omega))$. Then, using linear finite elements, the following estimate holds*

$$\begin{aligned}
\|u(t^n,\cdot,\cdot) - \mathbf{U}_h^n\|_{\mathcal{L}^1(0,a_\dagger;L^2(\Omega))} \leq\ & \left\|\mathbf{U}_h^0 - \Pi_h u_0\right\|_{\mathcal{L}^1(0,a_\dagger;L^2(\Omega))} \\
& + Ch\, \|u(t^n,\cdot,\cdot)\|_{\mathcal{L}^1(0,a_\dagger;H^1(\Omega))} + Ch \int_0^{t^n} \left\|\frac{\partial u}{\partial t}(t,\cdot,\cdot)\right\|_{\mathcal{L}^1(0,a_\dagger;H^1(\Omega))} dt \\
& + Ch \sum_{p=0}^{n} \Delta t \left\|\frac{\partial u}{\partial a}(t^p,\cdot,\cdot)\right\|_{L^1(0,a_\dagger;H^1(\Omega))} + C\,\Delta t \int_0^{t^n} \left\|\frac{\partial^2 u}{\partial t^2}(t,\cdot,\cdot)\right\|_{\mathcal{L}^1(0,a_\dagger;L^2(\Omega))} dt \\
& + C\,\Delta a \sum_{p=0}^{n} \Delta t \left\|\frac{\partial^2 u}{\partial a^2}(t^p,\cdot,\cdot)\right\|_{L^1(0,a_\dagger;L^2(\Omega))},
\end{aligned} \tag{11}$$

where the constant $C > 0$ is independent of h, Δa, and Δt. □

Proofs of the above propositions follow from a generalization of the results in [2], and will be detailed in a forthcoming paper [4].

7. Algorithm

Given $\mathbf{u}_h^{0,m}$ $(m = 1, \ldots, N_a)$, and $\mathbf{p}_h^0$, for $n = 1, \ldots, N_t$:

1. Compute the initial value $\mathbf{u}_h^{n,0}$ from the previous time step via a Simpson quadrature rule over two adjacent age intervals

$$\mathbf{u}_h^{n,0} = \sum_{l=1}^{N_a/2} \frac{\Delta a}{6} \left[m(a^{2(l-1)})\, \mathbf{u}_h^{n-1,2(l-1)} + 4\, m(a^{2l-1})\, \mathbf{u}_h^{n-1,2l-1} + m(a^{2l})\, \mathbf{u}_h^{n-1,2l}\right].$$

2. For $m = 1, \ldots, N_a$

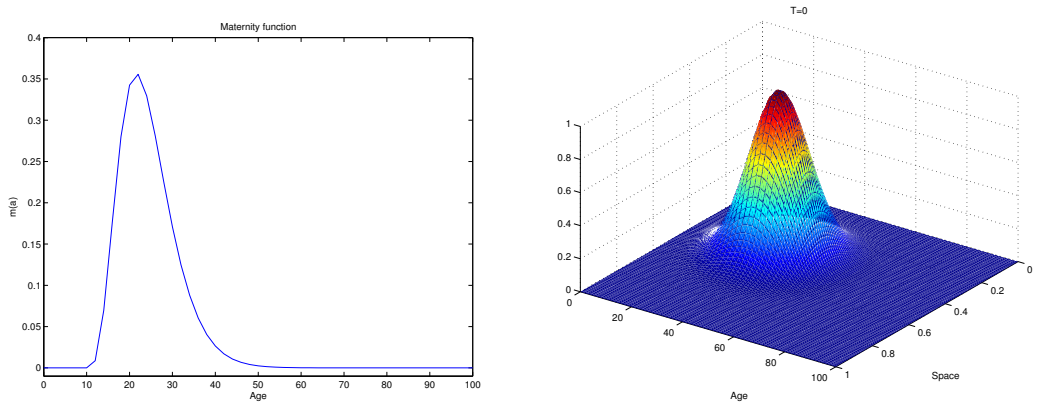


Figure 1: Maternity function (left) and age-space initial profile (right).

(a) Assemble the stiffness matrix $\mathcal{A}_m^{(n)}$ from the population at previous time step

$$\left[\mathcal{A}_m^{(n)}\right]_{ij} = A(\mathbf{p}_h^{n-1} : a^m; \varphi_j, \varphi_i)\,,$$

(b) solve

$$\begin{aligned}&\left[(\Delta t + \theta\Delta a)M + \theta\Delta t\Delta a\,\mathcal{A}_m^{(n)}\right]\mathbf{u}_h^{n,m}\\ &= \theta\Delta a\,M\,\mathbf{u}_h^{n-1,m} + \left[(\Delta t - (1-\theta)\Delta a)\,M - (1-\theta)\Delta t\Delta a\,\mathcal{A}_{m-1}^{(n)}\right]\mathbf{u}_h^{n,m-1}\\ &+ (1-\theta)\Delta a\,M\,\mathbf{u}_h^{n-1,m-1} + \Delta t\Delta a\Big[\theta\,\mathbf{f}^{n,m} + (1-\theta)\,\mathbf{f}^{n,m-1}\Big]\,.\end{aligned}$$

3. Update the total population $\mathbf{p}_h^n$ via a Simpson quadrature rule over two adjacent age intervals

$$\mathbf{p}_h^n = \sum_{l=1}^{N_a/2} \frac{\Delta a}{6}\left[\pi(a^{2(l-1)})\,\mathbf{u}_h^{n,2(l-1)} + 4\,\pi(a^{2l-1})\,\mathbf{u}_h^{n,2l-1} + m(a^{2l})\,\mathbf{u}_h^{n,2l}\right].$$

8. Numerical results

We present in this section some numerical results to show the effectivity of the method. The spatial domain is $\Omega = (0,1)$, the age interval is $[0,100]$, and we choose as maximal time $T = 10$. The computational domain is discretized by a uniform mesh in space, time and age, and we choose $\theta = 1$. The numerical simulations are run on a self developed code in Matlab® 7.8.

We consider a non-symmetric initial distribution of population (with respect to both space and age) given by

$$u_0(x,a) = \mathrm{e}^{-\left(\frac{(a-30)^2}{200} + 100(x-0.4)^2\right)},$$

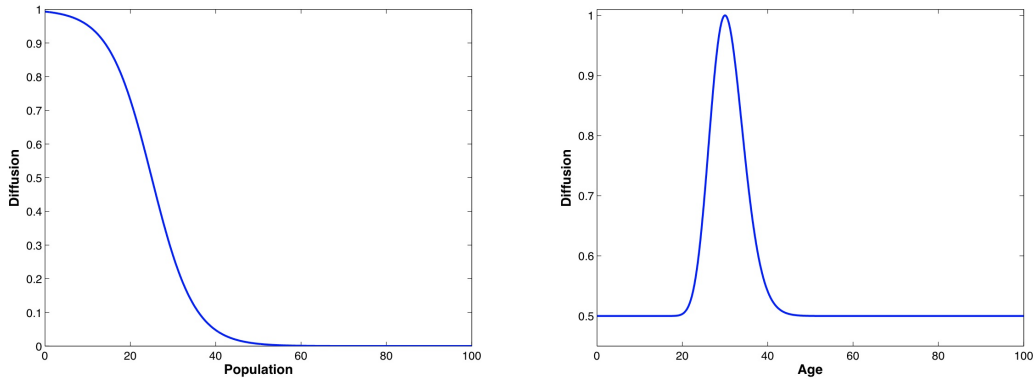


Figure 2: Diffusion coefficients: $k_p(p)$ (left) and $k_a(a)$ (right).

and we choose the mortality and fertility function as

$$\mu(a) = \frac{1}{a_\dagger - a}, \qquad \beta(a) = \begin{cases} 0 & \text{if } a \le a_1 \\ \dfrac{\beta(a-a_1)^{\alpha-1}\mathrm{e}^{-\frac{(a-a_1)}{\vartheta}}}{\vartheta^\alpha \Gamma(\alpha)} & \text{if } a_1 < a < a_2 \\ 0 & \text{if } a \ge a_2, \end{cases}$$

where we set $a_1 = 17$, $a_2 = 70$, $\beta = 7$, $\alpha = 5$, and $\vartheta = 3$. We plot in Figure 1 the resulting maternity function and the initial profile of the problem.

We consider a diffusion coefficient $k(p(t,x),a) = k_p(p) \times k_a(a)$, where we assume $k_p(p)$ to be a monotonic function of the total population $p(t,x)$. The rationale behind this choice is that the population is more keen to move in areas where a lower level of individuals is present, but a different behavior can be easily implemented. We choose in the tests

$$k_p(p) = 1 - \frac{1}{1+\exp\left(-\left(\frac{p}{5}-5\right)\right)} \qquad k_a(a) = 0.5 + 0.5 \times \exp\left(-\frac{(a-30)^2}{a}\right),$$

that we plot in Figure 2. With this choice of $k_a(a)$, youngster and old individuals are less mobile.

We investigate numerically the spatial convergence of the method. We consider diffusion coefficients depending on both population and age ($k = k_p \times k_a$), and population only ($k = k_p$): we plot in Figure 3 the corresponding diffusion coefficients in space and age at the initial time $t = 0$. We analyze the relative error $\frac{\|u(t^n,\cdot,\cdot) - U_h^n\|}{\|u(t^n,\cdot,\cdot)\|}$ in the discrete $\mathcal{L}^1(0, a_\dagger; L^2(\Omega))$ norm, with respect to a reference solution computed using a very fine grid in both age and time with $\Delta a = 2\Delta t = 0.1$ and $h = 1/1000$. In Figure 4 we show the work precision in h, for a uniform grid in age and time with $\Delta a = 2\Delta t = 0.2$ for both the case of a density dependent diffusion (left) and density and age dependent diffusion (right). Convergence appears to be robust with respect to the diffusion coefficients. In Figure 5 we plot, for $k = k_p \times k_a$, the age profile at $x = 0.4$ for different times, and the age-space contours at time $T = 5$.

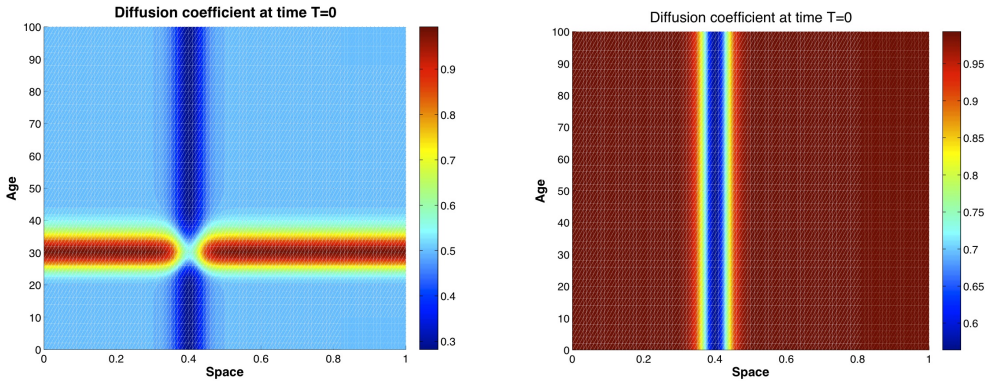


Figure 3: Diffusion coefficients at time $t = 0$. Left: $k = k_p \times k_a$. Right: $k = k_p$.

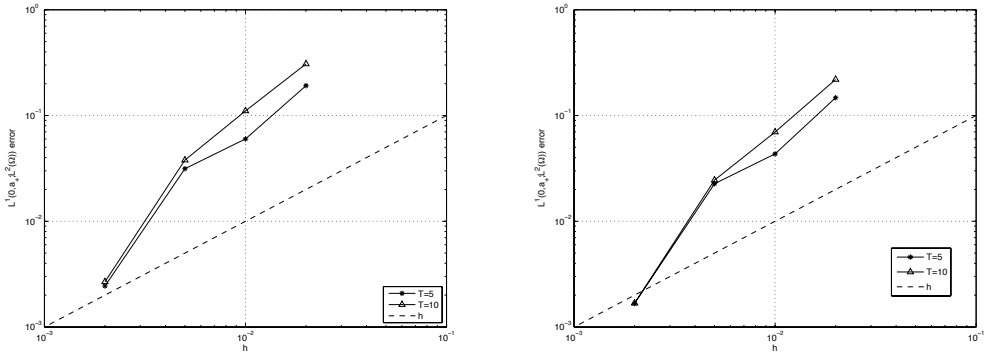


Figure 4: Convergence in $\mathcal{L}^1(0, a_\dagger; L^2(\Omega))$ norm: $k = k_p \times k_a$ (left) and $k = k_p$ (right).

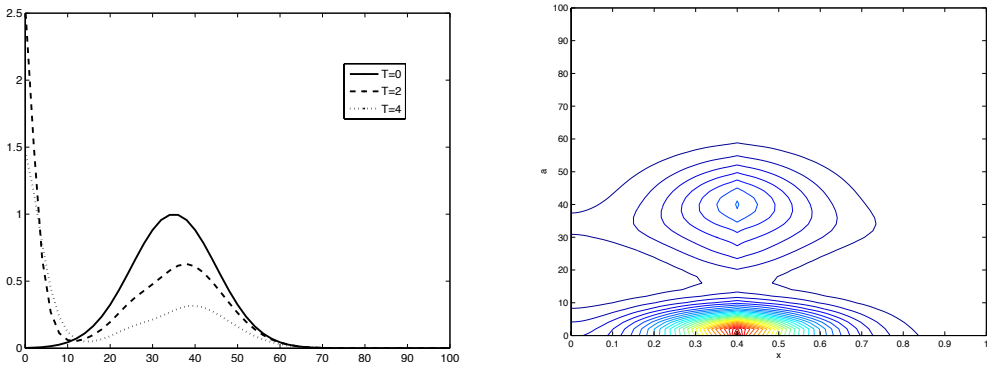


Figure 5: Age profile at $x = 0.4$ (left) and age-space profile at time $T = 5$ (right): $k = k_p \times k_a$.

9. Conclusions

We proposed a Galerkin type method for the numerical approximation of a density dependent diffusion dynamics of an age-structured population. The method is based on a finite elements discretization in space, on a semi-implicit discretization in time, and on the θ-method in age. The separate discretization of time and age, naturally allows for separate adaptivity, which can be necessary when dealing with practical ecological problems. Numerical results showed the effectiveness of the method, that will be analyzed in a more comprehensive way in a forthcoming paper [4].

References

[1] Ayati, B. P. and Dupont, T.: Galerkin methods in age and space for a population model with nonlinear diffusion. SIAM J. Numer. Anal., **40**(3) (2002), 1064–1076.

[2] Cusulin, C. and Gerardo-Giorda, L.: A numerical method for diffusion in age-structured populations. Numer. Methods Partial Differential Equations **26**(2) (2010), 253–273.

[3] de Roos, A. M.: Numerical methods for structured population models: the escalator boxcar train. Numer. Methods Partial Differential Equations **4** (1989), 173–195.

[4] Gerardo-Giorda, L.: Galerkin methods for nonlinear diffusion problems in age-structured population dynamics. In preparation.

[5] Iannelli, M.: Mathematical theory of age-structured population dynamics. Giardini editori e stampatori, Pisa, 1995.

[6] Kim, M.-Y.: Galerkin methods for a model of population dynamics with nonlinear diffusion. Numer. Methods Partial Differential Equations **12** (1996), 59–73.

[7] Kim, M.-Y. and Park, E-J.: Characteristic finite element methods for diffusion epidemic models with age-structured populations. Comput. Math. Appl. **97**, (1998), 55–70.

[8] Langlais, M.: A nonlinear problem in age-dependent population diffusion. SIAM J. Math. Anal. **16** (1985), 510–529.

[9] Milner, F. A.: A numerical method for a model of population dynamics with spatial diffusion. Comput. Math. Appl. **19**(31) (1990).

[10] Okubo, A. and Levin, S. A.: *Diffusion and ecological problems: modern perspectives*. Springer, New York, 2001.

[11] Quarteroni, A. and Valli, A.: *Numerical approximation of partial differential equations*. Springer-Verlag, Berlin, 1994.

Conference Applications of Mathematics 2013
in honor of the 70th birthday of Karel Segeth.
Jan Brandts, Sergey Korotov, Michal Křížek,
Jakub Šístek, and Tomáš Vejchodský (Eds.),
Institute of Mathematics AS CR, Prague 2013

ON THE COMPUTATION OF MOMENTS OF THE PARTIAL NON-CENTRAL CHI-SQUARE DISTRIBUTION FUNCTION

A. Gil[1], J. Segura[2], N. M. Temme[3]

[1]Departamento de Matemática Aplicada y Ciencias de la Computación
ETSI Caminos, Canales y Puertos, Universidad de Cantabria, 39005-Santander, Spain
amparo.gil@unican.es
[2]Departamento de Matemáticas, Estadística y Computación
Universidad de Cantabria, 39005 Santander, Spain
javier.segura@unican.es
[3]IAA, 1391 VD 18, Abcoude, The Netherlands*
nico.temme@cwi.nl

Abstract

Properties satisfied by the moments of the partial non-central chi-square distribution function, also known as Nuttall Q-functions, and methods for computing these moments are discussed in this paper. The Nuttall Q-function is involved in the study of a variety of problems in different fields, as for example digital communications.

1. Introduction

The non-central chi-square distribution function of probability appears in many applications. For example, in radar communications it appears when computing the detection of signals in noise using a square-law detector. Its cumulative distribution function is also known as the generalized Marcum $Q-$ function, which is defined by using the integral representation

$$Q_\mu(x,y) = x^{\frac{1}{2}(1-\mu)} \int_y^{+\infty} t^{\frac{1}{2}(\mu-1)} \mathrm{e}^{-t-x} I_{\mu-1}\left(2\sqrt{xt}\right)\, dt, \tag{1}$$

where $\mu > 0$ and $I_\mu(z)$ is the modified Bessel function.

In radar problems, if the signal-to-noise power ratio is x for the sum of μ independent samples of the output of a square-law detector, this integral gives the probability of that the sum will be y or more.

The complementary function of the generalized Marcum $Q-$ function is given by

$$P_\mu(x,y) = x^{\frac{1}{2}(1-\mu)} \int_0^{y} t^{\frac{1}{2}(\mu-1)} \mathrm{e}^{-t-x} I_{\mu-1}\left(2\sqrt{xt}\right)\, dt, \tag{2}$$

*Former address: CWI, 1098 XG Amsterdam, The Netherlands

and the following relation holds

$$P_\mu(x,y) + Q_\mu(x,y) = 1. \tag{3}$$

Methods and an algorithm for computing the functions $P_\mu(x,y)$ and $Q_\mu(x,y)$ are described in [2].

The ηth moment of the partial non-central chi-square distribution function is given by

$$Q_{\eta,\mu}(x,y) = x^{\frac{1}{2}(1-\mu)} \int_y^{+\infty} t^{\eta+\frac{1}{2}(\mu-1)} \mathrm{e}^{-t-x} I_{\mu-1}\left(2\sqrt{xt}\right)\, dt. \tag{4}$$

In this manuscript, we give properties satisfied by the moments of the partial non-central chi-square distribution functions and discuss methods for computing these moments, also known as Nuttall Q-functions [4]. There are several applications where these functions are involved as for example, the analysis of the outage probability of wireless communication systems with a minimum signal power constraint [5], to mention just one example within the telecommunications field.

2. Properties

The Maclaurin series for the modified Bessel function reads

$$I_\mu(z) = \left(\tfrac{1}{2}z\right)^\mu \sum_{n=0}^{\infty} \frac{\left(\frac{1}{4}z^2\right)^n}{n!\,\Gamma(\mu+n+1)}. \tag{5}$$

By substituting this expression in the integral representation, we obtain the series expansion for the ηth moment of the non-central chi-square distribution function:

$$Q_{\eta,\,\mu}(x,y) = \mathrm{e}^{-x} \sum_{n=0}^{\infty} \frac{x^n}{n!} \frac{\Gamma(\eta+\mu+n,\, y)}{\Gamma(\mu+n)}. \tag{6}$$

This expansion is given in terms of one of the standard incomplete gamma functions defined by

$$\Gamma(\mu,x) = \int_x^{+\infty} t^{\mu-1}\mathrm{e}^{-t}\, dt. \tag{7}$$

Introducing the factor $\Gamma(\eta+\mu+n)$ in (6), the expansion can be also given in terms of the incomplete gamma function ratio $Q_\mu(y)$, defined by

$$Q_\mu(x) = \frac{\Gamma(\mu,x)}{\Gamma(\mu)}, \tag{8}$$

and for which algorithms are given in [3].

The expansion for the ηth moment of the non-central chi-square distribution function in terms of incomplete gamma function ratios is given by

$$Q_{\eta,\,\mu}(x,y) = \mathrm{e}^{-x} \sum_{n=0}^{\infty} \frac{x^n}{n!} \frac{\Gamma(\eta+\mu+n)}{\Gamma(\mu+n)} Q_{\eta+\mu+n}(y). \tag{9}$$

The series representation can be computed by using the algorithms for the incomplete gamma ratios described in [3]. The recurrence relation

$$Q_{\eta+\mu+1}(y) = Q_{\eta+\mu}(y) + \frac{y^{\eta+\mu}\mathrm{e}^{-y}}{\Gamma(\eta+\mu+1)}, \tag{10}$$

is stable for $Q_\mu(y)$ in the forward direction, so the evaluation of the terms in the series for this function in (9) is rather easy.

A recurrence relation for the moments of the non-central chi-squared distribution function can be obtained considering integration by parts in the integral in (4), together with the relation $z^\mu I_{\mu-1}(z) = \frac{d}{dz}\left(z^\mu I_\mu(z)\right)$. This gives

$$Q_{\eta,\mu}(x,y) = Q_{\eta,\mu+1}(x,y) - \eta Q_{\eta-1,\mu+1} - \left(\frac{y}{x}\right)^{\mu/2} y^\eta \mathrm{e}^{-x-y} I_\mu(2\sqrt{xy}). \tag{11}$$

When $\eta = 0$, this recurrence reduces to a first order difference equation for the Marcum-Q function (see, for instance, [6] [1]). The recurrence relation given in (11) can be used for testing, and it can be also used for computation, as we describe later.

3. Computing moments using the series expansion

The series expansion given in (6) has been tested by using the recurrence relation of (11) written in the form

$$\frac{Q_{\eta,\mu+1}(x,y)}{Q_{\eta,\mu}(x,y) + \eta Q_{\eta-1,\mu+1} + \left(\frac{y}{x}\right)^{\mu/2} y^\eta \mathrm{e}^{-x-y} I_\mu(2\sqrt{xy})} = 1. \tag{12}$$

The deviations from 1 of the left-hand side of (12) (in absolute value) will measure the accuracy of the tested methods. The series expansion has been implemented in the Fortran 90 module **NuttallF**. This module uses another module (**IncgamFI**) for the computation of the gamma function ratios. We have tested the parameter region $(\eta,\, \mu,\, x,\, y) \in (1,\, 50) \times (1,\, 50) \times (0,\, 20) \times (0,\, 20)$. The tests show that an accuracy better than 10^{-12} in this region can be obtained with the series expansion.

When μ or $\mu + n$ are large, it is convenient to use approximations for the ratio of gamma functions appearing in the expression, in order to avoid the appearance of overflow problems sooner than expected. In the case $\mu + n \to \infty$ we have:

$$\frac{\Gamma(\eta+\mu+n)}{\Gamma(\mu+n)} \sim (\mu+n)^\eta. \tag{13}$$

The following table shows some values of moments of the chi-square distribution function computed with the series expansion and the corresponding values obtained

[1] We note that a factor e^{-y} is missing in [6, Eq. (1.4)].

η	μ	x	y	$Q_{\eta,\mu}(x,y)$	$Q_{\eta,\mu}(x,y)$ with Maple
1	1	0.1	1.5	0.6644091427683566	0.66440914276835656
5	10	0.1	1.5	252472.22699183668	252472.226991836658
50	30	0.1	1.5	$1.1944632251434243\,10^{+86}$	$1.19446322514344860\,10^{+86}$
1	1	1.2	5	0.5457546041478581	0.54575460414785805
5	10	1.2	5	419098.1927146542	419098.192714654143
50	30	1.2	5	$6.8093141960731 25\,10^{+86}$	$6.80931419607285639\,10^{+86}$
1	1	5	10	1.4822515303982464	1.48225153039824667
5	10	5	10	1654969.264263704	1654969.26426370245
50	30	5	10	$1.1734657613338925\,10^{+89}$	$1.17346576133388184\,10^{+89}$

Table 1: Values of the moments of the chi-square distribution function for different choices of the parameters η, μ x and y. The values shown are obtained with the series expansion and with the direct computation of the integral representation using Maple with 50 digits.

with the direct computation of the integral representation using Maple with 50 digits (the results shown in the table correspond to the first 18 digits obtained with these computations). The computation of the series expansion has been implemented in the double precision Fortran 90 module **NuttallF**. As can be seen, an agreement of minimum 14-15 digits is obtained in all cases, which is consistent with the expected accuracy of the double precision Fortran 90 module.

In some cases, Maple fails to compute the integral and acceleration can be obtained by suitably truncating the improper integral and changing the variable of integration. We notice that, as before commented, the modified Bessel function is exponentially increasing for large arguments and then the integrand in (4) can be estimated by $t^{\gamma}\mathrm{e}^{-(\sqrt{t}-\sqrt{x})^2}$, $\gamma=\eta+(\mu-1)/2$ which is related to a Gaussian centered $t=x$. The maximum value of this function is attained at $t=(\sqrt{x}+\sqrt{x+4\gamma})^2/4$ and integrating around this value with a sufficiently wide interval is enough. This truncated integral over finite interval $[a,b]$ can be then transformed with a linear change to an integral in $[-1,1]$ and the convergence is further accelerated by considering the change of variable $t=\tanh(u)$, particularly if the trapezoidal rule is used for evaluating the integral (see [1, §5.4.2]). These modifications are observed to speed up the computation of the integrals using Maple, particularly for the last value in Table 1 for which Maple does not appear to be able to converge to an accurate value.

4. Computing moments by recursion

If we write the recurrence relation (11) as

$$Q_{\eta,\mu+1}(x,y)=Q_{\eta,\mu}(x,y)+\eta Q_{\eta-1,\mu+1}+\left(\frac{y}{x}\right)^{\mu/2}y^{\eta}\mathrm{e}^{-x-y}I_{\mu}(2\sqrt{xy}), \qquad (14)$$

then it is clear that we have a numerically stable relation because all the terms in the right hand side are positive.

Now, assume that the moments of order zero (Marcum functions) $Q_{0,\mu}$ are known for $\mu = 1, 2, \ldots, N$ (or for a sequence of real values μ_i, $i=1, \ldots N$, with $\mu_{i+1} - \mu_i = 1$). If $Q(1, \mu)$ is also known, the relation (15) can be used to compute $Q(1, \mu + 1)$; therefore, starting from the value $Q(1, 1)$ we can compute $Q(1, \mu)$, $\mu = 1, 2, \ldots, N$ in a stable way. In the same way, after determining $Q(1, \mu)$, $\mu = 1, 2, \ldots, N$ and if $Q(2, 1)$ is know, we can compute $Q(1, \mu)$, $\mu = 1, 2, \ldots N$ and so on.

It is worth mentioning that the inhomogeneous recurrence has to be applied with care, particularly the inhomogeneous term. As x and/or y becomes large the Bessel function increases exponentially; therefore we have the product of a small exponential times an exponentially large function and because of the bad conditioning of the exponentials, this translates into larger relative errors; additionally, the exponentials may overflow/underflow. Part of this error can be avoided by considering the scaled Bessel function $\tilde{I}_\nu(x) = e^{-x} I_\nu(x)$. In terms of this function

$$Q_{\eta,\mu+1}(x, y) = Q_{\eta,\mu}(x, y) + \eta Q_{\eta-1,\mu+1} + \left(\frac{y}{x}\right)^{\mu/2} y^\eta \mathrm{e}^{-(\sqrt{x}-\sqrt{y})^2} \tilde{I}_\mu(2\sqrt{xy}). \tag{15}$$

An alternative way of computing with recurrences is considering a homogeneous equation, which we can be constructed from the inhomogeneous equation writing

$$\begin{aligned} Q_{\eta,\mu+2} - Q_{\eta,\mu+1} - \eta Q_{\eta-1,\mu+2} = \;& c_{\mu+1}(Q_{\eta,\mu+1} - Q_{\eta,\mu} - \eta Q_{\eta-1,\mu+1}), \\ & c_{\mu+1} = \sqrt{\frac{y}{x}} \frac{I_{\mu+1}(2\sqrt{xy})}{I_\mu(2\sqrt{xy})}. \end{aligned} \tag{16}$$

Then, if $Q(\eta - 1, \mu)$ is know $\mu = 1, 2, \ldots N$, we can compute $Q(\eta, \mu)$ $\mu = 1, 2, \ldots N$, starting from $Q(\eta, 1)$ and $Q(\eta, 2)$ with the recurrence

$$Q_{\eta,\mu+2} = (1 + c_{\mu+1}) Q_{\eta,\mu+1} - c_{\mu+1} Q_{\eta,\mu} + \eta Q_{\eta-1,\mu+2} - \eta c_{\mu+1} Q_{\eta-1,\mu+1}. \tag{17}$$

The advantage of this recurrence is that the overflow problems are reduced because ratios of Bessel functions appear instead of Bessel functions themselves. Also, for computing these ratios, continued fraction representations can be used. In Table 2 the use of the recurrence relation for computing $Q_{2,N}$ is tested for several values of N. The values of x and y are fixed to 2 and 3, respectively. The table shows the relative error obtained when comparing the value obtained with the recurrence relation and the direct computation using the series expansion of (9):

$$E_r = \left| 1 - \frac{Q^S_{2,N}(2, 3)}{Q^R_{2,N}(2, 3)} \right|. \tag{18}$$

The continued fraction for the ratio of Bessel functions is computed using the modified Lentz algorithm [7] and [1, §6.6.2].

N	Relative error (18)
10	$2.96\,10^{-16}$
20	$4.84\,10^{-16}$
30	$1.67\,10^{-15}$
40	$1.02\,10^{-15}$
50	$9.02\,10^{-15}$
60	$3.09\,10^{-15}$

Table 2: Test of the application of the recurrence relation given in (17). The relative errors are obtained when comparing the value obtained with the recurrence relation (17) and the direct computation using the series expansion of (9).

Acknowledgements

This work was supported by *Ministerio de Ciencia e Innovación*, project MTM2009-11686 and *Ministerio de Economía y Competitividad*, project MTM2012-34787.

References

[1] Gil, A., Segura, J., and Temme, N. M.: *Numerical methods for special functions*. SIAM, Philadelphia, PA, 2007.

[2] Gil, A., Segura, J., and Temme, N. M.: Computation of the Marcum Q-function (2012). Submitted.

[3] Gil, A., Temme, N. M., and Segura, J.: Efficient and accurate algorithms for the computation and inversion of the incomplete gamma function ratios. SIAM J. Sci. Comput. **34** (2012), A2965–A2981.

[4] Nuttall, A. H.: Some integrals involving the Q function. Naval Underwater Systems Center, New London Lab., New London, CT **4297** (1972).

[5] Simon, M. K.: The Nuttall $Q-$function-its relation to the Marcum $Q-$function and its application in digital communication performance evaluation. IEEE Trans. Commun. **50** (2002), 1712–1715.

[6] Temme, N. M.: Asymptotic and numerical aspects of the noncentral chi-square distribution. Comput. Math. Appl. **25** (1993), 55–63. doi:10.1016/0898-1221(93)90198-5. URL `http://dx.doi.org/10.1016/0898-1221(93)90198-5`.

[7] Thompson, I. and Barnett, A.: Coulomb and Bessel functions of complex arguments and order. J. Comput. Phys. **64** (1986), 490–509.

Conference Applications of Mathematics 2013
in honor of the 70th birthday of Karel Segeth.
Jan Brandts, Sergey Korotov, Michal Křížek,
Jakub Šístek, and Tomáš Vejchodský (Eds.),
Institute of Mathematics AS CR, Prague 2013

PATH-FOLLOWING THE STATIC CONTACT PROBLEM WITH COULOMB FRICTION

J. Haslinger[1,2], V. Janovský[1], R. Kučera[2]

[1] Department of Numerical Mathematics, Charles University, Prague
Sokolovská 83, 186 75 Prague 8, Czech Republic
hasling@karlin.mff.cuni.cz, janovsky@karlin.mff.cuni.cz
[2] Department of Mathematics and Descriptive Geometry, VŠB-TU Ostrava
17. listopadu 15, 708 33 Ostrava-Poruba, Czech Republic
radek.kucera@vsb.cz

Abstract

Consider contact problem with Coulomb friction on two planar domains. In order to find non-unique solutions we propose a new path following algorithm: Given a linear loading path we approximate the corresponding solution path. It consists of oriented piecewise linear branches connected by transition points. We developed a) predictor-corrector algorithm to follow oriented linear branches, b) branching and orientation indicators to detect transition points. The techniques incorporate semismooth Newton iterations and inactive/active set strategy on the contact zone.

1. Introduction

Consider deformable bodies in mutual contact. The relevant mathematical description consists in modeling both the non-penetration conditions and a friction law. The widely accepted Coulomb friction law represents a serious mathematical and numerical problem.

In particular, we consider 2D static contact problem with Coulomb friction. The problem is uniquely solvable, provided that the friction coefficient $\mathcal{F} > 0$ is sufficiently small, see [14, 6]. Since the seminal paper [14], no essential contribution was made concerning solvability of this problem for general data.

Obviously, engineers have always solved this important problem numerically, regardless unresolved theoretical issues. In a natural finite element (FEM) approximation, the discrete problem has always a solution, disregarding the size of $\mathcal{F}$, see [9, 8, 13]. Since the (discrete) problem is locally solvable, the idea was to apply the Implicit Function Theorem to follow the solution path, which was parameterized either by $\mathcal{F}$ or by a load increment. Nevertheless, lumped element models [11, 9, 13] indicate, that the particular solution points of interest should be those in which the Implicit Function Theorem *fails* to hold. They are *turning points* of the solution path. Actually, they are responsible for non-unique solvability of the problem.

The solution path is continuous, piecewise smooth, [8]. The classical numerical path following techniques, see e.g. [1], have to fail in principle. In [8], a special continuation algorithm was proposed to trace piecewise smooth solution curves. The algorithm was tested on lumped element models with just one or two points on the contact boundary, [12, 8].

In this paper, we present an improved continuation strategy and test it on a real FEM model. The outline is as follows: In Section 2, we define the state problem and its discretization. We recall the semi-smooth Newton method and apply it to the discrete state problem, see Section 3. The actual contribution is in Section 4, where a modified path following algorithm is presented. The substantial innovations consist in

1. application of *tangent continuation*, see [3], Algorithm 4.25,
2. introducing a robust *branching* and *orientation* indicator.

Note that due to material properties, the solution components are very uneven: The contact forces are within a range 10^6 $\mathrm{N\,kg^{-1}}$ while displacements are tiny.

2. State problem, FEM approximation

Let us consider two bodies Ω^1, Ω^2 in $\mathbb{R}^2$ with boundaries $\partial\Omega^k = \overline{\Gamma}_u^k \cup \overline{\Gamma}_p^k \cup \overline{\Gamma}_c^k$, $k = 1, 2$, see Figure 1. First, denote $\boldsymbol{u}^k$ the displacement field, $\boldsymbol{\sigma}(\boldsymbol{u}^k)$ the stress tensor, $\boldsymbol{f}^k$ the volume force, $\boldsymbol{p}^k$ the surface traction, $\boldsymbol{n}^k$ the outer normal vector to $\partial\Omega^k$, and $\lambda^k, \mu^k > 0$ material parameters. The state problem is defined by the Lamé equations in Ω^k, $k = 1, 2$,

$$\begin{aligned} -\mathrm{div}\,\boldsymbol{\sigma}(\boldsymbol{u}^k) &= \boldsymbol{f}^k, \\ \boldsymbol{\sigma}(\boldsymbol{u}^k) &= \lambda^k \mathrm{tr}(\boldsymbol{\epsilon}(\boldsymbol{u}^k))\boldsymbol{I} + 2\mu^k \boldsymbol{\epsilon}(\boldsymbol{u}^k), \\ \boldsymbol{\epsilon}(\boldsymbol{u}^k) &= \tfrac{1}{2}(\nabla \boldsymbol{u}^k + (\nabla \boldsymbol{u}^k)^\top), \end{aligned}$$

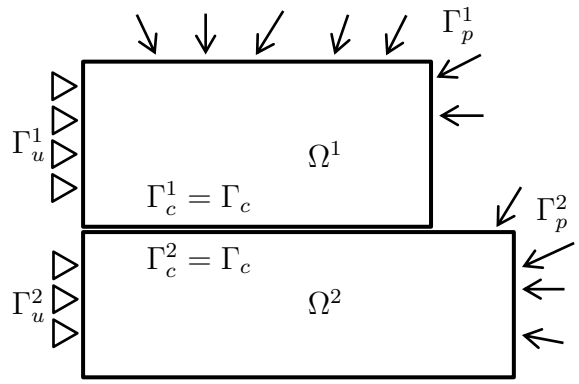


Figure 1: Geometry of the problem.

the Dirichlet and Neumann boundary conditions for $k = 1, 2$,

$$\begin{aligned} \boldsymbol{u}^k &= \boldsymbol{0} \quad \text{on } \Gamma_u^k, \\ \boldsymbol{\sigma}(\boldsymbol{u}^k)\boldsymbol{n}^k &= \boldsymbol{p}^k \quad \text{on } \Gamma_p^k, \end{aligned}$$

and by contact conditions on Γ_c:

- *unilateral contact law, Signorini problem:*

$$u_\nu \le 0,\ \sigma_\nu \le 0,\ \sigma_\nu u_\nu = 0 \quad \text{on } \Gamma_c,$$

  where $u_\nu = (\boldsymbol{u}^1 - \boldsymbol{u}^2)^\top \boldsymbol{n}$, $\sigma_\nu = \boldsymbol{n}^\top \boldsymbol{\sigma}(\boldsymbol{u}^1)\boldsymbol{n}$, and $\boldsymbol{n} = \boldsymbol{n}^1$,

- *transmission of contact stresses:*

$$\boldsymbol{\sigma}(\boldsymbol{u}^1)\boldsymbol{n} = \boldsymbol{\sigma}(\boldsymbol{u}^2)\boldsymbol{n} \quad \text{on } \Gamma_c,$$

- *the Coulomb friction law:*

$$\begin{aligned} &|\sigma_t| \le -\mathcal{F}\sigma_\nu, \\ &|\sigma_t| < -\mathcal{F}\sigma_\nu \quad \Rightarrow \quad u_t = 0, \\ &|\sigma_t| = -\mathcal{F}\sigma_\nu \quad \Rightarrow \quad \exists c_t \ge 0 :\ u_t = -c_t\sigma_t, \end{aligned}$$

  where $u_t = (\boldsymbol{u}^1 - \boldsymbol{u}^2)^\top \boldsymbol{t}$, $\sigma_t = \boldsymbol{t}^\top \boldsymbol{\sigma}(\boldsymbol{u}^1)\boldsymbol{n}$, $\boldsymbol{t}$ is orthogonal to $\boldsymbol{n}$, and $\mathcal{F} > 0$ is the coefficient of friction.

After FEM approximation we get the following *primal-dual* discrete state problem:

$$\boldsymbol{K}\boldsymbol{u} + \boldsymbol{N}^\top \boldsymbol{\lambda}_\nu + \boldsymbol{T}^\top \boldsymbol{\lambda}_t = \boldsymbol{f}, \tag{1}$$

$$\boldsymbol{N}\boldsymbol{u} \le 0,\ \boldsymbol{\lambda}_\nu \ge 0,\ \boldsymbol{\lambda}_\nu^\top \boldsymbol{N}\boldsymbol{u} = 0, \tag{2}$$

$$\left.\begin{aligned} &|\lambda_{t,i}| \le \mathcal{F}\lambda_{n,i}, \\ &|\lambda_{t,i}| < \mathcal{F}\lambda_{n,i} \Rightarrow (\boldsymbol{T}\boldsymbol{u})_i = \boldsymbol{0}, \\ &|\lambda_{t,i}| = \mathcal{F}\lambda_{n,i} \Rightarrow \exists\, c_{t,i} \ge 0 :\ (\boldsymbol{T}\boldsymbol{u})_i = c_{t,i}\lambda_{t,i}, \end{aligned}\right\} \quad i = 1, \ldots, m, \tag{3}$$

where $(\boldsymbol{u}, \boldsymbol{\lambda}_\nu, \boldsymbol{\lambda}_t) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^m$. Here $\boldsymbol{u}$ approximates the displacement field, $\boldsymbol{\lambda}_\nu$ and $\boldsymbol{\lambda}_t$ approximate normal and tangential stress components along the contact boundary Γ_c, m is the number of contact nodes. Data of the model: $\boldsymbol{K} \in \mathbb{R}^{n\times n}$ is positive definite stiffness matrix, $\boldsymbol{N}, \boldsymbol{T} \in \mathbb{R}^{m\times n}$ are full rank matrices (the actions of distributed contact forces along normal and tangential directions), $\boldsymbol{f} \in \mathbb{R}^n$ are nodal forces.

Next, we formulate inequalities (2)–(3) as a set of nonlinear equations using suitable projectors, see e.g. [7]. Let $P_{\mathbb{R}_+} : \mathbb{R} \mapsto \mathbb{R}_+$, $P_{\mathbb{R}_+}(x) = \max\{0, x\}$, $x \in \mathbb{R}$, be the projection onto $\mathbb{R}_+$. Let us define $P_{\mathbb{R}_+^m} : \mathbb{R}^m \mapsto \mathbb{R}_+^m$ for $\boldsymbol{x} = (x_1, \ldots, x_m)^\top$ by

$$P_{\mathbb{R}_+^m}(\boldsymbol{x}) = (P_{\mathbb{R}_+}(x_1), \ldots, P_{\mathbb{R}_+}(x_m))^\top.$$

Let $P_{[-g,g]} : \mathbb{R} \mapsto [-g, g]$, $P_{[-g,g]}(x) = \max\{0, x+g\} - \max\{0, x-g\} - g$, $x \in \mathbb{R}$, be the projection onto the interval $[-g, g]$, $g \geq 0$. Let us define $P_{[-\boldsymbol{g},\boldsymbol{g}]} : \mathbb{R}^m \mapsto [-\boldsymbol{g}, \boldsymbol{g}]$, where $[-\boldsymbol{g}, \boldsymbol{g}] = [-g_1, g_1] \times \cdots \times [-g_m, g_m]$, $\boldsymbol{g} = (g_1, \ldots, g_m)^\top$, $g_i \geq 0$, for $\boldsymbol{x} = (x_1, \ldots, x_m)^\top$ by

$$P_{[-\boldsymbol{g},\boldsymbol{g}]}(\boldsymbol{x}) = (P_{[-g_1,g_1]}(x_1), \ldots, P_{[-g_m,g_m]}(x_m))^\top.$$

The inequalities (2) and (3) can be equivalently written as

$$\boldsymbol{\lambda}_\nu - P_{\mathbb{R}^m_+}(\boldsymbol{\lambda}_\nu + \rho \boldsymbol{N}\boldsymbol{u}) = \boldsymbol{0} \quad \text{and} \quad \boldsymbol{\lambda}_t - P_{[-\mathcal{F}\boldsymbol{\lambda}_\nu, \mathcal{F}\boldsymbol{\lambda}_\nu]}(\boldsymbol{\lambda}_t + \rho \boldsymbol{T}\boldsymbol{u}) = \boldsymbol{0},$$

respectively, where $\rho > 0$ is arbitrary but fixed (e.g., $\rho = 1$). Therefore, solving (1)–(3) is equivalent to finding roots $\boldsymbol{y} = (\boldsymbol{u}, \boldsymbol{\lambda}_\nu, \boldsymbol{\lambda}_t) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^m$ of the equation

$$G(\boldsymbol{y}) \equiv \begin{pmatrix} \boldsymbol{K}\boldsymbol{u} + \boldsymbol{N}^\top \boldsymbol{\lambda}_\nu + \boldsymbol{T}^\top \boldsymbol{\lambda}_t \\ \boldsymbol{\lambda}_\nu - P_{\mathbb{R}^m_+}(\boldsymbol{\lambda}_\nu + \rho \boldsymbol{N}\boldsymbol{u}) \\ \boldsymbol{\lambda}_t - P_{[-\mathcal{F}\boldsymbol{\lambda}_\nu, \mathcal{F}\boldsymbol{\lambda}_\nu]}(\boldsymbol{\lambda}_t + \rho \boldsymbol{T}\boldsymbol{u}) \end{pmatrix} = \begin{pmatrix} \boldsymbol{f} \\ 0 \\ 0 \end{pmatrix}, \tag{4}$$

where $\boldsymbol{y} = (\boldsymbol{u}, \boldsymbol{\lambda}_\nu, \boldsymbol{\lambda}_t) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^m$.

The mapping $G : \mathbb{R}^{n+2m} \mapsto \mathbb{R}^{n+2m}$ is continuous and piecewise smooth. In particular, it is *piecewise affine*, see e.g. [16] for the notion.

3. The semi-smooth Newton method

To solve (4), we apply the Newton iterations. Due to the nature of the mapping G, semi-smooth methods are applicable [2]. Let us also refer to [10], where this technique was used for solving the Signorini problem.

Let $\mathcal{M} = \{1, 2, \ldots, m\}$ be the set of all indices of contact points. Given $\boldsymbol{y} = (\boldsymbol{u}, \boldsymbol{\lambda}_\nu, \boldsymbol{\lambda}_t) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^m$, we define the *inactive* sets $\mathcal{I}_\nu = \mathcal{I}_\nu(\boldsymbol{y})$, $\mathcal{I}_t^+ = \mathcal{I}_t^+(\boldsymbol{y})$, $\mathcal{I}_t^- = \mathcal{I}_t^-(\boldsymbol{y})$ by

$$\begin{aligned} \mathcal{I}_\nu &= \{i \in \mathcal{M} : \lambda_{\nu,i} + \rho(\mathbf{N}\mathbf{u})_i < 0\}, \\ \mathcal{I}_t^+ &= \{i \in \mathcal{M} : \lambda_{t,i} + \rho(\mathbf{T}\mathbf{u})_i - \mathcal{F}\lambda_{\nu,i} > 0\}, \\ \mathcal{I}_t^- &= \{i \in \mathcal{M} : \lambda_{t,i} + \rho(\mathbf{T}\mathbf{u})_i + \mathcal{F}\lambda_{\nu,i} > 0\}, \end{aligned}$$

and the *active* sets $\mathcal{A}_\nu = \mathcal{A}_\nu(\boldsymbol{y})$, $\mathcal{A}_t = \mathcal{A}_t(\boldsymbol{y})$ as their complements:

$$\mathcal{A}_\nu = \mathcal{M} \setminus \mathcal{I}_\nu, \quad \mathcal{A}_t = \mathcal{M} \setminus (\mathcal{I}_t^+ \cup \mathcal{I}_t^-).$$

Let us introduce the indicator matrix $\boldsymbol{D}_\mathcal{S} \in \mathbb{R}^{m \times m}$ of $\mathcal{S} \subset \mathcal{M}$ as follows:

$$\boldsymbol{D}_\mathcal{S} = diag(s_1, \ldots, s_m), \quad s_i = \begin{cases} 1, & i \in \mathcal{S}, \\ 0, & i \in \mathcal{M} \setminus \mathcal{S}. \end{cases}$$

We observe that

$$G(\boldsymbol{y}) = \begin{pmatrix} \boldsymbol{K}\boldsymbol{u} + \boldsymbol{N}^\top \boldsymbol{\lambda}_\nu + \boldsymbol{T}^\top \boldsymbol{\lambda}_t \\ \boldsymbol{\lambda}_\nu - \boldsymbol{D}_{\mathcal{A}_\nu}(\boldsymbol{\lambda}_\nu + \rho \boldsymbol{N}\boldsymbol{u}) \\ \boldsymbol{\lambda}_t - \boldsymbol{D}_{\mathcal{A}_t}(\boldsymbol{\lambda}_t + \rho \boldsymbol{T}\boldsymbol{u}) - \boldsymbol{D}_{\mathcal{I}_t^+}\mathcal{F}\boldsymbol{\lambda}_\nu + \boldsymbol{D}_{\mathcal{I}_t^-}\mathcal{F}\boldsymbol{\lambda}_\nu \end{pmatrix} = J(\boldsymbol{y})\,\boldsymbol{y}\,,$$

where

$$J(\boldsymbol{y}) \equiv \left(\begin{array}{c|c|c} \boldsymbol{K} & \boldsymbol{N}^\top & \boldsymbol{T}^\top \\ \hline -\rho \boldsymbol{D}_{\mathcal{A}_\nu}\boldsymbol{N} & \boldsymbol{D}_{\mathcal{I}_\nu} & \boldsymbol{0} \\ \hline -\rho \boldsymbol{D}_{\mathcal{A}_t}\boldsymbol{T} & \mathcal{F}(\boldsymbol{D}_{\mathcal{I}_t^-} - \boldsymbol{D}_{\mathcal{I}_t^+}) & \boldsymbol{D}_{\mathcal{I}_t^+ \cup \mathcal{I}_t^-} \end{array} \right). \tag{5}$$

Note that the matrix $J(\boldsymbol{y})$ can be interpreted as a generalized Jacobi matrix namely, the differential of a slanting function related to the mapping G at the point $\boldsymbol{y}$, see [2]. We apply the *semi-smooth Newton method* for finding roots of (4).

ALGORITHM *SSNM*: Denote $\boldsymbol{F} \in \mathbb{R}^{n+2m}$, $\boldsymbol{F} \equiv (\boldsymbol{f}, 0, 0) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^m$, the right-hand side of (4). Set the tolerance $\varepsilon > 0$. Let $\boldsymbol{y}^{(0)} \in \mathbb{R}^{n+2m}$, $\rho > 0$, $k := 1$.

(i) Define the inactive/active sets related to $\boldsymbol{y}^{(k-1)}$. Assembly the relevant $J(\boldsymbol{y}^{(k-1)})$.

(ii) Compute $\boldsymbol{y}^{(k)}$ by solving the linear system

$$J(\boldsymbol{y}^{(k-1)})\,\boldsymbol{y}^{(k)} \;=\; \boldsymbol{F}\,. \tag{6}$$

(iii) If $||\boldsymbol{y}^{(k)} - \boldsymbol{y}^{(k-1)}||/||\boldsymbol{y}^{(k)}|| \le \varepsilon$, return $\boldsymbol{y} := \boldsymbol{y}^{(k)}$.

(iv) Set $k := k + 1$ and go to step *(i)*.

In the case of convergence, we define

$$\boldsymbol{y} = \mathit{SSNM}(\boldsymbol{y}^{(0)}, \boldsymbol{f})$$

as a numerical solution of problem (4). We usually set the tolerance $\varepsilon = 10^{-6}$, referring to the observation at the end of Section 1.

It is readily seen that if $\boldsymbol{y} = \mathit{SSNM}(\boldsymbol{y}^{(0)}, \boldsymbol{f})$, $\boldsymbol{y} = (\boldsymbol{u}, \boldsymbol{\lambda}_\nu, \boldsymbol{\lambda}_t) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^m$, then

$$(\boldsymbol{N}\boldsymbol{u})_i = 0,\ i \in \mathcal{A}_\nu, \quad (\boldsymbol{T}\boldsymbol{u})_i = 0,\ i \in \mathcal{A}_t, \tag{7}$$

$$\lambda_{\nu,i} = 0,\ i \in \mathcal{I}_\nu, \quad \lambda_{t,i} + \mathcal{F}\lambda_{\nu,i} = 0,\ i \in \mathcal{I}_t^-, \quad \lambda_{t,i} - \mathcal{F}\lambda_{\nu,i} = 0,\ i \in \mathcal{I}_t^+. \tag{8}$$

As the active sets are complementary to the inactive sets, they define decoupling of contact nodes into two groups, i.e. the nodes with the Dirichlet conditions (7) and the nodes with the Neumann conditions (8).

Take another view: We may try to *guess* the inactive sets $\mathcal{I} = \{\mathcal{I}_\nu; \mathcal{I}_t^+; \mathcal{I}_t^-\}$ on the contact. Due to the dichotomy, it would imply the information concerning

$\mathcal{A} = \{\mathcal{A}_\nu; \mathcal{A}_t\}$. Hence, given $\mathcal{I} = \{\mathcal{I}_\nu; \mathcal{I}_t^+; \mathcal{I}_t^-\}$ on the contact, and given a load $\boldsymbol{f}$, find $(\boldsymbol{u}, \boldsymbol{\lambda}_\nu, \boldsymbol{\lambda}_t) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^m$ such that

$$\left(\begin{array}{c|c|c} \boldsymbol{K} & \boldsymbol{N}^\top & \boldsymbol{T}^\top \\ \hline -\rho \boldsymbol{D}_{\mathcal{A}_\nu} \boldsymbol{N} & \boldsymbol{D}_{\mathcal{I}_\nu} & \boldsymbol{0} \\ \hline -\rho \boldsymbol{D}_{\mathcal{A}_t} \boldsymbol{T} & \mathcal{F}(\boldsymbol{D}_{\mathcal{I}_t^-} - \boldsymbol{D}_{\mathcal{I}_t^+}) & \boldsymbol{D}_{\mathcal{I}_t^+ \cup \mathcal{I}_t^-} \end{array}\right) \left(\begin{array}{c} \boldsymbol{u} \\ \hline \boldsymbol{\lambda}_\nu \\ \hline \boldsymbol{\lambda}_t \end{array}\right) = \left(\begin{array}{c} \boldsymbol{f} \\ \hline \boldsymbol{0} \\ \hline \boldsymbol{0} \end{array}\right) . \tag{9}$$

System (9) can be interpreted as the discrete form of the Lamé equations (1) with the Dirichlet and Neumann boundary conditions (7) and (8), respectively. It motivates to define the linear operator

$$\boldsymbol{y} = \mathit{DirNeu}(\mathcal{I}, \boldsymbol{f}) , \quad \boldsymbol{y} = (\boldsymbol{u}, \boldsymbol{\lambda}_\nu, \boldsymbol{\lambda}_t) . \tag{10}$$

Note that due to the clamping along Γ_u^1 and Γ_u^2, see Figure 1, the system (9) is uniquely solvable. The matrix $J(\boldsymbol{y})$ of this system is regular. This justifies, by the way, that iterations (6) are well defined.

Remark 3.1 Let $\boldsymbol{y}^{(0)} = \mathit{DirNeu}(\mathcal{I}, \boldsymbol{f})$. Then $\boldsymbol{y}^{(1)} = \mathit{SSNM}(\boldsymbol{y}^{(0)}, \boldsymbol{f})$ and $\boldsymbol{y}^{(1)} = \boldsymbol{y}^{(0)}$ i.e., Algorithm *SSNM* converges in the first iteration. In other words, $\boldsymbol{y}^{(0)} = \mathit{SSNM}(\boldsymbol{y}^{(0)}, \boldsymbol{f})$ is a fixed point of the iterations (6). Conversely, if $\boldsymbol{y}^{(0)} \in \mathbb{R}^{n+2m}$, $\boldsymbol{y}^{(0)} = \mathit{SSNM}(\boldsymbol{y}^{(0)}, \boldsymbol{f})$, then defining $\mathcal{I} = \{\mathcal{I}_\nu; \mathcal{I}_t^+; \mathcal{I}_t^-\}$ to be the inactive sets of $\boldsymbol{y}^{(0)}$, we have $\boldsymbol{y}^{(0)} = \mathit{DirNeu}(\mathcal{I}, \boldsymbol{f})$. In that case, the solutions of the Dirichlet-Neumann problem (9) and the Coulomb friction problem (4) are identical.

Remark 3.2 In principle, we could find *all* roots $\boldsymbol{y}$ of (4) i.e., all fixed points $\boldsymbol{y}$ of the iterations (6). Given $\boldsymbol{f}$, make a trial choice of the inactive sets $\mathcal{I} = \{\mathcal{I}_\nu; \mathcal{I}_t^+; \mathcal{I}_t^-\}$ on the contact. Apply Remark 3.1: Let $\boldsymbol{y}^{(0)} = \mathit{DirNeu}(\mathcal{I}, \boldsymbol{f})$. The trial choice is successful, provided that $\boldsymbol{y}^{(0)} = \mathit{SSNM}(\boldsymbol{y}^{(0)}, \boldsymbol{f})$. The trouble is that we would have to check all $3 \sum_{j=0}^{m} \binom{m}{j} = 3 \cdot 2^m$ variants of the inactive sets $\mathcal{I} = \{\mathcal{I}_\nu; \mathcal{I}_t^+; \mathcal{I}_t^-\}$, which is not reasonable.

Remark 3.3 Let $\boldsymbol{y} = \mathit{DirNeu}(\mathcal{I}, \boldsymbol{f})$. The mapping G, see (4), is *not* differentiable at $\boldsymbol{y}$ provided that the active sets $\mathcal{A}_\nu$ and $\mathcal{A}_t$ have a special property: there exists a contact point $i \in \mathcal{M}$ such that

$$\text{either} \qquad \lambda_{\nu,i} + \rho(\mathbf{N}\mathbf{u})_i = 0 \tag{11}$$

$$\text{or} \qquad \lambda_{t,i} + \rho(\mathbf{T}\mathbf{u})_i - \mathcal{F}\lambda_{\nu,i} = 0 \tag{12}$$

$$\text{or} \qquad \lambda_{t,i} + \rho(\mathbf{T}\mathbf{u})_i + \mathcal{F}\lambda_{\nu,i} = 0 . \tag{13}$$

4. Continuation

Consider the Coulomb friction model (1)-(3), i.e. (4), assuming that $\boldsymbol{f} = \boldsymbol{f}(\alpha)$ depends on a scalar parameter α. We impose a continuous loading regime and seek for *continuous* response of the model. In particular, we consider a linear *loading path*

$$\boldsymbol{f}(\alpha) = (1-\alpha)\boldsymbol{f}_1 + \alpha \boldsymbol{f}_2, \quad \alpha \in \mathbb{R}, \tag{14}$$

where $\boldsymbol{f}_1 \in \mathbb{R}^n$ and $\boldsymbol{f}_2 \in \mathbb{R}^n$ are given. The resulting *solution path* is a curve in $\mathbb{R}\times\mathbb{R}^{n+2m}$, see a qualitative sketch in Figure 2. It consists of *oriented linear branches*, connected by *transition points*.

Each oriented linear branch connecting transition points $(\alpha^{k-1}, \boldsymbol{y}^{k-1}) \in \mathbb{R}\times\mathbb{R}^{n+2m}$ and $(\alpha^k, \boldsymbol{y}^k) \in \mathbb{R} \times \mathbb{R}^{n+2m}$ is parameterized by α, and defined as

$$\alpha \longmapsto (\alpha, \boldsymbol{y}(\alpha)) \in \mathbb{R} \times \mathbb{R}^{n+2m}\,, \quad \boldsymbol{y}(\alpha) = \mathit{DirNeu}(\mathcal{I}, \boldsymbol{f}(\alpha))\,. \tag{15}$$

Note that the same branch (15) can have two different orientations. In particular,

- if $\alpha^{k-1} < \alpha^k$, we consider the positive orientation, i.e., $\alpha^{k-1} < \alpha < \alpha^k$, as α is increasing,
- if $\alpha^{k-1} > \alpha^k$, we consider the negative orientation, i.e., $\alpha^{k-1} > \alpha > \alpha^k$, as α is decreasing.

Let us emphasize that the inactive set $\mathcal{I}$ does not depend on the position of α in the above intervals. In Subsection 4.1, we give a predictor/corrector algorithm to follow such branch numerically. We can define the *orientation* of a particular branch by setting

$$s \equiv \frac{\alpha^k - \alpha^{k-1}}{|\alpha^k - \alpha^{k-1}|}\,.$$

Hence, orientation s attains the value $s = 1$ (positive orientation) and $s = -1$ (negative orientation). The mentioned predictor/corrector algorithm follows a branch with the same orientation s.

Oriented linear branch terminates in a transition point $(\alpha^k, \boldsymbol{y}^k) \in \mathbb{R} \times \mathbb{R}^{n+2m}$. It is related to a fixed point $\boldsymbol{y}^k = \mathit{SSNM}(\boldsymbol{y}^k, \boldsymbol{f}(\alpha^k))$. Due to Remark 3.1, $\boldsymbol{y}^k = \mathit{DirNeu}(\mathcal{I}, \boldsymbol{f}(\alpha^k))$, where $\mathcal{I} = \{\mathcal{I}_\nu; \mathcal{I}_t^+; \mathcal{I}_t^-\}$ are the inactive sets of $\boldsymbol{y}^k$. It can be shown that in a transition point $(\alpha^k, \boldsymbol{y}^k) \in \mathbb{R}\times\mathbb{R}^{n+2m}$, the mapping G, see (4), is *not* differentiable. We refer to Remark 3.3 for the analysis. Note that our objective is not to localize transition points exactly. In fact, due to rounding errors it is not possible. Instead, we develop computationally stable *branching* and *orientation* indicators which are formally related to each of the transition points, see Subsection 4.2.

4.1. Continuation of an oriented linear branch

Data of a linear branch: The orientation s and the fixed inactive set $\mathcal{I}$. The continuation algorithm is defined as a one-step recurrence

$$(\alpha_{i-1}, \boldsymbol{y}(\alpha_{i-1})) \in \mathbb{R} \times \mathbb{R}^{n+2m} \to (\alpha_i, \boldsymbol{y}(\alpha_i)) \in \mathbb{R} \times \mathbb{R}^{n+2m}\,.$$

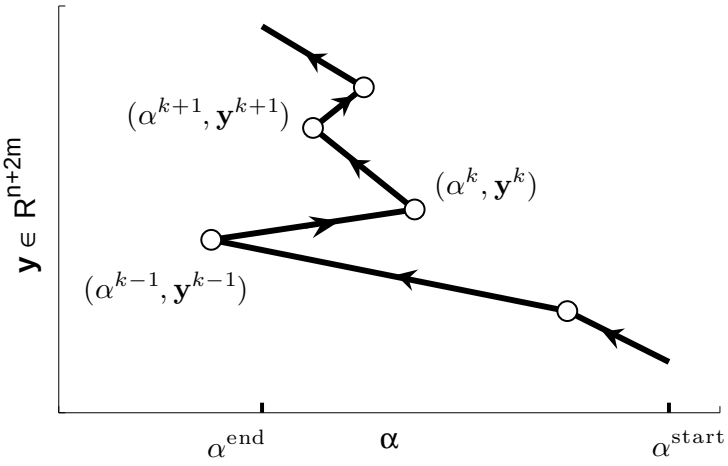


Figure 2: Solution path. For a fixed α, we may encounter up to five crossing points of the paths. They are related to five different solutions of equation (4) for the same right-hand side.

Parameters of the algorithm: The step-length h, in a range $0 < h_{\min} \le h \le h_{\max}$. The adaptive step-length strategy: Define c_s and c_p, $0 < c_s < 1 < c_p$, the *shortening* and the *prolongation* rates.

Let $(\alpha_{i-1}, \boldsymbol{y}(\alpha_{i-1})) \in \mathbb{R} \times \mathbb{R}^{n+2m}$ be given. Consider the following

PREDICTOR-CORRECTOR ALGORITHM:

(i) Predictor: $\alpha^{new} = \alpha_{i-1} + sh\,, \quad \boldsymbol{y}^{(0)} = DirNeu(\mathcal{I}, \boldsymbol{f}(\alpha^{new}))$.

(ii) Corrector:

`if` $\boldsymbol{y}^{(1)} = SSNM(\boldsymbol{y}^{(0)}, \boldsymbol{f}(\alpha^{new})) \quad \& \quad \boldsymbol{y}^{(1)} = \boldsymbol{y}^{(0)}$

`return` $\alpha_i := \alpha^{new}\,,\ \boldsymbol{y}(\alpha_i) := \boldsymbol{y}^{(1)}\,,\ i := i+1\,,\ h := \min(c_p h, h_{\max})$

`elseif` $h < h_{\min}$

`return` continuation failed, the last computed point of the branch: $(\alpha_{i-1}, \boldsymbol{y}(\alpha_{i-1}))$ with orientation s and the inactive set $\mathcal{I}$

`else` $h := \max(c_s h, h_{\min})\,,$ go to step *(i)*.

The algorithm returns either the new continuation point $(\alpha_i, \boldsymbol{y}(\alpha_i)) \in \mathbb{R} \times \mathbb{R}^{n+2m}$ with the same orientation s and the inactive set $\mathcal{I}$, or fails - the case which will be discussed in Subsection 4.2.

Note that the above algorithm can be characterized as a *tangent continuation*, see [3], Algorithm 4.25. The step-size control is inspired by [4].

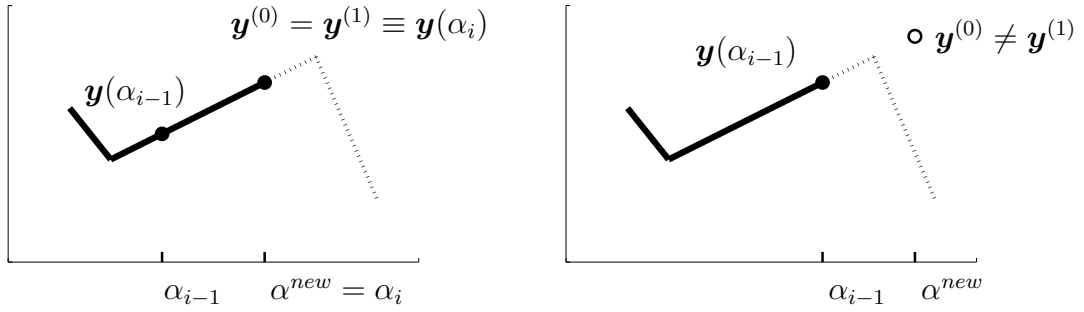


Figure 3: Oriented linear branch, predictor-corrector step. The corrector step is either accepted (on the left) or not accepted (on the right), and step-size h has to be shortened accordingly.

4.2. The branching and orientation indicators

Let $(\alpha_{i-1}, \boldsymbol{y}(\alpha_{i-1})) \in \mathbb{R} \times \mathbb{R}^{n+2m}$ be the last point of a linear branch with an orientation s and inactive set $\mathcal{I}$, see the failure of path following the linear branch in Subsection 4.1. Define a trial point $(\alpha^{\text{fail}}, \boldsymbol{y}^{\text{fail}}) \in \mathbb{R} \times \mathbb{R}^{n+2m}$ setting

$$\alpha^{\text{fail}} = \alpha_{i-1} + s h_{\text{fail}}\,, \quad \boldsymbol{y}^{\text{fail}} = \mathit{DirNeu}(\mathcal{I}, \boldsymbol{f}(\alpha^{\text{fail}}))\,, \tag{16}$$

where h_{fail} is the step-length related to the failure of continuation. Note that $\boldsymbol{y}^{\text{fail}} \neq \mathit{SSNM}(\boldsymbol{y}^{\text{fail}}, \boldsymbol{f}(\alpha^{\text{fail}}))$. Figure 4, the upper panel, suggests that $(\alpha_{i-1}, \boldsymbol{y}(\alpha_{i-1}))$ and $(\alpha^{\text{fail}}, \boldsymbol{y}^{\text{fail}})$ are close to a transition point. We may envisage two qualitatively different cases of the transition.

According to the generic scenario, we should indicate a change of $\mathcal{I}$: Let $\boldsymbol{u}$, $\boldsymbol{\lambda}_\nu$ and $\boldsymbol{\lambda}_t$ denote the solution components $\boldsymbol{y}(\alpha_{i-1}) = (\boldsymbol{u}, \boldsymbol{\lambda}_\nu, \boldsymbol{\lambda}_t) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^m$. Let

$$M = \min\left\{|\boldsymbol{\lambda}_\nu + \rho \boldsymbol{N}\boldsymbol{u}|,\ |\boldsymbol{\lambda}_t + \rho \boldsymbol{T}\boldsymbol{u} - \mathcal{F}\boldsymbol{\lambda}_\nu|,\ |\boldsymbol{\lambda}_t + \rho \boldsymbol{T}\boldsymbol{u} + \mathcal{F}\boldsymbol{\lambda}_\nu|\right\}\,. \tag{17}$$

The idea is that the minimizer of the above expression should indicate a transition point. We expect that just one component of the minimizer is significant. The transition is related to a transition between *inactive* and *active* sets. In this respect, the minimizer is interpreted as follows:

$$\left.\begin{array}{lll} \text{If } \ M = |\boldsymbol{\lambda}_\nu + \rho \boldsymbol{N}\boldsymbol{u}|_i, & \text{then } \ \mathcal{A}_\nu \overset{i}{\longleftrightarrow} \mathcal{I}_\nu \\ \text{else if } \ M = |\boldsymbol{\lambda}_t + \rho \boldsymbol{T}\boldsymbol{u} - \mathcal{F}\boldsymbol{\lambda}_\nu|_i, & \text{then } \ \mathcal{A}_t \overset{i}{\longleftrightarrow} \mathcal{I}_t^+ \\ \text{else if } \ M = |\boldsymbol{\lambda}_t + \rho \boldsymbol{T}\boldsymbol{u} + \mathcal{F}\boldsymbol{\lambda}_\nu|_i, & \text{then } \ \mathcal{A}_t \overset{i}{\longleftrightarrow} \mathcal{I}_t^- \end{array}\right\} \quad \mathcal{I}_{new} := \mathcal{I}. \tag{18}$$

The symbol "$\overset{i}{\longleftrightarrow}$" indicates a particular transition of the index i between the active and the inactive set. The procedure above results in an update of $\mathcal{I}$ denoted as $\mathcal{I}_{new}$.

We propose the following

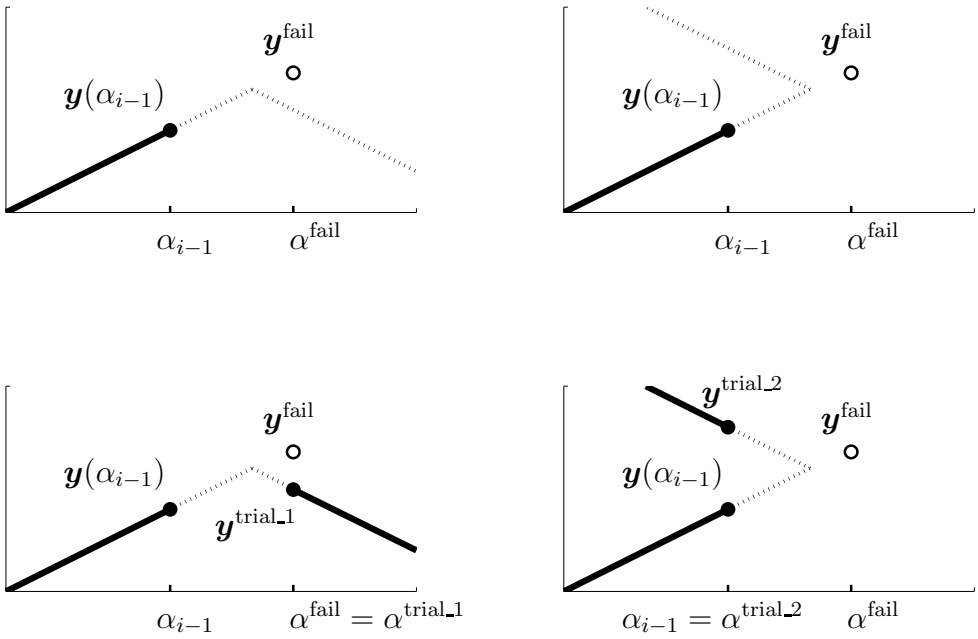


Figure 4: Let $(\alpha_{i-1}, \boldsymbol{y}(\alpha_{i-1})) \in \mathbb{R} \times \mathbb{R}^{n+2m}$ be the last point on a linear branch, continuation failure indicated on $(\alpha^{\text{fail}}, \boldsymbol{y}(\alpha^{\text{fail}})) \in \mathbb{R} \times \mathbb{R}^{n+2m}$. The upper panel, qualitative scenario envisaged: a) transversal transition on the left, b) fold (turning point) transition on the right. The lower panel: Branching due to the algorithm.

Branching Algorithm

Let $(\alpha_{i-1}, \boldsymbol{y}(\alpha_{i-1})) \in \mathbb{R} \times \mathbb{R}^{n+2m}$ be the last point of a linear branch with an orientation s and inactive set $\mathcal{I}$. Update $\mathcal{I}_{new}$ via (18).

Define $\alpha^{\text{trial\_1}} = \alpha_{i-1} + sh_{\text{fail}}$ and $\boldsymbol{y}^{\text{trial\_1}} = \mathit{DirNeu}(\mathcal{I}_{new}, \boldsymbol{f}(\alpha^{\text{trial\_1}}))$.

`If`

$\boldsymbol{y}^{\text{trial\_1}} = \mathit{SSNM}(\boldsymbol{y}^{\text{trial\_1}}, \boldsymbol{f}(\alpha^{\text{trial\_1}})$, set $(\alpha^{\text{trial\_1}}, \boldsymbol{y}^{\text{trial\_1}}) \in \mathbb{R} \times \mathbb{R}^{n+2m}$ to be the first point on a linear branch with the orientation $s := s$ and the inactive set $\mathcal{I} := \mathcal{I}_{new}$.

Comment: *transversal transition.*

`else`

Define $\alpha^{\text{trial\_2}} = \alpha_{i-1}$ and $\boldsymbol{y}^{\text{trial\_2}} = \mathit{DirNeu}(\mathcal{I}_{new}, \boldsymbol{f}(\alpha^{\text{trial\_2}}))$.

Set $(\alpha^{\text{trial\_2}}, \boldsymbol{y}^{\text{trial\_2}}) \in \mathbb{R} \times \mathbb{R}^{n+2m}$ to be the first point of a linear branch with orientation $s := -s$ and inactive set $\mathcal{I} := \mathcal{I}_{new}$.

Comment: *fold, turning point.*

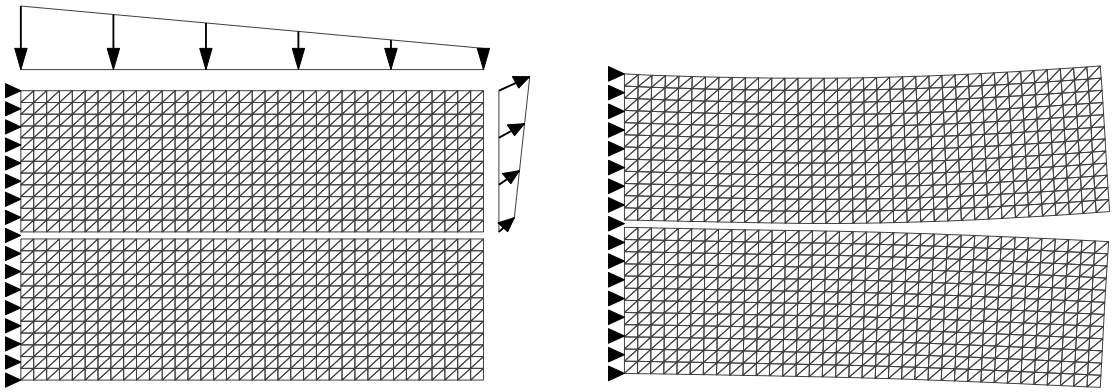


Figure 5: Contact of two elastic bodies Ω^1 (the upper body) and Ω^2, along the contact boundary Γ_c. The loading is due to the surface traction. On the right: Resulting deformation.

Set $i := i + 1$, and apply continuation of the oriented linear branch with orientation s and the inactive set $\mathcal{I}$.

The idea of the algorithm is indicated in the lower panel of Figure 4. The algorithm works provided that $h_{\min}$ is sufficiently small.

The ambition of the present paper is not to justify the branching scenario theoretically. Note only, that the transversal transition may be described using a proper version of the Implicit Function Theorem, see e.g. [15, 5] and [8] in the context of Coulomb friction. In case of the fold transition, we cannot quote (to our knowledge) a relevant analytical result immediately.

5. Numerical experiments

We consider a particular geometry, see Figure 5.

The actual computations are depicted in Figure 6. If $\mathcal{F}$ is sufficiently small then the solution path should contain transversal transition points only, see e.g. [8]. It pertains to Figure 6, upper left. For larger friction coefficients (e.g. $\mathcal{F} = 0.6$ and $\mathcal{F} = 30$), the path-following algorithm reveals non-unique solutions of the problem, see Figure 6, upper right and lower-left including the corresponding zoom. In particular, we can have up to three ($\mathcal{F} = 0.6$) and five ($\mathcal{F} = 30$) solutions for a fixed parameter α.

Acknowledgements

This work was supported by the grant GA CR P201/12/0671.

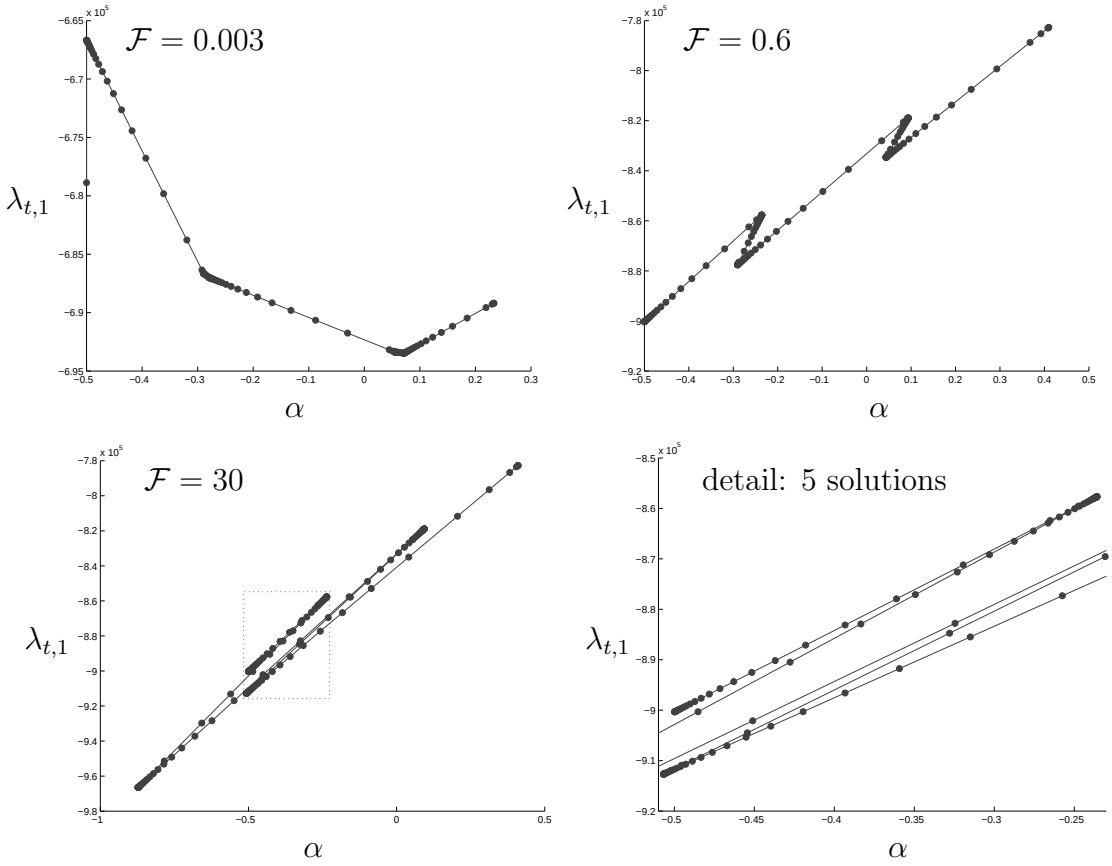


Figure 6: Discretization: $n = 1320$, $m = 30$. The stepsize control: $10^{-5} \leq h \leq 5$, $c_p = 1.3$, $c_s = 0.5$. Plots: Parameter α vs. the solution component $\lambda_{t,1}$, for selected friction coefficients $\mathcal{F}$.

References

[1] Allgower, E. L. and Georg, K.: *Numerical path following.* Handbook of Numerical Analysis, vol. V, Elsevier Science, New York, 1997.

[2] Chen, X., Nashed, Z., and Qi, L.: Smoothing methods and semismooth methods for nondifferential operator equations. SIAM J. Numer. Anal. **5** (2000), 1200–1216.

[3] Deuflhart, P. and Hohmann, A.: *Numerical analysis in modern scientific computing.* Texts in Applied Mathematics, Springer Verlag, New York, 2003.

[4] Dhooge, A., Govaerts, W., and Kuzetsov, Y. A.: MATCONT: A Matlab package for numerical bifurcation analysis of ODEs. ACM Trans. Math. Software **31** (2003), 141–164.

[5] Dontchev, A. L. and Rockafellar, R. T.: Robinson's implicit function theorem and its extensions. Math. Program. **117** (2009), 129–147.

[6] Eck, C. and Jarušek, J.: Existence results for the static contact problems with Coulomb friction. Math. Models Methods Appl. Sci. **8** (1997), 445–468.

[7] Facchinei, F. and Pang, J.: *Finite-dimensional variational inequalities and complementarity problems.* Springer Series in Operations Research xxxiii, Springer Verlag, New York, 2003.

[8] Haslinger, J., Janovský, V., and Ligurský, T.: Qualitative analysis of solutions to discrete static contact problems with Coulomb friction. Comput. Meth. Appl. Mech. Engrg. **205–208** (2012), 149–161.

[9] 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.

[10] Ito, K. and Kunisch, K.: Semi-smooth Newton methods for the Signorini problem. Appl. Math. **53** (2009), 455–468.

[11] Janovský, V.: Catastrophic features of Coulomb friction model. In: J. R. Whiteman (Ed.), *The Mathematics of Finite Elements and Applications IV*, pp. 259–264. Academic Press, New York, 1982.

[12] Janovský, V. and Ligurský, T.: Computing non unique solutions of the Coulomb friction problem. Math. Comput. Simulation **82** (2012), 2047–2061.

[13] Ligurský, T.: Theoretical analysis of discrete contact problems with Coulomb friction. Appl. Math. **57** (2012), 263–295.

[14] Nečas, J., Jarušek, J., and Haslinger, J.: On the solution of variational inequality to the Signorini problem with small friction. Bolletino U.M.I. **5** (1980), 796–811.

[15] Robinson, S.: An implicit-function theorem for a class of nonsmooth functions. Math. Oper. Res. **16** (1991), 292–309.

[16] Scholtes, S.: *Introduction to piecewise differentiable equations.* SpringerBriefs in Optimalization, Springer, Berlin, 2012.

Conference Applications of Mathematics 2013
in honor of the 70th birthday of Karel Segeth.
Jan Brandts, Sergey Korotov, Michal Křížek,
Jakub Šístek, and Tomáš Vejchodský (Eds.),
Institute of Mathematics AS CR, Prague 2013

FAST OPTICAL TRACKING OF DIFFUSION IN TIME-DEPENDENT ENVIRONMENT OF BRAIN EXTRACELLULAR SPACE

Jan Hrabě[1,2]

[1] Center for Advanced Brain Imaging, Nathan S. Kline Institute
140 Old Orangeburg Road, Orangeburg, NY 10962, USA
hrabe@nki.rfmh.org
[2] Department of Cell Biology, SUNY Downstate Medical Center
Brooklyn, NY 11203, USA

Abstract

An improved version of the Integrative Optical Imaging (IOI) method for diffusion measurements in a geometrically complex environment of the brain extracellular space has been developed. We present a theory for this Fast Optical Tracking Of Diffusion (FOTOD) which incorporates a time-dependent effective diffusion coefficient in homogeneous anisotropic media with time-dependent nonspecific linear clearance. FOTOD can be used to measure rapid changes in extracellular diffusion permeability that occur, e.g., during brain insults. The achievable time resolution is approximately one second, a ten fold improvement compared to the traditional IOI method.

1. Introduction

Brain cells (neurons and glia) are surrounded by an extracellular space (ECS) that facilitates diffusion transport of neuroactive substances, nutrients, metabolites and therapeutic agents. Our knowledge about the ECS in living brain tissue has largely been deduced from studying diffusion of extracellular marker molecules [2]. The ECS is a geometrically complex porous environment [4] characterized by two basic properties: volume fraction α and diffusion permeability θ, see [1]. Volume fraction is the proportion of brain tissue volume occupied by the ECS and primarily governs concentration of molecules released into the ECS. Diffusion permeability, a ratio of the effective diffusion coefficient to its value in an obstacle-free medium, describes how much a diffusion-mediated process is slowed down in the ECS by obstacles represented by the cells and their various appendages. One additional parameter, κ, accounts for small nonspecific clearance proportional to the concentration. It describes nonspecific loss of marker molecules over time, e.g., into blood stream.

We will address the physiologically important situation where the diffusion permeability depends on time, as is observed during brain insults, e.g., following a stroke.

Our assumption is that the brain ECS environment remains homogeneous, that is, the time-dependent changes are everywhere the same. However, we do allow the medium to be anisotropic, as typified by white matter fiber tracts.

The diffusion experiment consists of releasing a small amount of a fluorescent substance into the ECS from a glass micropipette and repeatedly recording the resulting diffusion cloud with a charge-coupled device (CCD) camera. Because the camera observes an image formed by a microscope, the optical properties of the imaging system (its point-spread function) must be taken into account.

2. Theory

We shall investigate concentration $c(\vec{r},t)$ of some extracellular marker as a function of position in space $\vec{r}=(x_1,x_2,x_3)$ and time t. In a geometrically complex ECS, all the diffusion parameters are defined as volume-averaged local quantities over a sufficiently large sampling volume. The concentration is related to the tissue volume rather than the ECS volume because the optical method does not distinguish between the tissue compartments. We assume that a homogeneous but anisotropic environment with time-dependent diffusion characteristics can be described by an effective diffusion tensor

$$D^{\star}_{ij}(t) = D\Theta_{ij}\,, \tag{1}$$

where D is the scalar free diffusion constant and Θ_{ij} is the diffusion permeability tensor. Both indices run from 1 to 3. In an environment where the loss of diffusing substance is proportional to the concentration, we also introduce linear time-dependent clearance $\kappa(t)$, which is also assumed to be homogeneous. The diffusion equation in this environment is

$$\frac{\partial c(\vec{r},t)}{\partial t} = D^{\star}_{ij}(t)\frac{\partial^2 c(\vec{r},t)}{\partial x_i \partial x_j} - \kappa(t)c(\vec{r},t)\,, \tag{2}$$

where we used Einstein's notation for sums ($a_i b_i = \mathbf{a}\cdot\mathbf{b} = \sum_{i=1}^{3} a_i b_i$). Equation (2) expresses the mass preservation when the diffusion flow $\vec{j}$ obeys Fick's law

$$j_i(\vec{r},t) = -D_{ik}(t)\frac{\partial c(\vec{r},t)}{\partial x_k}\,.$$

The initial concentration at time t_0 is represented by a function $c(\vec{r},t_0)$.

Equation (2) with its initial condition can be solved in the Fourier domain. Fourier transform of $c(\vec{r},t)$ with respect to the three spatial coordinates is defined as

$$\hat{c}(\vec{k},t) = \iiint_{-\infty}^{\infty} c(\vec{r},t)\exp(2\pi i k_j x_j)\,d\vec{r}\,,$$

and the inverse as

$$c(\vec{r},t) = \iiint_{-\infty}^{\infty} \hat{c}(\vec{k},t)\exp(-2\pi i k_j x_j)\,d\vec{k}\,.$$

The Fourier transform turns Eq. (2) into

$$\frac{\partial \hat{c}(\vec{k},t)}{\partial t} = -\left(4\pi^2 k_i D^{\star}_{ij}(t) k_j + \kappa(t)\right) \hat{c}(\vec{k},t) \tag{3}$$

with the initial condition $\hat{c}(\vec{k},t_0)$.

Solving Eq. (3) with respect to time yields

$$\hat{c}(\vec{k},t) = \hat{c}(\vec{k},t_0)\hat{c}_\delta(\vec{k},t)\,, \tag{4}$$

where

$$\hat{c}_\delta(\vec{k},t) = Q_\delta(t)\exp\left(-2\pi^2 k_i \Sigma_{ij}(t) k_j\right)\,, \tag{5}$$

$$Q_\delta(t) = \exp\left(-\int_{t_0}^{t} \kappa(t')\,dt'\right)\,, \tag{6}$$

and

$$\Sigma_{ij}(t) = 2\int_{t_0}^{t} D^{\star}_{ij}(t')\,dt'\,. \tag{7}$$

When the initial condition is Dirac's δ-function $\delta(\vec{r})$, its Fourier transform is unity. The inverse Fourier transform $c_\delta(\vec{r},t)$ of $\hat{c}_\delta(\vec{k},t)$ therefore describes the diffusion cloud initiated by the point source at time $t = t_0$:

$$c_\delta(\vec{r},t) = Q_\delta(t)\phi_\delta(\vec{r},t)\,, \tag{8}$$

where

$$\phi_\delta(\vec{r},t) = \frac{1}{(2\pi)^{\frac{3}{2}}\left[\det(\Sigma_{ij})\right]^{\frac{1}{2}}} \exp\left(-\frac{x_i \Sigma^{-1}_{ij}(t) x_j}{2}\right)\,. \tag{9}$$

This is a 3D Gaussian distribution with variance matrix $\Sigma_{ij}(t)$ and with the total amount of diffusing substance decreasing as $Q_\delta(t)$ from its initial value of $Q_\delta(t_0) = 1$.

Multiplication in the Fourier domain corresponds to a convolution in the spatial domain. The concentration distribution following an arbitrary initial condition can therefore be written as

$$c(\vec{r},t) = \iiint_{-\infty}^{\infty} c(\vec{r}\,',t_0) c_\delta(\vec{r}-\vec{r}\,',t)\,d\vec{r}\,'\,. \tag{10}$$

The total amount of diffusing substance is initially $Q(t_0) = \iiint_{-\infty}^{\infty} c(\vec{r},t_0)\,d\vec{r}$ and changes with time as

$$Q(t) = \iiint_{-\infty}^{\infty} c(\vec{r},t)\,d\vec{r} = Q(t_0)Q_\delta(t)\,, \tag{11}$$

where $Q_\delta(t)$ is substituted from Eq. (6). If a 3D measurement of concentration in time is available, the clearance $\kappa(t)$ can be computed from Eq. (11):

$$\kappa(t) = -\frac{d}{dt}\ln\left(\frac{Q(t)}{Q(t_0)}\right)\,. \tag{12}$$

Since the substance loss is homogeneous in space, the effect of nonzero clearance simply amounts to a global scaling of amplitude.

If the measured 3D concentration is normalized by the total amount $Q(t)$ of the diffusing substance at every time, a probability density function

$$\phi(\vec{r},t) = \frac{c(\vec{r},t)}{Q(t)} \tag{13}$$

can be constructed and the tensor of its second moments $\mu_{ij}(t)$ computed:

$$\begin{aligned}
\mu_{ij}(t) &= \iiint_{-\infty}^{\infty} x_i x_j \phi(\vec{r},t)\, d\vec{r} \\
&= \frac{1}{Q(t_0)} \iiint_{-\infty}^{\infty} c(\vec{r}\,',t_0) \left[\iiint_{-\infty}^{\infty} x_i x_j \phi_\delta(\vec{r}-\vec{r}\,',t)\, d\vec{r} \right] d\vec{r}\,' \\
&= \frac{1}{Q(t_0)} \iiint_{-\infty}^{\infty} c(\vec{r}\,',t_0) \left[\Sigma_{ij}(t) + x_i' x_j' \right] d\vec{r}\,' \\
&= \Sigma_{ij}(t) + \mu_{ij}(t_0)\,.
\end{aligned} \tag{14}$$

The components of the effective diffusion tensor are now easily extracted as time derivatives of these moments:

$$D^\star_{ij}(t) = \frac{1}{2}\frac{d\mu_{ij}(t)}{dt}\,. \tag{15}$$

Unfortunately, a complete 3D measurement of the concentration is not usually available. More common is an experimental setup with a traditional (non-confocal) microscope where a 2D image is recorded. Because the microscope's objective has a finite aperture, the system appears to be imaging a virtual object, constructed from the true object by a convolution with the point-spread function (PSF) $S(\vec{r})$ of the system. The effective width of the PSF limits the system resolution. Using the approximation for $S(\vec{r})$ suggested by [3], we can derive estimates for the effective "horizontal" and "vertical" resolutions Δ_h and Δ_v, respectively:

$$\Delta_h = 0.61\lambda \frac{\sqrt{n^2 - N_A^2}}{nN_A} \quad \text{and} \quad \Delta_v = 2\lambda \frac{n^2 - N_A^2}{nN_A^2}\,, \tag{16}$$

where λ is the wavelength, n is the index of refraction of the environment under the objective, and N_A is the numerical aperture. The horizontal resolution is typically smaller than the corresponding size of the recorded image pixel and the horizontal PSF effect can thus be safely ignored. Resolution Δ_v along the microscope optical axis is usually much lower and cannot be ignored. Under these assumptions, we can utilize the PSF approximation in the object space

$$S(\vec{r}) = \delta(x_1)\delta(x_2)S_v(x_3)\,, \tag{17}$$

where

$$S_v(x_3) = \frac{1}{\Delta_v} \operatorname{sinc}^2\left(\frac{\pi x_3}{\Delta_v}\right) \tag{18}$$

and $\operatorname{sinc}(x) = \sin(x)/x$. We shall see that the exact functional form of $S_v(x_3)$ is not important but the validity of the approximation given by Eq. (17) is.

If the PSF was very sharp ($S(\vec{r}) = \delta(\vec{r})$), the imaging system would simply record 2D image proportional to the concentrations $c(x_1, x_2, x_3 = z_0, t)$ in the plane of focus $x_3 = z_0$. Due to the PSF blurring effect, however, it instead appears to detect signal originating from concentration

$$\begin{aligned} c_s(x_1, x_2, z_0, t) &= \iiint_{-\infty}^{\infty} c(\vec{r}\,', t) S(\vec{r} - \vec{r}\,')\, d\vec{r}\,' \\ &= \int_{-\infty}^{\infty} c(x_1, x_2, x_3', t_0) S_v(z_0 - x_3')\, dx_3' \\ &= \iiint_{-\infty}^{\infty} c_\delta(\vec{r}\,', t) \int_{-\infty}^{\infty} c(x_1 - x_1', x_2 - x_2', \xi, t_0) S_v(z_0 - x_3' - \xi)\, d\xi\, d\vec{r}\,' \\ &= \iiint_{-\infty}^{\infty} c_\delta(\vec{r}\,', t) c_s(x_1 - x_1', x_2 - x_2', z_0 - x_3', t_0)\, d\vec{r}\,' \,. \end{aligned} \tag{19}$$

It can be seen that the effect of microscope's PSF in our approximation results in a simple modification (blurring along the x_3 axis) of the initial condition $c(\vec{r}, t_0)$ to $c_s(\vec{r}, t_0)$. After this modification, we can consider the microscope to perfectly follow the rules of geometrical optics, magnifying the image by a constant factor M and amplifying the image signal by another constant factor A. A single in-focus plane $x_3 = z_0$ through c_s is imaged.

Assuming that a 2D section at $x_3 = z_0$ of the concentration cloud elicited by the initial condition $c_s(\vec{r}, t_0)$ represents all the information that is available to us, let us extract as much as possible from it. The image intensity $I(x_1, x_2, t)$ expressed in the object coordinates after constant amplification A is

$$I(x_1, x_2, t) = A c_s(x_1, x_2, z_0, t)\,. \tag{20}$$

For the integrated total image intensity $Q_I(t)$ we get

$$\begin{aligned} Q_I(t) &= \iint_{-\infty}^{\infty} I(x_1, x_2, t)\, dx_1\, dx_2 \\ &= A Q_\delta(t) \iiint_{-\infty}^{\infty} c_s(\vec{r}\,', t_0) \phi_{\delta v}(z_0 - x_3', \Sigma_{33})\, d\vec{r}\,' \,, \end{aligned} \tag{21}$$

where

$$\phi_{\delta v}(\xi, \Sigma_{33}) = \frac{1}{\sqrt{2\pi\Sigma_{33}}} \exp\left(-\frac{\xi^2}{2\Sigma_{33}}\right) \tag{22}$$

is the "vertical" portion of the Gaussian distribution ϕ_δ. We have made the dependency on $\Sigma_{33} = \Sigma_{33}(t)$ explicit to emphasize it. In contrast to the 3D measurement, the time dependency is not fully determined by the clearance term $Q_\delta(t)$.

Finally, let us calculate the second moment $\mu_{JK}(t)$ of the image. Capitalized indices are introduced to distinguish their 2D range ($J, K = 1, 2$).

$$\begin{aligned}\mu_{JK}(t) &= \frac{1}{Q_I(t)} \iint_{-\infty}^{\infty} x_J x_K I(x_1, x_2, t)\, dx_1\, dx_2 \\ &= \frac{AQ_\delta(t)}{Q_I(t)} \iiint_{-\infty}^{\infty} c_s(\vec{r}\,', t_0) \left[\iint_{-\infty}^{\infty} x_J x_K \phi_\delta(\vec{r} - \vec{r}\,', t)\, dx_1\, dx_2 \right] d\vec{r}\,' .\end{aligned} \tag{23}$$

In the general anisotropic case with an arbitrarily rotated coordinate system, the result is rather complicated. However, it is usually possible to make one of the principle axes parallel to the x_3 axis of the imaging system. We then have

$$\Sigma_{ij}(t) = \begin{pmatrix} \Sigma_{11}(t) & \Sigma_{12}(t) & 0 \\ \Sigma_{21}(t) & \Sigma_{22}(t) & 0 \\ 0 & 0 & \Sigma_{33}(t) \end{pmatrix} .$$

This finally leads to

$$\mu_{JK}(t) = \Sigma_{JK}(t) + \frac{\iiint_{-\infty}^{\infty} x_J x_K c_s(\vec{r}, t_0) \phi_{\delta v}(z_0 - x_3, \Sigma_{33}(t))\, d\vec{r}}{\iiint_{-\infty}^{\infty} c_s(\vec{r}, t_0) \phi_{\delta v}(z_0 - x_3, \Sigma_{33}(t))\, d\vec{r}} . \tag{24}$$

3. Discussion

In Eq. (24), we have obtained a result useful for a biologist employing the FOTOD modification of the IOI method. It provides a means of extracting the effective diffusion coefficient as a function of time.

The experimentally accessible quantity is $\mu_{JK}(t)$. However, its time derivative yields the corresponding components of the diffusion tensor D_{JK} only if the last term of Eq. (24) is constant in time. Its time dependency is determined by the shape of the initial condition. When the initial condition is separable in the variable x_3, the blurred initial condition will also be separable, and the time dependency of this term will "cancel out". Therefore, for a commonly assumed Gaussian initial condition, which is obviously separable, the extraction of the diffusion tensor is straightforward. For a more realistic spherical initial condition though, the experimental curve of variance versus time will not be linear even when the diffusion tensor is constant in time.

Acknowledgments

This work was supported by NIH grants R56-NS047557 and R01-NS047557 (PI Sabina Hrabětová).

References

[1] Hrabě, J., Hrabětová, S., and Segeth, K.: A model of effective diffusion and tortuosity in the extracellular space of the brain. Biophys. J. **87** (2004), 1606–1617.

[2] Hrabětová, S. and Nicholson, C.: *Eletrochemical methods for neuroscience*. CRC, Boca Raton, Florida, 2007.

[3] Nicholson, C. and Tao, L.: Hindered diffusion of high molecular weight compounds in brain extracellular microenvironment measured with integrative optical imaging. Biophys. J. **65** (1993), 2277–2290.

[4] Syková, E. and Nicholson, C.: Diffusion in brain extracellular space. Physiol. Rev. **88** (2008), 1277–1340.

Conference Applications of Mathematics 2013
in honor of the 70th birthday of Karel Segeth.
Jan Brandts, Sergey Korotov, Michal Křížek,
Jakub Šístek, and Tomáš Vejchodský (Eds.),
Institute of Mathematics AS CR, Prague 2013

DETECTION CODES IN RAILWAY INTERLOCKING SYSTEMS

Lucie Kárná[1], Štěpán Klapka[2]

[1] Faculty of Transportation Sciences, Czech Technical University
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

This paper describes a model of influence of random errors on the safety of the communication. The role of the communication in railway safety is specified.

To ensure a safe communication, using of safety code is important. The most important parameter of the safety code is the maximal value of the probability of undetected error. Problems related with computing of this value are outlined in the article. As a model for the information transmission the binary symmetrical channel is introduced.

The usability of the concept of a 'proper' code is discussed.

1. Introduction

This article discuss safety of communication between components of a railway interlocking systems (for example level crossings). The term *safety* is defined as absence of unacceptable level of hazard. This definition is not quite understandable without an additional explanation. However, for purpose of this paper it is sufficient to consider the word "safety" in its common sense.

The safety of a system has two main aspects. A *functional safety* concerns the manner how the system reacts on various combinations of outer inputs and its inner states ("what does it do?"). A *safety integrity* means ability of the system to really perform required functions ("does it really work?"). The safety integrity concerns the resistance of the system against both systematic and random errors. Nevertheless, only the requirements on the integrity in relation to random errors can be quantified.

This paper focuses on only a small part of the safety issues, in particular on a model of influence of random errors on the safety integrity, namely on communication safety.

1.1. Communication safety

Let us introduce the term "safe communication". A safe communication must ensure the following requirements:

- a message originates from the intended source (message *authenticity*),
- received information is complete and unchanged (*integrity*),
- messages are delivered in the right time (*timeliness*), and
- in the right sequence (correct *ordering*).

Some applications require *confidentiality* as an additional safety service – that the information cannot be disclosed to unauthorized subjects.

Many techniques can be used to ensure the *safety services* introduced above. Since every from these techniques provides protection against separate elementary errors, usually combination of several of them is employed. We can add a sequence number, a time stamp, or source and recipient identifier to the message. We can check the maximum time delay between two messages. The receiver can send an acknowledge message back to the sender. We can introduce a more sophisticated procedure of identification of communication participants. We can secure the message by a safety code or by cryptographic techniques.

The safety code has a special position among defense techniques, as it is the unique method of protection of messages against corruption. A safety-responsible protocol layer then should implement safety code to ensure integrity of messages. International safety standards for various types of systems state the usage of safety codes as mandatory requirement (for example [1] for railway applications).

2. Safety codes

An *error detection code* is a code detecting presence of some amount of errors in received messages. Error detection codes are used to overcome or reduce the impact of communication channel errors. However, these codes cannot provide a perfect protection and some amount of residual errors passes through undetected. Quantification of probability of occurrence of a residual error is a keystone of the probabilistic safety integrity study. A *safety code* is an error detection code used as a means to ensure safety in safety relevant communication system.

2.1. Linear binary codes

The "code-related" terminology in this paper is based on terms used in mathematical coding theory (see for example [3]). In this article we restrict ourselves only to linear binary detection codes, with codewords of length n bits, and with k information bits, defined as follows:

A linear binary $(n, k)-$*code* $\mathbf{K}$ 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 $\mathbf{K}$

are *codewords*. In the $(n, k)-$code the codeword length is n, number of information bits is equal to k and number of redundant bits is equal to $n - k$.

The most simple example of the linear binary code is a parity check. The *even parity* code consists from all words of the given length n, in which the count of ones is even. This code has $n - 1$ information bits and 1 redundant bit. The parity check is used as a safety code in most harware and software applications.

2.2. Error detection

In transfer of the encoded information in the space (transmission and reception of the message) or in the time (usage of data storage medium to record and later restore the message) the message can be modified by various external influences. On the level of individual bits, a modification can manifest by missing or superfluous bit(s), or by altered bits with overall number of bits preserved. In this paper, we ignore the first type of modification (synchronization slip) and focus solely on the second type – modifications that do not change the number of bits.

Let us describe the mechanism of detecting these modifications. A source intends to send a k-bit message. The error detecting code generates an n-bit codeword u, and the source transfers this codeword. A target receives an n-bit word v from $\mathbf{Z}_2^n$, not necessarily a codeword. If the received word v is not a codeword, then the receiver detects an *error*.

The second possibility is that the received word v is a codeword. Then there are two possible scenarios: The received codeword v is equal to the original codeword u, because there were no modifications in transfer. Alternatively the received codeword v is different from the original codeword u, because a modification during transfer unfortunately creates some codeword. The receiver has no possibility to recognize, which one from this scenarios occurs. The second scenario is then bad and results in an *undetected error*. The probability of such undetected error of error detection codes used in safety relevant applications (including transportation control) is very important safety parameter.

The difference $v - u$ between the received word v and the original word u is called an *error word*. The undetected error words of a linear code are all nonzero codewords of the given code, due to its linearity (see for example [3]). This is a great advantage of using of linear codes, as this make probability calculations more feasible than for other types of codes.

2.3. Weight structure

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 codeword.

The minimal distance of a linear code sets the ability of the code to detect some classes of transmission errors. A code with minimal distance d will detect all errors with at most $d - 1$ modified bits in transmitted codeword (see [3]). Such a code will not detect all errors with d or more modified bits. Nevertheless, some cases of

modifications of d or more symbols will be detected. Various codes with equal n, k, and d differ in their capability of detecting modifications with d or more changes, and thus such codes differ in their undetected error probability.

For more detailed description of the code we define a *weight structure* of the code as a vector $A = (A_1, A_2, \ldots, A_n)$, where A_i denotes count of codewords with Hamming weight equal to i. For linear codes, the weight structure is fully sufficient for description of the ability of the code to detect modifications of d or more symbols, as we show in the following analysis.

2.4. Probability of undetected error

Let us derive a formula for the probability of undetected error of binary linear code. This probability is equal to probability of receiving a non-zero codeword if the zero codeword is transmitted (for details see [3]).

Consider a transmission of the zero codeword. Suppose we received a word with exactly i non-zero bits. The probability that the received word is a codeword is the ratio between the count of all codewords with i non-zero symbols (A_i from the weight structure of the code), and count of all words with i non-zero bits $\binom{n}{i}$. Denoting P_i the probability that received word has exactly i wrong bits, then the probability P_{ud} of an undetected error of the code is equal to

$$P_{ud} = \sum_{i=1}^{n} P_i \frac{A_i}{\binom{n}{i}}. \tag{1}$$

The probability P_i that exactly i bits are modified during transmission is independent of the code properties, and depends solely on conditions of the information transmission.

There are various models of communication channels, with varying characteristics. Choosing the right model of a communication channel that corresponds to real-world conditions, and produces useful results in our calculations is rather a difficult process.

2.5. Binary symmetrical channel

The most frequently used transmission channel model is a *memoryless binary symmetrical channel (BSC)*. This is a simple probability model based on a bit transmission, parametrized by a constant probability of bit modification p_e (*bit error rate*). In this model, a transmitted bit is modified during the transmission with the probability p_e, regardless of its original value and independently of other bits in the transmission.

In the BSC model, the probability that a word with n bits is received with i bits modified is equal to

$$P_i = \binom{n}{i} p_e^i (1 - p_e)^{n-i}. \tag{2}$$

Substituting (2) to the formula (1), for the probability of undetected error we get:

$$P_{ud}(p_e) = \sum_{i=1}^{n} {p_e}^i (1-p_e)^{n-i} A_i. \tag{3}$$

For a final calculation, it is necessary to know A_i, the quantities of codewords with the Hamming weight equal to i. A binary linear (n, k)-code is a k-dimensional linear subspace of $\mathbf{Z}_2^n$. It has exactly 2^k elements, and thus

$$A_0 + A_1 + \ldots + A_n = 2^k.$$

As every linear code contains a zero word, A_0 equals one. For minimal distance of the code equals to d, A_1, A_2, ..., A_{d-1} equals zero.

Other values of A_i are known only for a few types of specially constructed codes. The calculation of A_i for a general type of a code takes a lot of time and involves generation of 2^{n-k} codewords. (For more details see for example [3].) The generation of the codewords is not complicated and easily parallelizable, however the time spent is still enormous even for commonly used values $n - k$ (32, 48, 64 or 96).

2.6. 'Good' and 'proper' code

Because computing of the weight structure of a code is very troublesome, there are efforts to find some more manageable method of determination of the probability of undetected error of the code.

First, it is not necessary to know a complete course of the function $P_{ud}(p_e)$; for subsequent safety considerations its maximum value is sufficient.

Second, we need not consider all possible values of the bit error probability p_e. A channel with bit error probability $p_e = 1$ exactly inverts every transmitted message. Generally, channels with bit error probability higher than $1/2$ have tendency to invert messages rather than transmit them unchanged. If a code does not contain a word with all bits equal to one, then all inversions of a codeword are detected. Therefore, for such codes it is sufficient to consider values of the p_e in the interval $[0, 1/2]$.

Introducing the boundary value $p_e = 1/2$ into formula (3) for undetected error probability in the BSC, it follows

$$P_{ud}(1/2) = \frac{2^k - 1}{2^n} < 2^{k-n}.$$

We underline that this estimate is the same for every binary linear $(n, k)-$code. Moreover, the estimate $P_{ud}(p_e) < 2^{k-n}$ it fulfilled in some small neighbourhood of $1/2$, but not on the whole interval from zero to one.

In the case that the estimate $P_{ud}(p_e) < 2^{k-n}$ is valid on the whole interval $[0, 1/2]$, we need not examine the code any more. This is a motivation for following definitions of the terms 'good' and 'proper' code:

- A binary linear (n, k)-code is *'good'*, if for all $p_e \in [0, 1/2]$ the inequality $P_{ud}(p_e) < 2^{k-n}$ is valid.
- A binary linear (n, k)-code is *'proper'*, if the function $P_{ud}(p_e)$ is monotone increasing on the interval $[0, 1/2]$.

It is evident that a 'proper' code is always 'good' as well, and after then the term 'proper' seems to be redundant. However, this term has its sense: on less erroneous channels the 'proper' code has a lower failure probability than on more erroneous ones. This is reasonable behaviour.

The second reason for introducing the term 'proper' code is that the monotonicity of the function $P_{ud}(p_e)$ can be generally proven for some classes of codes. These codes are consequently 'good' and theirs probability of undetected error in the BSC is upperbounded by the known value 2^{k-n}, which can be used in following safety calculations of the whole system.

As consequence of this, it was widely supposed among safety engineers, that all "reasonable" codes are 'proper', or at least "almost proper". In fact, codes really used in practice very often are not 'proper' and their probability of undetected error exceeds the value 2^{k-n}, often very strongly. Usually this occurs for relatively low values of the bit error rate p_e. For example, we found a code with maximal value of the probability of undetected error more than thousand times higher than 2^{k-n}. Therefore, the execution of a probabilistic analysis using BSC is necessary in all cases.

The evaluation of the maximal value of the probability of undetected error has to be done numerically. When the code is not 'proper', the most successful procedure is based on the Newton's method with adaptive precision computation. As the function $P_{ud}(p_e)$ is almost constant on the most part of its rank, this calculation is complicated and time-consuming as well. For recognition that the code is not 'proper', it is very useful to use the binomial moments (for more details see [2]).

3. Conclusion

The calculation of the maximal value of probability of undetected error is a laboured and often very lengthy process. Nevertheless, it is essential for evaluation safety parameters of the whole systems and cannot be omitted. On the other hand, safety codes contribute to the overall safety of the railway traffic only by a small part.

The reasons of most railway accidents are simple. In the past, unpredictable technical failures was frequent. Presently accidents caused by neglecting of maintenance or by failure of operator predominate. Both of them are human factor failures. As far as we know, no one railway accident caused by failure of safety code was recorded.

Recent interlocking systems pass to unified interoperable communication interfaces, usually designated for employment in open transmission systems (European systems ERTMS/ETCS, GSM-R, EURORADIO protocol). Safety codes in this systems use cryptographic techniques. These cannot be evaluated by the above mentioned method.

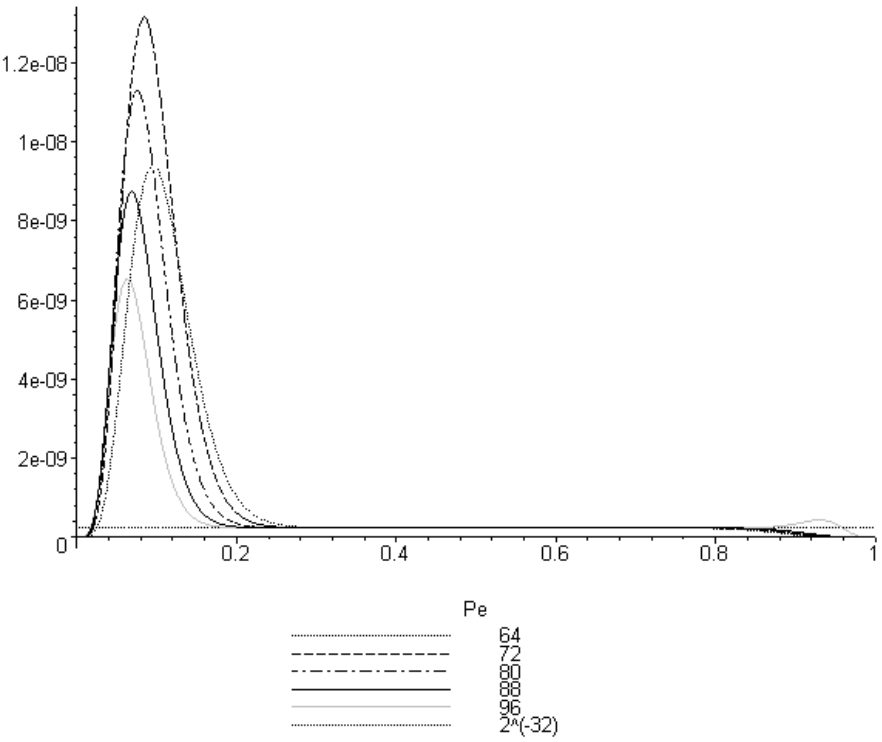


Figure 1: An example of codes, which are not 'proper'. These are five different codes, created by shortening of the same code. The codeword lengths n of these codes vary from 64 to 96 bits, number of redundant bits $k - n$ is 32 for all of them. The horizontal line near to the bottom edge of the graph is the constant 2^{-32}. The maximal value of probability of undetected error is more than 50-times higher than this value for the worst code with the codeword length 72 bits.

Another open problem is an ensuring of the independence between safety and transmission codes. Actually, there does not exist consensus even about the definition of this independence.

References

[1] EN 50159 Railway applications – Communication, signalling and processing systems – Safety-related communication in transmission systems, CENELEC, 2010.

[2] Dodunekova, R.: Extended binomial moments of a linear code and the undetected error probability. Probl. Peredachi Inf. **39**(3) (2003), 28–39. [Probl. Inf. Trans. (Engl. Transl.) **39**(3) (2003), 255–265].

[3] Huffman, W. C., Pless, V.: *Fundamentals of error-correcting codes*. Cambridge University Press, Cambridge, 2003, ISBN 0-521-78280-5.

Conference Applications of Mathematics 2013
in honor of the 70th birthday of Karel Segeth.
Jan Brandts, Sergey Korotov, Michal Křížek,
Jakub Šístek, and Tomáš Vejchodský (Eds.),
Institute of Mathematics AS CR, Prague 2013

ON SIMPLICIAL RED REFINEMENT IN THREE AND HIGHER DIMENSIONS

Sergey Korotov[1,2], Michal Křížek[3]

[1] BCAM – Basque Center for Applied Mathematics
Alameda de Mazarredo 14, E–48009 Bilbao
Basque Country, Spain
e-mail: korotov@bcamath.org
[2] IKERBASQUE, Basque Foundation for Science
E–48011, Bilbao, Spain
[3] Institute of Mathematics, Academy of Sciences
Žitná 25, CZ–115 67 Prague 1, Czech Republic
e-mail: krizek@math.cas.cz

Abstract

We show that in dimensions higher than two, the popular "red refinement" technique, commonly used for simplicial mesh refinements and adaptivity in the finite element analysis and practice, never yields subsimplices which are all acute even for an acute father element as opposed to the two-dimensional case. In the three-dimensional case we prove that there exists only one tetrahedron that can be partitioned by red refinement into eight congruent subtetrahedra that are all similar to the original one.

1. Introduction

In his speech at the International Congress of Mathematicians in Paris in 1900, David Hilbert formulated 23 open problems for the 20th century (see [22]). His 18th problem is concerned with tiling space with congruent polytopes [19]. Up to now, we do not know all space-filler polytopes.

In 1923, D. M. Y. Somerville in [21] discovered a special tetrahedral space-filler (which is now called after him the *Sommerville tetrahedron* T_1) having two opposite edges of length 2 and the other four of length $\sqrt{3}$ (see Figure 1). Thus, its mirror image is again T_1. Two of its dihedral angles at edges are right and the other four are 60°. In Theorem 1 below we show that T_1 is the only one tetrahedron up to similarity (i.e., rotation, translation, and dilatation, but no mirroring) that can be partitioned into 8 congruent subtetrahedra that are all similar to T_1 using a special technique which is called *red refinement* in the numerical analysis community. In such a partition all faces of T_1 are divided by midlines (cf. Figure 3). The tetrahedron T_1 can similarly be partitioned into $27, 64, 125, \ldots$ congruent subtetrahedra [13], but in

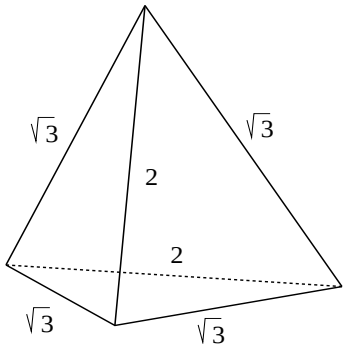


Figure 1: Sommerville tetrahedron T_1.

this work we shall only consider partitions which use the midpoints of edges (for any dimension, i.e. not only for $n = 3$).

For any $n \geq 1$ the convex hull of $n + 1$ points in $\mathbf{R}^n$ that do not lie in one hyperplane is called *n-simplex*. According to [7, p. 231], it is not known whether there exists a 4-simplex that would induce a monohedral tiling of $\mathbf{R}^4$, in general, not face-to-face. In Theorem 3 we prove that no 4-simplex has only Sommerville tetrahedral facets. In this paper we shall consider only face-to-face simplicial partitions of a given n-simplex $S \subset \mathbf{R}^n$, $n = 1, 2, \ldots$, see [3, 4].

If a domain is subdivided into congruent simplices, then we may calculate more easily entries of the stiffness matrix in the finite element method. This saves a lot of CPU time and moreover, some superconvergence phenomena can be achieved [14].

2. Red refinement

First, we will define "red refinement" of a simplex in higher dimension by a technique due to Freudenthal [9]. The term "red refinement" seems to appear first in [1] for two-dimensional triangulations. The regularity of a family of red refinements is investigated in [15] and [23].

The unit hypercube $K = [0, 1]^n$ can be partitioned into $n!$ simplices of dimension n defined as

$$S_\sigma = \{x \in \mathbf{R}^n \,|\, 0 \leq x_{\sigma(1)} \leq \cdots \leq x_{\sigma(n)} \leq 1\}, \tag{1}$$

where σ ranges over all $n!$ permutations of the numbers 1 to n.

The unit hypercube K can also be trivially partitioned into 2^n congruent sub-hypercubes. Each of the sub-hypercubes can be thus partitioned into $n!$ simplices as in (1). This will result in a face-to-face partition of K into $n!2^n$ subsimplices. All the subsimplices that are contained in the reference simplex

$$\hat{S} = \{x \in \mathbf{R}^n \,|\, 0 \leq x_1 \leq \cdots \leq x_n \leq 1\} \tag{2}$$

form a face-to-face partition which will be called to form the *red refinement* of $\hat{S}$. In this case the permutation σ is identity. The partition contains 2^n subsimplices (see Figure 2 for $n = 3$).

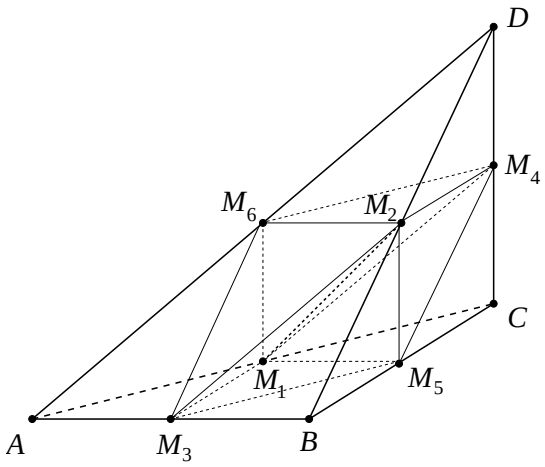


Figure 2: The red refinement of the reference simplex $\hat{S}$.

Definition 1 *Given an arbitrary n-simplex S, the reference n-simplex $\hat{S}$ can be mapped onto S by an affine transformation F. The 2^n subsimplices that form a red refinement of $\hat{S}$ are then mapped by F onto 2^n subsimplices in S, and we will call such a partition as a "red refinement" of S.*

It is clear that the above defined "red refinement" coincides with usual red refinements of triangles and tetrahedra (cf. [1, 13, 17] and Figure 3).

Remark 1 Because of possible permutations of simplex vertices, the red refinement of a given simplex is not uniquely determined except for the case $n = 1, 2$. For example, in the three-dimensional case we have 3 different possibilities how to perform a red refinement, since there are three possibilities to insert a new (interior) edge connecting the midpoints of two opposite edges (cf. [13]). To see this we denote the vertices of the reference tetrahedron $\hat{S}$ by $A = (0, 0, 0)$, $B = (1, 0, 0)$, $C = (1, 1, 0)$, and $D = (1, 1, 1)$ and let $M_1, \ldots, M_6$ be midpoints of its edges as marked in Figure 2. Now define the affine mapping $F : \hat{S} \to \hat{S}$ so that $F(A) = A$, $F(B) = C$, $F(C) = B$, and $F(D) = D$. Then the line segment M_1M_2 is mapped onto the line segment M_3M_4 yielding a different red refinement of the simplex $\hat{S}$ with the above permutation of vertices. Similarly we can define another affine transformation that maps M_1M_2 to M_5M_6.

Subsimplices that share a vertex with the original simplex are called *exterior* or *corner simplices*.

Remark 2 The $n + 1$ corner subsimplices are obviously similar to the original simplex S for any dimension n. Since F is affine, the volume of each subsimplex in the red refinement is $2^{-n}\text{vol}(S)$ and for each red refinement of S the associated refinements of its lower-dimensional facets are also red. According to [2], the red refinement algorithm produces at most $\frac{n!}{2}$ congruent classes for any initial n-simplex, no matter

how many subsequent refinements are performed (see also [23] for $n = 3$). Then the corresponding family of partitions is strongly regular in the sense of Ciarlet [6].

Remark 3 The red refinement of an arbitrary triangle produces only congruent subtriangles. However, the next theorem shows that is not true in the three-dimensional case.

Theorem 1 *There exists only one type of a tetrahedron T (up to similarity) whose red refinement produces eight congruent subtetrahedra similar to T. It is the Sommerville tetrahedron T_1.*

Proof: Let us consider such a tetrahedron T that its red refinement produces eight congruent subtetrahedra similar to T. Its faces are obviously partitioned into four congruent subtriangles. The four exterior subtetrahedra and the four interior subtetrahedra obtained by plane cuts passing through the midlines of its faces are shown in Figure 3. We show that T is similar to the Sommerville tetrahedron T_1.

Let o be the operator that assigns to a given edge of any tetrahedron its opposite edge and let us denote by a, b, c, d, e, f the edges of the front exterior subtetrahedron such that (see the lower part of Figure 3)

$$o(a) = b, \quad o(c) = d, \quad o(e) = f.$$

Parallel edges of the same length are denoted, for simplicity, by the same letters.

The exterior corner subtetrahedra are obviously similar to the original tetrahedron T. Denote by g the inner edge that is surrounded by all four interior subterahedra.

Consider the right interior and right exterior subtetrahedra. Their five edges are a, b, c, d, e. Since these two subtetrahedra are congruent, the remaining sixth edges must have the same length, i.e., $|f| = |g|$. Similarly, for the lower interior and lower exterior subtetrahedra we find that $|e| = |g|$.

Since the regular tetrahedron cannot be divided into eight congruent subtetrahedra, at least two edges of T have a different length. Without loss of generality, we may assume that $|a| \neq |e|$, since e, f, and g are in all cases opposite edges (otherwise we rename the edges a, b, c, and d).

Now consider four cases:

1. Let $|a| \notin \{|b|, |c|, |d|\}$. From the right exterior, right interior, and the lower interior subtetrahedron we see that $o(a) = b$, $o(a) = c$, and $o(a) = d$. Hence, $|b| = |c| = |d|$, since a is obviously mapped only on a during "congruence mapping". Consider the right interior subtetrahedron. If $|b| = |d| = |e| = |g|$, then the four dihedral angles at these edges have the same size. They cannot be nonacute, since any tetrahedron has at least three acute dihedral angles, see [12, p. 727]. Similarly we find that dihedral angles at g are acute for all four interior subtetrahedra, which is a contradiction. Thus, $|b| = |c| = |d| \neq |e| = |f| = |g|$, but then the right interior and

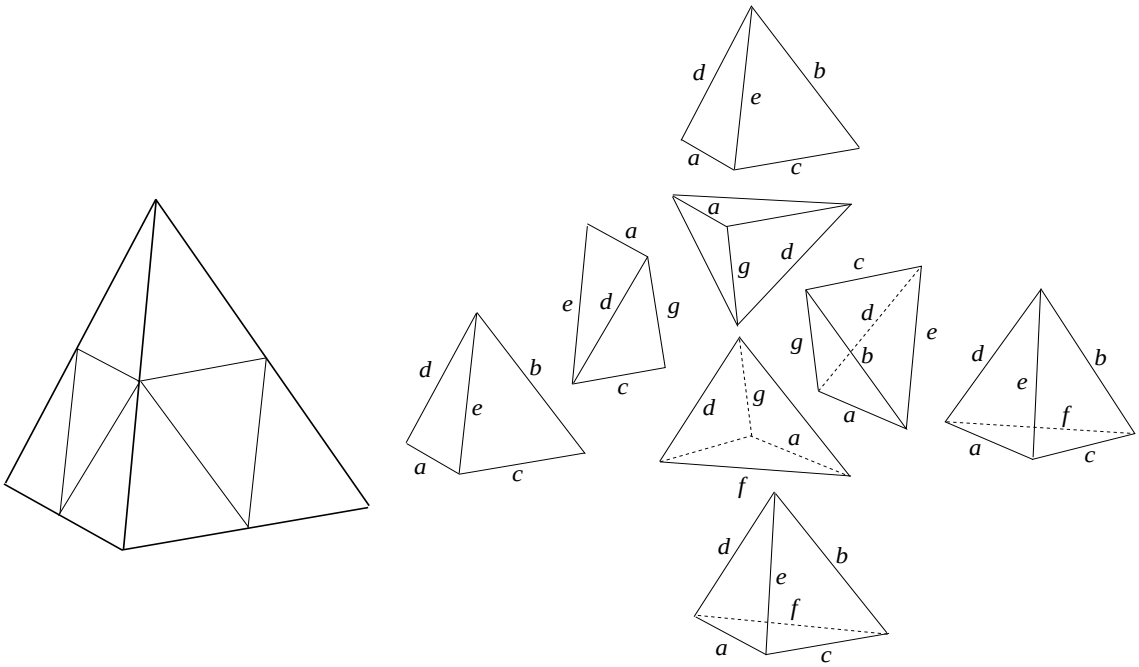


Figure 3: Red refinement of a tetrahedron T by plane cuts through midlines of its faces (left) and its exploded version (right).

right exterior subtetrahedron are not congruent (they are only indirectly congruent up to mirroring), which is a contradiction.

2. So let $|a| = |b|$. Then we easily find that $|b| = |c| = |d|$.

The cases 3. $|a| = |c|$ and 4. $|a| = |d|$ can be treated similarly. Therefore, altogether we obtain

$$|a| = |b| = |c| = |d|, \qquad |e| = |f| = |g|. \tag{3}$$

Due to the mirror image symmetry of T and its eight subtetrahedra, the edge e is perpendicular to the plane passing through the edges f and g. Similarly, the edge f is perpendicular to the plane of symmetry containing e and g. Hence, we find that (see Figure 3)

$$e \perp g \perp f \perp e.$$

Now applying the Parseval equality, we come to

$$(2|a|)^2 = |e|^2 + |g|^2 + |f|^2$$

and thus, (3) implies that

$$2|a| = \sqrt{3}|e|.$$

From this we see that T is the Sommerville tetrahedron T_1 up to similarity (cf. Figure 1) and there is no other type of a tetrahedron that can be partitioned into eight congruent subtetrahedra that are similar to the original one. ■

Red refinement of a tetrahedron that produces congruent subtetrahedra is treated also in [20]. Some authors allow mirroring of congruent tetrahedra. Zhang in [23] presents a different proof of Theorem 1. Dissection of simplices into congruent subsimplices is examined also in [10] and [18].

3. Nonobtuse red refinement

Opposite each vertex of an n-simplex lies a $(n-1)$-dimensional *facet.* For $n = 1$ facets are just points. For $n \geq 1$ the *dihedral angle* α between two facets is defined by means of the inner product of their outward unit normals ν_1 and ν_2,

$$\cos\alpha = -\nu_1 \cdot \nu_2.$$

If $n = 1$ these normals necessarily form an angle of 180° and thus $\alpha = 0$. Each simplex in $\mathbf{R}^n$ has $\binom{n+1}{2}$ dihedral angles.

Definition 2 *If all dihedral angles of a given simplex are less than* 90° *(less than or equal to* 90°*) we say that the simplex is acute (nonobtuse).*

For instance, the Sommerville tetrahedron (see Figure 1) is nonobtuse and the regular tetrahedron is acute.

Theorem 2 *If an* n*-simplex* T *for* $n \geq 2$ *is nonobtuse (acute), then any of its lower dimensional facets is also a nonobtuse (acute) simplex.*

For the proof see [8].

Definition 3 *The red refinement is said to be nonobtuse (acute) if all resulting subsimplices are nonobtuse (acute).*

Note that nonobtuse simplicial partitions lead to monotone stiffness matrices when solving elliptic problems by linear finite element methods, see e.g. [5, 11, 16].

Remark 4 We see that the inner diagonal, which is denoted by g in Figure 3 (or M_1M_2 in Figure 2), is surrounded by four tetrahedra. To get a nonobtuse red refinement, it is necessary that all dihedral angles sharing this edge are right. However, another more severe condition comes from the edges, which are denoted by e and f in Figure 3. Here the angle 180° is bisected and thus, the corresponding two dihedral angles sharing these edges have to be right. This yields a lot of restrictions on construction of nonobtuse red refinements. For instance, in the red refinement of the regular tetrahedron the dihedral angles at the edge g are all right, but one dihedral angle at edges e and f is greater than 109°. The red refinement of the (nonobtuse) cube corner terahedron with vertices $(0,0,0)$, $(1,0,0)$, $(0,1,0)$, and $(0,0,1)$, produces angles greater than 125° at e and f.

On the other hand, the red refinement of the path simplex yields only path subsimplices in any dimension $n \geq 2$ (cf. Figure 2). The path simplex in its basic

position can be stretched or shrinked along any coordinate axis x_i and we still get nonobtuse red refinement. If $n = 3$ then there are six path subtetrahedra T that are congruent with the original path tetrahedron. The remaining two are mirror images of T (see Figure 2 and Remark 2). The red refinement of the Sommerville tetrahedron also produces nonobtuse tetrahedra which follows from Theorem 1. This is due to the fact that the Sommerville tetrahedron is the union of 4 path subtetrahedra. In [12] we introduced the so-called yellow refinement which produces only nonobtuse subtetrahedra provided the initial tetrahedron is nonobtuse and contains the centre of its circumscribed ball.

Remark 5 Consider now a red refinement of a 4-simplex S, i.e., it is partitioned into 16 subsimplices. Then we get a situation which is a little bit difficult to imagine. Namely, we first cut off 5 congruent corner subsimplices that are similar to S. The remaining polytopic domain then has 10 three-dimensional facets and it is partitioned into $16 - 5 = 11$ subsimplices.

Theorem 3 *There is no 4-simplex whose three-dimensional facets are all Sommerville tetrahedra.*

Proof: From the well-known Euler-Poincaré formula we find that a 4-simplex has 5 vertices, 10 edges, 10 triangular faces, and there are 5 tetrahedral three-dimensional facets.

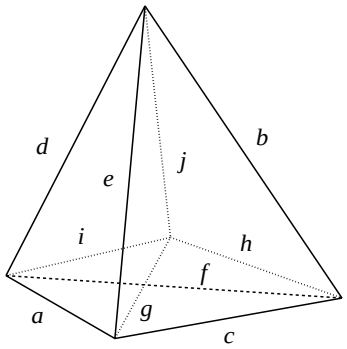


Figure 4: Schematic illustration of a 4-simplex and notation of its edges.

Now we show that there is no 4-simplex whose five facets are all the Sommerville tetrahedra T_1. Suppose to the contrary that such 4-simplex S exists. Denote its 10 edges by $a, b, c, d, e, f, g, h, i, j$ as indicated in Figure 4. Let one of its facets be the Sommerville tetrahedron T_1. Without lost of generality we may assume that its edges satisfy $|a| = |b| = |c| = |d| = \sqrt{3}$ and $|e| = |f| = 2$. Since e is opposite to h and i; and f is opposite to g and j, we get

$$|g| = |h| = |i| = |j| = 2.$$

However, this relation does not allow that all five facets are the Sommerville tetrahedra T_1, since the edges g, h, i, j contain a common point and thus their pairs are not opposite. This is a contradiction. ■

Theorem 4 *The red refinement of an acute simplex for $n > 2$ never yields subsimplices that would be all mutually congruent.*

Proof: Assume, on the contrary, that there exists an acute simplex whose red refinement produces mutually congruent subsimplicies, which should be then, obviously, acute as the exterior subsimplices are always similar to the father simplex. As the red refinement of the simplex implies by induction the red refinement of all its lower-dimensional facets (cf. Remark 2), any of its three-dimensional facets would be partitioned as in Figure 3. But then some nonacute angles between lower-dimensional faces appear, since the inner edge g is surrounded by four tetrahedra. This contradicts by Theorem 2 to the assumption that all subsimplicies are acute. ■

Remark 6 In fact, from the above proof we observe even a stronger result than the one stated in Theorem 4. The red refinement of n-simplex never produces only acute subsimplices for $n > 2$.

Acknowledgements

The first author was supported by Grant MTM2011–24766 of the MICINN, Spain. The second author was supported by the Institutional Research Plan RVO 67985840. Both authors are thankful to Jan Brandts for fruitful discussions.

References

[1] Bank, R. E., Sherman, A. H., and Weiser, A.: Refinement algorithms and data structures for regular local mesh refinement. In: R. Stepleman (Ed.), *Scientific Computing*, pp. 3–17. IMACS/North Holland, Amsterdam, 1983.

[2] Bey, J.: Simplicial grid refinement: on Freudenthal's algorithm and the optimal number of congruence classes. Numer. Math. **85** (2000), 1–29.

[3] Brandts, J., Korotov, S., and Křížek, M.: The strengthened Cauchy-Bunyakowski-Schwarz inequality for n-simplicial linear finite elements. In: Z. Li et al. (Eds.), *Proc. Third Internat. Conf. Numer. Anal. Appl. (NAA–2004), Rousse, Bulgaria*, pp. 203–210. Springer, 2005.

[4] Brandts, J., Korotov, S., and Křížek, M.: Simplicial finite elements in higher dimensions. Appl. Math. **52** (2007), 251–265.

[5] Brandts, J., Korotov, S., and Křížek, M.: The discrete maximum principle for linear simplicial finite element approximations of a reaction-diffusion problem. Linear Algebra Appl. **429** (2008), 2344–2357.

[6] Ciarlet, P. G.: *The finite element method for elliptic problems.* North-Holland, Amsterdam, 1978.

[7] Debrunner, H. E.: Tiling Euclidean d-space with congruent simplexes. Ann. N. Y. Acad. Sci. **440** (1985), 230–261.

[8] Fiedler, M.: Über qualitative Winkeleigenschaften der Simplexe. Czechoslovak Math. J. **7** (1957), 463–476.

[9] Freudenthal, H.: Simplizialzerlegungen von bescharaenkter. Flachheit. Ann. of Math. in Sci. and Engrg. **43** (1942), 580–582.

[10] Hertel, E.: Self-similar simplices. Beiträge Algebra Geom. **41** (2000), 589–595.

[11] 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.

[12] Korotov, S. and Křížek, M.: Acute type refinements of tetrahedral partitions of polyhedral domains. SIAM J. Numer. Anal. **39** (2001), 724–733.

[13] Křížek, M.: An equilibrium finite element method in three-dimensional elasticity. Apl. Mat. **27** (1982), 46–75.

[14] Křížek, M.: Superconvergence phenomena on three-dimensional meshes. Internat. J. Numer. Anal. Model. **2** (2005), 43–56.

[15] Křížek, M. and Strouboulis, T.: How to generate local refinements of unstructured tetrahedral meshes satisfying a regularity ball condition. Numer. Methods Partial Differential Equations **13** (1997), 201–214.

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

[17] Liu, A. and Joe, B.: Quality local refinement of tetrahedral meshes based on 8-subtetrahedron subdivision. Math. Comp. **65** (1996), 1183–1200.

[18] Matoušek, J. and Safernová, Z.: On the nonexistence of k-reptile tetrahedra. Discrete Comput. Geom. **46** (2011), 599–609.

[19] Milnor, J.: Hilbert's problem 18: On crystalographic groups, fundamental domains, and on sphere packing. In: F. E. Browder (Ed.), *Mathematical Developments Arising from Hilbert Problems, Part 2*, pp. 491–506. AMS Providence, Rhode Island, 1976.

[20] Moore, D. and Warren, J.: Adaptive simplicial mesh quadtrees. Houston J. Math. **21** (1995), 525–540.

[21] Sommerville, D. M. Y.: Space-filling tetrahedra in Euclidean space. Proc. Edinb. Math. Soc. **41** (1923), 49–57.

[22] Więsław, W. (Ed.): *Problemy Hilberta* (in Polish). Inst. Hist. Nauk PAN, Warsaw, 1997.

[23] Zhang, S.: Successive subdivisions of tetrahedra and multigrid methods on tetrahedral meshes. Houston J. Math. **21** (1995), 541–556.

Conference Applications of Mathematics 2013
in honor of the 70th birthday of Karel Segeth.
Jan Brandts, Sergey Korotov, Michal Křížek,
Jakub Šístek, and Tomáš Vejchodský (Eds.),
Institute of Mathematics AS CR, Prague 2013

A PARALLEL METHOD FOR POPULATION BALANCE EQUATIONS BASED ON THE METHOD OF CHARACTERISTICS

Yu Li[1], Qun Lin[1], Hehu Xie[2]

[1] LSEC, Institute of Computational Mathematics
Chinese Academy of Sciences, Beijing 100190, China
liyu@lsec.cc.ac.cn, linq@lsec.cc.ac.cn
[2] LSEC, NCMIS, Institute of Computational Mathematics
Chinese Academy of Sciences, Beijing 100190, China
hhxie@lsec.cc.ac.cn

Abstract

In this paper, we present a parallel scheme to solve the population balance equations based on the method of characteristics and the finite element discretization. The application of the method of characteristics transform the higher dimensional population balance equation into a series of lower dimensional convection-diffusion-reaction equations which can be solved in a parallel way. Some numerical results are presented to show the accuracy and efficiency.

1. Introduction

In this paper, we propose a parallel scheme to solve the population balance equation (PBE) based on the application of the method of characteristics and the finite element method. The PBEs aries from the model of the industrial crystallization process (see, e.g., [7, 11, 12] and the references cited therein). Recently, more and more researchers are interested in the numerical methods for PBEs (c.f. [1, 5, 6, 7]). In PBEs, besides the normal space and time variables, the distribution of entities also depends on their own properties which are referred to as internal coordinates. It is a high dimensional system of equations which is a big challenge from the computational point of view. In order to overcome this difficulty, we use the method of characteristics (c.f. [2, 4]) to transfer the original problem to a series of lower-dimensional convection-diffusion-reaction problems which are defined on the characteristics curves and the spatial directions. Based on the data structure for the method of characteristics, a parallel implementation can be applied to do the simulation process that can improve the computational efficiency.

So far, there exists the alternating direction (operator splitting) method for the PBE by decomposing the original problem into two unsteady subproblems of smaller

complexity (see, e.g., [1, 5, 6]). In the two subproblems, the ordering of the data for the solution needs to be different, since they are discretized in different direction (c.f. [1]). It is not so suitable for the parallel implementation and prevents the further improvement of the computation efficiency for the PBE.

In the present paper, we use the method of characteristics to transform the PBE into a series of convection-diffusion-reaction equations on the characteristic curves in each time step. Then the finite element method is applied to solve the series of convection-diffusion-reaction problems. Furthermore, based on the data structure of the numerical scheme, a parallel scheme is constructed to solve the PBE based on the distributed memory. Some numerical results are provided to check the efficiency of this parallel method.

The rest of the paper will go as follows: Section 2 introduces the model problem under consideration and defines some notation. In Section 3, we describe the method of characteristics for solving the PBE. The finite element discretization for the PBE is described in Section 4. Then Section 5 gives the parallel implementation way for the full discrete form of the PBE. The numerical results are given in Section 6 to validate the efficiency of the numerical method proposed in this paper. Some concluding remarks are given in the last section.

2. Model problem

Let $\Omega_{\mathbf{x}}$ be a simply connected domain in $\mathcal{R}^d$ ($d = 2$ or 3) with Lipschitz continuous boundary $\partial\Omega_{\mathbf{x}}$, $\Omega_\ell = [\ell_{\min}, \ell_{\max}] \subset \mathcal{R}$, and $T > 0$. The state of the individual particle in the PBE equation may consists of the external coordinate $\mathbf{x}$ ($\mathbf{x} = (x_1, \ldots, x_d)$), denoting its position in the physical space, and the internal coordinate ℓ, representing the properties of particles, such as size, volume, temperature etc. A PBE for a solid process such as crystallization with one internal coordinate can be described by the following partial differential equation:

Find $z : (0, T] \times \Omega_\ell \times \Omega_{\mathbf{x}} \to \mathcal{R}$ such that

$$\begin{cases} \partial z/\partial t + G(\ell)\partial_\ell z - \varepsilon\Delta_{\mathbf{x}} z + \mathbf{b}(\mathbf{x}) \cdot \nabla_{\mathbf{x}} z = f(t, \ell, \mathbf{x}) & \text{in } (0, T] \times \Omega_\ell \times \Omega_{\mathbf{x}}, \\ z(0, \ell, \mathbf{x}) = z_{\text{init}}(\ell, \mathbf{x}) & \text{in } \Omega_\ell \times \Omega_{\mathbf{x}}, \\ z(t, \ell_{\min}, \mathbf{x}) = z_{\text{bdry}}(t, \mathbf{x}) & \text{on } (0, T] \times \Omega_{\mathbf{x}}, \\ z(t, \ell, \mathbf{x}) = 0 & \text{on } (0, T] \times \Omega_\ell \times \partial\Omega_{\mathbf{x}}, \end{cases} \tag{1}$$

where the diffusion coefficient $\varepsilon > 0$ is a given constant, $\Delta_{\mathbf{x}}$ and $\nabla_{\mathbf{x}}$ denote the Laplacian and gradient with respect to $\mathbf{x}$, respectively, $\mathbf{b}$ is a given velocity and satisfies $\nabla_{\mathbf{x}} \cdot \mathbf{b} = 0$, and f is a source function. Here $G(\ell) > 0$ represents the growth rate of the particles that depends on ℓ but is independent of $\mathbf{x}$ and t. Furthermore, let us assume the data $G(\ell)$, $\mathbf{b}$, f, z_{init} and z_{bdry} are sufficiently smooth functions for our error estimate analysis.

Now we introduce some notation of the function spaces (see [2, 3]). Let $H^m(\Omega_{\mathbf{x}})$ denote the standard Sobolev space of functions with derivatives up to m in $L^2(\Omega_{\mathbf{x}})$ and the norm is defined by

$$\|v\|_{H^m(\Omega_{\mathbf{x}})} = \left(\int_{\Omega_{\mathbf{x}}} \sum_{0 \le |\alpha| \le m}^{m} \left| \frac{\partial^\alpha v}{\partial \mathbf{x}^\alpha} \right|^2 d\mathbf{x} \right)^{1/2},$$

where α denote a non-negative multi-index $\alpha = \{\alpha_1, \ldots, \alpha_d\}$, $|\alpha| = \sum_{1 \le j \le d} \alpha_j$, and

$$\frac{\partial^\alpha v}{\partial^\alpha \mathbf{x}} = \frac{\partial^{\alpha_1 \ldots \alpha_d} v}{\partial x_1^{\alpha_1} \ldots x_d^{\alpha_d}}.$$

We use $(\cdot, \cdot)_{\mathbf{x}}$ and $\|\cdot\|_{L^2(\Omega_{\mathbf{x}})}$ to denote the L^2-inner product and the associated norm in $\Omega_{\mathbf{x}}$, respectively, which are defined as follows

$$(v, w)_{\mathbf{x}} = \int_{\Omega_{\mathbf{x}}} vw d\mathbf{x} \quad \text{and} \quad \|v\|^2_{L^2(\Omega_{\mathbf{x}})} = (v, v)_{\mathbf{x}}.$$

Let X be a Banach space with the norm $\|\cdot\|_X$. Then we define

$$\begin{aligned}
C(\Omega_\ell; X) &= \Big\{ v : \Omega_\ell \to X : \ v \text{ is continuous} \Big\}, \\
W^{m,\infty}(\Omega_\ell; X) &= \Big\{ v : \Omega_\ell \to X : \ \Big\| \frac{\partial^j v}{\partial \ell^j} \Big\|_X < \infty, \ 0 \le j \le m \Big\}, \\
W^{m,\infty}((0,T]; X) &= \Big\{ v : (0,T] \to X : \ \Big\| \frac{\partial^j v}{\partial t^j} \Big\|_X < \infty, \ 0 \le j \le m \Big\},
\end{aligned}$$

where the derivatives $\partial^j v / \partial \ell^j$ and $\partial^j v / \partial t^j$ are understood in the sense of distributions on Ω_ℓ and $(0, T]$, respectively. The norms in the above defined spaces are given as follows

$$\begin{aligned}
\|v\|_{C(\Omega_\ell; X)} &= \sup_{\ell \in \Omega_\ell} \|v(\ell)\|_X, \\
\|v\|_{W^{m,\infty}(\Omega_\ell; X)} &= \max_{0 \le j \le m} \sup_{\ell \in \Omega_\ell} \Big\| \frac{\partial^j v}{\partial \ell^j} \Big\|_X, \\
\|v\|_{W^{m,\infty}((0,T]; X)} &= \max_{0 \le j \le m} \sup_{t \in (0,T]} \Big\| \frac{\partial^j v}{\partial t^j} \Big\|_X.
\end{aligned}$$

For spaces X, Y and Z, we use the short notation $Z(Y(X)) := Z((0, T]; (Y(\Omega_\ell; X))$ in this paper.

3. Method of characteristics

In this section, we describe the method of characteristics (c.f. [2, 4, 9]) for the PBE (1). The reason we adopt this method for the discretization in the product space

$(0,T]\times\Omega_\ell$ is that it has the suitable data structure for the parallel implementation which will be discussed in the following sections.

First we set

$$\psi(t,\ell) \;=\; (1+G(\ell)^2)^{1/2}.$$

Let the characteristic direction associated with the hyperbolic part of (1), $\partial z/\partial t + G(\ell)\partial z/\partial \ell$, be denoted by $s(t)$. Then

$$\frac{\partial}{\partial s} \;=\; \frac{1}{\psi}\frac{\partial}{\partial t} + \frac{G(\ell)}{\psi}\frac{\partial}{\partial \ell}. \tag{2}$$

Then (1) can be written as

$$\begin{cases} \psi\partial z/\partial s - \varepsilon\Delta_{\mathbf{x}} z + \mathbf{b}(\mathbf{x})\cdot\nabla_{\mathbf{x}} z = f & \text{in } (0,T]\times\Omega_\ell\times\Omega_{\mathbf{x}},\\ z(0,\ell,\mathbf{x}) = z_{\text{init}}(\ell,\mathbf{x}) & \text{in } \Omega_\ell\times\Omega_{\mathbf{x}},\\ z(t,\ell_{\min},\mathbf{x}) = z_{\text{bdry}}(t,\mathbf{x}) & \text{on } (0,T]\times\Omega_{\mathbf{x}},\\ z(t,\ell,\mathbf{x}) = 0 & \text{on } (0,T]\times\Omega_\ell\times\partial\Omega_{\mathbf{x}}. \end{cases} \tag{3}$$

We use uniform partitions for the time interval $(0,T]$ and the internal coordinate interval Ω_ℓ, respectively. Let $\tau = T/N$, $\iota = (\ell_{\max}-\ell_{\min})/M$, $t^n = n\tau$, $n = 0,1,\ldots,N$ and $\ell_m = \ell_{\min} + m\iota$, $m = 0,1,\ldots,M$. In order to satisfy the stability condition, we set

$$\tau \le \frac{\iota}{\max_{\ell_{\min}\le\ell\le\ell_{\max}} G(\ell)}. \tag{4}$$

Then starting with $z(0,\ell,\mathbf{x}) = z_{\text{init}}$, $z(t,\ell_{\min},\mathbf{x}) = z_{\text{bdry}}(t,\mathbf{x})$, the equation (3) can be discreted in each sub-intervals $(t^{n-1},t^n]\times(\ell_{m-1},\ell_m]\times\Omega_{\mathbf{x}}$ ($n = 1,2,\ldots,N$ and $m = 1,2,\ldots,M$) as follows.

First we compute

$$\check{\ell}_m \;=\; \ell_m - \tau G(\ell_m). \tag{5}$$

Actually, this is a first order discretization to obtain the approximation at the time level $t = t^{n-1}$ for the following characteristic ordinary differential equation (c.f. [4]):

$$\begin{cases} d\ell/dt \;=\; G(\ell) & \text{in } [t^{n-1},t^n),\\ \ell(t^n) \;=\; \ell_m. \end{cases} \tag{6}$$

From the condition (4), we have $\check{\ell}_m \ge \ell_{\min}$ for $m \ge 1$. Then we compute the direction differential $\psi\frac{\partial z}{\partial s}$ at the node (t^n,ℓ_m) in the following way

$$\begin{aligned} \psi(t^n,\ell_m)\frac{\partial z}{\partial s}(t^n,\ell_m,\mathbf{x}) \;&\approx\; \psi(t^n,\ell_m)\frac{z(t^n,\ell_m,\mathbf{x}) - \check{z}(t^{n-1},\check{\ell}_m,\mathbf{x})}{(\tau^2+(\ell_m-\check{\ell}_m)^2)^{1/2}}\\ &=\; \frac{z(t^n,\ell_m,\mathbf{x}) - \check{z}(t^{n-1},\check{\ell}_m,\mathbf{x})}{\tau}, \end{aligned} \tag{7}$$

where $\check{z}\left(t^{n-1},\check{\ell}_m,\mathbf{x}\right) := \alpha_m^n\, z\left(t^{n-1},\ell_{m-1},\mathbf{x}\right) + \left(1-\alpha_m^n\right) z\left(t^{n-1},\ell_m,\mathbf{x}\right)$ with $\alpha_m^n = (\ell_m - \check{\ell}_m)/\iota$.

In order to give the semi-discrete form of the PBE, we set $z_m^n(\mathbf{x}) \approx z(t^n,\ell_m,\mathbf{x})$. Then the semi-discrete form of the PBE can be defined as follows:

$$\begin{cases} \frac{z_m^n(\mathbf{x})-\check{z}_m^n(\mathbf{x})}{\tau} - \varepsilon\Delta_{\mathbf{x}} z_m^n(\mathbf{x}) + \mathbf{b}(\mathbf{x})\nabla_{\mathbf{x}} z_m^n(\mathbf{x}) = f_m^n(\mathbf{x}) & \text{in } \Omega_{\mathbf{x}}, \\ z_m^0(\mathbf{x}) = z_{\text{init}}^m(\mathbf{x}) & \text{for } x \in \Omega_{\mathbf{x}}, \\ z_0^n(\mathbf{x}) = z_{\text{bdry}}(t^n,\mathbf{x}) & \text{for } (0,T]\times\Omega_{\mathbf{x}}, \\ z_m^n(\mathbf{x}) = 0 \text{ for } m = 1,2,\ldots,M & \text{on } \partial\Omega_{\mathbf{x}}\,, \end{cases} \tag{8}$$

where $f_m^n(\mathbf{x}) = f(t^n,\ell_m,\mathbf{x})$, $\check{z}_m^n(\mathbf{x}) = \alpha_m^n z_{m-1}^{n-1}(\mathbf{x}) + (1-\alpha_m^n) z_m^{n-1}(\mathbf{x})$.

From the Taylor expansion method, we can derive the following error estimate for the semi-discrete form (8)

$$\|z(t^n,\ell_m,\mathbf{x}) - z_m^n(\mathbf{x})\|_{C(X)} \;\le\; C\tau \|z(t,\ell,\mathbf{x})\|_{W^{2,\infty}(W^{1,\infty}(X))}, \tag{9}$$

where the space X can be $L^2(\Omega_{\mathbf{x}})$ or $H^1(\Omega_{\mathbf{x}})$.

4. Finite element method

In this section, we give the fully discrete form of the PBE by the finite element method. Let V_h be a finite element subspace of $H_0^1(\Omega_{\mathbf{x}})$ which has the k-th order of accuracy (c.f. [2, 3]):

$$\inf_{v_h\in V_h} \|u - v_h\|_{H^1(\Omega_{\mathbf{x}})} \le Ch^k \|u\|_{H^{m+1}(\Omega_{\mathbf{x}})} \quad \forall u \in H^{m+1}(\Omega_{\mathbf{x}}). \tag{10}$$

and

$$\inf_{v_h\in V_h} \|u - v_h\|_{L^2(\Omega_{\mathbf{x}})} \le Ch^{k+1} \|u\|_{H^{m+1}(\Omega_{\mathbf{x}})} \quad \forall u \in H^{m+1}(\Omega_{\mathbf{x}}). \tag{11}$$

Based on the finite element space V_h, we can define the fully discrete form for the PBE as follows:

For the n-th time step $t = t^n$ and $m = 0,1,\ldots,M$, find $z_{m,h}^n \in V_h$ such that

$$\begin{cases} \left(\frac{z_{m,h}^n - \check{z}_{m,h}^n}{\tau}, v_h\right) + a(z_{m,h}^n, v_h) = (f_m^n(\mathbf{x}), v_h) & \forall v_h \in V_h, \\ a_0(z_{m,h}^0, v_h) = a_0(z_{\text{init}}(\ell_m,\mathbf{x}), v_h) & \forall v_h \in V_h,\ m = 1,\ldots,M, \\ a_0(z_{0,h}^n, v_h) = a_0(z_{\text{bdry}}(t^n,\mathbf{x}), v_h) & \forall v_h \in V_h, \end{cases} \tag{12}$$

where $\check{z}_{m,h}^n = \alpha_m^n z_{m-1,h}^{n-1} + (1-\alpha_m^n) z_{m,h}^{n-1}$ with α_m^n being defined in Section 3 and

$$\begin{aligned} a(u,v) &= \int_{\Omega_{\mathbf{x}}} \left(\varepsilon\nabla u\cdot\nabla v + \mathbf{b}(\mathbf{x})\cdot\nabla u\; v\right) d\mathbf{x}, \\ a_0(u,v) &= \int_{\Omega_{\mathbf{x}}} \nabla u\cdot\nabla v d\mathbf{x}. \end{aligned}$$

From the standard error estimate theory of the finite element method (c.f. [2, 3]), the fully discrete form (12) has the following error estimates

$$\max_{1\le m\le M}\|z(T,\ell_m,\mathbf{x})-z_{m,h}^N\|_{H^1(\Omega_{\mathbf{x}})}\le C(\tau+h^k)\|z\|_{W^{2,\infty}(W^{1,\infty}(H^{k+1}(\Omega_{\mathbf{x}})))} \tag{13}$$

and

$$\max_{1\le m\le M}\|z(T,\ell_m,\mathbf{x})-z_{m,h}^N\|_{L^2(\Omega_{\mathbf{x}})}\le C(\tau+h^{k+1})\|z\|_{W^{2,\infty}(W^{1,\infty}(H^{k+1}(\Omega_{\mathbf{x}})))}. \tag{14}$$

5. A parallel way

In this section, we present a parallel scheme to solve the PBE (1) based on the full discrete (12). Fortunately, from (12), we can find that the finite element equation is independent for each m in any time step t^n. Based on this property, we can construct a type of parallel scheme to implement the full discretization of the fully discrete form (12).

Assume we use P processors to compute the PBE. Decompose the set $\{0,1,2,\ldots,M\}$ into P subsets $\mathbf{m}_1,\mathbf{m}_2,\ldots,\mathbf{m}_P$ such that $\mathbf{m}_1=\{0,1,\ldots,m_1-1\}$, $\mathbf{m}_p=\{m_{p-1},m_{p-1}+1,\ldots,m_p-1\}$ $(p=2,\ldots,P-1)$ and $\mathbf{m}_P=\{m_{P-1},\ldots,m_P-1\}$. In the p-th processor, the equation (12) is solved on the sub-intervals $(t^{n-1},t^n]\times(\ell_{m_{p-1}},\ell_{m_p-1}]\times\Omega_{\mathbf{x}}$ $(n=1,2,\ldots,N,\ p=1,2,\ldots,P,\ \ell_0=\ell_{\min}$ and $\ell_M=\ell_{\max})$. Because the growth rate of the particles $G(\ell)$ is positive, the dependence of each point ℓ_m is on the left $(\ell<\ell_m)$. This means that the solution $z_{m_p-1,h}^{n-1}$ in the p-th processor as the initial condition for the $(p+1)$-th processor computing at the time step t^n.

We allocate the memory in the p-th processor $(p=1,\ldots,P)$ to save the solutions $z_{m_{p-1},h}^n,\ldots,z_{m_p-1,h}^n$ and the p-th processor $(p=1,\ldots,P-1)$ should send its saved solutions to the next $(p+1)$-th processor after each time step computation. Obviously, for $p=1$, we need to use the boundary condition $z_{\mathrm{bdry}}(t,\mathbf{x})$. Similarly for $p=P$, the sending of solutions is not required since it is the last processor. Based on this distribution of the memory and the computation of the scheme (12), we can construct the following parallel algorithm for the PBE.

Algorithm 5.1. *Parallel algorithm for PBE*

For $n=1,2,\ldots,N$ do

1. *On each processor, compute the solution $z_{m,h}^n$ for $m\in\mathbf{m}_p$ $(p=1,2,\ldots,P)$ in sub-interval $(t^{n-1},t^n]\times(\ell_{m_{p-1}},\ell_{m_p-1}]$.*
2. *For $p=1,2,\ldots,P-1$, send the solutions obtained in the p-th processor $z_{m,h}^n$ $(m\in\mathbf{m}_p)$ to the $(p+1)$-th processor.*
3. *If $n<N$, set $n:=n+1$ and go to Step 1. Else stop.*

6. Numerical results

In this section, we provide some numerical results to validate the numerical scheme proposed in this paper. Let $\Omega_{\mathbf{x}} = [0,1] \times [0,1]$, $\Omega_\ell = [0,1]$, $T = 1$, $\varepsilon = 1$, and $\mathbf{b}(\mathbf{x}) = (1,1)^T$. We chose the functions $f(t,\ell,\mathbf{x})$, $z_{\text{init}}(\ell,\mathbf{x})$ and $z_{\text{bdry}}(t,\mathbf{x})$ such that the exact solution is

$$z(t,\ell,x,y) = \mathrm{e}^{-at}\sin(\pi\ell)\sin(\pi x)\sin(\pi y)$$

with $a = 0.1$. The growth rate of the particles is $G(\ell) = \frac{1}{2} + 2(1-\ell)\ell$.

First, we check the convergence order for the error estimates

$$\|e\|_0 = \max_{1\le m\le M} \|z(T,\ell_m,\mathbf{x}) - z^n_{m,h}\|_{L^2(\Omega_{\mathbf{x}})} \tag{15}$$

and

$$\|e\|_1 = \max_{1\le m\le M} \|z(T,\ell_m,\mathbf{x}) - z^n_{m,h}\|_{H^1(\Omega_{\mathbf{x}})}. \tag{16}$$

The convergence order of the linear and quadratic finite element method for the discretization in $\Omega_{\mathbf{x}}$ is shown in Tables 1 and 2. We see that the experimental results of convergence approach to the theoretically predicated values both for linear and quadratic elements.

mesh size h	$\|e\|_0$		$\|e\|_1$	
	error	order	error	order
2^{-2}	4.5702E-01		2.6897E-00	
2^{-3}	1.4872E-01	1.6197	1.5128E-00	0.8302
2^{-4}	4.0481E-02	1.8773	7.8083E-01	0.9541
2^{-5}	1.0318E-02	1.9721	3.9359E-01	0.9883
2^{-6}	2.7230E-03	1.9219	1.9720E-01	0.9970

Table 1: Errors (15) and (16) and the corresponding rates of convergence for linear element with $\tau = \iota = h^2$.

mesh size h	$\|e\|_0$		$\|e\|_1$	
	error	order	error	order
2^{-1}	6.0137E-01		2.5073E-00	
2^{-2}	6.3958E-02	3.2331	8.5316E-01	1.5552
2^{-3}	7.4660E-03	3.0987	2.3528E-01	1.8584
2^{-4}	9.5200E-04	2.9713	6.0522E-02	1.9588

Table 2: Errors (15) and (16) and the corresponding rates of convergence for quadratic element with $\tau = \iota = h^3$.

size of internal coordinate ι	$\|e\|_0$ error	order	$\|e\|_1$ error	order
2^{-2}	6.3862E-01		2.8423E-00	
2^{-3}	3.4562E-01	0.8858	1.5382E-00	0.8858
2^{-4}	1.7650E-01	0.9695	7.8427E-01	0.9718
2^{-5}	8.8689E-02	0.9928	3.9404E-01	0.9930
2^{-6}	4.4398E-02	0.9983	1.9726E-01	0.9980

Table 3: Errors (15) and (16) and the corresponding rates of convergence in the internal coordinate for the quadratic element with $h = \iota$ and $\tau = \iota^2$.

number of processors	8	16	32	64	128
time (in seconds)	28103.01	13555.03	6832.26	3708.71	1840.43
rate of speed up	1.00	2.07	4.11	7.57	15.26

Table 4: Strong parallel test with linear element $h = 1/256$, $\tau = 1/512$ and $\iota = 1/512$.

number in ℓ	1	2	4	8	16
8	9.30	15.30	27.51	55.28	116.42
16	9.91	15.44	28.44	59.44	117.24
32	9.85	16.98	32.02	60.93	118.89
64	10.01	17.28	32.66	63.88	121.96
128	10.21	17.98	33.55	64.27	127.63

Table 5: Weak parallel test with linear element $h = 1/256$: average time in seconds.

number in ℓ	1	2	4	8	16
8	11.19	16.10	27.60	60.52	120.26
16	11.26	16.43	31.54	61.36	120.83
32	12.73	18.50	35.29	68.18	131.98
64	11.20	19.63	36.39	75.43	133.55
128	12.86	20.28	38.01	73.63	146.01

Table 6: Weak parallel test with linear element $h = 1/256$: maximum time in seconds.

We also check the convergence order for the method of characteristics developed in Section 3. The corresponding numerical result are provided in Table 3. From this table, we can find the convergence order is 1 which is the same as in (9).

Now we come to check the efficiency of the parallel scheme of Algorithm 5.1. For this aim, we set the discretization parameters $h = 1/256$, $\tau = \iota = 1/512$ and use the linear finite element method. The run-time (in seconds) is shown in Table 4. From Table 4, we can find that the parallel Algorithm 5.1 has a good expansibility.

We also check the run-time in each processor for different scale in each processor. For each test, we run 8 time steps ($N = 8$). Tables 5 and 6 show the corresponding run-time (in seconds) for the average time and maximum time, respectively, for all the processors. These two tables also show that Algorithm 5.1 has good parallel properties.

7. Concluding remarks

In this paper, we are concerned with the parallel numerical method for the PBEs with one internal coordinate posed on the domain $(0, T] \times \Omega_\ell \times \Omega_{\mathbf{x}}$ with the dimension $1 + 1 + d$. The parallel scheme is based on the method of characteristics and the finite element discretization. Some numerical results are also provided in Section 6 to demonstrate the efficiency of the proposed method.

Here, for the simplicity of the description of the numerical method, we assume the diffusion coefficient ε to be large enough such that the diffusion is dominated. For the convection dominated case (c.f. [1, 10, 13]), we will combine the method of characteristics and the stabilized finite element methods (c.f. [1, 2, 13, 10]) and this is our future work. Furthermore, the parallel method should also be applied to the simulation of the industrial crystallization process (c.f. [11, 12]) and other similar models (c.f. [7]).

Acknowledgements

This work is supported in part by the National Science Foundations of China (NSFC 11001259, 2011CB309703 and 2010DFR00700) and Croucher Foundation of Hong Kong Baptist University, the national Center for Mathematics and Interdisciplinary Science, CAS, the President Foundation of AMSS-CAS. The second author gratefully acknowledges the support from the MBF-Project SimParTurS under the grant 03TOPAA1 and the Institute for Analysis and Computational Mathematics, Otto-von-Guericke-University Magdeburg.

References

[1] Ahmed, N., Matthies, G., and Tobiska, L.: Finite element methods of an operator splitting applied to population balance equations. J. Comput. Appl. Math. **236** (2011), 1604–1621.

[2] Chen, Z.: *Finite element methods and their applications*. Springer, 2005.

[3] Ciarlet, P.: *The finite element method for elliptic problems*. North-Holland Amsterdam, 1978.

[4] Evanns, L.: *Partial differential equations*. American Mathematical Society, 1998.

[5] Ganesan, S.: Population balance equations, streamline-upwind Petrov-Galerkin finite element methods, operator-splitting method, backward Euler scheme, error analysis, Preprint 1531, WIAS, Berlin, 2010.

[6] Ganesan, S. and Tobiska, L.: Implementation of an operator splitting finite element method for high-dimensional parabolic problems. Preprint 11-04, Fakultät für Mathematik, Otto-von-Guericke-Universität Magdeburg, 2010.

[7] John, V., Roland, M., Mitkova, T., Sundmacher, K., Tobiska, L., and Voigt, A.: Simulations of population balance systems with one internal coordinate using finite element methods. Chem. Eng. Sci. **64** (2009), 733–741.

[8] Koch, J.: Effiziente Behandlung von Integraloperatoren bei populationsdynamischen Modellen. Ph.D Thesis, Otto-von-Guericke-Universität Magdeburg, Fakultät für Mathematik, 2005.

[9] Leveque, R.: *Finite volume methods for hyperbolic problems*. Cambridge University Press, 2002.

[10] Matthies, G., Skrzypacz, P., and Tobiska, L.: A unified convergence analysis for local projection stabilisations applied to the Oseen problem. Math. Model. Numer. Anal. **41** (2007), 713–742.

[11] Mersmann, A.: Batch precipitation of barium carbonate. Chem. Eng. Process. **38** (1993), 6177–6184.

[12] Mersmann, A.: Crystallization and precipitation. Chem. Eng. Process. **38** (1999), 345–353.

[13] Roos, H.-G., Stynes, M., and Tobiska, L.: Robust numerical methods for singularly perturbed differential equations. In: *Convection-diffusion-reaction and fow problems, Springer Series in Computational Mathematics*, vol. 24. Springer-Verlag, Berlin, Second ed., 2008.

Conference Applications of Mathematics 2013
in honor of the 70th birthday of Karel Segeth.
Jan Brandts, Sergey Korotov, Michal Křížek,
Jakub Šístek, and Tomáš Vejchodský (Eds.),
Institute of Mathematics AS CR, Prague 2013

PARALLEL PROGRAMMING AND OPTIMIZATION OF HEAT RADIATION INTENSITY

Jaroslav Mlýnek[1], Radek Srb[2]

[1] Department of Mathematics and Didactics of Mathematics
Technical University of Liberec
Studentská 2, 461 17 Liberec, Czech Republic
jaroslav.mlynek@tul.cz

[2] Institute of Mechatronics and Computer Engineering
Technical University of Liberec
Studentská 2, 461 17 Liberec, Czech Republic
radek.srb@tul.cz

Abstract

This article focuses on the practical possibilities of a suitable use of parallel programming during the computational processing of heat radiation intensity optimization across the surface of an aluminium or nickel mould. In practice, an aluminium or nickel mould is first preheated by infrared heaters located above the outer mould surface. Then the inner mould surface is sprinkled with a special PVC powder and the outer mould surface is continually warmed by infrared heaters. This is an energy-efficient way to produce artificial leathers in the car industry (e.g., the artificial leather on a car dashboard). It is necessary to optimize the location of the heaters to approximately ensure the same heat radiation intensity across the whole outer mould surface during the warming of the mould (to obtain a uniform material structure and color tone of the artificial leather). The problem of optimization is complicated (moulds used in production are often very rugged, during the process of optimization we avoid possible collisions of two heaters as well as a heater and the mould surface). Using of gradient methods is not suitable for solving the problem (minimized function contains many local extremes). A genetic algorithm is used to optimize the location of the heaters. The optimization computation procedure is demanding in terms of the number of numerical operations (especially when the mould volume is large and the number of used infrared heaters is higher). In this article practical results of parallel programming during the calculation process of the evaluation function of every created individual (one possible solution of optimization problem using genetic algorithm) to define its fitness are given. The numerical calculations were performed by a Matlab code written by the authors. Numerical experiments are focused exclusively on the opportunities to use parallel programming to accelerate the optimization procedure.

1. Introduction

This article focuses on the possibilities of parallel programming to accelerate computational optimization of heat radiation intensity on a mould surface. Our minimization problem has many local extremes. Using of gradient methods for finding

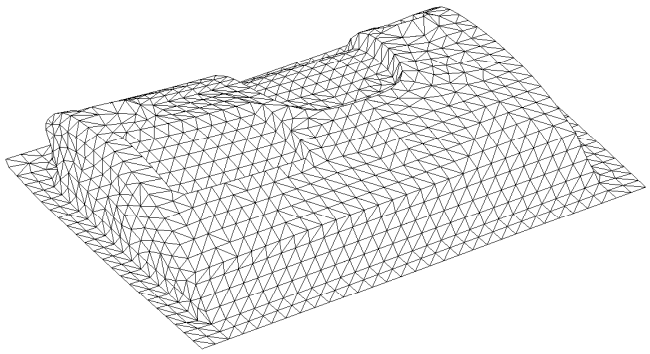


Figure 1: An aluminium mould of a passenger car dashboard part.

global minimum is therefore unsuitable, that is the reason why a genetic algorithm is used.

In practice, an aluminium or nickel mould is preheated by 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 mould surface by finding suitable locations of the heaters. In this way the same material structure and colour of the artificial leather are assured. Moulds of different proportions (often very complicated) and with weight of approximately 300 kilograms are used in production (see Figure 1). 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.

2. The model of heat radiation on the mould surface

In this chapter a simplified mathematical model of heat radiation produced by 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 $e_1 = (1, 0, 0)$, $e_2 = (0, 1, 0)$, $e_3 = (0, 0, 1)$.

Representation of a heater. A heater is represented by abscissa of length d (see Figure 2). The location of a heater is defined by the following parameters: (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 only the 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 every heater Z can be defined by the following 6 parameters

$$Z : (s_1, s_2, s_3, u_1, u_2, \varphi). \tag{1}$$

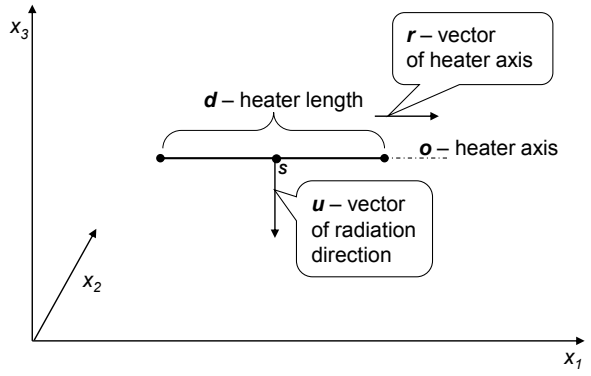


Figure 2: Schematic representation of the heater.

Representation of a mould. The outer mould surface P is described by elementary surfaces p_j, where $j \in \{1, 2, \ldots, N\}$. We have that $P = \cup p_j$, where $j \in \{1, 2, \ldots, N\}$ and int $p_i \cap$ int $p_j = \emptyset$ for $i \neq j$, $1 \leq i, j \leq N$. Every 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 faces "upwards"and therefore is defined through the first two components v_1^j and v_2^j), (iii) the area c_j of the elementary surface. Every elementary surface p_j thus can be defined by the following 6 parameters

$$p_j : (t_1^j, t_2^j, t_3^j, v_1^j, v_2^j, c_j). \tag{2}$$

Experimental measurement of 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 has not provided 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 was $Z : (0, 0, 0, 0, 0, 0)$ in accordance with relation (1), i.e., the centre S of the heater lay at the origin of the Cartesian coordinate system (O, x_1, x_2, x_3); the unit radiation vector had coordinates $u = (0, 0, -1)$ and the vector of the heater axis had 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 parameters in the elementary surface p_j given by relation (2)) and on the direction of the outer normal vector v_j at the 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 describe the position of the centre of gravity of fictitious elementary surface and

fourth and fifth parameters describe the direction of the outer normal vector in the point $[a_1, a_2, a_3]$). We can use measured values $I(a)$ of heat radiation intensity at the selected points a and 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. Interpolation formula is described in details in [2], p. 148.

The general case of a heater location. For a heater in 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 cross product of the vectors $-u$ and r (see more detail in [6], p. 6) 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 have unit length. Then we can define an orthonormal transformation matrix

$$\mathbf{A} = \begin{pmatrix} r_1 & n_1 & -u_1 \\ r_2 & n_2 & -u_2 \\ r_3 & n_3 & -u_3 \end{pmatrix}.$$

Let us 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 trio 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

$$\left(T_j^{'}\right)^T = \mathbf{A}^T \left(T_j - S\right)^T \quad \text{and} \quad \left(v_j^{'}\right)^T = \mathbf{A}^T v_j^T, \tag{3}$$

where $T_j^{'}$ and $v_j^{'}$ are the coordinates in $(S, r, n, -u)$. In this way, we transform the general case of heater location to the measured case and we can calculate heat radiation intensity by using linear interpolation described in previous paragraph "Experimental measurement of heater radiation intensity" (transformed point $T_j^{'}$ and vector $v_j^{'}$ correspond to the point b in previous paragraph).

Calculation of 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 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

$$I_j = \sum_{l \in L_j} I_{jl} \ . \tag{4}$$

The producer of artificial leathers recommends a constant value of heat radiation intensity across the whole outer mould surface. Let us denote this constant value as I_{rec}. We can define F, the aberration of the heat radiation intensity, by the relation

$$F = \frac{\sum_{j=1}^{N} |I_j - I_{rec}| c_j}{\sum_{j=1}^{N} c_j} \tag{5}$$

and the aberration $\tilde{F}$ by the relation

$$\tilde{F} = \sqrt{\sum_{j=1}^{N} (I_j - I_{rec})^2 c_j} \ . \tag{6}$$

We highlight that c_j denotes the area of the elementary surface p_j. We need to find the location of heaters such that value of aberration F (alternatively aberration $\tilde{F}$) will be within specified tolerance.

3. The optimization of the location of the heaters

Function F defined by (5) has many local extremes. Using gradient methods for finding minimum of the function F is not appropriate. If we use a gradient method, there is a high likelihood that we find only local minimum of function F. Therefore, we use a genetic algorithm for finding global minimum of function F (i.e., to optimization of the locations of the heaters). A disadvantage of genetic algorithms is its computational demand and slow convergence. A genetic algorithm is described in more details in [1] and [3]. Implementation of this algorithm for solution of our optimization problem is described in details in [4]. 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 genetic algorithm represents one possible location of the all $6M$ heaters. In the algorithm we successively construct populations of individuals. Every population includes Q individuals where every individual is a potential solution of our problem (in contrast with the gradient methods, where only one potential solution in each iteration exists). We use operators one-point crossover (operator that combines two individual to produce a new individual) and mutation (operator that alters one or more values in an individual) during the generation of new individuals. The generated individuals are saved in the matrix $\mathbf{B}_{Q\times 6M}$. Every row of this matrix represents one individual. We seek the individual $y_{\min} \in C$ satisfying the condition

$$F(y_{\min}) = \min\{F(y); y \in C\}, \tag{7}$$

where $C \subset E_{6M}$ is the searched set. 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 (7) is not realistic in practice. But we are able to determine an optimized solution y_{opt}. Now we describe particular steps of the genetic algorithm that is used.

Genetic algorithm
Input: the specimen y_1 (initial individual), ε_1 - the specified accuracy of the calculation.
Internal computation:
1. create an initial population of Q individuals,
2.a/ evaluate all the individuals of the population (calculate $F(y)$ for every individual y),
b/ sort values $F(y)$ of all individuals y into ascending order and organize individuals y accordingly,
c/ store the individuals y into the matrix $\mathbf{B}$,
3. *repeat until* $\min\{F(y); y \in \mathbf{B}\} < \varepsilon_1$,
a/ choose randomly between the crossover operation and the mutation operation,
b/ *if* the crossover operation is chosen *then*
randomly select (so-called roulette-wheel selection) a pair of individuals (parents), execute the crossover operation and create two new individuals
else
randomly select (roulette-wheel selection) an individual y, execute the mutation operation, create two new individuals
end if,
c/ calculate $F(y)$ for the two new individuals (penalize an individual in the case of the collision of heaters or the collision of a heater and the mould surface), d/ sort as in step 2.b/, e/ take the first Q individuals y with the smallest values $F(y)$ and store them in the matrix $\mathbf{B}$
end repeat.
Output: the first row of matrix $\mathbf{B}$ contains the best found individual.

4. Use of parallel programming during the calculation

Some numerical solutions to practical examples of heat radiation intensity optimization, including graphical representation of the locations of the infrared heaters above the outer mold surface, are published in articles [4] and [5].

This section focuses exclusively on the possibility to use parallel programming to accelerate the optimization procedure. Optimization of locations of heaters using the genetic algorithm requires a number of numerical operations. The calculation of aberration $F(y)$ or $\tilde{F}(y)$ given by the relations (5) and (6) respectively is computationally the most demanding part of the genetic algorithm and is performed for every created new individual y (where value $F(y)$ defines the fitness of the individual y). Before determining the value $F(y)$ or $\tilde{F}(y)$, we have to perform the following computational steps for a given individual y (one of the possible locations of heaters): (i) for every heater Z_i $(1 \leq i \leq M)$ determine the heat radiation intensity over all elementary surfaces p_j $(1 \leq j \leq N)$(use the relation (3) for p_j, interpolate the value of the heat radiation intensity of heater Z_i on the p_j using the interpolation formula, (ii) calculate the total heat radiation intensity I_j on p_j using the relation (4) for

all p_j. The calculations of heat radiation intensities of the heaters Z_i and Z_k $(i \neq k)$ for each and every elementary surfaces p_j are completely independent. The calculation time of $F(y)$ or $\tilde{F}(y)$ and thus the overall time of the optimization procedure can be significantly reduced by using the tools of parallel programming. For this experiment we used a PC with 3GB RAM, CPU 2x AMD Athlon 64 X2 Dual Core 2.81 GHz. We performed experiments for one, two and four processors.

The tests are carried out for the aluminium mould of a passenger car dashboard (see Figure 1). The volume of the mould was $0.6 \times 0.4 \times 0.12\,\mathrm{m}^3$, the number of elementary surfaces was $N = 2\,178$. The infrared heaters used were all the same type (capacity 1 600 W, length 15 cm, width 4 cm), the manufacturer of artificial leathers recommended heat radiation intensity $I_{rec} = 47\,\mathrm{kW/m}^2$. The calculations were performed using a Matlab code (including parallel programming) written by the authors. First, we focused on the real time of the calculation of the aberration $F(y)$ defined by relation (5). Real times of the calculation of $F(y)$ for different numbers of processors used and different numbers of heaters are presented in Table 1. The times in Table 1 required to the calculation of $F(y)$ were measured on the specified computer.

The total duration of the optimization procedure depends on the number of processors used, the number of heaters used and on the number of iterations of the genetic algorithm (two new individuals are generated in one iteration). The results are presented in Table 2. The maximum number of iterations in our tests was 100 000. We did not obtain better optimized solution after using a higher number of iterations. The times in Table 2 of total duration of optimization were measured as in Table 1 on the specified computer.

The duration of the optimization procedure can be significantly accelerated by using parallel programming to calculate $F(y)$ as is shown in Table 2. The acceleration of the optimization procedure is effective especially with higher number of infrared heaters and large number of elementary surfaces of the mould surface.

Number of applied heaters	Number of used processors		
	1	2	4
	Time of value $F(y)$ calculation [s]		
10	0.2510	0.1447	0.0915
20	0.4928	0.3129	0.2229
30	0.7853	0.4463	0.2769
40	1.0470	0.5951	0.3692
50	1.3088	0.7439	0.4615

Table 1: Time required for the calculation of value $F(y)$.

Number of applied heaters	Number of GA iterations	Number of used processors		
		1	2	4
		Time of optimization [h]		
10	20,000	1.3336	0.8366	0.5882
30	20,000	4.2792	2.6805	1.8812
50	20,000	7.2732	4.5499	3.1882
10	50,000	3.3340	2.0916	1.4704
30	50,000	10.6979	6.7016	4.7030
50	50,000	18.1831	11.3747	7.9706
10	100,000	6.6680	4.1832	2.9408
30	100,000	21.3958	13.4027	9.4061
50	100,000	36.3662	22.7495	15.9412

Table 2: Time required for the optimization procedure.

Acknowledgements

This work was supported by projects OP VaVpI CZ.1.05/2.1.00/01.0005 and IMSUCP SGS, No. 2012/7821 .

References

[1] Affenzeler, M., Winkler, S., Wagner, S., and Beham, A.: *Genetic algorithms and genetic programming*. Chapman and Hall/CRC, Boca Raton, 2009.

[2] Antia, H. M.: *Numerical methods for scientists and engineers*. Birkhäuser Verlag, Berlin, 2002.

[3] Chambers, L.: *Genetic algorithms*. Chapman and Hall/CRC, Boca Raton, 2002.

[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*, pp. 461–467. Digitaldruck Pirrot GmbH, Koblenz, Germany, May 2012.

[5] Mlýnek, J. and Srb, R.: The optimization of heat radiation intensity. In: J. Chleboun, K. Segeth, J. Šístek, T. Vejchodský (Eds.), *Proceedings of the 16th Conference Programs and Algorithms of Numerical Mathematics*, Horní Maxov, June 2012, pp. 142–148.

[6] Stoker, J. J.: *Differential geometry*. John Wiley & Sons, New York, 1989.

Conference Applications of Mathematics 2013
in honor of the 70th birthday of Karel Segeth.
Jan Brandts, Sergey Korotov, Michal Křížek,
Jakub Šístek, and Tomáš Vejchodský (Eds.),
Institute of Mathematics AS CR, Prague 2013

INTEGRAL TRANSFORMS – THE BASE OF RECENT TECHNOLOGIES

Vratislava Mošová

Moravian University College Olomouc
Jeremenkova 1142/42, 772 00 Olomouc, Czech Republic
vratislava.mosova@mvso.cz

Abstract

In this article, the attention is paid to Fourier, wavelet and Radon transforms. A short description of them is given. Their application in signal processing especially for repairing sound and reconstructing image is outlined together with several simple examples.

1. Introduction

In this survey paper we will deal with such integral transforms that are used in the image or sound processing. The transforms, we will speak about, were defined already long time ago. Joseph Fourier approximated 2π periodic functions by trigonometric series in 1778. The first wavelet basis – Haar wavelets – was proposed as an example of a countable orthonormal system in $L^2(R)$ in 1909. Johann Radon described the reconstruction of a function from its line integral values in his article in 1917. The entry of computers amplified the importance of these transforms in the second half of the past century, because the above named transforms became the theoretical base of algorithms that are used in signal processing or in computer tomography. In such a way they give possibilities to remove noise from the sound or visual recordings, to compress the image data before their transmission, to find the trend in given time series or to identify malignant tumors in a human body.

The outline of this article is as follows. Some basic information about the definition and construction of the Fourier transform together with examples is presented in Section 2. The wavelet transform is described and applied on the given data in Section 3. The Radon transform and its usage in medicine is discussed in Section 4.

2. Fourier transform

For $f \in L^1(R)$, the relation

$$F(\omega) = \int_{-\infty}^{\infty} f(x)\mathrm{e}^{2\pi \mathrm{i}\omega x}\,\mathrm{d}x,\ \omega \in R, \tag{1}$$

represents the continuous Fourier transform (FT) of a function f. The integral transform

$$F^{-1}(x) = \int_{-\infty}^{\infty} F(\omega) \mathrm{e}^{-2\pi \mathrm{i}\omega x} \, \mathrm{d}\omega, \ x \in R, \tag{2}$$

is the inverse Fourier transform.

The discrete analogies of the relations (1) and (2) are suitable for computer implementation. If sampled values $f_0, \ldots, f_{N-1}$ of a function f are given, the components of the discrete Fourier transform (DFT) are defined by

$$F_k = \sum_{j=0}^{N-1} f_j \mathrm{e}^{\frac{2\pi \mathrm{i} jk}{N}}, \ k = 0, \ldots, N-1, \tag{3}$$

and the components of the discrete inverse Fourier transform (DIFT)

$$f_j = \frac{1}{N} \sum_{k=0}^{N-1} F_k \mathrm{e}^{\frac{-2\pi \mathrm{i} jk}{N}}, \quad j = 0, \ldots, N-1. \tag{4}$$

The number of operations that are used for calculation of the DFT by relation (3) has the order $O(N^2)$. But there is an effective numerical algorithm of fast Fourier tramsform (FFT) that allows to reduce the number of used operations. This algorithm is based on the properties of exponential functions and on an ingenious arrangement of computation that is given in the next lemma (see [5]).

Lemma 1 (Danielson-Lanczos, 1942) Let N be even. Then

$$F_k = F_k^0 + W^k F_k^1, \ k = 0, \ldots, N-1, \tag{5}$$

where $F_k^0 = \sum_{j=0}^{\frac{N}{2}-1} W^{jk} f_{2j}$, $F_k^1 = \sum_{j=0}^{\frac{N}{2}-1} W^{jk} f_{2j+1}$, $W = \mathrm{e}^{\frac{2\pi \mathrm{i}}{N}}$ and $W^{jk} = \mathrm{e}^{\frac{2\pi \mathrm{i} jk}{N}}$.

Lemma 1 can be applied recurrently M times if $N = 2^M$. The FFT in the following way reduces the order of the number of operations that are necessary to compute the Fourier coefficients from $O(N^2)$ to $O(N \log N)$. Note that the schematic expression of the computation that is generated by relation (5) looks like a butterfly. This is the reason, why the name "butterfly" is used for one loop of the FFT process.

Example 1 The function $f(x) = \sin 1500x$ is impaired by random noise. The DFT for $N = 200$ values is computed and expressed in Figure 2.

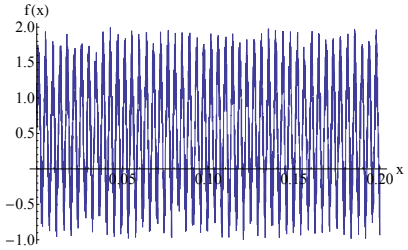


Figure 1: The original signal

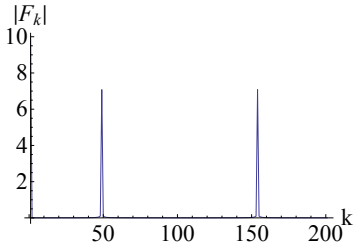


Figure 2: The signal after the FT

The signal is a quantity that depends on one or more variables. For example, a sound is a one dimensional signal that depends on time and a digital photograph is two dimensional signal over a matrix of pixels. While a signal gives information about variability with respect to independent variables, its FT gives information about frequencies that occur in the given signal. The knowledge of the frequency spectrum of a signal is important, because it helps to analyze this signal. The low frequencies are important for identification of the signal. The higher frequencies often represent the noise.

In the signal processing, the DFT is applied on the given data at first. This way the time depending function changes on the frequency depending function. Then, the received Fourier coeficients can be modified according to monitored aims. For instance, the noise can be removed from the given signal if the Fourier coefficients with frequency higher than the given treshold λ are put to zero. A signal is comprimed when the majority of Fourier coefficients is neglected. The DIFT is applied on the rest of the Fourier coefficients in the end.

The real part of the FT – the discrete cosine transform (DCT)

$$C_{km} = \sum_{j=0}^{N-1} \sum_{l=0}^{P-1} f_{jl} \cos \frac{2\pi \mathrm{i} jk}{N} \cos \frac{2\pi \mathrm{i} lm}{P}, \ k = 0, \ldots, N-1, \ m = 0, \ldots, P-1, \quad (6)$$

is the proper tool if some real 2D data are processed. For instance, the DCT is used for compression of an image in the JPEG format.

Example 2 Removing noise from the given data by hard tresholding (λ=0.5).

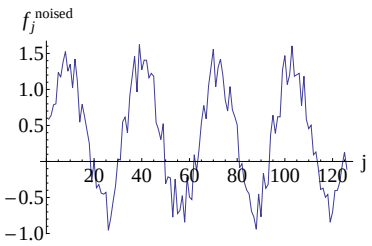


Figure 3: The noised data

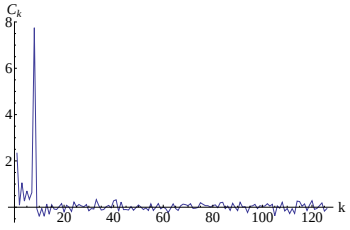


Figure 4: The DCT of the noised data

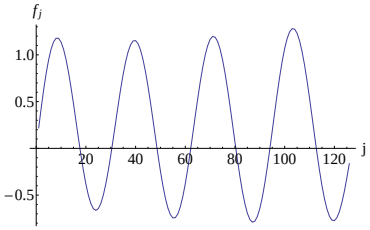


Figure 5: The data with removed noise

When the DFT is applied to data, the information about the frequency is received but the information about the time is lost. It means that the DFT is suitable for an analysis of stationary signals and it does not detect the jump changes and the trends that occur in non-stationary signals. The time localization of the signal can be reached if the short time Fourier transform (STFT) is used. The STFT is defined by

$$F(\omega, t) = \int_{-\infty}^{\infty} f(x) w_r \left(\frac{x-t}{r} \right) \mathrm{e}^{-\mathrm{i}\omega x} \, \mathrm{d}x, \quad (7)$$

where $w_r(x) = w(\frac{x}{r})$ is a window (a function smooth enough that is compactly supported). The parameter r allows to adjust the length of the analyzed signal segment. The size is the same for all windows in the discrete version of STFT.

3. Wavelet transform

Let f be in $L^2(R)$ and ψ be the wavelet (i.e. a function that can be imagined like a small wave that decreases quickly to 0 in $\pm\infty$). The wavelet transform is defined by

$$W_\psi(a,b) = \frac{1}{|a|}\int_{-\infty}^{\infty} f(x)\psi\left(\frac{x-b}{a}\right)\,\mathrm{d}x. \tag{8}$$

Here a is a scale[1] and b is a translation. If $a \in R$ and $b \in R$ we speak about the continuous wavelet transform (CWT).

Example 3 The CWT of the given signal using the Mexican hat wavelet $\psi(x) = \frac{2}{\sqrt{3}}\pi^{-1/4}(1-x^2)\mathrm{e}^{-x^2/2}$ is done. The corresponding scalogram (i.e. the graph in which the density of energy $E(a,b) = |(W_\psi f)(a,b)|^2$ for the scale a and for the position b is expressed) is given in Figure 7. Here, large absolute values of the wavelet coefficients are shown darker.

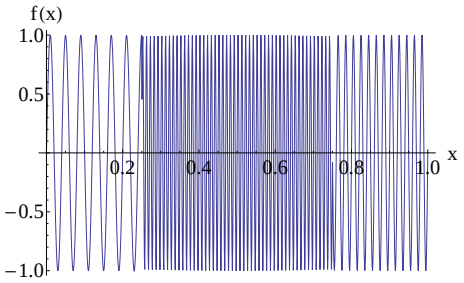


Figure 6: The original signal

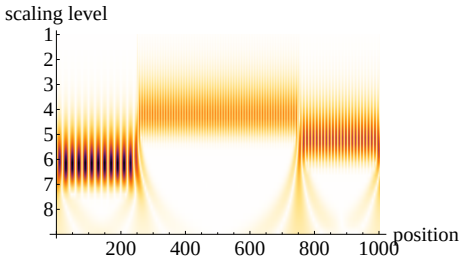


Figure 7: The scalogram

If a and b are discrete values we speak about the discrete wavelet transform (DWT). The diadic dilatation $a = 2^j$ and the translation $b = k$, where $j, k \in Z$, are used for the sake of the computation effectivity. The DWT has then the form

$$W_{j,k} = 2^{\frac{j}{2}}\int_{-\infty}^{\infty} f(x)\psi(2^j x - k)\,\mathrm{d}x. \tag{9}$$

The discrete reconstruction is realized by

$$f(x) = \sum_{j\in Z}\sum_{k\in Z} 2^{\frac{j}{2}} W_{j,k}\psi(2^j x - k). \tag{10}$$

But the system $2^{\frac{j}{2}}\psi(2^j x - k)$ does not need to be orthonormal for general functions ψ. One of possibilities how to receive an orthonormal basis in $L^2(R)$ is to use the multiresolution analysis (MRA)[2], where the spaces $V_j \subset L^2(R)$ $(j \in Z)$ that satisfy

$$V_j \subset V_{j+1};\ \bigcap_{j\in Z} V_j = \{0\};\ \bigcup_{j\in Z} V_j = L^2(R);$$

[1]Scales and the frequencies are connected: Higher scales correspond to lower frequencies.

[2]The construction of waveles by means of the MRA based on the existence of a scale function φ was proposed by Mallat in 1988.

$$\exists \varphi \in V_0 : \{\varphi(x-k)\}_{k\in Z} \text{ is a complete orthogonal set in } L^2(R);$$
$$f \in V_0 \Leftrightarrow f(2^j x) \in V_j$$

are constructed.

It follows from the properties of the spaces V_j given above that there exist the subspaces W_j orthogonal to V_j such that $V_{j+1} = V_j \oplus W_j$.

If $\{V_j\}$ is the MRA and φ is the scaling function that satisfies the dilatation equation

$$\varphi(x) = \sqrt{2} \sum_{k\in Z} u_k \varphi(2x-k), \tag{11}$$

then

$$\psi(x) = \sqrt{2} \sum_{k\in Z} v_k \varphi(2x-k), \text{ where } v_k = (-1)^{k-1} \overline{u_{1-k}} \tag{12}$$

is the associated wavelet correspondig to the MRA.

The spaces V_j resp. W_j are generated by functions that are dilatations and translations of the scaling function and the associated wavelet function[3]

$$V_j = \text{span}\{\varphi_{j,k}\}_{j,k\in Z}, \text{ where } \varphi_{j,k}(x) = 2^{j/2}\varphi(2^j x - k), \tag{13}$$

$$W_j = \text{span}\{\psi_{j,k}\}_{j,k\in Z}, \text{ where } \psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k). \tag{14}$$

The space V_{j+1} can be interpreted as an approximation space in $L^2(R)$ and $V_{j+1} = V_0 \oplus W_0 \oplus W_1 \oplus \cdots \oplus W_j$. It means that every function $f \in L^2(R)$ can be written as

$$f(x) = \sum_{k\in Z} a_{0,k}\varphi_{0,k}(x) + \sum_{j\geq 0}\sum_{k\in Z} b_{j,k}\psi_{j,k}(x), \tag{15}$$

where $a_{0,k}$ are the scaling coefficients and $b_{j,k}$ are the wavelet coefficients of f on the level j.

Let $\langle f, g\rangle$ be the inner product in $L^2(R)$. In what follow, we will denote the vectors of wavelet coefficients of f on the level j by

$$\mathbf{b}_j = (b_{j,k})_{k\in Z}, \text{ where } b_{jk} = \langle f, \psi_{j,k}\rangle, \tag{16}$$

and the vectors of scaling coefficients of f on the level j as

$$\mathbf{a}_j = (a_{j,k})_{k\in Z}, \text{ where } a_{j,k} = \langle f, \varphi_{j,k}\rangle. \tag{17}$$

Computation of wavelet coefficients is divided in two parts in the Mallat algorithm (see [3]).

[3]Note that multivariable wavelets are constructed in the form of the tensor product. For instance, a 2D MRA on the first level can be constructed from a decomposition

$$V_1^1 \oplus V_1^2 = (V_0^1 \otimes V_0^2) \oplus (V_0^1 \otimes W_0^2) \oplus (W_0^1 \otimes V_0^2) \oplus (W_0^1 \otimes W_0^2)$$

and the wavelet basis is given by $\{\varphi_{0,k} \otimes \varphi_{0,l}\}_{l\in Z} \cup \{\varphi_{0,k} \otimes \psi_{0,l}, \psi_{0,k} \otimes \varphi_{0,l}, \psi_{0,k} \otimes \psi_{0,l}\}_{l\in Z}$.

In the first one – decompositon, the wavelet coefficients are computed from the given data: The vector $\mathbf{a}_m$ of the scaling coefficients of function f is given for $m \in Z$ large enough. The wavelet transform $\mathbf{b}_{m-1}, \ldots, \mathbf{b}_{m-l}, \mathbf{a}_{m-l}$ of f is computed for a chosen $l \in N$ in the following way:

$$\mathbf{b}_j = D(\mathbf{a}_{j+1} * \tilde{\mathbf{v}}),\ \mathbf{a}_j = D(\mathbf{a}_{j+1} * \tilde{\mathbf{u}}),\ j = m-1, \ldots, m-l, \tag{18}$$

where $D(z_n) = z_{2n}$ is the downsampling operator, $\tilde{z}_n = \overline{z_{-n}}$ is the operator of conjugated reflexion and $\mathbf{b}_{j+1} * \tilde{\mathbf{u}}$ is the convolution of the vector $\mathbf{b}_{j+1}$ with the vector $\tilde{\mathbf{u}}$.

In the second part – reconstruction, the vector $\mathbf{a}_m$ is constructed from the received set $\mathbf{b}_{m-1}, \ldots, \mathbf{b}_{m-l}, \mathbf{a}_{m-l}$ in the following way:

$$\mathbf{a}_{j+1} = (U(\mathbf{a}_j)) * \mathbf{u} + (U(\mathbf{b}_j)) * \mathbf{v},\ j = m-l, \ldots, m-1, \tag{19}$$

where $U(z_n) = z_{n/2}$ for n even and zero while for n odd it is the operator of upsampling.

This process is realized by using proper quadratic mirror filters in signal processing. The given vector of values that represents a signal goes through the lowpass filter and highpass filter in the first phases of computation. The approximation coefficients a_j and detail coefficients b_j are received. Note that the approximation coefficients belong to low frequencies that represent trends and the details belong to high frequencies that can be interpreted as noise. The received outputs are downsampled and they can be filtered again. It is possible to express this process graphically in the form of the completing wavelet tree. In the second phasis the received approximation and details are upsampled and then they are filtered by conjugate filters.

Before reconstruction it is possible to modify the wavelet coefficients. For example, noise is removed from the given signal, if the wavelet coefficients $b_{i,j}$ that have smaller frequecy than the chosen treshold λ are set to zero. Also the soft thresholding with the modified cofficients $\tilde{b}_{j,k} = \begin{cases} 0 & \text{if } b_{j,k} < \lambda, \\ \operatorname{sgn} b_{j,k}\ |b_{j,k} - \lambda| & \text{in other cases} \end{cases}$ can be used.

Example 4 Removing noise from the given data by means of the wavelet transform. Here, the Daubechies wavelet Db4 was used.

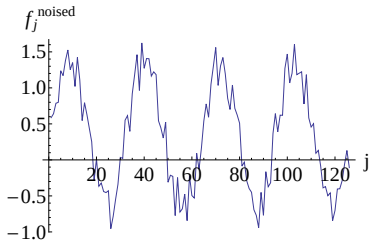


Figure 8: The noised data

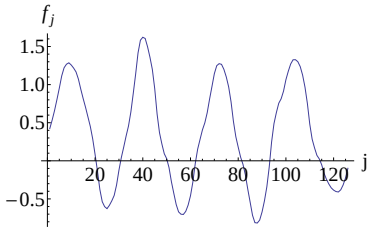


Figure 9: The data with removed noise

If a 2D visual signal has to be compressed, it is decomposed into horizontal, vertical, diagonal and approximation coefficients in the beginning. Only the approximation coefficients are used for the next decomposition, because only they hold the important information. The received details are cut on asked level. The DWT with the hard tresholding is used in the JPEG2000 format.

Example 5 The wavelet transform of the given image.

Figure 10: The original image

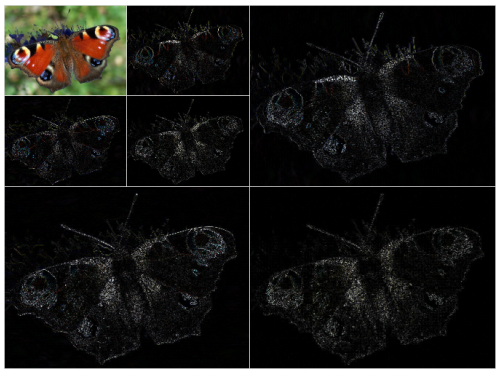


Figure 11: The decomposition of the image up to refinement level 2

4. Radon transform

Generally, the Radon transform[4] of a function f from the Schwartz space $S(R^n)$ is the integral transform

$$g(t,\theta) = \int\limits_{x\cdot\theta=t} f(x)\,\mathrm{d}m(x), \tag{20}$$

where $\{x \in R^n : x\cdot\theta = t\}$ is a hyperplane for a fixed $t \in R$ and $\theta \in \mathcal{S}^{n-1}$, $\mathcal{S}^{n-1} = \{x \in R^n : \|x\| = 1\}$ is a sphere, and $\mathrm{d}m$ is the Lebesgue measure.

At the beginning of the last century, the Austrian mathematician J. F. Radon found the way how to reconstruct the function f from the values g. If $n = 3$ the inverse Radon transform has the form

$$f(x) = -\frac{1}{8\pi^2}\Delta_x \int\limits_{S^3} g(\langle x,\theta\rangle,\theta)\,\mathrm{d}S^3_\theta \tag{21}$$

and if $n = 2$ the inverse Radon transform (IRT) is

$$f(x) = \frac{1}{4\pi^2}\int\limits_{S^2} \text{v.p.} \int\limits_{-\infty}^{\infty} \frac{g'_t(t,\theta)}{x\cdot\theta - t}\,\mathrm{d}t\,\mathrm{d}S^2_\theta, \tag{22}$$

[4]Note that the Radon transform is closely connected to the Fourier transform. The nD Fourier transform of f is the composition of the Radon transform of f and 1D Fourier transform.

where "v.p." means "within the meaning of the Cauchy principal value". The Radon transform (RT) of a function $f \in L^2(R^2)$ is given by the line integral

$$g(t,\varphi) = \int\limits_{\langle x,\theta\rangle = t} f(x)\,\mathrm{d}x, \quad \theta = (\cos\varphi, \sin\varphi)^T. \tag{23}$$

The inverse Radon transform (IRT) by

$$f(x) = \frac{1}{4\pi^2} \int\limits_0^{2\pi} \text{v.p.} \int\limits_{-\infty}^{\infty} \frac{g'_t(t,\varphi)}{x\cdot\theta - t}\,\mathrm{d}t\,\mathrm{d}\varphi. \tag{24}$$

The relations (23) and (24) became the theoretical basis of the computer tomography with the following basic idea: If a body is irradiated by X-rays or other type of waves, the intensity of the radiation I changes depending on the density distribution f of substances through which it passes. When the the initial intensity of radiation is I_0 and $l(x,\theta)$ is a line which the ray goes along, this change can be expressed as

$$\ln\frac{I_0}{I} = \int\limits_{l(x,\theta)} f(x)\,dx. \tag{25}$$

It means that the value $\ln\frac{I_0}{I}$ is equal to the Radon transform of f. When measurements for different directions of rays are realized, the inverse Radon transform can be used to determine the density distribution f in the studied plane.

The received results can be demonstrated graphically. The measured values for each ray are represented by the corresponding gray's shade. This allows to express graphically the density of distribution in the planar section. The space image arises by composition of the images from different planar sections.

Example 6 The Radon transform in R^2 of the given picture is done. Its graphical expression – sinogram – is given in Figure 13. Here, lighter color is assigned to higher values of the RT.

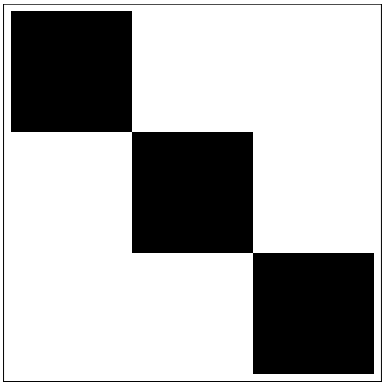


Figure 12: The original image

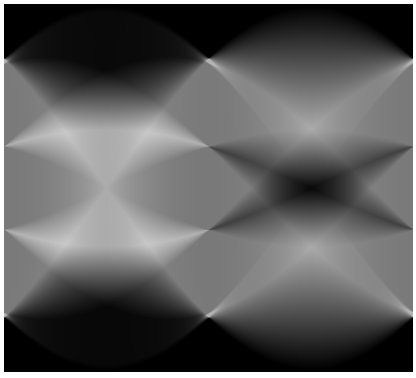


Figure 13: The sinogram

In practice, all measurements and evaluation are realized on the tomograph that consists of a scanner, computer and monitor. A program for the evaluation of the data provided by scanner is built in the computer. This program is based on a numerical algorithm. There are three basic types of algorithms in computer tomography that are used for reconstruction – convolution algorithms, algebraic algorithms and Fourier reconstruction. We focus only on one of the convolution algorithms that is used in medicine.

The formula (24) for inverse Radon transform is the base for the convolution reconstruction algorithms in the plane. But the form of algorithm depends on the design of the scanner (the formulas for parallel-ray geometry and divergent-ray geometry see [5]). Recent tomographic scanners are equipped with the 4^{th} generation of detectors placed around the circumference of a circle that moves along the source sending divergent rays.

Denote

D – the distance of the source from the origin of the coordinate system,

L – the distance of the reconstructed point (ρ, ψ) from the source,

β – the angular position of the source,

γ – the angle that gives the location of a ray within a fan,

γ' – the angle of the ray that passes through the reconstructed point (ρ, ψ) (see Figure 14).

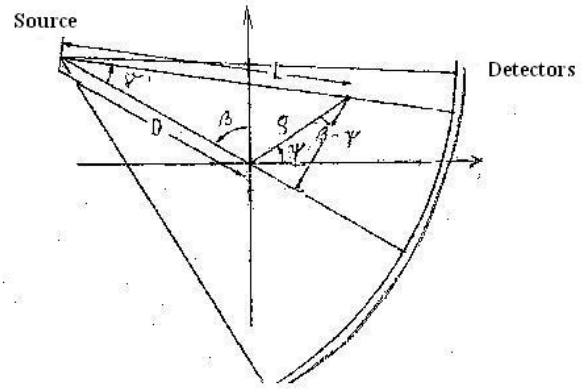


Figure 14: Measurement of data

The formula for the inverse Radon transform is converted into the form

$$f(\rho, \psi) = \frac{D}{2} \int_0^{2\pi} \int_{-\infty}^{\infty} v(L \sin(\gamma' - \gamma))\, g(\beta, \gamma) \cos\gamma \, \mathrm{d}\gamma \, \mathrm{d}\beta. \tag{26}$$

The derivation can be found in [2].

If the source is rotated p times about the same angle $\Delta\beta = \frac{2\pi}{p}$ and it always sends $2q$ rays that form an equal angle $\Delta\gamma = \frac{\pi}{2q}$, the values $g(\beta_j, \gamma_l)$ are received.

Now, the integral in equation (26) can be calculated by the trapezoidal rule

$$f(\rho,\psi) \sim \frac{D}{2}\Delta\beta\Delta\gamma \sum_{j=0}^{p-1}\sum_{l=-q}^{q} v(L\sin(\gamma_k-\gamma_l))g(\beta_j,\gamma_l)\cos\gamma_l. \tag{27}$$

The reconstruction of the function f is divided into two phases. First, the convolution of functions v and g (i.e. the sum inside the formula (27)) is calculated and, second, the back projection is performed.

5. Conclusion

The Fourier transform and the wavelet transform are used in signal processing, they allow to extract information from many different kinds of data, they can help to analyze voice or to compress pictures, they can also serve to analyze variability, to remove noise or to detect significant moments in the time series that are used in economy.

Also the tomographic methods have broad application. We can meet them not only in medical diagnostics, but they are also used in studying structure of materials (the study of composite materials), in prospecting (mapping oil deposits, the ocean floor), in pyrometry (temperature in the blast furnace) or in astronomy.

References

[1] Jelínek, J., Segeth, K., Overton, T. R.: Three-dimensional reconstruction from projections. Apl. Mat. **30** (1985), 92–109.

[2] Kak, A. C., Slaney, M.: *Principles of computerized tomographic imaging.* Society of Industrial and Applied Mathematics, IEEE Press New York 1988.

[3] Najzar, K.: *Základy teorie waveletů*, 1. vyd. Praha: Karolinum, 2004, 198 s. Učební texty Univerzity Karlovy v Praze.

[4] Radon, J.: *Über die Besstimmung von Functionen durch ihre Integralwertelangsgewisser Mannigfaltigkeiten.* Berichte Sachsissche Academie der Wissenschaften, 1917.

[5] Segeth, K.: *Numerický software I.* Karolinum, Praha, 1998.

Conference Applications of Mathematics 2013
in honor of the 70th birthday of Karel Segeth.
Jan Brandts, Sergey Korotov, Michal Křížek,
Jakub Šístek, and Tomáš Vejchodský (Eds.),
Institute of Mathematics AS CR, Prague 2013

ZERO POINTS OF QUADRATIC MATRIX POLYNOMIALS

Gerhard Opfer[1], Drahoslava Janovská[2]

[1] University of Hamburg
Bundesstraße 55, 20146 Hamburg, Germany
opfer@math.uni-hamburg.de
[2] Institute of Chemical Technology
Technická 5, 166 28 Prague 6, Czech Republic
janovskd@vscht.cz

Abstract

Our aim is to classify and compute zeros of the quadratic two sided matrix polynomials, i.e. quadratic polynomials whose matrix coefficients are located at both sides of the powers of the matrix variable. We suppose that there are no multiple terms of the same degree in the polynomial $\mathbf{p}$, i.e., the terms have the form $\mathbf{A}_j\mathbf{X}^j\mathbf{B}_j$, where all quantities $\mathbf{X}, \mathbf{A}_j, \mathbf{B}_j, j = 0, 1, \ldots, N$, are square matrices of the same size. Both for classification and computation, the essential tool is the description of the polynomial $\mathbf{p}$ by a matrix equation $\mathbf{P}(\mathbf{X}) := \mathbf{A}(\mathbf{X})\mathbf{X} + \mathbf{B}(\mathbf{X})$, where $\mathbf{A}(\mathbf{X})$ is determined by the coefficients of the given polynomial $\mathbf{p}$ and $\mathbf{P}$, $\mathbf{X}$, $\mathbf{B}$ are real column vectors. This representation allows us to classify five types of zero points of the polynomial $\mathbf{p}$ in dependence on the rank of the matrix $\mathbf{A}$. This information can be for example used for finding all zeros in the same class of equivalence if only one zero in that class is known. For computation of zeros, we apply Newtons method to $\mathbf{P}(\mathbf{X}) = \mathbf{0}$.

1. Introduction

In papers [4, 5] we have investigated quaternionic polynomials of the one-sided and the two-sided type. The one-sided type is described by terms of the form $a_j x^j$ or $x^j a_j$, whereas the two-sided type is described by terms of the form $a_j x^j b_j$, $j \geq 0$. In this paper we will consider matrix polynomials which have matrix coefficients and a matrix variable as well, i.e. the terms have the form $\mathbf{A}_j\mathbf{X}^j\mathbf{B}_j$. All quantities $\mathbf{X}$, $\mathbf{A}_j$, $\mathbf{B}_j$, $j = 0, 1, \ldots, N$, are square matrices of the same size.

We will use the notation $\mathbb{R}, \mathbb{C}$ for the field of real and complex numbers, respectively; $\mathbb{K}$ will stand for $\mathbb{R}$ or $\mathbb{C}$. The set of square matrices over $\mathbb{K}$ will be denoted by $\mathbb{K}^{n\times n}$, where n is the order of the matrix. By $\mathbf{I} \in \mathbb{K}^{n\times n}$ we will denote the identity matrix, the matrix $\mathbf{0} \in \mathbb{K}^{n\times n}$ is the zero matrix.

Since the general task is very complicated, in this paper we will restrict ourselves to quadratic matrix polynomials without multiple terms of the same degree: for

given $\mathbf{A}_0,\ \mathbf{A}_1,\ \mathbf{A}_2,\ \mathbf{B}_0,\ \mathbf{B}_1,\ \mathbf{B}_2 \in \mathbb{K}^{n\times n}$, we consider quadratic polynomial $\mathbf{p}$ in the form

$$\mathbf{p}(\mathbf{X}) = \mathbf{A}_0\mathbf{B}_0 + \mathbf{A}_1\mathbf{X}\mathbf{B}_1 + \mathbf{A}_2\mathbf{X}^2\mathbf{B}_2, \quad \text{where} \quad \mathbf{A}_0\mathbf{B}_0,\ \mathbf{A}_2,\ \mathbf{B}_2 \neq \mathbf{0}\,. \tag{1}$$

The condition $\mathbf{A}_0\mathbf{B}_0 \neq \mathbf{0}$ implies that $\mathbf{p}(\mathbf{0}) \neq \mathbf{0}$. The conditions $\mathbf{A}_2,\ \mathbf{B}_2 \neq \mathbf{0}$ imply that the term with the degree 2 is nonvanishing.

If the matrix $\mathbf{X}$ has the property $\mathbf{p}(\mathbf{X}) = \mathbf{0}$, we will call $\mathbf{X}$ a zero of $\mathbf{p}$.

As an example, let us consider matrices of the order $n = 2$. In this case the quadratic matrix polynomial can be formally transformed into a linear system of four equations (for $n = 2$, it is true for polynomials of any degree N) and we will classify the zeros of the polynomial in terms of the rank of the corresponding system.

In general, we transform the quadratic matrix polynomial $\mathbf{p}$ into a matrix equation $\mathbf{P}(\mathbf{X}) := \mathbf{A}(\mathbf{X})\mathbf{X} + \mathbf{B}(\mathbf{X})$, where $\mathbf{A}(\mathbf{X})$ is determined by the coefficients of the given polynomial $\mathbf{p}$ and $\mathbf{P},\ \mathbf{X},\ \mathbf{B}$ are real column vectors. Then we classify zeros by the rank of the matrix $\mathbf{A}$. We showed that in general there are five different types of zeros.

For computation of zeros, we apply Newton's method to the matrix equation $\mathbf{P}(\mathbf{X}) = \mathbf{0}$.

2. Preliminaries

This section contains basic facts from the theory of matrices. It can be found e. g. in Horn and Johnson, [2].

Let $\mathbb{A} \in \mathbb{K}^{n\times n}$. Then $\chi_{\mathbf{A}}(z) := \det(z\mathbf{I} - \mathbf{A}) = z^n + a_{n-1}^{(n)}z^{n-1} + \cdots + a_0^{(n)}$ is called the characteristic polynomial of $\mathbf{A}$. Cayley–Hamilton theorem says that the matrix $\mathbf{A}$ annihilates its characteristic polynomial,

$$\chi_{\mathbf{A}}(\mathbf{A}) = \mathbf{A}^n + \cdots + a_0^{(n)}\mathbf{I} = \mathbf{0}\,. \tag{2}$$

In particular, for $n = 2$ we have

$$\mathbf{A}^2 - \operatorname{tr}(\mathbf{A})\mathbf{A} + \det(\mathbf{A})\mathbf{I} = \mathbf{0}\,.$$

Let us recall that two matrices $\mathbf{A}$, $\mathbf{B}$ of the same order over $\mathbb{K}$ are similar if there is a nonsingular matrix $\mathbf{H}$ of the same order such that $\mathbf{A} = \mathbf{H}\mathbf{B}\mathbf{H}^{-1}$.

For fixed $\mathbf{A} \in \mathbb{K}^{n\times n}$ the set of matrices

$$[\mathbf{A}] = \{\,\mathbf{B}, \mathbf{B} = \mathbf{H}\mathbf{A}\mathbf{H}^{-1} \text{ for all nonsingular } \mathbf{H}\} \tag{3}$$

is called similarity class of $\mathbf{A}$. The similarity class is finite only for multiples of the identity matrix: if $\mathbf{A} = c\mathbf{I},\ c \in \mathbb{K}$, then $[\mathbf{A}] = \{\mathbf{A}\}$ consists only of one element.

There are two special cases of (1) worth mentioning. If we put $\mathbf{X} := z\mathbf{I} \in \mathbb{K}^{n\times n}$, where $z \in \mathbb{K}$, we obtain

$$\mathbf{p}(\mathbf{X}) = \mathbf{p}(z\mathbf{I}) = \mathbf{C}_0 + \mathbf{C}_1\,z + \mathbf{C}_2\,z^2, \quad \mathbf{C}_j = \mathbf{A}_j\mathbf{B}_j,\ j = 0, 1, 2\,. \tag{4}$$

If all coefficients have the special form $\mathbf{A}_j = \alpha_j \mathbf{I} \in \mathbb{K}^{n\times n}$, $\mathbf{B}_j = \beta_j \mathbf{I} \in \mathbb{K}^{n\times n}$, $\gamma_j := \alpha_j\beta_j$, $j = 0, 1, 2$, we obtain

$$\mathbf{p}(\mathbf{X}) = \gamma_0\mathbf{I} + \gamma_1\mathbf{X} + \gamma_2\mathbf{X}^2\,. \tag{5}$$

Both forms have their ranges in $\mathbb{K}^{n\times n}$, see also [7, 3].

Definition The set of matrices

$$\mathcal{C} := \{\mathbf{M} :\ \mathbf{M} = a\mathbf{I} \in \mathbb{K}^{n\times n}\} \tag{6}$$

is called the *center* of $\mathbb{K}^{n\times n}$.

Remark In general terms the center of a noncommutative (semi)group $\mathcal{G}$ is the set of all elements, which commute with all elements of $\mathcal{G}$.

If we want to find out whether an element of the center $\mathcal{C}$ is a zero of a given quadratic matrix polynomial $\mathbf{p}$, then, we have to use the form (4), namely

$$\mathbf{p}(z\mathbf{I}) = \mathbf{C}_0 + \mathbf{C}_1 z + \mathbf{C}_2 z^2 = \mathbf{0} \in \mathbb{K}^{n\times n}, \quad \mathbf{C}_j = \mathbf{A}_j\mathbf{B}_j,\ j = 0, 1, 2\,. \tag{7}$$

This matrix equation separates into n^2 standard polynomial equations: Let $\mathbf{C}_j := (c_{kl}^{(j)})$, $k, l = 1, 2, \ldots, n$, $j = 0, 1, 2$. Then (7) is equivalent to a system of n^2 equations

$$c_{kl}^{(0)} + c_{kl}^{(1)} z + c_{kl}^{(2)} z^2 = 0, \quad k, l = 1, 2, \ldots, n\,. \tag{8}$$

This allows us to assume, that in the sequel we are looking only for solutions $\mathbf{X} \notin \mathcal{C}$.

Lemma Let $\mathbf{p}$ be a quadratic polynomial defined by the coefficients $\mathbf{A}_i, \mathbf{B}_i \in \mathbb{K}^{n\times n}$, $i = 0, 1, 2$, and let $\mathbf{q}$ be a quadratic polynomial defined by the coefficients $\mathbf{H}^{-1}\mathbf{A}_i\mathbf{H}$, $\mathbf{H}^{-1}\mathbf{B}_i\mathbf{H}$, $i = 0, 1, 2$, for a fixed nonsingular matrix $\mathbf{H} \in \mathbb{K}^{n\times n}$. Then,

$$\mathbf{p}(\mathbf{X}) = \mathbf{0} \quad \Longleftrightarrow \quad \mathbf{q}(\mathbf{H}^{-1}\mathbf{X}\mathbf{H}) = \mathbf{0}. \tag{9}$$

Proof For the quadratic polynomial $\mathbf{q}$, we have

$$\begin{aligned}
\mathbf{q}(\mathbf{X}) =& (\mathbf{H}^{-1}\mathbf{A}_0\mathbf{H})\mathbf{X}^0(\mathbf{H}^{-1}\mathbf{B}_0\mathbf{H}) + \\
&+ (\mathbf{H}^{-1}\mathbf{A}_1\mathbf{H})\mathbf{X}^1(\mathbf{H}^{-1}\mathbf{B}_1\mathbf{H}) + (\mathbf{H}^{-1}\mathbf{A}_2\mathbf{H})\mathbf{X}^2(\mathbf{H}^{-1}\mathbf{B}_2\mathbf{H}) = \\
=& \mathbf{H}^{-1}\left(\mathbf{A}_0(\mathbf{H}\mathbf{X}^0\mathbf{H}^{-1})\mathbf{B}_0 + \mathbf{A}_1(\mathbf{H}\mathbf{X}^1\mathbf{H}^{-1})\mathbf{B}_1 + \mathbf{A}_2(\mathbf{H}\mathbf{X}\mathbf{H}^{-1})^2\mathbf{B}_2\right)\mathbf{H} = \\
=& \mathbf{H}^{-1}\mathbf{p}(\mathbf{H}\mathbf{X}\mathbf{H}^{-1})\mathbf{H},
\end{aligned}$$

which implies that $\mathbf{q}(\mathbf{H}^{-1}\mathbf{X}\mathbf{H}) = \mathbf{H}^{-1}\mathbf{p}(\mathbf{X})\mathbf{H}$. Or in other words $\mathbf{p}(\mathbf{X})$ is similar to $\mathbf{q}(\mathbf{H}^{-1}\mathbf{X}\mathbf{H})$ and (9) follows.

3. Quadratic matrix polynomial of order two

Let us assume that all occurring matrices have the order $n = 2$.

The following recursion was for the first time used by Horn and Johnson, see [2].

Theorem Let $\mathbf{X} \in \mathbb{K}^{2\times 2}$ and let $\chi_{\mathbf{X}}(z) := z^2 - \mathrm{tr}(\mathbf{X})z + \det(\mathbf{X})$ be its characteristic polynomial. Then, there are numbers $\alpha_j,\ \beta_j,\ j \geq 0$, such that

$$\mathbf{X}^j = \alpha_j \mathbf{X} + \beta_j \mathbf{I} \text{ for all } j = 0, 1, \ldots, \tag{10}$$

where

$$\begin{aligned} \alpha_0 &:= 0, \quad \beta_0 := 1, \\ \alpha_{j+1} &:= \mathrm{tr}(\mathbf{X})\alpha_j + \beta_j, \\ \beta_{j+1} &:= -\alpha_j \det(\mathbf{X}), \quad j \geq 0. \end{aligned}$$

In particular,

$$\begin{aligned} \alpha_1 &:= 1, & \beta_1 &:= 0, \\ \alpha_2 &:= \mathrm{tr}(\mathbf{X}), & \beta_2 &:= -\det(\mathbf{X}). \end{aligned}$$

If the coefficients of the characteristic polynomial are real, then also all $\alpha_j,\ \beta_j$ are real for all j.

Proof From the Cayley–Hamilton theorem we have

$$\mathbf{X}^2 = \mathrm{tr}(\mathbf{X})\mathbf{X} - \det(\mathbf{X})\mathbf{I}\,. \tag{11}$$

If we multiply (10) by $\mathbf{X}$ and replace $\mathbf{X}^2$ with the right-hand side of the equation (11), we obtain

$$\begin{aligned} \mathbf{X}^{j+1} &= \alpha_j(\mathrm{tr}(\mathbf{X})\mathbf{X} - \det(\mathbf{X})\mathbf{I}) + \beta_j\mathbf{X} = (\alpha_j \mathrm{tr}(\mathbf{X}) + \beta_j)\mathbf{X} - \alpha_j \det(\mathbf{X})\mathbf{I} = \\ &= \alpha_{j+1}\mathbf{X} + \beta_{j+1}\mathbf{I}\,, \end{aligned}$$

from which the desired recursion in (10) follows. □

The theorem says that a power $\mathbf{X}^j,\ j = 0, 1, \ldots$, of a matrix $\mathbf{X}$ of order 2, regardless of the power j, can always be expressed as a linear combination of the matrix $\mathbf{X}$ and the identity matrix $\mathbf{I}$.

Remark In general, for a matrix $\mathbf{X}$ of order n a power $\mathbf{X}^j$ can always be expressed as an element of the linear hull of matrices $\mathbf{X}^{\nu-1},\ \mathbf{X}^{\nu-2}, \ldots, \mathbf{I}$, where ν is the degree of the minimal polynomial of $\mathbf{X}$, see [2].

Remark The corresponding iteration given by Pogurui and Shapiro in [9] is three term recursion, whereas (10) is a two term recursion. Formally, they differ. In some cases, two term recursions are more stable than the corresponding three term recursions. For an example, see [8].

We apply formula (11). Then our quadratic polynomial $\mathbf{p}(\mathbf{X})$ in (1) has the form

$$\mathbf{p}(\mathbf{X}) = \mathbf{A}_1\mathbf{X}\mathbf{B}_1 + \mathrm{tr}(\mathbf{A})\mathbf{A}_2\mathbf{X}\mathbf{B}_2 + \mathbf{A}_0\mathbf{B}_0 - \det(\mathbf{X})\mathbf{A}_2\mathbf{B}_2\,. \tag{12}$$

Now, let $n \geq 2$ and let $\mathbf{X} \in \mathbb{K}^{n\times n}$, $\mathbf{X} := (x_{j,k})$, $j,k = 1,2,\ldots,n$. We define the operator

$$\mathrm{col} : \mathbb{K}^{n\times n} \to \mathbb{K}^{n^2\times 1}\,,$$

$$\mathrm{col}(\mathbf{X}) := \left(x_{11}, x_{21}, \cdots, x_{n1}, x_{12}, x_{22}, \cdots, x_{n2}, \cdots, x_{1n}, x_{2n}, \cdots, x_{nn}\right)^{\mathrm{T}}\,.$$

In particular for $\mathbf{X} \in \mathbb{K}^{2\times 2}$,

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{pmatrix}, \quad \text{we have} \quad \mathrm{col}(\mathbf{X}) := \left(x_{11}, x_{21}, x_{12}, x_{22}\right)^{\mathrm{T}}\,.$$

Let us note that col is an invertible linear mapping, $\mathrm{col} : \mathbb{K}^{n\times n} \longrightarrow \mathbb{K}^{n^2}$.

Let $\mathbf{A}, \mathbf{B}, \mathbf{X} \in \mathbb{K}^{n\times n}$. Let f be a linear mapping, $f : \mathbb{K}^{n\times n} \longrightarrow \mathbb{K}^{n\times n}$, defined as

$$f(\mathbf{X}) = \mathbf{AXB}\,, \tag{13}$$

represented by the Kronecker product in the form

$$\mathrm{col}(f(\mathbf{X})) = (\mathbf{B}^{\mathrm{T}} \otimes \mathbf{A})\mathrm{col}(\mathbf{X}). \tag{14}$$

Applying col to (12) and using (14), we obtain, see also [1],

$$\mathbf{P}(\mathbf{X}) := \mathrm{col}(\mathbf{p}(\mathbf{X})) = \mathbf{M}(\mathbf{X})\mathrm{col}(\mathbf{X}) + \mathbf{N}(\mathbf{X}), \tag{15}$$

where

$$\mathbf{M}(\mathbf{X}) = (\mathbf{B}_1^{\mathrm{T}} \otimes \mathbf{A}_1) + \mathrm{tr}(\mathbf{X})(\mathbf{B}_2^{\mathrm{T}} \otimes \mathbf{A}_2)\,, \tag{16}$$

$$\mathbf{N}(\mathbf{X}) = \mathrm{col}(\mathbf{A}_0\mathbf{B}_0 - \det(\mathbf{X})\mathbf{A}_2\mathbf{B}_2). \tag{17}$$

Let us remark that both $\mathbf{M}(\mathbf{X})$ and $\mathbf{N}(\mathbf{X})$ depend on $\mathbf{X}$ or more precisely on $\mathrm{tr}(\mathbf{X})$ and $\det(\mathbf{X})$. This means, that the matrices $\mathbf{M}(\mathbf{X})$ and $\mathbf{N}(\mathbf{X})$ are constant on the equivalence class $[\mathbf{X}]$.

Corollary Let $\mathbf{P}(\mathbf{X}) := \mathbf{M}(\mathbf{X})\mathrm{col}(\mathbf{X}) + \mathbf{N}(\mathbf{X}) = \mathbf{0}$. Then all (further) zeros $\mathbf{Y}$ of $\mathbf{P}$ in $[\mathbf{X}]$ can be determined by solving the linear 4×4 system

$$\mathbf{M}(\mathbf{X})\mathrm{col}(\mathbf{Y}) + \mathbf{N}(\mathbf{X}) = \mathbf{0}. \tag{18}$$

If the matrix $\mathbf{M}$ is nonsingular (we delete the arguments), then there is only one zero of $\mathbf{P}$ in $[\mathbf{X}]$. If the matrix $\mathbf{M}$ is the zero matrix, then $\mathbf{N} = \mathbf{0}$ and all matrices in $[\mathbf{X}]$ are zeros of $\mathbf{P}$. If $\mathbf{N} = \mathbf{0}$, then $\mathbf{M}$ is singular.

Since the zeros of $\mathbf{P}$ are eventually all solutions of the linear system (18), we can classify them according to the rank of $\mathbf{M}(\mathbf{X})$.

Definition Let $\mathbf{P}(\mathbf{X}) := \mathbf{M}(\mathbf{X})\mathrm{col}(\mathbf{X}) + \mathbf{N}(\mathbf{X}) = \mathbf{0}$ and let $\mathbf{X} \neq a\mathbf{I}$, $a \in \mathbb{R}$. We say that $\mathbf{X}$ is a zero of rank k if $\mathrm{rank}(\mathbf{M}(\mathbf{X})) = k$, $0 \leq k \leq 4$. A zero of rank 0 will be called spherical zero, a zero of rank 4 will be called isolated zero. If $\mathbf{X} = a\mathbf{I}$, $a \in \mathbb{R}$, the zero will also be called isolated.

Remark In [5], we have shown that for quaternionic polynomials zeros of all ranks, zero to four, exist. For the geometrical meaning of the term "spherical zeros" see [10].

As an example, let us have a special quadratic polynomial

$$\mathbf{p}(\mathbf{X}) := \mathbf{X}^2 + \alpha_1\mathbf{X} + \alpha_0\mathbf{I}, \quad \alpha_1,\ \alpha_0 \in \mathbb{K},\ \alpha_0 \neq 0, \quad \mathbf{X} \in \mathbb{K}^{2\times 2}, \tag{19}$$

which according to (12) can also be written as

$$\mathbf{P}(\mathbf{X}) = (\alpha_1 + \mathrm{tr}(\mathbf{X}))\mathrm{col}(\mathbf{X}) + (\alpha_0 - \det(\mathbf{X})) \begin{pmatrix} 1 \\ 0 \\ 0 \\ 1 \end{pmatrix}$$

or equivalently $\mathbf{p}(\mathbf{X}) = (\alpha_1 + \mathrm{tr}(\mathbf{X}))\mathbf{X} + (\alpha_0 - \det(\mathbf{X}))\mathbf{I}$.

Then, there are two cases for all zeros $\mathbf{X}$ of $\mathbf{p}$:

1. $\alpha_1 + \mathrm{tr}(\mathbf{X}) = \alpha_0 - \det(\mathbf{X}) = 0$,
2. $\alpha_1 + \mathrm{tr}(\mathbf{X}) \neq 0,\ \alpha_0 - \det(\mathbf{X}) \neq 0$.

All matrices which are not a real multiple of the identity matrix $\mathbf{I}$ and obey the equations of the first case are spherical zeros of the given polynomial, they form an equivalence class of spherical zeros. And there are no other spherical zeros. Put

$$\mathbf{X} := \begin{pmatrix} x_1 & x_3 \\ x_2 & x_4 \end{pmatrix}. \tag{20}$$

Then all spherical solutions have the form

$$\mathbf{X} := \begin{pmatrix} -\alpha_1 - x_4 & x_3 \\ x_2 & x_4 \end{pmatrix},$$

where x_2, x_3 are arbitrary and

$$x_4 := -\frac{1}{2}\left(\alpha_1 \pm \sqrt{\alpha_1^2 - 4(\alpha_0 + x_2x_3)}\right).$$

Let the second case be valid. In this case, there may exist other zeros than spherical ones, which are of rank four and which must have the form

$$\mathbf{X} = -\frac{\alpha_0 - \det(\mathbf{X})}{\alpha_1 + \mathrm{tr}(\mathbf{X})}\,\mathbf{I} =: a\mathbf{I}.$$

Since $\det(\mathbf{X}) = a^2,\ \mathrm{tr}(\mathbf{X}) = 2a$, we obtain

$$a := \frac{1}{2}\left(-\alpha_1 \pm \sqrt{\alpha_1^2 - 4\alpha_0}\right).$$

To summarize: Matrix polynomials (19) have always one spherical zero and in addition two isolated zeros (if $\alpha_1^2 - 4\alpha_0 \neq 0$) or one isolated zero (if $\alpha_1^2 - 4\alpha_0 = 0$). All in all, $\mathbf{p}$ has two or three zeros.

Example Consider the following quadratic polynomial with matrices of order $n = 2$:

$$\mathbf{p}(\mathbf{X}) := \mathbf{X}^2 - \mathbf{X} - \mathbf{I}, \tag{21}$$

i.e.

$$\alpha_1 = -1, \quad \alpha_0 = -1, \quad \alpha_1^2 - 4\alpha_0 = 5 \neq 0\,. \tag{22}$$

The matrix polynomial (21) has two isolated zeros

$$\mathbf{X}_1 = \frac{1}{2}\begin{pmatrix} 1+\sqrt{5} & 0 \\ 0 & 1+\sqrt{5} \end{pmatrix}, \qquad \mathbf{X}_2 = \frac{1}{2}\begin{pmatrix} 1-\sqrt{5} & 0 \\ 0 & 1-\sqrt{5} \end{pmatrix}$$

and there is also one spherical zero

$$\mathbf{X}_3 = \begin{pmatrix} 1-x_4 & x_3 \\ x_2 & x_4 \end{pmatrix},$$

where $x_4 = \frac{1}{2}(1 \pm \sqrt{5 - 4x_2x_3})$, x_2, x_3 arbitrary. Let us put, e. g., $x_2 = x_3 = 0$. We obtain

$$x_4^+ = \frac{1}{2}(1+\sqrt{5}), \qquad x_4^- = \frac{1}{2}(1-\sqrt{5})\,.$$

Accordingly, for the spherical root $\mathbf{X}_3$ we have

$$\mathbf{X}_3^+ = \frac{1}{2}\begin{pmatrix} 1-\sqrt{5} & 0 \\ 0 & 1+\sqrt{5} \end{pmatrix}, \qquad \mathbf{X}_3^- = \frac{1}{2}\begin{pmatrix} 1+\sqrt{5} & 0 \\ 0 & 1-\sqrt{5} \end{pmatrix}.$$

It is an easy exercise to show that $\mathbf{X}_3^+$ and $\mathbf{X}_3^-$ belong to the same equivalence class:

$$\mathbf{P}\mathbf{X}_3^+\mathbf{P}^{-1} = \frac{1}{2}\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} 1-\sqrt{5} & 0 \\ 0 & 1+\sqrt{5} \end{pmatrix}\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = \mathbf{X}_3^-\,.$$

Thus the polynomial $\mathbf{p}$ of (21) has altogether three zeros, one spherical and two isolated ones.

Lemma In order that the quadratic polynomial $\mathbf{p}$, defined in (12), has a spherical zero, it is necessary that

$$(\mathbf{B}_1^{\mathrm{T}} \otimes \mathbf{A}_1) = -\mathrm{tr}(\mathbf{X})(\mathbf{B}_2^{\mathrm{T}} \otimes \mathbf{A}_2) \quad \text{and} \quad \mathbf{A}_0\mathbf{B}_0 = -\det(\mathbf{X})\mathbf{A}_2\mathbf{B}_2\,.$$

Proof It follows directly from the definition of spherical zeros. □

Corollary Let $\mathbf{A}, \mathbf{B}$ be arbitrary nonvanishing matrices in $\mathbb{K}^{2\times 2}$. A necessary condition for spherical zeros to exist is that p has the form

$$\mathbf{p}(\mathbf{X}) := \mathbf{A}\mathbf{X}^2\mathbf{B} + \alpha_1\mathbf{A}\mathbf{X}\mathbf{B} + \alpha_0\mathbf{A}\mathbf{B}, \quad \mathbf{A}\mathbf{B} \neq \mathbf{0}, \tag{23}$$

for certain α_0, α_1.
On the other hand, not for each choice of α_0, α_1 does this lead to spherical zeros.

Remark Polynomials with order two matrices of any degree could be treated in a similar way as we did it here.

4. Numerical considerations for finding the zeros

Let us restrict ourselves to quadratic matrix polynomials with $n = 2$.

We apply Newton's method to

$$\mathbf{P}(\mathbf{X}) := \mathrm{col}(\mathbf{p}(\mathbf{X})) = \mathbf{0}\,, \quad \mathbf{X} = (x_{jk}), \quad j,k = 1,2\,,$$

i.e. we solve

$$\mathbf{P}(\mathbf{X}) + \mathbf{P}'(\mathbf{X})\mathbf{S} = \mathbf{0}\,, \quad \mathrm{col}(\mathbf{X}) := \mathrm{col}(\mathbf{X}) + \mathbf{S}\,, \tag{24}$$

where the matrix $\mathbf{P}'$ is the corresponding Jacobi matrix. The Jacobi matrix $\mathbf{P}$' can be found explicitly in a very simple way by using a technique described in [6], without employing partial derivatives.

In the following example, the computations were carried out with MATLAB.

Example We will treat a parameter dependent problem defined by

$$\mathbf{p}(\mathbf{X}(\lambda)) := \mathbf{A}_2\mathbf{X}^2\mathbf{B}_2 + \mathbf{A}_1\mathbf{X}\mathbf{B}_1 + \mathbf{C}(\lambda), \tag{25}$$

where

$$\mathbf{A}_2 := \begin{pmatrix} 1 & 3 \\ 2 & 4 \end{pmatrix}, \qquad \mathbf{B}_2 := \begin{pmatrix} 5 & 10 \\ 4 & 8 \end{pmatrix}, \tag{26}$$

$$\mathbf{A}_1 := \begin{pmatrix} 9 & 11 \\ 10 & 12 \end{pmatrix}, \quad \mathbf{B}_1 := \begin{pmatrix} 13 & 15 \\ 14 & 16 \end{pmatrix}, \tag{27}$$

$$\mathbf{C}(\lambda) := -\begin{pmatrix} 288 & 345 \\ 324 & 394+\lambda \end{pmatrix}, \quad \lambda \in [-1,1]. \tag{28}$$

Note, that $\mathbf{A}_2\mathbf{B}_2 + \mathbf{A}_1\mathbf{B}_1 + \mathbf{C}(\lambda) = \begin{pmatrix} 0 & 0 \\ 0 & -\lambda \end{pmatrix}$. If we denote the zeros by $\mathbf{X}(\lambda)$, we see that $\mathbf{X}(0) = \mathbf{I}$ is one of the zeros. The corresponding matrices $\mathbf{M}$, $\mathbf{N}$ from (16) and (17) for the zero $\mathbf{I}$ are

$$\mathbf{M} = \begin{pmatrix} 127 & 173 & 134 & 178 \\ 150 & 196 & 156 & 200 \\ 155 & 225 & 160 & 224 \\ 190 & 260 & 192 & 256 \end{pmatrix}, \quad \mathbf{N} = \begin{pmatrix} -305 \\ -350 \\ -379 \\ -446 \end{pmatrix}; \quad \mathbf{M}\mathrm{col}(\mathbf{I}) + \mathbf{N} = \mathbf{0} \text{ holds}\,.$$

In this case $\mathrm{rank}(\mathbf{M}) = 4$, i.e. in this case for $\lambda = 0$ the matrix $\mathbf{I}$ is the isolated zero.

However, there is another zero for $\lambda = 0$. For this zero the two matrices are

$$\mathbf{M} = \frac{1}{8}\begin{pmatrix} 931 & 1129 & 1004 & 1220 \\ 1030 & 1228 & 1112 & 1328 \\ 1070 & 1290 & 1144 & 1384 \\ 1180 & 1400 & 1264 & 1504 \end{pmatrix}, \ \mathbf{N} = \frac{1}{8}\begin{pmatrix} -2151 \\ -2358 \\ -2454 \\ -2684 \end{pmatrix}, \quad \begin{matrix} \mathbf{M}\mathrm{col}(\mathbf{I}) + \mathbf{N} = \mathbf{0} \\ \text{holds, too.} \end{matrix}$$

Here, $\mathrm{rank}(\mathbf{M}) = 3$, i.e. $\mathbf{I}$ is the zero of rank 3.

The general solution of $\mathbf{M}\mathrm{col}(\mathbf{X}) + \mathbf{N} = \mathbf{0}$ has the form $\mathrm{col}(\mathbf{X}) = \alpha\mathbf{x}_0 + \mathbf{x}_1$ for all $\alpha \in \mathbb{R}$, where

$$\mathbf{x}_1 = \frac{1}{11}\begin{pmatrix} -1 \\ 12 \\ 11 \\ 0 \end{pmatrix} \quad \text{and} \quad \mathbf{x}_0 = \begin{pmatrix} 0.52124669131568 \\ -0.52124669131568 \\ -0.47780946703938 \\ 0.47780946703938 \end{pmatrix}.$$

Acknowledgements

The research was supported by the German Science Foundation, DFG, under the contract number OP 33/19-1.

References

[1] Aramanovitch, L. I.: Quaternion non-linear filter for estimation of rotating body attitude. Math. Methods Appl. Sci. **18** (1995), 1239–1255.

[2] Horn, R. A. and Johnson, C. R.: *Matrix analysis*. Cambridge University Press, Cambridge, 1991, 561 p.

[3] Horn, R. A. and Johnson, C. R.: *Topics in matrix analysis*. Cambridge University Press, Cambridge, 1991, 607 p.

[4] Janovská, D. and Opfer, G.: A note on the computation of all zeros of simple quaternionic polynomials. SIAM J. Numer. Anal. **48**(244) (2010), 244–256.

[5] Janovská, D. and Opfer, G.: The classification and the computation of the zeros of quaternionic, two-sided polynomials. Numer. Math. **115**(1) (2010), 81–100.

[6] Janovská, D. and Opfer, G.: The algebraic Riccati equation for quaternions. Submitted to Advances in Applied Clifford Algebras, 2013.

[7] Lancaster, P. and Tismenetsky, M.: *The theory of matrices, 2nd ed, with applications*. Academic Press, Orlando, 1985, 570 p.

[8] Laurie, D. P.: Questions related to Gaussian quadrature formulas and two-term recursions. In: W. Gautschi, G. Golub, and G. Opfer (Eds.), *Applications and Computation of Orthogonal Polynomials, International Series of Numerical Mathematics (ISNM)*, vol. 131, pp. 133–144. Birkhäuser, Basel, 1999.

[9] Pogorui, A. and Shapiro, M.: On the structure of the set of zeros of quaternionic polynomials. Complex Variables and Elliptic Functions **49** (2004), 379–389.

[10] Yang, Y. and Qian, T.: On sets of zeroes of Clifford algebra-valued polynomials. Acta Math. Sci., Ser. B, Engl. Ed. **30**(3) (2010) 1004–1012.

Conference Applications of Mathematics 2013
in honor of the 70th birthday of Karel Segeth.
Jan Brandts, Sergey Korotov, Michal Křížek,
Jakub Šístek, and Tomáš Vejchodský (Eds.),
Institute of Mathematics AS CR, Prague 2013

FINITE ELEMENT MODELLING OF FLOW AND TEMPERATURE REGIME IN SHALLOW LAKES

Victor Podsechin[1], Gerald Schernewski[2]

[1] Department of Geophysics, University of Helsinki
P.O. Box 64, FI-00014 Helsinki, Finland
victor.podsechin@gmail.com
[2] Baltic Sea Research Institute Warnemünde (Institut für Ostseeforschung, IOW)
Seestraβe 15, D-18119 Warnemünde, Germany
gerald.schernewski@io-warnemuende.de

Abstract

A two-dimensional depth-averaged flow and temperature model was applied to study the circulation patterns in the Oder (Szczecin) Lagoon located on the border between Germany and Poland. The system of shallow water and temperature evolution equations is discretized with the modified Utnes scheme [4], which is characterized by a semi-decoupling algorithm. The continuity equation is rearranged to Helmholtz equation form. The upwinding Tabata method [3] is used to approximate convective terms. Averaged flow fields under prevailing wind conditions in August were calculated. The temperature variations were also simulated during the flood period in summer 1997. Simulation results are presented and limitations of the model are discussed.

1. Introduction

Wind induced flows in water bodies play an important role in dynamics of aquatic ecosystems. Water temperature, in turn, is one of the important physical parameter that affects limnological, biological processes. Especially in large shallow systems, like the lagoon, significant spatial differences in water temperature are possible. As a result biological and chemical processes may have different intensities in different regions of the lagoon. An accurate diagnosis and prediction of background physical processes, like currents and temperature variations is essential for correct understanding of aquatic ecosystems functioning. In shallow, well mixed water bodies a depth-averaged system of the so-called "shallow water" and temperature equations gives a rather good description of real systems.

2. Materials and methods

The dynamics of flow and temperature in shallow lakes can be described with vertically integrated equations of motion, continuity and heat transfer [1], given in vector form

$$\frac{\partial V}{\partial t} + (\nabla \cdot V)V + f \times V = -g\nabla\zeta + k\,|W|\,W - \frac{gn^2\,|V|\,V}{H^{4/3}} + \nu\Delta V, \tag{1}$$

$$\frac{\partial \zeta}{\partial t} + \nabla \cdot (HV) = 0, \tag{2}$$

$$\frac{\partial HT}{\partial t} + (\nabla \cdot V)HV = \mathrm{div}(\nu_T H \nabla T) + \frac{\alpha(T - T_a)}{\rho_0 C_p}. \tag{3}$$

In above, $V = (u, v)$ is a depth-averaged velocity vector, $H(x, y, t) = h(x, y) + \zeta(x, y, t)$ is the total water depth, ζ is the water surface elevation above a horizontal datum, h is the depth below datum, $W = (W_x, W_y)$ is the wind velocity vector, f is the Coriolis parameter, g is the acceleration due to gravity, n is the Manning roughness coefficient, k is the wind resistance coefficient, ν is the horizontal eddy viscosity, $\rho_0 = 10^3\,\mathrm{kg\,m^{-3}}$ is water density, $C_p = 4.18710^{-3}\,\mathrm{J\,kg^{-1}\ ^\circ C^{-1}}$ is a specific heat capacity of water, T is water temperature (°C), T_a is air temperature (°C), α is the bulk heat exchange coefficient ($\mathrm{W\ m^{-2}\ ^\circ C^{-1}}$). It is estimated using an empirical dependence on wind speed W, ($\mathrm{ms^{-1}}$), see [1]:

$$\alpha = 5.7 + 3.8W. \tag{4}$$

The boundary conditions of system (1)–(2) are as follows (cf. [1]):

$$\text{land boundary:} \quad V|_{B_1} = 0, \tag{5}$$

$$\text{liquid boundary:} \quad \zeta|_{B_2} = \zeta|_B(t). \tag{6}$$

When $\nu = 0$ the governing equations (1)–(2) constitute a system of quasi-linear hyperbolic partial differential equations. In this case the non-slip boundary condition (5) is replaced with the slip one

$$V \cdot n|_{B_1} = 0, \tag{7}$$

where n is the unit vector normal to the boundary of the solution domain. Zero initial conditions

$$V = \zeta|_{t=0} = 0 \tag{8}$$

are frequently used in practical applications to start the time integration.

For temperature equation (3) the no-flux boundary condition was applied along the solid boundary. The observations of time-varying inflowing water temperature in the river Oder altogether with estimated values of inflowing mean cross-section

velocity (T. Neumann, pers. com.) and time-series of wind in the Pomeranian Bight were used to drive the combined flow and temperature model.

Using a time-splitting algorithm [4] the momentum equation (1) is discretized as follows:

$$\frac{V^* - V^m}{\tau} + (\nabla \cdot V^m)V^* + f \times V^m = k \left|W^{m+1}\right| W^{m+1} - \frac{gn^2 \left|V^m\right| V^*}{H^{4/3}} + \nu \triangle V^*, \quad (9)$$

$$\frac{V^{m+1} - V^*}{\tau} = -g\nabla\zeta^{m+1}. \quad (10)$$

When the time derivative is approximated with the forward difference the continuity equation takes the form:

$$\zeta^{m+1} = \zeta^m - \tau\nabla \cdot HV^{m+1}. \quad (11)$$

Multiplying equation (10) by H, taking the divergence and substituting it in place of $\nabla \cdot HV^{m+1}$ into (11), the Helmholtz approximation of the semi-implicit continuity equation is obtained

$$[1 - \tau^2 g\nabla \cdot H\nabla]\zeta^{m+1} = \zeta^m - \tau\nabla \cdot HV^*. \quad (12)$$

The calculations are organized in the following way: an intermediate velocity V^* is calculated by (9), the water level elevation ζ^{m+1} is predicted by (12) and the corrected velocity V^{m+1} is obtained from (10).

The space domain Ω is divided into a sum of linear triangular elements. Unknown variables are approximated as series of basis functions

$$V \approx \sum_{j=1}^{N} V_j \cdot \varphi_j, \quad \zeta \approx \sum_{j=1}^{N} \zeta_j \cdot \varphi_j, \quad T \approx \sum_{j=1}^{N} T_j \cdot \varphi_j, \quad (13)$$

where N is the number of mesh nodes and φ_j are the global basis functions. After substituting the decompositions (13) to (3), (9), (10), and (12), multiplying according to the Galerkin method by the weight-functions φ_i^T, integrating over Ω and applying the Gauss theorem for the second-order terms, the system of linear algebraic equations is derived

$$(M + \tau(CONV + D + gn^2F))V^* = M(V^m - \tau(f \times V^m - k \left|W^{m+1}\right| W^{m+1})) + \tau \int_B \varphi_i \nu \frac{\partial V}{\partial n} dB, \quad (14)$$

$$(M + g\tau^2 K)\zeta^{m+1} = M\zeta^m - \tau G(HV)^* + g\tau^2 \int_B \varphi_i H \frac{\partial \zeta}{\partial n} dB, \quad (15)$$

$$MV^{m+1} = MV^m - \tau g G\zeta^{m+1}, \quad (16)$$

$$(M + \tau(CONV + D^*))HT^{m+1} = MHT^m + \tau\frac{\alpha}{\rho C_p} M(T - T_a)^m, \quad (17)$$

where the global matrices are compiled as follows:

$$M = \int_\Omega \varphi_i \cdot \varphi_j^T d\Omega, \quad D = \int_\Omega \nu \nabla\varphi \cdot \varphi_j^T d\Omega, \quad D^* = \int_\Omega \nu_T \nabla\varphi_i \cdot \nabla\varphi_j^T d\Omega, \tag{18}$$

$$F = \int_\Omega \varphi_i \cdot \varphi_j^T \cdot \frac{|V^m|}{H^{4/3}} d\Omega, \quad G = \int_\Omega \nabla\varphi_i \cdot \varphi_j^T d\Omega, \quad K = \int_\Omega \nabla\varphi_i \cdot H \nabla\varphi_j^T d\Omega. \tag{19}$$

The CONV denotes the global convective matrix, modified according to the upwinding Tabata scheme [3]. The systems of linear equations (14)–(17) are solved sequentially using a direct Gaussian elimination method [1].

3. Numerical results

In the first step a linear triangular mesh of 2240 nodes and 3845 elements covering the Oder lagoon was generated (Fig. 1) and linked to depth information. This grid density with a slightly simplified bathymetry was chosen to keep the computation time reasonable. During the Oder flood in summer 1997, the total simulation period of 20 days with simulation time steps of 5 minutes was used for flow field calculations. The Manning roughness coefficient of $0.015\,\mathrm{m}^{-1/3}\mathrm{s}$, the horizontal diffusion coefficient equal to $0.01\,\mathrm{m}^2\mathrm{s}^{-1}$, and the Coriolis parameter of $1.174 \cdot 10^{-4}\,\mathrm{s}^{-1}$ were applied.

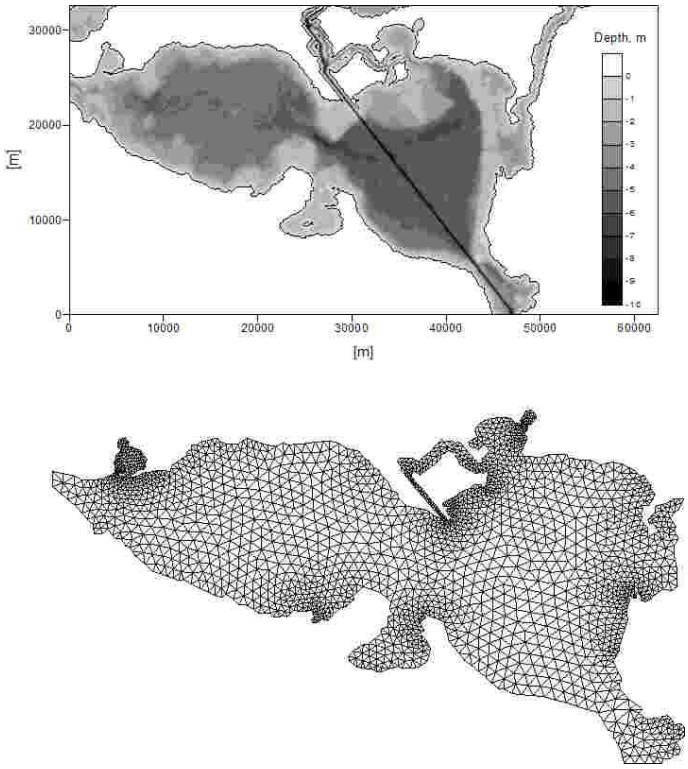


Figure 1: Bathymetric map and triangular mesh of the Oder Lagoon.

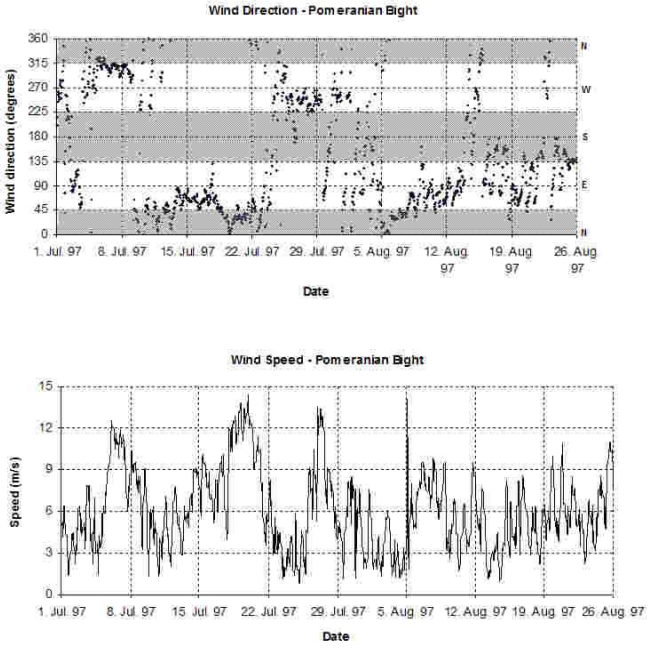


Figure 2: Wind direction and wind speed in the Pomeranian Bight during the Oder flood in July and August 1997.

In steady-state simulations a constant and spatial uniform wind field as well as constant water discharge was used. Wind data was available from an automatic recording station in the Oder Bight (Fig. 2) and for several periods from the centre of the lagoon, too. Wind speed from Oder Bight was adapted to the situation in the lagoon by multiplication with the empirical derived factor of 0.46 (Spiegel pers. com). According to Mohrholz and Lass [2] the discharge from the lagoon into the Baltic Sea varied depending on the prevailing wind direction (13–19 % Peene Strait, 8–14 % Dziwina Strait, 73 % Swina Strait) but was kept constant with time. Intrusion of seawater was neglected.

Depending on the general atmospheric situation wind from east and west is dominating during late summer. The flow field for these two typical wind situations were simulated assuming a river discharge into the lagoon of about $300\,\mathrm{m^3 s^{-1}}$ (Fig. 3). Under common late summer discharge situations the flow field is to a significant amount determined by wind conditions. In general, the simulated flow velocities are low and water masses need about 50 days to pass the lagoon and to enter the western bay (Kleines Haff). The flow in the deep channel, crossing the lagoon, shows some special behaviour, with reduced flow velocities. The water transport through the eastern and western shallower regions is significant faster. In the central parts of the Kleines Haff the flow velocities are low. But close to both shores a coastal jet with increased flow and transport speed occurs. Under these conditions the shores are to a higher degree affected by Oder water than other areas of the bay.

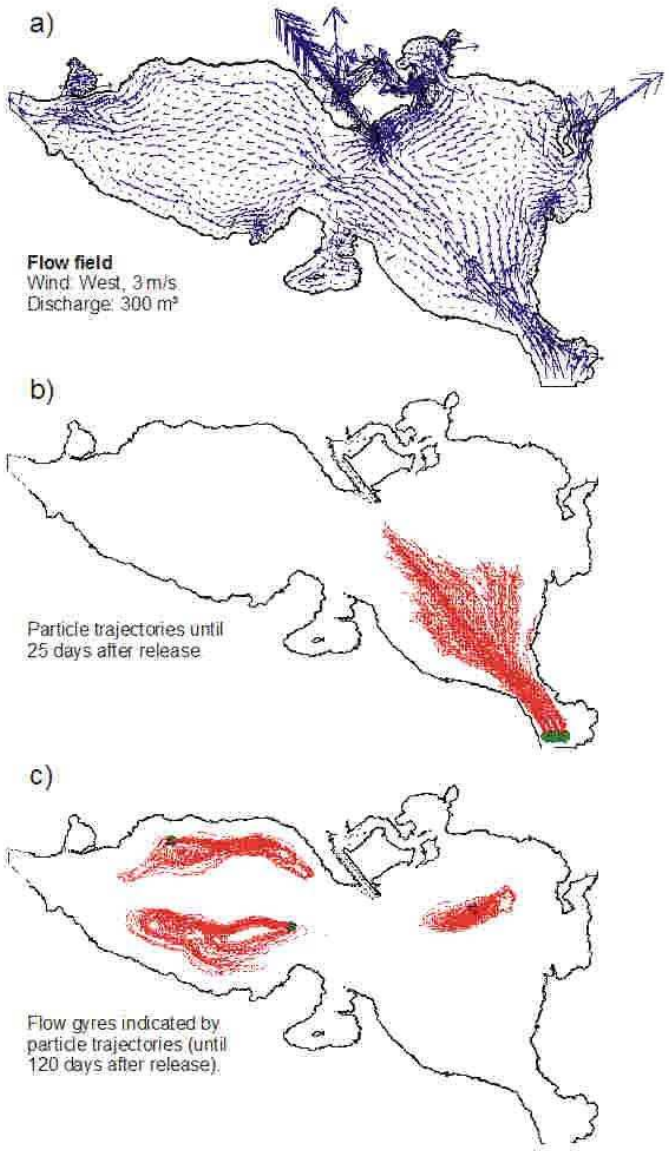


Figure 3: Average two-dimensional flow field under typical August weather conditions. An average Oder discharge into the lagoon of $300\,\mathrm{m^3 s^{-1}}$ and a mean wind of $3\,\mathrm{ms^{-1}}$ from west was applied. The trajectories of passive particles moving with currents are shown 25 and 120 days after release in different locations.

Despite these coastal jets water needs 70 days to pass this bay and 120 days to pass the whole lagoon from river mouth to the Peene strait. This slow transport and water exchange allow independent local ecological dynamic, like algal bloom in the bay. Under similar water discharge but west wind conditions a quite different flow pattern prevails (Fig. 3). The simulations show a relatively fast transport of 26 days across the lagoon via the deep channel. All shallower parts on the right- and left-hand

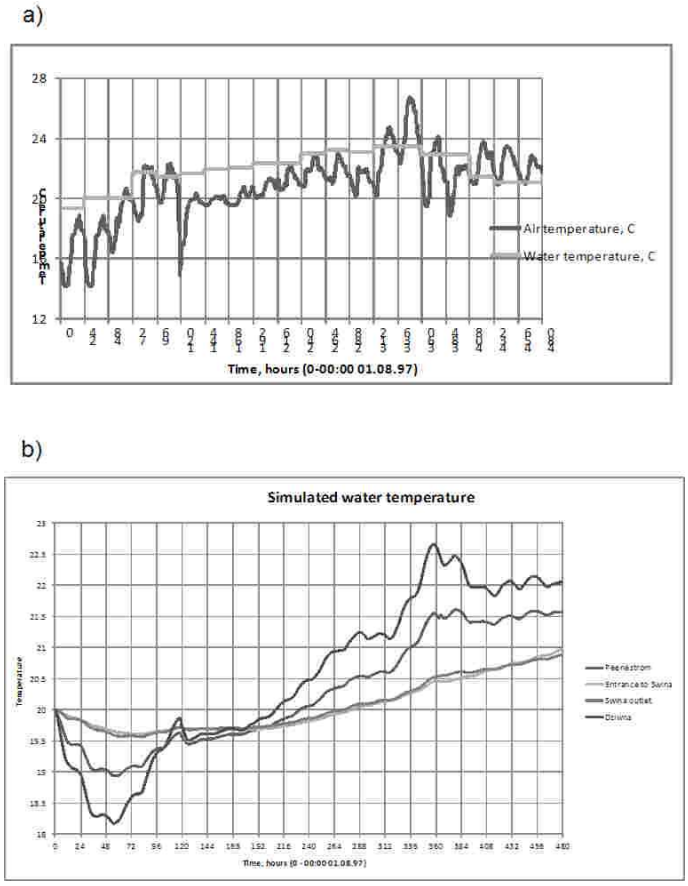


Figure 4: The Oder Lagoon between August 1 and August 20, 1997. a) Measured air temperature and simulated water temperature, b) Simulated water temperatures in different areas of the lagoon.

side show much lower current speeds. In several areas in the Kleines Haff as well as in the eastern part pronounced large eddies occur and limit the water exchange (Fig. 3c). Interpreting the figures, one has to keep in mind that the simulations yield depth-averaged flow velocities, that constant wind and discharge were applied and seawater intrusions were disregarded. These simplifications limit general statements.

Results of dynamic simulation under the time varying meteorological conditions show that the temperature regime of the Oder lagoon is strongly influenced by air temperature fluctuations. The dependence of the bulk heat transfer coefficient on wind speed (4) additionally accelerates the response of depth-averaged water temperature to changes in atmospheric conditions. This is clearly reflected in Fig. 4b showing simulated time-series of water temperature in different parts of the lagoon. Shallow near coastal areas exhibit higher and faster changes than deeper parts located in the vicinity of the navigational channel. During night the spatial temperature differences in the lagoon are significant lower than during day. In August 1997,

for example the spatial temperature difference exceeded 2°C during the day and was below 1°C in the night. Due to the flood, the flow velocity in the lagoon was much higher compared to other years and spatial water exchange increased. Under common summer conditions we can expect significant higher spatial temperature differences.

Acknowledgements

This work was supported by the DFG (Deutsche Forschungsgemeinschaft), Academy of Finland (von Humboldt Fund), Finnish Ministry of Environment, and Grant MTM2011–24766 of the MICINN (Spain).

References

[1] Benque, J.-P., Haugel, A., and Viollet, P.-L.: *Engineering applications of computational hydraulics*. Vol. II, Pitman Advanced Publishing Program, 1982.

[2] Mohrholz, V. and Lass U.: Transports between Oderhaff and Pomeranian Bight a simple barotrophic box model. Dt. Hydrogr. Z. **50** (1998), 371–383.

[3] Tabata, M.: A finite element approximation corresponding to the upwinding finite differencing. Mem. Numer. Math. **4** (1977), 46–63.

[4] Utnes, T.: A finite element solution of the shallow-water wave equations. Appl. Math. Model. **14** (1990), 20–29.

Conference Applications of Mathematics 2013
in honor of the 70th birthday of Karel Segeth.
Jan Brandts, Sergey Korotov, Michal Křížek,
Jakub Šístek, and Tomáš Vejchodský (Eds.),
Institute of Mathematics AS CR, Prague 2013

FUZZY SETS AND SMALL SYSTEMS

Jaroslav Považan[1], Beloslav Riečan[1,2]

[1]Department of Mathematics, Faculty of Natural Sciences, Matej Bel University
Tajovského 40, 974 01 Banská Bystrica, Slovakia
beloslav.riecan@umb.sk
[2]Mathematical Institute, Slovak Academy of Sciences
Štefánikova 49, 814 73 Bratislava, Slovakia

Abstract

Independently with [7] a corresponding fuzzy approach has been developed in [3–5] with applications in measure theory. One of the results the Egoroff theorem has been proved in an abstract form. In [1] a necessary and sufficient condition for holding the Egoroff theorem was presented in the case of a space with a monotone measure. By the help of [2] and [6] we prove a variant of the Egoroff theorem stated in [4].

1. Introduction

In [7] the notion of a fuzzy subset A of a space X has been defined as a mapping $A : X \to [0, 1]$. Especially, if $A : X \to \{0, 1\}$, then A can be identified with a classical set $B \subset X$ by the help of the equality $A = \chi_B$.

Almost at the same time the notion of a set of small measure has been characterized in [3–5] using a sequence $(\mathcal{N}_n)_{n=1}^{\infty}$ of subfamilies of a σ-algebra $\mathcal{S} \subset 2^X$ satisfying the following properties:

(i) $\emptyset \in \mathcal{N}_n, \mathcal{N}_{n+1} \subset \mathcal{N}_n$ for every $n \in \mathbb{N}$,

(ii) if $A \in \mathcal{N}_n, B \in \mathcal{S}$ and $B \subset A$, then $B \in \mathcal{N}_n$,

(iii) if $A, B, C \in \mathcal{N}_n$, then $A \cup B \cup C \in \mathcal{N}_{n-1}$,

(iv) if $A_i \supset A_{i+1}\, (i = 1, 2, \cdots)$ and $\bigcap_i A_i = \emptyset$, then to every $n \in \mathbb{N}$ there is i such that $A_i \in \mathcal{N}_n$.

The classical Egoroff theorem states that if a sequence $(f_n)_n$ of real measurable functions converges to a measurable function f almost everywhere, then it converges almost uniformly, i.e. $\forall \varepsilon > 0\ \exists A \in \mathcal{A}$ such that $\mu(A) < \varepsilon$ and $(f_n)_n$ converges uniformly to f on $X - A$.

Definition. We say that a sequence $(f_n)_n$ converges to f almost everywhere, if $\{x \in X; f_n(x)$ does not converge to $f(x)\} \in \mathcal{N}_n$ for every n. We say that $(f_n)_n$ converges to f almost uniformly, if for any $n \in \mathbb{N}$ there exists $A \in \mathcal{N}_n$ such that (f_n) converges uniformly to f on $X - A$.

2. Egoroff theorem

Theorem. Let $(\mathcal{N}_n)_n$ be a small system of subfamilies of a measurable space $(X, \mathcal{S})$. Let $(f_n)_n$ converges to f almost everywhere. Then $(f_n)_n$ converges to f almost uniformly.

Proof. First we use a result from [6]: If $(\mathcal{N}_n)_n$ satisfies (i)–(iv), then there exists a monotone continuous function $\mu : \mathcal{S} \to [0, 1]$ such that

$$\mathcal{N}_n = \{A \in \mathcal{S}; \mu(A) < 3^{-n}\},$$

$n = 1, 2, 3, \ldots$ In [1] the following theorem has been proved: A monotone function $\mu : \mathcal{S} \to [0, 1]$ satisfies the Egoroff theorem if and only if it satisfies the following condition (E):

For every double sequence $\left\{E_n^{(m)}\right\}$ of measurable sets which satisfies

$$E_n^{(m)} \searrow E^{(m)} \, (n \to \infty), \quad \mu\left(\bigcup_{m=1}^{\infty} E^{(m)}\right) = 0$$

there exist increasing sequences $\{n_i\}_{i=1}^{\infty}$ and $\{m_i\}_{i=1}^{\infty}$ of natural numbers such that

$$\lim_{k\to\infty} \mu\left(\bigcup_{i=k}^{\infty} E_{n_i}^{(m_i)}\right) = 0.$$

We are going to prove that the monotone continuous set function μ satisfies condition (E). Let $\left\{E_n^{(m)}\right\}$ is double sequence of measurable sets for which

$$E_n^{(m)} \searrow E^{(m)} \, (n \to \infty), \quad \mu\left(\bigcup_{m=1}^{\infty} E^{(m)}\right) = 0.$$

From the monotonicity it follows that

$$0 = \mu(\emptyset) \leq \mu\big(E^{(m_0)}\big) \leq \mu\left(\bigcup_{m=1}^{\infty} E^{(m)}\right) = 0.$$

We have proven that $\mu\big(E^{(m)}\big) = 0$ for arbitrary m. From this it follows that there is a natural number n_1 for which

$$\mu\big(E_{n_1}^{(1)}\big) \leq \frac{1}{3}.$$

Similarly there is a number $n_2 > n_1$ for which

$$\mu\big(E_{n_2}^{(2)}\big) \leq \frac{1}{3^2},$$

etc. Putting $m_i = i$, we get

$$\mu\left(\bigcup_{i=k}^{\infty} E_{n_i}^{(m_i)}\right) \leq \sum_{i=k}^{\infty} \frac{1}{3^i} = \frac{\frac{1}{3^k}}{1-\frac{1}{3}} = \frac{1}{2\cdot 3^{k-1}}.$$

From this it follows that

$$\lim_{k\to\infty} \mu\left(\bigcup_{i=k}^{\infty} E_{n_i}^{(m_i)}\right) = 0.$$

□

3. Conclusion

We presented a new proof of the Egoroff theorem for small systems [4]. It follows from a representation theorem in [6] and the Egoroff theorem for monotone measures in [2].

Acknowledgements

This paper was supported by Grant VEGA 1/0621/11.

References

[1] Li, J.: Convergence theorems in monotone measure theory. In: R. Mesiar et al. (Eds.), *Non-Classical Measures and Integrals*, pp. 88–91, 34th Linz Seminar on Fuzzy sets Theory, 2013.

[2] Li, J. and Yasuda, M.: Egoroff's theorems on monotone non-additive measure space. Fuzzy Sets and Systems **153** (2005), 71–78.

[3] Neubrunn, T.: On abstract formulation of absolute continuity and dominancy. Math. Čas. **19** (1969), 202–215.

[4] Riečan, B.: Abstract formulation of some theorems of measure theory. Math. Čas. **16** (1966), 268–273.

[5] Riečan, B.: Abstract formulation of some theorems of measure theory II. Math. Čas. **19** (1969), 138–141.

[6] Riečan, B. and Neubrunn, T.: *Integral, measure, and ordering.* Kluwer, Dordrecht, 1997.

[7] Zadeh, L. A.: Fuzzy sets. Inform. and Control **8** (1965), 336–358.

Conference Applications of Mathematics 2013
in honor of the 70th birthday of Karel Segeth.
Jan Brandts, Sergey Korotov, Michal Křížek,
Jakub Šístek, and Tomáš Vejchodský (Eds.),
Institute of Mathematics AS CR, Prague 2013

RIEMANN SOLUTION FOR HYPERBOLIC EQUATIONS WITH DISCONTINUOUS COEFFICIENTS

L. Remaki

BCAM – Basque Centre for Applied Mathematics
Mazarredo 14, 48009 Bilbao, Basque Country, Spain
lremaki@bcamath.org

Abstract

This paper deals with a Riemann solution for scalar hyperbolic equations with discontinuous coefficients. In many numerical schemes of Godunov type in fluid dynamics , electromagnetic and so on, usually hyperbolic problems are solved to estimate fluxes. The exact solution is generally difficult to obtain, but good approximations are provided in many situations like Roe and HLLC Riemann solvers in fluid. However all these solvers assumes that the acoustic waves speeds are continuous which is not true as we will show in this paper. A new Riemann solver is then proposed based on previous work of the author and an application to a gas-particle model for a 90 degree curved bend is performed.

1. Introduction

In many numerical methods such as the dual-mesh finite volume, DG methods, estimation of convective numerical fluxes at the (dual) interfaces is required. The accuracy of the method depends on the accuracy of the flux estimation. The most popular methods consists of using a Riemann approximation solver, because of its physical meaning. This approach was proposed first by Godunov [3] and then many Riemann solver approximations where developed. The most popular being Roe solver [8, 9] where the Jacobian matrix is averaged in such way that hyperbolicity, consistency with the exact Jacobian and conservation across discontinuities are fulfilled. For fluid application this solver has been modified [1, 2] to overcome the shortcoming for low-density flows. HLL Riemann solver [4] proposed to solve for the original flux, the major drawback of this solver due to the space averaging process, is that contact discontinuities, shear waves and material interfaces are not captured. To remedy to this problem, the HLLC solver [10] was proposed, by adding the missing wave to the structure. However, all these methods assume that the waves speeds are continuous through the interfaces (intercellular in the case of finite volume dual mesh) by applying diverse averaging process. This is not true in reality; typical situations are recirculation for turbulent flows and transitions from subsonic to supersonic for transonic regimes. To remedy to this situation and as a first step

a Riemann solver of scalar hyperbolic linear equation with discontinuous coefficient is developed, this is based on a first idea developed in [7]. This solver takes into account the discontinuities of waves speeds and shows physical behaviours that are missed by the existing solvers. Numerical proof of the proposed solution is provided and an application to a gas-particle model with a validation against experimental results for a 90^0 curved bend described in [5] is performed.

2. *Riemann* solution for hyperbolic equation with discontinuous coefficient

Consider the following scalar linear hyperbolic equation with discontinuous coefficient,

$$\begin{aligned} \tfrac{\partial}{\partial t}\varphi + \beta(x)\tfrac{\partial}{\partial x}\varphi = 0 \qquad & \text{on } \Omega \times [0,T]\,, \\ \varphi(0,x) = \varphi^0 = \begin{cases} \varphi_L & \text{if} \quad x < 0\,, \\ \varphi_R & \text{if} \quad x > 0\,, \end{cases} & \\ \beta(x) = \begin{cases} \beta_L & \text{if} \quad x < 0\,, \\ \beta_R & \text{if} \quad x > 0\,, \end{cases} & \end{aligned} \tag{1}$$

In this equation, the acoustic wave speed β is discontinuous which is not taken into account in the existing *Riemann* solvers, where acoustic waves speeds are assumed to be continuous in the vicinity of the origin. To build a solution let us first analyze the following different situations.

Case 1: $\beta_L > 0$ and $\beta_R > 0$ we have propagation of the discontinuity (of initial condition) to the right and we do not need to consider what happening within the fan defined by the two acoustic waves, because they will catch up if $\beta_L > \beta_R$ and if $\beta_L < \beta_R$ an expansion will appear.

Case 2: $\beta_L < 0$ and $\beta_R < 0$ similar the previous case with a propagation of the discontinuity to the left.

Case 3: $\beta_L < 0$ and $\beta_R > 0$ we have propagation of the discontinuity to the left and the right simultaneously, and we need to determine what happened within the fan defined by the two acoustics waves. We assume that a constant state appears and its expression will be given below.

Case 4: $\beta_L > 0$ and $\beta_R < 0$ in this case we have opposite acoustic wave speeds and then the discontinuity will remain blocked, which means there is no propagation.

Based on the above analysis, the *Riemann* solution of problem (4) is given by

$$\varphi(x,t) = \begin{cases} \varphi_L & \text{if} \quad \beta_L > 0 \quad \text{and} \quad \beta_R > 0, \\ \lambda & \text{if} \quad \beta_L < 0 \quad \text{and} \quad \beta_R > 0, \\ \varphi_R & \text{if} \quad \beta_L < 0 \quad \text{and} \quad \beta_R < 0, \\ \varphi^0 & \text{if} \quad \beta_L > 0 \quad \text{and} \quad \beta_R < 0, \end{cases} \tag{2}$$

where the expression of the constant λ is given by

$$\lambda = \frac{\frac{1}{\beta_L}\varphi_L + \frac{1}{\beta_R}\varphi_R}{\frac{1}{\beta_L} + \frac{1}{\beta_R}} \tag{3}$$

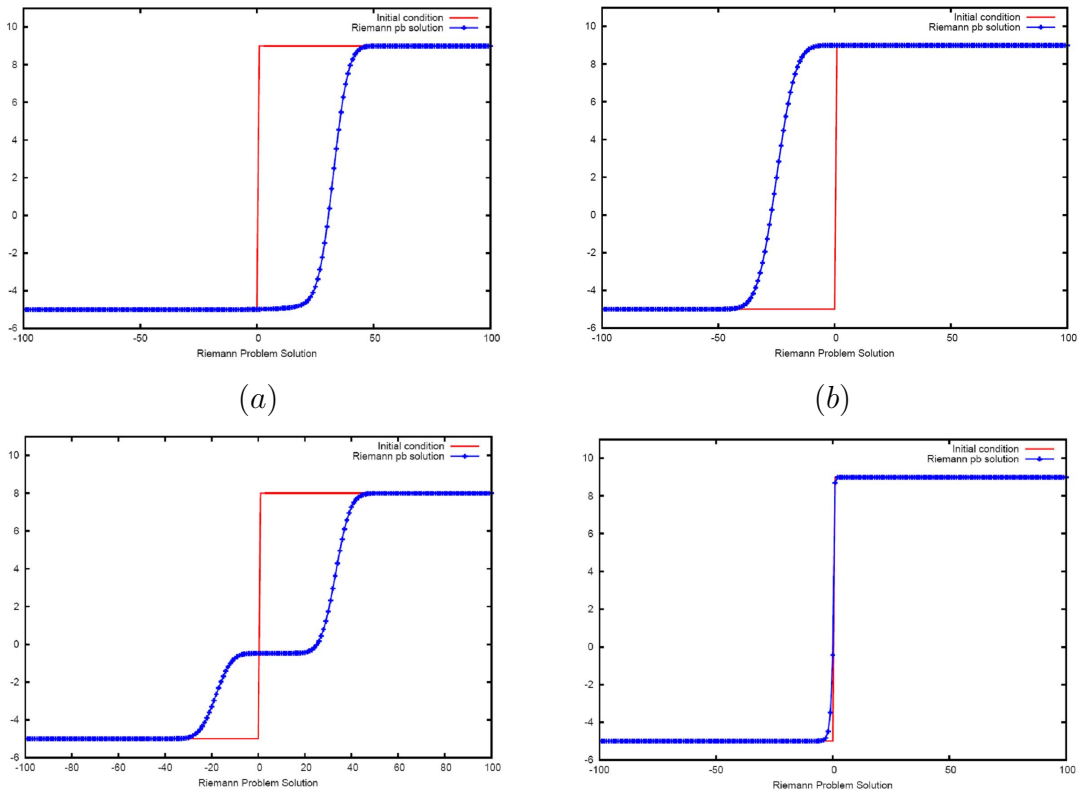


Figure 1: Initial condition and Riemann solution after 100 time iterations: a) Case 1, b) Case 2, c) Case 3, d) Case 4.

To prove solution (2) and formula (3) at least numerically, the *Riemann* problem (1) is solved using a centred finite volume scheme stabilized with a first order artificial viscosity, which is equivalent in this case to a finite difference scheme. Several initial conditions φ^0 and acoustic wave β values are tested. All tests confirm the proposed solution, two examples are shown in Figures 1 and 3. We can see in particular the solutions corresponding to cases 3 and 4, it is clear that they could not be obtained if the coefficient β is averaged as in the existing *Riemann* solvers. Figures 2 and 4 show the perfect agreement of the proposed analytical expression of λ with the predicted numerical value.

3. Application to gas-particle model discretization

In this section the proposed Riemann solver is applied to discretize part of gas-particle model that describes a motion of particles under the effect fluid drag forces. A 90 degree bend is then simulated and results are compared to experimental data. First let us recall equations governing three-dimensional unsteady viscous compressible flow coupled with a particle motion equation in Eulerian modelisation:

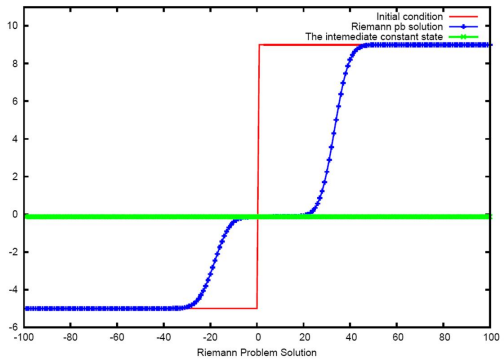


Figure 2: Initial condition and Riemann solution corresponding to Case 3 and the analytical value of λ.

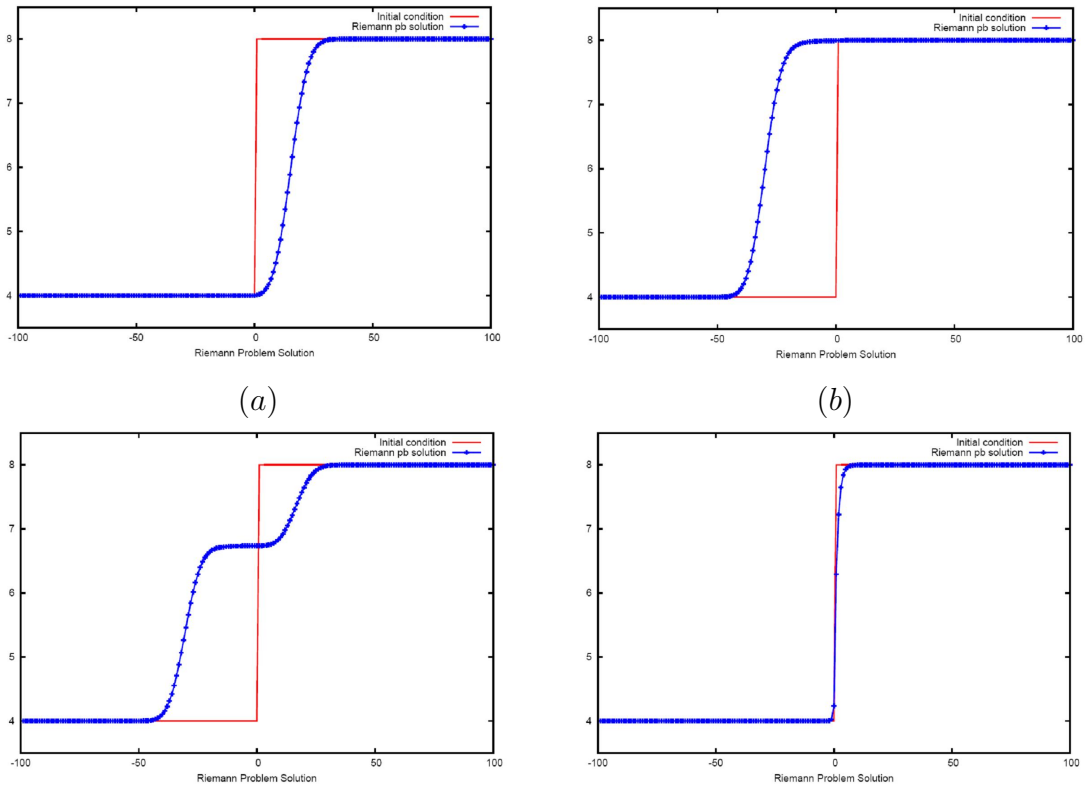


Figure 3: Initial condition and Riemann solution corresponding to Case 3 and the analytical value of λ.

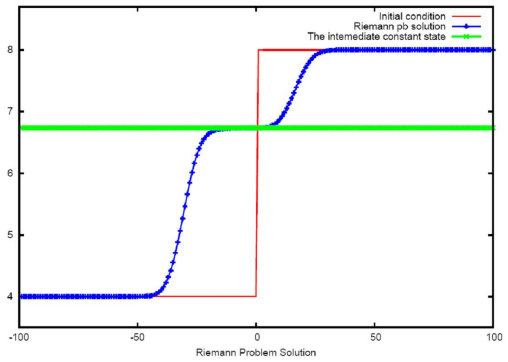


Figure 4: (a) Riemann Solution: a) Initial condition, b) Solution after 100 time iterations.

Fluid phase:

$$\frac{\partial}{\partial t}(\phi_g \rho_g) + \nabla(\phi_g \rho_g U_g) = 0 \qquad \text{on } \Omega \times [0,T] \ , \tag{4}$$

$$\frac{\partial}{\partial t}(\phi_g \rho_g U_g) + \nabla(\phi_g \rho_g U_g \otimes U_g) = -\nabla P_g + \nabla \tau_g - d_{fac} \frac{1}{\tau_p} \phi_g (U_g - U_p) \qquad \text{on } \Omega \times [0,T] \ . \tag{5}$$

Gas phase:

$$\frac{\partial}{\partial t}(\phi_p \rho_p) + \nabla(\phi_p \rho_p U_p) = 0 \qquad \text{on } \Omega \times [0,T] \tag{6}$$

$$\frac{\partial}{\partial t}(\phi_p \rho_p U_p) + \nabla(\phi_p \rho_p U_p \otimes U_p) = -\frac{\phi_p}{\rho_p} \nabla P_g + d_{fac} \frac{1}{\tau_p} \phi_p (U_g - U_p) + \phi_p (1 - \frac{\rho_g}{\rho_p}) \qquad \text{on } \Omega \times [0,T] \ , \tag{7}$$

where ϕ_g and ϕ_p are the gas and particle volume fraction satisfying the conservation condition $\phi_g + \phi_p = 1$, $d_{fac} = \begin{cases} 1 + 0.15 R_{e_0}^{0.687} & \text{if } R_{e_0} < 1000 \\ 0 \quad \text{else} \end{cases}$ is the drag coefficient, $R_{e_0} = \frac{D_p |U_p - U_g|}{\nu_g}$ is the particle Reynolds number, $\tau_p = \frac{\rho_p D_p^2}{18 \mu_g}$ is the particle response time and $\overrightarrow{g}$ is the gravity.

In this model only Drag and gravity forces are considered in the particle-gas interaction.

3.1. Numerical discretization

A finite volume method is used to discretize the overall gas-particle model. In the solid phase, the particle fraction volume variable vanishes, where the sand is absent which results in dividing by zero when computing the velocity vector in equation (7). This leads obviously to a severe instability of the scheme. To avoid this situation the non-conservative form of the momentum equation is considered, while the conservative form of the volume fraction is kept. The momentum equation takes the form:

$$\frac{\partial}{\partial t}(U_p) + \nabla(U_p \otimes U_p) = -\frac{1}{\rho_p}\nabla P_g + d_{fac}\frac{1}{\tau_p}(U_g - U_p) + (1 - \frac{\rho_g}{\rho_p}) \quad \text{on } \Omega \times [0, T]. \tag{8}$$

The inviscid fluxes of the fluid phase are estimated using a second order HLLC *Riemann* solver and the limiter developed in [6]. For the solid phase a centered scheme is used for the non-conservative momentum equations of the solid phase, and it is stabilized by adding a first order artificial viscosity of the form

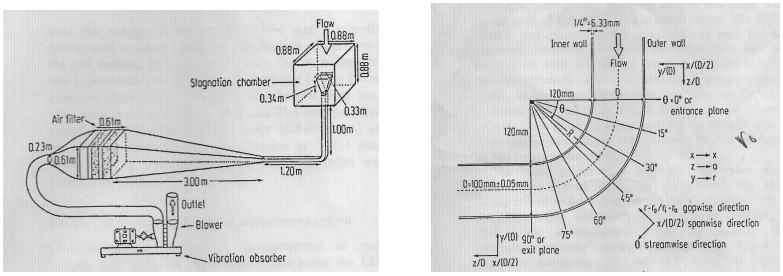
$$\triangle Q_i = \sum_{J \in N_I} \lambda_{IJ} \left(Q_J - Q_I\right). \tag{9}$$

With $\lambda_{IJ} = \frac{1}{1+\tan(\theta_{IJ})^2}$, and θ_{IJ} is the angle between the normal to the surface η_I and the velocity vector U_I. This allows diffusion to act mostly in the flow direction while minimizing the cross wind effects similar to streamline diffusion methods.

To estimate the volume fraction flux in the solid phase, we need to solve a *Riemann* problem of a hyperbolic equation with discontinuous coefficient of the type $\frac{\partial}{\partial t}(\phi_p) + q_{IJ}\frac{\partial}{\partial s}(\phi_p) = 0$, where $q_{IJ} = U_p^I \eta_{IJ}$ and U_p^I is the particle velocity at node I and η_{IJ} is the normal to the surface separating dual cells associated with nodes I and J. To take into account of such discontinuities, the proposed *Riemann* solver is used.

3.2. Validation

The numerical model is validated against experimental results of a 90° bend test case described in [5]. This test case was selected first because of the availability of experimental data and then for the wide use of curved ducts in industrial applications such as air-coal flows in coal combustion equipments, coal liquefaction-gasification pipe systems, gas-particle flows in turbo machinery, and contaminant particle flows in ventilation ducts. The apparatus and geometry of the test are shown on figure 5 scanned from reference [5]. The 90° duct has a square cross-section of $D = 0.1\,\text{m}$ and upstream and downstream duct lengths are 1 m and 1.2 m, respectively. Glass spherical particles with a material density of $2990\,\text{kg/m}^3$ and diameter size of $50\,\text{kg/m}^3$ are used. The inlet fluid and particles velocity is set to $52.19\,\text{m/s}$, for more details see [5]. For the numerical simulation a hybrid mesh is used as shown in Figure 6. The mesh contains 766614 tetrachordal elements and 1229952 prisms forming 12 boundary layers. The computational domain starts $10D$ upstream from the bend entrance



(a) (b)

Figure 5: Experimental apparatus: (a) General flow system, (b) Geometry of the curved square duct.

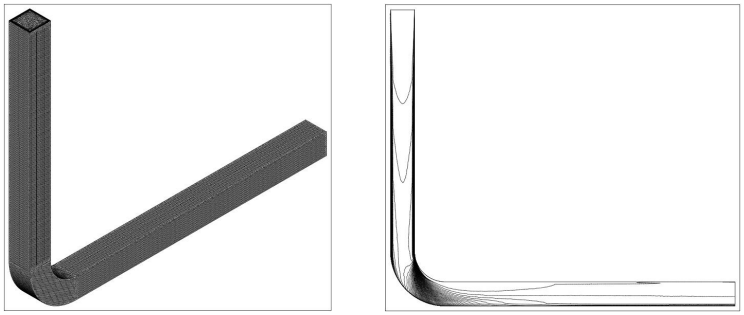


Figure 6: Hybrid used mesh and volume fraction profile for the curved duct.

and extends up to $12D$ downstream from the bend exit. The classical viscous flow boundary condition are imposed for the the fluid phase while a rebounding particle-wall conditions with normal and tangential restitution coefficients of 0.9 and 0.8 respectively, are considered for the solid phase. Finally, the turbulence features are captured using the Spalart-Allmaras turbulent model.

3.3. Results

Figure 6 shows the used mesh and the particles volume fraction profile. Figure 7 shows the residual convergence to the steady state. Figure 8 shows a good agreement of the fluid and particles mean stream velocity with experimental results for the different sections shown in 5-(b). This demonstrates the validity and the accuracy of the physical and numerical model.

4. Conclusions

The paper presented a new *Riemann* solver for scalar hyperbolic equations with discontinuous coefficient. A numerical proof of physical phenomena predicted by the proposed solver and missed by all existing *Riemann* solvers is provided. Appli-

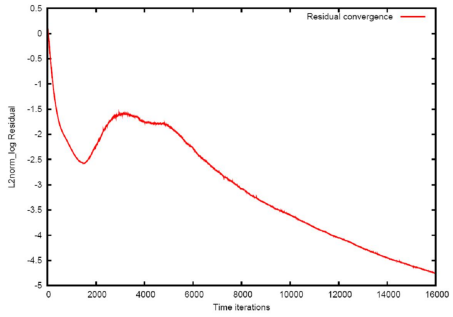
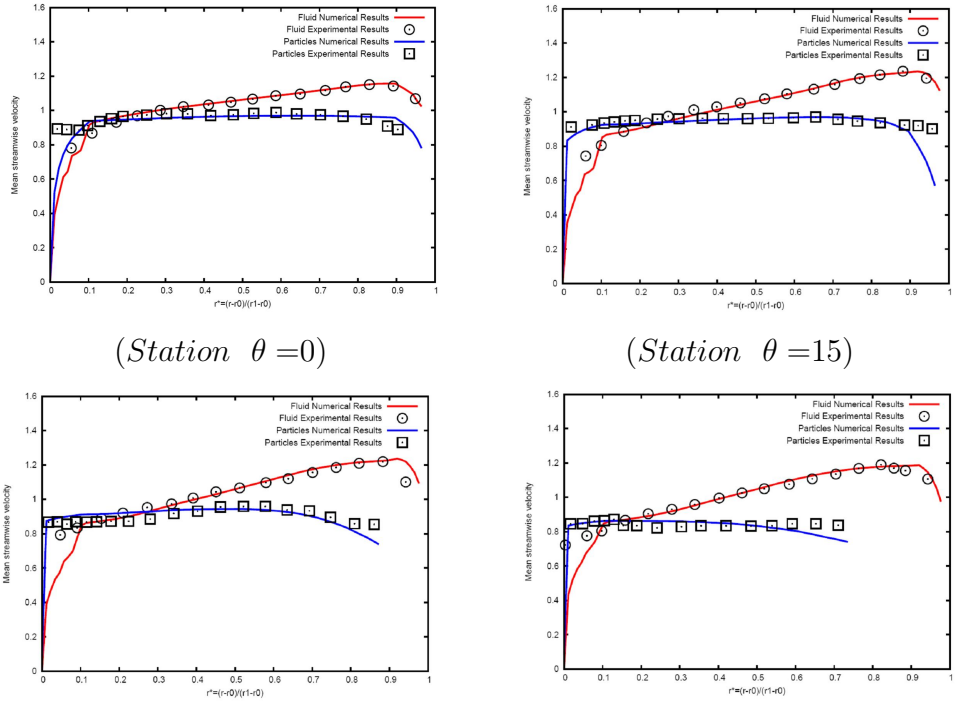


Figure 7: Bend case: Residual convergence.



(*Station* θ =0) (*Station* θ =15)

(*Station* θ =30) (*Station* θ =45)

Figure 8: 90 Bend case: Mean Stream fluid and particles velocity comparison to experimental results for different stations.

cation of the solver in a gas-particle model discretization is achieved and applied to a 90 curved bend with comparison to experimental data. This work is a first step toward a construction of a *Riemann* solver for systems with discontinuous coefficients and its application for inviscid fluxes estimation in the Navier-Stokes and electromagnetic equations discretization.

References

[1] Einfeldt, B.: On gudonov-type methods for gas dynamics. SIAM J. Numer. Anal. **25**(2) (1988), 294–318.

[2] Einfeldt, B., Munz, C., Roe, P., and Sjgreen, B.: On Gudonov-type methods near low densities. J. Comput. Phys. **92** (1991), 273–295.

[3] Godunov, S.: A finite difference method for the computation of discontinuous solutions of the equations of fluid dynamics. Comp. Math. (in Russian) **47** (1959), 357–393.

[4] Harten, A., Lax, P., and Van Leer, B.: On upstream diffrencing and Gudonov-type schemes for hyperbolic conservation laws. SIAM Review **25**(1) (1983), 35–61.

[5] Kliafas, Y. and Holt, M.: LDV measurements of a turbulent air-solid two-phase flow in a 90° bend. Experiments in Fluids **5** (1987), 73–85.

[6] Remaki, L., Hassan, O., and Kenneth, K.: New limiter and gradient reconstruction method for HLLC-finite volume scheme to solve Navier-Stokes equations. In: *TECCOMAS, the fifth European Congress on Computational in Fluid Dynamic, Lisbon, Portugal.*, 14–17 June 2010.

[7] Remaki, L.: Theoretical and numerical study of quasi-linear equations with discontinuous coefficients, and 2D linear acoustic. PhD Thesis, 1997, France.

[8] Roe, P.: Approximate riemann solvers, parameter vectors, and difference schemes. J. Comput. Phys. **43** (1981), 357–372.

[9] Roe, P.: Characteristic-based schemes for the Euler equations. Ann. Rev. Fluid Mech. **18** (1986), 337.

[10] Toro, E., Spruce, M., and Spearses, W.: Restoration of the contact surface in the HLL Riemann solver. Shock Waves **4** (1994), 25–34.

Conference Applications of Mathematics 2013
in honor of the 70th birthday of Karel Segeth.
Jan Brandts, Sergey Korotov, Michal Křížek,
Jakub Šístek, and Tomáš Vejchodský (Eds.),
Institute of Mathematics AS CR, Prague 2013

ON MATHEMATICAL MODELLING OF GUST RESPONSE USING THE FINITE ELEMENT METHOD

Petr Sváček[1], Jaromír Horáček[2]

[1] Czech Technical University in Prague, Faculty of Mechanical Engineering
Department of Technical Mathematics
Karlovo nám. 13, 121 35 Praha 2, Czech Republic
e-mail: Petr.Svacek@fs.cvut.cz
[2] Institute of Thermomechanics, Academy of Sciences of the Czech Republic
Dolejškova 5, 182 00 Praha 8, Czech Republic
email: jaromirh@it.cas.cz

Abstract

In this paper the numerical approximation of aeroelastic response to sudden gust is presented. The fully coupled formulation of two dimensional incompressible viscous fluid flow over a flexibly supported structure is used. The flow is modelled with the system of Navier-Stokes equations written in Arbitrary Lagrangian-Eulerian form and coupled with system of ordinary differential equations describing the airfoil vibrations with two degrees of freedom. The Navier-Stokes equations are spatially discretized by the fully stabilized finite element method. The numerical results are shown.

1. Introduction

The gust-response analysis is important in the aircraft wing design, typically the wing have to withstand a gust of certain intensity and profile. As the aeroelastic effects can have significant influence on the gust loads, the aeroelastic analysis of gust response is important, cf. [1]. Modern methods for dynamic gust analysis typically rely on panel-method aerodynamics, where the frequency domain formulations are being used. In this paper the dynamic gust-response analysis is performed with the aid of the developed finite element code, and particularly the gust response of a flexibly supported very light airfoil was numerically analyzed. The mathematical model consists of fluid flow described by the two-dimensional Navier-Stokes equations and the continuity equation coupled with the equations describing the airfoil motion. The incompressible flow is approximated by the finite element method (FEM). The couple of finite element velocity/pressure spaces satisfies the Babuška-Breezi condition, see e.g. [6]. The dominating convection is stabilized by the residual based stabilization, cf. [5]. The numerical solution is sought on adaptively refined meshes, cf. [4]. The motion of the computational domain is treated by the Arbitrary Lagrangian-Eulerian (ALE) method, cf. ([8, 7]).

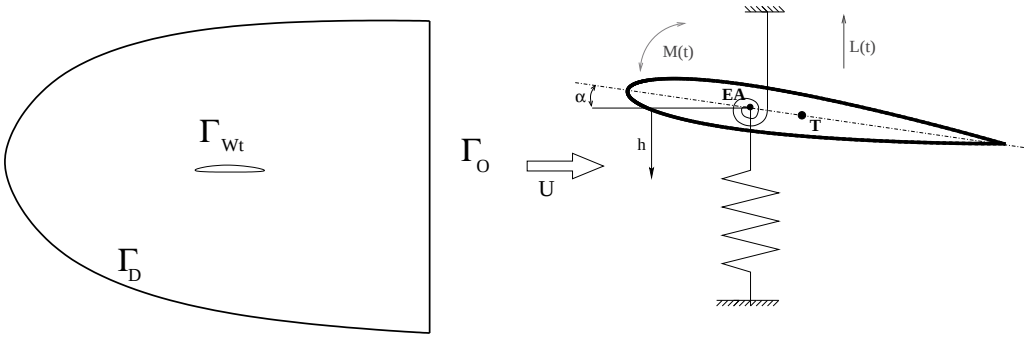


Figure 1: A sketch of the computational domain and its boundary (left). The elastic support of the airfoil on translational and rotational springs (right).

2. Mathematical model

Fluid flow. The flow in the two-dimensional computational domain Ω_t is described by the incompressible Navier-Stokes equations written in ALE form

$$\frac{D^{\mathcal{A}}\mathbf{u}}{Dt} + (\mathbf{u} - \mathbf{w}_D) \cdot \nabla \mathbf{u} + \nabla p - \nu \triangle \mathbf{u} = 0 \text{ in } \Omega_t, \qquad (1)$$
$$\nabla \cdot \mathbf{u} = 0 \text{ in } \Omega_t,$$

where $\frac{D^{\mathcal{A}}}{Dt}$ denotes the ALE derivative, $\mathbf{w}_D$ denotes the ALE domain velocity, $\mathbf{u} = (u_1, u_2)^T$ is the velocity vector, p is the kinematic pressure, and ν is the kinematic viscosity. The symbol $\mathcal{A}_t$ denotes a regular one-to-one Arbitrary Lagrangian-Eulerian (ALE) mapping of the reference configuration Ω_0 onto the current configuration Ω_t for any time instant $t \in I$. The boundary $\partial\Omega_t$ consists of mutually disjoint parts shown in Fig. 1, where Γ_D is the inlet part, Γ_O is the outlet boundary, and Γ_{Wt} is the moving surface of the airfoil, $\partial\Omega_t = \Gamma_D \cup \Gamma_O \cup \Gamma_{Wt}$. The system of equations (1) is completed with boundary conditions

$$\begin{aligned} &\text{a)} \quad \mathbf{u}(x,t) = \mathbf{u}_D + \mathbf{V}_g(t) \qquad \text{for } x \in \Gamma_D, \quad t \in I, \\ &\text{b)} \quad \mathbf{u}(x,t) = \mathbf{w}_D(x,t) \qquad \text{for } x \in \Gamma_{Wt}, \quad t \in I, \\ &\text{c)} \quad -\nu \tfrac{\partial \mathbf{u}}{\partial \mathbf{n}} + (p - p_{ref})\mathbf{n} = 0 \qquad \text{on } \Gamma_O, \end{aligned} \qquad (2)$$

where $\mathbf{u}_D = (U_\infty, 0)$ is the far field velocity, $\mathbf{V}_g(t)$ is the vertical gust velocity, p_{ref} is a reference mean value of pressure at the outlet part of boundary, and by an initial condition

$$\mathbf{u}(x,0) = \mathbf{u}_0(x), \qquad x \in \Omega_0. \qquad (3)$$

Structural model. The fluid flow is coupled with the motion of a flexibly supported airfoil, which can be vertically displaced and rotated. Fig. 1 shows the elastic support of the airfoil on translational and rotational springs. The pressure and viscous forces acting on the vibrating airfoil immersed in flow result in the lift force $L(t)$

and the torsional moment $M(t)$. The governing equations are written in the form (see [10])

$$\begin{aligned} m\ddot{h} + S_\alpha\,\ddot{\alpha} + k_{hh}h &= -L(t), \\ S_\alpha\ddot{h} + I_\alpha\ddot{\alpha} + k_{\alpha\alpha}\alpha &= M(t), \end{aligned} \tag{4}$$

where k_{hh} and $k_{\alpha\alpha}$ are the bending stiffness and torsional stiffness, respectively, m is the mass of the airfoil, S_α is the static moment around the elastic axis EA and I_α is the inertia moment around EA.

Coupling conditions. The aerodynamic lift force L acting in the vertical direction and the torsional moment M are defined by

$$L = -\,l\int_{\Gamma_{Wt}}\sum_{j=1}^{2}\tau_{2j}n_j dS, \qquad M = -\,l\int_{\Gamma_{Wt}}\sum_{i,j=1}^{2}\tau_{ij}n_j r_i^{\mathrm{ort}} dS, \tag{5}$$

where l is the depth of the considered airfoil section, and τ_{ij} are the components of the stress tensor defined by

$$\begin{aligned} \tau_{ij} &= \rho\left[-p\delta_{ij} + \nu\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right)\right], \\ r_1^{\mathrm{ort}} &= -(x_2 - x_{\mathrm{EA2}}),\ r_2^{\mathrm{ort}} = x_1 - x_{\mathrm{EA1}}. \end{aligned} \tag{6}$$

In Eq. (6) by δ_{ij} the Kronecker symbol is denoted, $\mathbf{n} = (n_1, n_2)$ is the unit outer normal vector to $\partial\Omega_t$ on Γ_{Wt} (pointing into the airfoil) and $x_{\mathrm{EA}} = (x_{\mathrm{EA1}}, x_{\mathrm{EA2}})$ is the position of the elastic axis. Relations (5) and (6) define the coupling of the fluid model with the structural model.

3. Finite element approximation

The straightforward application of FEM procedures is often not possible for the incompressible Navier-Stokes equations particularly due to the advection-diffusion character of the equations with the dominating advection, for which a case the Galerkin FEM leads to unphysical solutions if the grid is not fine enough in regions of strong gradients. In order to obtain physically admissible correct solutions it is necessary to apply suitable mesh refinement (e.g. anisotropically refined mesh, cf. [4]) combined with a stabilization technique, cf. [2, 9]. In this work, the FEM is stabilized with the aid of streamline upwind/pressure stabilizing Petrov-Galerkin (SUPG/PSPG) method (so called *fully stabilized scheme*, cf. [5]) modified for the application on moving domains (cf. [10]). In order to discretize the problem (1), we define the equidistant division of the time interval $[0, T]$ with the time step Δt, denote $t_n = n\Delta t$, and approximate the time derivative by second order backward difference formula:

$$\frac{D^{\mathcal{A}}\mathbf{u}}{Dt}(x,t) \approx \frac{3\mathbf{u}^{n+1} - 4\widehat{\mathbf{u}^n} + \widehat{\mathbf{u}^{n-1}}}{2\Delta t},$$

where $\mathbf{u}^{n+1}$ is the approximation of the flow velocity at time t^{n+1} defined on the computational domain Ω^{n+1}, and $\widehat{\mathbf{u}^k}$ is the transformation of the flow velocity at time t^k defined on Ω^k transformed onto Ω^{n+1}. Further, equation (1) is formulated weakly and the solution is sought on the couple of finite element spaces $W_\Delta \subset \mathbf{H}^1(\Omega^{n+1})$ and $Q_\Delta \subset L^2(\Omega^{n+1})$ for approximation of velocity components and pressure, respectively. Further, by $X_\Delta \subset W_\Delta$ the subspace of the test functions is denoted. Let us mention that the finite element spaces should satisfy the *Babuška–Brezzi (BB) condition* (see e.g. [6]). In practical computations we assume that the domain $\Omega = \Omega^{n+1}$ is a polygonal approximation of the region occupied by the fluid at time t^{n+1} and the finite element spaces are defined over a triangulation $\mathcal{T}_\Delta$ of the domain Ω_t as piecewise polynomial functions. In our computations, the well-known Taylor-Hood P_2/P_1 conforming finite element spaces are used for the velocity/pressure approximation.

The *stabilized discrete problem* at a time instant $t = t^{n+1}$ reads: Find $U = (\mathbf{u}, p) \in W_\Delta \times Q_\Delta$, $p := p^{n+1}$, $\mathbf{u} := \mathbf{u}^{n+1}$, such that $\mathbf{u}$ satisfies approximately the Dirichlet boundary conditions (2 a-b) and

$$a(U; U, V) + \mathcal{L}(U; U, V) + \mathcal{P}(U, V) = f(V) + \mathcal{F}(U; V) \tag{7}$$

holds for all $V = (\mathbf{z}, q) \in X_\Delta \times Q_\Delta$. Here, the Galerkin terms are defined for any $U = (\mathbf{u}, p)$, $V = (\mathbf{z}, q)$, $U^* = (\mathbf{u}^*, p^*)$ by

$$\begin{aligned} a(U^*; U, V) &= \frac{3}{2\Delta t}(\mathbf{u}, \mathbf{z})_\Omega + \frac{1}{Re}(\nabla \mathbf{u}, \nabla \mathbf{z})_\Omega + (\overline{\mathbf{w}} \cdot \nabla \mathbf{u}, \mathbf{z})_\Omega - (p, \nabla \cdot \mathbf{z})_\Omega + (\nabla \cdot \mathbf{u}, q)_\Omega, \\ f(\mathbf{u}, \mathbf{z}) &= \frac{1}{2\Delta t}(4\hat{\mathbf{u}}^n - \hat{\mathbf{u}}^{n-1}, \mathbf{z})_\Omega, \end{aligned}$$

where $\overline{\mathbf{w}} = \mathbf{u}^* - \mathbf{w}_D^{n+1}$, and the scalar product in $L^2(\Omega)$ is denoted by $(\cdot, \cdot)_\Omega$. Further, the SUPG/PSPG stabilization terms are used in order to obtain stable solution also for large values of Reynolds numbers,

$$\begin{aligned} \mathcal{L}(U^*; U, V) &= \sum_{K \in \mathcal{T}_\Delta} \delta_K \left(\frac{3\mathbf{u}}{2\tau} - \frac{1}{Re}\triangle \mathbf{u} + (\overline{\mathbf{w}} \cdot \nabla)\mathbf{u} + \nabla p, (\overline{\mathbf{w}} \cdot \nabla)\mathbf{v} + \nabla q \right)_K, \\ \mathcal{F}(U^*; V) &= \sum_{K \in \mathcal{T}_\Delta} \delta_K \left(\frac{4\hat{\mathbf{u}}^n - \hat{\mathbf{u}}^{n-1}}{2\tau}, (\overline{\mathbf{w}} \cdot \nabla)\mathbf{v} + \nabla q \right)_K, \end{aligned}$$

where $\overline{\mathbf{w}} = \mathbf{v}^* - \mathbf{w}^{n+1}$, and $(\cdot, \cdot)_K$ denotes the scalar product in $L^2(K)$. The term $\mathcal{P}(U, V)$ is the additional grad-div stabilization defined by

$$\mathcal{P}(U, V) = \sum_{K \in \mathcal{T}_\Delta} \tau_K (\nabla \cdot \mathbf{u}, \nabla \cdot \mathbf{z})_K.$$

Here, the choice of the parameters $\delta_K \approx h_K^2$ and $\tau_K \approx 1$ is carried out according to [5] or [9] on the basis of the local element length h_K.

Furthermore, the nonlinear stabilized weak formulation of Navier-Stokes system (7) is solved with the aid of Oseen linearization. The arising large system of linear equations is solved by a direct solver as UMFPACK (cf. [3]).

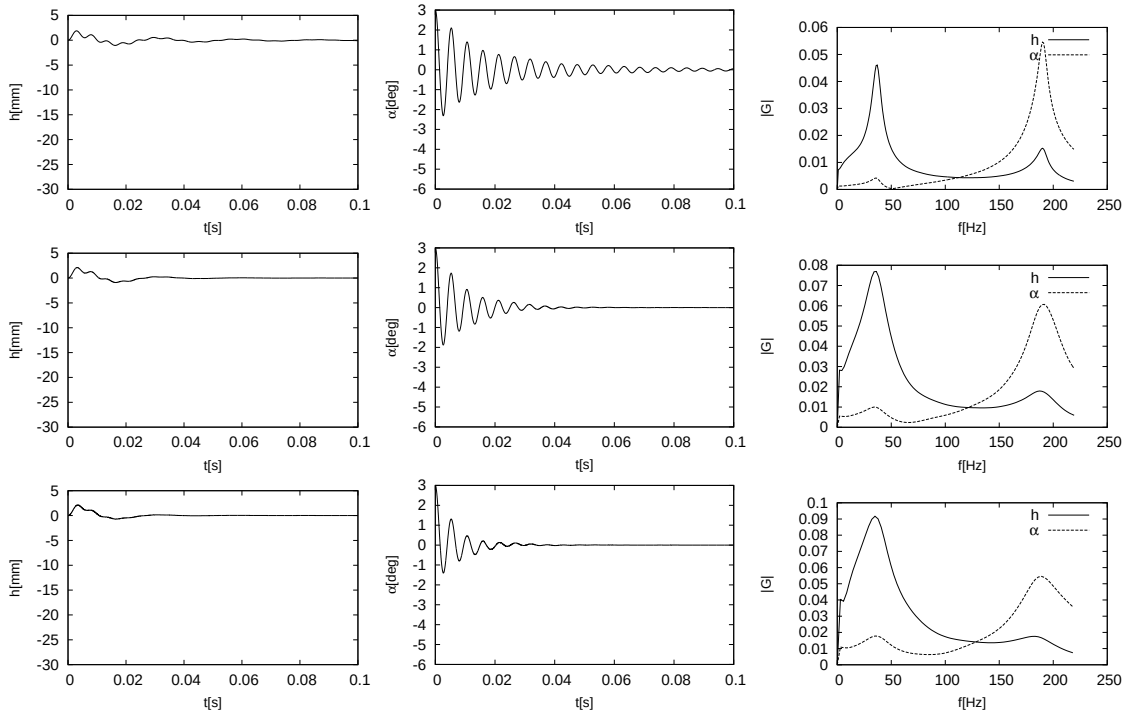


Figure 2: Aeroelastic response and its spectra for the far field velocity $U_\infty = 5\,\mathrm{m/s}$ (top), $U_\infty = 10\,\mathrm{m/s}$ (middle) and $U_\infty = 15\,\mathrm{m/s}$ (bottom).

The equations describing the motion of the flexibly supported airfoil are discretized in time by second order backward difference formula and the coupled fluid-structure model is solved using the partitioned strongly coupled algorithm. This means that per every time step the fluid flow and the structure motion are approximated repeatedly in order to converge to a solution which satisfies all interface conditions. In order to overcome the instability due to the coupling procedure, an underrelaxation is applied for the structural part of the problem.

4. Numerical results

The presented numerical method is applied for approximation of aeroelastic behaviour of a typical section, which is an idealized representation of a wing.

The structural parameters were chosen according to [1]. The aircraft wing structural arrangement is uniformly made of balsa wood (density $\rho = 150\,\mathrm{kg/m^3}$, Young modulus $E = 1.3\times10^9\,\mathrm{Pa}$, shear modulus $G = 6.2\times10^8\,\mathrm{Pa}$). The airfoil shape is given by the Karman-Trefftz conformal transformation, for details see [1]. The mass and inertial properties of the considered airfoil were $m = 2\times10^{-4}\,\mathrm{kg}$ and $I = 10^{-7}\,\mathrm{kg\ m^2}$. Thus $I_\alpha = 1.2\times10^{-7}\,\mathrm{kg\ m^2}$, $S_\alpha = 2\times10^{-6}\,\mathrm{kg\ m}$. The stiffness coefficients of the springs were $k_h = 25.4\,\mathrm{N/m}$, $k_\alpha = 0.272\,\mathrm{N\,m/rad}$. The airfoil chord was $c = 0.1\,\mathrm{m}$ and the depth of the section was $l = 0.03\,\mathrm{m}$. The air density was $\rho = 1.225\,\mathrm{kg\,m^{-3}}$ and the air kinematic viscosity was $\nu = 1.453\times10^{-5}\,\mathrm{m^2/s}$.

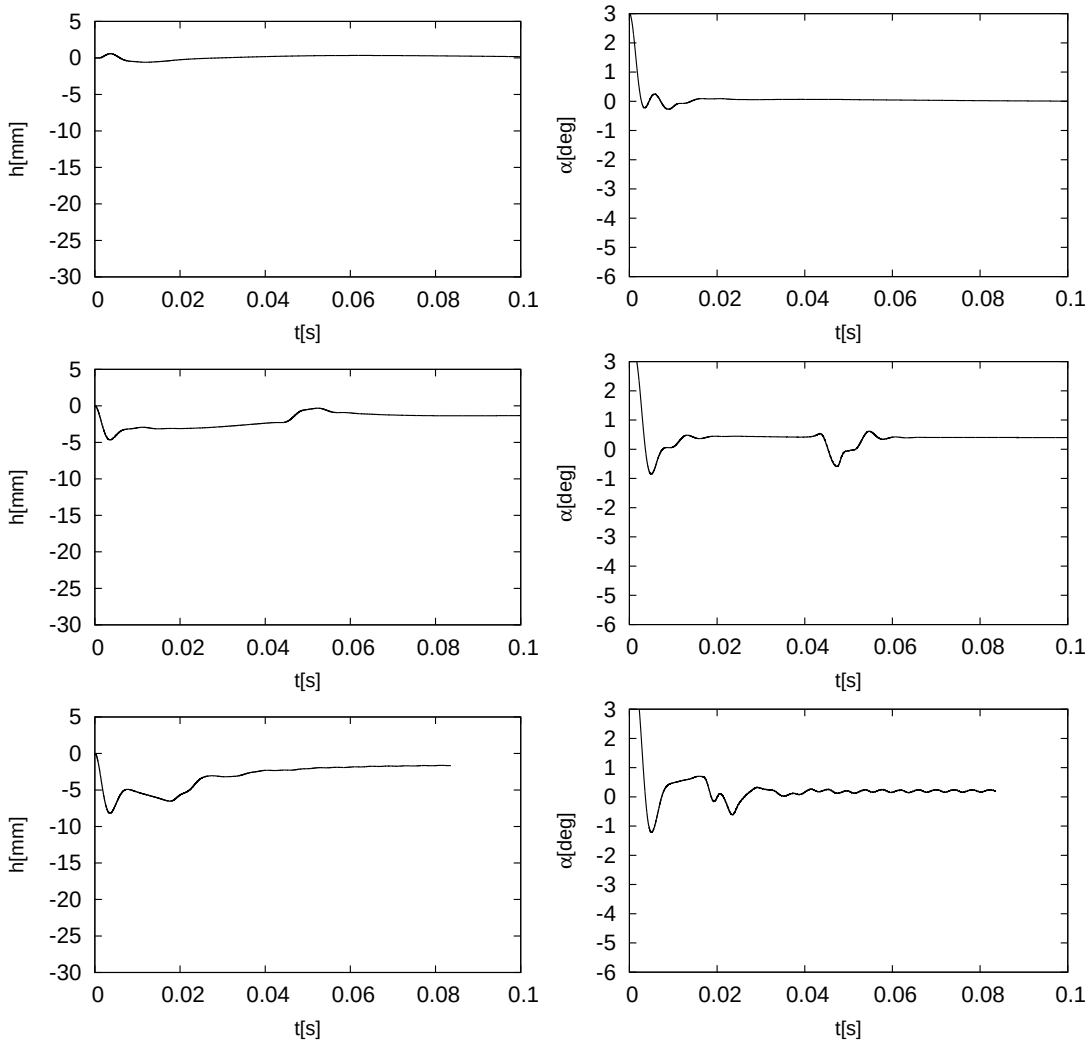


Figure 3: Aeroelastic response for the far field velocity $U_\infty = 30\,\mathrm{m/s}$ (top), $U_\infty = 50\,\mathrm{m/s}$ (middle) and $U_\infty = 60\,\mathrm{m/s}$ (bottom).

A vertical wind gust acts as the aerodynamic perturbation to the static equilibrium of the aeroelastic system and is introduced as a variation of the free-stream velocity prescribed on the inlet part of the boundary. A sinusoidal vertical gust of one-second duration is considered. The reference free stream velocity was chosen $U_\infty = 15\,\mathrm{m/s}$ and the gust intensity $V_G = 1.5\,\mathrm{m/s}$ for the light gust and $V_G = 5\,\mathrm{m/s}$ for the heavy gust case was considered.

The aeroelastic response of the considered airfoil computed for constant far field velocities and the initial conditions $\alpha(0) = 3°$, $h(0) = 0\,\mathrm{m}$, $\dot{h}(0) = 0\,\mathrm{m/s}$, $\dot{\alpha}(0) = 0°\mathrm{s}^{-1}$ are shown in Figs. 2 and 3. The spectra of the numerically simulated signals show the lower resonance frequency at about 36 Hz for predominantly vertical vibrations and at about 191 Hz for rotation of the airfoil. The damping of the system increases with higher flow velocities. The system is damped by aerodynamic forces and is stable for all the considered values of far field velocities up to $U_\infty = 60\,\mathrm{m/s}$.

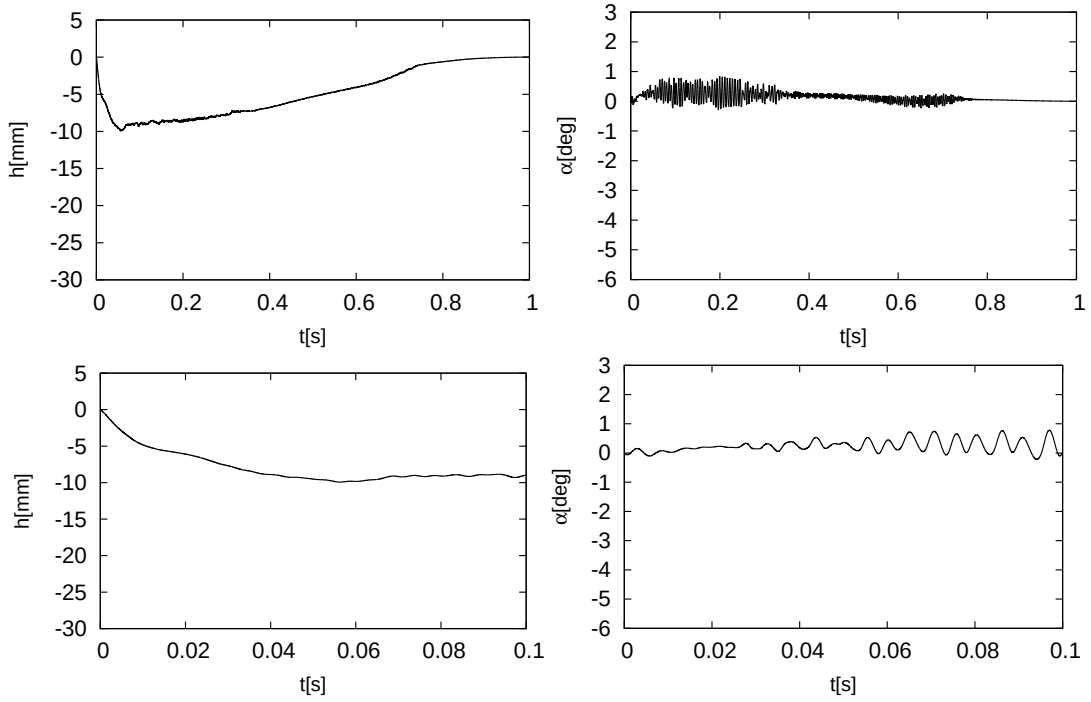


Figure 4: The numerical results for light gust. The aeroelastic response of the typical airfoil section for far field velocity $U_\infty = 15\,\mathrm{m/s}$ and gust velocity $V_G = 1.5\,\mathrm{m/s}$ (top). The detail of the response during the first 0.1 s (bottom).

The gust aeroelastic responses computed for the cases of either a light ($V_G = 1.5\,\mathrm{m/s}$) or a heavy gust ($V_G = 5\,\mathrm{m/s}$) are shown in Figs. 4 and 5. For the case shown in Fig. 5 the flow velocity patterns are shown in Fig. 6, where a very strong vorticity above the vibrating profile was developed after the airfoil loading by the heavy gust.

5. Discussion and conclusion

The gust response of a typical airfoil section has been investigated with the aid of a developed numerical scheme. The numerical method was described and the numerical results of a benchmark problem were presented.

The aeroelastic gust responses exhibit stronger oscillations than it was found by Berci et al. [1]. The maximum of the airfoil rotation amplitude for a light gust resulting from our computation ($\alpha \approx 0.8°$) was found nearly three time higher than the maximum rotation ($\alpha \approx 0.27°$) computed in [1], however, a maximum of a mean value for rotation ($\alpha \approx 0.3°$) is in a good agreement with the results by Berci et al. Similarly, the maximum value of the computed vertical displacement $h \approx 10\,\mathrm{mm}$ approximately correspond to a maximum $h \approx 8\,\mathrm{mm}$ found in [1]. A dominant oscillation frequency corresponds to the airfoil rotation.

Similar conclusions result from the computation of the airfoil response to a heavy gust. The maximum values for the horizontal displacement of about $h \approx 30\,\mathrm{mm}$

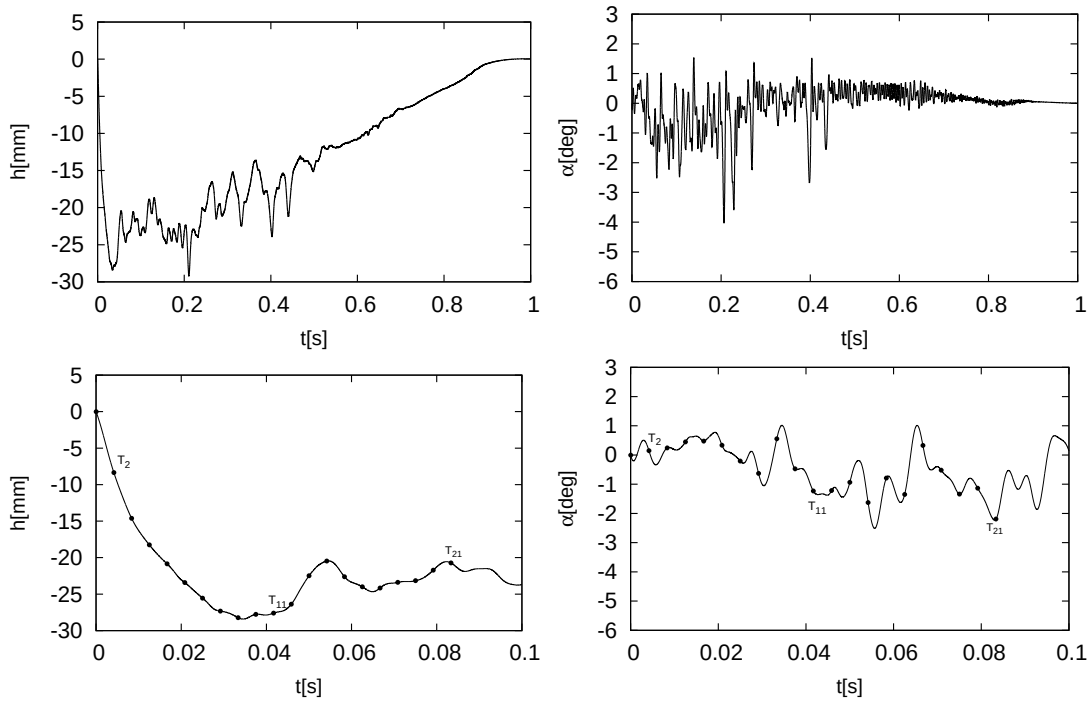


Figure 5: The numerical results for heavy gust. The aeroelastic response of the typical airfoil section for far field velocity $U_\infty = 15\,\mathrm{m/s}$ and gust velocity $V_G = 5\,\mathrm{m/s}$ (top). The detail of the response during the first 0.1 s (bottom).

computed by us correspond well to the maximum value $h \approx 27\,\mathrm{mm}$ obtained in [1]. The maxima of mean values computed for rotation ($\alpha \approx 1°$) are in a good agreement in both studies, however, the maximum value for rotation computed in our case ($\alpha \approx 4°$) is evidently higher than the maximum $\alpha \approx 2.5°$ found in [1].

The reason for the found differences between the two approaches of the numerical simulation can be mainly in the flow model. Berci et al. [1] considered Reynolds Average Navier Stokes equations (RANS) including a turbulence model for the flow and we considered the laminar flow, when the flow separation on the airfoil surface is becoming earlier and creation of the vortices is more frequent than in the case of turbulent boundary layer.

Acknowledgements

This work was supported by grant No. P101/11/0207 of the Czech Science Foundation.

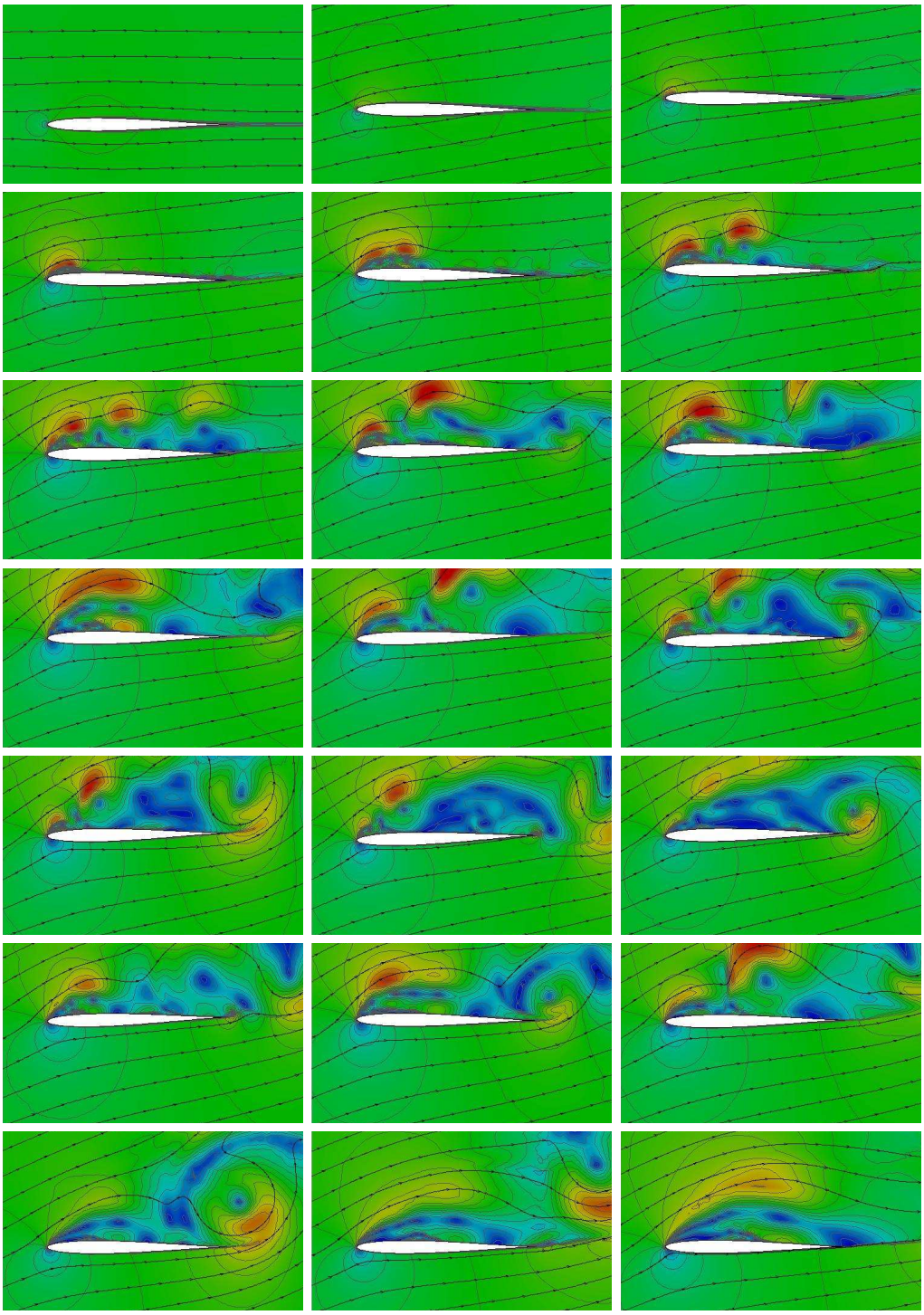


Figure 6: The flow pattern at the equidistant time instants $T_k = k\Delta T$ marked in Fig. 5, $\Delta T = 2 \times 10^{-4}$ s.

References

[1] Berci, M., Mascetti, S., Incognito, A., Gaskell, P. H., and Toropov, V. V.: Gust response of a typical section via CFD and analytical solutions. In: J. C. F. Pereira and A. Sequeira (Eds.), *ECCOMAS CFD 2010, V. European Conference on Computational Fluid Dynamics*, 2010.

[2] Codina, R.: Stabilization of incompressibility and convection through orthogonal sub-scales in finite element methods. Comput. Method Appl. Mech. Engrg. **190** (2000), 1579–1599.

[3] Davis, T. A. and Duff, I. S.: A combined unifrontal/multifrontal method for unsymmetric sparse matrices. ACM Trans. Math. Software **25** (1999), 1–19.

[4] Dolejší, V.: Anisotropic mesh adaptation technique for viscous flow simulation. East-West J. Numer. Math. **9** (2001), 1–24.

[5] Gelhard, T., Lube, G., Olshanskii, M. A., and Starcke, J. H.: Stabilized finite element schemes with LBB-stable elements for incompressible flows. J. Comput. Appl. Math. **177** (2005), 243–267.

[6] Girault, V. and Raviart, P. A.: *Finite element methods for the Navier-Stokes equations*. Springer, Berlin, 1986.

[7] Le Tallec, P. and Mouro, J.: Fluid structure interaction with large structural displacements. Comput. Method Appl. Mech. Engrg. **190** (2001), 3039–3067.

[8] Nomura, T. and Hughes, T. J. R.: An arbitrary Lagrangian-Eulerian finite element method for interaction of fluid and a rigid body. Comput. Method Appl. Mech. Engrg. **95** (1992), 115–138.

[9] Sváček, P. and Feistauer, M.: Application of a stabilized FEM to problems of aeroelasticity. In: *Numerical Mathematics and Advanced Application*, pp. 796–805. Springer, Berlin, 2004.

[10] Sváček, P., Feistauer, M., and Horáček, J.: Numerical simulation of flow induced airfoil vibrations with large amplitudes. J. Fluids and Structures **23** (2007), 391–411.

Conference Applications of Mathematics 2013
in honor of the 70th birthday of Karel Segeth.
Jan Brandts, Sergey Korotov, Michal Křížek,
Jakub Šístek, and Tomáš Vejchodský (Eds.),
Institute of Mathematics AS CR, Prague 2013

ON DOMAIN DECOMPOSITION METHODS FOR OPTIMAL CONTROL PROBLEMS

Minh-Binh Tran

Basque Center for Applied Mathematics
Mazarredo 14, 48009 Bilbao, Spain
tbinh@bcamath.org

Abstract

In this note, we introduce a new approach to study overlapping domain decomposition methods for optimal control systems governed by partial differential equations. The model considered in our paper is systems governed by wave equations. Our technique could be used for several other equations as well.

1. Introduction

The research about using domain decomposition methods to resolve optimal control problems started with the pioneering work of A. Bensoussan, R. Glowinski and P. L. Lions [8] in the 70's and B. Depres and J. D. Benamou in the early 90's [2, 1, 7, 6, 5, 4, 4, 3]. Since then, this research line has become very active with several works of J. E. Lagnese and G. Leugering [13, 11, 10, 9, 12]. However, most of the works on domain decomposition methods for optimal control of systems governed by partial differential equations are devoted to nonoverlapping algorithms, though overlapping algorithms are proved to be more stable and much faster [14]. One of the reasons is that there was no convergence proof of the overlapping algorithms. In the series of papers [17, 16, 18, 15], we develop a new technique to study the convergence of overlapping algorithms. The technique is proved to be applicable for the convergence study of domain decomposition algorithms for several kinds of partial differential equations. Within the frame of developing our new technique for different convergence problems, this note is devoted to the application of the technique to study an overlapping domain decomposition for optimal control systems governed by wave equations, which was studied in [1] but only for the nonoverlapping case. Our technique has the potential of being a new tool to extend many of the previous studies from nonoverlapping to overlapping algorithms. For the sake of simplicity, we only consider a decomposition with two subdomains, however, our technique could be extended to the multisubdomains case without any difficulty.

2. Model description and definition of the domain decomposition algorithm

Let Ω be a smooth bounded domain in $\mathbb{R}^N$. Similarly as in [1], we consider the following wave equation defined on $(0,T)\times\Omega$

$$\begin{cases} \partial_{tt}y(t,x)-\Delta y(t,x)=f(t,x)+v(t,x) \text{ on } (0,T)\times\Omega, \\ y(0,x)=y_0(x); \qquad \partial_t y(0,x)=y_1(x) \text{ on } \Omega, \\ y(t,x)=g(t,x) \text{ on } (0,T)\times\partial\Omega, \end{cases} \tag{1}$$

where $y_0, y_1 \in L^2(\Omega)$, $g \in L^2((0,T)\times\partial\Omega)$.

Let U be a convex subset of $L^2((0,T)\times\Omega)$ and define the function

$$J(v,y)=\frac{1}{2}\int_{(0,T)\times\Omega}(\gamma|y(x)|^2+\alpha|v(t,x)|^2)dxdt, \tag{2}$$

where α and γ are positive constants.

We consider the following optimization problem

$$\min_{v\in U} J(v,y(v)). \tag{3}$$

Following [1], we need to solve

$$\begin{cases} \partial_{tt}p\,(t,x)-\Delta p(t,x)=y(t,x) \text{ on } (0,T)\times\Omega, \\ p\,(T,x)=0; \qquad \partial_t p(T,x)=0 \text{ on } \Omega, \\ p\,(t,x)=0 \text{ on } (0,T)\times\partial\Omega, \\ \int_{(0,T)\times\Omega}(p+\alpha v)(w-v)dxdt\geq 0 \ \forall w\in U. \end{cases} \tag{4}$$

We now design an overlapping domain decomposition method to resolve the system (1) and (4). Divide the domain Ω into two overlapping subdomains Ω_1 and Ω_2 in the following sense

$$\Omega=\Omega_1\cup\Omega_2,$$

$$(\partial\Omega_1\backslash\partial\Omega)\cap(\partial\Omega_2\backslash\partial\Omega)=\emptyset.$$

The overlapping domain decomposition algorithm with Robin transmission condition now reads for $i\in\{1,2\}$

$$\begin{cases} \partial_{tt}y_i^{n+1}-\Delta y_i^{n+1}=f(t,x)+v_i^{n+1}(t,x) \text{ on } (0,T)\times\Omega_i, \\ y_i^{n+1}(0,x)=y_0(x), \qquad \partial_t y_i^{n+1}(0,x)=y_1(x) \text{ on } \Omega_i, \\ y_i^{n+1}(t,x)=g(t,x) \text{ on } (0,T)\times\partial\Omega_i, \end{cases}$$

$$\begin{cases} \partial_{tt} p_i^{n+1} - \Delta p_i^{n+1} = \gamma y_i^n(t,x) \text{ on } (0,T)\times\Omega_i, \\ p_i^{n+1}(T,x) = 0, \quad \partial_t p_i^{n+1}(T,x) = 0 \text{ on } \Omega_i, \\ p_i^{n+1}(t,x) = 0 \text{ on } (0,T)\times\partial\Omega_i, \end{cases}$$

$$\int_{(0,T)\times\Omega_i} (p_i^{n+1} + \alpha v_i^{n+1})(w_i - v_i^{n+1}) dxdt \geq 0,$$

with the transmission condition on $\partial\Omega_i \backslash \partial\Omega$

$$\begin{aligned} \partial_{\nu_i} y_i^{n+1} + r_i p_i^{n+1} &= \partial_{\nu_i} y_{3-i}^n + r_i p_{3-i}^n, \\ \partial_{\nu_i} p_i^{n+1} + r_i y_i^{n+1} &= \partial_{\nu_i} p_{3-i}^n + r_i y_{3-i}^n, \end{aligned}$$

where ν_i is the outward normal outward unit normal vector of Ω_i on the boundary $\partial\Omega_i \backslash \partial\Omega$ and r_i is a positive constant. At step 0, we choose an initial guess (y_i^0, p_i^0) in $C^2([0,T]\times\overline{\Omega})$. We can see that the algorithm is well-posed and $(y_i^n, p_i^n, v_i^n) \in L^2(0,T,H^2(\Omega_i)) \times L^2(0,T,H^2(\Omega_i)) \times L^2(0,T,H^2(\Omega_i))$.

3. Convergence of the algorithm

For $i \in \{1,2\}$ we define

$$\begin{aligned} \tilde{y}_i^{n+1} &= y_i^{n+1} - y, \\ \tilde{p}_i^{n+1} &= p_i^{n+1} - p, \\ \tilde{v}_i^{n+1} &= v_i^{n+1} - v, \end{aligned}$$

and get the following systems

$$\begin{cases} \partial_{tt} \tilde{y}_i^{n+1} - \Delta \tilde{y}_i^{n+1} = \tilde{v}_i^{n+1}(t,x) \text{ on } (0,T)\times\Omega_i, \\ \tilde{y}_i^{n+1}(0,x) = 0, \quad \partial_t \tilde{y}_i^{n+1}(0,x) = 0 \text{ on } \Omega_i, \\ \tilde{y}_i^{n+1}(t,x) = 0 \text{ on } (0,T)\times\partial\Omega_i, \end{cases}$$

$$\begin{cases} \partial_{tt} \tilde{p}_i^{n+1} - \Delta \tilde{p}_i^{n+1} = \gamma \tilde{y}_i^n(t,x) \text{ on } (0,T)\times\Omega_i, \\ \tilde{p}_i^{n+1}(T,x) = 0, \quad \partial_t \tilde{p}_i^{n+1}(T,x) = 0 \text{ on } \Omega_i, \\ \tilde{p}_i^{n+1}(t,x) = 0 \text{ on } (0,T)\times\partial\Omega_i, \end{cases}$$

with the transmission condition on $\partial\Omega_i \backslash \partial\Omega$

$$\begin{aligned} \partial_{\nu_i} \tilde{y}_i^{n+1} + r_i \tilde{p}_i^{n+1} &= \partial_{\nu_i} \tilde{y}_{3-i}^n + r_i \tilde{p}_{3-i}^n, \\ \partial_{\nu_i} \tilde{p}_i^{n+1} + r_i \tilde{y}_i^{n+1} &= \partial_{\nu_i} \tilde{p}_{3-i}^n + r_i \tilde{y}_{3-i}^n. \end{aligned}$$

We suppose that for any $n \in \mathbb{N}$, $\tilde{v}_i^n$ is extended by 0 in (T, ∞) and still denote by $\tilde{y}_i^{n+1}$ the solution of

$$\begin{cases} \partial_{tt}\tilde{y}_i^{n+1} - \Delta\tilde{y}_i^{n+1} = \tilde{v}_i^{n+1}(t,x) \text{ on } (0,\infty)\times\Omega_i, \\ \tilde{y}_i^{n+1}(0,x) = 0, \quad \partial_t\tilde{y}_i^{n+1}(0,x) = 0 \text{ on } \Omega_i, \\ \tilde{y}_i^{n+1}(t,x) = 0 \text{ on } (0,\infty)\times\partial\Omega_i. \end{cases}$$

Using the change of variable $t \to T - t$, we still denote by $\tilde{p}_i^{n+1}$ the solution of

$$\begin{cases} \partial_{tt}\tilde{p}_i^{n+1} - \Delta\tilde{p}_i^{n+1} = \gamma\tilde{y}_i^{n}(T-t,x) \text{ on } (0,\infty)\times\Omega_i, \\ \tilde{p}_i^{n}(0,x) = 0, \quad \partial_t\tilde{p}_i^{n+1}(0,x) = 0 \text{ on } \Omega_i, \\ \tilde{p}_i^{n+1}(t,x) = 0 \text{ on } (0,\infty)\times\partial\Omega_i, \end{cases}$$

with the assumption that $\tilde{y}_i^n(T-t,x) = 0$ for $t > T$. Let H be a positive constant to be chosen later. Define

$$\bar{y}_i^n = \left(\int_0^\infty |\tilde{y}_i^n| \exp\left(-\sqrt{H}t\right) dt\right) g_i^n; \quad \bar{p}_i^n = \left(\int_0^\infty |\tilde{p}_i^n| \exp\left(-\sqrt{H}t\right) dt\right) g_i^n,$$

with $g_i^n \in C^2(\mathbb{R}^N, \mathbb{R})$, $g_i^n > 0$ to be chosen later. For $F : \Omega \to \mathbb{R}$, we define the following norm

$$|||F||| = \left[\int_{\text{supp}(F)} \left|\int_0^\infty |F| \exp\left(-\sqrt{H}t\right) dt\right|^2 dx\right]^{1/2}.$$

Similarly as in [15], a simple calculation leads to

$$\begin{aligned} -\Delta\bar{y}_i^{n+1} + H\bar{y}_i^{n+1} + \left(-\sum_{\alpha=1}^N \frac{\partial_\alpha g_i^{n+1}}{g_i^{n+1}} + \frac{\nabla g_i^{n+1}}{g_i^{n+1}}\right)\bar{y}_i^{n+1} + \sum_{\alpha=1}^N \frac{2\partial_{x_\alpha} g_i^{n+1}}{g_i}\,\partial_{x_\alpha}\bar{y}_i^{n+1} \qquad (5) \\ = \int_0^T v_i^{n+1}\,\text{sign}\,(\tilde{y}_i^{n+1}) \exp\left(-\sqrt{H}t\right) dt \text{ on } \Omega_i, \end{aligned}$$

$$\begin{aligned} -\Delta\bar{p}_i^{n+1} + H\bar{p}_i^{n+1} + \left(-2\sum_{\alpha=1}^N \frac{\partial_{x_\alpha} g_i^{n+1}}{g_i^{n+1}} + \frac{\nabla g_i^{n+1}}{g_i^{n+1}}\right)\bar{p}_i^{n+1} + \sum_{\alpha=1}^N 2\frac{\partial_{x_\alpha} g_i^{n+1}}{g_i^{n}}\,\partial_{x_\alpha}\bar{p}_i^{n} \qquad (6) \\ = \gamma\int_0^T y_i^n(T-t)\,\text{sign}\,(\tilde{p}_i^{n+1}) \exp\left(-\sqrt{H}t\right) dt \text{ on } \Omega_i. \end{aligned}$$

Choosing g_i^n such that $\nabla g_i^n - r_i g_i^n = 0$ on $\partial\Omega_i \backslash \Omega$, the transmission condition become

$$\begin{aligned}
\partial_{\nu_i} \bar{y}_i^{n+1} =& \partial_{\nu_i} \left(\int_0^\infty |\tilde{y}_i^n| \exp\left(-\sqrt{H}t\right) dt g_i^n \right) \\
=& \left[\int_0^\infty (\partial_{\nu_i} |\tilde{y}_i^n| + r_i |\tilde{y}_i^n|) \exp\left(-\sqrt{H}t\right) dt \right] g_i^n \\
&+ \int_0^\infty |\tilde{y}_i^n| \exp\left(-\sqrt{H}t\right) dt (\partial_{\nu_i} - r_i) g_i^n \\
=& \frac{1}{\beta_i} \partial_{\nu_i} \bar{y}_i^{n+1} \text{ on } \partial\Omega_i \backslash \partial\Omega,
\end{aligned}$$

by choosing g_i^n and g_{3-i}^n, we can make β_i to be a very large positive constant. Similarly, we also have

$$\beta_i \partial_{\nu_i} \bar{p}_i^{n+1} = \partial_{\nu_i} \bar{p}_{3-i}^n.$$

Let φ_{3-i}^n be a function in $H^1(\Omega \backslash \overline{\Omega_i})$ and φ_i^{n+1} be a function in $H^1(\Omega_i)$ such that $\varphi_i^{n+1} = \varphi_{3-i}^n$ on $\partial\Omega_i \backslash \partial\Omega$ and use them as test functions for (5) and (6)

$$\begin{aligned}
&\int_{\Omega\backslash\Omega_i} \nabla \bar{y}_{3-i}^n \nabla \varphi_{3-i}^n dx + \int_{\Omega\backslash\Omega_i} \sum_{\alpha=1}^N 2 \frac{\partial_{x_\alpha} g_{3-i}}{g_{3-i}} \partial_{x_\alpha} \bar{y}_{3-i}^n \varphi_{3-i} dx \\
&+ \int_{\Omega\backslash\Omega_i} \left(\frac{\Delta g_{3-i}}{g_{3-i}} - 2 \sum_{\alpha=1}^N \frac{\partial_{x_\alpha} g_{3-i}}{g_{3-i}} \right) \bar{y}_{3-i}^n \varphi_{3-i}^n dx + \int_{\Omega\backslash\Omega_i} H \bar{y}_{3-i}^n \varphi_{3-i}^n dx \\
&- \int_{\Omega\backslash\Omega_i} \int_0^T v_{3-i}^n \text{sign}\,(\tilde{y}_{3-i}^n) \exp\left(-\sqrt{H}t\right) dt \varphi_{3-i}^n dx \\
=& -\beta_i \Bigg\{ \int_{\Omega_i} \nabla \bar{y}_i^{n+1} \nabla \varphi_i^{n+1} dx + \int_{\Omega_i} \sum_{\alpha=1}^N 2 \frac{\partial_{x_\alpha} g_i^{n+1}}{g_i^{n+1}} \partial_{x_\alpha} \bar{y}_i^{n+1} \varphi_i^{n+1} dx \\
&+ \int_{\Omega_i} \left(\frac{\Delta g_i^{n+1}}{g_i^{n+1}} - 2 \sum_{\alpha=1}^N \frac{\partial_{x_\alpha} g_i^{n+1}}{g_i^{n+1}} \right) \bar{y}_i^{n+1} \varphi_i^{n+1} dx + \int_{\bar{\Omega}_i} H \bar{y}_i^{n+1} \varphi_i^{n+1} dx \\
&- \int_{\Omega_i} \int_0^T v_i^{n+1} \text{sign}\,(\tilde{y}_i^{n+1}) \exp\left(-\sqrt{H}t\right) dt \varphi_i^{n+1} dx \Bigg\}.
\end{aligned} \tag{7}$$

In the above equation choose φ_i^{n+1} to be $\bar{y}_i^{n+1}$. Then there exists a function ρ such that ρ is defined on $\Omega \backslash \Omega_i$ and

$$\begin{aligned}
\|\rho\|_{H^1(\Omega\backslash\Omega_i)} &\le C_1 \|\bar{y}_i^{n+1}\|_{H^1(\Omega_i)}, \\
\|\rho\|_{L^2(\Omega\backslash\Omega_i)} &\le C_1 \|\bar{y}_i^{n+1}\|_{L^2(\Omega_i)},
\end{aligned}$$

where C_1 is a positive constant depending on Ω, Ω_1, and Ω_2. Choose φ_{3-i}^n to be ρ, then for H large enough, (7) implies

$$\begin{aligned}
&\sum_{i=1}^{2} C_2 \Bigg\{ \frac{1}{2} \int_{\Omega\setminus\Omega_i} |\nabla \bar{y}_{3-i}^n|^2 dx + \frac{H}{2} \int_{\Omega\setminus\Omega_i} |\bar{y}_{3-i}^n|^2 dx \\
&\quad - \int_{\Omega\setminus\Omega_i} \int_0^T v_{3-i}^n \mathrm{sign}\,(\tilde{y}_{3-i}^n) \exp\left(-\sqrt{H}t\right)\, dt \bar{y}_{3-i}^n dx \Bigg\} \\
\geq &\sum_{i=1}^{2} \beta_i \Bigg\{ \frac{1}{2} \int_{\Omega_i} |\nabla \bar{y}_i^{n+1}|^2 dx + \frac{H}{2} \int_{\Omega_i} |\bar{y}_i^{n+1}|^2 dx \\
&\quad - \int_{\Omega_i} \int_0^T v_i^{n+1} \mathrm{sign}\,(\tilde{y}_i^{n+1}) \exp\left(-\sqrt{H}t\right)\, dt \bar{y}_i^{n+1} dx \Bigg\},
\end{aligned} \tag{8}$$

where C_2 is some constants depending only on the structure of the equation. In a similar way, we have

$$\begin{aligned}
&\sum_{i=1}^{2} C_3 \Bigg\{ \frac{1}{2} \int_{\Omega\setminus\Omega_i} |\nabla \bar{p}_{3-i}^n|^2 dx + \frac{H}{2} \int_{\Omega\setminus\Omega_i} |\bar{p}_{3-i}^n|^2 dx \\
&\quad - \gamma \int_{\Omega\setminus\Omega_i} \int_0^T y_{3-i}^{n-1} \mathrm{sign}\,(\tilde{p}_{3-i}^n) \exp\left(-\sqrt{H}t\right)\, dt \bar{p}_{3-i}^n dx \Bigg\} \\
\geq &\sum_{i=1}^{2} \beta_i \Bigg\{ \frac{1}{2} \int_{\Omega_i} |\nabla \bar{p}_i^{n+1}|^2 dx + \frac{H}{2} \int_{\Omega_i} |\bar{p}_i^{n+1}|^2 dx \\
&\quad - \gamma \int_{\Omega_i} \int_0^T y_i^n \mathrm{sign}\,(\tilde{p}_i^{n+1}) \exp\left(-\sqrt{H}t\right)\, dt \phi_i^{n+1} dx \Bigg\},
\end{aligned}$$

where ϕ_i^{n+1} plays a similar role as the role of ϕ_i^{n+1} in the estimate of $\bar{y}_i^{n+1}$

$$\begin{aligned}
\|\phi_i^{n+1}\|_{H^1(\Omega\setminus\Omega_i)} &\leq C_1 \|\bar{p}_i^{n+1}\|_{H^1(\Omega_i)}, \\
\|\phi_i^{n+1}\|_{L^2(\Omega\setminus\Omega_i)} &\leq C_1 \|\bar{p}_i^{n+1}\|_{L^2(\Omega_i)}.
\end{aligned}$$

Similarly as [15], taking β_i and H to be very large, and using the equation (as in [1])

$$\int_{(0,T)\times\Omega_i} (p_i^{n+1} + \alpha v_i^{n+1})(w_i - v_i^{n+1})\, dxdt \geq 0,$$

we get

$$\lim_{n\to\infty} (|||\nabla y_i^n||| + |||y_i^n||| + |||\nabla p_i^n||| + |||p_i^n|||) = 0.$$

Notice that the fact $|||\nabla y_i^n|||$, $|||y_i^n|||$, $|||\nabla p_i^n|||$, $|||p_i^n|||$, $|||v_i^n|||$ are well-defined is also included in the convergence result.

Theorem 3.1 *The algorithm converges in the following sense:*

$$\lim_{n\to\infty}(|||\nabla y_i^n||| + |||y_i^n||| + |||\nabla p_i^n||| + |||p_i^n||| + |||v_i^n|||) = 0.$$

Acknowledgement

The author would like to thank the editors for a kind invitation to write this paper for the proceedings of the Appl. Math. Conference 2013. The author has been supported by Grant MTM2011-29306-C02-00, MICINN, Spain, ERC Advanced Grant FP7-246775 NUMERIWAVES, and Grant PI2010-04 of the Basque Government.

References

[1] Benamou, J.-D.: Domain decomposition, optimal control of systems governed by partial differential equations, and synthesis of feedback laws. J. Optim. Theory Appl. **102**(1) (1999), 15–36.

[2] Benamou, J.-D. and Després, B.: A domain decomposition method for the Helmholtz equation and related optimal control problems. J. Comput. Phys. **136**(1) (1997), 68–82.

[3] Benamou, J.-D.: Décomposition de domaine pour le réarrangement monotone d'applications vectorielles. C. R. Acad. Sci. Paris Sér. I Math. **315**(4) (1992), 469–474.

[4] Benamou, J.-D.: Décomposition de domaine pour le contrôle optimal de systèmes gouvernés par des équations aux dérivées partielles elliptiques. C. R. Acad. Sci. Paris Sér. I Math. **317**(2) (1993) 205–209.

[5] Benamou, J.-D.: A domain decomposition method for the optimal control of systems governed by the Helmholtz equation. In: *Mathematical and numerical aspects of wave propagation (Mandelieu-La Napoule, 1995)*, pp. 653–662. SIAM, Philadelphia, PA, 1995.

[6] Benamou, J.-D.: A domain decomposition method with coupled transmission conditions for the optimal control of systems governed by elliptic partial differential equations. SIAM J. Numer. Anal. **33**(6) (1996), 2401–2416.

[7] Benamou, J.-D.: Décomposition de domaine pour le contrôle de systèmes gouvernés par des équations d'évolution. C. R. Acad. Sci. Paris Sér. I Math. **324**(9) (1997), 1065–1070.

[8] Bensoussan, A., Glowinski, R., and Lions, J.-L.: Méthode de décomposition appliquée au contrôle optimal de systèmes distribués. In: *Fifth Conferences on Optimization Techniques (Rome, 1973), Part I, Lecture Notes in Comput. Sci.*, vol. 3, pp. 141–151. Springer, Berlin, 1973.

[9] Lagnese, J. E. and Leugering, G.: Time-domain decomposition of optimal control problems for the wave equation. Systems Control Lett. **48**(3–4) (2003), 229–242. Optimization and control of distributed systems.

[10] Lagnese, J. E. and Leugering, G.: Domain decomposition methods in optimal control of partial differential equations. *International Series of Numerical Mathematics*, vol. 148, Birkhäuser Verlag, Basel, 2004.

[11] Leugering, G.: Domain decomposition in optimal control problems for partial differential equations revisited. In: *Control theory of partial differential equations*, *Lect. Notes Pure Appl. Math.*, vol. 242, pp. 125–155. Chapman & Hall/CRC, Boca Raton, FL, 2005.

[12] Leugering, G.: Domain decomposition of optimal control problems for dynamic networks of elastic strings. Comput. Optim. Appl. **16**(1) (2000), 5–27.

[13] Leugering, G.: Domain decomposition of constrained optimal control problems for 2D elliptic system on networked domains: convergence and a posteriori error estimates. In: *Domain decomposition methods in science and engineering XVII*, *Lect. Notes Comput. Sci. Eng.*, vol. 60, pp. 119–130. Springer, Berlin, 2008.

[14] Toselli, A. and Widlund, O.: Domain decomposition methods – algorithms and theory. *Springer Series in Computational Mathematics*, vol. 34. Springer-Verlag, Berlin, 2005.

[15] Tran, M.-B.: Optimized overlapping domain decomposition: Convergence proofs. *Domain Decomposition Methods in Science and Engineering XX; Series: Lecture Notes in Computational Science and Engineering, Springer*. To appear.

[16] Tran, M.-B. Overlapping optimized Schwarz methods for parabolic equations in n-dimensions. *Proceedings of the American Mathematical Society*. To appear.

[17] Tran, M.-B. Parallel Schwarz waveform relaxation method for a semilinear heat equation in a cylindrical domain. C. R. Math. Acad. Sci. Paris **348**(13-14) (2010), 795–799.

[18] Tran, M.-B.: A parallel four step domain decomposition scheme for coupled forward-backward stochastic differential equations. J. Math. Pures Appl. (9) **96**(4) (2011), 377–394.

Conference Applications of Mathematics 2013
in honor of the 70th birthday of Karel Segeth.
Jan Brandts, Sergey Korotov, Michal Křížek,
Jakub Šístek, and Tomáš Vejchodský (Eds.),
Institute of Mathematics AS CR, Prague 2013

ON THE COMPUTATIONAL IDENTIFICATION OF TEMPERATURE-VARIABLE CHARACTERISTICS OF HEAT TRANSFER

Jiří Vala

Brno University of Technology, Faculty of Civil Engineering,
Department of Mathematics and Descriptive Geometry
Veveří 95, 602 00 Brno, Czech Republic
vala.j@fce.vutbr.cz

Abstract

The mathematical analysis of a heat equation and its solutions is a standard part of most textbook of applied mathematics and computational mechanics. However, serious problems from engineering practice do not respect formal simplifications of such analysis, namely at high temperatures, for phase-change materials, etc. This paper, motivated by the material design and testing of a high-temperature thermal accumulator, as a substantial part of the Czech-Swedish project of an original equipment for exploiting solar energy using optical fibres, demonstrates the possibility of both direct and inverse analysis, physically transparent and mathematically correct, paying attention to the set of basic temperature-variable characteristics of thermal transfer.

1. Introduction

Most textbooks, both from applied mathematics and computational mechanics, present a heat transfer equation as a slightly modified Poisson equation, supplied by standard Dirichlet or Neumann boundary conditions, with a few constant material characteristics. Consequently, some general analytic results, as [2], p. 184, or at least semi-analytic ones, making use of the Fourier method by [2], p. 219, can be derived. Applying the variational approach, the existence and uniqueness of solution of a linear equation can be verified using the Lax-Milgram theorem together with some basic facts from the variational calculus; moreover, for the convergence of sequences of approximate solutions, the proper error analysis both for space and time discretization, applying various approaches by [17], is available. However, it is not easy to find such ideal closed simple systems in the nature. All engineering applications, especially in the design of advanced materials, structures and technologies (where sufficiently long experience with their behaviour is missing) work with materials of complicated micro-structure, including potential phase changes. Their effective material characteristics cannot be evaluated in a simple way and may not exist at

all in any reasonable sense, at least in that using some standard (e.g. two-scale) periodic homogenization, as discussed in [5], p. 204, or its generalization (including non-periodic phenomena and stochastic analysis) by [8]. Even in the homogeneous and isotropic case, at least from the macroscopic point of view, the determination of material parameters, making use of incomplete data from available experiments, can generate non-trivial inverse problems, not covered by [11], p. 255. To compensate the usual lack of input data, the formulation of such identification problems should avoid all multi-physical considerations, as the hygro-thermo-chemo-mechanical ones in [21] (and in a lot of papers referenced there), based on the complete set of conservation laws of continuum thermomechanics by [3], p. 4, i. e. for mass, (linear and angular) momentum and energy (or enthalpy), related to particular material components, including their phase changes. Even in the case of reflective insulation layers with air gaps or layers, reviewed in [12] and [9], most authors try to avoid (as much as possible) any methods of computational fluid dynamics, to obtain some simplified formulae for energy conservation only. However, the need of knowledge of results from various research areas justifies the extensive list of references even in this paper.

Figure 1: Experiments with the exploitations of solar energy (Hudiksvall, Sweden).

The principal motivation for the deeper analysis of heat transfer phenomena, sketched in this paper, comes from the Czech-Swedish project of the advanced exploitation of solar energy using optical fibres (cf. *Acknowledgements*). The left-hand part of Fig. 1 illustrates the development of the needed technological equipment, whereas its right-hand part shows one model (a representative from several alternatives) of the heat accumulator, whose effective functionality at high temperatures (up to 1000 °C) is a crucial part of the whole system; more information (without technical details) can be found in [18]. Fig. 2 shows a hot-wire measurement for the identification of material characteristics under standard laboratory conditions at the Faculty of Civil Engineering of Brno University of Technology. This method is open to its upgrade to high temperatures (more expensive components for a measurement device are necessary); another active cooperation exists with PD-Refractories CZ (former Moravian Fire and Schistous Clay Works) in Velké Opatovice. Nevertheless,

Figure 2: A simple equipment for the non-stationary hot-wire measurements (Brno University of Technology, Czech Republic).

the computational approach of [1], related to this method (much better than valid European technical standards), based on the simplifying physical and geometrical assumptions and on the properties of Bessel functions, needs substantial improvements just in the case of high temperatures.

To demonstrate a (nearly) realistic computational problem without complicated notations and technical difficulties, we shall consider, apart from its material microstructure, a homogeneous and isotropic material, whose thermal behaviour can be studied using the energy balance in the solid phase by [3], p. 7, without any changes in geometrical configuration, in the 3-dimensional Euclidean space R^3 and at the time interval $I = \langle 0, \tau \rangle$ for some positive τ. Usually such material is surrounded by other layers from the measurement system, whose properties should be a priori known, as explained in [20]; here we shall consider only a separate material specimen, located in some open set Ω in R^3, with all boundary conditions prescribed on the boundary $\partial\Omega$ of Ω in R^3. The heat conduction in the specimen will be conditioned by the heat convection and radiation from its environment. We shall study i) how the temperature-dependent material characteristics can be inserted both to the direct calculations of the time development of unknown temperature fields for a priori known values of such characteristics, solving standard initial and boundary value problems, ii) how these characteristics can be evaluated in the case of overdetermined boundary conditions.

2. Direct problems

Let us consider some system of Cartesian coordinates $x = (x_1, x_2, x_3)$ in R^3 and the time variable $t \in I$; upper dot symbols will be reserved for the derivatives with respect to t, prime symbols for the ordinary derivatives of functions of one real variable, ∇ for $(\partial/\partial x_1, \partial/\partial x_2, \partial/\partial x_3)$ and $\cdot$ for scalar products of vectors from R^3. The most frequently used heat transfer equation in the literature is

$$c(\theta)\dot{\theta} - \nabla \cdot (\kappa(\theta)\nabla\theta) + f = 0 \qquad \text{on } \Omega \times I, \tag{1}$$

where $\theta(x,t)$ [K] is the unknown temperature, $\kappa(\theta)$ [W/(m·K)] the heat conductivity

(crucial for the thermal insulation ability of a material) $c(\theta)$ [J/m^3·K] the heat capacity (important for the thermal accumulation property) and f [W/m^3] the volume heat source. The thermal diffusivity $\alpha(\theta) := \kappa(\theta)/c(\theta)$ [m^2/s] occurs frequently, too.

The obvious initial condition is

$$\theta(.,0) = \theta_0 \qquad \text{on } \Omega\,; \tag{2}$$

θ_0 has to be prescribed. The boundary $\partial\Omega$ is supposed to contain a set Γ, where (in general nonlinear) boundary conditions of the type

$$\kappa(\theta)\nabla\theta\cdot\nu + \varphi(\theta,\theta_e)(|\theta|^{n-1}\theta - \theta_e^n) + g = 0 \qquad \text{on } \Gamma\times I \tag{3}$$

are satisfied; here g [W/m^2] is the surface heat source, $\nu := (\nu_1,\nu_2,\nu_3)$ denotes the unit normal vector to Γ (with the outside orientation), θ_e means the temperature of the environment and $\varphi(\theta,\theta_e)$ [W/(m^2·K^n)]refers to some boundary characteristic, related to a real $n \geq 1$ (mostly integer in technical applications), namely to that for interface convection by [6], p. 37, with $n = 1$, or that for interface radiation by [6], p. 116, with $n = 4$, well-known as the Stefan-Boltzmann law; the natural and simple generalization is to combine a finite number of additive terms of such type on the left-hand side of (3). Here we can see that even in the case of constant c we cannot substitute, like (1), κ by α totally, because it cannot be removed from (3) except the case of (practically) empty Γ. We shall also assume that

$$\theta = \theta_e \qquad \text{on } \Theta\times I, \tag{4}$$

where Θ is some part of the boundary $\partial\Omega$; in this section we shall consider disjoint Θ and Γ, whose closure covers the whole boundary $\partial\Omega$.

Nevertheless, following [3], p. 8, for the energy balance the most important quantity is the internal energy $\varepsilon(x,t)$ [W/kg]; thus we have

$$(\rho(\varepsilon)\varepsilon)^{\cdot} - \nabla\cdot(\sigma(\varepsilon)\nabla\varepsilon) + f = 0 \qquad \text{on } \Omega\times I, \tag{5}$$

where $\rho(\varepsilon)$ is the material density [kg/m^3] and $\sigma(\varepsilon)$ [kg/m] is the new material characteristic, expected to be replaced using $\kappa(\theta)$ from (3); from the point of view of practical measurements the values of θ can be obtained much easier than those of ε. Frequently $\rho(\varepsilon)\dot{\varepsilon}$ occurs instead of the first additive term in (5), referring to the mass conservation; however, some applications, e. g. [16], studying the early-age behaviour of concrete mixtures, require variable ρ due to the change of material structure, thus we are only allowed to define $\overline{\rho}(\varepsilon) := \rho(\varepsilon) + \rho'(\varepsilon)\varepsilon$ and write $\overline{\rho}(\varepsilon)\dot{\varepsilon}$ instead of the first additive term in (5). Dividing (5) by $\overline{\rho}(\varepsilon)$, assumed to be non-zero, we receive

$$\dot{\varepsilon} - \nabla\cdot(a(\varepsilon)\nabla\varepsilon) + b(\varepsilon)\nabla\varepsilon\cdot\nabla\varepsilon + \overline{f}(\varepsilon) = 0 \qquad \text{on } \Omega\times I, \tag{6}$$

with $a(\varepsilon) := \sigma(\varepsilon)/\overline{\rho}(\varepsilon)$ [m^2], $b(\varepsilon) := -\sigma(\varepsilon)\overline{\rho}'(\varepsilon)/\rho^2(\varepsilon)$ [m^2] and $\overline{f}(\varepsilon) := f/\overline{\rho}(\varepsilon)$ [W/kg].

Let us now introduce the following simplified notation: let $\widehat{\psi}(.)$ be an arbitrary real function with its derivative identical with some given real function $\psi(.)$ (defined up to an additive constant). Using such notation, we are able to set $\varepsilon = \widehat{c}_m(\theta)$,

where $c_m(\theta)$ [J/(kg·K)] denotes the heat capacity related to the unit mass (unlike c related to the unit volume); consequently $\theta = \widehat{c}_m^{-1}(\varepsilon)$. Thus we obtain $\sigma(\varepsilon)\nabla\varepsilon = \sigma(\widehat{c}_m(\theta))\nabla(\widehat{c}_m(\theta)) = \sigma(\widehat{c}_m(\theta))c_m(\theta)\nabla\theta = \kappa(\theta)\nabla\theta$, which implies $\sigma(\varepsilon) = \kappa(\theta)/\overline{\rho}(\varepsilon)$. Similarly $(\rho(\varepsilon)\varepsilon)^{\cdot} = \overline{\rho}(\varepsilon)\dot{\varepsilon} = \overline{\rho}(\varepsilon)c_m(\theta)\dot{\theta} = c(\theta)\dot{\theta}$ gives $c(\theta) = \overline{\rho}(\varepsilon)c_m(\theta)$. Consequently we are able to evaluate the thermal diffusivity from the (not very simple) formula $\alpha(\theta) = \kappa(\theta)/(c_m(\theta)\overline{\rho}(\widehat{c}_m(\theta)))$. Another important information is that for positive values of κ and $\overline{\rho}$ and negative values of $\overline{\rho}'$ (which is the physically realistic setting) both factors a and b in (6) remain positive.

The initial and boundary conditions, as a simple analogy to (2), (3), and (4), are

$$\varepsilon(.,0) = \varepsilon(\theta_0) \qquad \text{on } \Omega\,, \tag{7}$$

$$\sigma(\varepsilon)\nabla\varepsilon\cdot\nu + \varphi(c_m^{-1}(\varepsilon),\theta_e)(|c_m^{-1}(\varepsilon)|^{n-1}c_m^{-1}(\varepsilon) - \theta_e^n) + g = 0 \quad \text{on } \Gamma\times I\,, \tag{8}$$

$$\varepsilon(\theta) = \varepsilon(\theta_e) \qquad \text{on } \Theta\times I\,. \tag{9}$$

To find the solution, i. e. the space- and time- variable temperature field ε (and consequently to express θ, too), of (6) with the initial conditions (7) and the boundary conditions (8) and (9) in a reasonable sense, admitting its numerical analysis, in some appropriate space of mappings from I to Lebesgue and Sobolev spaces defined on Ω and $\partial\Omega$ is not easy because of the presence of various type of nonlinearities in (5) and (9). Some interesting ideas and partial existence and uniqueness results can be found in [14], referring to the former analysis of [7]. However, the set of formal simplifying assumptions hidden there does not enable to handle realistic engineering problems, as needed in this paper.

Seemingly it could be useful to formulate a similar problem to the just discussed one for θ directly, without any transformation using ε. Indeed, dividing (1) by $c(\theta)$ (whose values are positive usually), we receive

$$\dot{\theta} - \nabla\cdot(a_*(\theta)(\kappa(\theta)\nabla\theta)) + b_*(\theta)\nabla\cdot\nabla\theta + f_*(\theta) = 0 \qquad \text{on } \Omega\times I \tag{10}$$

with $a_*(\theta) := \kappa(\theta)/c(\theta)$, $b_*(\theta) := -\kappa(\theta)c'(\theta)/c^2(\theta)$, and $f_*(\theta) := f/c(\theta)$, thus we should find the solution of (10) with the boundary conditions (3) and (4) and the initial condition (2). The arguments on the positive values of a_* and b_* (instead of those related to a and b) can be repeated, but at least (10) is even more complicated than (6) and difficulties similar to those in [14] can be expected.

Some difficulties of the above mentioned type can be removed using the Kirchhoff transformation $u = \widehat{c}(\theta)$ [W/m^3], seemingly the slight modification of the discussed $\varepsilon = \widehat{c}_m(\theta)$; consequently $\theta = c^{-1}(u)$. Now we have $\dot{u} = \overline{\rho}(\varepsilon)\dot{\varepsilon}$ and, introducing $\beta(u) := \widehat{\kappa}(c^{-1}(u))$, also $\nabla\beta(u) = \beta'(u)\nabla u$. Then (1) can be converted to the form

$$\dot{u} - \nabla\cdot\nabla\beta(u) + f = 0 \qquad \text{on } \Omega\times I \tag{11}$$

and supplied by the initial and boundary conditions

$$u(.,0) = u_0 \qquad \text{on } \Omega, \tag{12}$$

$$\nabla\beta(u)\cdot\nu + \psi_e(u)(|\gamma(u)|^{n-1}\gamma(u) - \theta_e^n) + g = 0 \qquad \text{on } \Gamma\times I\,, \tag{13}$$

$$u = u_e \qquad \text{on } \Theta\times I, \tag{14}$$

with $u_0 := u(\theta_0)$, $u_e := u(\theta_e)$, $\gamma(u) := \widehat{\kappa}^{-1}(u)$ and $\psi_e(u) := \varphi(\kappa^{-1}(u),\theta_e)$.

For the sake of simplicity, let us now assume that Ω is a domain in R^3 with a sufficiently smooth boundary to satisfy theorems on the Sobolev and Lebesgue spaces introduced on Ω and $\partial\Omega$, namely the Sobolev (compact) imbedding, the Poincaré-Friedrichs and the trace theorems by [15], p. 17; much more general geometrical configurations (bringing unpleasant technical difficulties) are discussed in [13], pp. 62, 222, and 385. Moreoever, let β', γ and ψ_e be continuous real functions and Θ be an empty set (this last assumption will be removed soon). One can see that, even for $n = 1$ in (13), the classical theory of monotone operators by [10], p. 243, is not applicable, because the monotonicity is violated for any non-constant β' or γ, thus more general results on pseudomonotone or weakly continuous operators are needed. Applying the standard notation of Sobolev, Lebesgue and Bochner spaces, let us choose $V = W^{1,2}(\Omega)$ with its dual space V^* and consider $u_0 \in V$, $f \in L^2(I, L^{6/5}(\Omega))$ and $g, \theta_e^n \in L^2(I, L^{4/3}(\Gamma))$. In all following considerations, δ will be some positive constant (a priori known, small in practice). Let us suppose that $\beta'(r) \geq \delta$, $1/\delta \geq \gamma(r) \geq \delta$ and $1/\delta \geq \psi_e(r) \geq \delta$ for any $r \in R$. Then by [15], p. 237 (after rather long verification of abstract assumptions), thanks to the properties of quasilinear pseudomonotone mappings, the problem formulated by (11), (12) and (13) has a weak solution $u \in W^{1,2,2}(I, V, V^*)$ in the sense

$$\begin{aligned}
&\dot{u}(t)v(t) + \int_\Omega \beta'(u(x,t))\nabla u(x,t) \cdot \nabla v(x)\,\mathrm{d}x \\
&\qquad + \int_\Gamma \psi_e(u(x,t))(|\gamma(u(x,t))|^{n-1}\gamma(u(x,t)) - \theta_e^n(x,t))v(x)\,\mathrm{d}s(x) \qquad (15)\\
&\qquad = \int_\Omega f(x,t)v(x)\,\mathrm{d}x - \int_\Omega g(x,t)v(x)\,\mathrm{d}s(x) \\
&\text{for all } v \in V \text{ and almost every } t \in I
\end{aligned}$$

if $\beta'(r) \leq 1/\delta$ for any $r \in R$ and $n = 1$. Moreover, by [15], p. 241, thanks to the properties of quasilinear weakly continuous mappings, the same problem has a very weak solution $u \in L^2(I, V)$ in the sense

$$\begin{aligned}
&\int_\Omega (u(x,\tau)v(x,\tau) - u_0(x)v(x,0))\,\mathrm{d}x \\
&+ \int_I \int_\Omega \beta'(u(x,t))\nabla(u(x,t)) \cdot \nabla(v(x,t))\,\mathrm{d}x\,\mathrm{d}t \\
&+ \int_I \int_\Gamma \psi_e(u(x,t))(|\gamma(u(x,t))|^{n-1}\gamma(u(x,t)) - \theta_e^n(x,t))v(x,t)\,\mathrm{d}s(x)\,\mathrm{d}t \qquad (16)\\
&= \int_I \int_\Omega f(x,t)v(x,t)\,\mathrm{d}x\,\mathrm{d}t - \int_I \int_\Omega g(x,t)v(x,t)\,\mathrm{d}s(x)\,\mathrm{d}t \\
&\text{for all } v \in W^{1,\infty,\infty}(I, W^{1,\infty}(\Omega), L^{6/5}(\Omega))
\end{aligned}$$

if $\beta'(r) \leq (1 + |r|^{5/3-\delta})/\delta$ for any $r \in R$ and $n \leq 2$.

We can see that the very weak solution, unlike the weak one, admits e. g. the linear growth of $\beta'(r)$, which is useful in practice. However, the requirement $n \leq 2$ is not realistic, namely in the analysis of radiation effects. The remedy is to choose V as the space of all v from $W^{1,2}(\Omega)$, whose traces belong to $L^n(\Gamma)$; the properties

of such spaces are discussed in [15], pp. 64 and 253. Another needed generalization is to remove the assumption $\Gamma = \partial\Omega$. This can be done using the transformation $\widetilde{u} = u - u_*$, where some u_* from the same space, as required for u, satisfies (14) instead of u. Consequently we have only $\widetilde{u}(.,0) = u_0 - u_*(.,0)$ on Ω instead of (12), in addition to (11) and (13) in their slightly modified forms containing $\widetilde{u}$; then it is sufficient to take the subspace of all functions from V with zero values on Θ instead of V. However, the practical construction of u_* may not be easy.

Since the derivation of solutions (15) and (16) is based on the Rothe sequences and Galerkin approximations, the numerical construction of sequences of approximate solutions is available, although the verification of their convergence is not trivial because of the presence of non-linear terms in (15) and (16). However, in any algorithm of discretization in time, based on the Euler implicit, Crank-Nicholson or similar schemes, it is natural to take arguments of $\beta'(.)$, $\gamma(.)$, $\psi_e(.)$ and $|.|$ from the preceding time step, thus we obtain only linear systems; the proper convergence analysis then relies on various compactness theorems. Let us also notice that some our assumptions can be weakened, e. g. it is possible to work with arbitrary $f \in L^1(\Omega \times I)$; however, the derivation of relevant results, using accretive mappings and nonlinear semigroups, by [15], p. 291, does not seem friendly to the construction of simple computational algorithms.

3. Inverse problems

Due to the limited extent of this paper, we shall refer to the notations and considerations of the previous section as much as possible. The first step in the inverse analysis then admits the intersection $\Gamma \cup \Theta$ with non-zero measure on $\partial\Omega$, compensating the imperfect knowledge of β', γ and ψ_e. It is then useful to introduce $\Xi := \Theta \setminus \Gamma$ and $\Psi := \Gamma \setminus \Theta$ (in direct problems clearly $\Xi = \Theta$ and $\Psi = \Gamma$). Let P be a set of admissible parameters; its simplest choice can be a closed set in R^N with an integer number N of unknown parameters. Now we can consider $\beta'(r,p)$, $\gamma(r,p)$ and $\psi_e(r,p)$ as functions of $(r,p) \in R \times P$, instead as functions defined on R only. We shall suppose that all these functions satisfy assumptions of (15) or (16), taking into account their above sketched generalizations, too, for arbitrary $p \in P$.

Following [4], pp. 123 and 368, it is natural to define

$$F(p) = \int_I \int_\Xi |u(x,t,p) - u_e(x,t)|^\omega \, \mathrm{d}s(x) \, \mathrm{d}t, \tag{17}$$

where $1 \leq \omega \leq \infty$ (the well-known choice is the classical least-squares one, i. e. $\omega = 2$); $\Theta \times I$ in (14) must be reduced to $\Xi \times I$ in all direct problems (with fixed p). Minimizing F, which can be interpreted as an error in our overdetermined problem where $u(.,.,p) \approx u_e(.,.)$ on $\Xi \times I$, is required in some reasonable sense (the equality here is not realistic because of the inexact measurements of u_e and other input data, our physical and geometrical simplifying assumptions, disturbing effects from other physical processes, etc.). Let us notice that F is only a function of N real variables here, with respect to p. The setting of p enables us to identify all material characteristics

completely (although the corresponding algebraic manipulations may not be quite easy).

Another access is seemingly available, too: to define

$$G(p) = \int_I \int_\Xi |\nabla\beta(u(x,t,p)\cdot\nu(x) + \psi_e(u(x,t,p))(|\gamma(u(x,t,p))|^{n-1}\gamma(u(x,t,p)) - \theta_e^n(x,t)) + g(x,t)|^\omega \,\mathrm{d}s(x)\,\mathrm{d}t \tag{18}$$

similarly to (17); $\Gamma \times I$ in (13) must be reduced to $\Psi \times I$ in all direct problems. Minimizing G, also interpretable as an error of the (exactly zero) term $|.|$ in (18), analogous to that of F, but formulated (from the physical point of view) for the interface heat fluxes instead of the interface temperature, is possible, but rarely used in practice because i) the evaluation of G (and its derivatives) in (18) is much more difficult that that of F in (17) and ii) the reliability of recorded values of g is usually much lower than that of θ_e in most engineering applications, including that mentioned in *Introduction*.

Let us pay attention to (17) only. Let us assume that P is a closed bounded set in R^N, thus (because N is finite) it must be compact. To verify the existence of some minimum of (17), by [10], p. 191, it is then sufficient to prove its continuity. However, it is not quite simple, even in the case (15) and $\omega = 2$, although it seems to be easy i) to consider a sequence of $p_k \in P$ with $k \in \{1, 2, \ldots\}$ with the limit $p \in P$, ii) to derive a corresponding $u_k(.,.,p_k)$ by (15) to p_k, as well as $u(.,.,p)$ to p, iii) to insert $v(.) = u_k(.,.,p_k) - u(.,.p)$ into (15) with p_k and into (15) with p and calculate their diference, iv) to integrate the result over I to try to get estimates of $u_k(.,.,p_k) - u(.,.,p)$ in appropriate norms following [10], p. 264. The lack of monotonicity, crucial for iv), has to be overcome by more advanced tricks, inspired by the sequence of exercises from [15], p. 66.

The sketched approach gives us only one rough information on the uncertainty of identified characteristics: the minimal value of F. The further step of the inverse analysis, motivated by [22], then should be to interpret P as a sample space of elementary events, supplied by the minimal σ-algebra and by certain probability measure $\mathcal{P}$. Then, instead of (17), we should minimize

$$\Phi(p) = \int_P \int_I \int_\Xi |u(x,t,p) - u_e(x,t)|^\omega \,\mathrm{d}s(x)\,\mathrm{d}t\,\mathrm{d}\mathcal{P}\,, \tag{19}$$

with respect to all other modified conditions, improved by $\mathcal{P}$. Some preparatory results of such type for a linearized heat transfer problem, including much more references, remarks to direct, sensitivity and adjoint problems and to the convergence analysis of nonlinear conjugate gradient algorithms, generalizing the Newton-type ones, applicable to (17) (although the exact values of derivatives cannot be computed easily), to minimize Φ, have been presented in [19]. Unfortunately, the general case contains still open problems because of the absence of such lemmas, as the (generalized) Aubin-Lions one by [15], p. 194, crucial for the compactness results in the deterministic case, and corresponding interpolation ones; this makes it difficult to replace I from (17) by $I \times P$ from (19) with some probabilistic measure.

4. Conclusion

We have shown that the proper analysis of the heat transfer equation with temperature-variable characteristics, including the inverse problem of identification of such characteristics, open to the uncertainty estimates, too, brings substantial difficulties in comparison with the linearized model problems. However, these difficulties can be overcome by means of recent functional and numerical analysis. More detailed considerations (including complete proofs) should be published in the near future.

The further research is motivated by the design of thermal accumulator, mentioned in *Introduction*, although the deep mathematical analysis does not seem to be its most important part. Some original experimental devices and MATLAB-based software packages have been prepared; the complete technical equipment must be functional until the end of 2014.

Acknowledgements

This work was supported by grant No. 02021231 of the Technology Agency of the Czech Republic.

References

[1] André, S., Rémy, B., Pereira F. R., and Cella, N.: Hot wire method for the thermal characterization of materials: inverse problem application. Engenharia Térmica **4** (2003), 55–64.

[2] Barták, J., Herrmann, L., Lovicar, V., and Vejvoda, O.: *Partial differential equations of evolution*. Ellis Horwood, Chichester, 1991.

[3] Bermúdez de Castro, A.: *Continuum thermomechanics*. Birkhäuser, Basel, 2005.

[4] Bochev, P. B., and Gunzburger, M. D.: *Least-squares finite element methods*. Springer, New York, 2009.

[5] Cioranescu, D., and Donato, P.: *An introduction to homogenization*. Oxford University Press, Oxford, 1999.

[6] Davies, M. D.: *Building heat transfer*. J. Wiley & Sons, Hoboken, 2004.

[7] Feireisl, E., Petzeltová, H., and Simondon, F.: Admissible solutions for a class of nonlinear parabolic problems with non-negative data. Proceedings of the Royal Society in Edinburgh, Section A – Mathematics, **131** (2001), 857–883.

[8] Franců, J., and Svanstedt N. E. M.: Some remarks on two-scale convergence and periodic unfolding. Appl. Math. **57** (2012), 359–375.

[9] Fricker, J. M., and Yarbrough, D.: Review of reflective insulation estimation methods. In: *Proceedings of Building Simulation* in Sydney, 1989–1996. ARIAH (Australian Institute of Refrigeration, Air Conditioning and Heating), Sydney, 2011.

[10] Fučík, S., and Kufner, A.: *Nonlinear differential equations*. Elsevier, Amsterdam, 1980.

[11] Isakov, V.: *Inverse problems for partial differential equations*. Springer, New York, 2006.

[12] Levinson, R., Akbari, H., and Gartland, L. M.: Impact of temperature dependency of fiberglass insulation R-value on cooling energy use in buildings, 10.85–10.94. In: *Proceedings of ACEEE Summer Studies on Energy Efficiency in Buildings* in Pacific Grove (California, USA). ACEEE (American Council for an Energy-Efficient Economy), 1996.

[13] Maz'ya V. G.: *Prostranstva S. L. Soboleva*. Izdateľstvo Leningradskogo universiteta, Leningrad (St. Petersburg), 1985. (In Russian.)

[14] Rincon, M. A., Límaco, J., and Liu, I. -S.: Existence and uniqueness of solutions of a nonlinear heat equation. Tendências em Matemática Aplicada e Computacional **6** (2005), 273–284.

[15] Roubíček, T.: *Nonlinear partial differential equations with applications*. Birkhäuser, Basel, 2005.

[16] Schrefler, B. A., Pesavento, F., and Gawin, D.: Multiphase model for concrete: numerical solutions and applications. In: *Proceedings of WSEAS International Conference on Applied and Theoretical Mechanics* in Venice, pp. 108–116. WSEAS (World Scientific and Engineering Academy and Society), 2006.

[17] Segeth, K.: A review of some a posteriori error estimates for adaptive finite element methods. Math. Comput. Simulation **80** (2010), 1589–1600.

[18] Šťastník, S. and Vala, J.: Identification of thermal characteristics of a high-temperature thermal accumulator. In: *Proceedings of Thermophysics* in Podkylava (Slovak Republic), pp. 214–222. Slovak Technical University, Bratislava, 2012.

[19] Šťastník, S., and Vala, J.: Identifikace tepelných vlastností materiálu pro vysokoteplotní zásobník. In: *Sborník konference Maltoviny* in Brno, 21 pp. Brno University of Technology, Brno, 2012, to appear. (In Czech.)

[20] Vala, J.: Least-squares based technique for identification of thermal characteristics of building materials. International Journal of Mathematics and Computers in Simulation **5** (2011), 126–134.

[21] Vala, J.: Multiphase modelling of thermomechanical behaviour of early-age silicate composites. In: M. El-Amin (Ed.), *Mass Transfer in Multiphase Systems and its Applications*, Chap. 3. InTech, Rijeka, 2011.

[22] Zabaras, N.: Inverse problems in heat transfer. In: W. J. Minkowycz, E. M. Sparrow, and J. S. Murthy (Eds.), *Handbook on Numerical Heat Transfer*, Chap. 17. J. Wiley & Sons, Hoboken, 2004.

Conference Applications of Mathematics 2013
in honor of the 70th birthday of Karel Segeth.
Jan Brandts, Sergey Korotov, Michal Křížek,
Jakub Šístek, and Tomáš Vejchodský (Eds.),
Institute of Mathematics AS CR, Prague 2013

A DIRECT SOLVER FOR FINITE ELEMENT MATRICES REQUIRING $O(N \log N)$ MEMORY PLACES

Tomáš Vejchodský

Institute of Mathematics, Academy of Sciences
Žitná 25, CZ-115 61 Prague 1, Czech Republic
vejchod@math.cas.cz

Abstract

We present a method that in certain sense stores the inverse of the stiffness matrix in $O(N \log N)$ memory places, where N is the number of degrees of freedom and hence the matrix size. The setup of this storage format requires $O(N^{3/2})$ arithmetic operations. However, once the setup is done, the multiplication of the inverse matrix and a vector can be performed with $O(N \log N)$ operations. This approach applies to the first order finite element discretization of linear elliptic and parabolic problems in triangular domains, but it can be generalized to higher-order elements, variety of problems, and general domains. The method is based on a special hierarchical enumeration of vertices and on a hierarchical elimination of suitable degrees of freedom. Therefore, we call it hierarchical condensation of degrees of freedom.

1. Introduction

This paper is devoted to Prof Karel Segeth on the occasion of his 70th birthday. Karel stood at the very beginning of my scientific career as the supervisor of my Master thesis and since then we have continued to work together as collaborators and good friends until today. I am thankfull to him for many things he taught me, for a lot of help and constant support. Karel has been interested in several topics during his professional career. Efficient solution of large and sparse linear algebraic systems, which is the topic of this paper, is one of them [2, 14, 15, 18]. In addition, Karel studied the method of lines [13, 16, 19], higher-order finite elements [21], and hierarchical approaches [17, 20]. These techniques are utilized below as well.

Solvers of large and sparse linear algebraic systems stemming from discretizations of partial differential equations are considered as the bottleneck of scientific computing. Therefore, the efficiency of these solvers is of paramount importance. In this contribution we concentrate on the lowest-order triangular finite element discretization [3, 23], which is one of the most often used discretization methods that naturally yields large and sparse systems of linear algebraic equations. The sparse direct solvers and preconditioned iterative methods are two principal approaches how to solve such systems. The literature on this subject is vast. The interested reader can consult books [4, 5, 6, 10, 11, 12] and references therein.

In this contribution, we present a method that can be classified as a direct sparse solver. The idea is based on hierarchically applied static condensation of internal degrees of freedom (DOFs). The static condensation is often used in higher-order finite element methods [21], where so-called internal (or bubble) DOFs appear. These DOFs can be easily eliminated from the system in such a way that the resulting Schur complement system is of smaller dimension, better conditioned, and it keeps the original sparsity structure. See e.g. [25] for more details.

In this paper we consider the lowest-order finite element methods, where no internal DOFs exist. However, we propose to construct a hierarchy of nested meshes and consider certain DOFs of the finest mesh as internal with respect to elements of the coarser (parental) mesh. These internal DOFs can be eliminated out by the static condensation of internal DOFs. The remaining DOFs are associated with the parental mesh. Considering this mesh as the finest one, the same elimination procedure is repeated. We call this process the *hierarchical condensation of DOFs*.

During the hierarchical condensation certain auxiliary matrices are created. These matrices can be used to solve the original system with $O(N \log N)$ arithmetic operations. However, the setup of these auxiliary matrices has complexity $O(N^{3/2})$. On the other hand, they can be stored in asymptotically $O(N \log N)$ memory places. For these reasons, the hierarchical condensation of DOFs is especially useful for solving a sequence of systems with the same matrix and many different right-hand sides. For example, in the case of parabolic problems discretized in time by implicit methods.

For the sake of simplicity we present the approach using linear and symmetric parabolic problem. However, generalizations to other type of problems are possible. Generalizations to nonsymmetric, elliptic, Helmholtz, Maxwell, and similar type of problems are especially straightforward. Further, in order to simplify the description of the method, we consider triangular domains. However, generalization to arbitrary domains is not difficult. It suffices to consider an initial (coarse) mesh of the domain and apply the hierarchical condensation procedure to all triangular elements of the coarse mesh. Finally, let us note that this approach is especially advantageous in two spatial dimensions. In principal, it can be used in three and more spatial dimensions, but the resulting matrices are denser and both the memory requirements and computational complexity grow with the dimension.

The rest of this paper is organized as follows. A linear parabolic model problem is introduced in Section 2. Section 3 describes the hierarchical meshes and a special enumeration of DOFs. Section 4 forms the core of this paper and presents the hierarchical condensation of DOFs. Section 5 provides the algorithm and Section 6 computes its asymptotic complexity and memory requirements. Numerical experiments that compare the performance of various standard approaches and the hierarchical condensation of DOFs is presented in Section 7. Finally, Section 8 draws the conclusions.

2. Model problem

Let $\Omega \subset \mathbb{R}^2$ be a triangle and let $T > 0$ be fixed. We consider the following linear parabolic problem in Ω with homogeneous Dirichlet boundary conditions. Find $u = u(t, x)$ such that

$$\begin{aligned} \partial u/\partial t - \Delta u &= f \quad \text{in } (0,T) \times \Omega, \\ u(t,x) &= 0 \quad \text{for } t \in [0,T) \text{ and } x \in \partial\Omega, \\ u(0,x) &= u_0(x) \quad \text{for } x \in \Omega. \end{aligned} \tag{1}$$

In order to define the weak formulation of problem (1), we introduce the Sobolev space $V = H_0^1(\Omega)$ and assume $f \in L^2(\Omega)$ and $u_0 \in V$. The weak solution $u \in C([0,T], V)$ has the distributional time derivative $\dot{u} = \mathrm{d}u/\mathrm{d}t$ in $C([0,T], L^2(\Omega))$ and it satisfies

$$\int_\Omega \dot{u} v \,\mathrm{d}x + \int_\Omega \nabla u \cdot \nabla v \,\mathrm{d}x = \int_\Omega f v \,\mathrm{d}x \quad \forall v \in V,\ t \in (0,T), \tag{2}$$

and $u(0,x) = u_0(x)$ for a.a. $x \in \Omega$.

We discretize (2) by the method of lines, see e.g. [13, 16, 19] for the works of Karel Segeth on this topic. We use the usual first-order (piecewise linear) triangular finite elements for the space discretization. Hence, we consider a triangulation $\mathcal{T}_h$ of the domain Ω and we define a subspace $V_h \subset V$ of piecewise linear functions on $\mathcal{T}_h$ by

$$V_h = \{v_h \in V : v_h|_K \in P^1(K) \text{ for all } K \in \mathcal{T}_h\},$$

where $P^1(K)$ is the three-dimensional space of linear functions in a triangle $K \in \mathcal{T}_h$. Notice that all functions $v_h \in V_h$ are continuous in Ω.

The semidiscrete solution of (2) $u_h \in C^1([0,T], V_h)$ is given by

$$\int_\Omega \dot{u}_h v_h \,\mathrm{d}x + \int_\Omega \nabla u_h \cdot \nabla v_h \,\mathrm{d}x = \int_\Omega f v_h \,\mathrm{d}x \quad \forall v_h \in V_h,\ t \in (0,T), \tag{3}$$

and $u_h(0,x) = u_{0,h}(x)$ for $x \in \Omega$, where $u_{0,h} \in V_h$ is a suitable projection of the initial condition u_0.

Equality (3) yields a system of linear ordinary differential equations. Indeed, let us define the standard finite element *hat functions* $\varphi_1, \varphi_2, \ldots, \varphi_N$ [21, 23], where $N = \dim V_h$. Each hat function $\varphi_j \in V_h$ equals to one at a vertex x_j of the mesh $\mathcal{T}_h$ and vanishes at all the other vertices. If we expand the semidiscrete solution as $u_h(t,x) = \sum_{j=1}^N y_j(t)\varphi_j(x)$ then the expansion coefficients $y = (y_1, y_2, \ldots, y_N)^T$ are determined by the system of linear differential equations

$$M\dot{y} + Ay = F, \quad y(0) = y_0, \tag{4}$$

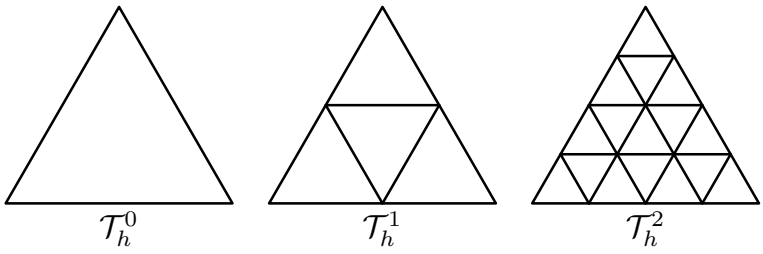


Figure 1: Triangulations of level 0, 1, and 2.

where the vector y_0 of the initial condition is determined by the expansion $u_{0,h} = \sum_{i=1}^{N} y_{0,i}\varphi_i$ of $u_{0,h}$ into the basis of V_h. Further, the mass matrix $M \in \mathbb{R}^{N\times N}$, the stiffness matrix $A \in \mathbb{R}^{N\times N}$, and the load vector $F \in \mathbb{R}^N$ have entries

$$M_{ij} = \int_\Omega \varphi_i\varphi_j \,\mathrm{d}x, \quad A_{ij} = \int_\Omega \nabla\varphi_i \cdot \nabla\varphi_j \,\mathrm{d}x, \quad \text{and} \quad F_i = \int_\Omega f\varphi_i \,\mathrm{d}x.$$

Solving system (4) by a suitable method for systems of ordinary differential equations, we finally arrive at a fully-discrete solution to (1). In this paper we use so-called θ-method [7, 9] with a fixed time step $\tau > 0$. This method yields a system of linear algebraic equations

$$Sy^{k+1} = b^k \tag{5}$$

for the approximation y^{k+1} of y at time $t = (k+1)\tau$, $k = 0, 1, 2, \ldots$. The matrix S and the right-hand side vector b^k are given as

$$S = M + \tau\theta A, \quad b^k = \tau F + (M - \tau(1-\theta)A)y^k, \quad k = 0, 1, 2, \ldots,$$

where $\theta \in [0, 1]$ is arbitrary and fixed. Let us note that the choices $\theta = 0$, $\theta = 1/2$, and $\theta = 1$ correspond to the explicit Euler method, Crank-Nicolson method, and implicit Euler method, respectively.

In the subsequent parts of the paper we will concentrate on the hierarchical condensation of DOFs, which is an efficient method for solving the sequence of linear algebraic problems (5). Let us emphasize that we restrict ourselves to the case of simple model problem (1) for the reason of clarity only. The hierarchical condensation of DOFs can be applied to a wide class of much more general problems.

3. Mesh construction and enumeration of DOFs

The hierarchical condensation of DOFs is based on a hierarchy of successively refined and nested triangular meshes. In this section we define the triangulation, introduce its hierarchical structure represented by levels, and present a special enumeration of vertices of the triangulation (and the corresponding DOFs) that enables relatively simple implementation of the method.

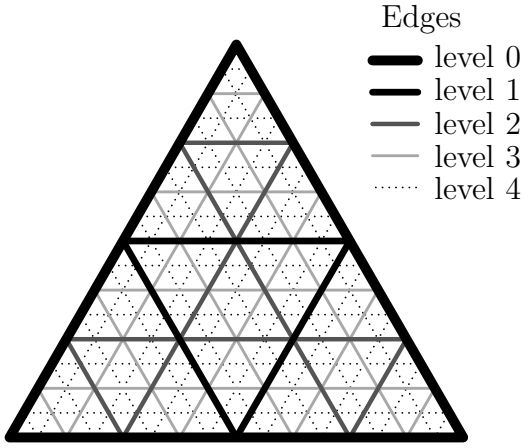


Figure 2: Levels of edges in the triangulation $\mathcal{T}_h^4$ of level 4.

The triangulation $\mathcal{T}_h$ of the triangle Ω is constructed hierarchically using *levels*. Triangulation $\mathcal{T}_h^0$ of level 0 consists of the single triangle Ω. Triangulation $\mathcal{T}_h^\ell$ of level ℓ is obtained by splitting all triangles in $\mathcal{T}_h^{\ell-1}$ into four similar subtriangles, see Figure 1. From now on we denote by $L > 0$ the fixed number of levels and we set $n = 2^L$ the number of subedges on an edge of Ω. The triangulation $\mathcal{T}_h = \mathcal{T}_h^L$ has n^2 elements, it contains $(n+2)(n+1)/2$ vertices from which $3n$ lay on the boundary $\partial\Omega$ and $(n-1)(n-2)/2$ lay in the interior of Ω. Thus, the number of DOFs is $N = \dim V_h = (n-1)(n-2)/2$, because we consider Dirichlet boundary conditions.

In order to describe the special enumeration of vertices we introduce a *level of an edge*. An edge in $\mathcal{T}_h = \mathcal{T}_h^L$ is of level $\ell = 0, 1, 2, \ldots, L$ if it lays on an edge of $\mathcal{T}_h^\ell$ but not on any edge of $\mathcal{T}_h^{\ell-1}$, $\mathcal{T}_h^{\ell-2}$, $\ldots$, $\mathcal{T}_h^0$. For example, edges of level 0 are those edges of $\mathcal{T}_h$ which lay on the boundary $\partial\Omega$. Levels of edges are indicated in Figure 2.

Level of a vertex is the smallest level of edges meeting at this vertex. Notice that any interior vertex of level $\ell = 1, 2, \ldots, L-1$ lays on two edges of level ℓ and on four edges of a higher level. Simply, all vertices of level ℓ lay on all edges of level ℓ. Since vertices correspond to the finite element basis functions and consequently to DOFs, we will naturally speak about levels of basis functions and DOFs.

The enumeration of vertices goes by levels. Since there are no vertices of level L, we first enumerate vertices of level $L-1$, then vertices of level $L-2$, etc. Finally, we enumerate vertices of level 1. Vertices of level 0 lay on the boundary of $\partial\Omega$, where we consider Dirichlet boundary conditions and hence there are no DOFs. Moreover, the enumeration of vertices of level $\ell = L-1, L-2, \ldots, 1$ goes in natural order. Precisely, there are always three interior edges of level ℓ in every element of $\mathcal{T}_h^{\ell-1}$. The enumeration of vertices of level ℓ goes by elements of $\mathcal{T}_h^{\ell-1}$. We first enumerate vertices of level ℓ on edges of level ℓ in the interior of the first element of $\mathcal{T}_h^{\ell-1}$ and then we proceed to enumerate in the same way vertices inside the second element of $\mathcal{T}_h^{\ell-1}$, etc. Figure 3 presents an example of enumeration of vertices for $L = 3$. The algorithm is as follows:

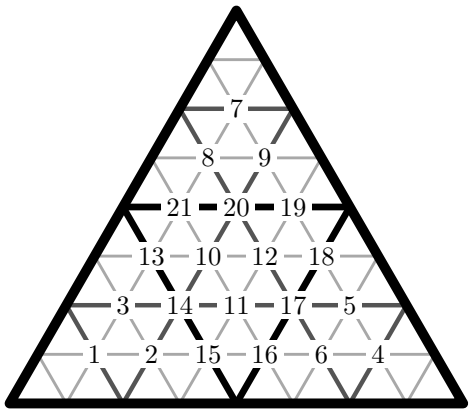


Figure 3: Triangulations $\mathcal{T}_h^3$ and enumeration of vertices. The vertices of level 2 are enumerated first (indices 1,2,...,12), then vertices of level 1 are enumerated (13,14,...,21). Notice that triplets of basis functions with indices 1,2,3; 4,5,6; 7,8,9; and 10,11,12 form bubbles in elements of triangulation $\mathcal{T}_h^1$.

for $\ell = L-1, L-2, \ldots, 1$ do
  for all elements K in $\mathcal{T}_h^{\ell-1}$ do
    enumerate all vertices lying on the three interior edges of level ℓ in K
    (there are $2^{L-\ell}-1$ vertices on each such edge and all are of level ℓ)
  end (loop through elements)
end (loop through levels)

4. Hierarchical condensation of DOFs

To describe the hierarchical condensation of DOFs, we introduce the notion of *bubble functions* or shortly *bubbles*. A function is called a bubble in element K if it is supported solely in K. The basis functions of level $L-1$ form bubbles in elements of $\mathcal{T}_h^{L-2}$. We use the static condensation to eliminate these bubbles, i.e., we eliminate all DOFs of level $L-1$. In the remaining (Schur complement) system, the DOFs of level $L-2$ correspond to bubbles in elements of $\mathcal{T}_h^{L-3}$ and they can be eliminated in the same way. This procedure continues until we traverse the whole hierarchy of meshes.

Now, we describe details of this procedure. The goal is to solve linear system

$$S_{(0)} y_{(0)} = b_{(0)}, \tag{6}$$

where $S_{(0)} = S$, $y_{(0)} = y^k$, and $b_{(0)} = b^k$ come from (5) for a fixed k. Recall that the number of DOFs (i.e. the size of this system) is $N_{(0)} = N = (n-2)(n-1)/2$. The algorithm goes in $L-2$ steps for $m = 1, 2, \ldots, L-2$.

Step 1: ($m = 1$) In this step we eliminate DOFs of level $L-1$, which form bubbles in elements of triangulation $\mathcal{T}_h^{L-2}$. The triangulation $\mathcal{T}_h^{L-2}$ consists of $(n/4)^2 = 2^{2(L-2)}$

elements and there are three vertices of level $L-1$ inside of all these elements. Thus, there are $M_{(0)} = 3 \cdot 2^{2(L-2)}$ vertices of level $L-1$. The DOFs corresponding to these vertices were enumerated first and therefore their indices are $1, 2, \ldots, M_{(0)}$. This yields the following block structure of $S_{(0)} \in \mathbb{R}^{N_{(0)} \times N_{(0)}}$, $y_{(0)} \in \mathbb{R}^{N_{(0)}}$, and $b_{(0)} \in \mathbb{R}^{N_{(0)}}$:

$$S_{(0)} = \begin{pmatrix} A_{(1)} & B_{(1)}^T \\ B_{(1)} & D_{(1)} \end{pmatrix}, \quad y_{(0)} = \begin{pmatrix} x_{(1)} \\ y_{(1)} \end{pmatrix}, \quad \text{and} \quad b_{(0)} = \begin{pmatrix} F_{(1)} \\ G_{(1)} \end{pmatrix}, \tag{7}$$

where $A_{(1)} \in \mathbb{R}^{M_{(0)} \times M_{(0)}}$, $B_{(1)} \in \mathbb{R}^{(N_{(0)} - M_{(0)}) \times M_{(0)}}$, $D_{(1)} \in \mathbb{R}^{(N_{(0)} - M_{(0)}) \times (N_{(0)} - M_{(0)})}$, $F_{(1)} \in \mathbb{R}^{M_{(0)}}$, and $G_{(1)} \in \mathbb{R}^{N_{(0)} - M_{(0)}}$. The matrix $A_{(1)}$ corresponds to bubble functions and therefore it is blockdiagonal, consisting of $(n/4)^2 = 2^{2(L-2)}$ blocks of size 3×3.

The bubble DOFs, i.e. the unknowns $x_{(1)}$, can be efficiently eliminated. The remaining DOFs, i.e. the unknowns $y_{(1)}$, are then given by a Schur complement system. To be more specific, we compute the block-wise inverse $A_{(1)}^{-1}$ and use it to obtain the Schur complement $S_{(1)}$ and the complement load $b_{(1)}$ as

$$S_{(1)} = D_{(1)} - B_{(1)} A_{(1)}^{-1} B_{(1)}^T \quad \text{and} \quad b_{(1)} = G_{(1)} - B_{(1)} A_{(1)}^{-1} F_{(1)}.$$

The two components of the coefficient vector $y_{(0)} = (x_{(1)}, y_{(1)})^T$ are then given by

$$S_{(1)} y_{(1)} = b_{(1)} \quad \text{and} \quad x_{(1)} = A_{(1)}^{-1}(F_{(1)} - B_{(1)}^T y_{(1)}).$$

Thus, as soon as the vector $y_{(1)}$ is known, the vector $x_{(1)}$ can be easily and efficiently computed. In order to compute $y_{(1)}$, we have to solve a system with matrix $S_{(1)}$. The Schur complement $S_{(1)}$ has a similar structure as the original matrix $S_{(0)}$ in the sense that vertices of level $L-2$ correspond to bubbles in the triangulation $\mathcal{T}_h^{L-3}$. Consequently, DOFs of level $L-2$ can be eliminated from $S_{(1)}$ in the same way as DOFs of level $L-1$ were eliminated from $S_{(0)}$. As a result, we obtain a Schur complement $S_{(2)}$ and the whole procedure can be repeated. In general, the m-th step of the algorithm is as follows.

Step m: (Elimination of DOFs of level $L-m$.) Put $N_{(m-1)} = N_{(m-2)} - M_{(m-2)}$ and set $M_{(m-1)} = 3(2^m - 1)(n/2^{m+1})^2 = 3(2^m - 1)2^{2(L-m-1)}$. The matrix $S_{(m-1)}$ is of size $N_{(m-1)} \times N_{(m-1)}$ and the coefficient vector $y_{(m-1)}$ and the load vector $b_{(m-1)}$ are of length $N_{(m-1)}$. There is $M_{(m-1)}$ bubble functions corresponding to vertices of level $L-m$. Thanks to the special enumeration of vertices from Section 3 the DOFs corresponding to these bubble functions have indices $1, 2, \ldots, M_{(m-1)}$ with respect to $S_{(m-1)}$. This naturally introduces the block structure

$$S_{(m-1)} = \begin{pmatrix} A_{(m)} & B_{(m)}^T \\ B_{(m)} & D_{(m)} \end{pmatrix}, \quad y_{(m-1)} = \begin{pmatrix} x_{(m)} \\ y_{(m)} \end{pmatrix}, \quad b_{(m-1)} = \begin{pmatrix} F_{(m)} \\ G_{(m)} \end{pmatrix}, \tag{8}$$

where $A_{(m)} \in \mathbb{R}^{M_{(m-1)} \times M_{(m-1)}}$, $B_{(m)} \in \mathbb{R}^{(N_{(m-1)} - M_{(m-1)}) \times M_{(m-1)}}$, etc. The matrix $A_{(m)}$ corresponds to bubble DOFs and it is block diagonal with $(n/2^{m+1})^2 = 2^{2(L-m-1)}$

blocks of size $3(2^m-1)\times 3(2^m-1)$. We invert $A_{(m)}$ and compute the Schur complement as well as the complement load as

$$S_{(m)} = D_{(m)} - B_{(m)}A_{(m)}^{-1}B_{(m)}^T \quad \text{and} \quad b_{(m)} = G_{(m)} - B_{(m)}A_{(m)}^{-1}F_{(m)}. \tag{9}$$

The components of the coefficient vector $y_{(m-1)} = (x_{(m)}, y_{(m)})^T$ are determined by

$$S_{(m)}y_{(m)} = b_{(m)} \quad \text{and} \quad x_{(m)} = A_{(m)}^{-1}(F_{(m)} - B_{(m)}^T y_{(m)}).$$

Step $L-1$: After $L-2$ steps (for $m = 1, 2, \ldots, L-2$), we are left with system

$$S_{(L-2)}y_{(L-2)} = b_{(L-2)} \tag{10}$$

with fully populated matrix $S_{(L-2)} \in \mathbb{R}^{3(n/2-1)\times 3(n/2-1)}$. We can solve this system by a standard approach such as the Cholesky decomposition for instance. As a result, we obtain the coefficients $y_{(L-2)}$ and we can compute the remaining ones by backward substitution.

Backward substitution: The remaining vectors of unknowns $x_{(m)}$, $m = L-2$, $L-3, \ldots, 1$, are easily computed as

$$x_{(m)} = A_{(m)}^{-1}(F_{(m)} - B_{(m)}^T y_{(m)}) \quad \text{and} \quad y_{(m-1)} = \begin{pmatrix} x_{(m)} \\ y_{(m)} \end{pmatrix}. \tag{11}$$

Once the matrix $S_{(0)}$ is hierarchically decomposed by the above algorithm, the next linear system with matrix $S_{(0)}$ and a different right-hand side $b_{(0)}$ can be solved very efficiently. It suffices to store matrices $Q_{(m)} = B_{(m)}A_{(m)}^{-1}$, $A_{(m)}^{-1}$, for $m = 1, 2, \ldots, L-2$, and $S_{(L-2)}^{-1}$. The given right-hand side $b_{(0)}$ is then hierarchically split into vectors $F_{(m)}$, $m = 1, 2, \ldots, L-2$, and vector $b_{(L-2)}$ using matrices $Q_{(m)}$, see (9). The final Schur complement system (10) is then solved using the stored matrix $S_{(L-2)}^{-1}$. Finally, the backward substitution (11) is performed utilizing matrices $A_{(m)}^{-1}$ and $Q_{(m)}^T$ for $m = L-2, L-1, \ldots, 1$.

Let us note that storing matrices $Q_{(m)}$ instead of $B_{(m)}$ increases the efficiency of the entire procedure significantly. On the other hand the matrix $Q_{(m)}$ has more nonzero entries than $B_{(m)}$ and its storage requires more memory. However, the difference is not large and asymptotically both these matrices have $O(N)$ nonzero entries. For details see Section 6 below.

5. Algorithm

In this section we rigorously describe the algorithm of hierarchical static condensation of DOFs with the emphasis on many linear algebraic systems with the same matrix and different right-hand side vectors. The rigorous formulation of the algorithm will be utilized in Section 6 to compute its complexity and memory requirements.

The algorithm consists of setup and solve phases. We consider the enumeration of DOFs from Section 3 and use $M_{(m-1)} = 3(2^m - 1)2^{L-m-1}$ to denote the number of bubble DOFs of level $L - m$, $m = 1, 2, \ldots, L - 2$.

First, we describe the setup phase. Its input is the matrix $S_{(0)} \in \mathbb{R}^{N_{(0)} \times N_{(0)}}$ that comes from the finite element discretization, see (6). The output consists of matrices $Q_{(m)}$, $A^{-1}_{(m)}$ for $m = 1, 2, \ldots, L-2$, and $S^{-1}_{(L-2)}$ that are needed in the solve phase.

Setup phase:

1. For $m = 1, 2, \ldots, L - 2$ do the following:
   (a) Split the matrix $S_{(m-1)}$ into blocks $A_{(m)}$, $B_{(m)}$, and $D_{(m)}$ as in (8).
   (b) Matrix $A_{(m)}$ is block-diagonal with 2^{L-m-1} blocks of size $3(2^m - 1) \times 3(2^m - 1)$. Use block-wise inversion to compute $A^{-1}_{(m)}$.
   (c) Perform the sparse matrix multiplication $Q_{(m)} = B_{(m)}A^{-1}_{(m)}$.
   (d) Compute the Schur complement matrix $S_{(m)} = D_{(m)} - Q_{(m)}B^T_{(m)}$, see (9).
   (e) Update $N_{(m)} = N_{(m-1)} - M_{(m-1)}$.
2. Compute the inverse $S^{-1}_{(L-2)}$ of the fully populated matrix $S_{(L-2)}$.
3. Output matrices $Q_{(m)}$, $A^{-1}_{(m)}$ for $m = 1, 2, \ldots, L - 2$, and $S^{-1}_{(L-2)}$.

Second, we present the solve phase. Its input data consist of a vector $b_{(0)} \in \mathbb{R}^{N_{(0)}}$, matrices $Q_{(m)}$, $A^{-1}_{(m)}$ for $m = 1, 2, \ldots, L-2$, and matrix $S^{-1}_{(L-2)}$. The output is a vector $y_{(0)}$ that solves system (6).

Solve phase:

1. For $m = 1, 2, \ldots, L - 2$ do the following:
   (a) Split vector $b_{(m-1)}$ into two blocks $F_{(m)}$ and $G_{(m)}$ as in (8).
   (b) Compute $b_{(m)} = G_{(m)} - Q_{(m)}F_{(m)}$, see (9).
   (c) Update $N_{(m)} = N_{(m-1)} - M_{(m-1)}$.
2. Solve the Schur complement problem: $y_{(L-2)} = S^{-1}_{(L-2)}b_{(L-2)}$.
3. Perform the backward substitution. For $m = L-2, L-3, \ldots, 1$ do the following:
   (a) Compute $x_{(m)} = A^{-1}_{(m)}F_{(m)} - Q^T_{(m)}y_{(m)}$.
   (b) Update $y_{(m-1)} = (x_{(m)}, y_{(m)})^T$.
4. Output vector $y_{(0)}$.

6. Computational complexity and memory requirements

In this section we compute the complexity and the memory requirements of the setup and the solve phase of the algorithm from Section 5. By the complexity we understand the asymptotic number of arithmetic operations need to perform the algorithm. The memory requirements are represented by the asymptotic number of memory places needed to store the data structures. We recall that L stands for the fixed number of levels, $n = 2^L$ denotes the number of mesh-edges on one edge of Ω, and $N = (n-2)(n-1)/2$ is the number of DOFs (the size of matrix S).

Theorem 1. *The complexity of the setup phase is $O(N^{3/2})$.*

Proof. For each $m = 1, 2, \ldots, L-2$ we invert the block diagonal matrix $A_{(m)}$. The number of arithmetic operations needed to invert a block diagonal matrix is proportional to the number of blocks multiplied by the size of each block cubed: $N_{\mathrm{op}}\left(A_{(m)}^{-1}\right) \approx 2^{2(L-m-1)} \cdot [3(2^m-1)]^3$. Further, we have to invert a dense matrix $S_{(L-2)}$. This requires $N_{\mathrm{op}}\left(S_{(L-2)}^{-1}\right) \approx [3(2^{L-1}-1)]^3$ operations. The number of arithmetic operations needed for the other steps of the setup phase is asymptotically minor with respect to $N_{\mathrm{op}}\left(A_{(m)}^{-1}\right)$ and $N_{\mathrm{op}}\left(S_{(L-2)}^{-1}\right)$. Thus, the complexity of the setup phase is

$$N_{\mathrm{op}}\left(S_{(L-2)}^{-1}\right) + \sum_{m=1}^{L-2} N_{\mathrm{op}}\left(A_{(m)}^{-1}\right) \approx (2^L)^3 = n^3 \approx N^{3/2}.$$

□

Theorem 2. *The complexity of the solve phase is $O(N \log N)$.*

Proof. The most significant operation in step 1 of the solve phase is the matrix-vector multiplication $Q_{(m)}F_{(m)}$. This multiplication requires a number of operations proportional to the number of nonzero entries in $Q_{(m)}$. It is at most twice the number of vertices of levels less then $L-m$ times the number of vertices of level $L-m$ inside one element of mesh $\mathcal{T}_h^{L-m-1}$. Thus, this number can in general reach the value up to

$$N_{\mathrm{NZ}}\left(Q_{(m)}\right) = 2 \times (N_{(m-1)} - M_{(m-1)}) \times 3(2^m - 1), \tag{12}$$

where $N_{(m-1)} = 3(2^m-1)2^{L-m-1}(2^{L-m}-1)$ is the number of vertices of levels less than or equal to $L-m$ and $M_{(m-1)} = 3(2^m-1)2^{2(L-m-1)}$ is the number of vertices of level $L-m$. Consequently, $N_{(m-1)} - M_{(m-1)} = 3(2^m-1)(2^{2(L-m-1)} - 2^{L-m-1})$. Since the matrix-vector multiplication $Q_{(m)}F_{(m)}$ is performed for $m = 1, 2, \ldots, L-2$, the complexity of step 1 is proportional to

$$\sum_{m=1}^{L-2} N_{\mathrm{NZ}}\left(Q_{(m)}\right) = \sum_{m=1}^{L-2} 18(2^m-1)^2\left(2^{2(L-m-1)} - 2^{L-m-1}\right) = (9L-42)2^{2L-1}$$

$$+ (18L+9)2^L + 12 = \frac{n^2}{2}(9\log_2 n - 42) + n(18\log_2 n + 9) + 12 \approx N \log N. \tag{13}$$

In step 2 we multiply the vector $b_{(L-2)}$ by the fully populated matrix $S^{-1}_{(L-2)}$ of size $3(2^{L-1}-1)\times 3(2^{L-1}-1)$. The complexity of this operation is

$$3^2(2^{L-1}-1)^2 \approx n^2 \approx N.$$

In step 3 we perform matrix-vector multiplications with matrices $A^{-1}_{(m)}$ and $Q^T_{(m)}$ for $m = L-2, L-3, \ldots, 1$. The complexity of multiplication by matrix $A^{-1}_{(m)}$ is proportional to the number of its nonzero entries, which is less than the number of nonzero entries of $Q^T_{(m)}$. Multiplications by the matrices $Q^T_{(m)}$ and $Q_{(m)}$ are of the same complexity proportional to $N_{\mathrm{NZ}}\left(Q_{(m)}\right)$, see (12). Thus, the complexity of step 3 is proportional to $N\log N$ as in step 1. Consequently, the total complexity of the solve phase is $O(N\log N)$. □

Theorem 3. *The memory requirements to store matrices $Q_{(m)}$, $A^{-1}_{(m)}$ for $m = 1, 2, \ldots, L-2$, and matrix $S^{-1}_{(L-2)}$ are $O(N\log N)$.*

Proof. The fully populated matrix $S^{-1}_{(L-2)}$ contains $N_{\mathrm{NZ}}\big(S^{-1}_{(L-2)}\big) = (3(n/2-1))^2 = 9/4n^2 - 9n + 9$ entries. The number of nonzero entries in matrix $A^{-1}_{(m)}$ is equal to the number of its blocks times the size of the block squared, i.e. $N_{\mathrm{NZ}}\big(A^{-1}_{(m)}\big) = 2^{2(L-m-1)}[3(2^m-1)]^2$. Thus, for all $m = 1, 2, \ldots, L-2$ we have

$$N_{\mathrm{NZ}}\left(A^{-1}\right) = \sum_{m=1}^{L-2} N_{\mathrm{NZ}}\big(A^{-1}_{(m)}\big) = 2^{2L-2}(9L-33) + 18\cdot 2^L - 12$$
$$= \frac{n^2}{4}(9\log_2 n - 33) + 18n - 12.$$

The total number of nonzero entries in all matrices $Q_{(m)}$ for $m = 1, 2, \ldots, L-2$ was computed above, see (13). Hence, we can conclude that the total amount of memory places needed to store matrices $Q_{(m)}$, $A^{-1}_{(m)}$ for $m = 1, 2, \ldots, L-2$, and matrix $S^{-1}_{(L-2)}$ is asymptotically proportional to $N\log N$. □

Let us note that the original stiffness matrix $S_{(0)}$ contains $N_{\mathrm{NZ}}\left(S_{(0)}\right) = 7(n-2)(n-1)/2 - 6(n-2) = (7n^2-33n+38)/2$ nonzero entries. Making rough estimates and considering a sufficiently high number of levels L, we may say that the total memory requirements to store matrices $Q_{(m)}$, $A^{-1}_{(m)}$ for $m = 1, 2, \ldots, L-2$, and $S^{-1}_{(L-2)}$ are about $2(L-4)$ times higher than $N_{\mathrm{NZ}}\left(S_{(0)}\right)$.

7. Numerical experiments

In this section we compare the performance of the above described hierarchical condensation of DOFs with standard methods. The numerical tests are done in Matlab and the hierarchical condensation is compared with standard Matlab implementations of fully populated matrix inversion, sparse Cholesky factorization,

conjugate gradients (CG) preconditioned by the incomplete Cholesky factorization and an optimized direct sparse solver (backslash command in Matlab).

We consider the parabolic problem (1) in a triangle Ω with vertices $[0,0]$, $[1,0]$, and $[0.7,0.8]$ for $t \in (0,100)$. The right-hand side and the initial conditions are chosen as $f = 1$ and $u_0 = \lambda_1^4 \lambda_2 \lambda_3$, where $\lambda_1, \lambda_2, \lambda_3$ are barycentric coordinates in Ω. This problem is discretized as described in Section 2. We use the time step $\tau = 0.1$ and the Crank-Nicolson method ($\theta = 1/2$) for time discretization. The space discretization is done on a sequence of uniform and successively refined meshes. We construct the meshes as described in Section 3 for the number of levels $L = 4, 5, \ldots, 10$.

During the time evolution, it is necessary to solve system (5) in total 1000 times (the final time is $T = 100$ and the time step is $\tau = 0.1$). Before the first system (5) is solved, we perform a setup phase for the given matrix S, store the necessary data, and then we run the solve phase 1000 times.

The first of the tested methods is to compute the fully populated matrix inverse in the setup phase, store the inverse, and then just multiply the right-hand side by this inverse in the solve phases. This method is very inappropriate for the presented problem, because it ignores the sparsity of matrix S. We include it in the test in order to illustrate the magnitude of this inappropriateness.

The second method is a sparse Cholesky factorization with the approximate minimum degree permutation trying to minimize the fill-in. A suitable permutation and the Cholesky factor of permuted S are computed in the setup phase. The stored permutation and the Cholesky factor are then used in the solve phases.

The third method is the CG method preconditioned by incomplete Cholesky factorization with no fill-in. In the setup phase we construct the preconditioner, store it, and use it in the solve phases. The initial approximation is taken from the previous time step and the CG iterations are stopped as soon as the relative residual decreases below 10^{-6}. This was always happening in a few iterations.

The fourth method is the backlash command of Matlab. It is a highly optimized procedure combining various sparse direct solvers for various types and sizes of matrices. We perform no setup phase and use the backslash command directly in the solve phases. Finally, the fifth method is the hierarchical condensation of DOFs as described above.

Figure 4 presents the CPU-times required to perform the setup phase (left panel) and the CPU-times for 1000 solve phases (right panel). We see that the hierarchical condensation has the fastest solve phase of all tested methods. It is more than two times faster than the sparse Cholesky factorization. The setup phase of the hierarchical condensation is the second fastest after the sparse Cholesky factorization. However, the asymptotic complexity of the setup phases of the hierarchical condensation and sparse Cholesky factorization seems to be the same in this example. The other methods are not competitive with the exception of the preconditioned CG. Its performance during the solve phase is relatively improving with growing number of DOFs, however the complexity of its setup phase is considerably higher than the complexity of the setup phase of both sparse Cholesky factorization and hierarchi-

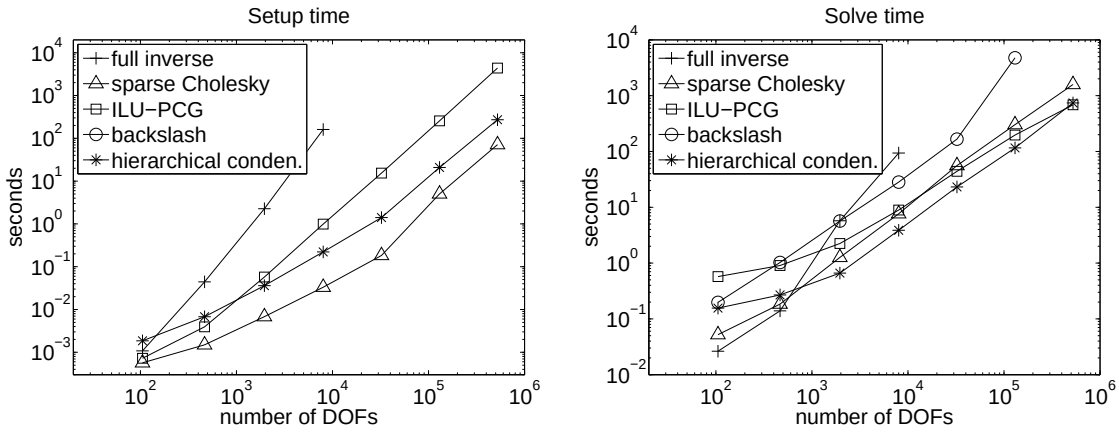


Figure 4: CPU-times for the setup phase (left) and for the 1000 of solve phases (right). Notice the logarithmic scales and zero setup time for the backslash solver.

cal condensation. Let us note that the full-inverse approach runs over the available memory starting from the number of levels $L = 8$.

Further, let us note that no special effort was made to optimize the Matlab code of hierarchical condensation for speed. The code is relatively simple, it contains a few short for-cycles and the majority of CPU-time is spent by various sparse matrix operations. Therefore, we assume that the differences due to the compiled codes (inversion, Cholesky factorization, and backlash) and interpreted codes (preconditioned CG and hierarchical condensation) are not fundamental.

8. Conclusions

In this paper we presented a hierarchical condensation of DOFs, which is a direct sparse method for solving linear algebraic systems. We prove that the setup phase requires $O(N^{3/2})$ arithmetic operations, the resulting data are stored in $O(N \log N)$ memory places, and the solve phase takes $O(N \log N)$ operations, where N stands for the number of DOFs.

The method was presented using a simple model problem, a triangular domain, and the lowest-order finite element method. However, generalizations to more general problems and domains are straightforward as well as generalization to higher-order finite elements. Generalizations to higher spatial dimensions are possible as well.

A clear bottleneck of this approach is the setup phase which has a suboptimal complexity, because the expected optimal complexity would be $O(N)$, see e.g. multigrid approaches [1, 24] or optimal methods in special domains [8, 22].

Nevertheless, the hierarchical condensation of DOFs provides an insight into the structure of the inverse of the finite element matrices. We believe that this insight can be fruitful if it enables to modify the setup phase such that it is performed approximately and fast. This approximate inverse can then serve as an efficient and hopefully optimal preconditioner.

References

[1] Briggs, W. L., Henson, V. E., and McCormick, S. F.: *A multigrid tutorial*, 2nd ed. SIAM, Philadelphia, PA, 2000.

[2] Červ, V. and Segeth, K.: A comparison of the accuracy of the finite-difference solution to boundary value problems for the Helmholtz equation obtained by direct and iterative methods. Apl. Mat. **27** (1982), 375–390.

[3] Ciarlet, P. G.: *The finite element method for elliptic problems.* North-Holland Publishing Co., Amsterdam, 1978.

[4] Duff, I., Erisman, A., and Reid, J.: *Direct methods for sparse matrices.* Clarendon Press, Oxford, 1986.

[5] Greenbaum, A.: *Iterative methods for solving linear systems.* SIAM, Philadelphia, PA, 1997.

[6] Hackbusch, W.: *Iterative solution of large sparse systems of equations. Transl. from the German.* Springer-Verlag, New York, NY, 1994.

[7] Hairer, E. and Wanner, G.: *Solving ordinary differential equations. II: Stiff and differential-algebraic problems. 2nd rev. ed.* Springer, Berlin, 1996.

[8] Hockney, R.: A fast direct solution of Poisson's equation using Fourier analysis. J. Assoc. Comput. Mach. **12** (1965), 95–113.

[9] Lambert, J.: *Numerical methods for ordinary differential systems: the initial value problem.* John Wiley & Sons, Chichester, 1991.

[10] Meurant, G.: *Computer solution of large linear systems.* Elsevier, Amsterdam, 1999.

[11] Osterby, O. and Zlatev, Z.: *Direct methods for sparse matrices.* Springer-Verlag, Berlin, 1983.

[12] Saad, Y.: *Iterative methods for sparse linear systems*, 2nd ed. SIAM, Philadelphia, PA, 2003.

[13] Segeth, K.: A posteriori error estimation with the finite element method of lines for a nonlinear parabolic equation in one space dimension. Numer. Math. **83** (1999), 455–475.

[14] Segeth, K.: On the choice of iteration parameters in the Stone incomplete factorization. Apl. Mat. **28** (1983), 295–306.

[15] Segeth, K.: Numerical experiments with the Stone incomplete triangular decomposition. In: *Mathematical models in physics and chemistry and numerical methods of their realization (Visegrád, 1982)*, *Teubner-Texte Math.*, vol. 61, pp. 226–236. Teubner, Leipzig, 1984.

[16] Segeth, K.: A posteriori error estimates for parabolic differential systems solved by the finite element method of lines. Appl. Math. **39** (1994), 415–443.

[17] Šolín, P. and Segeth, K.: A new sequence of hierarchic prismatic elements satisfying de Rham diagram on hybrid meshes. J. Numer. Math. **13** (2005), 295–317.

[18] Šolín, P. and Segeth, K.: Performance of various ODE solvers on FV-semidiscretized nonstationary compressible Euler equations. Acta Tech. CSAV **47** (2002), 47–66.

[19] Šolín, P. and Segeth, K.: Application of the method of lines to unsteady compressible Euler equations. Internat. J. Numer. Methods Fluids **41** (2003), 519–535.

[20] Šolín, P. and Segeth, K.: Hierarchic higher-order Hermite elements on hybrid triangular/quadrilateral meshes. Math. Comput. Simulation **76** (2007), 198–204.

[21] Šolín, P., Segeth, K., and Doležel, I.: *Higher-order finite element methods*. Studies in Advanced Mathematics, Chapman & Hall/CRC, Boca Raton, FL, 2004.

[22] Swarztrauber, P. N.: The methods of cyclic reduction, Fourier analysis and the FACR algorithm for the discrete solution of Poisson's equation on a rectangle. SIAM Rev. **19** (1977), 490–501.

[23] Szabó, B. and Babuška, I.: *Finite element analysis*. A Wiley-Interscience Publication, John Wiley & Sons Inc., New York, 1991.

[24] Trottenberg, U., Oosterlee, C. W., and Schüller, A.: *Multigrid. with guest contributions by A. Brandt, P. Oswald, K. Stüben.* Academic Press, Orlando, FL, 2001.

[25] Vejchodský, T. and Šolín, P.: Static condensation, partial orthogonalization of basis functions, and ILU preconditioning in the *hp*-FEM. J. Comput. Appl. Math. **218** (2008), 192–200.

Conference Applications of Mathematics 2013
in honor of the 70th birthday of Karel Segeth.
Jan Brandts, Sergey Korotov, Michal Křížek,
Jakub Šístek, and Tomáš Vejchodský (Eds.),
Institute of Mathematics AS CR, Prague 2013

SPHERICALLY SYMMETRIC SOLUTIONS TO A MODEL FOR INTERFACE MOTION BY INTERFACE DIFFUSION

Peicheng Zhu[1,2]

[1] Department of Mathematics, University of the Basque Country
E-48940 Leioa, Spain
[2] IKERBASQUE, Basque Foundation for Science
E-48011 Bilbao, Spain
peicheng.zhu@ehu.es

Abstract

The existence of spherically symmetric solutions is proved for a new phase-field model that describes the motion of an interface in an elastically deformable solid, here the motion is driven by configurational forces. The model is an elliptic-parabolic coupled system which consists of a linear elasticity system and a non-linear evolution equation of the order parameter. The non-linear equation is non-uniformly parabolic and is of fourth order. One typical application is sintering.

1. Introduction

A central tenet in materials science is that many properties of materials are determined by microstructure. Microstructure can be defined as the totality of all thermodynamic non-equilibrium lattice defects on a space scale ranging from Ångstrøms to meters. By their dimension, defects can be arranged in the following hierarchy: i) zero-, ii) one-, iii) two-, iv) three-dimensional defects. Their typical examples are, respectively, point defects, dislocations, grain boundaries and voids. The driving forces for the evolution of defects are of the Eshelby type that is radically different to the Newton type.

We shall study, in this paper, the evolution of two-dimensional defects, taking grain boundary as an example, by employing a phase-field approach that is still young but has been shown powerful and important for both theoretical and numerical investigations, especially for multi-dimensional problems, see, e.g. [8, 10, 13]. Starting from a sharp interface model based on a formula of configurational forces in terms of the Eshelby tensor, Alber and Zhu [1, 2] have formulated a new phase-field model which differs from the famous Cahn-Hilliard model (see [7]) by a non-smooth gradient term. An application of our model is to describe sintering, a technique for making a material from powders.

To state the new model we now introduce some notations. Let Ω be an open subset in $\mathbb{R}^3$. It stands for the set of material points of a solid body. The different

phases of a solid are indicated by an order parameter $S(t,x) \in \mathbb{R}$: That S takes values near to zero or one means the solid is in phase γ or γ'. Other unknowns are the displacement $u(t,x) \in \mathbb{R}^3$ of the material point x at time t and the Cauchy stress tensor $T(t,x) \in \mathcal{S}^3$. Here $\mathcal{S}^3$ denotes the set of symmetric 3×3-matrices. We shall investigate the quasi-static process, the unknowns thus must satisfy the following equations

$$\begin{aligned} -\mathrm{div}_x\, T(t,x) &= b(t,x), && (1)\\ T(t,x) &= D\big(\varepsilon(\nabla_x u) - \bar{\varepsilon} S\big)(t,x), && (2)\\ S_t(t,x) &= c\,\mathrm{div}_x \Big(\nabla_x\big(\psi_S(\varepsilon(\nabla_x u), S) - \nu\Delta_x S\big)|\nabla_x S|\Big)(t,x), && (3) \end{aligned}$$

for $(t,x) \in (0,\infty) \times \Omega$, and the boundary and initial conditions

$$\begin{aligned} u(t,x) &= \gamma(t,x), \qquad (t,x) \in [0,\infty) \times \partial\Omega, && (4)\\ \frac{\partial}{\partial n} S(t,x) &= 0, \qquad (t,x) \in [0,\infty) \times \partial\Omega, && (5)\\ \frac{\partial}{\partial n}\big(\psi_S(\varepsilon, S) - \nu\Delta_x S\big)\,|\nabla_x S|(t,x) &= 0, \qquad (t,x) \in [0,\infty) \times \partial\Omega, && (6)\\ S(0,x) &= S_0(x), \qquad x \in \bar{\Omega}. && (7) \end{aligned}$$

Here n is the unit outward normal vector, $\nabla_x u$ denotes the 3×3-matrix of first order derivatives of u, the deformation gradient, and

$$\varepsilon\,(\nabla_x u) = \frac{1}{2}\big(\nabla_x u + (\nabla_x u)^T\big)$$

is the strain tensor, where $(\nabla_x u)^T$ denotes the transposed matrix. Further, $\bar{\varepsilon} \in \mathcal{S}^3$ is a given matrix, the transformation strain. The elasticity tensor $D : \mathcal{S}^3 \to \mathcal{S}^3$ is a linear, symmetric, positive definite mapping, and ψ_S is the derivative with respect to S of the free energy

$$\psi^*(\varepsilon, S, \nabla_x S) = \psi(\varepsilon, S) + \frac{\nu}{2}|\nabla_x S|^2 = \frac{1}{2}\big(D(\varepsilon - \bar{\varepsilon} S)\big) \cdot (\varepsilon - \bar{\varepsilon} S) + \hat{\psi}(S) + \frac{\nu}{2}|\nabla_x S|^2, \quad (8)$$

where for $\hat{\psi} : \mathbb{R} \to [0,\infty)$ we choose a double well potential with minima at points $S \le 0$ and $S \ge 1$. The scalar product of two matrices is $A \cdot B = \sum_{i,j=1}^3 a_{ij} b_{ij}$. Thus,

$$\psi_S(\varepsilon, S) = -T \cdot \bar{\varepsilon} + \hat{\psi}'(S). \quad (9)$$

Given are the positive constant c, the small positive constant ν, the volume force $b : [0,\infty) \times \Omega \to \mathbb{R}^3$ and the boundary and initial data $\gamma : [0,\infty) \times \partial\Omega \to \mathbb{R}^3$, $S_0 : \Omega \to \mathbb{R}$.

We thus complete the formulation of an initial-boundary value problem. The equations (1) and (2) differ from the system of linear elasticity only by the term $\bar{\varepsilon} S$,

which couples this system to equation (3). The evolution equation (3) for the order parameter S is non-uniformly parabolic because of the term $\mathrm{div}_x\,(|\nabla_x S|\nabla_x\Delta_x S)$.

Statement of the main result. Since we shall look for spherically symmetric solutions to problem (1)–(7), we can make suitable assumptions to reduce the problem to its one space dimensional form. To this end we now assume that the body force boundary and initial data and the unknowns, which are defined in the domain $\Omega\times(0,T_e)$, have the following form

$$b(t,x)=\hat{b}(t,r)\frac{x}{r},\ \gamma(t,x)=\hat{\gamma}(t,r),\ S_0(x)=\hat{S}_0(r)$$

and

$$u(t,x)=\hat{u}(t,r)\frac{x}{r},\ S(t,x)=\hat{S}(t,r),$$

respectively, where T_e is a positive constant which denotes the life-span of weak solutions, $r=|x|$, $\Omega=\{x\in\mathbb{R}^3\mid a<r<d\}$ for two positive constant $a,\ d$ satisfying $a<d$, and $\hat{b},\ \hat{\gamma},\ \hat{S}_0$ are given functions and $\hat{u},\ \hat{S}$ are scalar functions to be determined, which depend only on $t,\ r$. We write

$$x=(x_i),\ u=(u_i),\ T=(T_{ij}),\ D=(D^{ij}_{kl}),$$

hereafter, $i,j,k,l=1,2,3$, and we assume that D satisfies the properties of symmetry: $D^{ij}_{kl}=D^{kl}_{ij}=D^{ij}_{lk}=D^{ji}_{kl}$. Moreover we assume that the material is isotropic, namely we have

$$D^{ij}_{kl}=\mu_1\delta_{ik}\delta_{jl}+\frac{\mu_2}{3}\,\delta_{ij}\delta_{kl}, \tag{10}$$

where δ_{ij} is the Kronecker delta, and $\mu_1>0$, $\mu_2\geq 0$ are constants. For $\bar{\varepsilon}$, we assume that

$$\bar{\varepsilon}_{ij}=\lambda\delta_{ij}. \tag{11}$$

Then it follows that

$$D\varepsilon=\mu_1\varepsilon+\frac{\mu_2}{3}\,\mathrm{Trace}(\varepsilon)I,\quad D\bar{\varepsilon}=\mu_1\lambda I+\frac{\mu_2}{3}\,\mathrm{Trace}(\lambda I),\quad I=(\mu_1+\mu_2)I, \tag{12}$$

here for a matrix A, $\mathrm{Trace}(A)$ denotes the trace of A. Hence,

$$D\varepsilon\cdot\varepsilon=\mu_1\varepsilon\cdot\varepsilon+\frac{\mu_2}{3}\,(\mathrm{Trace}(\varepsilon))^2>0\quad\forall\varepsilon\neq 0. \tag{13}$$

Under these assumptions, equations (1)–(3) are reduced to

$$\hat{u}_{rr}+\frac{2}{r}\hat{u}_r-\frac{2}{r^2}\hat{u}\ =\ \mathcal{G}, \tag{14}$$

$$\frac{\partial}{\partial t}\,\hat{S}+c\,\frac{\partial}{\partial r}\,\Big((\nu\hat{S}_{rrr}+\mathcal{F}_2)|\hat{S}_r|\Big)\ =\ -\frac{2c}{r}\,(\nu\hat{S}_{rr}+\mathcal{F}_1)|\hat{S}_r|\,, \tag{15}$$

with $\mathcal{F}_1, \mathcal{F}_2, \mathcal{G}$ being nonlinear functions defined by

$$\begin{aligned} \mathcal{G} &= \mathcal{G}(\hat{S}_r, \hat{b}) = \frac{\lambda}{\mu}\,\hat{S}_r + \frac{\hat{b}}{\mu}, && (16)\\ \mathcal{F}_1 &= \mathcal{F}_1\Big(\hat{u}, \hat{u}_r, \hat{u}_{rr}, \hat{S}, \hat{S}_r\Big) \\ &= \lambda\Big(\hat{u}_r + \frac{2}{r}\,\hat{u}\Big) + \frac{2\nu}{r}\,\hat{S}_r - D\bar{\varepsilon}\cdot\bar{\varepsilon}\hat{S} - \hat{\psi}'(\hat{S}), && (17)\\ \mathcal{F}_2 &= \mathcal{F}_2\Big(\hat{u}, \hat{u}_r, \hat{u}_{rr}, \hat{S}, \hat{S}_r, \hat{S}_{rr}\Big) = \mathcal{F}_{1,r}. && (18) \end{aligned}$$

Since eq. (14) is linear, the inhomogeneous Dirichlet boundary condition for $\hat{u}$ can be reduced to the homogeneous one. So we may assume for simplicity that $\hat{\gamma} = 0$. Hence, simple computations show that (14) can be rewritten as

$$\hat{u}_r + \frac{2}{r}\,\hat{u} = \frac{\lambda}{\mu}\,\hat{S} + \frac{1}{\mu}\int_a^r \hat{b}(t,y)dy + C(t), \tag{19}$$

here, $C(t)$ is a constant depending on t and $\hat{\gamma}(t,r)$ which is zero by assumption. It thus follows from formula (19) and the boundary conditions for $\hat{u}$ that

$$\begin{aligned} \hat{u} &= \frac{1}{r^2}\left(\frac{\lambda}{\mu}\int_a^r y^2\hat{S}(t,y)dy + \frac{1}{\mu}\int_a^r x^2\int_a^x \hat{b}(t,y)dydx\right) \\ &\quad -\frac{1}{r^2}\,\frac{r^3-a^3}{d^3-a^3}\left(\frac{\lambda}{\mu}\int_a^d y^2\hat{S}(t,y)dy + \frac{1}{\mu}\int_a^d x^2\int_a^x \hat{b}(t,y)dydx\right). \end{aligned} \tag{20}$$

Therefore, (14)–(15) can be reduced to the following single equation

$$\frac{\partial}{\partial t}\,(r^2\hat{S}) + c\,\frac{\partial}{\partial r}\left(r^2(\nu\hat{S}_{rrr} + \mathcal{F})|\hat{S}_r|\right) = 0, \tag{21}$$

with

$$\mathcal{F} = \frac{\lambda}{\mu}\,(\lambda\hat{S}_r + \hat{b}) + \left(\frac{2\nu}{r}\,\hat{S}_r - D\bar{\varepsilon}\cdot\bar{\varepsilon}\hat{S} - \hat{\psi}'(\hat{S})\right)_r. \tag{22}$$

The boundary and initial conditions become

$$(\nu\hat{S}_{rrr} + \mathcal{F})|\hat{S}_r| = 0,\ (t,r) \in [0,T_e]\times\partial\Omega, \tag{23}$$

$$\hat{S}(0,r) = \hat{S}_0(r),\ r \in \Omega. \tag{24}$$

Consequently, the existence of spherically symmetric solutions to problem (1)–(7) is equivalent to solvability of problem (21)–(24), since $\hat{u}$ can be obtained from formula (20) once $\hat{S}$ is known.

The domain Ω is reduced to an interval: $\Omega = (a,d)$ is a bounded open interval with constants $a < d$. We write $Q_{T_e} := (0,T_e)\times\Omega$, where T_e is a positive constant.

To state the existence result for this problem we need two definitions. For $\mathcal{A} \subset Q_{T_e}$, $g : \mathcal{A} \to V \subset \mathbb{R}$ and $t \in [0,T_e]$ let

$$\mathcal{A}(t) = \{x \mid (t,x) \in \mathcal{A}\} \quad \text{and} \quad g(t) : \mathcal{A}(t) \to V,\ g(t)(x) = g(t,x).$$

Definition 1.1 *Let $\mathcal{A} \subset Q_{T_e}$ such that $\mathcal{A}(t)$ is open for almost all $t \in [0, T_e]$, and let $\alpha \in \mathbb{N}_0$. We call $g : \mathcal{A} \to \mathbb{R}$ the α-th local weak derivative of $S \in L^2(Q_{T_e})$ with respect to x in $\mathcal{A}$, if for almost all $t \in [0, T_e]$ the function $g(t)$ belongs to $L^{2,\mathrm{loc}}(\mathcal{A}(t))$ and is the local weak derivative of S in the usual sense:*

$$g(t) = \partial_x^\alpha S(t)|_{\mathcal{A}(t)}, \tag{25}$$

and if moreover there exists a sequence $\{\mathcal{A}_n\}_n$ of measurable sets $\mathcal{A}_n \subset \mathcal{A}$ with $g|_{\mathcal{A}_n} \in L^2(\mathcal{A}_n)$ for all $n \in \mathbb{N}$, such that

$$\mathrm{meas}\left(\mathcal{A} \setminus \bigcup_{n=1}^{\infty} \mathcal{A}_n\right) = 0.$$

Remark 1. The uniqueness of local weak derivatives in the sense of this definition is obvious because of (25), and it is clear that if $\mathcal{A}$ is open and if S has the local weak derivative $\partial_x^\alpha S$ in the usual sense in $\mathcal{A}$, then $\partial_x^\alpha S$ is also a local weak derivative in the sense of our definition. So Definition 1.1 generalizes the ordinary definition; this allows us to use the same name and the same notation $\partial_x^\alpha S$ as for ordinary local weak derivatives.

For a function $S \in L^2(0, T_e; H_N^2(\Omega))$, where $H_N^2(\Omega) = \left\{f \in H^2(\Omega) \mid \frac{\partial}{\partial n} f = 0, \text{ on } \partial\Omega\right\}$, let

$$\mathcal{A}^S = \{(t, r) \in Q_{T_e} \mid |S_r(t, r)| > 0\}.$$

By the Sobolev embedding theorem we see that $S_r(t)$ is continuous for almost all $t \in (0, T_e)$. This implies that $\mathcal{A}^S(t)$ is open for almost all t.

Definition 1.2 *Let $\hat{b} \in L^\infty(0, T_e; L^2(\Omega))$ and $\hat{S}_0 \in L^2(\Omega)$. A function $\hat{S}$ with*

$$\hat{S} \in L^2(0, T_e; H^2(\Omega)) \cap L^\infty(Q_{T_e}), \quad \hat{S}_r(t) \in H_0^1(\Omega) \text{ a.e. in } (0, T_e), \tag{26}$$

is a weak solution of the problem (21) – (24), if and only if $\hat{S}$, with local weak derivative $\hat{S}_{rrr}$ in $\mathcal{A}^{\hat{S}}$ and $|\hat{S}_r|\hat{S}_{rrr} \in L^1(\mathcal{A}^{\hat{S}})$, satisfies that

$$(r^2\hat{S}, \varphi_t)_{Q_{T_e}} + c\left(\nu r^2 \hat{S}_{rrr}|\hat{S}_r|, \varphi_r\right)_{\mathcal{A}^{\hat{S}}} + c\left(r^2\mathcal{F}|\hat{S}_r|, \varphi_r\right)_{Q_{T_e}} + (r^2\hat{S}_0, \varphi(0))_\Omega = 0 \tag{27}$$

holds for all $\varphi \in C_0^\infty((-\infty, T_e) \times \mathbb{R})$.

For the function $\hat{\psi}$, we need the following

Assumptions A. *The function $\hat{\psi}(S)$ is a smooth double-well potential, and it has two local minima at S_- and S_+ with $S_- < S_+$, one local maximum at S_* satisfying $S_- < S_* < S_+$; and $\hat{\psi}'(S) > 0$ for $S_- < S < S_*$ and $\hat{\psi}'(S) < 0$ for $S_* < S < S_+$. For simplicity, we assume further that*

$$\hat{\psi}^{(k)}(S_+) = 0 \quad \text{for} \quad 1 \le k \le 2m_1 - 1, \hat{\psi}^{(2m_1)}(S_+) > 0,$$
$$\hat{\psi}^{(k)}(S_-) = 0 \quad \text{for} \quad 1 \le k \le 2m_2 - 1, \hat{\psi}^{(2m_2)}(S_-) > 0.$$

and that $\hat{\psi}(S) \sim S^{2\ell_1}$ *as* $S \to \infty$, $\hat{\psi}(S) \sim S^{2\ell_2}$ *as* $S \to -\infty$, *where* m_1, m_2, ℓ_1, *and* ℓ_2 *are positive integers. Let* $\ell = \max\{\ell_1, \ell_2\}$. *Assume that* $\ell > 1$.

Remark 2. One typical example of $\hat{\psi}$ which satisfies assumptions A is $\hat{\psi}(S) = (S(1-S))^2$, with $S_+ = 1$, $S_- = 0$, $\ell = \ell_1 = \ell_2 = 2$ and $m_1 = m_2 = 1$.

We are now in a position to state the main result of this paper.

Theorem 1.3 *Assume that the double-well potential* $\hat{\psi}$ *satisfies assumptions A. Then to all* $\hat{S}_0 \in H^1(\Omega)$ *and* $\hat{b} \in L^2(Q_{T_e})$ *with* $\hat{b}_t \in L^2(Q_{T_e})$ *there exists a weak solution* $\hat{S}$ *to* (21)–(24), *which in addition to* (26) *satisfies* (20) *and*

$$\hat{S} \in L^\infty(0, T_e; H^1(\Omega)), \qquad \hat{S}_t \in L^{\frac{4}{3}}(0, T_e; W^{-1,\frac{4}{3}}(\Omega)), \tag{28}$$

$$|\hat{S}_r|\hat{S}_{rrr} \in L^{\frac{4}{3}}(Q_{T_e}), \tag{29}$$

where we defined $|\hat{S}_r|\hat{S}_{rrr} = 0$ *on* $Q_{T_e} \setminus \mathcal{A}^{\hat{S}}$.

The main difficulties of the proof of this theorem are caused by the term $|S_r|$ which results in that eq. (21) is degenerate and its coefficients are non-smooth. The coefficient of the principal term in (21) contains $|S_r|$, so this principal term can only be defined over a domain $\mathcal{A}^{\hat{S}}$ which may be not open. This leads to the difficulty of definition of weak derivatives $\hat{S}_{rrr}$.

Related results are Alber and Zhu [1] – [6], Kawashima and Zhu [12], and those for the degenerate Cahn-Hilliard equation and for the equation of thin film $S_t = -\mathrm{div}_x(m(S)\nabla_x\Delta_x S)$, where $m(S)$ vanishes at zero. We refer to [9, 11] and the references therein. However, the mathematical properties of (3) containing the term $|\nabla_x S|$ differ essentially from the ones of these equations.

2. Sketch of the proof of the main result

The proof of Theorem 1.3 consists of the following three steps. For simplicity, we drop the upper-script $\hat{}$, i.e. change $\hat{S}, \cdots$ back to $S, \cdots$.

Step 1. Construction of approximate solutions

To construct approximate solutions to (21)–(24) we prove that there exist weak solutions to the following initial-boundary value problem

$$\begin{aligned}
(r^2 S)_t + c\left(r^2(\nu S_{rrr} + \mathcal{F}_\kappa)|S_r|_\kappa\right)_r &= 0 \qquad \text{in } Q_{T_e}, && (1)\\
S_r &= 0 \qquad \text{on } [0,T_e]\times\partial\Omega, && (2)\\
(\mathcal{F}_\kappa + \nu S_{rrr})|S_r|_\kappa &= 0 \qquad \text{on } [0,T_e]\times\partial\Omega, && (3)\\
S|_{t=0} &= S_0 \qquad \text{in } \Omega, && (4)
\end{aligned}$$

where κ is a fixed positive constant, $|y|_\kappa$ is defined by $|y|_\kappa = \sqrt{|y|^2 + \kappa^2}$, and $\mathcal{F}_\kappa$ is the smoothed $\mathcal{F}$ in which b is replaced by its smooth approximation b^κ.

Eq. (1) is quasi-linear, uniformly parabolic over a domain that S_r is bounded. However it is not easy to prove the existence of classical solution to problem (1)–(4), whence we consider the weak solutions to this problem. By definition, $S \in L^2(0, T_e; H^1(\Omega)$ with $S_{rrr} \in L^2(Q_{T_e})$ is a weak solution of (1)–(4) if and only if for all $\varphi \in C_0^\infty((-\infty, T_e) \times \mathbb{R})$

$$-(r^2 S, \varphi_t)_{Q_{T_e}} = (r^2 S_0, \varphi(0))_\Omega + c\big(r^2(\nu S_{rrr} + \mathcal{F}_\kappa)|S_r|_\kappa, \varphi_r\big)_{Q_{T_e}}. \tag{5}$$

Step 2. Main a-priori estimates

Lemma 2.1 *There is a constant C, independent of κ, such that for any $t \in [0, T_e]$*

$$\| S_r^\kappa\|^2_{H^1(\Omega)} + \int_{Q_t} (|S_r^\kappa|_\kappa + \kappa)|S_{rrr}^\kappa|^2 d(\tau, y) \leq C, \tag{6}$$

$$\| |S_r^\kappa|_\kappa S_{rrr}^\kappa\|_{L^{\frac{4}{3}}(Q_t)} \leq C. \tag{7}$$

Step 3. Limits

To investigate the limits of approximate solutions constructed in Step 1, we need the Egorov theorem.

Theorem 2.2 (Egorov) *Let (Γ, Σ, μ) be a measure space with $\mu(\Gamma) < \infty$, let $f, f^1, f^2, f^3, \cdots$ be real valued, measurable functions on Γ, and assume that $f^j(x) \to f(x)$ as $j \to \infty$ for almost every $x \in \Gamma$.*

Then, for every $\varepsilon > 0$ there is a subset $M_\varepsilon \subset \Gamma$ with $\mu(M_\varepsilon) > \mu(\Gamma) - \varepsilon$ such that $f^j(x)$ converges to $f(x)$ uniformly on M_ε. That is, for every $\delta > 0$ there is an N_δ such that when $j > N_\delta$ we have that for every $x \in M_\varepsilon$

$$|f^j(x) - f(x)| < \delta.$$

With the help of this theorem we can get the local weak derivative S_{rrr} as follows. Decompose the set $\hat{\mathcal{A}}_n = \{(t, r) \in Q_{T_e} \mid |S_r(t, r)| > \frac{1}{n}\}$ into a set $\mathcal{A}_n$ (on which the sequence S_r^κ converges uniformly to S_r and thus satisfies $|S_r^\kappa| \geq \frac{1}{2n}$ for sufficiently small κ) and the set $\hat{\mathcal{A}}_n \setminus \mathcal{A}_n$ (which has small measure). Using the uniform estimate $\int_{Q_{T_e}} (|S_r^\kappa|_\kappa + \kappa)|S_{rrr}^\kappa|^2 \, d(\tau, r) \leq C$, we can then show that S_{rrr}^κ converges in $L^2(\mathcal{A}_n)$ to S_{rrr}. Finally, we apply the fact that $\mathcal{A}^S$ differs from $\bigcup_{n=1}^\infty \mathcal{A}_n$ only by a set of measure zero. We then have the following key lemma.

Lemma 2.3 *The limit function S has the local weak L^2–derivative S_{rrr} on $\mathcal{A}^S$ in the sense of Definition 1.1. Moreover, there exists a subsequence S^κ such that $|S_r^\kappa|_\kappa S_{rrr}^\kappa \rightharpoonup \chi$, weakly in $L^{\frac{4}{3}}(Q_{T_e})$, where the function $\chi = \chi(t, r)$ in $L^{\frac{4}{3}}(Q_{T_e})$ is given by $\chi = 0$, if $S_r = 0$, and $= |S_r| S_{rrr}$, if $S_r \neq 0$.*

Acknowledgements

This work was supported by grant No. MTM2011-24054, Ministerio de Ciencia e Innovación of the Spanish Government.

References

[1] Alber, H.-D. and Zhu, P.: Evolution of phase boundaries by configurational forces. Arch. Rational Mech. Anal. **1**85 (2007), 235–286.

[2] Alber, H.-D. and Zhu, P.: Solutions to a model for interface motion by interface diffusion. Proc. Royal Soc. Edinburgh, **1**38A (2008), 923–955.

[3] Alber, H.-D. and Zhu, P.: Interface motion by interface diffusion driven by bulk energy: justification of a diffusive interface model. Continuum Mech. Thermodyn., **2**3 (2011), 139–176.

[4] Alber, H.-D. and Zhu, P.: Solutions to a model with Neumann boundary conditions for phase transitions driven by configurational forces. Nonlinear Anal. RWA **1**2 (2011), 1797–1809.

[5] 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, Online July 2012.

[6] Alber, H.-D. and Zhu, P.: Spherically symmetric solutions to a model for grain boundary motion. Submitted, 2013.

[7] Cahn, J. and Hilliard, J.: Free energy of a nonuniform system. I. Interfacial free energy. J. Chem. Phys. **2**8 (1958), 258–267.

[8] Chen, L.: Phase-fieldmodels for microstructure evolution. Annu. Rev. Mater. Res. **3**2 (2002), 113–140.

[9] Elliott, C. and Garcke, H.: On the Cahn-Hilliard equation with degenerate mobility. SIAM J. Math. Anal. **2**7 (1996), 404–423.

[10] Emmerich, H.: The diffuse interface approach in materials science. *Lecture Notes in Physics, Springer*, Heidelberg 2003.

[11] Garcke, H.: On Cahn-Hilliard systems with elasticity. Proc. R. Soc. Edinb., Sect. A, Math. **1**33(2) (2003), 307 – 331.

[12] Kawashima, S. and Zhu, P.: Traveling waves for models of phase transitions of solids driven by configurational forces. Disc. Conti. Dyna. Systems B **1**5 (2011), 309–323.

[13] Zhu, P.: *Solid-solid phase transitions driven by configurational forces: A phase-field model and its validity.* Lambert Academic Publishing, Germany, 2011.

Conference Applications of Mathematics 2013
in honor of the 70th birthday of Karel Segeth.
Jan Brandts, Sergey Korotov, Michal Křížek,
Jakub Šístek, and Tomáš Vejchodský (Eds.),
Institute of Mathematics AS CR, Prague 2013

APPLICATION OF RICHARDSON EXTRAPOLATION WITH THE CRANK–NICOLSON SCHEME FOR MULTI-DIMENSIONAL ADVECTION

Zahari Zlatev[1], Ivan Dimov[2], István Faragó[3], Krassimir Georgiev[2], Ágnes Havasi[3], Tzvetan Ostromsky[2]

[1] Department of Environmental Science, Aarhus University
Roskilde, Denmark
zz@dmu.dk
[2] Institute of Information and Communication Technologies
Bulgarian Academy of Sciences
Sofia, Bulgaria
ivdimov@bas.bg, georgiev@parallel.bas.bg, ceco@parallel.bas.bg
[3] Department of Applied Analysis and Computational Mathematics
Eötvös Loránd University
and MTA-ELTE Numerical Analysis and Large Networks Research Group
Budapest, Hungary

Abstract

Multi-dimensional advection terms are an important part of many large-scale mathematical models which arise in different fields of science and engineering. After applying some kind of splitting, these terms can be handled separately from the remaining part of the mathematical model under consideration. It is important to treat the multi-dimensional advection in a sufficiently accurate manner. It is shown in this paper that high order of accuracy can be achieved when the well-known Crank–Nicolson numerical scheme is combined with the Richardson extrapolation.

1. Multi-dimensional advection equations

Consider the multi-dimensional advection equation:

$$\frac{\partial c}{\partial t} = -\sum_{q=1}^{Q} u_q \frac{\partial c}{\partial x_q} \tag{1}$$

with $Q \geq 0$, $t \in [a, b]$ and $x_q \in [a_q, b_q]$ for $q = 1, 2, \ldots, Q$. It is assumed that the coefficients $u_q = u_q(t, x_1, x_2, \ldots x_Q)$, $q = 1, 2, \ldots, Q$, before the spatial partial derivatives in the right-hand-side of the partial differential equation (1) are some given functions.

Let D be the domain in which the independent variables involved in (1) vary and assume that:

$$(t, x_1, x_2, \ldots x_Q) \in D \Rightarrow t \in [a, b] \ \wedge \ x_0 \in [a_q, b_q] \text{ for } q = 1, 2, \ldots, Q. \tag{2}$$

By applying the definition proposed in (2), it is assumed here that the obtained domain D is rather special (being a multi-dimensional parallelepiped), but this assumption is done only for the sake of simplicity. In fact, many of the results will also be valid for some considerably more complicated domains.

It will always be assumed that the unknown function $c = c(t, x_1, x_2, \ldots, x_Q)$ is continuously differentiable up to some order $2p$ with $p \geq 1$ in all points of the domain D and with respect to all independent variables. Here p is the order of the numerical method which will be used in order to obtain some approximations of the unknown function at the points of some grid, which is appropriately selected (see below) in the domain defined in (2).

For some of the proofs, see [8], it will also be necessary to assume that continuous derivatives up to order two of all given functions u_q exist with respect of all independent variables.

The multi-dimensional advection equation (1) must always be considered together with some initial and boundary conditions.

The following notation in connection with some given positive increments h_q is useful in the proofs (see also [6]):

$$\overline{x} = (x_1, x_2, \ldots, x_Q), \tag{3}$$

$$\overline{x}^{(+q)} = (x_1, x_2, \ldots, x_{q-1}, x_q + h_q, x_{q+1}, \ldots, x_Q), \ q = 1, 2, \ldots, Q, \tag{4}$$

$$\overline{x}^{(-q)} = (x_1, x_2, \ldots, x_{q-1}, x_q - h_q, x_{q+1}, \ldots, x_Q), \ q = 1, 2, \ldots, Q, \tag{5}$$

$$\overline{x}^{(+0.5q)} = (x_1, x_2, \ldots, x_{q-1}, x_q + 0.5h_q, x_{q+1}, \ldots, x_Q), \ q = 1, 2, \ldots, Q, \tag{6}$$

$$\overline{x}^{(-0.5q)} = (x_1, x_2, \ldots, x_{q-1}, x_q - 0.5h_q, x_{q+1}, \ldots, x_Q), \ q = 1, 2, \ldots, Q. \tag{7}$$

2. Expanding the unknown function in Taylor series

The following result is very important in the efforts (see Section 6 and the conclusions in Section 7) to establish the order of accuracy which can be achieved when the Crank–Nicolson scheme is combined with the Richardson extrapolation.

Theorem 2.1 *Consider the multi-dimensional advection equation (1). Assume that $(t, \overline{x}) \in D$ is an arbitrary but fixed point and introduce the increments $k > 0$ and $h_q > 0$ such that $t + k \in [a, b]$, $x_q - h_q \in [a_q, b_q]$ and $x_q + h_q \in [a_q, b_q]$ for all $q = 1, 2, \ldots, Q$. Assume furthermore that the unknown function $c = c(t, \overline{x})$ is continuously differentiable up to some order $2p$ with regard to all independent variables. Then there exists an expansion in Taylor series of the unknown function $c = c(t, \overline{x})$ around the point $(t + 0.5k, \overline{x})$ which contains terms involving only even degrees of the increments k and h_q, $q = 1, 2, \ldots, Q$.*

Proof. The main ideas of the proof are quite straightforward (the unknown function c must be expanded in Taylor series and the series should be truncated after the first $2p$ terms), but it is rather long and complicated.
The full proof of Theorem 2.1 can be found in [8]. More precisely, the following equality is proved there:

$$\begin{aligned}\frac{c(t+k,\overline{x})-c(t,\overline{x})}{k} &= -\sum_{q=1}^{Q} u_q(t+0.5k,\overline{x})\frac{c(t+k,\overline{x}^{(+q)})-c(t+k,\overline{x}^{(-q)})}{4h_q} \\ &\quad -\sum_{q=1}^{Q} u_q(t+0.5k,\overline{x})\frac{c(t,\overline{x}^{(+q)})-c(t,\overline{x}^{(-q)})}{4h_q} \\ &\quad +\sum_{s=1}^{p} k^{2s}K^{(2s)}+\mathcal{O}(k^{2p+1}),\end{aligned} \tag{8}$$

where $K_t^{(2s)}$ and $K_q^{(2s)}$ are some constants and

$$K^{(2s)}=K_t^{(2s)}+\sum_{q=1}^{Q}\frac{h_q^{2s}}{k^{2s}}K_q^{(2s)}. \tag{9}$$

It should be noted here that it is assumed that all ratios h_q/k, $q=1,2,\ldots,Q$, remain constants when $k\to 0$ (which can easily be achieved; for example by reducing all spatial increments h_q by a factor of two when the time-increment k is reduced by a factor of two).

3. Designing a second-order numerical method

Consider the grids:

$$G_t=\{t_n, n=0,1,\ldots,N_t \mid t_0=a, t_n=t_{n-1}+k, n=1,2,\ldots,N_t, k=\frac{b-a}{N_t}, t_{N_t}=b\} \tag{10}$$

and (for $q=1,2,\ldots,Q$ and $h_q=(b_q-a_q)/N_q$)

$$G_x^{(q)}=\{x_q^{i_q}, i_q=0,1,\ldots,N_q \mid x_q^0=a_q, x_q^{i_q}=x_q^{i_q-1}+h_q, i=1,2,\ldots,N_q, x_q^{N_q}=b_q\}. \tag{11}$$

Introduce the following notations:

$$\tilde{x}=(x_1^{i_1},x_2^{i_2},\ldots,x_Q^{i_Q}), \tag{12}$$

$$\tilde{x}^{(+q)}=(x_1^{i_1},x_2^{i_2},\ldots,x_{q-1}^{i_{q-1}},x_q^{i_q}+h_q,x_{q+1}^{i_{q+1}},\ldots,x_Q^{i_Q}), \tag{13}$$

$$\tilde{x}^{(-q)}=(x_1^{i_1},x_2^{i_2},\ldots,x_{q-1}^{i_{q-1}},x_q^{i_q}-h_q,x_{q+1}^{i_{q+1}},\ldots,x_Q^{i_Q}), \tag{14}$$

where $x_q^{i_q}\in G_x^{(q)}$ for $q=1,2,\ldots,Q$.

In this notation the following numerical method can be defined:

$$\frac{\tilde{c}(t_{n+1},\tilde{x})-\tilde{c}(t_n,\tilde{x})}{k} \tag{15}$$
$$=-\sum_{q=1}^{Q}u_q(t_n+0.5k,\tilde{x})\frac{\tilde{c}(t_{n+1},\tilde{x}^{(+q)})-\tilde{c}(t_{n+1},\tilde{x}^{(-q)})+\tilde{c}(t_n,\tilde{x}^{(+q)})-\tilde{c}(t_n,\tilde{x}^{(-q)})}{4h_q}.$$

The computational device introduced by the finite difference equation (15) is often called the Crank–Nicolson scheme (see, for example, [5]). It is clear that (15) can be obtained from (8) by neglecting the terms in the last line and by assuming additionally that an arbitrary inner point of the grids defined by (10) and (11) is considered.

The quantities $\tilde{c}(t_n,\tilde{x})$ can be considered as approximations of the exact values of the unknown function $c(t_n,\tilde{x})$ at the grid-points from the grids defined by (10) and (11). It can easily be shown that the method introduced in (15) is of order two in respect to **all** independent variables.

Assume that the values of $\tilde{c}(t_n,\tilde{x})$ have been calculated for all grid-points of (11). Then the values $\tilde{c}(t_{n+1},\tilde{x})$ of the unknown function at the next time-point $t_{n+1}=t_n+k$ can be obtained by solving a huge system of linear algebraic equations of dimension $\tilde{N}$ where $\tilde{N}$ is defined by

$$\tilde{N}=\prod_{q=1}^{Q}(N_q-1). \tag{16}$$

4. Application of Richardson extrapolation

Consider (15) with $\tilde{c}$ replaced by z when $t=t_{n+1}$:

$$\frac{z(t_{n+1},\tilde{x})-\tilde{c}(t_n,\tilde{x})}{k} \tag{17}$$
$$=-\sum_{q=1}^{Q}u_q(t_n+0.5k,\tilde{x})\frac{z(t_{n+1},\tilde{x}^{(+q)})-z(t_{n+1},\tilde{x}^{(-q)})+\tilde{c}(t_n,\tilde{x}^{(+q)})-\tilde{c}(t_n,\tilde{x}^{(-q)})}{4h_q}.$$

Suppose that $0.5k$ and $0.5h_q$ are considered instead of k and h_q $(q=1,2,\ldots,Q)$, respectively. Consider, as in formulae (5) and (6) but in the grid-points of the grids (10) and (11), the two vectors $\tilde{x}^{(+0.5q)}=(x_1^{i_1},x_2^{i_2},\ldots,x_{q-1}^{i_{q-1}},x_q^{i_q}+0.5h_q,x_{q+1}^{i_{q+1}},\ldots,x_q^{i_{N_q}})$ and $\tilde{x}^{(-0.5q)}=(x_1^{i_1},x_2^{i_2},\ldots,x_{q-1}^{i_{q-1}},x_q^{i_q}-0.5h_q,x_{q+1}^{i_{q+1}},\ldots,x_q^{i_{N_q}})$ for $q=1,2,\ldots,Q$.

Perform now additionally two small steps:

$$\frac{w(t_n+0.5k,\tilde{x})-\tilde{c}(t_n,\tilde{x})}{0.5k} \tag{18}$$

$$\begin{aligned}
&= -\sum_{q=1}^{Q} u_q(t_n+0.25k,\tilde{x})\frac{w(t_n+0.5k,\tilde{x}^{(+0.5q)})-w(t_n+0.5k,\tilde{x}^{(-0.5q)})}{4(0.5h_q)} \\
&\quad -\sum_{q=1}^{Q} u_q(t_n+0.25k,\tilde{x})\frac{\tilde{c}(t_n,\tilde{x}^{(+0.5q)})-\tilde{c}(t_n,\tilde{x}^{(-0.5q)})}{4(0.5h_q)}
\end{aligned}$$

$$\frac{w(t_n+k,\tilde{x})-w(t_n+0.5k,\tilde{x})}{0.5k} \tag{19}$$

$$\begin{aligned}
&= -\sum_{q=1}^{Q} u_q(t_n+0.75k,\tilde{x})\frac{w(t_n+k,\tilde{x}^{(+0.5q)})-w(t_n+k,\tilde{x}^{(-0.5q)})}{4(0.5h_q)} \\
&\quad -\sum_{q=1}^{Q} u_q(t_n+0.75k,\tilde{x})\frac{w(t_n+0.5k,\tilde{x}^{(+0.5q)})-w(t_n+0.5k,\tilde{x}^{(-0.5q)})}{4(0.5h_q)}\,.
\end{aligned}$$

The Richardson extrapolation can now be calculated by using the following formula (exploiting here the fact that the order of the underlying numerical method is of order of accuracy two in regard to all independent variables):

$$\tilde{c}(t_{n+1},\tilde{x}) = \frac{4w(t_{n+1},\tilde{x})-z(t_{n+1},\tilde{x})}{3}. \tag{20}$$

If the order of accuracy of the underlying method is not 2 but p, then the numbers 4 and 3 in (20) should be replaced by 2^p and 2^p-1, respectively.

5. Several general remarks on the Richardson extrapolation

Assume now that an arbitrary method of order p is used. Then, as mentioned above, (20) can be written as

$$\tilde{c}(t_{n+1},\tilde{x}) = \frac{2^p w(t_{n+1},\tilde{x})-z(t_{n+1},\tilde{x})}{2^p-1}. \tag{21}$$

The exact solution $c(t_{n+1},\tilde{x})$ can be expressed in the following two ways, where K is some constant and k is the time-increment:

$$c(t_{n+1},\tilde{x}) = z(t_{n+1},\tilde{x}) + k^p K + \mathcal{O}(k^{p+1}), \tag{22}$$

$$c(t_{n+1},\tilde{x}) = w(t_{n+1},\tilde{x}) + (0.5k)^p K + \mathcal{O}(k^{p+1}). \tag{23}$$

Eliminating the terms containing K in (22) and (23) gives:

$$c(t_{n+1},\tilde{x}) = \frac{2^p w(t_{n+1},\tilde{x})-z(t_{n+1},\tilde{x})}{2^p-1} + \mathcal{O}(k^{p+1}). \tag{24}$$

Denote:

$$\tilde{c}(t_{n+1},\tilde{x}) = \frac{2^p w(t_{n+1},\tilde{x}) - z(t_{n+1},\tilde{x})}{2^p - 1}. \tag{25}$$

It is clear that the approximation $\tilde{c}(t_{n+1},\tilde{x})$, being of order $p+1$, will be more accurate than both $z(t_{n+1},\tilde{x})$ and $w(t_{n+1},\tilde{x})$ when the stepsize k is sufficiently small. Thus, the Richardson extrapolation can be used in the efforts to improve the accuracy.

The Richardson extrapolation can also be used in an attempt to evaluate the leading term of the local error of the approximation $w(t_{n+1},\tilde{x})$. Subtract (22) from (23), neglect the rest terms $\mathcal{O}(k^{p+1})$ and solve for K. The result is:

$$K = \frac{2^p[w(t_{n+1},\tilde{x}) - z(t_{n+1},\tilde{x})]}{k^p(2^p - 1)}. \tag{26}$$

Substitute K from (26) in (23):

$$c(t_{n+1},\tilde{x}) - w(t_{n+1},\tilde{x}) = \frac{w(t_{n+1},\tilde{x}) - z(t_{n+1},\tilde{x})}{2^p - 1} + \mathcal{O}(k^{p+1}), \tag{27}$$

which means that the quantity:

$$E_n = \frac{w(t_{n+1},\tilde{x}) - z(t_{n+1},\tilde{x})}{2^p - 1} \tag{28}$$

can be used as an evaluation of the local error of the approximation $w(t_{n+1},\tilde{x})$ when the time-increment k is sufficiently small. If the evaluation of the local error computed by using (28) is not acceptable, then E_n can also be used to determine a new stepsize k which will hopefully give an acceptable error. Assume that the requirement for the accuracy imposed by the user is TOL. Then the new, hopefully better, time-increment k_{new} can be calculated by

$$k_{\text{new}} = \gamma \frac{TOL}{E_n} k, \tag{29}$$

where $\gamma < 1$ is used as a precaution factor, see for example [4]. Thus, the Richardson extrapolation can be applied in codes with automatic stepsize control.

The use of the Richardson extrapolation for stepsize control is relatively easy when systems of ordinary differential equations are solved numerically. The procedure becomes difficult when systems of partial differential equations are to be handled, because of the introduction of the assumption made in (9). This assumption implies that if k is multiplied by the factor $\gamma TOL/E_n$, then all h_q must be multiplied by the same factor in order to keep the ratios h_q/k constant. This difficulty can be avoided in the special case where all $K_q^{(2s)}$ are much smaller than $K_t^{(2s)}$. Then the Richardson extrapolation can be slightly modified so that all h_q are kept constant and only the time-stepsize k is to be controlled (thus, the situation becomes in principle the same as that appearing when systems of ordinary differential equations are treated).

It must be emphasized here that the Richardson extrapolation does not depend too much on the particular method used. It can be utilized both when classical numerical algorithms are applied in the solution of differential equations and when more advanced numerical methods which are combination of splitting procedures and classical numerical algorithms are devised and used. Two issues are important: (a) the large time-increment and the two small time-increments must be handled by the same numerical method and (b) the order p of the selected method should be known. If the Richardson extrapolation is used in connection with the multi-dimensional equation (1), then it is appropriate, see (9), to assume that for all values of q the ratios h_q/k are constants.

Much more useful details about different application issues related to the introduction of the Richardson extrapolation and its stability properties can be found in [2, 3, 9, 7, 10].

The above analysis shows that the accuracy order is as a rule increased by one. In the next section it will be shown that the application of the Richardson extrapolation in connection with the numerical method derived in Section 3 gives better accuracy when applied in the solution of (1).

6. Accuracy of Richardson extrapolation

Theorem 6.1 *Consider the multi-dimensional advection equation* (1). *Assume that the coefficients u_q before the spatial derivatives in* (1) *are continuously differentiable with respect to all independent variables and continuous derivatives of the unknown function c up to order four exist, again with respect to all variables. Then the combination of the numerical method* (15) *and the Richardson extrapolation is of order of accuracy four.*

Proof. The ideas, on which the proof is to be based, are quite clear. One must apply the result proved in Theorem 2.1 for $p = 2$ under an assumption that the numerical method defined by (15) is used at the grid-points of (10) and (11). However, the actual poof is very long and rather complicated. It can be found in [8]. It should be noted (see also Section 5) that in general the use of the Richardson extrapolation is leading to an increase of the accuracy order of the underlying numerical method by one. For the second-order numerical method (15) the accuracy order is increased by **two** (from order two to order four) when it is applied to the multi-dimensional advection equation (1) together with the Richardson extrapolation.

7. Conclusions

The result proved in this paper is a generalization of the result proved in [6], where the much simpler one-dimensional advection is handled.

It will be interesting to investigate whether the result proved in Theorem 6.1 for equation (1) can be extended for the more general multi-dimensional advection equation:

$$\frac{\partial c}{\partial t} = -\sum_{q=1}^{Q} \frac{\partial (u_q c)}{\partial x_q}, \ x_q \in [a_q, b_q] \text{ for } q = 1, 2, \ldots, Q \text{ with } Q \geq 1, \ t \in [a, b]. \quad (30)$$

Richardson extrapolation can be repeatedly applied (see, for example, [1]). Theorem 6.1 indicates that when this is done, the order of accuracy will be increased by two after each successive application of the Richardson extrapolation. This remark explains why Theorem 2.1 is proved for an arbitrary value of p and not only for $p = 2$ as required in Theorem 6.1.

Acknowledgements

The research of the Bulgarian authors was partly supported by the Bulgarian National Science Fond under Grants DFNI I01/0005, DCVP 02/1, and DTK 02/44. The research of I. Faragó was supported by Hungarian National Research Fund OTKA No. K67819.

References

[1] Christiansen, E. and Petersen, H. G.: Estimation of convergence orders in repeated Richardson Extrapolation. BIT Numer. Math. **20**(1) (1989), 48–59.

[2] Faragó, I., Havasi, Á., and Zlatev, Z.: Richardson extrapolated sequential splitting and its application. J. Comp. Appl. Math. **226**(2) (2009), 218–227.

[3] Faragó, I., Havasi,Á., and Zlatev, Z.: Efficient implementation of stable Richardson Extrapolation algorithms. Comput. Math. Appl. **60**(8) (2010), 2309–2325.

[4] Shampine, L. F.: *Numerical solution of ordinary differential equations.* Chapman and Hall, New York, London, 1994.

[5] Strikwerda, J. C.: *Finite difference schemes and partial differential equations.* SIAM (Society of Industrial and Applied Mathematics), Philadelphia, 1989.

[6] Zlatev, Z., Dimov, I., Faragó, I., Georgiev, K., Havasi, Á., and Ostromsky, T.: Solving advection equations by applying the Crank–Nicolson scheme combined with the Richardson Extrapolation. International Journal of Differential Equations, doi:10.1155/2011/520840, 2011 (open access article).

[7] Zlatev, Z., Dimov, I., Faragó, I., Georgiev, K., Havasi, Á., and Ostromsky, T.: Richardson Extrapolated numerical methods for treatment of one-dimensional advection equations. In: I. Dimov, S. Dimova, and N. Kolkovska (Eds.) *Numerical Methods and Applications, Lecture Notes in Computer Science*, vol. 6046, pp. 198–206. Springer, Berlin, 2011.

[8] Zlatev, Z., Dimov, I., Faragó, I., Georgiev, K., Havasi, Á., and Ostromsky, T.: On the accuracy of an application of the Richardson Extrapolation in the treatment of multi-dimensional advection equations. http://www.cs.elte.hu/∼faragois/Richardson_Extrapolation_for_multidimensional advection_equations.new.pdf, 2013.

[9] Zlatev, Z., Faragó, I., and Havasi, Á.: Stability of the Richardson Extrapolation applied together with the θ-method. J. Comp. Appl. Math. **235**(2) (2010), 507–517.

[10] Zlatev, Z., Faragó, I., and Havasi, Á.: Richardson Extrapolation combined with the sequential splitting and the θ-method. Cent. Eur. J. Math. **10**(1) (2012), 159–172.

LIST OF AUTHORS

LIST OF PARTICIPANTS

Hana Bílková
Academy of Sciences, Prague, Czech Republic, hanka@cs.cas.cz

Jan Brandts
University of Amsterdam, Netherlands, janbrandts@gmail.com

Pavel Burda
Czech Technical University, Prague, Czech Republic, Pavel.Burda@fs.cvut.cz

Jim Byrnes
Prometheus Inc., Newport, Rhode Island, USA, jim@prometheus-us.com

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

Vít Dolejší
Charles University, Prague, Czech Republic, dolejsi@karlin.mff.cuni.cz

Zdeněk Dostál
Technical University of Ostrava, Czech Republic, zdenek.dostal@vsb.cz

István Faragó
Eötvös Loránd University, Budapest, Hungary, faragois@cs.elte.hu

Leandro Farina
Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brazil, and Basque Center for Applied Mathematics, Bilbao, Basque Country, Spain,
farina@mat.ufrgs.br

Ruszlán Farzan
Szechényi István University, Győr, Hungary, farzan@sze.hu

Miloslav Feistauer
Charles University, Prague, Czech Republic, feist@karlin.mff.cuni.cz

Miroslav Fiedler
Academy of Sciences, Prague, Czech Republic, fiedler@cs.cas.cz

Pavla Fraňková
University of West Bohemia, Pilsen, Czech Republic, frankova@kma.zcu.cz

Antti Hannukainen
Aalto University, Finland, antti.hannukainen@tkk.fi

Milan Hanuš
University of West Bohemia, Pilsen, Czech Republic, mhanus@kma.zcu.cz

Drahoslava Janovská
Institute of Chemical Technology, Prague, Czech Republic,
Drahoslava.Janovska@vscht.cz

Vladimír Janovský
Charles University, Prague, Czech Republic, janovsky@karlin.mff.cuni.cz

Josef Ježek
Charles University, Prague, Czech Republic, jezek@prfdec.natur.cuni.cz

János Karátson
Eötvös Loránd University, Hungary, karatson@cs.elte.hu

Lucie Kárná
Czech Technical University, Prague, Czech Republic, karna@fd.cvut.cz

Jaroslav Kautský
Flinders University, Australia, jardakau@internode.on.net

Hana Kopincová
University of West Bohemia, Pilsen, Czech Republic, kopincov@kma.zcu.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
Academy of Science, Prague, Czech Republic, krizek@math.cas.cz

Roman Kužel
University of West Bohemia, Pilsen, Czech Republic, rkuzel@kma.zcu.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

Vratislava Mošová
Moravian University College Olomouc, Czech Republic, Vratislava.Mosova@mvso.cz

Šárka Nečasová
Academy of Sciences, Prague, Czech Republic, matus@math.cas.cz

Miroslav Pospíšek
ANECT, Prague, Czech Republic, mpospisek@anect.com

Milan Práger
Academy of Sciences, Prague, Czech Republic, prager@math.cas.cz

Jiří Rákosník
Academy of Sciences, Prague, Czech Republic, rakosnik@math.cas.cz

Beloslav Riečan
Matej Bel University in Banská Bystrica, Slovakia, Beloslav.Riecan@umb.sk

Hans-Goerg Roos
Technical University of Dresden, Germany, hans-goerg.roos@tu-dresden.de

Miroslav Rozložník
Academy of Sciences, Prague, Czech Republic, miro@cs.cas.cz

Karel Segeth
Academy of Sciences, Prague, Czech Republic, segeth@math.cas.cz

Carmen Simerská
Czech Technical University, Prague, Czech Republic, Carmen.Simerska@vscht.cz

Jakub Šístek
Academy of Sciences, Prague, Czech Republic, sistek@math.cas.cz

Lawrence Somer
The Catholic University of America, Washington, D.C., USA, 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

Jiří Vala
Brno University of Technology, Czech Republic, Vala.J@fce.vutbr.cz

Petr Vaněk
University of West Bohemia, Pilsen, Czech Republic, petrvanek09@seznam.cz

Zbyněk Vastl
University of West Bohemia, Pilsen, Czech Republic, zvastl@kma.zcu.cz

Tomáš Vejchodský
Academy of Sciences, Prague, Czech Republic, vejchod@math.cas.cz

Emil Vitásek
Academy of Sciences, Prague, Czech Republic, vitas@math.cas.cz

Vítězslav Vít Vlček
CA Technologies, Prague, Czech Republic, vitezslav@vvvlcek.info

Hehu Xie
Chinese Academy of Sciences, Beijing, China, hhxie@lsec.cc.ac.cn

Shuhua Zhang
Tianjin University of Finance and Economics, China, shuhua55@126.com

Jan Zítko
Charles University, Prague, Czech Republic, zitko@karlin.mff.cuni.cz

PROGRAM OF THE CONFERENCE

Wednesday, May 15

13.00 – 14.00 Registration

14.00 – 14.30 Opening
Presentation of the Medal of the Czech Mathematical Society to Karel Segeth

14.30 – 15.00 Jan Brandts
Counting triangles that share their vertices with the unit n-cube

15.00 – 15.30 Coffee Break

15.30 – 16.00 Miloslav Feistauer
Analysis of the discontinuous Galerkin method for elliptic problems with boundary singularities

16.00 – 16.30 Sergey Korotov
Nonobtuse refinements of tetrahedral FE meshes

16.30 – 17.00 Antti Hannukainen
Pseudospectral analysis of preconditioners for the Helmholtz equation

17.00 – 17.20 Pavel Burda
Application of analytical solution of rotationally symmetric Stokes flow near corners

18.00 – 22.00 Welcome Party, Blue Hall

Thursday, May 16

9.00 – 9.30 Vít Dolejší
hp-anisotropic mesh adaptation technique based on interpolation error estimates

9.30 – 10.00 János Karátson
Discrete maximum principles for nonlinear PDEs

10.00 – 10.30 Beloslav Riečan
Small systems and fuzzy sets

10.30 – 11.00 Coffee Break

11.00 – 11.30 Zdeněk Dostál
Scalable algorithm and variationally consistent discretization for contact problems

11.30 – 12.00 István Faragó
Convergence and stability constant in the theta-method

12.00 – 14.00 Lunch Break

14.00 – 14.20 Miroslav Fiedler
A quasicondition number of matrices

14.20 – 14.40 Šárka Nečasová
Incompressible limits of fluids excited by moving domains

14.40 – 15.00 Drahoslava Janovská
Zero points of quadratic matrix polynomials

15.00 – 15.20 Lucie Kárná
Detection codes in railway interlocking systems

15.20 – 15.40 Coffee Break

15.40 – 16.00 Jiří Vala
On the computational identification of temperature variable characteristics of heat transfer

16.00 – 16.20 Petr Vaněk
Multigrid with aggressive coarsening and polynomial smoothing 1

16.20 – 16.40 Roman Kužel
Multigrid with aggressive coarsening and polynomial smoothing 2

16.40 – 17.00 Pavla Fraňková
Multigrid with aggressive coarsening and polynomial smoothing 3

18.00 – 23.00 Conference Dinner, U Seminaristy Restaurant, Spálená St. 45

Friday, May 17

9.00 – 9.30 Shuhua Zhang
The valuation of weather derivatives and corresponding numerical methods

9.30 – 10.00 Karel Kozel
Numerical solution of 2D and 3D unsteady flow in a channel with low upstream velocity

10.00 – 10.30 Hehu Xie
The lower and upper bounds of eigenvalues by the finite element method

10.30 – 11.00 Coffee Break

11.00 – 11.30 Petr Sváček
On numerical modelling of gust response of an airfoil section

11.30 – 12.00 Vladimír Janovský
Path-following the static contact problem with Coulomb friction

12.00 – 14.00 Lunch Break

14.00 – 14.20 Torsten Linss
A posteriori error estimation for parabolic PDEs

14.20 – 14.40 Zbyněk Vastl
Multigrid with aggressive coarsening and polynomial smoothing 4

14.40 – 15.00 Hana Kopincová
Multigrid with aggressive coarsening and polynomial smoothing 5

15.00 – 15.20 Milan Hanuš
Multigrid with aggressive coarsening and polynomial smoothing 6

15.20 – 15.40 Coffee Break

15.40 – 16.00 Vratislava Mošová
Integral transformations – the base of recent technologies

16.00 – 16.20 Jaroslav Mlýnek
Parallel programming and optimization of heat radiation intensity

16.20 – 16.40 Jan Chleboun
A random set approach applied to definite integration of uncertain functions

16.40 – 17.00 Tomáš Vejchodský
Two-sided bounds of eigenvalues with applications to trace inequalities

17.00 – 17.20 Jakub Šístek
Parallel adaptive-multilevel BDDC method

Saturday, May 18

9.00 – 12.00 A walk through the Elephant Valley in the zoological garden