Technical University of Liberec
Academic Coordination Centre at the Euroregion Neisse (ACC)
Akademické koordinační středisko v Euroregionu Nisa
Akademisches Koordinierungszentrum in der Euroregion Neiße
Akademickie Centrum Koordynacyjne w Euroregionie Nysa
LIBEREC ZITTAU/GÖRLITZ JELENIA GÓRA

ACC JOURNAL
XVIII
4/2012

Issue D

Natural Sciences and Technology

TECHNICKÁ UNIVERZITA V LIBERCI

INTERNATIONALES HOCHSCHULINSTITUT ZITTAU

HOCHSCHULE ZITTAU/GÖRLITZ

UNIWERSYTET EKONOMICZNY WE WROCŁAWIU
WYDZIAŁ EKONOMII, ZARZĄDZANIA I TURYSTYKI W JELENIEJ GÓRZE

KARKONOSKA PAŃSTWOWA SZKOŁA WYŻSZA W JELENIEJ GÓRZE

TECHNICKÁ UNIVERZITA V LIBERCI
Studentská 1402/2, Liberec 1, 461 17, Česká republika
www.tul.cz

INTERNATIONALES HOCHSCHULINSTITUT ZITTAU
Markt 23, 02763 Zittau, Deutschland
www.ihi-zittau.de

HOCHSCHULE ZITTAU/GÖRLITZ
Theodor- Körner- Allee 16, 02763 Zittau, Deutschland
www.hszg.de

UNIWERSYTET EKONOMICZNY WE WROCŁAWIU
Wydział Ekonomii, Zarządzania i Turystyki w Jeleniej Górze
ul. Nowowiejska 3, 58-500 Jelenia Góra, Polska
www.ue.wroc.pl/grit

KARKONOSKA PAŃSTWOWA SZKOŁA WYŻSZA W JELENIEJ GÓRZE
ul. Lwówecka 18, 58-503 Jelenia Góra 5, Polska
www.kpswjg.pl

Technical University of Liberec
Academic Coordination Centre at the Euroregion Neisse (ACC)
Akademické koordinační středisko v Euroregionu Nisa
Akademisches Koordinierungszentrum in der Euroregion Neiße
Akademickie Centrum Koordynacyjne w Euroregionie Nysa
LIBEREC ZITTAU/GÖRLITZ JELENIA GÓRA

ACC JOURNAL
XVIII
4/2012

Issue D

Natural Sciences and Technology

TECHNICKÁ UNIVERZITA V LIBERCI

INTERNATIONALES HOCHSCHULINSTITUT ZITTAU

HOCHSCHULE ZITTAU/GÖRLITZ

UNIWERSYTET EKONOMICZNY WE WROCŁAWIU
WYDZIAŁ EKONOMII, ZARZĄDZANIA I TURYSTYKI W JELENIEJ GÓRZE
KARKONOSKA PAŃSTWOWA SZKOŁA WYŻSZA W JELENIEJ GÓRZE

ISSN 1803-9782

ACC JOURNAL je mezinárodní časopis, jehož vydavatelem je Technická univerzita v Liberci. Na jeho tvorbě se podílí šest vysokých škol sdružených v Akademickém koordinačním středisku v Euroregionu Nisa (ACC). Ročně vycházejí zpravidla dvě čísla.

ACC JOURNAL je periodikum publikující původní recenzované vědecké práce, vědecké studie, příspěvky ke konferencím a výzkumným projektům. První vydání obsahuje příspěvky z oblasti přírodních věd a techniky, druhé je zaměřeno na oblast společenských věd a ekonomiky.

ACC JOURNAL má charakter recenzovaného časopisu. Jeho edice navazuje na sborník „Vědecká pojednání“, který vycházel v letech 1995-2008. Od roku 2010 je ACC JOURNAL v databázi Rady pro vědu, výzkum, vývoj a inovace (Seznam recenzovaných neimpaktovaných časopisů vydávaných v České republice) hodnocených v RIVu.

ACC JOURNAL is an international journal. It is published at the Technical University of Liberec. Six high schools united in the Academic Coordination Centre in the Euroregion Nisa participate in its creation. There are usually two issues of the journal annually.

In the ACC JOURNAL original reviewed scientific studies, conference presentations and research project reports are published. The first issue focuses on natural sciences and technology; the second issue deals with social sciences and economics.

ACC JOURNAL is a reviewed one. It is building upon the tradition of the "Scientific Treaties" published between 1995 and 2008. The ACC JOURNAL has been in the database of the Research and Development Council since 2010 (List of reviewed non-impact journals published in the Czech Republic) recorded and evaluated in the Information Register of R&D results.

Acknowledgements:

This edition of ACC JOURNAL, Issue D, has been cofinanced from the European regional development fund by the Euroregion Neisse-Nisa-Nysa and by the project ESF No. CZ.1.07/2.3.00/09.0155 "Constitution and improvement of a team for demanding technical computations on parallel computers at TU Liberec".

This volume contains selected papers that have been presented at the International Conference Presentation of Mathematics 2012. The Conference was organized by the Department of Mathematics and Didactics of Mathematics under auspices of the dean of Faculty of Science, Humanities and Education of Technical University of Liberec doc. RNDr. Miroslav Brzezina, CSc., on Juni 21 and Juni 22, 2012.

The organizers are grateful to invited speakers, contributors, participants, and reviewers for their interest in the Conference. Also the support from the Ministry of Education Youth and Sports via the ESF Project Nr. CZ.1.07/2.3.00/09.0155 "Constitution and improvement of a team for demanding technical computations on parallel computers at TU Liberec" is gratefully acknowledged.

On behalf of the Organizing and Scientific Committee,

Václav Finěk

Contents

PARALLEL APPROACH TO THE SOLUTION OF STATIONARY REACTION-DIFFUSION PROBLEM

Daniela Bímová

Technical University of Liberec
Faculty of Science, Humanities and Education
Department of Mathematics and Didactics of Mathematics
Studentská 2, 461 17, Liberec, Czech Republic
daniela.bimova@tul.cz

Abstract

The paper is devoted to the parallel solution of the two-dimensional stationary reaction-diffusion problem. By the usage of parallel approach to the linear algebra representation we create the parallel algorithm for computing a numerical solution of the two-dimensional stationary reaction-diffusion problem. We compare calculation times of computing the approximate solution of the system of (linear) difference equations for different sizes of the system matrix by the numerical conjugate gradient method on 1, 2, 3, and 4 processors, respectively.

Keywords: Stationary reaction – diffusion problem, parallel linear algebra, finite difference method, conjugate gradient method.

Introduction

Our model problem is represented by the stationary reaction-diffusion equation, which discretization via the finite difference method leads to the system of linear algebraic equations.

We choose the numerical conjugate gradient method for solving the system of linear algebraic equations in this paper and we try to apply this method in parallel algorithm implemented in Fortran. The aim of this paper is to find how we save the calculating time computing the system of linear algebraic equations on 2, 3, or 4 processors instead of on one processor. We perform our calculation in numerical experiment in which we compute the vector of the approximate solution (for different mesh sizes of the grid) of the system of linear difference equations by the conjugate gradient method.

We have discussed the analogous problem using the method of steepest descent in [1].

1. Setting of Stationary Reaction-Diffusion Problem

We consider the two-dimensional boundary value problem in this chapter. It is the analogy of a one-dimensional boundary value problem mentioned in [2].
The task is to find a function u: $\Omega \to \mathbf{R}$ fulfilling the stationary reaction-diffusion equation

$$-\varepsilon\, \Delta u + k \cdot u = f \text{ in } \Omega = \langle a, b\rangle \times \langle c, d\rangle,\ a, b, c, d \in \mathbf{R},\ a < b,\ c < d,\ \varepsilon > 0,\ k \in \mathbf{R} \quad (1)$$

$$u\,|_{\delta\Omega} = g \text{ on } \delta\Omega, \quad (2)$$

where f: $\Omega \to \mathbf{R}$ is the given function and where g: $\delta\Omega \to \mathbf{R}$ represents a Dirichlet boundary condition.

For $m, n \in \mathbf{N}$ we bring up $(m+1)\times(n+1)$ of evenly spaced points $X_{i,j} = [x_i, y_j] = [a + i \cdot h, c + j \cdot h]$, where $i = 0, 1, \ldots, m$, $j = 0, 1, \ldots, n$ and where h is the spatial step. We denote $U_{i,j}$ the approximate solution at the points $X_{i,j}$, i. e. $U_{i,j} \approx u(x_i, y_j) \approx u(X_{i,j})$, and we put $F_{i,j} = f(x_i, y_j) = f(X_{i,j})$.

For every regular knot $X_{i,j}$ we use three-point stencil for approximation the second derivatives, i.e.

$$\frac{\delta^2 u}{\delta x^2}(x_i, y_j) \approx \frac{U_{i+1,j} - 2U_{i,j} + U_{i-1,j}}{h^2}, \quad \frac{\delta^2 u}{\delta y^2}(x_i, y_j) \approx \frac{U_{i,j+1} - 2U_{i,j} + U_{i,j-1}}{h^2}.$$

Applying this to (1) we obtain the difference equations for the regular knots $X_{i,j} = [x_i, y_j]$, it means we have the equations of the form

$$-U_{i-1,j} - U_{i,j-1} + (4 + k \cdot h^2) \cdot U_{i,j} - U_{i+1,j} - U_{i,j+1} = \frac{h^2}{\varepsilon} F_{i,j}, \qquad \begin{array}{l} i = 1, 2, \ldots, m-1, \\ j = 1, 2, \ldots, n-1, \\ k \in \mathbf{R}. \end{array}$$

This is the standard five-point scheme. The approximate values of the sought function *u* in knots of the given grid are represented by the numerical solution of the system of (linear) algebraic equations

$$\mathbf{AU} = \mathbf{F} \tag{3}$$

with an unknown vector

$$U = (U_{1,1}, \ldots, U_{1,n-1}, U_{2,1}, \ldots, U_{2,n-1}, \ldots, U_{m-1,1}, \ldots, U_{m-1,n-1})^{\mathrm{T}} \in \mathbf{R}^{(m-1)\times(n-1)}$$

and a matrix **A**, which we can write in the following way

$$\mathbf{A} = \begin{pmatrix} \mathbf{L} & -\mathbf{I} & \mathbf{0} & \cdots & \mathbf{0} \\ -\mathbf{I} & \mathbf{L} & \ddots & \ddots & \vdots \\ \mathbf{0} & \ddots & \ddots & \ddots & \mathbf{0} \\ \vdots & \ddots & \ddots & \mathbf{L} & -\mathbf{I} \\ \mathbf{0} & \cdots & \mathbf{0} & -\mathbf{I} & \mathbf{L} \end{pmatrix},$$

where

$$\mathbf{L} = \begin{pmatrix} 4 + k \cdot h^2 & -1 & 0 & \cdots & 0 \\ -1 & 4 + k \cdot h^2 & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & 4 + k \cdot h^2 & -1 \\ 0 & \cdots & 0 & -1 & 4 + k \cdot h^2 \end{pmatrix},$$

I is the identity matrix, and **0** is the zero matrix. As well as the vector of the right side of the mentioned system of (linear) difference equations we can write as follows

$$\mathbf{F} = \begin{pmatrix} \frac{h^2}{\varepsilon} F_{1,1} + U_{0,1} + U_{1,0} \\ \frac{h^2}{\varepsilon} F_{1,2} + U_{0,2} \\ \vdots \\ \frac{h^2}{\varepsilon} F_{1,n-1} + U_{0,n-1} + U_{1,n} \\ \frac{h^2}{\varepsilon} F_{2,1} + U_{2,0} \\ \frac{h^2}{\varepsilon} F_{2,2} \\ \vdots \\ \frac{h^2}{\varepsilon} F_{2,n-1} + U_{2,n} \\ \\ \vdots \\ \\ \frac{h^2}{\varepsilon} F_{m-1,1} + U_{m,1} + U_{m-1,0} \\ \frac{h^2}{\varepsilon} F_{m-1,2} + U_{m,2} \\ \vdots \\ \frac{h^2}{\varepsilon} F_{m-1,n-1} + U_{m,n-1} + U_{m-1,n} \end{pmatrix}.$$

According to (2) we consider the following Dirichlet boundary conditions:

$$U_{i,0} = g\left(X_{i,0}\right),\ U_{i,n} = g\left(X_{i,n}\right) \quad \forall\, i = 1, 2, \ldots, m-1,$$
$$U_{0,j} = g\left(X_{0,j}\right),\ U_{m,j} = g\left(X_{m,j}\right) \quad \forall\, j = 1, 2, \ldots, n-1.$$

2. Parallel approach to the solution of the system of linear algebraic equations

2.1. Matrix mapping in parallel linear algebra

The way, in which the matrix is mapped on the net of processors, determines efficiency and elegance of an algorithm in most the cases. We have chosen the striped mapping cyclically by rows of a matrix for our stationary reaction-diffusion problem.

2.2. The striped mapping of the matrix cyclically by rows

In this paragraph we describe how we map a matrix using MPI in parallel algorithms.

Let us assume that there is given the matrix **A** of the mentioned shape and of the type (*n*, *n*). In the parallel algorithm we work generally on *p* processors. The first task is to map the matrix **A** onto the particular processors. The matrix **A** will be mapped by the striped mapping cyclically by rows onto the particular processors. The striped mapping assumes that the processors are connected into the linear virtual array and that they are numbered 0, 1, 2, …, $p - 1$, where *p* is the number of used processors, in general. The matrix **A** (denoted A_global) is generated and known only by so called master processor. The master processor distributes data (the row vectors of the matrix **A)** into the particular processors. We obtain new matrices (denoted A_local) of the type ($\lceil n/p \rceil$, *n*), where *n* is the number of rows of original matrix and where *p* is the number of used processors, by the data distribution. Matrices "A_local" consist of the appropriate rows of the matrix **A**.

The analogous situation is true for mapping the vector **F** (of the right side of the system of linear algebraic equations). There is one small difference, only one element (not the whole row) is sent in every step of the distribution.

2.3. Product of a matrix and of a vector in parallel algorithm

We use the iterative methods for solving the system of linear algebraic equations. Part of every iterative method is the product of a matrix and of a vector. That is why we describe product of a matrix **A** and of a vector **U** from the matrix equation

$$\mathbf{AU} = \mathbf{F}$$

using MPI in parallel algorithms in this paragraph.

Every processor knows only a part of the vector **U**, with whom we want to multiply the matrix **A**, on the basis of the previous mapping of a vector onto the particular processors. (We assume that the given vector **U** is mapped in the same way in which the vector **F** was mapped). That is the reason why every processor has to find the rest part of the vector **U**, which it does not know. Every processor can call the command "MPI_ALLGATHER" by which it can find the missing information about the rest parts of the vector **U**.

If all the processors know all the elements of the vector **U**, we can perform dot product and calculate the appropriate element of the result vector.

2.4. Scalar product of two vectors in parallel algorithm

Except of the product of a matrix and of a vector, the scalar product of two vectors occurs in the iterative methods, too. In this paragraph we describe process of computing a scalar product of two vectors in parallel algorithms.

Every processor knows only a part of the vector **U** as well as of the vector **V** (which we want to multiply scalarly) on the basis of the previous mapping (the striped mapping cyclically by rows) of a vector onto the particular processors. That is why we firstly multiply scalarly the appropriate elements of the vectors **U**, **V**. Secondly, we call the command "MPI_ALLREDUCE" that sums over all processors and distributes the resulting sum onto all the processors. All the processors finally know value of the scalar product of the vectors **U**, **V**.

3. Numerical experiment of stationary reaction-diffusion problem

We consider the two-dimensional stationary reaction-diffusion problem

$$-\varepsilon\,\Delta u + k\cdot u = f \ \text{ in } \Omega = [0,1]^2, \tag{4}$$

$$u\,|_{\delta\Omega} = 0 \ \text{ on } \ \Omega. \tag{5}$$

We set $\varepsilon = 1, k = 1$ and cover the domain Ω by the grid of knots with the spatial step h in the direction of both the coordinate axes x and y. We choose function f in the problem (4) so that the function

$$u(x, y) = 64\cdot x\cdot y\cdot(1-x)\cdot(1-y)\cdot(y-x) \tag{6}$$

is the exact solution.

The approximate values of the solution u in all the inner regular knots of the given grid are the numerical solution of the problem (4) – (5). We find the numerical solution by the conjugate gradient method with prescribed tolerance 10^{-5} for the norm of residue.

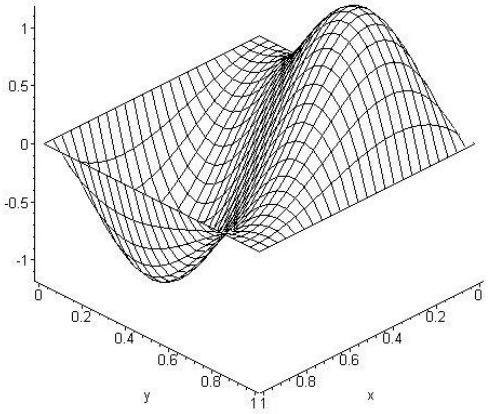
Table 1 illustrates the development of calculation times according to the number of used processors with respect to spatial step.

Tab. 1. *Calculation times needed for computing the vector* **U** *of the approximate solution for different mesh sizes of the grid on one, two, three, and four processors*

Step	Number of knots	Number of iterations	Condition number	1 processor	2 processors	3 processors	4 processors
h = 1/25	576	56	260	**1** s	**1** s	**1** s	**1** s
h = 1/49	2 304	110	963	**15** s	**10** s	**7** s	**5** s
h = 1/73	5 184	163	2 112	**2** min **35** s	**1** min **13** s	**49** s	**35** s
h = 1/97	9 216	217	3 704	**13** min **01** s	**5** min **50** s	**3** min **45** s	**2** min **57** s

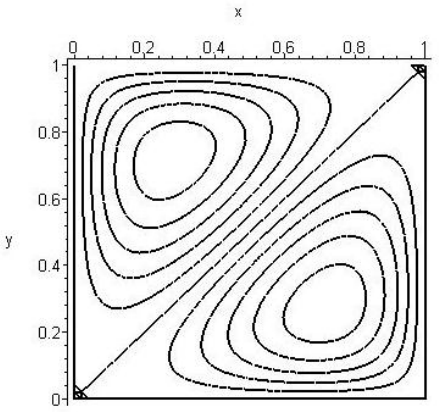
Source: Own based on computations on the cluster

There is shown the graph of the exact solution of the problem (4) – (5) in Fig. 1a and there are drawn the contours of the exact solution of the problem (4) – (5) in Fig. 1b.



Source: Own based on exact solution of (6)

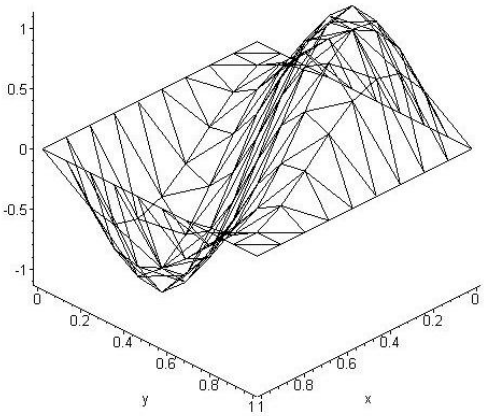
Fig. 1a. *Exact solution*



Source: Own based on exact solution of (6)

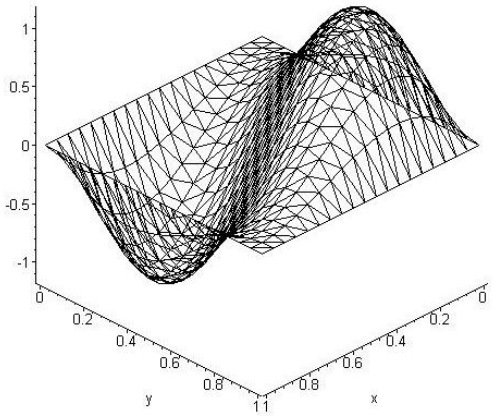
Fig. 1b. *Contours of the exact solution*

Furthermore, there are drawn the graphs of the approximate solutions of the problem (4) – (5) for the different mesh sizes in Fig 2a – 2d. The graph of the approximate solution for the mesh size $h = \frac{1}{9}$ is sketched in Fig. 2a, for the mesh size $h = \frac{1}{21}$ in Fig. 2b, for the mesh size $h = \frac{1}{41}$ in Fig. 2c, and finally for the mesh size $h = \frac{1}{61}$ in Fig. 2d.



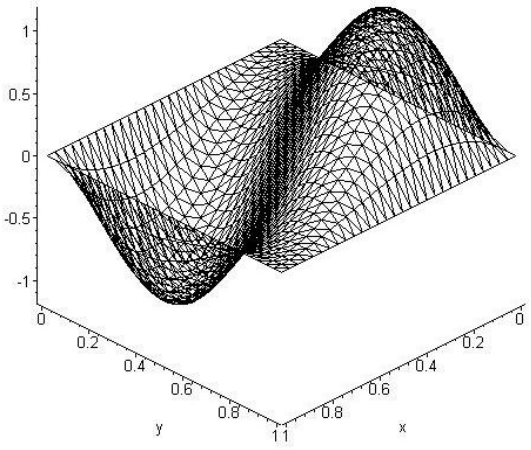
Source: Own based on numerical experiment

Fig. 2a. *Approximate solution, mesh size* $h = \frac{1}{9}$



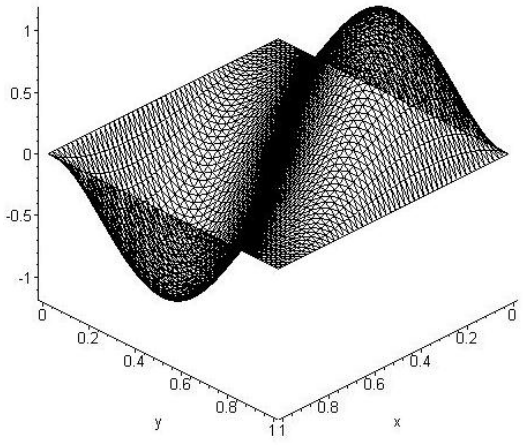
Source: Own based on numerical experiment

Fig. 2b. *Approximate solution, mesh size* $h = \frac{1}{21}$



Source: Own based on numerical experiment

Fig. 2c. *Approximate solution, mesh size* $h = \frac{1}{41}$



Source: Own based on numerical experiment

Fig. 2d. *Approximate solution, mesh size* $h = \frac{1}{61}$

Conclusion

We discussed the two-dimensional stationary reaction-diffusion problem. We described the striped mapping of a matrix cyclically by rows in parallel programming. We showed the possibilities of applying the basic principles of linear algebra – product of a matrix and of a vector, scalar product of two vectors – in parallel algorithm. Finally, we commented the numerical experiment in which the mentioned theory is applied in practice.

We can summarize on the basis of the numerical experiment results that computing the approximate solution of the system of linear difference equations is approximately two times (three times, resp. four times) quicker, if we calculate the system of equations on two (three, resp. four) processors instead of on one processor. Mentioned results are best seen from the calculation times for bigger mesh sizes of the grid. The results are written in the Tab. 1.

Acknowledgements

The paper was supported by the project ESF, no. CZ.1.07/2.3.00/09.0155, "Constitution and improvement of a team for demanding technical computations on parallel computers at TU Liberec".

Literature

[1] BÍMOVÁ, D.: *Parallel solution of Poisson Equation.* In: Proceedings of Aplimat 2012. Bratislava, Slovak University of Technology 2012, pp. 233-243. ISBN 978-80-89313-57-0.

[2] DOLEJŠÍ, V.; KNOBLOCH, P.; KUČERA, V.; VLASÁK, M.: *Finite element methods: theory, applications and implementation*. TUL, Liberec 2011. ISBN 978-80-7372-728-4.

Mgr. Daniela Bímová, Ph.D.

PARALELNÍ PŘÍSTUP K ŘEŠENÍ STACIONÁRNÍHO REAKČNĚ-DIFÚZNÍHO PROBLÉMU

Článek je věnován paralelnímu řešení dvoudimenzionálního stacionárního reakčně-difúzního problému. Pomocí paralelního přístupu k reprezentaci lineární algebry vytvoříme paralelní algoritmus pro výpočet numerického řešení dvoudimenzionálního stacionárního reakčně-difúzního problému. Porovnáme časy potřebné k výpočtu přibližného řešení systému (lineárních) diferenciálních rovnic pro různě velkou matici soustavy numerickou metodou sdružených gradientů na 1, 2, 3 a 4 procesorech.

PARALLERER ANSATZ ZUR LÖSUNG EINES STATIONÄREN REAKTIONSDIFFUSEN PROBLEMS

Dieser Artikel beschäftigt sich mit der parallelen Lösung eines zweidimensionalen stationären reaktionsdiffusen Problems. Mit Hilfe eines parallelen Ansatzes zur Repräsentation linearer Algebra bilden wir einen parallelen Algorithmus für die Berechnung einer numerischen Lösung eines zweidimensionalen stationären reaktionsdiffusen Problems. Wir vergleichen die Zeiten, die zur Berechnung einer annähernden Lösung des Systems (linearer) differenzialer Gleichungen einer großen Matrix eines Systems durch die numerische Methode dualer Gradienten auf 1, 2, 3 und 4 Prozessoren nötig sind.

PARALELNE PODEJŚCIE DO ROZWIĄZYWANIA STACJONARNEGO PROBLEMU REAKCYJNO-DYFUZYJNEGO

Artykuł poświęcony jest paralelnemu rozwiązywaniu dwuwymiarowego stacjonarnego problemu reakcyjno-dyfuzyjnego. Przy pomocy podejścia paralelnego do reprezentacji algebry liniowej stworzymy algorytm równoległy do wyliczenia numerycznego rozwiązania dwuwymiarowego stacjonarnego problemu reakcyjno-dyfuzyjnego. Porównamy czasy niezbędne do wyliczenia przybliżonego rozwiązania układu równań różniczkowych (liniowych) dla macierzy układu o różnej wielkości przy pomocy metody numerycznej gradientów sprzężonych na 1, 2, 3 i 4 procesorach.

ALGORITHM FOR ELIMINATION OF SYSTEMS WITH SPARSE ASYMMETRIC REDUCIBLE MATRICES

Daniela Bittnerová

Faculty of Science, Humanities and Education
Department of Mathematics and Didactics of Mathematics
Studentská 2, 461 17, Liberec, Czech Republic
daniela.bittnerova@tul.cz

Abstract

In the paper, an algorithm for finding an optimal or almost optimal permutation for an ordering of elements of a matrix, which is sparse, asymmetric and reducible, is suggested. Using this algorithm we can solve large sparse systems of linear equations more efficiently. The algorithm is a modification of the algorithm presented in [6], therefore the same indications and symbols are use.

Keywords: Asymmetric reducible sparse matrix, algorithm.

Introduction

A matrix can be considered as sparse if the number of its zero elements is much greater than the number of nonzero ones. Sparse matrices occur in many fields of applied mathematics, for example at partial differential equations solving, but also in many other fields of science, like structural analysis, management analysis, power system analysis, surveying etc. Solving a practical problem we obtain large systems of linear equations with sparse matrices. Hand in hand with a development of a parallel programming for large quantity data, the importance of an optimization of sparse matrices increases. There exist lots of ways how to do it. Some of them go out from graph structures of associated matrices. In the paper, one such algorithm is suggested. It is a modification of the algorithm presented in [6], where symmetric sparse irreducible matrices were discussed, for generally Boolean asymmetric reducible sparse matrices. Using such algorithm we are able to solve large sparse systems of linear equations by some direct method. Indications and symbols are kept from the paper [6].

1. Basic Theory and Assumptions

Let $\mathbf{A} = (a_{i,j})$ be a sparse matrix of order n, $\mathbf{V} = (v_{i,j})$, $\mathbf{W} = (w_{i,j})$ be matrices of order n, where

$$v_{i,j} \in \{0;1\}, \quad \forall i, j = 1,\cdots n \;\; v_{i,j} = 1 \Leftrightarrow a_{i,j} \neq 0$$

$$w_{i,j} \in \{0;1\}, \quad \forall i, j = 1,\cdots n \;\; w_{i,j} = 1 \Leftrightarrow (a_{i,j} \neq 0 \;\vee\; a_{j,i} \neq 0).$$

The matrix **A** is called **Boolean symmetric** if **V** = **W**. In the opposite case, it is said **Boolean asymmetric**.

The graph { **G**(**A**) = (**X**, **E**) / **G**$^{\mathbf{0}}$(**A**) = (**X**, **E**$^{\mathbf{0}}$) } is called { the **non-oriented** / **oriented** } graph associated to the matrix **A**,

if $\mathbf{X} = \{x_1, \cdots, x_n\}$ is a set of nodes of the graph $\begin{bmatrix} \mathbf{G}(\mathbf{A}) \\ \mathbf{G^0}(\mathbf{A}) \end{bmatrix}$ such that the node x_i corresponds to the row *i* of **A**, and the set $\begin{bmatrix} \mathbf{E} \text{ is a set of all} \\ \mathbf{E^0} \end{bmatrix} \begin{bmatrix} \text{non-oriented edges of the graph} \\ \text{oriented} \end{bmatrix} \begin{bmatrix} \mathbf{G}(\mathbf{A}), \\ \mathbf{G^0}(\mathbf{A}) \end{bmatrix}$

i.e. $\begin{bmatrix} \mathbf{E} = \{\{x_i; x_j\}; x_i \in \mathbf{X} \wedge x_j \in \mathbf{X} \wedge w_{i,j} = 1, i < j\}. \\ \mathbf{E^0} = \{[x_i; x_j]; x_i \in \mathbf{X} \wedge x_j \in \mathbf{X} \wedge v_{i,j} = 1, i \neq j\} \end{bmatrix}$

If **G** is a connected graph associated to the matrix **A**, it implies the matrix **A** is irreducible. Let us denote:

$$\mathbf{N}(x) = \{y \in \mathbf{X}; \{x; y\} \in \mathbf{E}\} \cup \{x\}$$
$$\mathbf{N_1}(x) = \{y \in \mathbf{X}; [x; y] \in \mathbf{E^0}\} \cup \{x\}$$
$$\mathbf{N_2}(x) = \{y \in \mathbf{X}; [y; x] \in \mathbf{E^0}\} \cup \{x\}$$
$$\mathbf{D}(x) = \{\{y; z\} \notin \mathbf{E}; y \in \mathbf{N}(x) \wedge z \in \mathbf{N}(x) \wedge y \neq z\}$$
$$\mathbf{D_1}(x) = \{[y; z] \notin \mathbf{E^0}; y \in \mathbf{N_1}(x) \wedge z \in \mathbf{N_1}(x) \wedge y \neq z\}$$
$$\mathbf{D_2}(x) = \{[y; z] \notin \mathbf{E^0}; y \in \mathbf{N_2}(x) \wedge z \in \mathbf{N_2}(x) \wedge y \neq z\}$$

Evidently,

N is the set of all successors (neighbours) of the node *x* including *x* in the graph **G**,

$\mathbf{N_1}(x)$ is the set of all $\begin{bmatrix} \text{heads (ended nodes) of all oriented edges} \\ \text{tails (starting nodes)} \end{bmatrix} \begin{bmatrix} \text{from } x \text{ in } \mathbf{G^0} \text{ including } x. \\ \text{to} \end{bmatrix}$
$\mathbf{N_2}(x)$

The sets $\mathbf{D}(x)$, $\mathbf{D_1}(x)$, $\mathbf{D_2}(x)$ describe relations of successors in the graph $\mathbf{G^0}$ each other, in relation to the fill in of matrices **L** and **U** in the **LU**-decomposition of the matrix **A (**and/or to the fill in of matrices **L**, **D** in the $\mathbf{LDL^T}$-decomposition). The numbers of elements of the sets $\mathbf{N}(x_i)$, $\mathbf{N_1}(x_i)$, $\mathbf{N_2}(x_i)$, $\mathbf{D}(x_i)$, $\mathbf{D_1}(x_i)$, $\mathbf{D_2}(x_i)$ are denoted by the symbols $n_i(\mathbf{G})$, $n_i^1(\mathbf{G^0})$, $n_i^2(\mathbf{G^0})$, $d_i(\mathbf{G})$, $d_i^1(\mathbf{G^0})$, $d_i^2(\mathbf{G^0})$, respectively. The number $d_i(\mathbf{G})$ (called "fill in") determines a number of new edges after elimination of the node x_i.

The graph $\begin{bmatrix} \mathbf{G}_{x_i} = (\mathbf{X} - \{x_i\},\ \mathbf{E}(\mathbf{X} - \{x_i\}) \cup \mathbf{D}(x_i)) \\ \mathbf{G}^0_{x_i} = (\mathbf{X} - \{x_i\},\ \mathbf{E^0}(\mathbf{X} - \{x_i\}) \cup \mathbf{D}_1(x_i) \cup \mathbf{D}_2(x_i)) \end{bmatrix}$ is called the graph

produced from the graph $\begin{bmatrix} \mathbf{G} \text{ by an elimination of the node } x_i. \\ \mathbf{G^0} \end{bmatrix}$

With respect to the found algorithm, let us denote $\mathbf{G_0} = \mathbf{G}$, $\mathbf{G_0^0} = \mathbf{G^0}$,

$\begin{bmatrix} \mathbf{G}_k \\ \mathbf{G}^0_k \end{bmatrix}$ be a graph produced from $\begin{bmatrix} \mathbf{G}_{k-1} \\ \mathbf{G}^0_{k-1} \end{bmatrix}$ by an elimination of the node x_i, $k = 1, 2, \cdots, n-1$.

2. Algorithms

Our goal is to find an optimal, respective almost optimal, permutation to change the given ordering of rows and columns of the matrix **A** so that the elimination of a system with the matrix **A** is as efficient as possible. In [6], two algorithms are presented. They are defined for irreducible sparse symmetric systems. The following algorithms are determined for reducible sparse generally asymmetric systems. The Algorithm 1 deals with matrices as if symmetric, Algorithm 2 deals with asymmetric matrices.

Algorithm 1

Step 0:

Put $\mathbf{S^0} = \varnothing$, $\mathbf{G_0} = \mathbf{G}$.

Step $k = 1, \cdots, n-1$:

1. Put $\mathbf{R^k} = \mathbf{S^{k-1}}$ for $\mathbf{S^{k-1}} \neq \varnothing$ else $\mathbf{R^k} = \left\{ x_j ; x_j \in \mathbf{G_{k-1}} \right\}$.
2. $\mathbf{S_1^k} = \left\{ x_j \in \mathbf{R^k} ; d_j\left(\mathbf{G_{k-1}}\right) = \min\limits_{x_q \in \mathbf{R^k}} d_q\left(\mathbf{G_{k-1}}\right) \right\}$.
3. $i_k = \min\limits_{x_j \in \mathbf{S_1^k}} \{ j \}$.
4. Eliminate the node x_{i_k} from the graph $\mathbf{G_{k-1}}$ (i.e. the index i_k constructed in the permutation *p* is equal to the integer *k*).
5. $\mathbf{S^k} = \left\{ x_j \in \mathbf{N}(x_{i_k}) \cap \mathbf{G_k} \right\}$

Algorithm 2

Step 0:

Put $\mathbf{S^0} = \varnothing$, $\mathbf{G_0^0} = \mathbf{G^0}$.

Step $k = 1, \cdots, n-1$:

1. Put $\mathbf{R^k} = \mathbf{S^{k-1}}$ for $\mathbf{S^{k-1}} \neq \varnothing$ else $\mathbf{R^k} = \left\{ x_j ; x_j \in \mathbf{G_{k-1}^0} \right\}$.
2. $\mathbf{S_1^k} = \left\{ x_j \in \mathbf{R^k} ; d_j\left(\mathbf{G_{k-1}^0}\right) = \min\limits_{x_q \in \mathbf{R^k}} d_q\left(\mathbf{G_{k-1}^0}\right) \right\}$.
3. $i_k = \min\limits_{x_j \in \mathbf{S_1^k}} \{ j \}$.
4. Eliminate the node x_{i_k} from the graph $\mathbf{G_{k-1}^0}$ (i.e. the index i_k constructed in the permutation *p* is equal to the integer *k*).
5. $\mathbf{S^k} = \left\{ x_j \in \left(\mathbf{N_1}(x_{i_k}) \cup \mathbf{N_2}(x_{i_k}) \right) \cap \mathbf{G_k^0} \right\}$

Sometimes, Boolean symmetric matrices of linear systems contain a row or column, of which the number of non-zero elements approximates to *n*.

Then it is suitable to define a maximal permissible number of elements $d_{\max}$ of the set $\mathbf{D}(x_i)$, and transport nodes x_i, for them $d_i(\mathbf{G}) > d_{\max}$, resp. $d_i(\mathbf{G^0}) > d_{\max}$, in the desired permutation to the last positions. We will eliminate these nodes from the graph **G**, resp. $\mathbf{G^0}$, and Algorithm 1, resp. Algorithm 2, deals with remaining nodes.

Example 1:

Let

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 \\ 2 & 1 & 0 & 0 & 2 \\ 3 & 0 & -1 & 0 & 1 \\ 4 & 0 & 0 & 2 & 0 \\ 5 & 2 & 1 & 0 & 3 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 1 & -1 & 2 & 0 & 3 \\ -4 & 1 & 0 & 0 & 0 \\ 0 & 0 & 3 & 0 & 4 \\ 2 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 4 \end{bmatrix}$$

be matrices. Applying the Algorithm 1 above, we have got the permutation matrix **P** = (2, 5, 3, 1, 4) and then

$$\mathbf{P} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}, \quad \mathbf{P^T} = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 \end{bmatrix},$$

$$\mathbf{PAP^T} = \begin{bmatrix} 1 & 2 & 0 & 2 & 0 \\ 2 & 3 & 1 & 5 & 0 \\ 0 & 1 & -1 & 3 & 0 \\ 2 & 5 & 3 & 1 & 4 \\ 0 & 0 & 0 & 4 & 2 \end{bmatrix}, \quad \mathbf{PBP^T} = \begin{bmatrix} 1 & 0 & 0 & -4 & 0 \\ 1 & 4 & 0 & 0 & 0 \\ 0 & 4 & 3 & 0 & 0 \\ -1 & 3 & 2 & 1 & 0 \\ 0 & 0 & 0 & 2 & 1 \end{bmatrix}.$$

The products **PA** and **PB** exchange rows of the matrix **A** and **B**, $\mathbf{AP^T}$ and $\mathbf{BP^T}$ exchange columns of **A** and **B**. Algorithms 1 manipulate with matrices as they are Boolean symmetric. It removes (almost all) non-zero elements of diagonals and produces (almost all) non-zero vectors perpendicular to it. Let us look at the structures of the matrices **A** and **B (**non-zero elements are indicated by the symbol **x**, the symbol • denotes non-zero elements which we must fill in **B** to obtain the symmetric matrix **B**):

$$\mathbf{A} = \begin{bmatrix} \mathbf{x} & \mathbf{x} & \mathbf{x} & \mathbf{x} & \mathbf{x} \\ \mathbf{x} & \mathbf{x} & & & \mathbf{x} \\ \mathbf{x} & & \mathbf{x} & & \mathbf{x} \\ \mathbf{x} & & & \mathbf{x} & \\ \mathbf{x} & \mathbf{x} & \mathbf{x} & & \mathbf{x} \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} \mathbf{x} & \mathbf{x} & \mathbf{x} & \bullet & \mathbf{x} \\ \mathbf{x} & \mathbf{x} & & & \bullet \\ \bullet & & \mathbf{x} & & \mathbf{x} \\ \mathbf{x} & & & \mathbf{x} & \\ \bullet & \mathbf{x} & \bullet & & \mathbf{x} \end{bmatrix}$$

Then the matrix **B** has the same structure as **A**.

Using the algorithms 1, we have got

$$\mathbf{PAP}^{\mathbf{T}} = \begin{bmatrix} \mathbf{x} & \mathbf{x} & & \mathbf{x} & \\ \mathbf{x} & \mathbf{x} & \mathbf{x} & \mathbf{x} & \\ & \mathbf{x} & \mathbf{x} & \mathbf{x} & \\ \mathbf{x} & \mathbf{x} & \mathbf{x} & \mathbf{x} & \mathbf{x} \\ & & & \mathbf{x} & \mathbf{x} \end{bmatrix}, \qquad \mathbf{PBP}^{\mathbf{T}} = \begin{bmatrix} \mathbf{x} & \bullet & & \mathbf{x} & \\ \mathbf{x} & \mathbf{x} & \bullet & \bullet & \\ & \mathbf{x} & \mathbf{x} & \bullet & \\ \mathbf{x} & \mathbf{x} & \mathbf{x} & \mathbf{x} & \bullet \\ & & & \mathbf{x} & \mathbf{x} \end{bmatrix}.$$

Example 2:

In the Table 1, we see how the almost optimal permutation is changing depending on the choice of the number $d_{\max}$.

Tab. 1. *The dependance of the permutation on* $d_{\max}$

Matrices A, B		$d_{\max} = 0$			$d_{\max} = 1$			$d_{\max} = 2$		
i		D(i)	inv_i	per_i	D(i)	inv_i	per_i	D(i)	inv_i	per_i
1	1,2,3,4,5	[2,3]	**5**	2	[2,3], [2,3]	**5**	2	[2,3], [2,4], [3,4], [4,5]	**5**	2
2	1,2,5	∅	1	3	∅	1	5	∅	1	5
3	1,3,5	∅	2	4	∅	3	3	∅	3	3
4	1,4	∅	3	5	∅	4	4	∅	4	4
5	1,2,3,5	[2,3]	**4**	1	[2,3]	2	1	[2,3]	2	1

Source: Own based on computations

The highlighted node i is the node, for that the number of elements of the set D(i) is greater than the chosen number $d_{\max}$ and that is transport to last free positions in the permutation **P**.

If $d_{\max} = 2$ then the permutation matrix **P** is of the form $\mathbf{P} = (2,5,3,4,1)$ and

$$\mathbf{PAP}^{\mathbf{T}} = \begin{bmatrix} 1 & 2 & 0 & 0 & 2 \\ 2 & 3 & 1 & 0 & 5 \\ 0 & 1 & -1 & 0 & 3 \\ 0 & 0 & 0 & 2 & 4 \\ 2 & 5 & 3 & 4 & 1 \end{bmatrix}, \quad \mathbf{PBP}^{\mathbf{T}} = \begin{bmatrix} 1 & 0 & 0 & 0 & -4 \\ 1 & 4 & 0 & 0 & 0 \\ 0 & 4 & 3 & 0 & 0 \\ 0 & 0 & 0 & 1 & 2 \\ -1 & 3 & 2 & 0 & 1 \end{bmatrix}.$$

The structures of non-zero elements are:

$$\mathbf{PAP}^{\mathbf{T}} = \begin{bmatrix} \mathbf{x} & \mathbf{x} & & & \mathbf{x} \\ \mathbf{x} & \mathbf{x} & \mathbf{x} & & \mathbf{x} \\ & \mathbf{x} & \mathbf{x} & & \mathbf{x} \\ & & & \mathbf{x} & \mathbf{x} \\ \mathbf{x} & \mathbf{x} & \mathbf{x} & \mathbf{x} & \mathbf{x} \end{bmatrix}, \qquad \mathbf{PBP}^{\mathbf{T}} = \begin{bmatrix} \mathbf{x} & \bullet & & \mathbf{x} & \mathbf{x} \\ \mathbf{x} & \mathbf{x} & \bullet & & \bullet \\ & \mathbf{x} & \mathbf{x} & & \bullet \\ & & & \mathbf{x} & \mathbf{x} \\ \mathbf{x} & \mathbf{x} & \mathbf{x} & \bullet & \mathbf{x} \end{bmatrix}.$$

The advantage of these algorithms is a less number of computing operations for a solving of systems of linear equations since we calculate with almost all non-zero elements of given matrices of systems.

Conclusion

In a science papers and books, there exist lots of algorithms and methods, how to solve large linear systems with a sparse matrix. Principles of computations are different. Some of them use graphic structures of matrices. The algorithm above is one of them. The other ways could be found, for example in [2] -[6].

Acknowledgements

The paper was supported by the project ESF, no. CZ.1.07/2.3.00/09.0155, „Constitution and Improvement of a Team for Demanding Technical Computations on Parallel Computers at TU Liberec.“

Literature

[1] Duff, I. S. – Erisman, A. M. – Reid, J. K.: *Direct Methods for Sparse Matrices*. Oxford, Clarendon Press 1990, pp. 341. ISBN 0-19-853421-3

[2] Ersavas, B. F.: *Sparse Matrix Ordering and Gaussian Elimination.* In: ECE 3652 Fundamentals of Computer Engineering 2002, pp. 1-12.

[3] Gilbert, A. - Li, Y. - Porat, E. - M. Strauss, M.: *Approximate Sparse Recovery: Optimizing time and measurements*. Manuscript, 2009.

[4] Gilbert, J. R. – Ng, E. G.: *Predicting Structure in Nonsymetric Sparse Matrix Factorizations*. New York 1993.

[5] Kapre, N. – DeHon, A.: *Parallelizing Sparse Matrix Solve for Spice Circuit Simulation Using FPGAS*. In: Proceedings of IEEE International Conference on Field-Programmable Technology (FPT 2009), 2009, pp. 1-9.

[6] Segethová, J.: *Elimination on Sparse Symmetric Systems of a Special Structure.* Aplikace matematiky 17, 6, 1972, pp. 447–460.

RNDr. Daniela Bittnerová, CSc.

ALGORITMUS PRO ELIMINACI SOUSTAV S ŘÍDKÝMI NESYMETRICKÝMI ROZLOŽITELNÝMI MATICEMI

V článku je uveden algoritmus určený k nalezení optimálního či skoro optimálního uspořádání prvků matice, která je řídká, nesymetrická a rozložitelná (reducibilní). Pomocí tohoto algoritmu můžeme efektivněji řešit velké řídké soustavy lineárních rovnic. Uvedený algoritmus je modifikací známého algoritmu pro symetrické nerozložitelné matice.

EIN ALGORITHMUS FÜR DIE ELIMINIERUNG DER SYSTEME MIT SELTENEN ASYMMETRISCHEN ZERLEGBAREN MATRIXEN

Im Artikel wird ein Algorithmus vorgestellt, der zu einer optimalen oder fast optimalen Anordnung der Matrixelemente führen soll. Diese Matrix ist selten, asymmetrisch und zerlegbar (reduzierbar). Mit Hilfe dieses Algorithmus können wir große seltene Systeme linearer Gleichungen effektiver lösen. Der angeführte Algorithmus ist eine Modifikation eines bekannten Algorithmus für eine symmetrische, nicht zerlegbare Matrix.

ALGORYTM DO ELIMINACJI UKŁADÓW Z RZADKIMI NIESYMETRYCZNYMI MACIERZAMI ROZKŁADALNYMI

W artykule przedstawiono algorytm przeznaczony do znalezienia optymalnego lub prawie optymalnego układu elementów macierzy, która jest rzadka, niesymetryczna i rozkładalna. Przy pomocy tego algorytmu można bardziej efektywnie rozwiązywać duże rzadkie układy równań liniowych. Podany algorytm stanowi modyfikację znanego algorytmu dla symetrycznych macierzy nierozkładalnych.

ITERATION-DISCRETIZATION METHODS FOR SOME VARIATIONAL INEQUALITY

Andrzej Cegielski
***Christian Grossmann**

University of Zielona Góra
Faculty of Mathematics
ul. Szafrana 4a, Pl-65-514 Zielona Góra, Poland
a.cegielski@wmie.uz.zgora.pl

*Technical University Dresden
Institute of Numerical Mathematics
D-01062 Dresden, Germany
christian.grossmann@tu-dresden.de

Abstract

Iterative methods for solving variational inequalities in infinite dimensional Hilbert spaces as a rule require some discretization. This leads to variational inequalities over families of spaces. In the present paper this problem is addressed by an iterative method with only a finite number of steps at each discretization level. First, abstract methods are studied and later an optimal control problem with elliptic state equations and some bound on the controls is considered. The discretization technique rests upon a nested family of piecewise linear C^0-elements conforming finite element discretizations.

Keywords: Optimal control; iteration-discretization; projection algorithm; elliptic state equation; finite element discretization.

Introduction

The solution of variational inequalities in function spaces often requires discretizations as well as iteration methods for solving the obtained finite dimensional problems. Similarly, the practical application of iteration methods in function spaces as a rule needs some finite dimensional approximation of the iteration procedure. Both processes, i.e. discretization and iteration, are not finite. The aim of our paper is to provide a sketch of the ideas developed in [3] for a new iteration-discretization method for variational inequality problems over the fixed point set of a quasi-nonexpansive operator which bases on appropriate extensions to families of problems and mappings. We give partially an overview over the problem setting and illustrate underlying analytical results. The related proofs can be found in [3].

1 Variational Inequalities over Sets of Fixed Points

Let $\mathscr{H}$ denote a real Hilbert space with the inner product $\langle \cdot, \cdot \rangle$ and the related norm $\| \cdot \|$ and $T : \mathscr{H} \to \mathscr{H}$ be a quasi-nonexpansive operator, i.e. an operator with $\operatorname{Fix} T \neq \emptyset$ and

$$\|Tu - z\| \leq \|u - z\| \qquad \text{for all } u \in \mathscr{H}, \quad \text{for all } z \in \operatorname{Fix} T,$$

where

$$\operatorname{Fix} T := \{ v \in \mathscr{H} : Tv = v \}.$$

Further, let a mapping $\mathscr{F} : \mathscr{H} \to \mathscr{H}$ be given which is κ-Lipschitz continuous and η-strongly monotone over $\mathscr{H}$, i.e. with some $\kappa \geq \eta > 0$ holds

$$\|\mathscr{F}(u) - \mathscr{F}(v)\| \leq \kappa \|u - v\| \qquad \text{for all } u, v \in \mathscr{H} \tag{1}$$

and

$$\langle \mathscr{F}(u) - \mathscr{F}(v), u - v \rangle \geq \eta \|u - v\|^2 \qquad \text{for all } u, v \in \mathscr{H}. \tag{2}$$

We consider the problem, called $VIP(\mathscr{F}, \operatorname{Fix} T)$,

Find $\bar{u} \in \operatorname{Fix} T$ *with*

$$\langle \mathscr{F}(\bar{u}), u - \bar{u} \rangle \geq 0 \text{ for all } u \in \operatorname{Fix} T. \tag{3}$$

Since $\operatorname{Fix} T$ is non-empty, closed and convex there exists a unique solution $\bar{u}$ of the problem $VIP(\mathscr{F}, \operatorname{Fix} T)$. Problems of this type are considered in e.g. [1], [5], [8] and the book [2] provides a comprehensive discussion of them.

Having infinite-dimensional Hilbert spaces $\mathscr{H}$ in mind, the numerical treatment of $VIP(\mathscr{F}, \operatorname{Fix} T)$ requires an appropriate discretization. Let $\mathscr{H}_k \subset \mathscr{H}$ denote a family of nested closed subspaces and $T_k : \mathscr{H} \to \mathscr{H}_k$ a family of quasi-nonexpansive operators that satisfy $\operatorname{Fix} T_k \subset \operatorname{Fix} T$ and $\bigcap_{k=0}^{\infty} \operatorname{Fix} T_k \neq \emptyset$. Further, let $\{\mathscr{F}_k\}_{k=0}^{\infty}$ with $\mathscr{F}_k : \mathscr{H} \to \mathscr{H}_k$ be a sequence of operators which are κ-Lipschitz continuous and η-strongly monotone over $\mathscr{H}_k$. Now, instead of $VIP(\mathscr{F}, \operatorname{Fix} T)$ we will consider the following sequence of problems $VIP(\mathscr{F}_k, \operatorname{Fix} T_k)$

Find $\bar{u}^k \in \operatorname{Fix} T_k$ *with*

$$\langle \mathscr{F}_k(\bar{u}^k), u - \bar{u}^k \rangle \geq 0 \text{ for all } u \in \operatorname{Fix} T_k. \tag{4}$$

The corresponding conditions for the approximation will be presented in the further part of the paper (see Theorem 1). As for $VIP(\mathscr{F}, \operatorname{Fix} T)$ the made assumption guarantees that each of the problems $VIP(\mathscr{F}_k, \operatorname{Fix} T_k)$ possess a unique solution $\bar{u}^k$. However, we will not solve each problem $VIP(\mathscr{F}_k, \operatorname{Fix} T_k)$, but propose and analyze an iteration-discretization procedure that approximate the solution of $VIP(\mathscr{F}, \operatorname{Fix} T)$ by simultaneously performing iteration steps and refining the discretization.

2 Iteration Methods Based upon Families of Operators

With $\mu \in (0, \frac{2\eta}{\kappa^2})$ and $\{\lambda_k\}_{k=0}^{\infty} \subset (0, \mu]$ we consider the following iteration for solving $VIP(\mathscr{F}, \operatorname{Fix} T)$:

$$u^{k+1} = (I - \lambda_k \mathscr{F}_k) T_k u^k, \tag{5}$$

where I denotes the identity operator and $u^0 \in \mathscr{H}$ is arbitrarily chosen. An alternative recursion is defined by

$$u^{k+1} = T_k (I - \lambda_k \mathscr{F}_k) u^k, \tag{6}$$

which we study later under the additional assumption that the operators T_k are projections on closed convex sets. Both iterations can be seen equivalent after transformations. Despite the fact that the sequences generated either by (5) or by (6) do not coincide we use the same notation since in the sequel we clearly distinguish which of the iterations is applied. First, we analyze the iteration (5). Denote

$$S_k = (I - \lambda_k \mathscr{F}_k) T_k. \tag{7}$$

Iteration (5) can be shortly written in the form

$$u^{k+1} = S_k u^k.$$

Let $\mu \in \left(0, \frac{2\eta}{\kappa^2}\right)$, define

$$\tau := 1 - \sqrt{1 + \mu^2\kappa^2 - 2\mu\,\eta}. \tag{8}$$

Since $\eta \leq \kappa$, consequently, τ is well defined and we have $\tau \in (0,1]$.

Lemma 1 *Let $\mu \in \left(0, \frac{2\eta}{\kappa^2}\right)$. Then the operators $G_k := I - \mu\mathscr{F}_k$ satisfy*

$$\|G_k x - G_k y\| \leq (1-\tau)\|x-y\| \qquad \text{for all} \quad x, y \in \mathscr{H}_k.$$

From the lemma above with (7) we obtain

Corollary 1 *Let $\mu \in \left(0, \frac{2\eta}{\kappa^2}\right)$ and $\lambda_k \in [0,\mu]$. Then*

$$\|S_k u - S_k v\| \leq \left(1 - \frac{\lambda_k \tau}{\mu}\right)\|T_k u - T_k v\| \qquad \text{for all} \quad u, v \in \mathscr{H}_k.$$

For the further analysis we make the assumption:

$$\text{There exist } z \in \bigcap_{k=0}^{\infty} \operatorname{Fix} T_k \text{ and some } c \in \mathbb{R} \text{ such that } \|\mathscr{F}_k z\| \leq c \tag{9}$$

for all $k \geq 0$. The κ-Lipschitz continuity of $\mathscr{F}_k$ for any $r > 0$ yields the boundedness of $\|\mathscr{F}_k u\|$ for any $u \in B(z,r) := \{v \in \mathscr{H}_k : \|v - z\| < r\}$.

Lemma 2 *Let $u^0 \in \mathscr{H}$ be arbitrary and $\{u^k\}_{k=0}^{\infty}$ be generated by (5). Then the sequences $\{u^k\}_{k=0}^{\infty}$, $\{\mathscr{F}_k T_k u^k\}_{k=0}^{\infty}$ and $\{T_k u^k - P_{\operatorname{Fix} T}\, u^k\}_{k=0}^{\infty}$ are bounded, where $P_{\operatorname{Fix} T}$ denotes the metric projection onto* $\operatorname{Fix} T$.

Define

$$\alpha_k := \frac{\lambda_k \tau}{\mu} \tag{10}$$

and

$$\beta_k := \alpha_k \frac{\mu^2}{\tau^2}\left(\|\mathscr{F}_k \vec{u}^k\|^2 + 2\langle \mathscr{F}_k T_k u^k - \mathscr{F}_k \vec{u}^k, \mathscr{F}_k \vec{u}^k\rangle\right) \tag{11}$$

It is clear that $\alpha_k \in (0,1]$.

Lemma 3 *Let u^{k+1} be given by (5) and $\vec{u}^k \in \operatorname{Fix} T_k$ be the unique solution of VIP$(\mathscr{F}_k, \operatorname{Fix} T_k)$. Then*

$$\|u^{k+1} - \vec{u}^k\|^2 \leq (1-\alpha_k)\|u^k - \vec{u}^k\|^2 + \alpha_k \beta_k \tag{12}$$

Before we turn to the convergence theorem we provide the following auxiliary result.

Lemma 4 *Let $\{a_k\}_{k=0}^{\infty} \subset \mathbb{R}_+$ be a sequence satisfying the inequality*

$$a_{k+1} \leq (1-\alpha_k)a_k + \alpha_k\beta_k + \gamma_k, \tag{13}$$

where $\{\alpha_k\}_{k=0}^{\infty} \subset [0,1]$, $\{\beta_k\}_{k=0}^{\infty} \subset \mathbb{R}_+$, $\{\gamma_k\}_{k=0}^{\infty}, \subset \mathbb{R}_+$. If $\sum_{k=0}^{\infty}\alpha_k = +\infty$, $\limsup_k \beta_k \leq 0$ and $\sum_{k=0}^{\infty}\gamma_k < +\infty$ then

$$\lim_{k\to\infty} a_k = 0.$$

Theorem 1 *Let $\bar{u} \in \operatorname{Fix} T$ be the unique solution of VIP$(\mathscr{F}, \operatorname{Fix} T)$ and let $\bar{u}^k$ be the unique solution of VIP$(\mathscr{F}_k, \operatorname{Fix} T_k)$. Suppose that $\{\mathscr{F}_k\bar{u}^k\}_{k=0}^{\infty}$ is bounded and that*

$$\sum_{k=0}^{\infty} \|\bar{u}^k - \bar{u}\| < +\infty. \tag{14}$$

Let $\{\lambda_k\}_{k=0}^{\infty} \subset (0,\mu]$ be a sequence with

$$\lim_{k\to\infty}\lambda_k = 0 \quad \textit{and} \quad \sum_{k=0}^{\infty}\lambda_k = +\infty. \tag{15}$$

Then for any $u^0 \in \mathscr{H}$ the sequence $\{u^k\}_{k=0}^{\infty}$ generated by (5) converges strongly to $\bar{u}$.

3 Iterations by a Sequence of Contraction Mappings

Now, we study the convergence behavior of the alternative iteration process (6), i.e. the case where $\{u^k\}_{k=0}^{\infty}$ is generated by the iteration

$$u^{k+1} = V_k u^k := T_k\,(I - \lambda_k\mathscr{F}_k)\,u^k. \tag{16}$$

For this type of iteration we apply metric-projections as operators T_k. These operators are non-expansive. Unlike in the preceeding section here the parameters λ_k have not to tend to zero. For the operators $V_k = T_k\,(I - \lambda_k\mathscr{F}_k)$ with Lemma 1 for $\lambda_k := \mu$ we obtain

$$\|V_k u - V_k v\| \leq \sigma\|u - v\| \qquad \text{for all} \quad u, v \in \mathscr{H}_k, \tag{17}$$

where $\sigma := 1 - \tau < 1$. Further, the map $V := T\,(I - \lambda\mathscr{F})$ is assumed to satisfy

$$\|Vu - Vv\| \leq \sigma\|u - v\| \qquad \text{for all} \quad u, v \in \mathscr{H}. \tag{18}$$

As a consequence the operators V_k and V possess unique fixed points $\bar{u}^k$ and $\bar{u}$, respectively, i.e.,

$$\bar{u}^k = V_k\bar{u}^k, \quad k = 0,1,2,\ldots \qquad \text{and} \qquad \bar{u} = V\bar{u}. \tag{19}$$

Theorem 2 *Let the conditions (17), (18) be satisfied and let $\bar{u}$ and $\bar{u}^k$ denote the unique fixed points of the operator V and V_k, respectively. Assume that the approximation property*

$$\sum_{k=0}^{\infty} \|\bar{u}^k - \bar{u}\| < +\infty \tag{20}$$

holds. Then for any $u^0 \in \mathscr{H}_0$ the sequence $\{u^k\}_{k=0}^{\infty}$ generated by (16) converges to the fixed point $\bar{u}$ of V.

Proof With the fixed point property of $\bar{u}^k$ for the operator V_k holds

$$\|u^{k+1} - \bar{u}^k\| = \|V_k u^k - V_k \bar{u}^k\| \leq \sigma \|u^k - \bar{u}^k\|.$$

With the triangle inequality this yields

$$\|u^{k+1} - \bar{u}^{k+1}\| \leq \|u^{k+1} - \bar{u}^k\| + \|\bar{u}^k - \bar{u}^{k+1}\| \leq \sigma \|u^k - \bar{u}^k\| + \|\bar{u}^k - \bar{u}^{k+1}\|.$$

and we obtain

$$\|u^{k+1} - \bar{u}^{k+1}\| \leq \sigma \|u^k - \bar{u}^k\| + \|\bar{u}^k - \bar{u}\| + \|\bar{u}^{k+1} - \bar{u}\|, \quad k \geq 0. \tag{21}$$

Let define $a_k := \|u^k - \bar{u}^k\|$, $\alpha_k := 1 - \sigma$, $\beta_k := 0$ and

$$\gamma_k := \|\bar{u}^k - \bar{u}\| + \|\bar{u}^{k+1} - \bar{u}\|.$$

Then (21) can be expressed by (13). Trivially $\sum_{k=0}^{\infty} \alpha_k = +\infty$ and $\limsup_k \beta_k \leq 0$. The made assumption (20) yields $\sum_{k=0}^{\infty} \gamma_k < +\infty$. Now, we can apply Lemma 4 and obtain $\lim_{k \to \infty} \|u^k - \bar{u}^k\| = 0$. With (20) this completes the proof. ■

4 An Optimal Control Model Problem

Next, we apply the iteration method introduced above to some variational inequality problem that arises from the optimality criterion of an optimal control problem with bounds upon the controls.

Let $\Omega \subset \mathbb{R}^2$ be some open convex polyhedron and Γ its boundary. Consider the optimal control problem

$$\tilde{J}(y,u) := \frac{1}{2} \int_\Omega (y-d)^2 + \frac{\alpha}{2} \int_\Omega u^2 \to \min ! \tag{22a}$$

$$\text{s.t. } -\Delta y = u \text{ in } \Omega, \quad y + \frac{\partial y}{\partial n} = 0 \text{ on } \Gamma, \quad u \in Q := \{u : u \leq b \text{ a.e. in } \Omega\} \tag{22b}$$

with given $\alpha > 0$, $d \in L_2(\Omega)$ and $b \in \mathbb{R}$. The state equation is understood as weak formulation. Let $Y := H^1(\Omega)$ be the Sobolev space of functions over Ω that possess the first order weak derivatives in $L_2(\Omega)$. Further, we let $U = L_2(\Omega)$. Define $a(\cdot,\cdot) : Y \times Y \to \mathbb{R}$ by

$$a(u,v) := \int_\Omega \nabla u \cdot \nabla v + \int_\Gamma uv \qquad \text{for all } u, v \in Y.$$

Now, the weak formulation of (22b) is given by

$$y \in Y : \qquad a(y,v) = \langle u, v \rangle \qquad \text{for all } v \in Y. \tag{23}$$

The Lax-Milgram Lemma (cf. [4]) guarantees that for any $u \in U$ the equation (23) possesses a unique solution. This defines a linear operator $S : U \to Y$ by

$$Su \in Y : \qquad a(Su,v) = \langle u, v \rangle \qquad \text{for all } v \in Y \tag{24}$$

and with the ellipticity constant $\gamma > 0$ this satisfies

$$\|Su\|_1 \leq \frac{1}{\gamma}\|u\|_0. \tag{25}$$

Using the operator S we obtain the reduced form of (22)

$$J(u) := \frac{1}{2}\langle Su-d, Su-d\rangle + \frac{\alpha}{2}\langle u,u\rangle \to \text{ min!} \quad \text{ s.t. } \quad u \in Q. \tag{26}$$

This problem has a unique solution $\bar{u} \in Q$ and this solution can be characterized by a variational inequality that satisfies the general assumptions made above for the abstract problem (cf. [4], [7]).

Theorem 3 *The problem (26) possesses a unique solution $\bar{u} \in Q$. There $\bar{u} \in U$ forms the solution of (26) if and only if*

$$\langle J'(\bar{u}), u-\bar{u}\rangle \geq 0 \qquad \textit{for all } u \in Q \tag{27}$$

holds.

The structure of $J(\cdot)$ yields

$$J'(u) = S^*(Su-d) + \alpha u, \tag{28}$$

where S^* denotes the adjoint of S given by

$$a(z,w) = \langle z, v\rangle \quad \forall z \in Y \qquad \text{and} \qquad S^*v := w.$$

Now, we can show that (27) is of the considered abstract type. First, the Hilbert space $\mathscr{H}$ is just the space $U = L_2(\Omega)$. The set $Q \subset U$ is closed and convex. Thus, the metric-projection $P_Q : U \to Q$ is well defined by

$$P_Q u \in Q: \qquad \|P_Q u - u\|_0 \leq \|v-u\|_0 \qquad \text{for all } v \in Q,$$

and we have

$$u \in Q \quad \Longleftrightarrow \quad u \in \operatorname{Fix} P_Q.$$

From $Q \neq \emptyset$ and from the nonexpansivity of P_Q we obtain that $T := P_Q$ is a quasi-nonexpansive operator. Further, the operator $\mathscr{F} : U \to U$ is defined by

$$\mathscr{F}u := S^*(Su-d) + \alpha u.$$

Trivially, it is Lipschitz continuous because of $U \hookrightarrow H^2(\Omega)$ this yields

$$\|\mathscr{F}u - \mathscr{F}v\| \leq (\|S^*\|\,\|S\| + \alpha)\,\|u-v\| \qquad \text{for all } u, v \in U$$

and with (25) we obtain

$$\|\mathscr{F}u - \mathscr{F}v\| \leq \left(\frac{1}{\gamma^2} + \alpha\right)\|u-v\| \qquad \text{for all } u, v \in U. \tag{29}$$

Further, we have

$$\langle \mathscr{F}u - \mathscr{F}v, u-v\rangle \geq \alpha\,\|u-v\|^2 \quad \text{ for all } u, v \in U.$$

This proves that the problem (27) satisfies all assumptions made for the general problem. As a consequence some parameter $\lambda > 0$ can be found such that the iteration

$$u^{k+1} = T(u^k - \lambda \mathscr{F}u^k), \quad k \geq 0. \tag{30}$$

for any $u^0 \in U$ converges to the optimal solution $\bar{u} \in Q$ of the considered control problem. However, the iteration (30) is in the function space U and requires an appropriate discretization.

5 Families of Conforming Discretizations

Let $U_k \subset U$, $Y_k \subset Y$ we apply a piecewise linear C^0-discretization over nested families $\{\mathscr{T}_k\}$ of uniformly regular triangulations (see [4]). With h_k we denote the maximal diameter of the triangles in $\{\mathscr{T}_k\}$. We denote the related grid points by $\Omega_k := \{x^{k,j}\}_{j=1}^{N_k}$. We have

$$\Omega_k \subset \Omega_{k+1}, \quad k \geq 0. \tag{31}$$

Let $\phi_{k,j} \in C(\bar{\Omega})$ be piecewise linear over $\mathscr{T}_k$ with

$$\phi_{k,i}(x^{k,j}) = \delta_{ij}, \qquad i,j = 1,\ldots,N_k, \quad k \geq 0.$$

This means we assume $\{\phi_{k,j}\}$ to form a Lagrangian basis of U_k as well as of Y_k, where

$$U_k := Y_k := \operatorname{span}\left\{\phi_{k,j}\right\}_{j=1}^{N_k}.$$

With (31) this implies $\quad U_k \subset U_{k+1} \subset \cdots \subset U, \qquad Y_k \subset Y_{k+1} \subset \cdots \subset Y.$
For given $u \in U$ the conforming discretization of the state equations have the form:

$$y_k \in Y_k: \qquad a(y_k, v) = \langle u, v\rangle \qquad \text{for all } v \in Y_k. \tag{32}$$

Again, the Lax-Milgram Lemma implies that for any $u \in U$ problem (32) possesses a unique solution. Thus, $S_k u := y_k$ defines linear mappings $S_k : U \to Y_k \subset Y$. Let

$$Q_k := \{u \in U_k : \ u \leq b\}.$$

This yields the following discrete problems

$$J_k(u) := \frac{1}{2}\langle S_k u - d, S_k u - d\rangle + \frac{\alpha}{2}\langle u,u\rangle \to \text{ min!} \quad \text{ s.t. } \quad u \in Q_k\,. \tag{33}$$

Problem (33) has a unique solution $\bar{u_k} \in Q_k$. Analogously to the continuous case we define $\mathscr{F}_k : U \to Y_k$ by

$$\mathscr{F}_k u := S_k^*(S_k u - d) + \alpha u \qquad \text{for all } u \in U$$

and $T_k := P_{Q_k}$, the metric-projection onto Q_k. Then

$$u_{j+1} = T_k(u_j - \lambda_k \mathscr{F}_k u_j), \quad j \geq 0. \tag{34}$$

forms an iteration technique on the discretization level k. For sufficiently small $\lambda_k > 0$ the sequence generated by (34) converges to the unique solution $\bar{u}^k$ of the discrete control problem (33). As mentioned in Section 3, there are two different types of iterations. Instead of (34) we may also apply iteration-discretization method

$$u_{k+1} = T_k(u_k - \lambda_k \mathscr{F}_k u_k), \quad k \geq 0. \tag{35}$$

with appropriate parameters $\lambda_k > 0$. This recursion differs from (34) by acting on a family of discrete problems, where on each discretization level only a maximal number of iteration steps is performed. In particular, if the discretization changes in each step then (35) means that only one iteration step is performed per discretization.

Now, we derive that (35) fulfills the assumptions made in Section 3. For the proofs we refer again to [3].

Lemma 5 *Let $U_0 \neq \emptyset$ then* $\operatorname{Fix} T_k \neq \emptyset$, $k \geq 0$ *holds. Further, we have*

$$\operatorname{Fix} T_k \subset \operatorname{Fix} T_{k+1} \subset \operatorname{Fix} T \qquad k \geq 0.$$

Lemma 6 *The continuous optimal control problem (26) has a unique solution $\bar{u}$ and for any $k \geq 0$ the discrete control problem (33) possesses a unique solution $\bar{u}^k \in Q_k$. Further, there is a constant $c > 0$ with*

$$\|\mathscr{F}_k \bar{u}^k\| \leq c, \quad k \geq 0.$$

Theorem 4 *Let $\bar{u}^k \in Q_k$ denote the solution of the discrete problem (33) and $\bar{u} \in Q$ the solution the original continuous problem (26). Then there exists a constant $c > 0$ such that*

$$\|\bar{u}^k - \bar{u}\| \leq c\, h_k. \tag{36}$$

In principle, the estimate given in Theorem 4 could be refined to

$$\|\bar{u}^k - \bar{u}\| \leq c\, h_k^{3/2}$$

if the technique proposed in [6] is applied to the discretization of the elliptic control problem under consideration. In our application, however, the bound (36) is already sufficient to ensure the convergence for simple refinement strategies. Indeed, if $\mathscr{T}_k$ is generated by subdivisions of all triangles using the midpoints of all edges then we obtain $h_{k+1} = h_k/2$ and consequently $\sum_{k=0}^{\infty} h_k < +\infty$ holds. Thus, the proposed iteration-discretization technique converges.

Conclusion

The present paper provides some theoretical approach to the simultaneous refinement of the discretzation of function spaces and iterations for solving variational inequality problems over the fixed point set of a quasi-nonexpansive operator in Hilbert spaces. This approach forms some basis for implementable algorithms because it avoids nested infinite processes. The increase the efficiency of the discussed methods further improvements, e.g. preconditioning, have to be investigated in the future.

Literature

[1] BAUSCHKE, H. H.; COMBETTES, P. L.: A weak to strong convergence principle for Fejér-monotone methods in Hilbert spaces. *Math. Operations Res.*, 26 (2001), pp. 248–264.

[2] CEGIELSKI, A.: Iterative methods for fixed point problems of nonexpansive operators in Hilbert spaces. *Lecture Notes in Mathematics* 2057, Springer Heidelberg 2012. (in print)

[3] CEGIELSKI, A.; GROSSMANN, C.: *Iteration-discretization methods for variational inequality over fixed point sets*. (submitted)

[4] GROSSMANN, C.; ROOS, H.-G.; STYNES, M.: *Numerical treatment of partial differential equations*. Springer, Berlin, 2007.

[5] OGURA, N.; YAMADA, I.: Nonstrictly convex minimization over the bounded fixed point set of a nonexpansive mapping. *Numer. Funct. Anal. Opt.*, 24 (2003), pp. 129–135.

[6] RÖSCH, A.: Error estimates for parabolic optimal control problems with control constraints. *Z. Anal. Anwendungen*, 23 (2004), pp. 353–376.

[7] TRÖLTZSCH, F.: Optimal control of partial differential equations. Theory, methods and applications. *Graduate Studies in Mathematics*, 112. AMS, Providence, RI, 2010.

[8] YAMADA, I.; OGURA, N.: Hybrid steepest descent method for variational inequality problem over the fixed point set of certain quasi-nonexpansive mapping. *Numer. Funct. Anal. Opt.*, 25 (2004), pp. 619–655.

Iteračně–diskretizační metody pro variační nerovnost

Iterační metody pro řešení variačních nerovností v Hilbertových prostorech nekonečné dimenze vyžadují diskretizaci. To vede k řešení posloupnosti variačních nerovností v prostorech konečné dimenze. Tato práce se věnuje iteračním metodám, které vyžadují pouze konečný počet kroků na každé diskretizační úrovni. Nejprve je studována abstraktní úloha a následně konkrétní úloha optimálního řízení s eliptickou stavovou rovnicí a s omezeními na řídící proměnnou. Diskretizace je provedena pomocí posloupnosti do sebe vnořených po částech lineárních, spojitých, konformních konečných prvků.

Ein Iterations-Diskretisierungs-Verfahren für eine Variationsungleichung

Iterationsverfahren zur Behandlung von Variationsungleichungen in undendlichdimensionalen Hilbert-Räumen erfordern in der Regel eine Diskretisierung. Diese führt auf Variationsungleichungen über einer Familie von Räumen. In der vorliegenden Arbeit wird dieses Problem durch ein Iterationsverfahren mit einer nur endlichen Zahl von Schritten je Diskretisierungsniveau behandelt. Zunächst werden abstrakte Methoden untersucht und später auf ein Problem der optimalen Steuerung mit elliptischen Zustandsgleichungen und Steuerrestriktionen angewandt. Die Diskretisierung erfolgt durch eine sich verfeinernde Familie stückweise linearer, konformer C^0 finiter Elemente.

Metoda iteracyjno-dyskretyzacyjna dla nierówności wariacyjnej

Metody iteracyjne dla nierówności wariacyjnych w nieskończenie-wymiarowych przestrzeniach Hilberta zazwyczaj wymagają dyskretyzacji. Ta z kolei prowadzi do nierówności wariacyjnych określonych na rodzinie przestrzeni. W niniejszej pracy dla tego problemu stosujemy metodę iteracyjną, w której na każdym poziomie dyskretyzacji przeprowadzanych jest skończenie wiele kroków. Najpierw badamy metody abstrakcyjne, które następnie stosujemy do problemu sterowania optymalnego z eliptycznym równaniem stanu i z ograniczeniami na sterowanie. Przy dyskretyzacji stosujemy rodzinę zagęszczających się elementów skończonych, które są kawałkami liniowe, ciągłe i konforemne.

ADAPTIVE WAVELET SCHEME FOR CONVECTION–DIFFUSION EQUATIONS

Dana Černá
***Václav Finěk**

Technical University of Liberec
Faculty of Science, Humanities and Education
Department of Mathematics and Didactics of Mathematics
Studentská 2, 461 17, Liberec, Czech Republic
dana.cerna@tul.cz

*Technical University of Liberec
Faculty of Science, Humanities and Education
Department of Mathematics and Didactics of Mathematics
Studentská 2, 461 17, Liberec, Czech Republic
vaclav.finek@tul.cz

Abstract

The paper is concerned with theoretical and computational issues of a numerical resolution of the convection-diffusion equation. We use an implicit scheme for the time discretization and an adaptive wavelet-based method for a spatial discretization. We use a well-conditioned cubic spline-wavelet basis and a method for an inexact multiplication of wavelet stiffness matrix with a vector which we have recently proposed in [1, 2]. The theoretical advantages of our scheme as well as numerical examples will be presented.

Keywords: Convection-diffusion equation; spline; wavelet; adaptivity.

Introduction

In this paper we consider a numerical approximation of the convection-diffusion equation

$$\frac{\partial u}{\partial t} = \mu \Delta u - b \operatorname{div} u - cu + f \quad \text{for} \quad x \in \Omega, \quad t \in (0,T), \tag{1}$$

with initial and boundary conditions

$$u(x,0) = u^0(x) \quad \text{for} \quad x \in \Omega, \quad u(x,t) = 0 \quad \text{for} \quad x \in \partial\Omega.$$

We assume that $\mu > 0$, b and c are constants, $u : \Omega \times (0,T) \to \mathbb{R}$, $f, u^0 \in L^2(\Omega)$. We consider only the domain $\Omega = (0,1)^n$. It is well-known that the solution of (1) typically contains layers and that the numerical solution by the Galerkin method on uniform mesh suffers from the Gibbs phenomenon. Thus, it is convenient to solve the problem adaptively. In this paper, we use a modification of the wavelet-based adaptive method from [6] for a spatial discretization, because it has several advantages, namely:

- The adaptivity in the context of a wavelet discretization is simple. It consists in keeping the large wavelet coefficients and discarding the smaller ones.
- The algorithm is asymptotically optimal. It means that the number of floating point operations depends linearly on the number of nonzero wavelet coefficients.

- The condition numbers of stiffness matrices are uniformly bounded.

The paper is organized as follows: In Section 2 we use the backward Euler formula for a discretization in time and derive a variational formulation for the given time level. We define an equivalent l^2-problem and propose iterations for solving this problem in Section 3. In Section 4 we introduce an implementable version of the iterations and discuss computational issues. Numerical examples are presented in Section 5.

1 Discretization

Let $h > 0$, $t_k := kh$ and u^k be an approximation of $u(\cdot, t_k)$. For simplicity, we use the backward Euler method for a discretization in time:

$$\frac{u^{k+1} - u^k}{\tau} = \mu \Delta u^{k+1} - b \operatorname{div} u^{k+1} - c u^{k+1} + f.$$

Then the variational formulation of our problem at the time level t_k is

$$a_k\left(u^{k+1}, v\right) = f^k(v), \quad v \in H_0^1(\Omega), \tag{2}$$

where a continuous bilinear form $a_k : H_0^1(\Omega) \times H_0^1(\Omega) \to \mathbb{R}$ and $f^k \in H_0^{-1}(\Omega)$ are given by

$$a_k(u,v) := \mu \int_\Omega \nabla u \cdot \nabla v \, \mathrm{d}x + b \int_\Omega v \operatorname{div} u \, \mathrm{d}x + \left(\frac{1}{\tau} + c\right) \int_\Omega uv \, \mathrm{d}x$$

$$f^k(v) := \frac{1}{\tau} \int_\Omega u^k v \, \mathrm{d}x + \int_\Omega f v \, \mathrm{d}x,$$

where $H_0^1(\Omega)$ denotes the subspace of all functions from a Sobolev space $H^1(\Omega)$ with zero traces on $\partial\Omega$, $H_0^{-1}(\Omega)$ denotes its dual. In the sequel, we solve (2) by wavelet-based method. For this reason we propose the definition of a wavelet basis of Sobolev space $H_0^s(\Omega)$.

Family $\Psi := \{\psi_\lambda, \lambda = (j,k) \in \mathscr{J}\}$ for an infinite set $\mathscr{J} = \mathscr{J}_\Phi \cup \mathscr{J}_\Psi$, $\#\mathscr{J}_\Phi < \infty$, is called the *wavelet basis* of $H_0^1(\Omega)$, if

i) Ψ is a Riesz basis of $H_0^1(\Omega)$, that means Ψ generates $H_0^1(\Omega)$ and there exist constants $c, C \in (0, \infty)$ such that for all $\mathbf{b} := \{b_\lambda\}_{\lambda \in \mathscr{J}} \in l^2(\mathscr{J})$, where $\lambda = (j,k)$ and $|\lambda| = j$ denotes the level, holds

$$c \|\mathbf{b}\| \le \left\| \sum_{\lambda \in \mathscr{J}} b_\lambda 2^{-|\lambda|} \psi_\lambda \right\|_{H^s(\Omega)} \le C \|\mathbf{b}\|.$$

ii) Functions are local in the sense that $\operatorname{diam}(\Omega_\lambda) \le C 2^{-|\lambda|}$ for all $\lambda \in \mathscr{J}$, where Ω_λ is support of ψ_λ.

iii) Functions have cancellation properties of the order m, i.e.

$$|\langle v, \psi_\lambda \rangle| \le 2^{-m|\lambda|} |v|_{H^m(\Omega_\lambda)}, \quad \lambda \in \mathscr{J}_\psi, \quad v \in H_0^1(\Omega),$$

where $\langle \cdot, \cdot \rangle$ denotes the $L^2(\Omega)$ inner product.

In this paper, Ψ will be a cubic spline wavelet basis adapted to homogeneous Dirichlet boundary conditions from [1] that has a cancellation property of the order six. The basis functions are local and also their duals are local. Its structure is

$$\Psi = \Phi_4 \cup \bigcup_{j=4}^{\infty} \Psi_j, \tag{3}$$

where Φ_4 contains scaling functions and Ψ_j is formed by wavelets on the level j. Also the smoothness of basis functions is an important property. In this case, the Sobolev exponent of smoothness is 2.5.

2 Wavelet Method

In this section we propose a method for solving (2) based on wavelets. Let $\mathbf{D}_k$ be a diagonal matrix with diagonal elements $\sqrt{a_k(\psi_\lambda, \psi_\lambda)}$. Then $\mathbf{D}_k^{-1}\Psi$ is a wavelet basis in $H_0^1(\Omega)$ and the equation (2) can be reformulated as an equivalent biinfinite matrix equation

$$\mathbf{A}^k \mathbf{u}^k = \mathbf{f}^k, \tag{4}$$

where $\mathbf{A}^k = \mathbf{D}_k^{-1} a_k(\Psi, \Psi)\mathbf{D}_k^{-1}$ is a diagonally preconditioned stiffness matrix, $u^k = \left(\mathbf{u}^k\right)^T D_k^{-1}\Psi$ and $\mathbf{f}^k = \mathbf{D}_k^{-1} f^k(\Psi)$.

Then, u^k solves (2) if and only if $\mathbf{u}^k$ solves the matrix equation (4). Moreover, the matrix $\mathbf{A}^k$ satisfies

$$\operatorname{cond} \mathbf{A}^k \leq C < +\infty. \tag{5}$$

While the classical adaptive methods use refining and derefining a given mesh according to a-posteriori local error indicators, the wavelet approach is somewhat different and follows a paradigm which comprises the following steps:

1. One starts with a variational formulation but instead of turning to a finite dimensional approximation, using the suitable wavelet basis the continuous problem is transformed into an infinite-dimensional l^2-problem.

2. One then tries to devise convergent iterations for the l^2-problem.

3. Finally, one derives a practical version of this idealized iteration. All infinite-dimensional quantities have to be replaced by finitely supported ones and the routine for the application of the biinfinite-dimensional matrix $\mathbf{A}$ approximately have to be designed.

We solve the discrete infinite-dimensional problem (4) approximately. For notational simplicity we omit the index k in this section. Matrix $\mathbf{A} := \mathbf{A}^k$ is not symmetric positive definite, but one can obtain a symmetric positive definite formulation by squaring: $\mathbf{L} := \mathbf{A}^T\mathbf{A}$, $\mathbf{g} := \mathbf{A}^T\mathbf{f}$. Then (4) is equivalent to $\mathbf{Lu} = \mathbf{g}$. There exists some relaxation weight ω with

$$\|\mathbf{I} - \omega\mathbf{L}\| \leq \rho < 1. \tag{6}$$

Thus simple iterations of the form

$$\mathbf{u}^{n+1} = \mathbf{u}^n + \omega(\mathbf{g} - \mathbf{L}\mathbf{u}^n) \tag{7}$$

converge. Neither we can evaluate the generally infinite array $\mathbf{g}$ exactly, nor we can compute $\mathbf{Lu}$, even when $\mathbf{u}$ has a finite support. Thus, we need to approximate $\mathbf{g}$ and product $\mathbf{Lu}$ with some given precision.

3 Adaptive Wavelet Scheme

We use the following implementable version of the ideal iteration (7).

SOLVE [L, g, ε] $\to \mathbf{u}_\varepsilon$
Let ω satisfy (6) and $K \in \mathbb{N}$ be fixed such that $\rho^K < 1/10$.
Set $j := 0$, $\mathbf{u}_0 := 0$, $\varepsilon_0 := \left\|\mathbf{L}^{-1}\right\| \|\mathbf{g}\|$.
While $\varepsilon_j > \varepsilon$ do
$j := j+1$, $\varepsilon_j := 10\rho^K \varepsilon_{j-1}$, $\mathbf{g}_j := \mathbf{RHS}[\mathbf{g}, \frac{\varepsilon_j}{20\omega K}]$, $\mathbf{z}_0 := \mathbf{u}_{i-1}$,
$l := l+1$, $\mathbf{res}_l := 0$,
If $l \le K$ and $\|\mathbf{res}_l\| > \left(\left\|\mathbf{L}^{-1}\right\| - \frac{1}{10\omega K}\right)\varepsilon_j$ do
$l := l+1$,
$\mathbf{res}_l := \mathbf{g}_j - \mathbf{APPLY}[\mathbf{L}, \mathbf{z}_{l-1}, \frac{\varepsilon_j}{20\omega K}]$,
$\mathbf{z}_l := \mathbf{z}_{l-1} + \omega\, \mathbf{res}_l$,
end if,
$\mathbf{u}_j := \mathbf{COARSE}[\mathbf{z}_l, 0.7\varepsilon_j]$,
end while,
$\mathbf{u}_\varepsilon := \mathbf{u}_j$.

Let us comment particular subroutines. The subroutine **COARSE** computes a vector $\mathbf{v}_\varepsilon$ which is close to the vector of $\mathbf{v}^N$ of N-largest coefficients of $\mathbf{v}$ for which $\left\|\mathbf{v} - \mathbf{v}^N\right\| < \varepsilon$. Since sorting of all elements of $\mathbf{v}$ requires $\mathcal{O}(M \log M)$ operations, where M is the length of $\mathbf{v}$, the procedure **COARSE** uses so called binning [8].

COARSE[v, ε] $\to \mathbf{v}_\varepsilon$
1. Set $q := \left\lceil \log\left((\#\mathrm{supp}\ \mathbf{v})^{1/2} \|\mathbf{v}\|_{l^2} / \varepsilon \right) \right\rceil$.
2. Regroup the elements of $\mathbf{v}$ into the sets $B_0, \ldots, B_q$, where $\mathbf{v}_\lambda \in B_i$ if and only if

$$2^{-(i+1)} \|\mathbf{v}\|_{l^2} < |\mathbf{v}_\lambda| \le 2^{-i} \|\mathbf{v}\|_{l^2}, \quad 0 \le i < q. \tag{8}$$

Possible remaining elements are put into the set B_q.

3. Create $\mathbf{v}_\varepsilon$ by collecting nonzero entries from B_0 and when it is exhausted from B_1 and so forth until

$$\|\mathbf{v} - \mathbf{v}_\varepsilon\| \le \varepsilon. \tag{9}$$

The subroutine

$$\mathbf{RHS}\,[\mathbf{g}, \varepsilon] \to \mathbf{g}_\varepsilon \tag{10}$$

approximates the right-hand side $\mathbf{g}$ by the finitely supported vector $\mathbf{g}_\varepsilon$ such that

$$\|\mathbf{g} - \mathbf{g}_\varepsilon\| \le \varepsilon. \tag{11}$$

It can be realized by computing in a preprocessing step a highly accurate approximation to $\mathbf{g}$ in the dual basis along with the corresponding coefficients and then applying of **COARSE** to this finitely supported array of coefficients.

In [2], we proposed the improved matrix-vector multiplication in the context of adaptive wavelet methods. Unlike [3, 9], we are not searching for 2^k greatest vector entries in absolute value, but instead we trace actual decay of matrix and vector entries. Let us denote

$$(\mathbf{L}_k)_{\lambda,\lambda'} := \begin{cases} \mathbf{L}_{\lambda,\lambda'}, & ||\lambda| - |\lambda'|| \le k, \\ 0, & \textit{otherwise}. \end{cases} \tag{12}$$

and

$$S_{\mathbf{L}k} := \max \left| (\mathbf{L}_k)_{\lambda,\lambda'} \right|, \lambda, \lambda' \in \mathscr{J} \tag{13}$$

for $k \geq 0$, and $\mathbf{v}_k$ contains all vector entries greater than a given tolerance divided by $S_{\mathbf{L}_k}$. Then we can compute

$$\mathbf{Lv} \approx \mathbf{w}_K := \sum_{k=0}^{K} \mathbf{L}_k \mathbf{v}_k.$$

We can choose K such that

$$\|\mathbf{w}_K - \mathbf{Lv}\| \leq \varepsilon. \tag{14}$$

In [2], it was shown that the number of floating point operations needed for computation of the approximate product **LV**, depends linearly on the number of nonzero elements of $\mathbf{w}_K$.

Theorem 1. *Under the above assumptions, for any $\varepsilon > 0$, the approximations $\mathbf{u}_\varepsilon$ produced by* **SOLVE** *satisfy*

$$\left\| u - \mathbf{u}_\varepsilon^T \mathbf{D}^{-1} \Psi \right\| \lesssim \varepsilon \tag{15}$$

and

$$\#flops \leq C \# \operatorname{supp} \mathbf{u}_\varepsilon, \tag{16}$$

where the constant C does not depend on ε.

Proof. Since the used subroutines are asymptotically optimal, the asymptotical optimality of **SOLVE** can be proven by similar way as in [6]. □

As an alternative to the Richardson iterations the steepest descent approach was studied in [7].

4 Numerical examples

In this section, we present two numerical examples.

Example 1. *Let us consider the equation*

$$\mu u'' + u' - u = 1, \tag{17}$$

with homogeneous Dirichlet boundary conditions

$$u(0) = u(1) = 0. \tag{18}$$

It is known that the exact solution is given by

$$u(x) = \frac{\left(e^l - 1\right) e^{kx} - \left(1 - e^k\right) e^{lx}}{e^k - e^l} - 1, \tag{19}$$

where

$$k = \frac{-1 + \sqrt{1 + 4\mu}}{2\mu}, \quad l = \frac{-1 - \sqrt{1 + 4\mu}}{2\mu}. \tag{20}$$

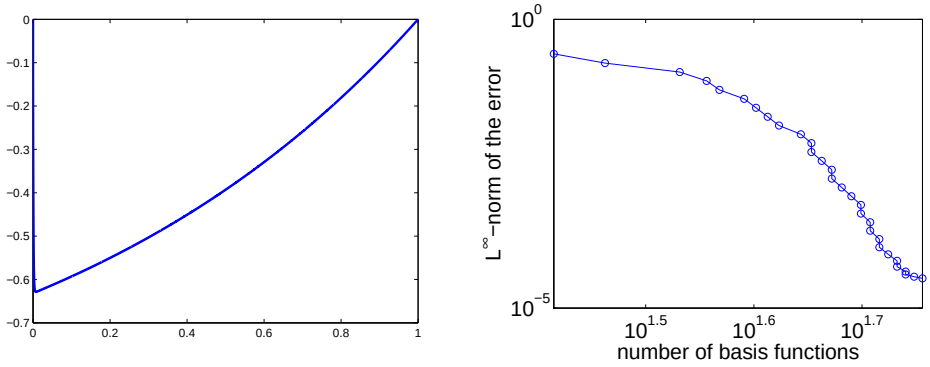
The condition numbers of stiffness matrices $\mathbf{A}^s$ for this equation corresponding to the multi-scale wavelet basis with s levels of wavelets are displayed in Table 1. It can be seen that they are indeed uniformly bounded.

In this case a boundary layer occurs near the point 0. *The graph of u for $\mu = 10^{-3}$ is displayed at Figure 1. We solved this equation by an adaptive wavelet scheme from Section 3. Convergence history is shown in Figure 1. Only* 54 *basis function were used for achieving the error* $3.27 \cdot 10^{-5}$ *in the L^∞-norm.*

Tab. 1. *The condition numbers of stiffness matrices* $\mathbf{A}^s$ *corresponding to the multi-scale wavelet basis with s levels of wavelets*

s	cond $\mathbf{A}^s$
1	75.1
2	96.7
3	101.2
4	101.2
5	101.8
6	101.8
7	101.8
8	101.8

Source: Own



Source: Own

Fig. 1. *The graph of the exact solution (left) and convergence history (right) for Example 1. for* $\mu = 10^{-3}$

Example 2. *Let us consider the equation*

$$\frac{\partial u}{\partial t} = \mu u'' + u' - e^t, \tag{21}$$

with initial and boundary conditions

$$u(x,0) = u(x) \quad \text{for} \quad x \in [0,1], \quad u(0,t) = u(1,t) = 0 \quad \text{for} \quad t \in [0,1], \tag{22}$$

where $u(x)$ *is given by (19). Then the exact solution is*

$$u(x,t) = u(x)\, e^t. \tag{23}$$

For $\mu = 10^{-3}$ *and* $\tau = 0.05$ *we need only* 355 *basis functions to obtain an error less than* 10^{-3} *at the time* $T = 1$.

Conclusion

In this paper we proposed an adaptive wavelet-based method for numerical resolution of the convection-diffusion equation and we briefly reviewed some aspects of a wavelet discretization in connection with our numerical scheme. Computational details and theoretical results one can find in [1, 2, 4, 5, 6].

Acknowledgments

This work was supported by the project ESF "Constitution and improvement of a team for demanding technical computations on parallel computers at TU Liberec" No. CZ.1.07/2.3.00/09.0155.

Literature

[1] ČERNÁ, D.; FINĚK, V.: Construction of optimally conditioned cubic spline wavelets on the interval. *Adv. Comput. Math.* 34, pp. 519-552, 2011.

[2] ČERNÁ, D.; FINĚK, V.: *Approximate multiplication in adaptive wavelet methods.* Submitted.

[3] COHEN, A.; DAHMEN, V.; DEVORE, R.: Adaptive Wavelet Schemes for Elliptic Operator Equations - Convergence Rates. *Math. Comput.* 70, pp. 27-75, 2001.

[4] COHEN, A.: *Numerical Analysis of Wavelet Methods*. Elsevier Science, Amsterdam, 2003.

[5] COHEN, A.; DAHMEN, V.; DEVORE, R.: Adaptive wavelet techniques in numerical simulation. *Encyclopedia of Computational Mathematics* 1, pp. 157-197, 2004.

[6] COHEN, A.; DAHMEN, V.; DEVORE, R.: Adaptive wavelet methods II - Beyond the elliptic case. *Foundations of Computational Mathematics* 2, pp. 203-245, 2002.

[7] DAHLKE, S.; FORNASIER, M.; RAASCH, T.; STEVENSON, R.; WERNER, M.: Adaptive Frame Methods for Elliptic Operator Equations: The Steepest Descent Approach. *IMA J. Numer. Anal.* 27, pp. 717-740, 2007.

[8] STEVENSON, R.: Adaptive Solution of Operator Equations Using Wavelet Frames. *SIAM J. Numer. Anal.* 41, pp. 1074-1100, 2003.

[9] DIJKEMA, T.J.; SCHWAB, Ch.; STEVENSON, R.: An adaptive wavelet method for solving high-dimensional elliptic PDEs. *Constructive approximation* 30, pp. 423-455, 2009.

RNDr. Dana Černá, Ph.D., RNDr. Václav Finěk, Ph.D.

Adaptivní waveletové schéma pro konvektivně-difůzní rovnici

Článek se zabývá teoretickými a výpočetními aspekty numerického řešení konvektivně-difůzních rovnic. Použijeme implicitní schéma pro časovou diskretizaci a adaptivní waveletovou metodu pro prostorovou diskretizaci. Použijeme dobře podmíněnou kubickou spline-waveletovou bázi a metodu pro přibližné násobení waveletové matice tuhosti s vektorem, které jsme nedávno navrhli. Budou prezentovány teoretické výhody našeho schématu a také numerické příklady.

Das adaptive Wavelet-Schema für konvektiv-diffuse Gleichungen

Dieser Artikel befasst sich mit theoretischen und rechnerischen Aspekten der numerischen Lösung konvektiv-diffuser Gleichungen. Dazu verwenden wir ein implizites Schema für die zeitliche Diskretisierung und die adaptive Wavelet-Methode für eine räumliche Diskretisierung. Wir benutzen die gut bedingte kubische Spline-Wavelet-Basis und -Methode für eine annähernde Multiplikation der Zähigkeits-Wavelet-Matrix mit einem Vektor, den wir vor nicht langer Zeit entworfen haben. Es werden theoretische Vorteile unseres Schemas und auch numerische Beispiele vorgestellt.

Adaptacyjny schemat falkowy dla równania konwekcyjno-dyfuzyjnego

W artykule przedstawiono teoretyczne i obliczeniowe aspekty numerycznego rozwiązywania równań konwekcyjno-dyfuzyjnych. Zastosowano schemat niejawny (dyskretny) dla czasowej dyskretyzacji oraz adaptacyjną metodę falkową (waveletową) dla dyskretyzacji przestrzennej. Wykorzystujemy dobrze uwarunkowaną kubiczną bazę splajno-falkową oraz metodę przybliżonego mnożenia falkowej macierzy sztywności z wektorem, które niedawno zaproponowaliśmy. Zostały zaprezentowane teoretyczne zalety naszego schematu oraz przykłady numeryczne.

ON A SPARSE REPRESENTATION OF LAPLACIAN

Dana Černá
***Václav Finěk**
****Zdeňka Ondračková**

Technical University of Liberec
Faculty of Science, Humanities and Education
Department of Mathematics and Didactics of Mathematics
Studentská 2, 461 17, Liberec, Czech Republic
dana.cerna@tul.cz

*Technical University of Liberec
Faculty of Science, Humanities and Education
Department of Mathematics and Didactics of Mathematics
Studentská 2, 461 17, Liberec, Czech Republic
vaclav.finek@tul.cz

**Technical University of Liberec
Faculty of Science, Humanities and Education
Department of Mathematics and Didactics of Mathematics
Studentská 2, 461 17, Liberec, Czech Republic
zdenka.ondrackova@tul.cz

Abstract

One of the most important part of adaptive wavelet methods is an efficient approximate multiplication of stiffness matrices with vectors in wavelet coordinates. Although there are known algorithms to perform it in linear complexity, the application of them is relatively time consuming and its implementation is very difficult. Therefore, it is necessary to develop a well-conditioned wavelet basis with respect to which both the mass and stiffness matrices are sparse in the sense that the number of nonzero elements in any column is bounded by a constant. Then, matrix-vector multiplication can be performed exactly with linear complexity. We present here a wavelet basis on the interval with respect to which both the mass and stiffness matrices corresponding to the one-dimensional Laplacian are sparse. Consequently, the stiffness matrix corresponding to the n-dimensional Laplacian in tensor product wavelet basis is also sparse. Moreover, the constructed basis has an excellent condition number. In this contribution, we shortly review this construction and show several numerical tests.

Keywords: Wavelet; Hermite cubic splines; sparse representations.

Introduction

A general concept for solving of operator equations by means of wavelets was proposed by A. Cohen, W. Dahmen and R. DeVore in [2, 3]. The aim of this concept consists in the approximation of the unknown solution u which should correspond to the best N-term approximation, and the associated computational work should be proportional to the number of unknowns. It consists of the following steps: transformation of the variational formulation into the well-conditioned infinite-dimensional problem in the space l^2, finding of the convergent iteration process for the l^2 problem and finally derivation of its computable version. The essential step to achieve this goal is an efficient approximate multiplication of quasi-sparse wavelet matrices

with vectors.

In [2], authors exploited an off-diagonal decay of entries of the wavelet stiffness matrices and designed a numerical routine **APPLY** which approximates the exact matrix-vector product with the desired tolerance ε and that has linear computational complexity, up to sorting operations. The idea of **APPLY** is following: To truncate $\mathbf{A}$ in scale by zeroing $a_{i,j}$ whenever $\delta(i,j) > k$ (δ represents the level difference of two functions in the wavelet expansion) and denote the resulting matrix by $\mathbf{A_k}$. At the same time to sort vector entries $\mathbf{v}$ with respect to the size of their absolute values. One obtains $\mathbf{v_k}$ by retaining 2^k greatest coefficients in absolute values of $\mathbf{v}$ and setting all other equal to zero. The maximum value of k should be determined to reach a desired accuracy of approximation. Then one computes an approximation of $\mathbf{Av}$ by

$$\mathbf{w} := \mathbf{A_k v_0} + \mathbf{A_{k-1}}(\mathbf{v_1} - \mathbf{v_0}) + \ldots + \mathbf{A_0}(\mathbf{v_k} - \mathbf{v_{k-1}}) \tag{1}$$

with the aim to balance both accuracy and computational complexity at the same time. In [6], binning and approximate sorting strategy was used to eliminate these sorting costs and then an asymptotically optimal algorithm was obtained. The idea is following: Reorder the elements of $\mathbf{v}$ into the sets $V_0, \ldots, V_q$, where $v_\lambda \in V_i$ if and only if

$$2^{-i-1}\|\mathbf{v}\|_{l^2} < v_\lambda < 2^{-i}\|\mathbf{v}\|_{l^2}, \quad 0 \le i < q.$$

Eventual remaining elements are put into the set V_q. And subsequently to generate vectors $\mathbf{v_k}$ by successively extracting 2^k elements from $\bigcup_i V_i$, starting from V_0 and when it is empty continuing with V_1 and so forth. Finally, the scheme (1) is applied. Further improvements of this scheme were proposed in [1, 4].

The described scheme was shown to possess the best possible rate of convergence in linear complexity and since tensor product wavelets are applied, this rate is independent of the space dimension [4]. Although it has optimal computational complexity, the application of the **APPLY** routine is relatively time consuming and moreover it is not easy to implement it efficiently. Therefore, it is necessary to develop a well-conditioned wavelet basis with respect to which both the mass and stiffness matrix are sparse. This means that the number of nonzero elements in any column as well as the condition number of stiffness matrices are bounded independent of the matrix size which is not generally true using a wavelet discretization. Then, a matrix-vector multiplication can be performed exactly with linear complexity.

The remainder of this paper is organized as follows. In the next section, we present a construction of a wavelet basis on the interval based on Hermite multiwavelets with respect to which both the mass and stiffness matrices corresponding to the one-dimensional Laplacian are sparse and very well-conditioned. We also identify several free parameters in the construction of the second two wavelets which can be used to improve properties of a constructed basis. In the last section, we present several numerical tests to compare a proposed basis with the basis proposed in [5].

1 Hermite Multiwavelets

We start with Hermite cubic splines as the primal scaling bases on the interval. They are defined by

$$\phi_1(x) = \begin{cases} (x+1)^2(1-2x) & -1 \le x \le 0 \\ (1-x)^2(2x+1) & 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}, \quad \phi_2(x) = \begin{cases} (x+1)^2 x & -1 \le x \le 0 \\ (1-x)^2 x & 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}.$$

For $n \geq 1$, let V_n be the space of piecewise cubic splines $v \in C^1(0,1) \cap C[0,1]$ for which $v(0) = v(1) = 0$. The dimension of V_n is 2^{n+1} and the set

$$\Phi_n := \{\phi_1(2^n x - j) : j = 1, ..., 2^n - 1\} \cup \{\phi_2(2^n x - j)|_{[0,1]} : j = 0, ..., 2^n\}$$

is the basis for V_n. Let W_n be the complement of V_n in V_{n+1} then we have the following decomposition of the space $H_0^1(0,1)$

$$H_0^1(0,1) = V_1 + W_1 + W_2 + W_3 \ldots.$$

We follow the construction proposed in [5] and construct four wavelets. Wavelets from the space W_{n+1} are orthogonal to the scaling functions from the space V_n for $n \geq 1$. This property ensures that both the mass and stiffness matrix corresponding to the one-dimensional Laplacian have at most three wavelet blocks of nonzero elements in any column and then the number of nonzero elements in any column is bounded independent of the matrix size. The first two wavelets have supports in $[-1,1]$ and are uniquely determined by the above orthogonality condition and by imposing that the first one is odd and the second one is even. The second two wavelets have supports in $[-2,2]$. And we impose on them the orthogonality condition and again one of them should be odd and the second one even.

Then the space W_n is defined by

$$\begin{aligned}\Psi_n \;:=\; & \{\psi_1(2^n x - 2j - 1),\, \psi_2(2^n x - 2j - 1) : j = 0, ..., 2^{n-1} - 1\} \\ & \cup \{\psi_3(2^n x - 2j) : j = 1, ..., 2^{n-1} - 1\} \cup \{\psi_4(2^n x - 2j)|_{[0,1]} : j = 0, ..., 2^{n-1}\}.\end{aligned}$$

There remain several free parameters in the construction of the second two wavelets. In [5], these parameters were used to prescribe the orthogonality to the first two wavelets. We try to use these parameters to obtain better conditioned basis and alternatively also more sparse stiffness matrices. Some preliminary results are presented in the next section.

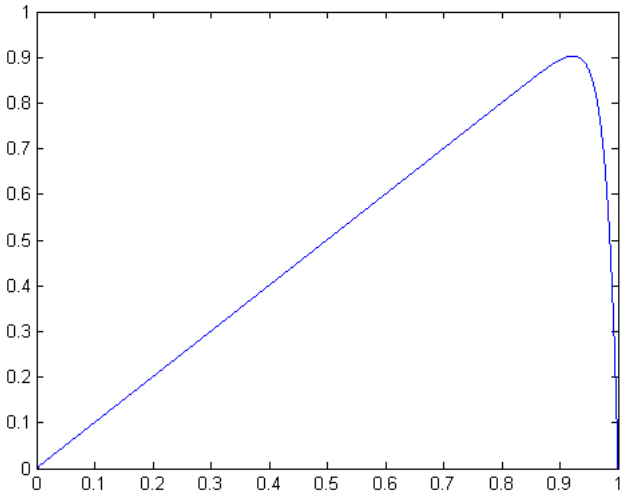
2 Numerical Experiments

We will limit ourselves to the equation $-pu'' + qu = f$ with the Dirichlet boundary conditions $u(0) = u(1) = 0$ and with positive constant coefficients. The corresponding Galerkin approximation problem is the following: Find $u_n = \sum_{i=1}^{2^{n+2}} c_i v_i$ such that

$$\int_0^1 p\, u_n' v' + q\, u_n v\, dx = \int_0^1 f v\, dx \qquad \forall v \in V_n.$$

To solve arising system of equations, we use the standard wavelet diagonal preconditioning and then apply the conjugate gradient method. The iterations are terminated if the difference of two consecutive iterations is less than $10^{-n-2}/\mathrm{cond}(A_n)$, where $\mathrm{cond}(A_n)$, denotes the condition number of the corresponding stiffness matrix.

First, we solve the above problem with $p = q = 1$ and with $p = 1$, $q = 0$. In both examples, the exact solution is $u = x(1 - e^{50x-50})$ which exhibits a steep gradient near the point 1. See Figure 1. Obtained results are summarized in Table 1. In this table, NIT represents the number of iterations, NEW denotes new wavelets, and finally DS denotes wavelets proposed in [5]. The achieved approximation error was the same in all cases.



Source: Own

Fig. 1. *The exact solution*

Tab. 1. *Obtained results*

		$p=1,\ q=0$		$p=q=1$	
n	$\|\|u_n-u\|\|_{L_2}$	NIT DS	NIT NEW	NIT DS	NIT NEW
1	6.0e-02	5	4	5	4
2	1.7e-02	7	6	7	6
3	2.6e-03	9	7	8	7
4	2.7e-04	10	9	10	9
5	2.3e-05	13	10	12	11
6	1.8e-06	14	12	14	12
7	1.2e-07	16	13	15	13
8	8.3e-09	17	15	16	15
9	5.3e-10	19	16	19	16
10	3.4e-11	19	17	19	18

Source: Own

Conclusion

The presented numerical results affirm that the proposed wavelet basis is better conditioned then the wavelet basis proposed in [5] which resulted in slightly smaller numbers of required iterations. Moreover, the numbers of nonzero elements in stiffness matrices corresponding to Poisson equation were substantially smaller. We believe that it will be possible to further improve the proposed construction.

Acknowledgments

D. Černá and V. Finěk have been supported by the project ESF No. CZ.1.07/2.3.00/09.0155 "Constitution and improvement of a team for demanding technical computations on parallel computers at TU Liberec" and V. Finěk and Z. Ondračková have been supported by the SGS

project “Construction of multiwavelets for numerical solution of differential equations”.

Literature

[1] ČERNÁ, D.; FINĚK, V.: *Approximate Multiplication in Adaptive Wavelet Methods*, submitted, 2010.

[2] COHEN, A.; DAHMEN W.; DEVORE, R.: Adaptive Wavelet Schemes for Elliptic Operator Equations - Convergence Rates, *Math. Comput.* 70, no. 233, pp. 27-75, 2001.

[3] COHEN, A.; DAHMEN W.; DEVORE, R.: Adaptive Wavelet Methods II - Beyond the elliptic case, *Found. Math.* 2, no. 3, pp. 203-245, 2002.

[4] DIJKEMA, T. J.; SCHWAB, Ch.; STEVENSON, R.: An adaptive wavelet method for solving high-dimensional elliptic PDEs, *Constructive approximation*, 30 (2009), no. 3, pp. 423-455.

[5] DIJKEMA, T. J.; STEVENSON, R.: A sparse Laplacian in tensor product wavelet coordinates. *Numer. Math.*, vol. 115, no. 3 (2010), pp. 433-449.

[6] STEVENSON, R.: Adaptive solution of operator equations using wavelet frames, *SIAM J. Numer. Anal.* 41, no. 3, pp. 1074-1100, 2003.

O ŘÍDKÉ REPREZENTACI LAPLACIÁNU

Jednou z nejdůležitějších částí waveletových adaptivních metod je efektivní přibližné násobení matice tuhosti s vektory ve waveletových souřadnicích. Přestože jsou známy algoritmy pro přibližné násobení s lineární složitostí, jejich aplikace je relativně časově náročná a jejich implementace velice obtížná. Proto je důležité vyvíjet dobře podmíněné waveletové báze, pro které jsou jak matice hmotnosti, tak matice tuhosti řídké ve smyslu, že počet nenulových prvků v libovolném sloupci je omezený konstantou. Potom je totiž možné násobit matici s vektorem přesně s lineární přesností. V tomto příspěvku prezentujeme waveletovou bázi adaptovanou na interval, vzhledem ke které jsou jak matice hmotnosti, tak matice tuhosti odpovídající jednodimenzionálnímu Laplaciánu řídké. Následně je rovněž matice tuhosti odpovídající n - dimenzionálnímu Laplaciánu ve waveletové bázi, která vznikne tenzorovým součinem jednodimenzionálních bází, řídká. Navíc jsou konstruované báze velmi dobře podmíněné. V tomto příspěvku krátce představíme tuto konstrukci a ukážeme několik numerických testů.

ÜBER DIE SELTENE REPRÄSENTATION VON LAPLAZIAN

Eine der der wichtigsten Teile der adaptiven Wavelet-Methoden besteht in der annähernden Multiplikation der Zähigkeitsmatrix mit Vektoren in den Wavelet-Koordinaten. Obschon Algorithmen fr eine annähernde Multiplikation mit linearer Komplexität bekannt sind, ist deren Anwendung zeitlich gesehen relativ aufwändig und ihre Implementierung sehr beschwerlich. Daher ist es wichtig, eine gut bedingte Wavelet-Basis zu entwickeln, für welche sowohl die Massen- als auch die Zähigkeitsmatrix insofern selten sind, als die Anzahl der Null-Elemente in einer beliebigen Kolumne durch die Konstante begrenzt ist. Hernach ist es nämlich möglich, die Matrix mit dem Vektor genau mit linearer Genauigkeit zu multiplizieren. In diesem Beitrag präsentieren wir die ans Intervall angepasste Wavelet-Basis, hinsichtlich derer sowohl die Massen- als auch die dem eindimensionalen Laplazian entsprechende Zähigkeitsmatrix selten sind. Auch die dem n-dimensionalen Laplazian in der Wavelet-Basis entsprechende Zähigkeitsmatrix, die durch das Tensorprodukt eindimensionaler Basen entsteht, ist selten. Darüber hinaus werden die konstruierten Basen sehr gut bedingt. In diesem Beitrag stellen wir kurz diese Konstruktion vor und zeigen einige numerische Tests.

O RZADKIEJ REPREZENTACJI LAPLASJANU

Jedną z najważniejszych części falkowych metod adaptacyjnych jest efektywne przybliżone mnożenie macierzy sztywności z wektorami we współrzędnych waveletowych. Chociaż znane są algorytmy służące do przybliżonego mnożenia o złożoności liniowej, ich stosowanie jest stosunkowo czasochłonne a ich wdrażanie bardzo trudne. Dlatego ważne jest opracowywanie dobrze uwarunkowanej bazy falkowej, dla której macierze masy, jak również macierze sztywności są rzadkie, to znaczy liczba elementów niezerowych w dowolnej kolumnie nie jest ograniczona stałą. Wówczas bowiem można macierz z wektorem mnożyć z dokładnością liniową. W niniejszym artykule zaprezentowano bazę falkową zaadaptowaną do interwału, w stosunku do której macierze masy, jak również macierze sztywności odpowiadające jednowymiarowemu laplasjanowi są rzadkie. Następnie także macierz sztywności odpowiadająca n-wymiarowemu lapsjanowi w bazie falkowej (waveletowej), która powstaje w wyniku iloczynu tensorowego baz jednowymiarowych, jest rzadka. Ponadto konstruowane bazy są bardzo dobrze uwarunkowane. W niniejszym artykule krótko przedstawiono taką konstrukcję oraz pokazano kilka testów numerycznych.

MULTIWAVELETS BASED ON HERMITE CUBIC SPLINES

Dana Černá
***Václav Finěk**
****Gerta Plačková**

Technical University of Liberec
Faculty of Science, Humanities and Education
Department of Mathematics and Didactics of Mathematics
Studentská 2, 461 17, Liberec, Czech Republic
dana.cerna@tul.cz

*Technical University of Liberec
Faculty of Science, Humanities and Education
Department of Mathematics and Didactics of Mathematics
Studentská 2, 461 17, Liberec, Czech Republic
vaclav.finek@tul.cz

**Technical University of Liberec
Faculty of Science, Humanities and Education
Department of Mathematics and Didactics of Mathematics
Studentská 2, 461 17, Liberec, Czech Republic
gerta.plackova@tul.cz

Abstract

First multiwavelets have appeared around the early 1990s. The basic idea behind multiwavelets is simple: to replace the single scaling function ϕ by the multiscaling function $\mathbf{\Phi}$ to have some additional desired properties. It seems to be an interesting trade off because multiwavelets provide higher order approximation with shorter support than single scaling function. Moreover, it is possible to have both symmetric and orthogonal multiwavelets while this is not possible for single wavelets. In recent years, several simple constructions of wavelet bases based on Hermite cubic splines were proposed. In this contribution, we shortly review these constructions, use these wavelets to solve numerically differential equations, and compare their performance.

Keywords: Wavelet; Hermite cubic splines; elliptic differential equations.

Introduction

A vector-valued function

$$\mathbf{\Phi}(x) = (\phi_1(x), \phi_2(x), \ldots, \phi_r(x))^T, \qquad \phi_1(x), \phi_2(x), \ldots, \phi_r(x) \in L^2(\mathbb{R})$$

will be called a multiscaling function if it satisfies the following refinement equation

$$\mathbf{\Phi}(x) = \sum_{k=s}^{t} \mathbf{H_k} \mathbf{\Phi}(ax-k),$$

where $s, t \in \mathbb{Z}$, $s < t$, $\mathbf{H_k}$ are some $r \times r$ matrices, $a \in \mathbb{N}$ and $a \geq 2$. The multiscaling function generates a multiresolution analysis of $L^2(\mathbb{R})$ in a similar way as for scalar wavelets. Subspaces V_j are defined

$$V_j = \mathrm{clos}_{L^2(\mathbb{R})}\{\phi_{l,j,k} = a^{j/2}\phi_l(a^j x - k) : 1 \leq l \leq r, k \in \mathbb{Z}\}, \qquad j \in \mathbb{Z}.$$

Then W_j, $j \in \mathbb{Z}$ denotes the orthogonal complementary subspaces of V_j in V_{j+1} and the vector-valued function

$$\mathbf{\Psi}(x) = (\psi_1(x), \psi_2(x), \ldots, \psi_{(a-1)r}(x))^T, \qquad \psi_1(x), \psi_2(x), \ldots, \psi_{(a-1)r}(x) \in L^2(\mathbb{R}).$$

Therefore there exist matrices $\mathbf{Q_k}$ such that

$$\mathbf{\Psi}(x) = \sum_{k=s}^{t} \mathbf{Q_k} \mathbf{\Phi}(ax-k),$$

and fast decomposition and reconstruction algorithms can be constructed like for scalar wavelets. Moreover, multiwavelets usually have shorter support than corresponding single wavelets, it is possible to construct both symmetric and orthogonal multiwavelets while this is not possible for single wavelets, and finally, it is possible to have a basis formed by functions with different orders (see Example 1). However, multiwavelets have also some disadvantages: the discrete multiwavelet transform usually requires preprocessing and postprocessing, and their construction is also more complicated. The reason, why preprocessing is necessary, is in the fact that for scalar wavelets, we have

$$2^{-n/2} f\left(2^{-n}k\right) \approx \int f(x)\, \phi_{n,k}(x)\, dx$$

and this does not hold in general for multiwavelets. Multiwavelets which do not require preprocessing are for instance the so called balanced multiwavelets. They were constructed in [7, 8]. For more details on multiwavelets, we refer to [6].

Example 1 *Simple example of piecewise linear multiwavelets is taken from [1]:*

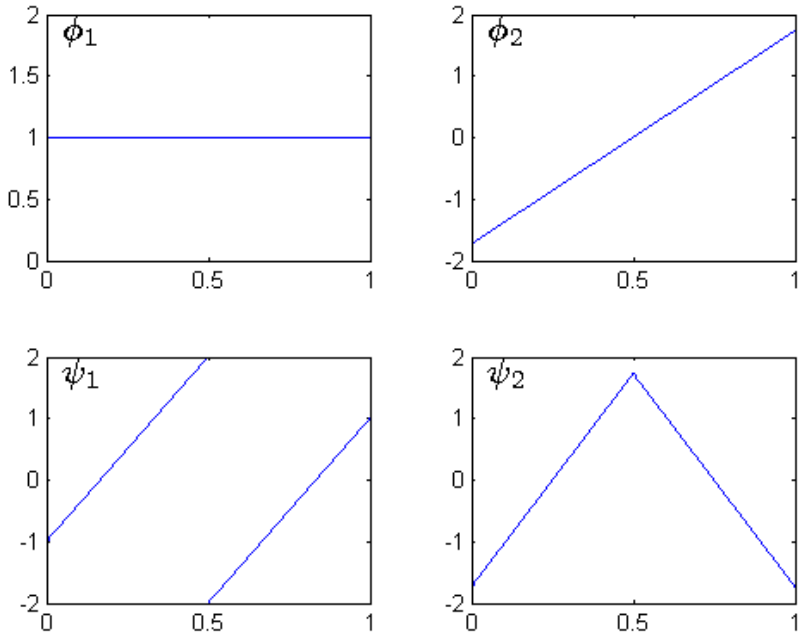
$$\phi_1(x) = \begin{cases} 1 & 0 \le x \le 1 \\ 0 & \textit{otherwise} \end{cases}, \qquad \phi_2(x) = \begin{cases} 2\sqrt{3}\left(x-\frac{1}{2}\right) & 0 \le x \le 1 \\ 0 & \textit{otherwise} \end{cases},$$

$$\psi_1(x) = \begin{cases} 6x-1 & 0 \le x < \frac{1}{2} \\ 6x-5 & \frac{1}{2} \le x \le 1 \\ 0 & \textit{otherwise} \end{cases}, \qquad \psi_2(x) = \begin{cases} 2\sqrt{3}\left(2x-\frac{1}{2}\right) & 0 \le x < \frac{1}{2} \\ -2\sqrt{3}\left(2x-\frac{3}{2}\right) & \frac{1}{2} \le x \le 1 \\ 0 & \textit{otherwise} \end{cases}.$$

The remainder of this paper is organized as follows. In the next section, we review several simple constructions of wavelets based on Hermite cubic splines. Specifically, wavelets with two vanishing wavelet moments proposed in [5], a hierarchical basis based on Hermite cubic splines, wavelets with four vanishing wavelet moments proposed in [2], and wavelets with respect to which both the mass and stiffness matrices corresponding to the one-dimensional Laplacian are sparse [3]. In the last section, we use these wavelets to solve numerically differential equations and compare their performance.

1 Hermite Cubic Spline Wavelets

In the year 2000, W. Dahmen et al. [4] proposed a construction of biorthogonal multiwavelets adapted to the interval $[0,1]$ on the basis of Hermite cubic splines. They started with Hermite cubic splines as the primal scaling bases on $\mathbb{R}$. Then, they constructed dual scaling bases on $\mathbb{R}$ consisting of continuous functions with small supports and with polynomial exactness of order 2. Consequently, they derived primal and dual boundary scaling functions retaining the polynomial exactness. This ensures vanishing moments of the corresponding wavelets. Finally,



Source: Own

Fig. 1. *Piecewise linear multiwavelets*

they applied the method of stable completions to construct the corresponding primal and dual multiwavelets on the interval.

These Hermite cubic splines are defined by

$$\phi_1(x) = \begin{cases} (x+1)^2(1-2x) & -1 \le x \le 0 \\ (1-x)^2(2x+1) & 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}, \phi_2(x) = \begin{cases} (x+1)^2 x & -1 \le x \le 0 \\ (1-x)^2 x & 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}.$$

They possess the following interpolation property:

$$\phi_1(0) = 1, \quad \phi_1'(0) = 0, \quad \phi_2(0) = 0, \quad \phi_2'(0) = 1$$

and then for any function $f \in C^1(\mathbb{R})$

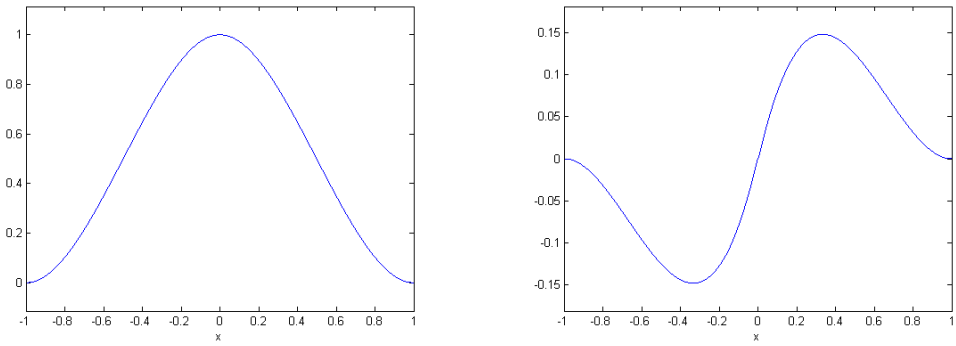
$$u := \sum_{j\in\mathbb{Z}} f(j)\phi_1(x-j) + \sum_{j\in\mathbb{Z}} f'(j)\phi_2(x-j)$$

is a Hermite interpolant to f on $\mathbb{Z}$.

1.1 Wavelets proposed by R. Q. Jia and S. T. Liu

An interesting construction was proposed in [5]. The authors constructed multiwavelets based on Hermite cubic splines with continuous derivatives and with a support on the interval $[-1,1]$. One of the constructed wavelets is symmetric, and the second one is antisymmetric. They also adapted them to the interval $[0,1]$ and their construction of boundary wavelets is very simple unlike the construction from [4]. In comparison with the semi-orthogonal wavelets, the wavelets at different levels are orthogonal with respect to $\langle u', v' \rangle$ instead of $\langle u, v \rangle$. They are given by

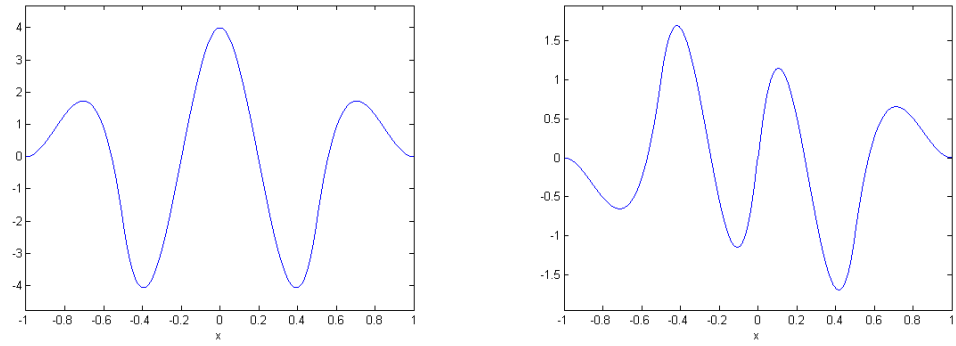
$$\psi_1(x) = -2\phi_1(2x+1) + 4\phi_1(2x) - 2\phi_1(2x-1) - 21\phi_2(2x+1) + 21\phi_2(2x-1),$$



ϕ_1 ϕ_2

Source: Own

Fig. 2. *The Hermite cubic spline*



ψ_1 ψ_2

Source: Own

Fig. 3. *Wavelets proposed by R. Q. Jia and S. T. Liu*

$$\psi_2(x) = \phi_1(2x+1) - \phi_1(2x-1) + 9\phi_2(2x+1) + 12\phi_2(2x) + 9\phi_2(2x-1).$$

Both wavelets are supported on $[-1,1]$ and its adaptation to the interval $[0,1]$ is performed by the restriction of the antisymmetric wavelet to the interval $[0,1]$. For $n \geq 1$, let V_n be the space of piecewise cubic splines $v \in C^1(0,1) \cap C[0,1]$ for which $v(0) = v(1) = 0$. The dimension of V_n is 2^{n+1} and the set

$$\Phi_n := \{\phi_1(2^n x - j) : \ j = 1, \ldots, 2^n - 1\} \cup \left\{\phi_2(2^n x - j)|_{[0,1]} : \ j = 0, \ldots, 2^n\right\}$$

is the basis for V_n. Let W_n be the complement of V_n in V_{n+1} with the basis defined by

$$\Psi_n := \{\psi_1(2^n x - j) : \ j = 1, \ldots, 2^n - 1\} \cup \left\{\psi_2(2^n x - j)|_{[0,1]} : \ j = 0, \ldots, 2^n\right\}.$$

Then, we have the following decomposition of $H_0^1(0,1)$

$$H_0^1(0,1) = V_1 + W_1 + W_2 + W_3 \ldots.$$

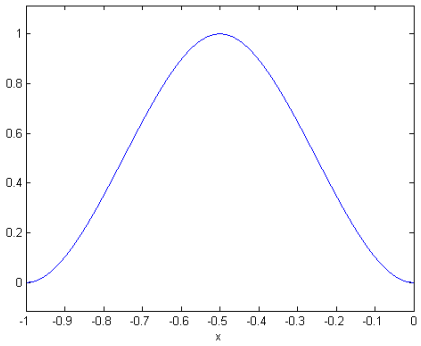
1.2 Hierarchical Basis

The second possibility to define complement wavelet spaces, we use the following hierarchical approach:

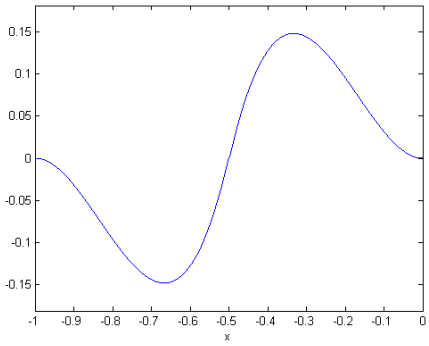
$$\psi_1(x) = \phi_1(2x+1) \quad \text{and} \quad \psi_2(x) = \phi_2(2x+1).$$

Both functions are supported on $[-1,0]$ and therefore no boundary functions are necessary. The complement space W_n is then given by

$$\Psi_n := \{\psi_1(2^n x - j) : j = 1,...,2^n\} \cup \{\psi_2(2^n x - j) : j = 1,...,2^n\}.$$



ψ_1



ψ_2

Source: Own

Fig. 4. *Hierarchical wavelets*

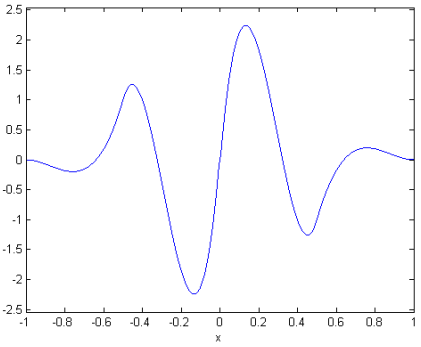
1.3 Wavelets with Moments

The third possibility is to define wavelets to have maximal number of vanishing moments for the given support $[-1,1]$. Wavelets are given then by

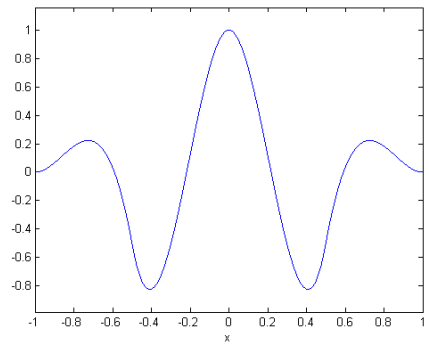
$$\psi_1(x) = \phi_1(2x+1) - \phi_1(2x-1) + \frac{39}{7}\phi_2(2x+1) + \frac{132}{7}\phi_2(2x) + \frac{39}{7}\phi_2(2x-1),$$

and

$$\psi_2(x) = -\frac{1}{2}\phi_1(2x+1) + \phi_1(2x) - \frac{1}{2}\phi_1(2x-1) - \frac{15}{4}\phi_2(2x+1) + \frac{15}{4}\phi_2(2x-1).$$



ψ_1



ψ_2

Source: Own

Fig. 5. *Wavelets with the maximal number of vanishing moments*

Boundary wavelets are constructed to have the same number of vanishing moments as inner wavelets. For more details, see [2]. Again let W_n be the complement of V_n in V_{n+1} with the basis defined by

$$\begin{aligned}\Psi_n \;:=\;& \{\psi_1(2^n x-j),\, \psi_2(2^n x-j) :\, j=1,...,2^n-1\}\\ & \cup\psi_3(2^n x-1)|_{[0,1]} \cup \psi_4(2^n(x-1)+1)|_{[0,1]}.\end{aligned}$$

1.4 Wavelets Proposed by T. J. Dijkema and R. Stevenson

Further construction was proposed in [3]. They first proved that it is not possible to construct continuous piecewise smooth wavelets which have a compact support, form Riesz basis for $L^2(I)$, properly scaled wavelets form the Riesz basis for $H^1(I)$, and with

$$\left\langle \psi'_\lambda, \psi'_\mu \right\rangle = 0, \quad \text{and} \quad \left\langle \psi_\lambda, \psi_\mu \right\rangle = 0 \quad \forall \lambda, \mu.$$

Consequently, they constructed cubic Hermite wavelets with the continuous first derivative such that among the primal multiresolution spaces V_j and the dual multiresolution spaces $\tilde{V}_j$ the following relation holds

$$V_j + V_j'' \subset \tilde{V}_{j+1}.$$

As a consequence

$$\left\langle \psi'_\lambda, \psi'_\mu \right\rangle = 0, \quad \left\langle \psi_\lambda, \psi_\mu \right\rangle = 0, \quad \forall \lambda, \mu : |\lambda| > |\mu| + 1.$$

Their wavelets are then given by

$$\psi_1(x) = \phi_1(2x+1) - \phi_1(2x-1) + \frac{39}{7}\phi_2(2x+1) + \frac{132}{7}\phi_2(2x) + \frac{39}{7}\phi_2(2x-1),$$

$$\psi_2(x) = -\frac{1}{2}\phi_1(2x+1) + \phi_1(2x) - \frac{1}{2}\phi_1(2x-1) - \frac{15}{4}\phi_2(2x+1) + \frac{15}{4}\phi_2(2x-1),$$

$$\psi_3(x) = \sum_{i=-3}^{3} c_i\phi_1(2x-i) + d_i\phi_2(2x-i), \quad \psi_4(x) = \sum_{i=-3}^{3} e_i\phi_1(2x-i) + f_i\phi_2(2x-i)$$

with

$$c = \left[-\frac{4595}{13728}, \frac{7}{65}, -\frac{18737}{68640}, 1, -\frac{18737}{68640}, \frac{7}{65}, -\frac{4595}{13728}\right],$$

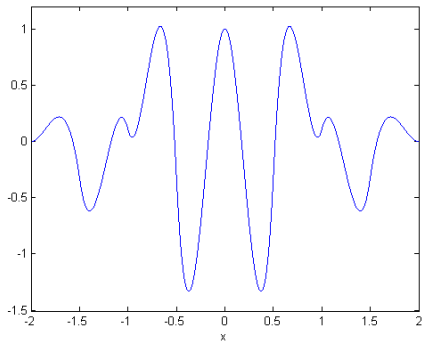
$$d = \left[-\frac{68741}{22880}, -\frac{69}{40}, -\frac{204701}{22880}, 0, \frac{204701}{22880}, \frac{69}{40}, \frac{68741}{22880}\right],$$

$$e = \left[\frac{417}{22880}, -\frac{7}{2340}, \frac{5443}{205920}, 0, -\frac{5443}{205920}, \frac{7}{2340}, -\frac{417}{22880}\right],$$

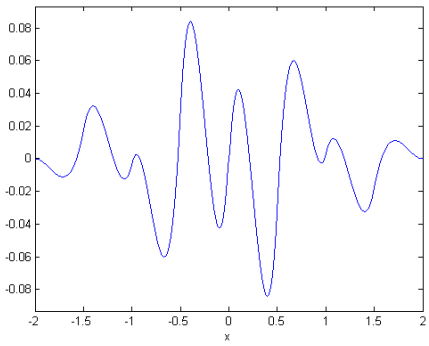
$$f = \left[\frac{723}{4576}, \frac{1}{8}, \frac{8153}{13728}, \frac{1}{2}, \frac{8153}{13728}, \frac{1}{8}, \frac{723}{4576}\right].$$

Again W_n is the complement of V_n in V_{n+1} with the basis defined by

$$\begin{aligned}\Psi_n \;:=\;& \left\{\psi_1(2^n x-2j-1),\, \psi_2(2^n x-2j-1) :\, j=0,...,2^{n-1}-1\right\}\\ & \cup\left\{\psi_3(2^n x-2j) :\, j=1,...,2^{n-1}-1\right\} \cup \left\{\psi_4(2^n x-2j)|_{[0,1]} :\, j=0,...,2^{n-1}\right\}.\end{aligned}$$



ψ_3



ψ_4

Source: Own

Fig. 6. *Wavelets proposed by T. J. Dijkema and R. Stevenson*

2 Numerical Experiments

In this section, the wavelets introduced in the previous section are used to solve numeri-cally differential equations. We will restrict ourselves to the equation $-pu'' + qu = f$ with the Dirichlet boundary conditions $u(0) = u(1) = 0$ and with positive constant coefficients. The corresponding Galerkin approximation problem is the following: Find $u_n = \sum_{i=1}^{2^{n+2}} c_i v_i$ such that

$$\int_0^1 p\,u_n' v' + q\,u_n v\,dx = \int_0^1 f v\,dx \qquad \forall v \in V_n.$$

By the Lax-Milgram lemma, this approximation problem has the unique solution. We also use the standard wavelet preconditioning consisting in normalizing each basis function with respect to the above bilinear form. To solve the arising system of linear equations, we use the conjugate gradient method. The iterations are terminated if the difference of two consecutive iterations is less than $10^{-n-2}/\mathrm{cond}(A_n)$, where $\mathrm{cond}(A_n)$ denotes the condition number of the corresponding stiffness matrix.

Tab. 1. *Obtained results for the problem with* $p = q = 1$.

		JL		H		M		DS	
n	$\|\|u_n - u\|\|_{L_2}$	NZ	IT	NZ	IT	NZ	IT	NZ	IT
1	6.0e-02	48	4	40	7	54	7	54	5
2	1.7e-02	172	5	128	13	162	11	154	7
3	2.6e-03	552	6	368	20	418	16	410	9
4	2.7e-04	1580	7	976	31	1010	21	986	10
5	2.3e-05	4168	8	2448	41	2306	24	2202	13
6	1.8e-06	10396	10	5904	52	5042	25	4698	14
7	1.2e-07	24936	10	13840	74	10690	28	9754	16
8	8.3e-09	58283	11	31760	84	22194	31	19967	17
9	5.4e-10	135302	12	71696	110	46963	33	41627	19

Source: Own

Tab. 2. *Obtained results for the problem with $p = 1$ and $q = 0$.*

		JL		H		M		DS	
n	$\|\|u_n - u\|\|_{L_2}$	NZ	IT	NZ	IT	NZ	IT	NZ	IT
1	6.0e-02	24	3	36	7	50	7	54	5
2	1.7e-02	56	5	116	12	154	11	154	7
3	2.6e-03	128	6	340	20	406	16	410	8
4	2.7e-04	280	6	916	31	994	20	986	10
5	2.3e-05	592	8	2324	41	2286	24	2202	12
6	1.8e-06	1224	10	5652	52	5018	25	4698	14
7	1.2e-07	2847	10	13332	69	10677	28	9762	15
8	8.3e-09	8818	10	30740	84	22347	30	20019	16
9	5.4e-10	35863	12	69652	110	46613	33	41421	19

Source: Own

First we solve the above problem with $p = q = 1$ and the exact solution $u = x(1 - e^{50x-50})$ which exhibits a steep gradient near the point 1. Results are summarized in Table 1. Then we solve the above problem with $p = 1,\ q = 0$ and with the exact solution $u = x(1 - e^{50x-50})$ which exhibits a steep gradient near the point 1. Results are summarized in Table 2. In both tables, NZ is the number of nonzero elements in stiffness matrices, IT represents the number of iterations, JL denotes wavelets proposed in [5], H denotes hierarchical basis, M denotes wavelets proposed in [2], and finally DS denotes wavelets proposed in [3]. Achieved approximation error was the same for all bases.

Conclusion

Presented results affirm that wavelets proposed in [5] have excellent condition number, and especially for the Poisson equation, the arising stiffness matrices are very sparse. Wavelets proposed in [3] are best suited for general differential equations with constant coefficients. Results of tested hierarchical basis confirm the well known fact that these basis do not form the Riesz basis and therefore they need some additional preconditioning. Concerning the basis proposed in [2], we suppose that it will be better (than tested bases) suited for non-constant differential equations solved adaptively. Presented results also suggest that there is probably some space to improve the condition number of the basis proposed in [2].

Acknowledgments

This work has been supported by the project ESF No. CZ.1.07/2.3.00/09.0155 “Constitution and improvement of a team for demanding technical computations on parallel computers at TU Liberec”.

Literature

[1] ALPER, B. K.; ROKHLIN, V.: A fast algorithm for the evaluation of Legendre expansions. *Siam J. Sci. Statist. Comput.*, 12 (1991), pp. 158-179.

[2] ČERNÁ, D.; FINĚK, V.: *Hermite cubic spline wavelets with moments on the interval*. In preparation, 2012.

[3] DIJKEMA, T. J.; STEVENSON, R.: A sparse Laplacian in tensor product wavelet coordinates. *Numer. Math.*, vol. 115, no. 3 (2010), pp. 433-449.

[4] DAHMEN, W.; HAN, B.; JIA, R.-Q.; KUNOTH, A.: Biorthogonal Multiwavelets on the Interval: Cubic Hermite Splines. *CONSTR. APPROX.*, vol. 16, (1998), pp. 221-259.

[5] JIA, R.-Q.; LIU, S.-T.: Wavelet bases of Hermite cubic splines on the interval. *Adv. in Comp. Math.*, 25 (2006), pp. 23-39.

[6] KEINERT, F.: *Wavelets and Multiwavelets*. CHAPMAN & HALL/CRC, 2004.

[7] LEBRUN, J.; VETTERLI, M.: Balanced multiwavelets theory and design. *IEEE Trans. Signal Process.*, 46 (1998), pp. 1119-1125.

[8] LEBRUN, J.; VETTERLI, M.: High-order balanced multiwavelets: theory, factorization, and design. *IEEE Trans. Signal Process.*, 49 (2001), pp. 1918-1930.

MULTIWAVELETY ZALOŽENÉ NA HERMITOVSKÝCH KUBICKÝCH SPLINECH

První multiwaveletová báze se objevila kolem roku 1990. Základní myšlenka multiwaveletů je jednoduchá: nahradíme jednu škálovou funkci jednou multiškálovou, abychom získali lepší vlastnosti bázových funkcí. Multiwavelety totiž umožňují aproximace vyššího řádu s kratším nosičem než klasické škálové funkce. Navíc je možné zkonstruovat symetrické ortogonální multiwavelety, což pro klasické wavelety není možné. V minulých letech se objevilo několik jednoduchých konstrukcí waveletových bází založených na hermitovských kubických splinech. V tomto příspěvku stručně představíme tyto konstrukce, použijeme tyto wavelety k numerickému řešení diferenciálních rovnic a srovnáme jejich vlastnosti.

AUF KUBISCHEN HERMITE-SPLINES ANGELEGTE MULTIWAVELETS

Die erste Multiwaveletbasis erschien um das Jahr 1990. Der Grundgedanke der Multiwavelets ist einfach: Wir ersetzen eine Skalenfunktion durch eine Multiskalenfunktion, um bessere Eigenschaften von Basenfunktionen zu erhalten. Die Multiwavelets ermöglichen nämlich eine Annäherung höherer Ordnung mit einem kürzeren Träger als die klassischen Skalenfunktionen. Außerdem ist es möglich, symmetrische orthogonale Multiwavelets zu konstruieren, was für die klassischen Wavelets nicht möglich ist. In den vergangenen Jahren erschienen einige einfache Konstruktionen von Wavelet-Basen, die auf kubischen Hermite-Splinen beruhen. In diesem Beitrag stellen wir kurz diese Konstruktionen vor. Wir benutzen die Wavelets zur numerischen Lösung von Differenzialgleichungen und vergleichen deren Eigenschaften.

FALKI WIELOKROTNE OPARTE NA HERMITOWSKICH SPLAJNACH KUBICZNYCH

Pierwsza falka wielokrotna (ang. multiwavelets) pojawiła się około 1990 roku. Podstawowa idea falek wielokrotnych jest prosta: zastępujemy jedną funkcję skalującą jedną wieloskalującą, aby pozyskać lepsze cechy funkcji bazowych. Falki wielokrotne umożliwiają bowiem aproksymację wyższego rzędu z krótszą bazą w porównaniu z klasyczną funkcją skalującą. Ponadto możliwe jest skonstruowanie symetrycznych ortogonalnych falek wielokrotnych, co nie jest możliwe w przypadku klasycznych falek. W poprzednich latach pojawiło się kilka prostych konstrukcji baz falkowych opartych na hermitowskich splajnach kubicznych. W niniejszym artykule krótko przedstawiono takie konstrukcje, omawiane falki zastosowano do numerycznego rozwiązywania równań różniczkowych oraz porównano ich cechy.

hp-DISCONTINUOUS GALERKIN METHOD FOR NONLINEAR PROBLEMS

Vít Dolejší

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

Abstract

We deal with a numerical solution of nonlinear convection-diffusion problems with the aid of the discontinuous Galerkin finite element (DGFE) method. We propose a new *hp*-adaptation technique, which is based on a combination of a residuum-nonconformity estimator and a regularity indicator. The residuum-nonconformity estimator consists of two building blocks (the residuum error indicator and the value of the nonconformity). The estimator marks mesh elements for a refinement. The regularity indicator decides if the marked elements will be refined by *h*- or *p*-technique. The residuum-nonconformity estimator as well as the regularity indicator are easily computable quantities. Moreover, the same technique estimates an algebraic error arising from an iterative solution of the corresponding nonlinear algebraic system. The performance of the proposed *hp*-DGFE method is demonstrated by several numerical examples.

Keywords: *hp*-discontinuous Galerkin finite element method; residuum-nonconformity indicator; regularity estimator; algebraic error.

Introduction

Our aim is to develop a sufficiently robust, efficient and accurate numerical scheme for the simulation of viscous compressible flows. The *discontinuous Galerkin finite element* (DGFE) methods have become very popular numerical techniques for the solution of the compressible Navier-Stokes equations. Recent progress of the use of the DG method for compressible flow simulations can be found in [11].

In this paper, we deal with a model problem represented by a scalar nonlinear convection-diffusion equation. We focus on a *hp*-adaptation strategy of DGFE methods which significantly increase the accuracy and efficiency of the computation, see, e.g. [2, 6, 8, 12, 13].

The proposed strategy is based on a combination of a residuum-nonconformity estimator and a regularity indicator. The *residuum-nonconformity estimator* gives a lower estimate of the error measure consisting of the error measured in a dual norm and the quantity measuring a violation of the conformity of the solution. This estimator is locally defined for each mesh element, it is easily computable and its implementation is very simple. The *regularity indicator* is based on the integration of interelement jumps of the approximate solution over the element boundary. Taking into account results from a priori error analysis, we define the regularity indicator. If this value is smaller than one then we apply a *p*-refinement otherwise we use a *h*-refinement. Both refinements (*h* and *p*) are only isotropic, an anisotropic adaptation will be a subject of the further research.

1 Problem Formulation

1.1 Governing Equations

We consider a formal nonlinear *partial differential equation*

$$\begin{aligned} Lu &= 0 \quad \text{in } \Omega, && \text{(1a)}\\ u &= u_D \quad \text{on } \partial\Omega_D, && \text{(1b)}\\ Nu &= g_N \quad \text{on } \partial\Omega_N, && \text{(1c)} \end{aligned}$$

where $u : \Omega \to \mathbb{R}$ is the unknown scalar function defined on $\Omega \in \mathbb{R}^d$, $d = 2,3$, L is a formal second order differential operator and (1b) and (1c) formally represent the Dirichlet and Neumann boundary conditions on the parts of boundary $\partial\Omega_D$ and $\partial\Omega_N$, respectively. We assume that there exits a unique weak solution of (1) which we denote again by u.

Let $\mathscr{T}_h$ ($h > 0$) be a partition of the closure $\overline{\Omega}$ of the domain Ω into a finite number of closed d-dimensional simplicies K with mutually disjoint interiors. We call $\mathscr{T}_h = \{K\}_{K\in\mathscr{T}_h}$ a *triangulation* of Ω and do not require the conforming properties from the finite element method.

Over the triangulation $\mathscr{T}_h$ we define the so-called *broken Sobolev space*

$$H^s(\Omega,\mathscr{T}_h) := \{v; v|_K \in H^s(K)\ \forall K \in \mathscr{T}_h\}, \quad s \geq 0 \tag{2}$$

with the seminorm $|v|_{H^s(\Omega,\mathscr{T}_h)} := \left(\sum_{K\in\mathscr{T}_h} |v|^2_{H^s(K)}\right)^{1/2}$, where $|\cdot|_{H^s(K)}$ denotes the seminorm of the Sobolev space $H^s(K)$, $K \in \mathscr{T}_h$.

Moreover, to each $K \in \mathscr{T}_h$, we assign a positive integer p_K (=local polynomial degree). Then we define the set

$$\mathsf{p}_h := \{p_K, K \in \mathscr{T}_h\}. \tag{3}$$

We define the finite dimensional subspace of $H^1(\Omega,\mathscr{T}_h)$ which consists of discontinuous piecewise polynomial functions associated with the vector p_h by

$$S_{h\mathsf{p}} = \{v;\ v \in L^2(\Omega),\ v|_K \in P_{p_K}(K)\ \forall K \in \mathscr{T}_h\}, \tag{4}$$

where $P_{p_K}(K)$ denotes the space of all polynomials on K of degree $\leq p_K$, $K \in \mathscr{T}_h$.

We introduce a formal discretization of (1) with the aid of DGFE method. Hence, let

$$\tilde{c}_h(u,v) : H^2(\Omega,\mathscr{T}_h) \times H^2(\Omega,\mathscr{T}_h) \to \mathbb{R} \tag{5}$$

be the corresponding form which is nonlinear with respect to its first argument and linear with respect to the second one. We say that the function $u_h \in S_{h\mathsf{p}}$ is an *approximate solution* of (1), if

$$\tilde{c}_h(u_h, v_h) = 0 \qquad \forall v_h \in S_{h\mathsf{p}}. \tag{6}$$

Moreover, we define the form $\mathscr{N}_h : H^1(\Omega,\mathscr{T}_h) \to \mathbb{R}$ by

$$\mathscr{N}_h(v) := \left(2\sum_{\Gamma\in\mathscr{F}_h^I}\int_\Gamma h_\Gamma^{-1}[\![v]\!]^2\,\mathrm{d}S + \sum_{\Gamma\in\mathscr{F}_h^D}\int_\Gamma h_\Gamma^{-1}(v-u_D)^2\,\mathrm{d}S\right)^{1/2}, \tag{7}$$

where u_D is from (1b), $\mathscr{F}_h^I$ denotes the set of all interior faces of $\mathscr{T}_h$, $\mathscr{F}_h^D$ denotes the set of all faces of $\mathscr{T}_h$ lying on $\partial\Omega_D$, $[\![\cdot]\!]$ is the jump of a function from $H^1(\Omega,\mathscr{T}_h)$ and h_Γ is the diameter of a face Γ.

Finally, we characterise the weak solution of (1).

Lemma 1.1 *The following implications are valid:*

i) Let $u \in H^2(\Omega)$ *be the weak solution of* (1) *then*

$$\tilde{c}_h(u,v) = 0 \quad \forall v \in H^2(\Omega,\mathscr{T}_h), \tag{8a}$$
$$\mathscr{N}_h(u) = 0. \tag{8b}$$

ii) If $u \in H^2(\Omega,\mathscr{T}_h)$ *satisfies both conditions of* (8) *then u is the weak solution of* (1).

The discrete problem (6) represents a system of $N_h = \dim S_{h\mathsf{p}}$ nonlinear algebraic equations. We solve it with the aid of a Newton-like iterative method which gives the solution $\tilde{u}_h \in S_{h\mathsf{p}}$ such that $\tilde{c}_h(\tilde{u}_h, v_h) \approx 0 \; \forall v_h \in S_{h\mathsf{p}}$.

2 Residuum Estimator

In this section we investigate the discretization error $u - u_h$ and the algebraic error $\tilde{u}_h - u_h$ in a suitable (dual) norm and define estimators giving some information about these errors.

2.1 Error Measure

Similarly as in [3], our proposed error measure consists of *two building blocks*, which are motivated by Lemma 1.1, namely relations (8a) and (8b). Let $X := H^2(\Omega,\mathscr{T}_h)$ and $\|\cdot\|_X$ is a norm defined on X, which will be specified later. The *first building block* is given by

$$\mathscr{R}_h(u_h) := \sup_{0\neq v\in X} \frac{\tilde{c}_h(u_h,v)}{\|v\|_X} \tag{9}$$

which defines the *residuum error in the dual norm* of the approximate solution $u_h \in S_{h\mathsf{p}} \subset X$ and it measures a violation of (8a). However, it is impossible to evaluate $\mathscr{R}_h(u_h)$, since the supremum is taken over an infinite-dimensional space. Therefore, in our approach, we seek the maximum over some sufficiently large but finite dimension subspace of X, which is presented in Section 2.2.

The *second building block* is based on (8b), which characterises a violation of the conformity of the weak solution and a violation of the Dirichlet boundary condition. It is represented by the value $\mathscr{N}_h(u_h) \geq 0$ given by (7) which we call the *nonconformity* of the approximate solution. In contrast to $\mathscr{R}_h(u_h)$ the quantity $\mathscr{N}_h(u_h)$ is directly computable from (7). Finally, our *error measure* is the sum of squares of the *residuum error* and *nonconformity*, i.e.,

$$\mathscr{E}_h(u_h) := (\mathscr{R}_h(u_h)^2 + \mathscr{N}_h(u_h)^2)^{1/2}. \tag{10}$$

Due to Lemma 1.1 we simply observe that $\mathscr{E}_h(u_h) = 0$ if and only if $u_h = u$.

2.2 Global and Element Residuum Estimators

In the previous section, we introduced the error measure $\mathscr{E}_h(u_h) = \sqrt{\mathscr{R}_h(u_h)^2 + \mathscr{N}_h(u_h)^2}$ of the approximate solution $u_h \in S_{h\mathsf{p}} \subset X$. Whereas $\mathscr{N}_h(u_h)$ is easy to evaluate, the quantity $\mathscr{R}_h(u_h)$ has to be approximated in a suitable way, which is presented in this section. For each $K \in \mathscr{T}_h$ and each integer $p \geq 0$, we define the spaces

$$S_K^p := \{\phi_h \in X,\ \phi_h|_K \in P^p(K),\ \phi_h|_{\Omega\setminus K} = 0\} \tag{11}$$

and

$$S_{hp}^{+} := \{\phi \in X; \phi = \sum_{K\in\mathscr{T}_h} c_K \phi_K,\ c_K \in \mathbb{R},\ \phi_K \in S_K^{p_K+1},\ K \in \mathscr{T}_h\}. \tag{12}$$

Obviously, $S_{hp} \subset S_{hp}^{+} \subset X$.

Now, we define the *element residuum estimator*

$$\rho_{h,K}(u_h) := \sup_{0\neq\psi_h\in S_K^{p_K+1}} \frac{\tilde{c}_h(u_h,\psi_h)}{\|\psi_h\|_X} = \sup_{\psi_h\in S_K^{p_K+1},\|\psi_h\|_X=1} \tilde{c}_h(u_h,\psi_h), \quad u_h \in X, \tag{13}$$

for each $K \in \mathscr{T}_h$ and the *global residuum estimator*

$$\rho_h(u_h) := \sup_{0\neq\psi_h\in S_{hp}^{+}} \frac{\tilde{c}_h(u_h,\psi_h)}{\|\psi_h\|_X} = \sup_{\psi_h\in S_{hp}^{+},\|\psi_h\|_X=1} \tilde{c}_h(u_h,\psi_h), \quad u_h \in X, \tag{14}$$

which are easily computable quantities if $\|\cdot\|_X$ is suitably chosen, see Section 2.4.

Obviously, if $u \in X$ is the exact solution of (1) then consistency (8a) implies $0 = \rho_h(u) = \rho_{h,K}(u)$, $K \in \mathscr{T}_h$. Moreover, we have immediately a lower bound

$$\rho_h(u_h) \le \mathscr{R}_h(u_h), \tag{15}$$

since ρ_h is the supremum over subspace $S_{hp}^{+} \subset X$. However, it is open if there exists an upper bound, i.e., $\mathscr{R}_h(u_h) \le C\rho_h(u_h)$, where $C > 0$. This will be the subject of a further research.

2.3 Algebraic Residuum Estimators

Similarly as in the previous section, we define the estimator corresponding to the *algebraic error residuum*. Let $\tilde{u}_h \in S_{hp}$ be the output of the iterative process used for the solution of (6). We define the *algebraic residuum estimator*

$$\rho_h^A(\tilde{u}_h) := \sup_{0\neq\psi_h\in S_{hp}} \frac{\tilde{c}_h(\tilde{u}_h,\psi_h)}{\|\psi_h\|_X} = \sup_{\psi_h\in S_{hp},\|\psi_h\|_X=1} \tilde{c}_h(u_h,\psi_h), \tag{16}$$

which measures the algebraic error (the difference between $\tilde{u}_h$ and u_h), since $\rho_h^A(u_h) = 0$ due to (6). The relations (14) and (16) give $\rho_h^A(\tilde{u}_h) \le \rho_h(\tilde{u}_h)$, since in (14) the supremum is taken over the larger space.

We stop the iterative process for the solution of the nonlinear algebraic problem (6) if

$$\rho_h^A(\tilde{u}_h) \le \beta\rho_h(\tilde{u}_h), \tag{17}$$

where $\beta \in (0,1)$. In numerical experiments we put $\beta \approx 0.01$.

2.4 Choice of the Norm $\|\cdot\|_X$

In order to ensure a fast evaluation of estimators ρ_h and ρ_h^A, we need to choose the norm $\|\cdot\|_X$ in a suitable way. We employ

$$\|\cdot\|_X := \left(\delta\|\cdot\|^2_{L^2(\Omega)} + \varepsilon|\cdot|^2_{H^1(\Omega,\mathscr{T}_h)}\right)^{1/2}, \tag{18}$$

which guarantees that

$$\rho_h(u_h)^2 = \sum_{K\in\mathscr{T}_h} \rho_{h,K}(u_h)^2. \tag{19}$$

Here δ and ε denote the size of "convection" and "diffusion", respectively.

Therefore, it is sufficient to evaluate the element residuum estimators $\rho_{h,K}$, $K \in \mathscr{T}_h$. This is a standard task of seeking a constrain extrema over $S_K^{p_K+1}$ with the constraint $\|\psi_h\|_X = 1$. This can be done directly very fast since the dimension of $S_K^{p_K+1}$, $K \in \mathscr{T}_h$ is small.

2.5 Residuum-Nonconformity Estimators

We have already mentioned that the second building block of the error measure is given by the nonconformity $\mathscr{N}_h(u_h)$ defined by (7). For the purpose of a mesh adaptation, we define its local variant

$$\mathscr{N}_{h,K}(v) := \left(\sum_{\Gamma \in \mathscr{F}_h^I \cap \partial K} \int_\Gamma h_\Gamma^{-1} [\![v]\!]^2 \,\mathrm{d}S + \sum_{\Gamma \in \mathscr{F}_h^D \cap \partial K} \int_\Gamma h_\Gamma^{-1} (v - u_D)^2 \,\mathrm{d}S \right)^{1/2}, \quad v \in H^1(\Omega, \mathscr{T}_h). \tag{20}$$

Obviously, from (7) and (20), we have $\mathscr{N}_h(v)^2 = \sum_{K \in \mathscr{T}_h} \mathscr{N}_{h,K}(v)^2$. Finally, we define the *local* and *global residuum-nonconformity estimators* of the approximate solution $u_h \in S_{hp}$ by

$$\eta_{h,K}(u_h) := \left(\rho_{h,K}(u_h)^2 + \mathscr{N}_{h,K}(u_h)^2\right)^{1/2}, \quad K \in \mathscr{T}_h, \tag{21a}$$

and

$$\eta_h(u_h) := \left(\rho_h(u_h)^2 + \mathscr{N}_h(u_h)^2\right)^{1/2} = \left(\sum_{K \in \mathscr{T}_h} \eta_{h,K}(u_h)^2 \right)^{1/2}, \tag{21b}$$

respectively. In virtue of (10), (15) and (21), we expect that the global residuum-nonconformity estimator $\eta_h(u_h)$ approximates the error measure $\mathscr{E}_h(u_h)$. In the following we introduce an adaptation technique which produces a *hp*-mesh and the corresponding approximate solution such that the estimator $\eta_h(u_h)$ is under a given tolerance.

3 *hp*-Adaptation Process

In this section, we present a new *hp*-adaptive DG technique for the solution of (6). In Section 2, we defined the element and global residuum-nonconformity estimators $\eta_{h,K}$ and η_h, respectively. We employ the norm $\|\cdot\|_X$ given by (18) which guarantees that equality (19) is valid. As already mentioned, our interest is to find the solution $\tilde{u}_h \in S_{hp}$ such that

$$\eta_h(\tilde{u}_h) \le \omega, \tag{22}$$

where $\omega > 0$ is a given tolerance.

Let $\mathscr{T}_h$ be a given mesh and $\tilde{u}_h$ the corresponding approximation of (6). We require that

$$\eta_{h,K}(\tilde{u}_h) \le \frac{\omega}{\sqrt{\#\mathscr{T}_h}} \qquad \forall K \in \mathscr{T}_h, \tag{23}$$

where $\#\mathscr{T}_h$ denotes the number of elements of $\mathscr{T}_h$. Obviously, if (23) is satisfied then, due to (21b), condition (22) is valid and the adaptation process stops. Otherwise, we mark for refinement all $K \in \mathscr{T}_h$ violating (23).

Furthermore, all marked elements will be refined either by *h*- or by *p*-adaptation, namely, either we split a given mother element K into four daughter elements or we increase the degree of polynomial approximation for a given element. Thus the new mesh $\mathscr{T}_{\hat{h}}$ and new set $\{\hat{p}_K, K \in \mathscr{T}_{\hat{h}}\}$ are created. We interpolate the old solution on a new mesh and perform the next adaptation step till (22) is valid.

3.1 Regularity Indicator

The estimation of the regularity of the solution is an essential key of any *hp*-adaptation strategy. Our approach is based on a measure of inter-element jumps which is the base of the jump indicator from [5] and the shock capturing technique from [7].

We propose the *regularity indicator*

$$g_K(u_h) := \frac{\int_{\partial K\cap\Omega}[\![u_h]\!]^2\,\mathrm{d}S}{|K|h_K^{2p_K-2}}, \quad K\in\mathscr{T}_h, \tag{24}$$

where $|K|$ is the area of $K\in\mathscr{T}_h$. If the exact solution is sufficiently regular, i.e., $s_K \geq p_K+1$, then $g_K(u_h)\approx O\left(h_K^{2p_K+1}/h_K^2 h_K^{2p_K-2}\right) = O(h_K^1)$. On the other hand, if the exact solution is not sufficiently regular, i.e., $s_K < p_K+1$ ($\Leftrightarrow s_K \leq p_K$), then $g_K(u_h)\approx O\left(h_K^{2s_K-1}/h_K^2 h_K^{2p_K-2}\right) = O(h_K^{2\delta-1})$, where $\delta = s_K - p_k \leq 0$. Then we use the following strategy

$$\begin{array}{ll} g_K(u_h)\leq 1 & \Rightarrow \text{solution is regular} \quad \Rightarrow \text{p-refinement},\\ g_K(u_h)>1 & \Rightarrow \text{solution is irregular} \quad \Rightarrow \text{h-refinement}, \end{array} \qquad K\in\mathscr{T}_h. \tag{25}$$

4 Numerical Experiments

In the previous sections, we introduced and developed the adaptive *hp*-DGFE method. We demonstrate its performance in this section by several numerical examples. Let $\tilde{u}_h$ be the approximate solution resulting from an iterative method, i.e., the solution influenced by the algebraic error. We deal with two following numerical examples:

(E1) nonlinear convection-diffusion equation with a corner singularity from [9],

(E2) linear convection-diffusion equation with the strong interior layer and the exponential boundary layer from [10].

For the first one, we know the exact solution and therefore we are able to evaluate the computational error. We carried out a *hp*-adaptive algorithm starting on a mesh with the step h_0 and P_1 polynomial approximation. We evaluate $\|u-\tilde{u}_h\|_X$, $\mathscr{N}_h(\tilde{u}_h)$ and $\rho_h(\tilde{u}_h)$ with the corresponding experimental orders of convergence (EOC) with respect to the number of degree of freedom N_h defined by

$$\mathrm{EOC} = \frac{\log e_{h_{l+1}} - \log e_{h_l}}{\log(1/\sqrt{N_{h_{l+1}}}) - \log(1/\sqrt{N_{h_l}})}, \qquad l=1,2,\ldots, \tag{26}$$

Moreover, we evaluate the *"effectivity index"*

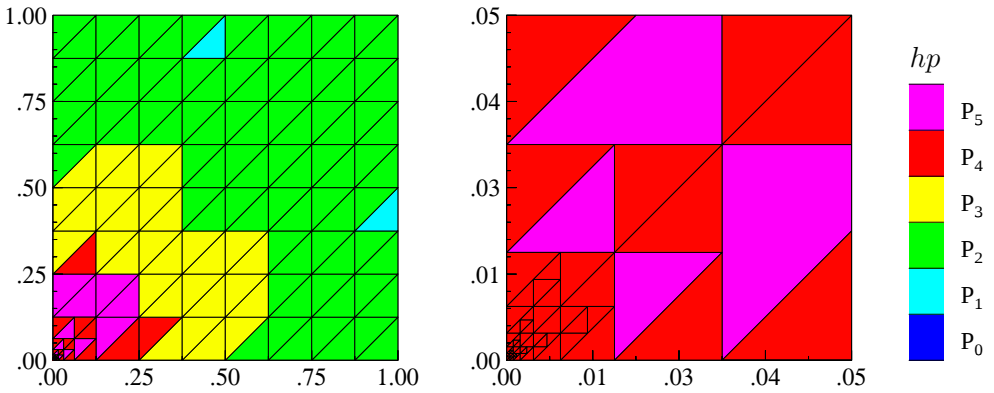
$$i_{\mathrm{eff}} := \frac{\eta_h(\tilde{u}_h)}{\left(\|u-\tilde{u}_h\|_X{}^2 + \mathscr{N}_h(\tilde{u}_h)^2\right)^{1/2}} = \frac{\left(\rho_h(\tilde{u}_h)^2 + \mathscr{N}_h(\tilde{u}_h)^2\right)^{1/2}}{\left(\|u-\tilde{u}_h\|_X{}^2 + \mathscr{N}_h(\tilde{u}_h)^2\right)^{1/2}}. \tag{27}$$

Let us note that the index i_{eff} is not a standard effectivity index since $\rho_h(\tilde{u}_h)$ is the approximation of $\mathscr{R}_h(\tilde{u}_h)$ and not of $\|u-\tilde{u}_h\|_X$. Obviously, if $\mathscr{N}_h(\tilde{u}_h)$ dominate $\rho_h(\tilde{u}_h)$ and $\|u-\tilde{u}_h\|_X$ then the index i_{eff} is close to one. In this case it would be also interesting to evaluate the ratio $\rho_h(\tilde{u}_h)/\|u-\tilde{u}_h\|_X$.

4.1 (E1): Nonlinear Convection-Diffusion Equation with a Corner Singularity

We consider the scalar nonlinear convection-diffusion equation

$$-\nabla\cdot(K(u)\nabla u) - \frac{\partial u^2}{\partial x_1} - \frac{\partial u^2}{\partial x_2} \;=\; g \quad \text{in } \Omega := (0,1)^2, \tag{28}$$



Source: Own

Fig. 1. *Example (E1) given by* (28) – (30) *with* $\alpha = -3/2$*: the final grid with the corresponding degrees of polynomial approximation, the whole domain (left) and its detail* $(0,1/20)\times(0,1/20)$ *(right)*

where $K(u)$ is the nonsymmetric matrix given by

$$K(u) = \varepsilon \begin{pmatrix} 2+\arctan(u) & (2-\arctan(u))/4 \\ 0 & (4+\arctan(u))/2 \end{pmatrix}. \tag{29}$$

The parameter $\varepsilon > 0$ plays a role of an amount of diffusivity and we put $\varepsilon = 10^{-3}$. We prescribe a Dirichlet boundary condition on $\partial\Omega$ and set the source term g such that the exact solution is

$$u(x_1,x_2) = (x_1^2+x_2^2)^{\alpha/2} x_1 x_2 (1-x_1)(1-x_2), \quad \alpha \in \mathbb{R}. \tag{30}$$

We present two choices: $\alpha = 4$ and $\alpha = -3/2$. It is possible to show (see [1]) that $u \in H^\kappa(\Omega)$, $\kappa \in (0, 3+\alpha)$, where $H^\kappa(\Omega)$ denotes the Sobolev-Slobodetskii space of functions with "non-integer derivatives". Whereas the choice $\alpha = 4$ gives sufficiently regular solution, the choice $\alpha = -3/2$ leads to the solution with a singularity at $x_1 = x_2 = 0$. Numerical examples presented in [4], carried out for a little different problem, show that this singularity avoids to achieve an order of convergence better than $O(h^{3/2})$ in the L^2-norm and $O(h^{1/2})$ in the H^1-seminorm for any degree of polynomial approximation. Nevertheless, the exact solution is regular outside of the singularity.

Table 1 shows the results for problem (28) – (30) with $\alpha = 4$ and $\alpha = -3/2$, namely the values of the error $\|u-\tilde{u}_h\|_X$, nonconformity $\mathscr{N}_h(\tilde{u}_h)$, residuum error estimate $\rho_h(\tilde{u}_h)$ with the corresponding EOC, index i_{eff} and the computational times in seconds. We observe that the computational error $\|u-\tilde{u}_h\|_X$ converge exponentially for $\alpha = 4$ and significantly faster than $O(h^{1/2})$ for $\alpha = -3/2$. Furthermore, the index i_{eff} is very close to one for increasing N_h which supports the accuracy of the method. A small increase of i_{eff} in Table 1 for $\alpha = 4$ for the last adaptation level is caused by the fact that we are close to the machine accuracy.

Furthermore, Figure 1 shows the final hp-grid obtained with the aid of the hp-DGFE algorithm for $\alpha = -3/2$. (The case $\alpha = 4$ is not interesting since only p-refinement is carried out due to the regularity of the exact solution.) We observe that the h-adaptation was carried out in a small region near the singularity. On the other hand, the p-adaptation appears in regions where the solution is regular.

Tab. 1. *Example (E1) given by* (28) – (30)*: error* $\|u-\tilde{u}_h\|_X$*, nonconformity* $\mathscr{N}_h(\tilde{u}_h)$*, residuum error estimate* $\rho_h(\tilde{u}_h)$ *with the corresponding EOC, index* i_{eff} *and the computational time in seconds*

$\alpha = 4$:

lev	$\#\mathscr{T}_h$	N_h	$\|u-\tilde{u}_h\|_X$	EOC	$\mathscr{N}_h(\tilde{u}_h)$	EOC	$\rho_h(\tilde{u}_h)$	EOC	i_{eff}	CPU(s)
0	128	384	6.45E-03	–	1.99E-02	–	5.58E-03	–	0.99	0.4
1	128	705	4.56E-04	8.72	3.07E-03	6.15	6.60E-04	7.03	1.01	0.7
2	128	384	6.45E-03	8.72	1.99E-02	6.15	5.58E-03	7.03	0.99	0.4
3	128	768	4.43E-04	7.73	3.01E-03	5.45	6.35E-04	6.27	1.01	0.7
4	128	1280	3.87E-05	9.54	2.75E-04	9.36	5.22E-05	9.78	1.01	1.0
5	128	1920	2.75E-06	13.05	1.73E-05	13.65	3.05E-06	14.02	1.00	1.4
6	128	2688	1.10E-07	19.12	7.04E-07	19.02	1.13E-07	19.59	1.00	2.2
7	128	3584	2.49E-09	26.35	1.62E-08	26.24	2.42E-09	26.72	1.00	3.4
8	128	4608	2.98E-11	35.22	2.23E-10	34.08	3.00E-11	34.92	1.00	5.3
9	128	5760	3.17E-15	82.03	2.02E-14	83.51	1.56E-14	67.83	1.25	9.0

$\alpha = -3/2$:

lev	$\#\mathscr{T}_h$	N_h	$\|u-\tilde{u}_h\|_X$	EOC	$\mathscr{N}_h(\tilde{u}_h)$	EOC	$\rho_h(\tilde{u}_h)$	EOC	i_{eff}	CPU(s)
0	128	384	1.32E-02	–	1.41E-01	–	4.52E-02	–	1.05	0.5
1	128	759	5.98E-03	2.32	6.70E-02	2.18	1.26E-02	3.75	1.01	0.8
2	128	919	5.50E-03	0.87	6.36E-02	0.55	6.26E-03	7.31	1.00	1.1
3	128	969	4.30E-03	9.31	5.52E-02	5.36	4.34E-03	13.81	1.00	1.4
4	134	1089	2.98E-03	6.29	3.96E-02	5.69	3.09E-03	5.86	1.00	1.6
5	140	1191	2.10E-03	7.81	2.75E-02	8.14	2.07E-03	8.91	1.00	1.9
6	152	1371	1.49E-03	4.82	1.93E-02	4.99	1.45E-03	5.03	1.00	2.1
7	158	1476	1.07E-03	9.07	1.37E-02	9.33	1.05E-03	8.72	1.00	2.5
8	164	1578	7.71E-04	9.78	9.78E-03	10.13	7.97E-04	8.34	1.00	2.9
9	176	1758	5.65E-04	5.75	7.04E-03	6.09	6.35E-04	4.21	1.00	3.3
10	188	1938	4.26E-04	5.81	5.15E-03	6.41	5.37E-04	3.43	1.00	3.8
11	200	2118	3.35E-04	5.41	3.87E-03	6.41	4.81E-04	2.48	1.00	4.3

Source: Own

4.2 (E2): Linear Convection-Diffusion Equation with the Strong Interior Layer and the Exponential Boundary Layer

We consider the example from [10], given by

$$-\varepsilon\triangle u + b_1\frac{\partial u}{\partial x_1} + b_2\frac{\partial u}{\partial x_2} = 0 \quad \text{in } \Omega := (0,1)^2, \tag{31}$$

where $\varepsilon = 10^{-8}$ is a constant diffusion coefficient and $(b_1,b_2) = (\cos(-\pi/3),\sin(-\pi/3))$ is the convection. We prescribe the Dirichlet boundary condition on $\partial\Omega$ by

$$u_D(x_1,x_2) = \begin{cases} 0 & \text{for } x_1 = 1 \text{ or } x_2 \le 0.7, \\ 1 & \text{otherwise.} \end{cases} \tag{32}$$

The solution possesses an interior layer in the direction of the convection starting in $(x_1,x_2) = (0,0.7)$ and contains two boundary layers along $x_1 = 0$ and $x_2 = 0$. On the boundary $x_1 = 1$ and on the right part of the boundary $x_2 = 0$, exponential layers are developed. The width of layers are proportional to ε.

For this case, the exact solution is piecewise constant except thin regions along the boundary and interior layer, however, its analytical expression is unknown. Therefore the computational

Tab. 2. *Example (E2) given by* (31) – (32) *with* $\varepsilon = 10^{-8}$*: the approximation of the error* $\|\overline{u} - \tilde{u}_h\|_X$*, nonconformity* $\mathcal{N}_h(\tilde{u}_h)$*, residuum error estimate* $\rho_h(\tilde{u}_h)$ *with the corresponding EOC, index* i_{eff} *and the computational time in seconds*

lev	$\#\mathcal{T}_h$	N_h	$\|\overline{u} - \tilde{u}_h\|_X$	EOC	$\mathcal{N}_h(\tilde{u}_h)$	EOC	$\rho_h(\tilde{u}_h)$	EOC	i_{eff}	CPU(s)
0	128	384	1.25E-01	–	3.58E+00	–	1.03E+00	–	1.04	0.3
1	128	635	9.56E-02	1.08	3.57E+00	0.01	7.60E-01	1.20	1.02	0.6
2	167	1211	8.06E-02	0.53	5.03E+00	-1.06	7.60E-01	-0.00	1.01	1.0
3	245	1984	7.50E-02	0.30	7.09E+00	-1.39	5.85E-01	1.06	1.00	2.2
4	467	3921	6.35E-02	0.49	1.00E+01	-1.02	5.95E-01	-0.05	1.00	4.0
5	1025	9751	6.12E-02	0.08	1.42E+01	-0.76	6.89E-01	-0.32	1.00	9.4
6	2189	21663	4.51E-02	0.76	2.01E+01	-0.87	7.32E-01	-0.15	1.00	18.2
7	4910	54765	3.17E-02	0.76	2.84E+01	-0.75	6.75E-01	0.17	1.00	50.8
8	7733	106285	2.28E-02	0.99	2.86E+01	-0.02	5.38E-01	0.68	1.00	196.6
9	14240	221582	1.76E-02	0.71	2.86E+01	-0.00	5.20E-01	0.09	1.00	505.8
10	19103	310336	1.62E-02	0.50	2.86E+01	-0.00	5.16E-01	0.04	1.00	1046.7
11	17849	272175	1.61E-02	-0.05	2.86E+01	0.00	5.16E-01	-0.00	1.00	1163.8
12	17426	250641	1.61E-02	-0.01	2.86E+01	-0.00	5.16E-01	0.00	1.00	1226.7
13	17366	240586	1.61E-02	0.03	2.86E+01	-0.01	5.16E-01	0.00	1.00	1248.5
14	17288	236116	1.61E-02	-0.16	2.86E+01	0.02	5.16E-01	0.01	1.00	1290.4
15	17207	233366	1.61E-02	0.12	2.86E+01	-0.05	5.16E-01	0.01	1.00	1311.4

Source: Own

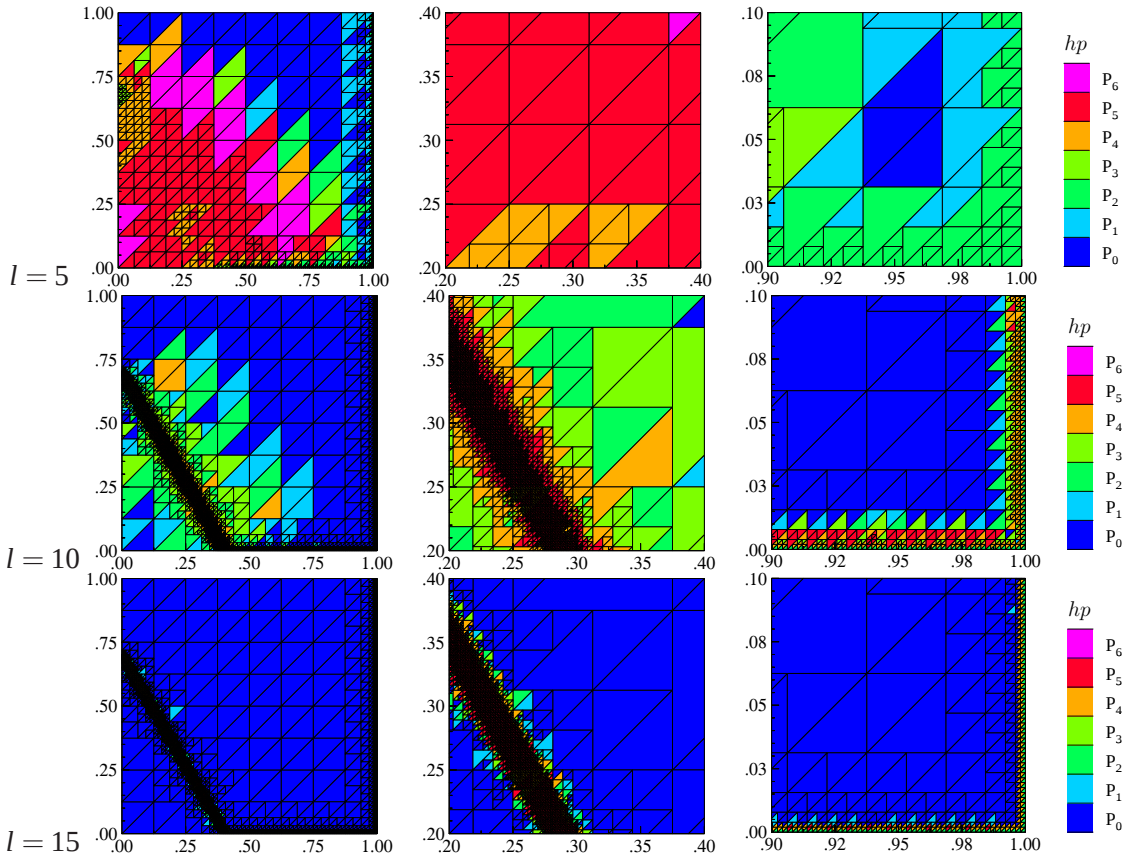
error $\|u - \tilde{u}_h\|_X$ is approximated by $\|\overline{u} - \tilde{u}_h\|_X$ where $\overline{u}$ is piecewise constant function corresponding to the exact solution of (31) in the limit with $\varepsilon \to 0$. (The boundary conditions have to be modified of course.)

Table 2 shows the results of computations (E2) for problem (31) – (32), namely the values of the approximation of the error $\|\overline{u} - \tilde{u}_h\|_X$, nonconformity $\mathcal{N}_h(\tilde{u}_h)$, residuum error estimate $\rho_h(\tilde{u}_h)$ with the corresponding EOC, index i_{eff} and the computational times in seconds. The value $\|\overline{u} - \tilde{u}_h\|_X$ does not tend to zero, probably due to the difference between the exact solution u and its approximation $\overline{u}$. Moreover, the nonconformity $\mathcal{N}_h(\tilde{u}_h)$ is high, it will decrease with an additional mesh adaptation, however, here we face a problem with a too high number of degree of freedom and the limits of our computer. It would be more efficient to apply an anisotropic mesh adaptation. Nevertheless, the results show that the presented hp-method works reasonably, both layers are well captured.

Furthermore, Figure 2 shows the final hp-grid obtained with the aid of the hp-DGFE algorithm after 5, 10 and 15 adaptive cycles. We observe that the h-adaptation was carried out in regions near both layers. In regions, where the solution is constant, the P_0 approximation is finally used. We also observe that the algorithm is able to decrease N_h when the thin layers are well localized. Finally, Figure 3 shows the detail of the diagonal cut $[0,0] \to [1,1]$ of the approximate solution after $l = 5$, $l = 10$ and $l = 15$ adaptation cycles.

Conclusion

We presented a hp-adaptive numerical method for the solution of the second order boundary value problem. This approach is based on a heuristic approximation of the error measured in a dual norm. Although a theoretical justification of this technique is missing, numerical experiments show reasonable computational properties.



Source: Own

Fig. 2. *Example (E2) given by* (31) – (32) *with* $\varepsilon = 10^{-8}$*: the grids after* $l = 5, 10, 15$ *adaptation cycles with the corresponding degrees of polynomial approximation, the whole domain (left) and its details* $(0.2, 0.4) \times (0.2, 0.4)$ *(centre) and* $(0.9, 1) \times (0, 0.1)$ *(right)*

Acknowledgments

This work is a part of the research project Grant No. 201/08/0012 of the Czech Science Foundation.

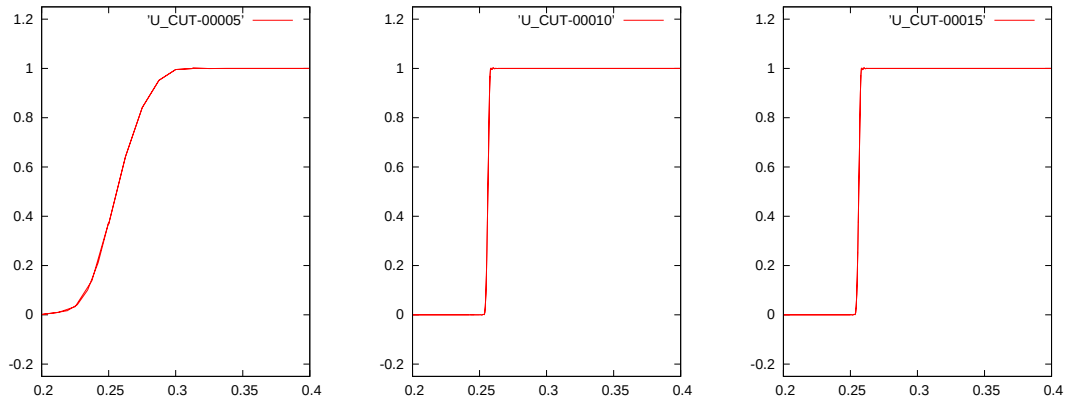
Literature

[1] BABUŠKA, I.; SURI, M.: The *p*- and *hp*- versions of the finite element method. An overview. *Comput. Methods Appl. Mech. Eng.*, 80 (1990), pp. 5–26.

[2] BABUŠKA, I.; SURI, M.: The $p-$ and $hp-$FEM a survey. *SIAM Review*, 36 (1994), pp. 578–632.

[3] DOLEJŠÍ, V.; ERN, A.; VOHRALÍK, M.: A framework for robust a posteriori error control in unsteady nonlinear advection-diffusion problems. *SIAM J. Numer. Anal.*, (submitted for publication).

[4] DOLEJŠÍ, V.; FEISTAUER, M.; KUČERA, V.; SOBOTÍKOVÁ, V.: An optimal $L^\infty(L^2)$-error estimate of the discontinuous Galerkin method for a nonlinear nonstationary convection-diffusion problem. *IMA J. Numer. Anal.*, 28 (2008), pp. 496–521.



Source: Own

Fig. 3. *Example (E2) given by* (31) – (32) *with* $\varepsilon = 10^{-8}$*: detail of the diagonal cut* $[0,0] \to [1,1]$ *of the approximate solution* $l = 5$ *(left),* $l = 10$ *(centre) and* $l = 15$ *(right) adaptation cycles*

[5] DOLEJŠÍ, V.; FEISTAUER, M.; SCHWAB, C.: On some aspects of the discontinuous Galerkin finite element method for conservation laws. *Math. Comput. Simul.*, 61 (2003), pp. 333–346.

[6] EIBNER, T.; MELENK, J. M.: An adaptive strategy for *hp*-FEM based on testing for analyticity. *Comput. Mech.*, 39 (2007), pp. 575–595.

[7] FEISTAUER, M.; KUČERA, V.: On a robust discontinuous Galerkin technique for the solution of compressible flow. *J. Comput. Phys.*, 224 (2007), pp. 208–221.

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

[9] HOZMAN, J.: *Discontinuous Galerkin Method for Convection-Diffusion Problems*. PhD thesis, Charles University Prague, Faculty of Mathematics and Physics, 2009.

[10] JOHN, V.; KNOBLOCH, P.: On spurious oscillations at layer diminishing (SOLD) methods for convection–diffusion equations: Part I – A review. *Comput. Methods Appl. Mech Engrg.*, 196 (2007), pp. 2197–2215.

[11] KROLL, N.; BIELER, H.; DECONINCK, H.; COUALLIER, V.; VAN DER VEN, H.; SORENSEN, K., eds.: *ADIGMA A European Initiative on the Development of Adaptive Higher-Order Variational Methods for Aerospace Applications*. vol. 113 of Notes on Numerical Fluid Mechanics and Multidisciplinary Design, Springer Verlag, 2010.

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

[13] ŠOLÍN, P.; DEMKOWICZ, L.: Goal-oriented *hp*-adaptivity for elliptic problems. *Comput. Methods Appl. Mech. Engrg.*, 193 (2004), pp. 449–468.

prof. RNDr. Vít Dolejší, Ph.D., DSc.

hp-NESPOJITÁ GALERKINOVA METODA PRO NELINEÁRNÍ PROBLÉMY

Zabýváme se numerickým řešením nelineárních konvektivně-difusních rovnic pomocí nespojité Galerkinovy metody. Navrhujeme novou *hp*-adaptivní metodu, která je založena na residuálně-nekonformním odhadu chybu a indikátoru regularity řešení. Residuálně-nekonformní odhad se skládá ze dvou částí, residuálním odhadu chyby a tzv. hodnotě nekonformity. Pomocí odhadu chyby hledáme elementy sítě, které se mají zjemnit a indikátor regularity rozhoduje, zda-li mají být označené elementy zjemněny technikou *h* nebo *p*. Residuálně-nekonformní odhad a i indikátor regularity jsou snadno spočitatelné veličiny. Stejná technika může rovněž pomoci odhadnout chybu vzniklou z nepřesného řešení příslušné nelineární algebraické soustavy rovnic. Možnosti metody jsou dokumentovány pomocí několika numerických experimentů.

DIE *hp*-DISKONTINUIERLICHE GALERKIN-METHODE FÜR NICHTLINEARE PROBLEME

Wir beschäftigen uns mit de numerischen Lösung nichtlinearer konvektiv-diffuser Gleichungen mit Hilfe der diskontinuierlichen Galerkin-Methode. Wir entwerfen eine neue *hp*-adaptive Methode, die auf einer Kombination einer residual-nichtkonformen Fehlerschätzung und einem Regularitätsindikator beruht. Die residual-nichtkonforme Schätzung besteht aus zwei Teilen, der residualen Fehlerschätzung und dem so genannten Nichtkonformitätswert. Mit Hilfe der Fehlerschätzung suchen wir Netzelemente, die verfeinert werden sollen, und der Regularitätsindikator entscheidet, ob die bezeichneten Elemente mit der Technik *h* oder *p* verfeinert werden sollen. Die residual-nichtkonforme Schätzung und auch der Regularitätsindikator sind wohl berechenbare Größen. Die gleiche Technik kann ebenfalls bei der Schätzung eines Fehlers helfen, der aus einer ungenauen Lösung des zugehörigen nichtlinearen algebraischen Gleichungssystems hervorgegangen ist. Die Möglichkeiten der Methode werden mit Hilfe einiger numerischer Experimente dokumentiert.

hp-NIECIĄGŁA METODA GALERKINA DLA PROBLEMÓW NIELINIOWYCH

Zajmujemy się numerycznym rozwiązywaniem nieliniowych równań konwekcyjno-dyfuzyjnych przy pomocy nieciągłej metody Galerkina. Proponujemy nową metodę *hp*-adaptacyjną, opartą na rezydualnie niekonformicznym szacunku błędu i wskaźniku regularności rozwiązania. Szacunek rezydualno niekonformiczny składa się z dwóch części, rezydualnym szacunku błędu oraz tzw. wartości niezgodności. Za pomocą szacunku błędu poszukujemy elementów sieci, które mają zostać złagodzone a wskaźnik regularności decyduje o tym, czy zaznaczone elementy mają być złagodzone przy pomocy techniki *h* czy *p*. Szacunek rezydualno- niekonformiczny oraz wskaźnik regularności są wielkościami łatwymi do wyliczenia. Ta sama technika może pomóc także przy szacowaniu błędu powstałego w wyniku niedokładnego rozwiązania danego nieliniowego algebraicznego układu równań. Możliwości metody pokazano na przykładzie kilku eksperymentów numerycznych.

ADAPTIVE INEXACT NEWTON METHODS WITH A POSTERIORI STOPPING CRITERIA

Alexandre Ern
***Martin Vohralík**

Université Paris-Est, CERMICS
Ecole des Ponts ParisTech
77455 Marne la Vallée, France
ern@cermics.enpc.fr

*Université Paris 6
Laboratoire Jacques-Louis Lions
75005 Paris, France
vohralik@ann.jussieu.fr

Abstract

We consider nonlinear algebraic systems arising from numerical discretizations of nonlinear partial differential equations of diffusion type. In order to solve them, some iterative nonlinear solver, and, on each step of this solver, some iterative linear solver are used. We propose adaptive stopping criteria for both these solvers, based on an a posteriori error estimate which distinguishes the different error components, namely the discretization, linearization, and algebraic ones. Our estimates give a guaranteed error upper bound and also a robust error lower bound. Numerical experiments for the nonlinear Laplace equation, nonconforming finite element discretization, Newton linearization, and conjugate gradients algebraic solver illustrate the theory.

Keywords: Nonlinear algebraic system; adaptive linearization, adaptive algebraic solution; stopping criterion; a posteriori error estimate.

Introduction

Consider a system of nonlinear algebraic equations written in the form: find a vector $U \in \mathbb{R}^N$, $N \geq 1$, such that

$$\mathscr{A}(U) = F, \tag{1}$$

where $\mathscr{A} : \mathbb{R}^N \to \mathbb{R}^N$ is a nonlinear operator and $F \in \mathbb{R}^N$ a given vector. We describe in this contribution an adaptive version of the inexact Newton method, cf. [2], for problem (1).

Our method is driven by stopping criteria based on a posteriori error estimators distinguishing three main error components, namely the discretization, linearization, and algebraic ones. On a nonlinear solver step k, $k \geq 1$, and linear solver step i, $i \geq 1$, these estimators are respectively denoted by $\eta_{\mathrm{disc}}^{k,i}$, $\eta_{\mathrm{lin}}^{k,i}$, and $\eta_{\mathrm{alg}}^{k,i}$. A notion of an algebraic remainder estimator $\eta_{\mathrm{rem}}^{k,i}$ also appears. All the estimators are fully computable quantities on each iteration step; their precise form depends on the nonlinear problem, numerical discretization, and nonlinear solver at hand, but is independent of the linear solver. Examples are given below, whereas the detailed forms can be found in [1].

Let γ_{rem}, γ_{alg}, and γ_{lin} be positive user-given weights, typically of order 0.1, related to the maximum percentage part of the given error component in the total error. The algorithm reads:

Algorithm 1 (Adaptive inexact Newton method)**.**

1. *Choose an initial vector $U^0 \in \mathbb{R}^N$. Set $k := 1$.*

2. *From U^{k-1}, define a matrix $\mathbb{A}^k \in \mathbb{R}^{N,N}$ and a vector $F^k \in \mathbb{R}^N$. Consider the following system of linear algebraic equations:*

$$\mathbb{A}^k U^k = F^k. \tag{2}$$

3. (a) *Define $U^{k,0} := U^{k-1}$ and set $i := 1$.*

   (b) *Perform a step of a chosen iterative linear solver for the solution of the linear system* (2)*, starting from the vector $U^{k,i-1}$. This yields an approximation $U^{k,i}$ to U^k which satisfies*

$$\mathbb{A}^k U^{k,i} = F^k - R^{k,i}, \tag{3}$$

   where $R^{k,i} \in \mathbb{R}^N$ is the algebraic residual vector on step i.

   (c) *Perform $\nu > 0$ additional steps of the iterative linear solver yielding an approximation $U^{k,i+\nu}$ to U^k which satisfies*

$$\mathbb{A}^k U^{k,i+\nu} = F^k - R^{k,i+\nu}, \tag{4}$$

   where $R^{k,i+\nu} \in \mathbb{R}^N$ is the algebraic residual vector on step $i+\nu$. The parameter ν is progressively increased until

$$\eta^{k,i}_{\mathrm{rem}} \le \gamma_{\mathrm{rem}} \max\{\eta^{k,i}_{\mathrm{disc}}, \eta^{k,i}_{\mathrm{lin}}, \eta^{k,i}_{\mathrm{alg}}\}. \tag{5}$$

   (d) *Check the convergence criterion for the linear solver in the form*

$$\eta^{k,i}_{\mathrm{alg}} \le \gamma_{\mathrm{alg}} \max\{\eta^{k,i}_{\mathrm{disc}}, \eta^{k,i}_{\mathrm{lin}}\}. \tag{6}$$

   If satisfied, set $U^k := U^{k,i}$. If not, set $i := i+\nu$ and go back to step 3b.

4. *Check the convergence criterion for the nonlinear solver in the form*

$$\eta^{k,i}_{\mathrm{lin}} \le \gamma_{\mathrm{lin}} \eta^{k,i}_{\mathrm{disc}}. \tag{7}$$

   If satisfied, finish. If not, set $k := k+1$ and go back to step 2.

A prominent example of linearization is the Newton one, giving (2) with

$$\mathbb{A}^k_{ij} := \frac{\partial \mathscr{A}_i}{\partial U_j}(U^{k-1}), \qquad F^k := F - \mathscr{A}(U^{k-1}) + \mathbb{A}^k U^{k-1}, \tag{8}$$

but other linearizations like the fixed point one are also allowed. As for the iterative algebraic solver yielding (3), we also do not make any requirement.

1 A Nonlinear Partial Differential Equation and its Numerical Approximation

The nonlinear systems (1) typically arise from some numerical approximation of a nonlinear partial differential equation. Let $\Omega \subset \mathbb{R}^d$, $d \ge 2$, be a polygonal (polyhedral) domain (open, bounded, and connected set). We consider the following model partial differential equation: find $u : \Omega \to \mathbb{R}$ such that

$$-\nabla\cdot\boldsymbol{\sigma}(u,\nabla u) = f \qquad \text{in } \Omega, \tag{9a}$$

$$u = 0 \qquad \text{on } \partial\Omega, \tag{9b}$$

where $\boldsymbol{\sigma} : \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}^d$ is a nonlinear flux function and $f : \Omega \to \mathbb{R}$ a source term. The scalar-valued unknown function u is termed the *potential*, and, given a potential u, the vector-valued function $-\boldsymbol{\sigma}(u, \nabla u)$ is termed the *flux*. We assume that $f \in L^q(\Omega)$, $q > 1$, and set $p := \frac{q}{q-1}$ so that $\frac{1}{p} + \frac{1}{q} = 1$. The energy space is $V := W_0^{1,p}(\Omega)$, i.e., the space of $L^p(\Omega)$ functions whose weak derivatives are in $L^p(\Omega)$, with the zero trace on $\partial\Omega$.

The exact solution u lies in the space V. Let $u_h^{k,i}$ be a numerical approximation on a mesh $\mathcal{T}_h$ of Ω, linearization step $k \geq 1$, and algebraic solver step $i \geq 1$, corresponding to the algebraic vector $U^{k,i}$ of (3). We suppose $u_h^{k,i} \in V(\mathcal{T}_h)$, where

$$V(\mathcal{T}_h) := \{v \in L^p(\Omega),\, v|_K \in W^{1,p}(K) \quad \forall K \in \mathcal{T}_h\}. \tag{10}$$

Remark that $V(\mathcal{T}_h) \not\subset V$, so that $u_h^{k,i}$ can be nonconforming. Let $\mathcal{E}_K$ regroup the faces e of an element $K \in \mathcal{T}_h$, denote by h_e the diameter of the face e, and by $[\![\cdot]\!]$ the jump operator, yielding the difference of (the traces of) the argument from the two mesh elements that share e on interfaces and the actual trace if e is a boundary face. The error between the exact solution u of (9) and the approximate solution $u_h^{k,i}$ is measured as

$$\mathscr{J}_u(u_h^{k,i}) := \mathscr{J}_{u,\mathrm{F}}(u_h^{k,i}) + \mathscr{J}_{u,\mathrm{NC}}(u_h^{k,i}), \tag{11}$$

where

$$\mathscr{J}_{u,\mathrm{F}}(u_h^{k,i}) := \sup_{\varphi \in V;\, \|\nabla\varphi\|_p = 1} \big(\boldsymbol{\sigma}(u, \nabla u) - \boldsymbol{\sigma}(u_h^{k,i}, \nabla u_h^{k,i}), \nabla\varphi\big), \tag{12a}$$

$$\mathscr{J}_{u,\mathrm{NC}}(u_h^{k,i}) := \left\{ \sum_{K \in \mathcal{T}_h} \sum_{e \in \mathcal{E}_K} h_e^{1-q} \|[\![u - u_h^{k,i}]\!]\|_{q,e}^q \right\}^{\frac{1}{q}}. \tag{12b}$$

The quantity $\mathscr{J}_{u,\mathrm{F}}(u_h^{k,i})$ measures the error in the approximation of the exact flux $-\boldsymbol{\sigma}(u, \nabla u)$ by the approximate one $-\boldsymbol{\sigma}(u_h^{k,i}, \nabla u_h^{k,i})$ and represents the dual norm of the residual of (9); $\mathscr{J}_{u,\mathrm{NC}}(u_h^{k,i})$ then measures the nonconformity of the discrete potential $u_h^{k,i}$, i.e., the departure of $u_h^{k,i}$ from the space V. Therein $\|\cdot\|_r$ stands for the Lebesgue norm in L^r. Importantly, there holds $\mathscr{J}_u(u_h^{k,i}) = 0$ if and only if $u_h^{k,i} = u$. The error measure $\mathscr{J}_u(u_h^{k,i})$ is not easily computable for a known exact solution u; the Hölder inequality, however, gives

$$\mathscr{J}_u(u_h^{k,i}) \leq \mathscr{J}_u^{\mathrm{up}}(u_h^{k,i}) := \|\boldsymbol{\sigma}(u, \nabla u) - \boldsymbol{\sigma}(u_h^{k,i}, \nabla u_h^{k,i})\|_q + \mathscr{J}_{u,\mathrm{NC}}(u_h^{k,i}), \tag{13}$$

which is simple to evaluate, being based on the $[L^q(\Omega)]^d$-difference of the fluxes, plus the nonconformity term. Our numerical experiments indicate that both error measures $\mathscr{J}_u(u_h^{k,i})$ and $\mathscr{J}_u^{\mathrm{up}}(u_h^{k,i})$ exhibit a very close behavior.

2 A posteriori Error Estimates and their Efficiency

Recall the estimators $\eta_{\mathrm{disc}}^{k,i}$, $\eta_{\mathrm{lin}}^{k,i}$, $\eta_{\mathrm{alg}}^{k,i}$, and $\eta_{\mathrm{rem}}^{k,i}$ introduced in the Introduction. These are quantities that are fully computable from $u_h^{k,i}$. Their precise forms for the model problem (9), various numerical discretizations, and various linearizations are given in [1]. Let moreover $\eta_{\mathrm{quad}}^{k,i}$ be a quadrature and $\eta_{\mathrm{osc}}^{k,i}$ a data oscillation estimator. The following theorem has been shown in [1]:

Theorem 1 (A posteriori error estimate distinguishing the different error components)**.** *Let $u \in V$ solve* (9) *and let $u_h^{k,i} \in V(\mathscr{T}_h)$. Then*

$$\mathscr{J}_u(u_h^{k,i}) \le \eta_{\mathrm{disc}}^{k,i} + \eta_{\mathrm{lin}}^{k,i} + \eta_{\mathrm{alg}}^{k,i} + \eta_{\mathrm{rem}}^{k,i} + \eta_{\mathrm{quad}}^{k,i} + \eta_{\mathrm{osc}}^{k,i}. \tag{14}$$

Theorem 1 gives an overall error control on each step k of the linearization and i of the algebraic solver. This control is tight in the sense of the following result, shown in [1]:

Theorem 2 (Global efficiency and robustness)**.** *Let $u \in V$ solve* (9) *and let $u_h^{k,i} \in V(\mathscr{T}_h)$. Let the global stopping and balancing criteria* (5), (6), (7) *be satisfied. Then there exists a generic constant C independent of the mesh size h, the domain Ω, the nonlinear function $\boldsymbol{\sigma}$, and the Lebesgue exponent q such that*

$$\eta_{\mathrm{disc}}^{k,i} + \eta_{\mathrm{lin}}^{k,i} + \eta_{\mathrm{alg}}^{k,i} + \eta_{\mathrm{rem}}^{k,i} \le C(\mathscr{J}_u(u_h^{k,i}) + \eta_{\mathrm{quad}}^{k,i} + \eta_{\mathrm{osc}}^{k,i}). \tag{15}$$

3 Numerical Illustration

This section gives a quick numerical illustration of the theoretical developments. We consider (9) with $\boldsymbol{\sigma}(u,\nabla u) = |\nabla u|^{p-2}\nabla u$, which is the so-called p-Laplacian. Here we only consider the value $p = 10$. We take $\Omega := (0,1)\times(0,1)$, $f := 2$, and prescribe an inhomogeneous Dirichlet boundary condition by the exact solution

$$u(\mathbf{x}) = -\frac{p-1}{p}|\mathbf{x}-(0.5,0.5)|^{p/(p-1)} + \frac{p-1}{p}\left(\frac{1}{2}\right)^{p/(p-1)}.$$

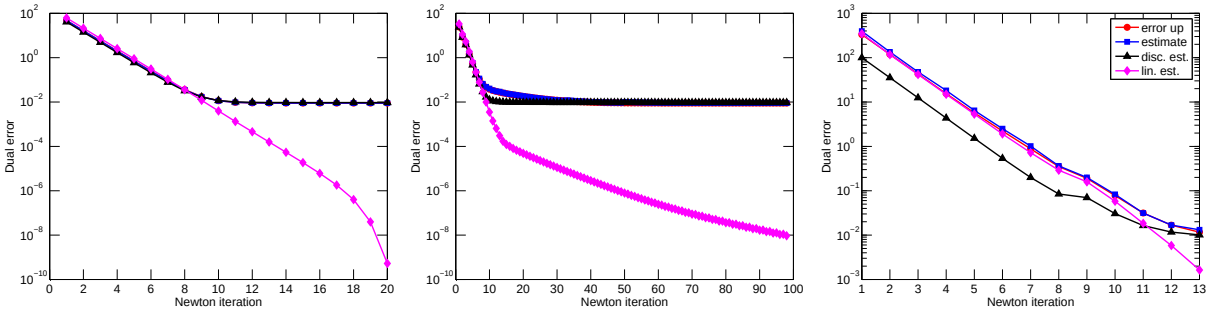
We employ the Crouzeix–Raviart nonconforming finite element method for the discretization, the Newton linearization (8), and the conjugate gradient algebraic solver with a diagonal preconditioning.

We compare three different stopping criteria in Algorithm 1, leading to three different solution approaches:

- In the *Full Newton (FN) method*, both the nonlinear and linear solvers are iterated to "almost" convergence, with the global stopping criteria $\eta_{\mathrm{alg}}^{k,i} \le 10^{-8}$ and $\eta_{\mathrm{lin}}^{k,i} \le 10^{-8}$. The balancing criterion (5) is employed with $\gamma_{\mathrm{rem}} = 0.1$.
- In the *Inexact Newton (IN) method*, the only difference with FN is that a fixed number of preconditioned CG iterations is performed on each Newton linearization step. These values were chosen respectively as 2, 3, 5, 8, 10, 15 on each level of the uniform mesh refinement.
- Finally, in the *Adaptive Inexact Newton (AIN) method* that we propose, we rely on the global stopping criteria (5), (6), and (7) with $\gamma_{\mathrm{rem}} = \gamma_{\mathrm{alg}} = \gamma_{\mathrm{lin}} = 0.3$.

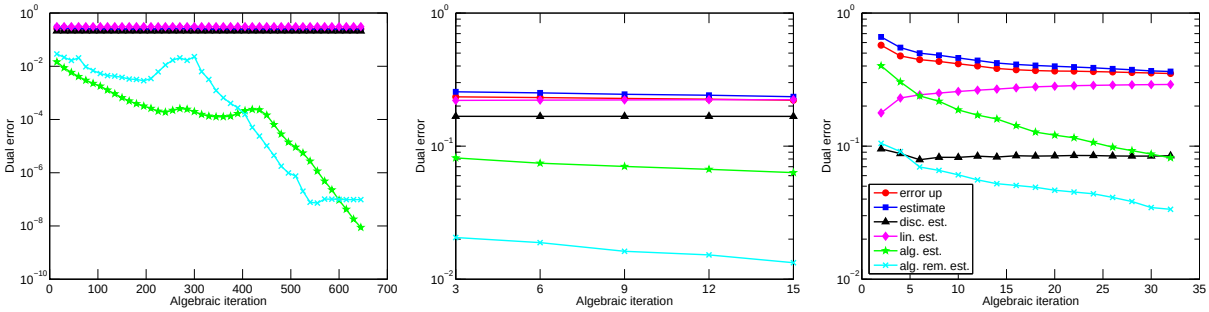
Figure 1 focuses on the 6th level uniformly refined mesh and tracks the dependence of the error measure $\mathscr{J}_u^{\mathrm{up}}(u_h^{k,i})$, the overall error estimator, and the discretization and linearization estimators $\eta_{\mathrm{disc}}^{k,i}$ and $\eta_{\mathrm{lin}}^{k,i}$ of Theorem 1 on the Newton iterations. Typically, the error and all the estimators except $\eta_{\mathrm{lin}}^{k,i}$ start to stagnate after the linearization error ceases to dominate. This is precisely the point where the nonlinear iteration is stopped in AIN using (7), whereas both FN and IN perform many unnecessary additional iterations.

Figure 2 further analyzes the situation on one chosen Newton iteration from Figure 1. To be in a region with similar error measure $\mathscr{J}_u^{\mathrm{up}}(u_h^{k,i})$, we have chosen the 6th iteration for FN and



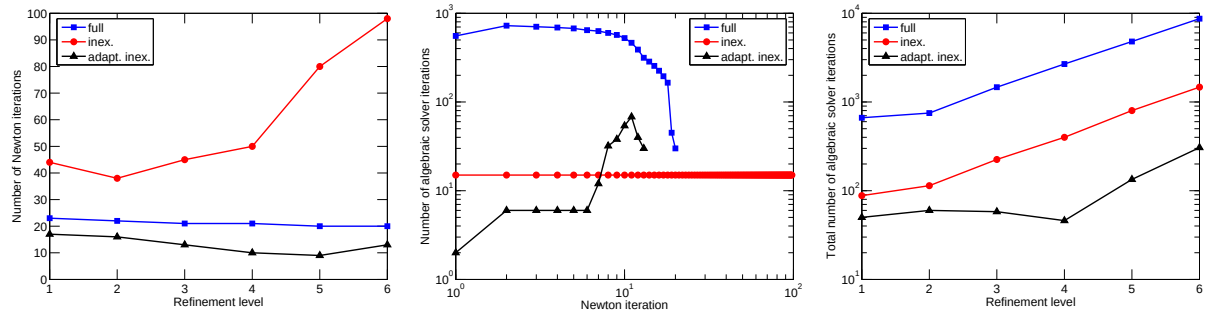
Source: Own

Fig. 1. *Error and estimators as a function of Newton iterations, 6th level mesh. Newton (left), inexact Newton (middle), and adaptive inexact Newton (right)*



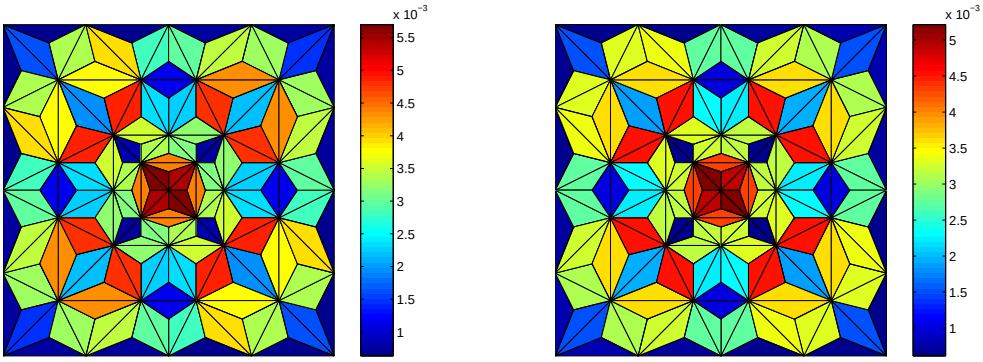
Source: Own

Fig. 2. *Error and estimators as a function of preconditioned CG iterations, 6th level mesh. Newton, 6th step (left), inexact Newton, 6th step (middle), and adaptive inexact Newton, 8th step (right)*



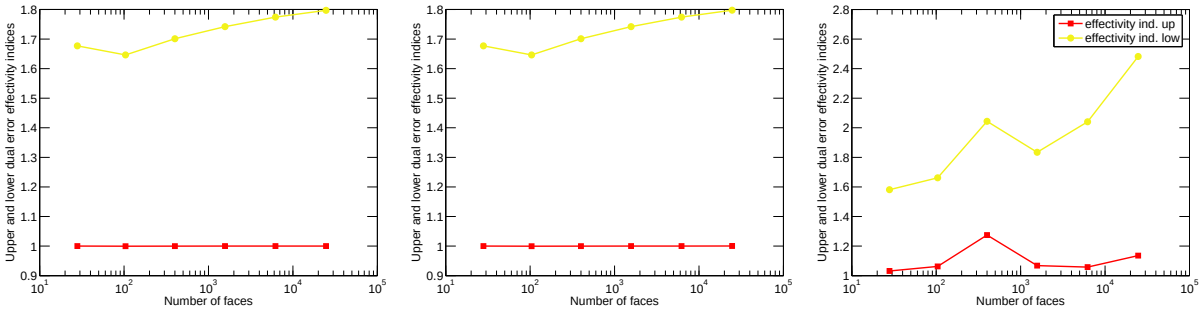
Source: Own

Fig. 3. *Number of Newton iterations per refinement level (left), number of linear solver iterations per Newton step on 6th level mesh (middle), and total number of linear solver iterations per refinement level (right)*



Source: Own

Fig. 4. *Estimated (left) and actual (right) error distribution, 2nd level uniformly refined mesh, adaptive inexact Newton*



Source: Own

Fig. 5. *Upper and lower effectivity indices. Newton (left), inexact Newton (middle), and adaptive inexact Newton (right)*

IN and the 8th iteration for AIN. We see that almost no decrease of the error measure $\mathscr{J}_u^{\mathrm{up}}(u_h^{k,i})$ can be observed during the almost 650 iterations of the preconditioned CG method in the FN case. The fixed 15 CG iterations in the IN case are, on the contrary, not sufficient to decrease significantly the error. In our approach, just the sufficient, "online-decided" number of CG iterations is performed and timely stopped using (6).

Figure 3 illustrates the overall performance of three approaches. We can see that the number of Newton iterations per refinement level is stable around 20 for FN. It increases significantly for IN, whereas it is still reduced for AIN. On one Newton iteration, the number of CG iterations also varies significantly among three approaches. Many iterations are necessary in the FN case and fixed 15 iterations in the IN case, whereas AIN picks up the number that is "just necessary." Remark that this number is equal to two on the first Newton step; from here, the error is "lagged" as a function of Newton iterations in the AIN case, cf. Figure 1. The total number of necessary CG iterations per refinement level is displayed in the right part of Figure 3. On the last mesh, AIN only needs 306 total iterations, whereas IN needs 1470, and FN 8690 iterations. Thus, our approach yields an economy by a factor of roughly 5 with respect to IN and roughly 30 with respect to FN in terms of total iterations.

Figure 4 displays the distribution of the overall error estimator and of the error measure $\mathscr{J}_u^{\mathrm{up}}(u_h^{k,i})$ on the 2nd level uniformly refined mesh for AIN. We see that even in presence of algebraic and linearization errors, the overall error distribution is very well predicted. Finally,

let $\mathscr{J}_u^{\mathrm{low}}(u_h^{k,i})$ be an easily computable lower bound for the error measure $\mathscr{J}_u(u_h^{k,i})$ obtained by estimating the supremum in (12a) just with one φ. We define the upper and lower effectivity indices respectively as $\mathscr{I}^{\mathrm{up}} := \eta^{k,i} / \mathscr{J}_u^{\mathrm{up}}(u_h^{k,i})$ and $\mathscr{I}^{\mathrm{low}} := \eta^{k,i} / \mathscr{J}_u^{\mathrm{low}}(u_h^{k,i})$ and observe that the effectivity index for the original error measure $\mathscr{J}_u(u_h^{k,i})$, defined as $\mathscr{I} := \eta / \mathscr{J}_u(u_h^{k,i})$, lies between $\mathscr{I}^{\mathrm{up}}$ and $\mathscr{I}^{\mathrm{low}}$. For the three methods (FN, IN, and AIN), all these indices are close to the optimal value of 1 and in particular $\mathscr{I}^{\mathrm{up}}$ takes values very close to 1, see Figure 5.

Conclusion

In this work, we have presented an inexact Newton method with a posteriori error control and adaptive stopping criteria. Numerical experiments illustrate that tight error bound and important computational gains can be achieved by our approach. Details on all the presented developments can be found in [1].

Acknowledgments

This work was partly supported by the Groupement MoMaS (PACEN/CNRS, ANDRA, BRGM, CEA, EdF, IRSN) and by the ERT project “Enhanced oil recovery and geological sequestration of CO_2: mesh adaptivity, a posteriori error control, and other advanced techniques” (LJLL/IFPEN).

Literature

[1] ERN, A.; VOHRALÍK, M.: Adaptive inexact Newton methods with a posteriori stopping criteria for nonlinear diffusion PDEs. *HAL Preprint 00681422v2*, submitted for publication, 2012.

[2] QUARTERONI, A.; VALLI, A.: *Numerical approximation of partial differential equations*. Springer-Verlag, Berlin, 1994.

prof. Alexandre Ern, Ph.D., Ing. Martin Vohralík, Ph.D.

Adaptivní nepřesné Newtonovy metody s a posteriorními zastavovacími kritérii

V této práci uvažujeme systémy nelineárních algebraických rovnic vznikající při numerické diskretizaci nelineárních parciálních diferenciálních rovnic difúzního typu. K jejich (přibližnému) řešení uvažujeme nelineární iterační metodu a, na každém jejím kroku, iterační řešič systému lineárních algebraických rovnic. Navrhujeme adaptivní uzpůsobení počtu kroků obou iteračních řešičů. Obě zastavovací kritéria jsou založena na a posteriorních odhadech, které rozlišují různé složky celkové chyby, v daném případě algebraickou chybu, linearizační chybu a diskretizační chybu. Naše a posteriorní odhady poskytují zaručenou horní hranici na celkovou chybu mezi přibližným a přesným řešením a zároveň robustnostní hranici spodní. Numerické experimenty pro nelineární Laplaceovu rovnici, nekonformní metodu konečných prvků, Newtonovu linearizaci a metodu sdružených gradientů pro řešení soustav lineárních algebraických rovnic ilustrují teoretické výsledky.

Die adaptive ungenaue Newton-Methode mit A-posteriori-Stopp-Kriterien

In dieser Arbeit betrachten wir die Systeme nichtlinearer algebraischer Gleichungen, die bei der numerischen Diskretisation nichtlinearer partieller Differenzialgleichungen entstehen. Zu deren - annäherungsweisen - Lösung betrachten wir die nicht lineare iterative Methode und - auf jedem ihrer Schritte - den iterativen Löser des Systems der linearen algebraischen Gleichungen. Wir schlagen eine adaptive Angleichung der Schrittzahl beider iterativer Löser vor. Beide Stopp-Kriterien gründen sich auf A-posteriori-Schätzungen, welche verschiedene Bestandteile des Gesamtfehlers unterscheiden, im vorliegenden Fall einen algebraischen Fehler, einen linearisierenden Fehler und einen Diskreditisationsfehler. Unsere A-posteriori-Schätzungen gewährleisten eine garantierte Obergrenze für den Gesamtfehler zwischen der Näherungs- und der genauen Lösung und gleichzeitig eine Robustitätsgrenze nach unten. Die numerischen Experimente für die nichtlineare Laplace-Gleichung, die nichtkonforme Methode der Endelemente, die Newton'sche Linearisation sowie die Methode der dualen Gradienten zur Lösung der Systeme linearer algebraischer Gleichungen illustrieren die theoretischen Ergebnisse.

Adaptacyjne niedokładne metody Newtona z kryteriami stopującymi a posteriori

W niniejszym opracowaniu przedstawiono systemy nieliniowych równań algebraicznych powstające podczas numerycznej dyskretyzacji nieliniowych parcjalnych równań różniczkowych o charakterze dyfuzyjnym. Do ich (przybliżonego) rozwiązania zastosowano nieliniową metodę iteracji a na każdym jej etapie iteracyjny sposób rozwiązania układu liniowych równań algebraicznych. Zaproponowano adaptacyjne dostosowanie liczby etapów obu sposobów. Oba kryteria stopujące oparte są na szacunkach a posteriori, które odróżniają różne elementy ogólnego błędu, w tym przypadku algebraiczne, linearyzacyjne i dyskretyzacyjne. Nasze oraz a posteriori szacunki zapewniają górną granicę ogólnego błędu pomiędzy przybliżonym a dokładnym rozwiązaniem oraz dolną stałą granicę błędu. Eksperymenty numeryczne dla nieliniowego równania Laplace'a, niekonformiczną metodę elementów skończonych, linearyzację Newtona oraz metodę gradientów sprzężonych do rozwiązywania układów liniowych równań algebraicznych przedstawiono w postaci teoretycznych wyników.

ON THE PROBLEM OF VARIABILITY OF INTERVAL DATA

Václav Finěk
***Ctirad Matonoha**

Technical University of Liberec
Faculty of Science, Humanities and Education
Department of Mathematics and Didactics of Mathematics
Studentská 2, 461 17, Liberec, Czech Republic
vaclav.finek@tul.cz

*Institute of Computer Science
Academy of Sciences of the Czech Republic
Pod Vodárenskou věží 2, 182 07 Prague, Czech Republic
matonoha@cs.cas.cz

Abstract

In this contribution we will deal with the problem of variability of interval data. The issue is to find lower and upper bounds for the interval of possible values for the variance of given interval data. This leads to the problem of minimizing/maximizing the sum of squares of the distance between the components of the $n-$dimensional vector and their average value. As the maximization is more difficult than the minimization, we present here some theoretical results concerning the solution. Furthermore, we introduce preliminary algorithms for solving both problems which take into consideration their special structure.

Keywords: Interval data; computation of variance; theoretical analysis.

Introduction

In some cases, we have only intervals $[a_i, b_i]$ of possible values of x_i instead of the actual value x_i. For example, measured values usually include some measurement error with known upper bounds. Then, the actual value of x_i is unknown and we only know that its value is located within the interval determined by the upper bound of the measurement error. Therefore, we should work rather with these intervals than with these single values. Consequently, possible values of their average and their variance are also intervals. For more details on interval data see in [1]. While the computation of lower and upper bounds for the average of interval data is straightforward, the computation of lower and upper bounds for their variance is significantly complicated. In [4], it was proved that computing the variance for interval data is NP-hard. They also proposed algorithms with quadratic complexity for computing the lower bound of variance and for computing its upper bound in some special cases. Further interesting results were recently published in [2, 3, 7]. In this contribution, we present some theoretical results concerning the solution of a maximization problem and introduce preliminary algorithms for solving both problems which use their special structure.

We consider n intervals $I_i = [a_i, b_i]$ and define

$$K = I_1 \otimes I_2 \otimes I_2 \otimes \cdots \otimes I_n. \tag{1}$$

We would like to find:

$$\mathbf{x}^{min} = \arg\min_{\mathbf{x}\in K} \frac{1}{n} F(x), \tag{2}$$

$$\mathbf{x}^{max} = \arg\max_{\mathbf{x}\in K} \frac{1}{n} F(x), \tag{3}$$

where $F(x) = \sum_{i=1}^{n} (x_i - \overline{\mathbf{x}})^2$ with $\overline{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and the tensor product K of the given intervals forms the feasible set of optimization problems (2) and (3).
Note that the function F can be written in the following forms:

$$F(x) = \sum_{i=1}^{n} (x_i - \overline{\mathbf{x}})^2 = \sum_{i=1}^{n} x_i^2 - \frac{(x_1 + \cdots + x_n)^2}{n} = \sum_{i=1}^{n} x_i^2 - n\overline{\mathbf{x}}^2 = \frac{1}{n}\left(\sum_{j<i} (x_i - x_j)^2\right)$$

Problem (2) possesses a convex objective function F, thus any local solution is also a global one. In problem (3), the function F is concave and the solution has to be a vertex of K. This results in a rather different behavior. The analysis of this second case is discussed in Section 3. There are several apparent properties:

- If $\cap_i I_i \neq \emptyset$ then there exist either one or infinitely many solutions of (2). All elements of the solution are the same and belong to $\cap_i I_i$.
- The solution of (3) lies on the vertex of K.
- It holds $\sum_{i=1}^{n} (x_i - \overline{\mathbf{x}})^2 < \sum_{i=1}^{n} (x_i - d)^2$ for any $d \neq \overline{\mathbf{x}}$.

In this contribution we give some theoretical results of problems (2) and (3), respectively, and outline the algorithms for seeking the solutions. We will consider the following structure of those problems:

$$\begin{aligned} & \mathbf{x}^{min} = \arg\min_{x} F(x), && (4) \\ \text{subject to} \quad & x_i \in [a_i, b_i], \quad i = 1 \ldots n \end{aligned}$$

$$\begin{aligned} & \mathbf{x}^{max} = \arg\max_{x} F(x), && (5) \\ \text{subject to} \quad & x_i \in [a_i, b_i], \quad i = 1 \ldots n \end{aligned}$$

The following quantities are used throughout the paper:

$$c_i = \frac{a_i + b_i}{2}, \quad d_i = b_i - a_i > 0, \quad i = 1 \ldots n,$$

$$p_a = \frac{a_1 + \cdots + a_n}{n}, \quad p_b = \frac{b_1 + \cdots + b_n}{n}.$$

1 The Problem of Minimization

Both problems (4) and (5), respectively, are classical optimization problems with simple bounds that can be solved by a suitable optimization method, e.g. a variable metric method or trust-region method, see e.g. [6]. The basic optimization method is an iteration process starting from an initial point $x^{(0)}$ and generating a sequence of points $x^{(1)}, x^{(2)}, \ldots$ such that

$$x^{(k+1)} = x^{(k)} + \alpha^{(k)} d^{(k)},$$

where $d^{(k)}$ is a direction vector and $\alpha^{(k)} > 0$ is a step-length. The direction vector is determined on the basis of values $x^{(j)}, F(x^{(j)}), F'(x^{(j)}), F''(x^{(j)}),\ 0 \le j \le k,$ and the step-length is determined on the basis of behavior of the function F in the neighborhood of $x^{(k)}$.

Unfortunately, the Hessian matrix of F is dense which makes difficult to solve problems (4) and (5), respectively, for large n by standard optimization methods. We will use of a special structure of the problem and introduce algorithms that take into consideration this structure.

The problem of minimization (4) is simple and no significant difficulties arise. We need not to know anything special about the solution and the following algorithm works well on the test problems.

Algorithm 1 *Solving problem (4).*

1. **Initiation:** Set

$$p_a = øa_i, \quad p_b = øb_i, \quad p = \frac{p_a + p_b}{2}, \quad A^{\max} = \max\{a_i\}, \quad B_{\min} = \min\{b_i\},$$

   $s_i = 0\ \forall i$ (indicator of the solution x_i^*: 0 - not found, 1 - found).

2. **Test on intersection of all intervals:** If $A^{\max} \le B_{\min},$ then the solution $x_i^*\ \forall i$ is an arbitrary number in $[A^{\max}, B_{\min}]$ and $F = 0$.

3. Until we have not found all x_i^* (i.e. $\exists i$ such that $s_i = 0$):

   (a) **Iteration process:** For $i = 1 \ldots n$ such that $s_i = 0$:
   If $p \in [a_i, b_i],$ set $x_i = p;$ else if $p < a_i,$ set $x_i = a_i;$ else set $x_i = b_i$.

   (b) **Reducing of all intervals:** For $i = 1 \ldots n$ such that $s_i = 0$:
   If $a_i < p_a \le b_i,$ set $a_i = p_a;$ if $b_i > p_b \ge a_i,$ set $b_i = p_b$.

   (c) **Test on components of the solution:** For $i = 1 \ldots n$ such that $s_i = 0$:
   If $[a_i, b_i] \cap [p_a, p_b] = \emptyset$ or a_i or $b_i,$ set $a_i = b_i = x_i^* = x_i$ and $s_i = 1$.

4. **Computation of new values:** Set $p_a^{old} = p_a,\ p_b^{old} = p_b$ and update p_a, p_b.

5. **Test on termination:** If $\max\{|p_a - p_a^{old}|, |p_b - p_b^{old}|\} > \varepsilon$ (e.g. $= 10^{-6}$), update p and goto Step 3. Otherwise set $x_i^* = x_i$ for i such that $s_i = 0$.

2 The Problem of Maximization

The maximization problem is much harder than the minimization problem, so we will focus on the theoretical analysis and find useful information concerning the solution. The solution x^* to (5) lies on the vertex of K, see the second apparent property above or Lemma 1 below. Thus in the subsequent analysis we assume that x^* has components x_i^* equal to either a_i or b_i. First, we will study what the solution must satisfy.

Lemma 1 *The solution x^* of (5) lies on the vertex of K, i.e. $x^* = (x_1^* \ldots x_n^*)$, where $x_i^* = a_i$ or $x_i^* = b_i$.*

Lemma 2 *The solution x^* of (5) has the property that there exists at least one i and at least one j, $i \ne j$ such that $x_i^* = a_i$ and $x_j^* = b_j$.*

The following lemma says that no component x_i^* of the solution x^* is equal to the average value $\overline{\mathbf{x}}^*$.

Lemma 3 *Let* $x \in \mathbb{R}^n$ *and* $\bar{\mathbf{x}}$ *be the average value of componenets* $x_1 \dots x_n$. *Suppose that there exists an index* j *such that* $x_j = \bar{\mathbf{x}}$. *Then if we take a component* $x_j + \tau$ *instead of* x_j *for some* $\tau \neq 0$, *we obtain a greater function value, i.e.* $F(x) < F(x_\tau)$, *where*

$$x_\tau = (x_1 \dots x_{j-1},\ x_j + \tau,\ x_{j+1} \dots x_n).$$

The following analysis concerns the differences among function values. The first lemma stands for the most general case.

Lemma 4 *Let* x, y *be arbitrary points. Denote* p_{xy} *the average value of all points* x_i, y_i, $i = 1 \dots n$, *and define the following sets:*

$$N_{aa} = \{i :\ x_i = a_i,\ y_i = a_i\},$$

$$N_{ab} = \{i :\ x_i = a_i,\ y_i = b_i\},$$

$$N_{ba} = \{i :\ x_i = b_i,\ y_i = a_i\},$$

$$N_{bb} = \{i :\ x_i = b_i,\ y_i = b_i\}.$$

It is evident that $N_{aa} \cup N_{ab} \cup N_{ba} \cup N_{bb} = \{1 \dots n\}$. *Then it holds*

$$F(x) - F(y) = 2 \left[\sum_{i \in N_{ba}} d_i (c_i - p_{xy}) - \sum_{i \in N_{ab}} d_i (c_i - p_{xy}) \right]$$

Suppose that we have some combination of points $x_1 \dots x_n$. Now we fix some j and ask if the function value is greater for $x_j = a_j$ or $x_j = b_j$.

Lemma 5 *Let* $x \in \mathbb{R}^n$ *and take an arbitrary index* j. *Denote* p_{x_i} *the average value of points* $\{x_i\}_{i \neq j}$. *Then*

$$F(\{x_i\}_{i \neq j}, b_j) - F(\{x_i\}_{i \neq j}, a_j) = 2 \frac{n-1}{n} d_j (c_j - p_{x_i})$$

The consequence is that:

$$F(\{x_i\}_{i \neq j}, b_j) > F(\{x_i\}_{i \neq j}, a_j) \quad \Leftrightarrow \quad c_j > p_{x_i},$$

$$F(\{x_i\}_{i \neq j}, b_j) < F(\{x_i\}_{i \neq j}, a_j) \quad \Leftrightarrow \quad c_j < p_{x_i},$$

$$F(\{x_i\}_{i \neq j}, b_j) = F(\{x_i\}_{i \neq j}, a_j) \quad \Leftrightarrow \quad c_j = p_{x_i}.$$

The special case of this lemma is the following result.

Lemma 6 *Consider sets* $\{a_i\}, \{b_i\}$ *and let* j *be an arbitrary index. Then*

$$F(\{a_i\}_{i \neq j}, b_j) - F(\{a_i\}) = 2 d_j (c_j - p_a - \frac{1}{2n} d_j),$$

$$F(\{b_i\}_{i \neq j}, a_j) - F(\{b_i\}) = 2 d_j (p_b - c_j - \frac{1}{2n} d_j).$$

Both numbers on the right-hand side are equal if and only if $c_j = p := \frac{p_a + p_b}{2}$.
In general, if we compare both numbers on the right-hand side, we will derive relations

$$F(\{a_i\}_{i \neq j}, b_j) - F(\{a_i\}) > F(\{b_i\}_{i \neq j}, a_j) - F(\{b_i\}) \quad \Leftrightarrow \quad c_j > p,$$

$$F(\{a_i\}_{i\neq j}, b_j) - F(\{a_i\}) < F(\{b_i\}_{i\neq j}, a_j) - F(\{b_i\}) \quad \Leftrightarrow \quad c_j < p,$$
$$F(\{a_i\}_{i\neq j}, b_j) - F(\{a_i\}) = F(\{b_i\}_{i\neq j}, a_j) - F(\{b_i\}) \quad \Leftrightarrow \quad c_j = p.$$

As Lemma 2 says that the solution must contain at least one a_i and at least one b_j, the consequence is that

a_j can be replaced with b_j if and only if $p_a < c_j - \frac{1}{2n} d_j$,

b_j can be replaced with a_j if and only if $p_b > c_j + \frac{1}{2n} d_j$.

The following lemma gives the comparison of function values on each side of the set K.

Lemma 7 *It holds*

$$F(a_1, \ldots, a_n) < F(b_1, \ldots, b_n) \quad \Leftrightarrow \quad p < \frac{\sum(c_i d_i)}{\sum d_i},$$
$$F(a_1, \ldots, a_n) > F(b_1, \ldots, b_n) \quad \Leftrightarrow \quad p > \frac{\sum(c_i d_i)}{\sum d_i},$$
$$F(a_1, \ldots, a_n) = F(b_1, \ldots, b_n) \quad \Leftrightarrow \quad p = \frac{\sum(c_i d_i)}{\sum d_i}.$$

Lemma 8 *Let j be an arbitrary index. Then*

$$F(\{a_i\}_{i\neq j}, b_j) < F(\{b_i\}_{i\neq j}, a_j) \quad \Leftrightarrow \quad (c_j - p)d_j < \textstyle\sum_{i\neq j}[(c_i - p)d_i],$$
$$F(\{a_i\}_{i\neq j}, b_j) > F(\{b_i\}_{i\neq j}, a_j) \quad \Leftrightarrow \quad (c_j - p)d_j > \textstyle\sum_{i\neq j}[(c_i - p)d_i],$$
$$F(\{a_i\}_{i\neq j}, b_j) = F(\{b_i\}_{i\neq j}, a_j) \quad \Leftrightarrow \quad (c_j - p)d_j = \textstyle\sum_{i\neq j}[(c_i - p)d_i].$$

Another approach to determine Lemma 5 consists in that we fix a set of points $\{x_i\}_{i\neq j}$ and study the function values on $[a_j, b_j]$. Denote

$$x_j = a_j + \tau_j(b_j - a_j) = a_j + \tau_j d_j, \quad \tau_j \in [0,1].$$

Then we have

$$f(\tau_j) \equiv F(\{x_i\}_{i\neq j}, x_j) = \sum_{i\neq j} x_i^2 + (a_j + \tau_j d_j)^2 - \frac{1}{n}\left(\sum_{i\neq j} x_i + a_j + \tau_j d_j\right)^2$$

$$f'(\tau_j) = 2d_j(a_j + \tau_j d_j) - \frac{2}{n} d_j \left(\sum_{i\neq j} x_i + a_j + \tau_j d_j\right)$$

$$f'(\tau_j) = 0 \quad \Leftrightarrow \quad \tau_j^* = \frac{p_{x_i} - a_j}{b_j - a_j}$$

$$f''(\tau_j) = 2\frac{n-1}{n} d_j^2 \quad \Rightarrow \quad f(\tau_j^*) = \min f(\tau_j)$$

From here we obtain properties mentioned in Lemma 5, that is

$$\tau_j^* < 0.5 \quad \Leftrightarrow \quad c_j > p_{x_i} \quad \Leftrightarrow \quad F(\{x_i\}_{i\neq j}, b_j) > F(\{x_i\}_{i\neq j}, a_j)$$
$$\tau_j^* > 0.5 \quad \Leftrightarrow \quad c_j < p_{x_i} \quad \Leftrightarrow \quad F(\{x_i\}_{i\neq j}, b_j) < F(\{x_i\}_{i\neq j}, a_j)$$
$$\tau_j^* = 0.5 \quad \Leftrightarrow \quad c_j = p_{x_i} \quad \Leftrightarrow \quad F(\{x_i\}_{i\neq j}, b_j) = F(\{x_i\}_{i\neq j}, a_j)$$

Remark 1 *Function* $f(\tau_j)$ *satisfies*

$$f(\tau_j) = \frac{n-1}{n} d_j^2 \tau_j^2 - 2\frac{n-1}{n} d_j (p_{x_i} - a_j)\tau_j + \sum_{i \neq j} x_i^2 + a_j^2 - \frac{1}{n}\left(\sum_{i \neq j} x_i + a_j\right)^2.$$

From Lemma 2 and Lemma 5 we can immediately determine some components of the solution. Denote

$p_j^a =$ the average value of $\{a_i\}_{i \neq j},\ j = 1 \ldots n$

$p_j^b =$ the average value of $\{b_i\}_{i \neq j},\ j = 1 \ldots n$

$p_\alpha =$ the average value of $\left\{\{a_i\}_{i \neq k}, b_k\right\}$, where k is such that $d_k = \min_i\{d_i\}$.

$p_\beta =$ the average value of $\left\{\{b_i\}_{i \neq k}, a_k\right\}$, where k is such that $d_k = \min_i\{d_i\}$.

Then it holds

If $c_j < \max\{p_j^a, p_\alpha\}$, then $x_j^* = a_j$.

If $c_j > \min\{p_j^b, p_\beta\}$, then $x_j^* = b_j$.

If $c_j \in P_j = [\max\{p_j^a, p_\alpha\}, \min\{p_j^b, p_\beta\}]$, an iteration process must be performed.

Now we give a preliminary algorithm for solving the maximization problem (5).

Algorithm 2 *Solving problem (5).*

1. **Initiation:** Set

   $$p_a = \text{ø} a_i, \quad p_b = \text{ø} b_i, \quad p = \frac{p_a + p_b}{2}, \quad c_i = \frac{a_i + b_i}{2}, \quad d_i = \frac{b_i - a_i}{2n}, \quad d = 2\min d_i,$$

   $s_i = 0\ \forall i$ (indicator of the solution x_i^*: 0 - not found, 1 - found).

2. For $i = 1 \ldots n$ such that $s_i = 0$:

   (a) **Test on components of the solution:**
   If $c_i > p_b - d$, then $x_i^* = a_i = b_i$ and $s_i = 1$.
   If $c_i < p_a + d$, then $x_i^* = b_i = a_i$ and $s_i = 1$.

   (b) Otherwise – **Iteration process:**
   If $c_i < p$, then $x_i = a_i$ and $b_i = \min\{b_i, p_a\}$.
   If $c_i > p$, then $x_i = b_i$ and $a_i = \max\{a_i, p_b\}$.

   (c) Otherwise – **Possible reduction of intervals:**
   If $a_i < p_a$ and $p_b < b_i$, then $a_i = p_b$ or $b_i = p_a$.

3. **Computation of new values:** Set $p^{old} = p$ and update p_a, p_b, p, c_i, d_i, d.

4. **Test on termination:** If $|p - p^{old}| > \varepsilon$ (e.g. $= 10^{-6}$), goto Step 2. Otherwise set $x_i^* = x_i$ for i such that $s_i = 0$.

Example 1 Let $n = 3$ and

$$[a_i,\, b_i] = [-3,\, 1],\, [-9/2,\, 3],\, [-1,\, 1/2]$$

It holds that

$$c_1 = -1, \quad d_1 = 4, \quad p_1^a = -11/4, \quad p_1^b = 7/4$$
$$c_2 = -3/4, \quad d_2 = 15/2, \quad p_2^a = -2, \quad p_2^b = 3/4$$
$$c_3 = -1/4, \quad d_3 = 3/2, \quad p_3^a = -15/4, \quad p_3^b = 2$$
$$p_\alpha = -7/3, \quad p_\beta = 1, \quad p = -2/3$$

We do not have immediately any component of the solution because $c_j \in P_j\ \forall j$. The solution satisfies

$$x^* = (-3,\, 3,\, -1), \quad \text{ø}x_i^* = -1/3, \quad F(x^*) = 6.\overline{2}$$

and the point just on the other sides of all interval satisfies

$$\tilde{x} = (1,\, -9/2,\, 1/2), \quad \text{ø}\tilde{x}_i = -1, \quad F(\tilde{x}) = 6.1\overline{6}$$

The function values are located close together which makes problems for the algorithm to identify the right solution.

Example 2 This example shows efficiency of our method and also that using a standard optimization method (in this case the line-search approach from the UFO system [5]) the right solution has not to be obtained. Let $n = 40$ and $[a_i,\, b_i] =$

$$[-47.5,\, 28.75],\, [47.25,\, 91.75],\, [38.5,\, 81.5],\, [-53.5,\, 45],\, [-46.5,\, 93.5],$$
$$[-98.25,\, -40.75],\, [-34.5,\, -28],\, [51,\, 64],\, [-88.5,\, -71.5],\, [-93.75,\, -33],$$
$$[40.75,\, 95],\, [-47,\, 30.5],\, [-95.75,\, -25.25],\, [-34,\, -28.25],\, [-1.25,\, 34.5],$$
$$[16.5,\, 81.25],\, [-21.75,\, 39.75],\, [30.25,\, 80.25],\, [-14.5,\, -6],\, [-22.5,\, 1.25],$$
$$[-10,\, 7.75],\, [-81.5,\, -72.25],\, [-94,\, -18.75],\, [-65.5,\, 22],\, [-3.25,\, 83.25],$$
$$[-90.25,\, -77],\, [-24.75,\, -1.25],\, [42.5,\, 79.75],\, [-11,\, -5],\, [-19.5,\, 6.75],$$
$$[27.75,\, 57],\, [49.75,\, 85.5],\, [12.75,\, 90.75],\, [18.5,\, 85.5],\, [-93,\, -83],$$
$$[-95.75,\, -52],\, [-66,\, -61.25],\, [-77.5,\, -42.75],\, [-32.5,\, -13.5],\, [-42.25,\, 20].$$

In the following table, we present the obtained results with differences in bold.

x^* using Algorithm 2	x^* using standard optimization method
-47.5, 91.75, 81.5, 45, 93.5, -98.25, -34.5, 64, -88.5, -93.75, 95, **-47**, -95.75, -34, 34.5, 81.25, 39.75, 80.25, -14.5, -22.5, 7.75, -81.5, -94, **-65.5**, 83.25, -90.25, -24.75, 79.75, -11, **-19.5**, 57, 85.5, 90.75, 85.5, -93, -95.75, -66, -77.5, -32.5, **-42.25**	**28.75**, 91.75, 81.5, 45, 93.5, -98.25, -34.5, 64, -88.5, -93.75, 95, **30.5**, -95.75, -34, 34.5, 81.25, 39.75, 80.25, -14.5, -22.5, 7.75, -81.5, -94, **22**, 83.25, -90.25, -24.75, 79.75, -11, **6.75**, 57, 85.5, 90.75, 85.5, -93, -95.75, -66, -77.5, -32.5, **20**
$F = 4911.99902$, #iterations $= 2$	$F = 4709.79625$, #iterations $= 41$

Note that we have found the same solution as in [3] in the second iteration while their genetic algorithm used several hundred iterations.

Conclusion

In this contribution we have presented some theoretical results and preliminary algorithms for computing variance of interval data described in the introduction. Although the problem can be solved using various optimization methods combining direction vectors and the step-length, developing a special algorithm is advantageous. Concerning the maximization problem, the solution satisfies useful properties which allow us to identify directly some components of the solution. The main task is to properly deal with the case when $c_j \in P_j$ and develop more robust algorithm to identify the right solution.

Acknowledgments

This work has been supported by the project ESF No. CZ.1.07/2.3.00/09.0155 “Constitution and improvement of a team for demanding technical computations on parallel computers at TU Liberec” and by the long-term strategic development financing of the Institute of Computer Science (RVO:67985807).

Literature

[1] ALEFELD, G. - HERZBERGER, J.: *Introduction to interval computations.* Academic Press, New York, 1983.

[2] ANTOCH, J.; BRZEZINA, M.; MIELE, R.: A note on variability of interval data. *Comput. Stat.* 25, No. 1, pp. 143-153, 2010.

[3] ANTOCH, J.; MIELE, R.: Use of Genetic Algorithms When Computing Variance of Interval Data. *Classification and Multivariate Analysis for Complex Data Structures*, Springer, pp. 371-377, 2011, ISBN 978-3-642-13311-4.

[4] FERSON, S.; GINZBURG, L.; KREINOVICH, V.; LONGPRE, L.; and AVILES, M.: Computing Variace for Interval Data is NP-Hard. *ACM SIGACT News* 33, pp. 108-118, 2002.

[5] LUKŠAN, L.; TŮMA, M.; VLČEK, J.; RAMEŠOVÁ, N.; ŠIŠKA, M.; HARTMAN, J.; MATONOHA, C.: *UFO 2011 - Interactive system for universal functional optimization*. Technical Report V-1151, ICS AS CR, Prague 2011.

[6] NOCEDAL, J.; WRIGHT, S.J.: *Numerical Optimization. Second Edition.* Springer, New York, 2006. ISBN 978-0387-30303-1; 0-387-30303-0.

[7] XIANG, G.: Fast Algorithm for Computing the Upper Endpoint of Sample Variance for Interval Data: Case of Sufficiently Accurate Measurements. *Reliable Computing* 12, pp. 59-64, 2006.

Problém variability intervalových dat

V příspěvku se zabýváme problémem variability intervalových dat. Pro daná intervalová data se jedná o nalezení dolních a horních mezí intervalu možných hodnot jejich rozptylu. Toto vede na problém minimalizace/maximalizace součtu čtverců rozdílu složek *n*-dimensionálního vektoru a jejich průměrné hodnoty. Jelikož je problém maximalizace mnohem obtížnější než problém minimalizace, uvedeme některé teoretické výsledky týkající se řešení. Také uvedeme předběžné algoritmy pro nalezení řešení obou problémů, které využívají jejich speciální vlastnosti.

Das Problem der Variabilität von Intervalldaten

In diesem Beitrag befassen wir uns mit dem Problem der Variabilität von Intervalldaten. Bei den vorliegenden Intervalldaten handelt es sich um die Auffindung der Unter- und Obergrenzen der möglichen Werte ihrer Zerstreuung. Dies führt zum Problem der Minimalisierung/ Maximalisierung der Summe der Quadrate des Unterschieds der Komponenten eines *n*-dimensionalen Vektors und deren Durchschnittswert. Da das Problem der Maximalisierung viel schwieriger ist das Problem der Minimalisierung, führen wir einige theoretische Ergebnisse an, welche mit der Lösung zu tun haben, ebenso vorläufige Algorithmen zur Findung einer Lösung für beide Probleme, welche deren speziellen Eigenschaften nutzen.

Problem zmienności danych interwałowych

W artykule przedstawiono problem zmienności danych interwałowych. W przypadku określonych danych interwałowych chodzi o znalezienie dolnych i górnych granic interwału możliwych wartości ich rozproszenia. W związku z tym pojawia się problem minimalizacji/maksymalizacji sumy kwadrarów różnicy elementów *n*-wymiarowego wektora oraz ich przeciętnej wartości. Problem maksymalizacji jest o wiele trudniejszy w porównaniu z problemem minimalizacji, dlatego wskazano niektóre teoretyczne wnioski dotyczące rozwiązania. Ponadto przedstawiono wstępne algorytmy służące do znalezienia rozwiązania obu problemów, które wykorzystują ich szczególne cechy.

OPTIMIZATION PROBLEMS UNDER TWO-SIDED $(\max,\min)-$LINEAR INEQUALITIES CONSTRAINTS

Mahmoud Gad

Charles University in Prague
Faculty of Mathematics and Physics
Sokolovská 83, Praha 8, 186 75, Czech Republic

*Sohag University
Faculty of Science
Sohag, Egypt
Mahmoud_Attya_Or@yahoo.com

Abstract

Systems of so called two-sided $(\max,\min)-$linear inequalities with variables on both sides will be studied. Optimization problems, the objective function of which is equal to the maximum of a finite number of continuous functions of one variable are considered. The set of feasible solutions in described by a system of two-sided $(\max,\min)-$linear inequalities with variables on both sides. A finite algorithm for finding the optimal solution of the problem is proposed.

Keywords: Two-sided $(\max,\min)-$linear inequalities system; lower and upper bounds; max-min optimization problems.

Introduction

The algebraic structures in which $(\max,+)$ or $(\max,\min)$ replace addition and multiplication of the classical linear algebra have been appeared in the literature approximately since the sixties of the last century (see e.g. [1], [3], and [8]). In recently published book [2] readers can find the latest results concerning theory and algorithms for $(\max,+)-$linear systems of equations. A polynomial method for finding the maximum solution of the $(\max,\min)$-linear system has been proposed in [5]. A finite algorithm for finding the optimal solution of the optimization problems under $(\max,+)-$linear constraints has been introduced in [10]. A survey of some of the recent results concerning the $(\max,\min)$-linear systems of equations and inequalities and optimization problems under the constraints described by such systems of equations and inequalities is presented in [6]. Algorithm for optimization problems under one-sided $(\max,\min)-$linear equality constraints is introduced in [4]. Maximal solutions of two-sided linear systems in max-min algebra have been given in [7]. A note on application of two-sided systems of $(\max,\min)-$linear equations and inequalities to some fuzzy set problems has been given in [9].

In this contribution, we will study systems of so called $(\max,\min)-$linear (or using an alternative notation $(\max,\wedge)-$linear) inequalities with variables on both sides. We consider optimization problems, the objective function of which is equal to the maximum of a finite number of continuous and unimodal functions of one variable. The set of feasible solutions is described by a system of $(\max,\wedge)-$linear inequalities with variables on both sides. Let us note that if we have variables x on the left hand sides and different variables y on the right hand sides, the system can be processed like the one-sided system considered e.g. in [6]. Including lower and upper bounds on x, y is only a technical problem.

We can consider the practical problem, in which transportation means of different size are

transporting goods from places $i \in I$ to one terminal T. The goods are unloaded in T and the transportation means (possibly with other goods are uploaded in T) have to return to i. We assume that the connection between i and T is only possible via one of the places (e.g. cities) $j \in J$ the roads between i and j are one-way roads, and the capacity of the road between $i \in I$ and $j \in J$ is equal to a_{ij}. We have to join places j with T by a two-way road with a capacity x_j in both directions. The total capacity of the connection between i and T is therefore equal to $\max_{j\in J}(a_{ij} \wedge x_j)$. The transport from T to i is carried out via other one-way roads between places $j \in J$ and $i \in I$ with (in general, different) capacities between j and i are equal to b_{ij}. Since the roads between T and j are two-way roads, the total capacity of the connection between T and i is equal to $\max_{j\in J}(b_{ij} \wedge x_j)$, for all $i \in I$. We assume that the transportation means can only pass through some roads with the capacity which is not smaller than the capacity of the transportation mean and our task is to choose appropriate capacities $x_j, j \in J$. In order that each of the transportation means may return to i, we may e.g. require for each i that the maximal attainable capacity of connections between i and T via j is greater than or equal to maximal attainable capacity of connections between T and i on the way back. In other words, we have to choose $x_j, j \in J$, which satisfy relation (1) below. In what follows, assume that we have the same variables on the left hand sides and right hand sides of the inequality system.

1 Systems of $(\max, \min)-$Linear Inequalities

Let us consider the following system of inequalities:

$$a_i(x) \geq b_i(x), i \in I, \tag{1}$$

where $a_i(x) = \max_{j\in J}(a_{ij} \wedge x_j)$, $b_i(x) = \max_{j\in J}(b_{ij} \wedge x_j)$, and a_{ij}, $b_{ij} \in R$, $i \in I$, $j \in J$ be given numbers. Let $M^{\geq}$ denote the set of all solutions of system (1). We will set for any $x, y \in R^n : x \leq y \Leftrightarrow x_j \leq y_j \ \ \forall j \in J$. Let us set $M^{\geq}(\underline{x}, \overline{x}) = \{x \ ; \ x \in M^{\geq} \ \& \ \underline{x} \leq x \leq \overline{x}\}$ for any finite $\underline{x} \leq \overline{x}$ and let $x^{\max}$ denote the maximum element of $M^{\geq}(\underline{x}, \overline{x})$. So that $M^{\geq}(\underline{x}, \overline{x}) \subset M^{\geq}$, and $M^{\geq}(\underline{x}, x^{\max}) \subset M^{\geq}$, also it is clear $M^{\geq}(\underline{x}, x^{\max}) \subseteq M^{\geq}(\underline{x}, \overline{x})$. To prove $M^{\geq}(\underline{x}, \overline{x}) \subseteq M^{\geq}(\underline{x}, x^{\max})$ there are two cases: the first one, if $\overline{x} \notin M^{\geq}$, then $x^{\max} < \overline{x}$. Therefore $\forall\, x \in M^{\geq}(\underline{x}, \overline{x})$, the inequality $x \leq x^{\max}$ verified, i.e. $x_j \leq x_j^{\max} \ \forall \ j \in J$ and if $x^* \in (x^{\max}, \overline{x}]$, (i.e. $x^{\max} < x^* \leq \overline{x}$, i.e. $x_{j_0}^{\max} < x_{j_0}^* \leq \overline{x}_{j_0}$ for at least one $j_0 \in J$ and $x_j^{\max} \leq x_j^* \leq \overline{x}_j$ for $j \in J \ \& \ j \neq j_0$) then $x^* \notin M^{\geq}$, otherwise x^* is the maximum element of $M^{\geq}(\underline{x}, \overline{x})$, but this contradicts the hypothesis $x^{\max}$ is the maximum element of $M^{\geq}(\underline{x}, \overline{x})$. So that for any $x \in M^{\geq}(\underline{x}, \overline{x})$, we have $x \leq x^{\max}$, and $x \in M^{\geq}(\underline{x}, x^{\max})$, then $M^{\geq}(\underline{x}, \overline{x}) \subseteq M^{\geq}(\underline{x}, x^{\max})$. The second case, if $\overline{x} \in M^{\geq}$, then $x^{\max} = \overline{x}$. Then we have $M^{\geq}(\underline{x}, x^{\max}) = M^{\geq}(\underline{x}, \overline{x}) \subset M^{\geq}$. In this section we will propose an algorithm, which find the maximum element of the set $M^{\geq}(\underline{x}, \overline{x})$, and calculates the maximum solution of system (1), take in account $\underline{x} \leq x \leq \overline{x}$. Note that, since any equation can be replaced by two inequalities, therefor we can use the next algorithm to find the maximum element of the set $M^{=}(\underline{x}, \overline{x})$, which is the set of all solutions of a system of equations, $(a_i(x) = b_i(x), i \in I)$.

Algorithm 1

0 Input I, J, $\overline{x}$, a_{ij} and b_{ij} for all $i \in I$ and $j \in J$.

1 Find $I^{<}(\overline{x}) \equiv \{i \in I \ ; \ a_i(\overline{x}) < b_i(\overline{x})\}$.

2 If $I^{<}(\overline{x}) = \emptyset$, then $x^{\max} := \overline{x}$, STOP.

3 Find $\alpha(\overline{x}) \equiv \min_{i\in I^{<}(\overline{x})} a_i(\overline{x})$.

4 Find $I^{<}(\alpha(\bar{x})) \equiv \{i \in I^{<}(\bar{x}) \; ; \; a_i(\bar{x}) = \alpha(\bar{x})\}$.

5 Find $H_i^{<}(\bar{x}) \equiv \left\{ j \in J \; ; b_{ij} \wedge \bar{x}_j > \alpha(\bar{x}) \right\}, \; \forall \, i \in I^{<}(\alpha(\bar{x}))$.

6 Set $H^{<}(\bar{x}) := \bigcup_{i \in I^{<}(\alpha(\bar{x}))} H_i^{<}(\bar{x})$.

7 Set $\bar{x}_j := \alpha(\bar{x})$ for all $j \in H^{<}(\bar{x})$ go to 1.

We will illustrate the performance of this algorithm by the following small numerical example.

Example 1. *: Let $J = \{1,2,3,4\}$, $I = \{1,2,3\}$, $\bar{x} = (10,10,10,10)$, and consider system (1) of inequalities where a_{ij} & b_{ij} $\forall$ $i \in I$ and $j \in J$ are given by the matrices A and B as follows:*

$$A = \begin{pmatrix} 7 & 5 & 3 & 0 \\ 4 & 3 & 1 & 2 \\ 10 & 20 & 10 & -1 \end{pmatrix}, \quad B = \begin{pmatrix} 6 & 13 & 10 & -1 \\ 8 & 0 & 3 & 1 \\ 1 & 1 & 1 & -8 \end{pmatrix}$$

By substitution for these values in system (1) and using Algorithm 1:

***Iteration 1**:*

1 $I^{<}(\bar{x}) = \{1,2\}$.

2 $I^{<}(\bar{x}) \neq \emptyset$.

3 $\alpha(\bar{x}) = \min(7,4) = 4$.

4 $I^{<}(\alpha(\bar{x})) = \{2\}$.

5 $H_2^{<}(\bar{x}) = \{1\}$.

6 $H^{<}(\bar{x}) = \{1\}$.

7 $\bar{x}_1 = 4$, $\bar{x} = (4,10,10,10)$ *go to* 1.

***Iteration 2**:*

1 $I^{<}(\bar{x}) = \{1\}$.

2 $I^{<}(\bar{x}) \neq \emptyset$.

3 $\alpha(\bar{x}) = 5$.

4 $I^{<}(\alpha(\bar{x})) = \{1\}$.

5 $H_i^{<}(\bar{x}) = \{2,3\}$.

6 $H^{<}(\bar{x}) = \{2,3\}$.

7 $\bar{x}_2 = 5, \bar{x}_3 = 5$, $\bar{x} = (4,5,5,10)$ *go to* 1.

***Iteration 3**:*

1 $I^{<}(\bar{x}) = \emptyset$, *then* $x^{\max} = (4,5,5,10)$ *STOP.*

In the next part of this section we will introduce a method which finds the minimum upper bound $\tilde{x}$ for solution of system (1) such that $\tilde{x} \geq \underline{x}$. In other words $\tilde{x}$ has the properties $\tilde{x} \in M^{\geq}(\underline{x}, x^{\max})$ and if $\underline{x} \leq \tilde{x}, \underline{x} \neq \tilde{x}$, then there exists $x^* \in M^{\geq}(\underline{x}, x^{\max})$ such that $x^* \not\leq \tilde{x}$. It will be clear that $\tilde{x} \in M^{\geq}(\underline{x}, x^{\max})$ and this element is suitable to find the optimal solution of the minimization problem as we will see in the next section. In what follows to simplify the notation we set for any $\alpha, \beta \in R : \alpha \vee \beta = \max(\alpha, \beta)$. Let us set

$$T_{ij} = \{x_j \, ; x_j \leq x_j^{\max} \& \; a_{ij} \wedge x_j \; \geq \; b_i(\underline{x}) \; \vee \; \underline{x}_j\}, \; \forall \, i \in I, \; j \in J.$$

Note that if $i_1, \; i_2$ are two different indices of $I, \; j \in J$, and $b_{i_2}(\underline{x}) \vee \underline{x}_j \leq b_{i_1}(\underline{x}) \vee \underline{x}_j$, then evidently $T_{i_1 j} \subseteq T_{i_2 j}$. It follows that for any subset of r indices of I, there exists such permutation $i_1, \; , \ldots, \; i_r$ of these indices that the inclusions $T_{i_1 j} \; \subseteq \; T_{i_2 j} \; \subseteq \; \ldots \; \subseteq T_{i_r j}$ hold so that $\bigcap_{h=1}^{r} T_{i_h j} = T_{i_1 j}$. Sets T_{ij} have the following properties:

$$T_{ij} \neq \emptyset \; \Leftrightarrow \; a_{ij} \geq b_i(\underline{x}) \vee \underline{x}_j,$$

$$T_{ij} \neq \emptyset \; \Rightarrow \; T_{ij} = [b_i(\underline{x}) \vee \underline{x}_j \; , \; x_j^{\max}].$$

Since we assumed that $\underline{x} \leq x^{\max}$, set $M^{\geq}(\underline{x}, x^{\max})$ is nonempty. Let us note that for any $x \in M^{\geq}(\underline{x}, x^{\max})$ and any $i \in I$, the inequalities $b_i(x) \geq b_i(\underline{x}) \; \& \; x_j \geq \underline{x}_j \; \forall \; j \in J$ hold and further there exists for each $i \in I$ an index $j(i) \in J$ such that $T_{ij(i)} \neq \emptyset$ (otherwise set $M^{\geq}(\underline{x}, x^{\max})$ would be empty, because we would have $a_{ij} < b_i(\underline{x}) \vee \underline{x}_j \; \forall j \in J$ and therefore $a_i(x) < b_i(x)$ for any $x \in R^n$ and we have $\underline{x} \leq x^{\max}$ so that $M^{\geq}(\underline{x}, x^{\max}) \neq \emptyset$). Let us note further, that if $a_{ij} \wedge x_j < b_i(\underline{x}) \vee \underline{x}_j \; \forall j \in J$, then we have $a_i(x) < b_i(\underline{x})$ and thus $x \notin M^{\geq}(\underline{x}, x^{\max})$. If for some fixed $j \in J$ the inequalities $a_{ij} < b_i(\underline{x}) \vee \underline{x}_j$ hold , then $a_{ij} \wedge x_j < b_i(\underline{x}) \vee \underline{x}_j \; \forall \, x_j \in R$ so that $T_{ij} = \emptyset$ and x_j will never be "active" in $a_i(x)$ or $b_i(x)$ if $x \in M^{\geq}$ (i.e. it will never determine the values of $a_i(x)$ or $b_i(x)$). We will exclude such variables from our considerations and assume that for each $j \in J$ there exists at least one "row" index $i \in I$ such that $a_{ij} \geq b_i(\underline{x}) \vee \underline{x}_j$. We define sets $V_j, \;\; j \in J$

$$V_j = \{i \in I; a_{ij} \geq b_i(\underline{x}) \vee \underline{x}_j\},$$

and denote $\max_{k \in V_j}(b_k(\underline{x})) = b_{k(j)}(\underline{x})$. A vector $\tilde{x}$ will be defined as follows:

$$\tilde{x}_j = \max_{k \in V_j}(b_k(\underline{x})) \vee \underline{x}_j = b_{k(j)}(\underline{x}) \vee \underline{x}_j \;\; \forall j \in J. \tag{2}$$

The element $\tilde{x}$ defined by (2) has the following properties:

(1) $M^{\geq}(\underline{x}, \tilde{x}) \neq \emptyset, \; \& \; \tilde{x} \in M^{\geq}(\underline{x}, \tilde{x})$.

(2) $\xi \in M^{\geq}(\underline{x}, \tilde{x}) \Rightarrow \underline{x} \leq \xi \leq \tilde{x}$.

(3) There may exist elements $\eta \in M^{\geq}(\underline{x}, \tilde{x})$ such that $\eta \neq \tilde{x}$.

If $\tilde{x}$ is the minimum element of $M^{\geq}(\underline{x}, x^{\max})$, then it would be $\tilde{x} \in M^{\geq}(\underline{x}, x^{\max})$ and for any $x \in M^{\geq}(\underline{x}, x^{\max}) \Rightarrow x \geq \tilde{x}$. Therefore, because of the property (3) $\tilde{x}$ is not the minimum element of $M^{\geq}(\underline{x}, x^{\max})$, but we can say that $\tilde{x}$ is the minimum upper bound of $M^{\geq}(\underline{x}, x^{\max})$ such that $M^{\geq}(\underline{x}, \tilde{x}) \neq \emptyset$. Let us choose $\tau \leq x^{\max}$, $\& \; \tau \neq x^{\max}$, and $\check{x} \in M^{\geq}(\underline{x}, \tau) \; \Rightarrow \; \check{x} \leq x^{\max}$ and $a_i(\check{x}) \geq b_i(\check{x}) \; \forall \, i \in I$ and $\underline{x} \leq \check{x} \leq \tau$. Let $H = \left\{x^{\max}(\tau) \mid x^{\max}(\tau) \; \textit{is the maximum element of} \; M^{\geq}(\underline{x}, \tau)\right\}$, then $\tilde{x}$ is the minimum element of H.

Theorem 1. *: Let $\tilde{x}$ be defined as in (2). Then $\tilde{x} \in M^{\geq}(\underline{x}, x^{\max})$.*

Proof: Since evidently $\tilde{x} \geq \underline{x}$, we have to prove that only $a_i(\tilde{x}) \geq b_i(\tilde{x}),\ \forall i \in I$. Let $i \in I$ be arbitrarily chosen. We have

$$b_i(\tilde{x}) = \max_{j\in J}(b_{ij} \wedge \tilde{x}_j) = \max_{j\in J}(b_{ij} \wedge (\max_{k\in V_j}(b_k(\underline{x}) \vee \underline{x}_j))) = \max_{j\in J}(b_{ij} \wedge (b_{k(j)}(\underline{x}) \vee \underline{x}_j))$$

Let us assume that

$$b_i(\tilde{x}) = \max_{j\in J}(b_{ij} \wedge \tilde{x}_j) = b_{ij(i)} \wedge \tilde{x}_{j(i)}.$$

Since in this case $i \in V_{j(i)}$, we have $a_{ij(i)} \geq \tilde{x}_{j(i)}$ and we obtain $a_i(\tilde{x}) \geq a_{ij(i)} \wedge \tilde{x}_{j(i)} = \tilde{x}_{j(i)} \geq b_{ij(i)} \wedge \tilde{x}_{j(i)} = b_i(\tilde{x})$. Since $i \in I$ was arbitrarily chosen, the theorem is proved.

□

Element $\tilde{x}$ defined by (2) shows that the given lower bound $\underline{x}$ might not be an element of $M^{\geq}(\underline{x}, x^{\max})$. Moreover we obtained an explicit dependence of $\tilde{x}$ on the given lower bound $\underline{x}$ (compare (2)), which can be used for sensitivity analysis of the set $M^{\geq}(\underline{x}, \overline{x})$ or for a post optimal analysis of optimization problems, the set of feasible solutions equal to $M^{\geq}(\underline{x}, x^{\max})$. The properties of $\tilde{x}$ enable us to solve some of the optimization problems mentioned above explicitly.

2 Optimization Problems under Two-Sided $(\max, \min)-$Linear Inequalities Constraints

In this section we consider an optimization problem that is a combination of the problems solved in the above chapters but with a different feasible set. In other words, let us consider for instance the optimization problem:

$$f(x) \equiv \max_{j\in J} f_j(x_j) \quad \longrightarrow \quad \min \tag{3}$$

subject to $x \in M^{\geq}(\underline{x}, x^{\max})$, where $f_j,\ j \in J$ are increasing functions. Let indices $j(i) \in J$ will be chosen for each $i \in I$ such that $\min_{j\in J} f_j(x_j^{(i)}) = f_{j(i)}(x_{j(i)})$, where $f_j(x_j^{(i)}) = \min_{x_j \in T_{ij}} f_j(x_j)$. Let $\tilde{x}$ be defined as in (2) and then we have to proceed as follows:

$$\tilde{T}_{ij} = \begin{cases} \emptyset & \text{if} \quad a_{ij} < b_i(\underline{x}), \\ b_i(\underline{x}) & \text{if} \quad a_{ij} > b_i(\underline{x}), \\ [\underline{x}_j, \tilde{x}] & \text{if} \quad a_{ij} = b_{ij}. \end{cases}$$

Set $f_j(\tilde{x}_j^{(i)}) = \min_{x_j \in \tilde{T}_{ij}} f_j(x_j)$, (if $\tilde{T}_{ij} = \emptyset$, we set minimum equal to $+\infty$). Let us set

$$\min_{j\in J} f_j(\tilde{x}_j^{(i)}) = f_{j(i)}(\tilde{x}_{j(i)}^{(i)}).$$

And $\tilde{R}_j = \{i \in I \mid j(i) = j\},\ \forall\ j \in J$, (it may be $\tilde{R}_j = \emptyset$ for some j). Then we have

$$f_k(x_k^{opt}) = \max_{i\in \tilde{R}_k} f_k(\tilde{x}_k^{(i)}),$$

if $\tilde{R}_k \neq \emptyset$, but when $\tilde{R}_k = \emptyset$, we set

$$f_k(x_k^{opt}) = f_k(\underline{x}_k).$$

The proof can be carried out in the same way as in the one sided case in [6]. We mentioned above that a system of inequalities can be transformed to a system of equations by making use of slack variables. Let us note that the other way round, systems of equations considered can be solved alternatively by the methods in this section, if we replace the equation system by the system of inequalities of the form

$$a_i(x) \geq b_i(x),\ i \in I$$
$$b_i(x) \geq a_i(x),\ i \in I$$
$$x_j \geq \underline{x}_j,\ j \in J.$$

We will describe now the corresponding algorithm explicitly step by step.

Algorithm 2

0 Input $m, n, \underline{x}, \overline{x}, A, B, f(x)$.

1 Find $x^{\max} \in M^{\geq}(\underline{x}, \overline{x})$.

2 If $\underline{x} \not\leq x^{\max}$, then $M^{\geq}(\underline{x}, \overline{x}) = \emptyset$, STOP.

3 $V_j := \{i \in I\ ;\ a_{ij} > b_i(\underline{x}) \vee \underline{x}_j\}\ \ \forall j \in J$.

4 $x_j^{(i)} := (b_i(\underline{x}) \vee \underline{x}_j)\ \ \forall i \in V_j$ for all $j \in J$ such that $V_j \neq \emptyset$.

5 Set $\tilde{x}_j := \max_{i \in V_j}(x_j^{(i)})$ if $V_j \neq \emptyset$, $\tilde{x}_j := \underline{x}_j$ if $V_j = \emptyset$.

6 $Q := \{k \in J\ ;\ f(\tilde{x}) = f_k(\tilde{x}_k)\},\ \ P := \{j \in J\ ;\ \tilde{x}_j = \underline{x}_j\}$.

7 If $Q \cap P \neq \emptyset$, then set $x^{opt} := \tilde{x}$, STOP.

8 $P_k := \{i \in I\ ;\ \tilde{x}_k = x_k^{(i)}\}\ \ \forall k \in Q$.

9 $V_k := V_k \setminus P_k\ \ \forall k \in Q$.

10 If $\bigcup_{j \in J} V_j = I$, go to 4.

11 Set $x^{opt} := \tilde{x}$, STOP.

We will illustrate the performance of this algorithm by the following numerical examples.

Example 2. *: Let $J = \{1, 2, \ldots, 5\}$, $I = \{1, 2, 3\}$, $\overline{x} = (10, 10, 10, 10, 10)$, $\underline{x} = (0, 3, 0, 0, 1)$ and consider system (1) of inequalities where a_{ij} & b_{ij} $\forall$ $i \in I$ and $j \in J$ are given by the matrices A and B as follows:*

$$A = \begin{pmatrix} -10 & 10 & 15 & -9 & -8 \\ 5 & -8 & 10 & 20 & 7 \\ 3 & 4 & -18 & 19 & 11 \end{pmatrix}, \quad B = \begin{pmatrix} 7 & 2 & -10 & -20 & 6 \\ 8 & 9 & -15 & -25 & 5 \\ 13 & -17 & 12 & 10 & 9 \end{pmatrix}$$

and consider the objective function $f(x) = \max(x_1, x_2 - 3, x_3, x_4, x_5)$. By substitution for these values in system (1) and using Algorithm 1 and Algorithm 2:

1 $x^{\max} = \overline{x} = (10, 10, 10, 10, 10)$.

2 $\underline{x} \leq x^{\max}$.

3 $V_1 = \{2\},\ V_2 = \{1, 3\}, V_3 = \{1, 2\}, V_4 = \{2, 3\}, V_5 = \{2, 3\}$.

$\boxed{4}$ $x_1^{(1)}=2,\quad x_2^{(1)}=3,\quad x_3^{(1)}=2,\quad x_4^{(1)}=2,\quad x_5^{(1)}=2, x_1^{(2)}=3,\quad x_2^{(2)}=3,\quad x_3^{(2)}=3,$
$x_4^{(2)}=3,\quad x_5^{(2)}=3, x_1^{(3)}=1,\quad x_2^{(3)}=3,\quad x_3^{(3)}=1,\quad x_4^{(3)}=1,\quad x_5^{(3)}=1.$

$\boxed{5}$ $\tilde{x}=(3,3,2,3,3).$

$\boxed{6}$ $Q=\{1,4,5\},, f(\tilde{x})=3,\ P=\{2\}$ *then* $Q\cap P=\emptyset.$

$\boxed{8}$ $P_1=\{2\},\ P_2=\{1,2,3\},P_3=\{1\},P_4=\{2\},P_5=\{2\}.$

$\boxed{9}$ $V_1=\emptyset,\ V_2=\emptyset,V_3=\{2\},V_4=\{3\},V_5=\{3\}.$

$\boxed{10}$ $\bigcup_{j\in J}V_j=\{2,3\}\neq I.$

$\boxed{11}$ $x^{opt}=\tilde{x}$, *STOP.*

Then $x^{opt}=(3,3,2,3,3)$ *is the optimal solution of the set* $M^{\geq}(\underline{x},\overline{x})$ *and* $f(x^{opt})=\max(3,0,2,3,3)$, *then the objective function is equal to* 3.

Example 3. *: Let* $J=\{1,2,\ldots,5\}$, $I=\{1,2,\ldots,6\}$, $\overline{x}=(20,20,20,20,20)$, $\underline{x}=(0,3,0,0,0)$ *and consider system (1) of inequalities where* a_{ij} & b_{ij} $\forall$ $i\in I$ *and* $j\in J$ *are given by the matrices A and B as follows:*

$$A=\begin{pmatrix}2&2&6&0&13\\8&11&10&7&7\\4&3&0&13&8\\14&3&3&13&2\\1&3&13&4&2\\12&15&7&3&14\end{pmatrix},\quad B=\begin{pmatrix}0&10&9&-1&5\\3&-3&1&-6&-7\\4&-8&2&-14&11\\14&-7&7&-3&4\\6&-8&12&2&0\\0&-11&2&-3&5\end{pmatrix}$$

and consider the objective function $f(x)=\max_{j\in J}(f_j(x_j)$, *where* $f_j(x_j)=c_jx_j+d_j$, $c=(6,3,7,3,7)$ *and* $d=(10,0,5,1,7)$. *By substitution for these values in system (1) and using Algorithm 1 and Algorithm 2:*

$\boxed{1}$ $x^{\max}=\overline{x}=(20,20,20,20,20).$

$\boxed{2}$ $\underline{x}\leq x^{\max}.$

$\boxed{3}$ $V_1=\{2,3,4,5,6\},\ V_2=\{2,6\},V_3=\{1,2,4,5,6\},V_4=\{2,3,4,5,6\},V_5=\{1,2,3,4,5,6\}.$

$\boxed{4}$ *find* $x_j^{(i)}$.

$\boxed{5}$ $\tilde{x}=(0,3,3,0,3).$

$\boxed{6}$ $Q=\{5\},f(\tilde{x})=28,\ P=\{1,2,4\}$ *then* $Q\cap P=\emptyset.$

$\boxed{10}$ $\bigcup_{j\in J}V_j=\{1,2,3,4,5,6\}=I$ *go to* $\boxed{4}$.

$\boxed{4}$ *find* $x_j^{(i)}$.

$\boxed{5}$ $\tilde{x}=(0,3,3,0,0).$

$\boxed{6}$ $Q = \{3\}, f(\tilde{x}) = 26,\ P = \{1,2,4,5\}$ *then* $Q \cap P = \emptyset$.

$\boxed{10}$ $\bigcup_{j \in J} V_j = \{1,2,4,5,6\} \neq I$.

$\boxed{11}$ $x^{opt} = \tilde{x}$, *STOP.*

Then $x^{opt} = (0,3,3,0,0)$ *is the optimal solution of the set* $M^{\geq}(\underline{x}, \overline{x})$ *and* $f(x^{opt}) = \max(10,9,26,1,7)$, *then the objective function is equal to* 26.

Conclusion

We can summarize the properties of the systems of $(\max, \min)$-linear inequalities studied in this paper as follows:
(1) Any system of two-sided $(\max, \min)$-linear inequalities is solvable and has a unique maximum element $x^{\max}(A,B)$ depending on the matrices $A,\ B$ with finite elements $a_{ij},\ b_{ij}$ (note that including infinite elements can cause nonsolvability of the system).
(2) If we include an additional requirement $x \leq \overline{x}$, then the system is also solvable and has the maximum element $x^{\max}(A,\ B,\ \overline{x}) \leq x^{\max}(A,B)$.
(3) The system with a finite lower bound on variables (i.e. with an additional constraint $x \geq \underline{x}$) is solvable if and only if $\underline{x} \leq x^{\max}(A,B)$, or in case of the additional upper bound $\overline{x}$ if and only if $\underline{x} \leq x^{\max}(A,\ B,\ \overline{x})$.

Acknowledgments

This work was supported by The Ministry of Higher Education and Scientific Research of the Arab Republic of Egypt. The author also offers sincere thanks to Prof. Karel Zimmermann for his continuing support and eternally great advice.

Literature

[1] BUTKOVIČ, P.; HEGEDÜS, G.: An Elimination Method for Finding All Solutions of the System of Linear Equations over an Extremal Algebra, *Ekonomicko–matematický obzor* **20**, 1984, pp. 203-215.

[2] BUTKOVIČ, P.: *Max-linear Systems: Theory and Algorithms*, Springer Monographs in Mathematics, 267 p., Springer-Verlag, London, 2010.

[3] CUNINGHAME-GREEN, R. A.: *Minimax Algebra*. Lecture Notes in Economics and Mathematical Systems **166**, Springer-Verlag, Berlin 1979.

[4] GAD, M.: Optimization problems under one-sided $(\max, \min)-$linear inequalities constraints (to appear).

[5] GAVALEC, M.; ZIMMERMANN, K.: Solving Systems of Two-Sided $(\max, \min)$-Linear Equations, *Kybernetika* **46**, 2010, pp. 405-414.

[6] GAVALEC, M.; GAD, M.; ZIMMERMANN, K.: Optimization problems under $(\max, \min)-$linear equation and/or inequality constraints (to appear).

[7] KRBÁLEK, P.; POZDÍLKOVÁ, A.: Maximal solutions of two-sided linear systems in max-min algebra, *Kybernetika* **46**, 2010, pp. 501-512.

[8] VOROBJOV, N. N.: Extremal Algebra of positive Matrices, *Datenverarbeitung und Kybernetik* **3**, 1967, pp. 39-71 (in Russian).

[9] ZIMMERMANN, K.: A Note on Application of Two-sided Systems of $(\max,\min)-$Linear Equations and Inequalities to Some Fuzzy Set Problems, *Acta Univ. Palacki. Olomuc., Fac. rer. nat., Mathematica 50,* 2, 2011, pp. 129-135.

[10] ZIMMERMANN, K.; GAD, M.: Optimization Problems under $(\max,+)-$ linear Constraints, *International conference presentation of mathematics '11 (ICPM '11),* Liberec, 11, pp. 159-165, 2011. ISBN 978-80-7372-773-4.

Optimalizační problémy při omezeních ve tvaru soustav dvoustranných $(\max, \min)$−lineárních nerovností

Zkoumají se soustavy tzv. dvoustranných $(\max, \min)$−lineárních nerovností s proměnnými na obou stranách těchto nerovností. Zabýváme se optimalizačními úlohami, jejichž účelová funkce je rovna maximu konečného počtu spojitých funkcí jedné proměnné. Množina přípustných řešení těchto úloh je popsána soustavou dvoustranných $(\max, \min)$−lineárních nerovností. Je navržen konečný algoritmus pro nalezení optimálního řešení zkoumaného optimalizačního problému.

Optimalisierungsprobleme bei Begrenzungen in der Form von zweiseitigen $(\max, \min)$−linearer Ungleichheitssystemen

Es werden sog. zweiseitige $(\max, \min)$−lineare Ungleichheitssysteme mit Variablen auf beiden Seiten dieser Ungleichheiten untersucht. Wir befassen uns mit Optimalisierungsaufgaben, deren Zweckfunktion dem Maximum einer finiten Anzahl kontinuierlicher Funktionen einer Variablen gleich ist. Die Menge der zulässigen Lösungen dieser Aufgaben wird durch zweiseitige $(\max, \min)$−lineare Ungleichheitssysteme beschrieben. Es wird ein finiter Algorithmus zur Auffindung einer optimalen Lösung der untersuchten Optimalisierungsprobleme vorgeschlagen.

Problemy optymalizacji przy ograniczeniach w postaci układów dwustronnych $(\max, \min)$−liniowych nierówności

Badaniem objęto układy tzw. dwustronnych $(\max, \min)$−nierówności liniowych ze zmiennymi po obu stronach tych nierówności. W artykule przedstawiono zadania optymalizacyjne, których funkcja celowa jest równa maksymum skończonej liczby funkcji ciągłych jednej zmiennej. Zbiór możliwych rozwiązań tych zadań opisano przy pomocy układu dwustronnych $(\max, \min)$−nierówności liniowych. Zaproponowano ostateczny algorytm służący do znalezienia optymalnego rozwiązania badanego problemu optymalizacji.

FINITE ELEMENT AND BOUNDARY ELEMENT SOLUTION OF NUCLEAR WASTE REPOSITORY THERMAL DIMENSIONING PROBLEM

Milan Hokr
*** Josef Novák**

Technical University of Liberec
Centre for Nanomaterials, Advanced Technologies and Innovation
Studentská 2, 46117, Liberec, Czech Republic
milan.hokr@tul.cz

* Technical University of Liberec
Centre for Nanomaterials, Advanced Technologies and Innovation
Studentská 2, 46117, Liberec, Czech Republic
josef.novak@tul.cz

Abstract

We solve the thermal dimensioning problem of the deep geological spent nuclear fuel repository, which means to estimate the maximum temperature in the repository caused by the heat generation of the spent fuel. We use a combination of the boundary element method for the exterior problem of heat discharge to the infinity and finite element method for a near-field thermal problem with a boundary condition expressed by the far-field problem solution. This combination is implemented within the simulation software ANSYS as the "far-field element". The far-field element solution confirmed to well represent the heat discharge, in comparison with the variant of a standard FEM far-field problem solution and conventional boundary conditions (constant temperature or zero heat flow).

Keywords: heat conduction, numerical simulation, far-field element, multiscale, ANSYS

Introduction

The solved problem comes from analysis of the project of geological disposal of a spent nuclear fuel – the spent fuel is put to steel canisters which are placed deep to a stable rock massif, further protected by the buffer layer of compacted bentonite [3]. One of the issues is heat dissipation of the fuel, not to reach a certain safe maximum temperature for the construction, while keeping the repository sufficiently small for economical reasons. For numerical simulation, the challenge is in multiscale character of the problem: we need to take into account details of canister and buffer shape in scale of tens of centimeters while there are several tens or hundreds of such boreholes/canisters and the extent of thermal influence of whole repository is in scale of hundreds of meters.

The methods used in the literature [2,4,6] are e.g. a superposition of solutions of the single borehole/canister or multiscale model with line sources instead of real canister geometry, either analytically or numerically. But for changing heat power and particular canister geometry, the numerical solution is necessary anyway. Except of the analytical solution, all approaches require solution of the heat conduction problem in much larger domain than the actual volume influenced, to allow defining a realistic but simple boundary condition (e.g. undisturbed temperature). For the solution of problems in infinite domain, the boundary element method is well suited, as solution of the exterior problem. To include both correct

solution of heat conduction to infinity and particular complicated geometry of the canister, buffer, and disposal borehole, the methods can be combined together – the finite element method for the local problem and the boundary element method for the infinite (exterior) problem. Several variants are described in literature as a general concept not limited to the particular application [5].

The FEM-BEM approach is also implemented in ANSYS commercial multiphysical simulation software, in the form of the "far-field element" [1,5], i.e. the standard finite element formulation is extended with a special element attached to the boundary of the local problem, expressing the interaction of the boundary with "infinity", equivalent to the real solution of the heat conduction in the infinite domain. Thus the problem can be solved in the much smaller computational domain and much less degrees of freedom, with all flexibility of finite elements on the local scale. In this paper we show how this approach can be efficiently used to the particular problem of thermal field of the spent nuclear fuel repository and we compare the "far-field element" solution with the solution with conventional boundary conditions.

1. Model description and data

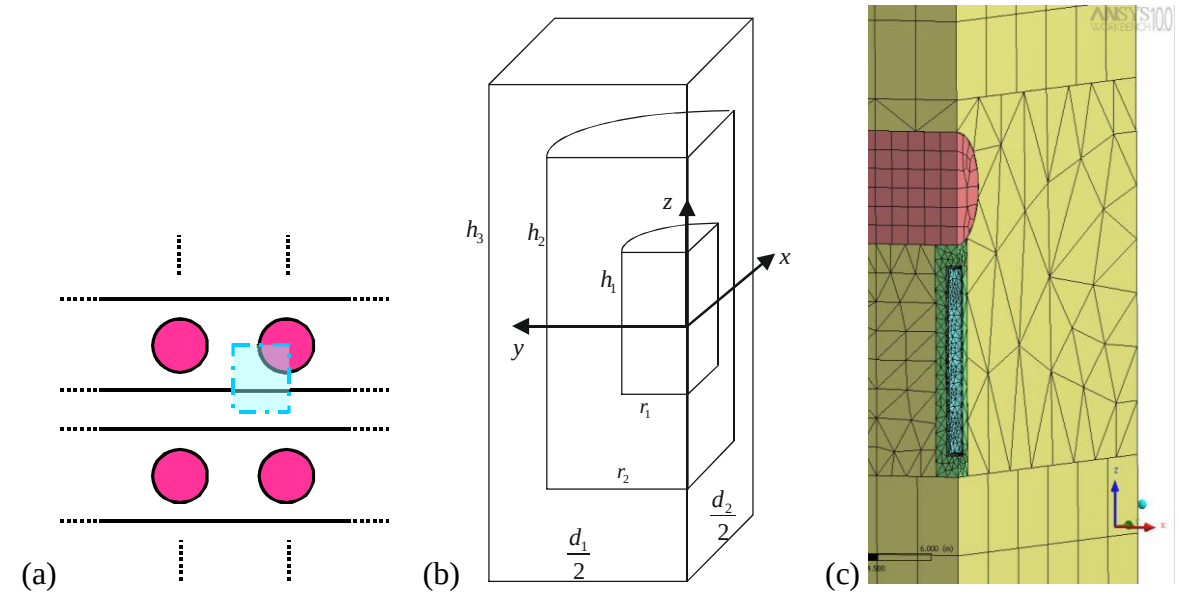
The structure of the planned repository [3,7] is a set of horizontal access tunnels with a certain spacing, with vertical disposal boreholes at the bottom of the tunnels, with a certain spacing. There is a cylindrical fuel canister in each borehole, with a layer of bentonite from top, bottom and vertical sides. We consider a periodic symmetry for the model formulation – in the plan view, the rectangle of half the tunnel spacing and half of the borehole spacing is the representative for the whole periodic structure (Fig. 1). In the vertical direction we define the domain extent according to needs of the heat influence and choice of a boundary condition. The dimensions are specified in Fig. 1b and Tab. 1.

We apply some simplifications concerning the geometry and materials. The access tunnels are filled with "backfill" material in the standard repository concept. The expected material is typically a mixture of clay and other rock, so the used thermal properties for bentonite (buffer) are not far from the possible values for the real backfill, moreover former tests have shown almost no effect of the backfill thermal properties on the buffer temperatures. We neglect the heat distribution inside the canister since the thermal dimensioning task is usually motivated by the buffer material stability, not by the phenomena inside the canister. The canister shell behaves as a perfect heat conductor distributing the heat according to the outer buffer material properties and geometry. The homogeneous volume of steel is a good approximation for the same behavior of the canister in our model. The real temperatures in the canister would be higher due to air gaps between the fuel rods and the steel skeleton/shell, but this is not a subject of this study.

Tab. 1. *List of parameters describing model geometry (reference to Fig. 1b)*

Parameter	Notations	Reference value
Borehole spacing (between axes)	d_1	9m
Tunnel sparing (between axes)	d_2	25m
Canister height (outer steel)	h_1	5.05m
Borehole height (outer buffer)	h_2	6.15m
Canister radius (steel/buffer interface)	r_1	0.35m
Borehole radius (buffer/rock interface)	r_2	0.66m
Vertical size of the model	h_3	60m
Tunnel diameter (circular)		3m

Source: Own compilation of data from [7,8]



Source: Own.

Fig. 1. *(a) plan view of tunnels and boreholes (red) with symmetrical segment (blue), (b) dimensions of the problem parts, (c) discretisation mesh of the model with distinguished materials (a cut-out detail about 1/5 of the model height).*

Tab. 2. *Material coefficients of the heat conduction problem, for different components of the model (see Fig 1.c).*

Material	Density	Heat capacity	Heat conductivity
	$kg \cdot m^{-3}$	$kJ \cdot kg^{-1} \cdot K^{-1}$	$W \cdot m^{-1} \cdot K^{-1}$
Buffer (partly saturated compacted bentonite)	2000	2500	1
Host rock (granite)	2700	850	2.7
Canister (steel)	7800	45	460
Tunel backfill = same as buffer (simplified)	2000	2500	1

Source: Own compilation of data from [8]

The input data of the model are material coefficients, boundary and initial conditions, and prescribed heat power changing in time. The heat conductivity, heat capacity and density are listed in Tab 2. The initial condition is the constant temperature 10°C. The heat power is calculated from the nuclear decay equations, for this study we got a table form of time/power dependence, possible to fit with a sum of three exponential functions [8]. The heat power at the beginning is 2137W, which means the power density 1858W/m^3 of the homogeneous volume source mentioned above.

The boundary conditions on symmetry planes (all vertical) are no heat flow. For the boundary on the horizontal planes on top and bottom of the model, we consider the following variants (to be compared between each other):

- The “Far-field element” instead of the strict-sense boundary condition, i.e. solution of the unbounded problem.

- Constant temperature 10°C (equal to initial): this can possibly overestimate heat dissipation and underestimate temperature (keeping the temperature constant in presence of heating requires additional cooling of boundary).
- No heat flow on the bottom and the "far-field element" on the top: at the bottom side, the heat flow is underestimated and therefore the temperature is overestimated.

To resume, we solve the problem of transient 3D linear heat conduction, with either standard finite element method with mixed block and tetrahedral elements and linear or bilinear base functions, or with combination of the finite element and the boundary element method, where the solution of the exterior problem is set as a special element instead of boundary condition on particular model side. We use ANSYS v11 academic license for all calculations.

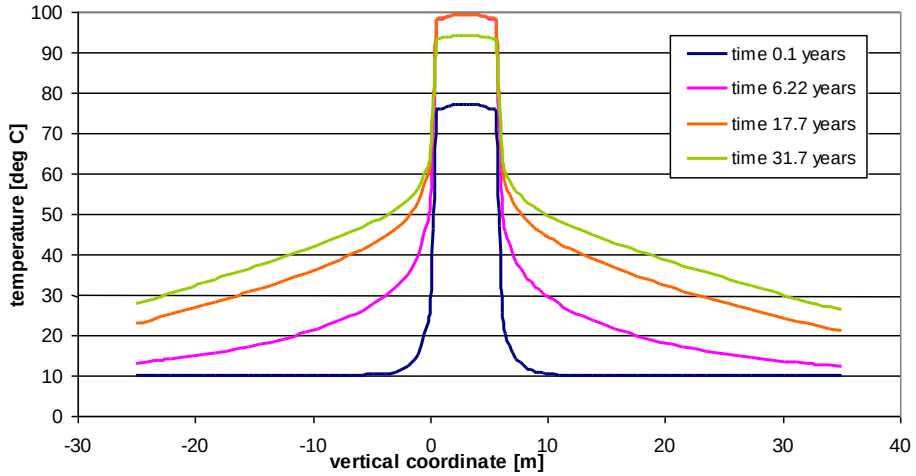
2. Results

We analyze the results in form of vertical temperature profiles, in order to precisely distinguish the behaviour near the boundary, which is the main subject of interest in relation to the far-field element demonstration. For the thermal dimensioning of the repository, the primary evaluated result is the maximum temperature. Surprisingly, the maximum temperature (in both space and time) is not much influenced by the choice of the boundary condition or far-field element. The reason can be that the heat power starts to decrease in similar time scale as the heat reaches the boundary (only after this time the variant of boundary can have effect on the maximum temperature). But the maximum temperature of the single profile in particular time after 10 or 20 years is then visibly influenced by the variant of the model boundary.

The profiles changing in time are presented in Fig. 2, for the far-field element variant. The quick rise of the temperature in canister and buffer and then the spread of heat towards boundaries and slower rise of the canister temperature is visible. The time development is similar to all variants of boundary. In Fig. 3 (sooner time) and Fig. 4 (later time) the profiles between model variants are compared: in the first case, the profiles are almost the same (the effect of the boundary did not yet happen) while in the second case there are differences as expected. The steepest decrease of temperature and the smallest canister temperature is for the constant temperature boundary (artificial cooling), the slowest decrease and the larger temperature is for zero heat flow (all heat kept inside), and the far-field element solution ("correct" heat conduction to the infinity) is between – the cooling is given by the infinite volume of rock around. We note that only left part of profile (bottom part of the model) is relevant, because the top of the "no heat flow" model also uses the far-field element. Also, the slope at the no-flow boundary is not ideally zero, because the model is slightly larger for this variant.

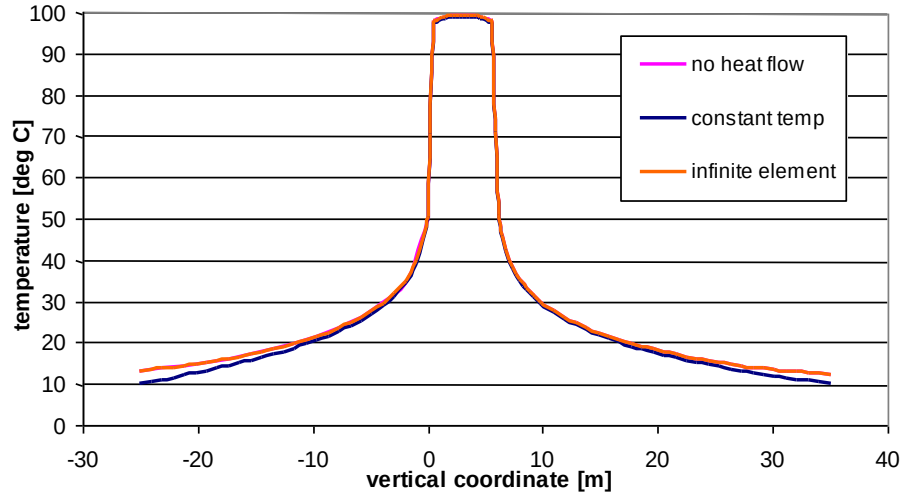
Conclusion

The solution with the far-field element (FEM-BEM method) representing the heat conduction in the infinite domain (without need of its actual discretisation) confirmed to behave as expected. In comparison the temperature is between the results of the constant-temperature and the no-heat-flow boundary. In contrast to e.g. empirical fitting of the third-type boundary, we get the more accurate solution for almost the same computing price. The presented model is in many aspects simplified against the needs for realistic prediction of the thermal condition in the repository, but this analysis gives good reference for choice of the model geometry and boundaries in the further more precise studies.



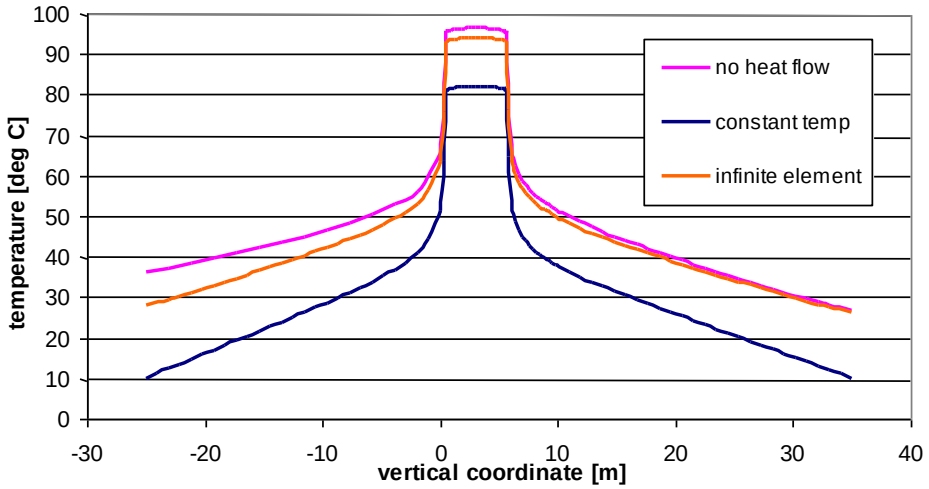
Source: Own

Fig. 1. *Development of vertical axial temperature profiles in time for the model with far-field element (infinite boundary).*



Source: Own

Fig. 2. *Comparison of temperature profiles among the model variants for a sooner time – 6.22 years (identical for "no heat flow" and "infinite element").*



Source: Own

Fig. 3. *Comparison of temperature profiles among the model variants for later time – 31.7 years.*

Acknowledgements

This project is realized under the state subsidy of the Czech Republic within the research and development project FR-TI/362 "Research of properties of material for safe disposal of radioactive waste and development of procedures for their assessment" supported by the Ministry of Industry and Trade.

The paper was supported in part by the Project OP VaVpI Centre for Nanomaterials, Advanced Technologies and Innovation CZ.1.05/2.1.00/01.0005.

Literature

[1] ANSYS, Inc.: *Release 12.0 Documentation for ANSYS*, ANSYS Incorporated, 2009.

[2] HÖKMARK H., FÄLTH B.: *Thermal dimensioning of the deep repository. Influence of canister spacing, canister power, rock thermal properties and nearfield design on the maximum canister surface temperature*, SKB report nr. TR-03-09, 2003.

[3] IAEA: *Scientific and Technical Basis for the Geological Disposal of Radioactive Wastes*, Int.Atom. En. Ag., Technical Reports Series No. 413, 2003.

[4] IKONEN, K.: *Thermal Analyses of Spent Nuclear Fuel Repository*, technical report, POSIVA 2003-04.

[5] KALJEVIC, I., SAIGAL, S., ALI, A.: *An Infinite Boundary Element Formulation for Three-Dimensional Potential Problems*, International Journal for Numerical Methods in Engineering, Vol. 35, No. 10, pp. 2079-2100 (1992) .

[6] KOHUT R, BLAHETA R, KOLCUN A, MALÍK J AND STARÝ J, *Computations of thermo-elastic responses of rocks in the vicinity of geological deposition of the spent nuclear fuel*, In: Proceedings of SIMONA 2006, Technical University of Liberec, 2006, pp. 70-77.

[7] SÚRAO: *Zadávací bezpečnostní zpráva, Příloha č. 1, Referenční projekt hlubin-ného úložiště v hostitelském prostředí granitových hornin*, SURAO-EGP Invest, Uh. Brod 1999.

[8] ÚJV ŘEŽ a.s.: *Shromáždění a aktualizace vstupních údajů o vývoji tepla vyhořelých palivových souborů*, Výstup 1.2, Ev. č. zhotovitele: 5 SMP 156, Nuclear research institute Řež, March 2006.

ŘEŠENÍ ÚLOHY TEPELNÉHO DIMENZOVÁNÍ ÚLOŽIŠTĚ VYHOŘELÉHO JADERNÉHO PALIVA POMOCÍ KONEČNÝCH PRVKŮ A HRANIČNÍCH PRVKŮ

Řešíme úlohu tepelného dimenzování hlubinného geologického ukládání vyhořelého jaderného paliva, což spočívá v odhadu maximální teploty v úložišti vlivem tepla generovaného uloženým palivem. Používáme kombinaci metody hraničních prvků pro vnější úlohu odvodu tepla v nekonečném prostoru a metody konečných prvků (MKP) pro tepelnou úlohu v blízkém poli s okrajovou podmínkou vyjádřenou řešením vnější úlohy. Tato kombinace je implementována v simulačním softwaru ANSYS jako "far-field element". Řešení touto metodou potvrdilo, že dobře reprezentuje odvod tepla, v porovnání s variantami modelu se standardní MKP pro úlohu vzdáleného pole nebo s volbou základních typů okrajových podmínek (konstantní teplota nebo nulový tepelný tok).

LÖSUNG DER AUFGABE DER WÄRMEDIMENSIONIERUNG DER LAGESTÄTTE FÜR ATOMAREN ABFALL MIT HILFE FINITER ELEMENTE UND GRENZELEMENTE

Hier wird die Aufgabe der Wärmedimensionierung der Tieflagerstätte von ausgebranntem Atombrennstoff gelöst. Dies ist die Schätzung der Maximaltemperatur in der Lagerstätte, die durch den auf Grund der Wärme verbrannten Brennstoff verursacht wurde. Wir benutzen eine Kombination aus der Methode der Grenzelemente für die äußere Wärmeableitung in einen unendlichen Raum (entfernte Felder) und der Methode der finiten Elemente für die Leitung der Wärme auf dem nahen Feld mit der Randbedingung, welche durch die Lösung im entfernten Feld ausgedrückt wird. Diese Kombination wird in der Simulationssoftware ANSYS als „far-field element“ implementiert. Die Lösung unter Verwendung dieses Elements bestätigte die Fähigkeit, die Wärme korrekt abzuführen, und das im Vergleich mit der Lösung des entfernten Feldes durch finite Elemente und Standardrandbedingungen (konstante Temperatur oder kein Wärmefluss).

ROZWIĄZANIE ZADANIA ZAGOSPODAROWANIA ENERGII CIEPLNEJ ZE SKŁADOWISKA ODPADÓW JĄDROWYCH PRZY POMOCY ELEMENTÓW SKOŃCZONYCH I ELEMENTÓW BRZEGOWYCH

W artykule przedstawiono rozwiązanie zadania zagospodarowania energii cieplnej pochodzącej z głębinowego składowiska wypalonego paliwa jądrowego, opartą na oszacowaniu maksymalnej temperatury w składowisku spowodowanej energią cieplną generowaną przez wypalone paliwo. Zastosowano połączenie metody elementów brzegowych dla zewnętrzego zadania odprowadzenia energii cieplnej w nieskończoną przestrzeń (oddalone pole) oraz metody elementów skończonych dla odprowadzenia energii cieplnej na pobliskie pole z warunkiem krańcowym wyrażonym rozwiązaniem na oddalonym polu. Połączenie to wprowadzono do oprogramowania symulacyjnego ANSYS jako "far-field element" (element oddalonego pola). Rozwiązanie z wykorzystaniem tego elementu umożliwiło prawidłową prezentację odprowadzenia energii cieplnej, w porównaniu z rozwiązaniem w postaci oddalonego pola przy wykorzystaniu elementów brzegowych oraz standardowych warunków krańcowych (stała temperatura lub zerowy przepływ energii cieplnej).

Numerical Solution of the MEW Equation by the Semi-Implicit Numerical Scheme

Jiří Hozman

Technical University of Liberec
Faculty of Science, Humanities and Education
Department of Mathematics and Didactics of Mathematics
Studentská 2, 461 17, Liberec, Czech Republic
jiri.hozman@tul.cz

Abstract

In this paper we deal with the development of a numerical method for the solution of the modified equal width wave (MEW) equation – a very important equation with a cubic nonlinearity describing a large number of physical phenomena. The crucial idea of introduced approach is based on the discretization of the MEW equation with the aid of a combination of the discontinuous Galerkin (DG) method for the space semi-discretization and the backward Euler method for the time discretization. The appended numerical experiments investigate the conservative properties of the MEW equation related to mass, momentum and energy, and illustrate the potency of this scheme, consequently.

Keywords: Discontinuous Galerkin method; modified equal width wave equation; semi-implicit scheme; solitary wave.

Introduction

Our aim is to present a sufficiently robust, accurate and efficient numerical method for the solution of scalar nonlinear partial differential equations. As a model problem, we consider a modified equal width wave (MEW) equation, describing the various phenomena in physical disciplines. The MEW equation contains a cubic nonlinearity and exhibits a pulse-like solitary wave having the same width with both positive and negative amplitudes. Several numerical methods have been introduced in the literature for the solution of the MEW equation, see [1], [6], [9] and references cited therein.

In this paper, we proposed a semi-implicit scheme for the numerical solution of the MEW equation. The discontinuous Galerkin (DG) methods have become a very popular numerical technique for the solution of nonlinear problems. The DG space semi-discretization uses higher order piecewise polynomial discontinuous approximation on arbitrary meshes, for a survey, see [2], [3], [4]. Among several variants of DG methods we prefer the so-called interior penalty Galerkin (IPG) discretizations. The discretization in the time coordinate is performed with the aid of a linearization and the backward Euler method, sidetracking the time step restriction well-known from the explicit schemes. Consequently, the fully discrete problem is represented by the system of linear algebraic equations.

The rest of the paper is organized as follows. The problem formulation and its variational reformulation are given in Section 1. The discretization including a space semi-discretization and fully time space discretization is considered in Section 2. Some numerical results are provided in Section 3.

1 Problem Formulation

Let $\Omega = (a,b) \subset \mathbb{R}$ be a bounded open interval, we consider the following MEW equation:

Problem (I): Find $u(x,t) : Q_T = \Omega \times (0,T) \to \mathbb{R}$ such that, for all $T > 0$,

$$\begin{aligned} &\text{(a)} \quad \frac{\partial u}{\partial t} + \varepsilon u^2 \frac{\partial u}{\partial x} - \mu \frac{\partial}{\partial t}\left(\frac{\partial^2 u}{\partial x^2}\right) = 0 \quad \text{in } Q_T, \\ &\text{(b)} \quad u(a,t) = u_D^a(t) \text{ and } u(b,t) = u_D^b(t), \quad t \in (0,T) \\ &\text{(c)} \quad u(x,0) = u^0(x), \quad x \in \Omega, \end{aligned} \tag{1}$$

where positive parameters ε and μ represent the amplitude of the wave and long-wavelength, respectively. The initial boundary value problem (1) is equipped with the initial condition $u^0 : \Omega \to \mathbb{R}$ and the Dirichlet boundary conditions $u_D^a, u_D^b : (0,T) \to \mathbb{R}$ prescribed at both endpoints of the domain Ω.

In order to obtain a variational formulation of (1) we introduce the standard notation for function spaces. Let $k \geq 0$ be an integer and $p \in [1,\infty]$. We use the well-known Lebesgue and Sobolev spaces $L^p(\Omega)$, $H^k(\Omega)$, Bochner spaces $L^p(0,T;X)$ of functions defined in (0, T) with values in the Banach space X and the spaces $C^k([0,T];X)$ of k-times continuously differentiable mappings of the interval [0, T] with values in X. Further, we denote the inner product in $L^2(\Omega)$ by $(\cdot,\cdot)$, and let $H_0^1(\Omega) = \{v \in H^1(\Omega) : v(a) = v(b) = 0\}$.

Now, we are ready to introduce the following *weak formulation* of problem (1):

Problem (II): Find $u \in C^1(0,T;H^1(\Omega))$ such that, for all $t \in [0,T]$,

$$\begin{aligned} &\text{(a)} \quad u - u^* \in C^1(0,T;H_0^1(\Omega)) \\ &\text{(b)} \quad \frac{\mathrm{d}}{\mathrm{d}t}(u(t),v) + \varepsilon b(u(t),v) + \mu \frac{\mathrm{d}}{\mathrm{d}t} a(u(t),v) = 0 \quad \forall v \in H_0^1(\Omega) \\ &\text{(c)} \quad u(0) = u^0 \quad \text{in } \Omega, \quad u^0 \in L^2(\Omega), \end{aligned} \tag{2}$$

where symbol $u(t)$ stands for the function on Ω such that $u(t)(x), x \in \Omega$ and the bilinear form $a(\cdot,\cdot)$ and the nonlinear form $b(\cdot,\cdot)$ are defined as

$$a(u(t),v) = \int_\Omega \frac{\partial u(t)}{\partial x} v' \,\mathrm{d}x, \qquad \text{(diffusion term)} \tag{3}$$

$$b(u(t),v) = \int_\Omega \frac{\partial f(u(t))}{\partial x} v \,\mathrm{d}x \quad \text{with } f(u) = \frac{1}{3}u^3 \quad \text{(convection term).} \tag{4}$$

Function $f(u)$ in (4) represents the physical flux.

2 Discretization

Let $\mathscr{T}_h$ ($h > 0$) be a family of the partitions of the closure $\overline{\Omega} = [a,b]$ of the domain Ω into N closed mutually disjoint subintervals $I_k = [x_{k-1}, x_k]$ with length $h_k := x_k - x_{k-1}$ and the symbol $\mathscr{J}$ stands for an index set $\{1,\ldots,N\}$. Then we call $\mathscr{T}_h = \{I_k, k \in \mathscr{J}\}$ a *triangulation* with the spatial step $h := \max_{k \in \mathscr{J}} h_k$ and the interval I_k an *element*. Let $\mathscr{E}_h = \{x_0 = a, x_1, \ldots, x_{N-1}, x_N = b\}$. Further, we label by $\mathscr{E}_h^I$ the set of all inner nodes. Obviously, $\mathscr{E}_h = \mathscr{E}_h^I \cup \{a,b\}$.

The DG method can handle different polynomial degrees over elements. Therefore, we assign a positive integer p_k as a *local polynomial degree* to each $I_k \in \mathscr{T}_h$. Then we set the vector $\mathsf{p} = \{p_k, I_k \in \mathscr{T}_h\}$. Over the triangulation $\mathscr{T}_h$ we define the finite dimensional space of discontinuous piecewise polynomial functions

$$S_{h\mathsf{p}} \equiv S_{h\mathsf{p}}(\Omega, \mathscr{T}_h) = \{v; v|_{I_k} \in P_{p_k}(I_k)\ \forall k \in \mathscr{J}\}, \tag{5}$$

where $P_{p_k}(I_k)$ denotes the space of all polynomials of degree $\leq p_k$ on I_k, $I_k \in \mathscr{T}_h$. Consequently, the approximate solution of the continuous problem (1) is sought in the space S_{hp}.

For each $x \in \mathscr{E}_h^I$ there exist two elements I_k, $I_{k+1} \in \mathscr{T}_h$ such that $I_k \cap I_{k+1} = \{x\}$. Let us denote

$$v(x^+) = \lim_{\varepsilon \to 0+} v(x+\varepsilon) \quad \text{and} \quad v(x^-) = \lim_{\varepsilon \to 0+} v(x-\varepsilon) \tag{6}$$

the *traces* of v at inner points of Ω. Moreover,

$$[v(x)] = v(x^-) - v(x^+), \qquad \langle v(x) \rangle = \frac{1}{2}\left(v(x^-) + v(x^+)\right), \tag{7}$$

denote the *jump* and *mean value* of the function v at points $x \in \mathscr{E}_h^I$, respectively. By convention, we also extend the definition of the jump and mean value for endpoints of domain Ω, i.e.

$$[v(x_0)] = -v(x_0^+), \quad \langle v(x_0) \rangle = v(x_0^+), \quad [v(x_N)] = v(x_N^-), \quad \langle v(x_N) \rangle = v(x_N^-) \tag{8}$$

In case that $x \in \mathscr{E}_h$ are arguments of $v(x^-)$ or $v(x^+)$, we usually omit these arguments x^-, x^+ and write simply v^- and v^+, respectively.

2.1 Space Semi-Discrete DG Scheme

Now, we recall the space semi-discrete DG scheme presented in [8]. The crucial item of the DG formulation of model problem is the treatment of the convection part. The convection terms are approximated with the aid of the following numerical flux $H(\cdot,\cdot)$ through node $x \in \mathscr{E}_h$ in the positive direction (i.e. outer normal is equal to one):

$$H\Big(u(x^-), u(x^+)\Big) = \begin{cases} f\left(u(x^-)\right), & \text{if } A \geq 0 \\ f\left(u(x^+)\right), & \text{if } A < 0 \end{cases}, \quad \text{where } A = f'\left(\frac{u(x^-)+u(x^+)}{2}\right), \tag{9}$$

which is based on the concept of *upwinding*, for more details see [7].

Due to the cubic form of physical flux $f(u) = \frac{1}{3}u^3$, the derivative $f'(u) = u^2$ is nonnegative everywhere in Ω and the numerical flux H defined in (9) has simpler form:

$$H\Big(u(x^-), u(x^+)\Big) = f\left(u(x^-)\right), \quad x \in \mathscr{E}_h \tag{10}$$

The choice of $u(x^-), u(x^+)$ for boundary points $\{a,b\}$ is necessary to specify. Here we use:

$$u(x_0^-) = u(a^-) = u_D^a \quad \text{and} \quad u(x_N^+) = u(b^+) = u_D^b. \tag{11}$$

A particular attention should be also paid to the treatment of the diffusion terms, which include artificially added *stabilization* in order to guarantee the stability of the resulting numerical scheme. Furthermore, in order to replace the inter-element discontinuities, the semi-discrete scheme is completed with *penalty* vanishing for the continuous solution.

Therefore, we can define the *semi-discrete solution* u_h of the problem (1).

Problem (III): Find $u_h \in C^1(0,T;S_{hp})$ such that, for all $t \in [0,T]$,

$$\text{(a)} \qquad \frac{\mathrm{d}}{\mathrm{d}t}\Big\{(u_h(t),v_h) + \mu a_h^{\Theta}(u_h(t),v_h) + \mu J_h^{\sigma}(u_h(t),v_h)\Big\} + \varepsilon\, b_h(u_h(t),v_h) = 0 \quad \forall v_h \in S_{hp}, \tag{12}$$

$$\text{(b)} \qquad (u_h(0),v_h) = (u^0,v_h)\ \forall v_h \in S_{hp}$$

where

$$a_h^{\Theta}(u(t),v) = \sum_{k\in\mathscr{J}} \int_{I_k} \frac{\partial u(t)}{\partial x}\cdot v'\,\mathrm{d}x - \sum_{x\in\mathscr{E}_h} \left\langle \frac{\partial u(t)}{\partial x} \right\rangle [v] + \Theta \sum_{x\in\mathscr{E}_h^I} \langle v' \rangle\, [u(t)], \tag{13}$$
$$+\Theta v'(x_0^+)\cdot\Big(u_D^a(t) - u(x_0^+,t)\Big) + \Theta v'(x_N^-)\cdot\Big(u(x_N^-,t) - u_D^b(t)\Big)$$

(diffusion form)

$$b_h(u(t),v) = -\sum_{k\in\mathscr{J}} \int_{I_k} f(u(t))\cdot v'\,\mathrm{d}x + \sum_{x\in\mathscr{E}_h^I} f(u^-(t))\,[v] \tag{14}$$
$$-f(u_D^a(t))\cdot v(x_0^+) + f(u(x_N^-,t))\cdot v(x_N^-)$$

(convection form)

$$J_h^{\sigma}(u(t),v) = \sum_{x\in\mathscr{E}_h^I} \sigma[u(t)]\,[v] + \sigma(x_0)\cdot\Big(u_D^a(t) - u(x_0^+,t)\Big)\cdot v(x_0^+) \tag{15}$$
$$+\sigma(x_N)\cdot\Big(u(x_N^-,t) - u_D^b(t)\Big)\cdot v(x_N^-)$$

(penalty form)

According to value of the parameter Θ, we speak of *symmetric* ($\Theta = -1$), *incomplete* ($\Theta = 0$) or *nonsymmetric* ($\Theta = 1$) variants of stabilization of the DG method, i.e., we generally consider three variants of the diffusion form a_h^{Θ}. The penalty parameter function $\sigma : \mathscr{E}_h \to \mathbb{R}$ in (15) is defined in spirit of [5] as

$$\sigma(x) = \frac{C_W}{d(x)} \quad \text{with} \quad d(x) = \begin{cases} h_1/{p_1}^2 & , x = a, \\ \min\left(h_k/p_k^2, h_{k+1}/p_{k+1}^2\right) & , x \in \mathscr{E}_h^I \wedge \{x\} = I_k \cap I_{k+1}, \\ h_N/{p_N}^2 & , x = b, \end{cases} \tag{16}$$

where $C_W > 0$ is a suitable constant depending on the used variant of scheme and on the degree of polynomial approximation.

In order to simplify the notation we introduce the form

$$\mathscr{A}_h^{\Theta,\mu}(u(t),v) := (u(t),v) + \mu\, a_h^{\Theta}(u(t),v) + \mu J_h^{\sigma}(u(t),v), \quad u(t), v \in S_{hp},\ t \in (0,T), \tag{17}$$

which is bilinear due to (13) and (15). Consequently, the equation (12a) can be rewritten as

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathscr{A}_h^{\Theta,\mu}(u_h(t),v_h) + \varepsilon\, b_h(u_h(t),v_h) = 0 \qquad \forall v_h \in S_{hp}, \forall t \in [0,T], \tag{18}$$

The problem (18) represents a system of ordinary differential equations (ODEs) for $u_h(t)$ which has to be discretized in time by a suitable method.

2.2 Fully Time-Space Discrete DG Scheme

There exists a wide range of approaches for the time discretization of ODE systems resulting from the DG semi-discretization. In practical computations, the simplest time discretization is via *explicit scheme* (e.g. Euler forward scheme and Runge-Kutta methods). These schemes suffer from a strong limitation on the time step due to a *CFL-stability condition*. However, their main advantage is easy implementation. On the other hand, in order to avoid the strong time step restriction of explicit DG schemes, it is suitable to use an *implicit* time discretization.

Let $0 = t_0 < t_1 < \cdots < t_M = T$ be a partition of the interval $[0,T]$, time steps $\tau_l \equiv t_{l+1} - t_l$, and u_h^l stands for the *approximate solution* of $u_h(t_l)$, $t_l \in [0,T]$, $l = 0,\ldots,M$. The fully discrete

solution of problem (12) via an implicit approach with the *backward Euler method* is defined in following way.

Problem (IV): Find $u_h^l \in S_{h\mathsf{p}}$, such that, for $l = 1,\dots,M$,

$$\text{(a)} \quad \frac{1}{\tau_l}\left(\mathscr{A}_h^{\Theta,\mu}(u_h^{l+1},v_h) - \mathscr{A}_h^{\Theta,\mu}(u_h^l,v_h)\right) + \varepsilon b_h\left(u_h^{l+1},v_h\right) = 0 \quad \forall v_h \in S_{h\mathsf{p}}, \tag{19}$$

$$\text{(b)} \quad u_h^0 \text{ is } S_{h\mathsf{p}} \text{ approximation of } u^0.$$

The main drawback inhibiting from the fully implicit treatment in (19a) is the nonlinearity of the convection form $b_h(\cdot,\cdot)$ which leads to a nonlinear system of algebraic equations at each time step the solution of which is rather expensive and more complicated.

Therefore, in our case, the proposed *semi-implicit* approach, sidetracking the time step restriction as well as the nonlinearity of the convection form is generally based on a suitable *linearization* of this form. The nonlinearity appearing in the convection form (14) comes from the physical flux $f(u)$ the linearization of which is crucial for the later linear treatment of the form $b_h(\cdot,\cdot)$.

We use a linearization of $f(u)$ with the aid of the Taylor expansion as

$$f(u(t+\tau)) = \frac{1}{3}u^3(t+\tau) = \frac{1}{3}u^3(t) + \frac{\partial}{\partial t}\frac{1}{3}u^3(t)\tau + O(\tau^2), \quad \tau > 0,\ t \in (0,T). \tag{20}$$

Hence, differentiation and omitting of higher order terms $O(\tau^2)$ in (20) lead to approximation

$$f(u(t+\tau)) \approx \frac{1}{3}u^3(t) + u^2(t)\frac{\partial u(t)}{\partial t}\tau \approx \frac{1}{3}u^3(t) + u^2(t)\left(u(t+\tau) - u(t)\right), \tag{21}$$

the last approximation in (21) comes from the Taylor expansion of u, i.e.

$$u(t+\tau) = u(t) + \frac{\partial u(t)}{\partial t}\tau + O(\tau^2) \approx u(t) + \frac{\partial u(t)}{\partial t}\tau, \quad \tau > 0,\ t \in (0,T). \tag{22}$$

Finally, from (21) we get

$$f(u(t+\tau)) \approx u^2(t)u(t+\tau) - \frac{2}{3}u^3(t) = u^2(t)u(t+\tau) - 2f(u(t)), \quad \tau > 0,\ t \in (0,T) \tag{23}$$

and by substituting u_h^{l+1} for $u(t+\tau)$ and u_h^l for $u(t)$ in (23), respectively

$$f(u_h^{l+1}) \approx \left(u_h^l\right)^2 u_h^{l+1} - 2f(u_h^l), \quad l = 1,\dots,M. \tag{24}$$

Next, we can proceed to the linearization of an implicit treatment of b_h as

$$b_h(u_h^{l+1},v_h) = -\sum_{k\in\mathscr{J}}\int_{I_k} f(u_h^{l+1})\cdot v_h'\,\mathrm{d}x + \sum_{x\in\mathscr{E}_h^I\cup\{x_N\}} f((u_h^{l+1})^-)\,[v_h] - f(u_D^a(t_{l+1}))\cdot v_h(x_0^+) \tag{25}$$

From (24) and (25) we find that

$$\begin{aligned} b_h(u_h^{l+1},v_h) \;=\; & \underbrace{-\sum_{k\in\mathscr{J}}\int_{I_k}(u_h^l)^2\cdot u_h^{l+1}\cdot v_h'\,\mathrm{d}x + \sum_{x\in\mathscr{E}_h^I\cup\{x_N\}}((u_h^l)^-)^2\cdot(u_h^{l+1})^-\cdot[v_h] - f(u_D^a(t_{l+1}))\cdot v_h(x_0^+)}_{=:b_{hL}(u_h^l,u_h^{l+1},v_h)} \\ & \underbrace{-2\left\{-\sum_{k\in\mathscr{J}}\int_{I_k} f(u_h^l)\cdot v_h'\,\mathrm{d}x + \sum_{x\in\mathscr{E}_h^I\cup\{x_N\}} f((u_h^l)^-)\,[v_h]\right\}}_{=:b_h^*(u_h^l,v_h)} \end{aligned} \tag{26}$$

where the form $b_{hL}(\cdot,\cdot,\cdot)$ is linear with respect to its second and third components and the form $b_h^*(\cdot,\cdot)$ is in fact the original convection form (14) with homogeneous Dirichlet boundary conditions.

Since linear parts of b_{hL} are treated implicitly and the nonlinear ones together with the form b_h^* explicitly, the semi-implict treatment preserves a linear algebraic problem at each time step. In this way we arrive at the following semi-implicit method.

Problem (V): Find $u_h^l \in S_{h\mathrm{p}}$, such that, for $l = 1,\ldots,M$,

$$\text{(a)} \quad \mathscr{A}_h^{\Theta,\mu}(u_h^{l+1},v_h) + \tau_l \varepsilon b_{hL}\left(u_h^l, u_h^{l+1}, v_h\right) = \mathscr{A}_h^{\Theta,\mu}(u_h^l,v_h) + 2\tau_l \varepsilon b_h^*\left(u_h^l, v_h\right) \quad \forall v_h \in S_{h\mathrm{p}},$$

$$\text{(b)} \quad u_h^0 \text{ is } S_{h\mathrm{p}} \text{ approximation of } u^0. \tag{27}$$

The discrete problem (27) is equivalent to a system of linear algebraic equations at each time instant $t_l \in [0,T]$. The resulting method has a high order of accuracy with respect to the space coordinates and the first order of accuracy with respect to time.

3 Numerical Experiments

In this section we present the results obtained from the semi-implicit method (27) proposed for the numerical solution of the problem (1) for a propagation of a single solitary wave. The MEW equation has three conservation quantities corresponding to mass, momentum, and energy

$$I_M(u) = \int_\Omega u\,\mathrm{d}x, \qquad I_{MM}(u) = \int_\Omega (u^2 + \mu(u')^2)\,\mathrm{d}x, \qquad I_E(u) = \int_\Omega u^4\,\mathrm{d}x, \tag{28}$$

respectively, and will be monitored to check the conservation properties of the proposed algorithm.

Let us consider the following analytical solution of (1)

$$u(x,t) = A\,\mathrm{sech}\,(k(x - x_0 - pt)) \tag{29}$$

which represents a single solitary wave of amplitude $A = \sqrt{6p/\varepsilon}$, where p is the velocity of the wave and $k = \sqrt{1/\mu}$. The boundary and initial conditions are extracted from the exact solution (29).

In order to compare our semi-implicit approach to the schemes given in [1] and [6] we set the parameters values $A = 0.25$, $x_0 = 30.0$, $\varepsilon = 3.0$ and $\mu = 1.0$. The run of the algorithm is carried up to time $T = 20.0$ over the problem domain $[0,80]$ with the constant mesh size $h = 0.1$ and the time step $\tau = 0.05$. The computations were performed by piecewise cubic approximations with $\Theta = 0$ (incomplete variant).

Figure 1 captures the development of approximation solutions of single solitary wave from an initial condition to different time instants. Table 1 records the invariant quantities and compares obtained results with several previous schemes given in [1] and [6]. We obtained satisfactory results and quite good agreement was already achieved for a piecewise cubic approximation with reference results from [1].

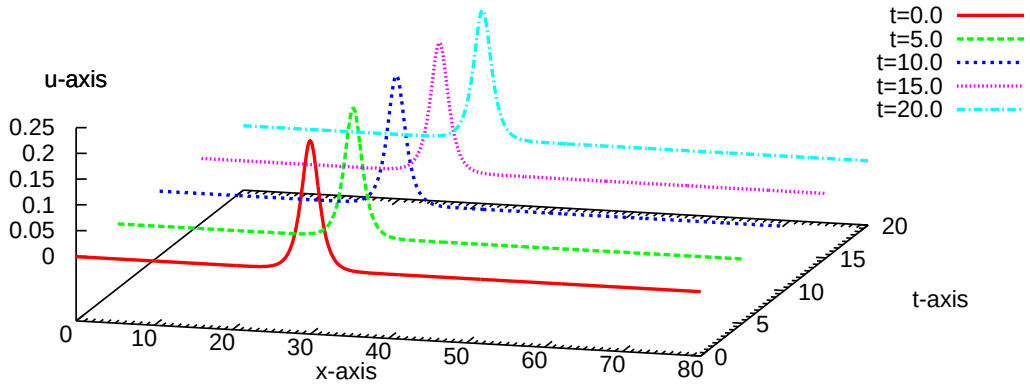
Conclusion

We dealt with the numerical solution of the MEW equation via semi-implicit scheme based on the discontinuous Galerkin method and the backward Euler method for the space and time discretization, respectively. A preliminary numerical example produced satisfactory results and illustrated the potency of the resulting scheme.

Tab. 1. *Computed invariant quantities for single solitary wave*

method	l time	$I_M(u_h^l)$	$I_{MM}(u_h^l)$	$I_E(u_h^l)$
present method	0.0	0.785398	0.166666	0.0052083
	5.0	0.785398	0.166666	0.0052083
	10.0	0.785398	0.166667	0.0052084
	15.0	0.785398	0.166668	0.0052084
	20.0	0.785398	0.166668	0.0052085
ref. method [6]	20.0	0.785398	0.166474	0.0052083
ref. method [1]	20.0	0.785398	0.166667	0.0052083

Source: Own based on [1] and [6]



Source: Own based on [1]

Fig. 1. *Development of approximate solutions of single solitary wave*

Acknowledgments

This work was supported by the ESF Project No. CZ.1.07/2.3.00/09.0155 “Constitution and improvement of a team for the demanding technical computations on parallel computers at TU of Liberec”.

Literature

[1] ALI, A.; HAQ, S.; ISLAM, S.: A Numerical Meshfree Technique for the Solution of the MEW Equation. *Computer Modeling in Engineering and Sciences*, vol. 38(1), pp. 1–23, 2008.

[2] ARNOLD, D. N.; BREZZI, F.; COCKBURN, B.; MARINI, L. D.: Unified analysis of discontinuous Galerkin methods for elliptic problems. *SIAM J. Numer. Anal.*, vol. 39(5), pp. 1749–1779, 2002.

[3] COCKBURN, B.: Discontinuous Galerkin methods for convection dominated problems. In T. J. Barth and H. Deconinck, editors. *High–Order Methods for Computational Physics*,

Lecture Notes in Computational Science and Engineering 9, pp. 69–224. Springer, Berlin, 1999.

[4] COCKBURN, B., KARNIADAKIS, G. E.; SHU, C.–W., editors: *Discontinuous Galerkin Methods*. Springer, Berlin, 2000.

[5] DOLEJŠÍ, V.; HOZMAN, J.: A priori error estimates for DGFEM applied to non-stationary nonlinear convectiondiffusion equation. In G. Kreiss et. Al. Eds., *Numerical Mathematics and Advanced Applications, ENUMATH 2009*, pp. 459–468, Springer, 2010.

[6] ESEN, A.; KUTLUAY, S.: Solitary wave solutions of the modified equal width wave equation. *Comm. Nonlinear Science and Numer. Simul.*, vol. 13, pp. 1538–1546, 2008.

[7] FEISTAUER, M.; FELCMAN, J.; STRAŠKRABA, I.: *Mathematical and Computational Methods for Compressible Flow*. Oxford University Press, Oxford, 2003.

[8] RIVIÈRE, B.: *Discontinuous Galerkin Methods for Solving Elliptic and Parabolic Equations: Theory and Implementation*. Frontiers in Applied Mathematics. Society for Industrial and Applied Mathematics, Philadelphia, 2008.

[9] ZAKI, S. I.: Solitary wave interactions for the modified equal width equation. *Comput. Phys. Comm.*, vol. 126, pp. 219–231, 2000.

RNDr. Jiří Hozman, Ph.D.

Numerické řešení MEW rovnice prostřednictvím semi-implicitního numerického schématu

V tomto článku se zabýváme vývojem numerické metody pro řešení MEW rovnice – velmi významné rovnice s kubickou nelinearitou popisující rozsáhlé množství fyzikálních jevů. Klíčová myšlenka uvedeného přístupu je založena na diskretizaci MEW rovnice pomocí kombinace nespojité Galerkinovy metody pro prostorovou semi-diskretizaci a zpětné Eulerovy metody pro časovou diskretizaci. Připojené numerické experimenty zkoumají vlastnosti zachování pro MEW rovnici spojené s hmotností, hybností a energií a následně tak dokládají potenci tohoto schématu.

Numerische Lösung der MEW-Gleichung mittels eines halbimpliziten numerischen Schemas

In diesem Artikel beschäftigen wir uns mit der Entwicklung einer numerischen Methode zur Lösung einer MEW-Gleichung. Das ist eine sehr bedeutsame Gleichung mit einer kubischen Nichtlinearität, welche eine umfangreiche Menge physikalischer Erscheinungen beschreibt. Der Schlüsselgedanke des angeführten Ansatzes gründet sich auf der Diskretisierung einer MEW-Gleichung mit Hilfe einer Kombination aus der nichtkontinuierlichen Galerkin-Methode für die räumliche Semi-Diskretisierung und impliziten Euler-Verfahren für die zeitliche Diskretisierung. Die angeschlossenen numerischen Experimente untersuchen die Eigenschaften der Erhaltung für die MEW-Gleichung, die mit der Masse, der Beweglichkeit und der Energie verbunden ist, und belegen so die Potenz dieses Schemas.

Numeryczne rozwiązanie równania MEW za pośrednictwem semi-dyskretnego schematu numerycznego

W niniejszym artykule uwagę poświęcono rozwojowi numerycznej metody do rozwiązywania równania MEW - bardzo ważnego równania nieliniowego trzeciego rzędu opisującego wiele zjawisk fizykalnych. Kluczowa idea wymienionego podejścia oparta jest na dyskretyzacji równania MEW przy pomocy połączenia nieciągłej metody Galerkina dla przestrzennej semi-dyskretyzacji oraz zwrotne metody Eulera dla dyskretyzacji czasowej. Przedstawione eksperymenty numeryczne mają na celu zbadanie cech zachowania równania MEW związanych z masą, ruchem i energią. Ponadto przedstawiaja potencjałtakiego schematu.

APPLICATION OF RECONSTRUCTION OPERATORS IN THE DISCONTINUOUS GALERKIN METHOD

Václav Kučera

Charles University in Prague
Faculty of Mathematics and Physics
Department of Numerical Mathematics
Sokolovská 83, Praha 8, 186 75, Czech Republic
vaclav.kucera@email.cz

Abstract

This paper gives an overview of the main ingredients needed to incorporate reconstruction operators, as known from higher order finite volume (FV) and spectral volume (SV) schemes, into the discontinuous Galerkin (DG) method. Such an operator constructs higher order approximations from the lower order DG scheme, increasing the order of convergence, while leading to a more efficient numerical scheme than the corresponding higher order DG scheme itself. We discuss theoretical, as well as implementational aspects and numerical experiments.

Keywords: Higher order reconstruction; discontinuous Galerkin; finite volumes.

Introduction

During the last decades, the finite volume (FV) method has gained the position of the standard method used by the engineering community for the discretization of equations governing conservation laws and convection-dominated problems. This method has many advantages, however one of its drawbacks is its low order accuracy. The standard approach to overcome this problem is the introduction of higher order reconstruction operators into the FV scheme. Although such an approach has not yet been theoretically justified, it gives excellent results in practice. However, one usually uses at most piecewise quadratic reconstructions, since higher orders are impractical and cumbersome from the implementational point of view.

A somewhat different approach to higher order schemes is the discontinuous Galerkin (DG) method, which combines concepts form the finite element and finite volume methods. Again, this method has many advantages, but one major drawback. While arbitrary orders of convergence can be achieved, the number of degrees of freedom needed grows very fast.

In the presented paper, we give an overview of a relatively new idea, originating in [2, 5], to combine the DG method with reconstruction operators to obtain a numerical scheme of very high orders of accuracy, which, we demonstrate, is computationally more efficient than the DG scheme itself. We introduce concepts needed to introduce such a scheme, discuss implementation and numerical results, as well as some theoretical considerations.

1 Problem Formulation and Notation

For simplicity, we shall be concerned with a scalar hyperbolic equation, although the same arguments basically hold for any time-dependent PDE. We treat a nonlinear nonstationary scalar hyperbolic equation in a bounded domain $\Omega \subset \mathbb{R}^d$ with a Lipschitz-continuous boundary $\partial\Omega$. We seek $u : \Omega \times [0,T] \to \mathbb{R}$ such that

$$\frac{\partial u}{\partial t} + \mathrm{div}\,\mathbf{f}(\mathrm{u}) = 0 \quad \text{in } \Omega \times (0,\mathrm{T}) \tag{1}$$

along with an appropriate initial and boundary condition. Here $\mathbf{f} = (f_1, \cdots, f_d)$ and $f_s, s = 1, \ldots, d$ are Lipschitz continuous fluxes in the direction $x_s, s = 1, \ldots, d$.

Let $\mathscr{T}_h$ be a partition (triangulation) of the closure $\overline{\Omega}$ into a finite number of closed simplices $K \in \mathscr{T}_h$. In general we do not require the standard conforming properties of $\mathscr{T}_h$ used in the finite element method (i.e. we admit the so-called hanging nodes). We shall use the following notation. By ∂K we denote the boundary of an element $K \in \mathscr{T}_h$ and set $h_K = \mathrm{diam}(K)$, $h = \max_{K \in \mathscr{T}_h} h_K$.

Let $K, K' \in \mathscr{T}_h$. We say that K and K' are *neighbours*, if they share a common *face* $\Gamma \subset \partial K$. By $\mathscr{F}_h$ we denote the system of all faces of all elements $K \in \mathscr{T}_h$. Further, we define the set of all interior and boundary faces, by $\mathscr{F}_h^I$ and $\mathscr{F}_h^B$, respectively.

For each $\Gamma \in \mathscr{F}_h$ we define a unit normal vector $\mathbf{n}_\Gamma$, such that for $\Gamma \in \mathscr{F}_h^B$ the normal $\mathbf{n}_\Gamma$ has the same orientation as the outer normal to $\partial\Omega$.

Over a triangulation $\mathscr{T}_h$ we define the *broken Sobolev spaces*

$$H^k(\Omega, \mathscr{T}_h) = \{v;\, v|_K \in H^k(K), \forall K \in \mathscr{T}_h\}.$$

For each face $\Gamma \in \mathscr{F}_h^I$ there exist two neighbours $K_\Gamma^{(L)}, K_\Gamma^{(R)} \in \mathscr{T}_h$ such that $\Gamma \subset K_\Gamma^{(L)} \cap K_\Gamma^{(R)}$. We use the convention that $\mathbf{n}_\Gamma$ is the outer normal to $K_\Gamma^{(L)}$. For $v \in H^1(\Omega, \mathscr{T}_h)$ and $\Gamma \in \mathscr{F}_h^I$ we introduce the following notation:

$$v|_\Gamma^{(L)} = \text{ trace of } v|_{K_\Gamma^{(L)}} \text{ on } \Gamma, \quad v|_\Gamma^{(R)} = \text{ trace of } v|_{K_\Gamma^{(R)}} \text{ on } \Gamma, \quad [v]_\Gamma = v|_\Gamma^{(L)} - v|_\Gamma^{(R)}.$$

On boundary edges we define $v|_\Gamma^{(R)} = [v]_\Gamma := v|_\Gamma^{(L)}$.

Let $n \geq 0$ be an integer. We define the space of discontinuous piecewise polynomial functions

$$S_h^n = \{v;\, v|_K \in P^n(K), \forall K \in \mathscr{T}_h\},$$

where $P^n(K)$ is the space of all polynomials on K of degree $\leq n$. Specifically,

- S_h^0: is the space of piecewise constant functions as known from the FV method,
- $S_h^n,\ n \geq 0$: the DG solution lies in this space of piecewise nth degree polynomials,
- $S_h^N,\ N > n$: the higher order reconstructed DG solution will lie in this space.

2 Discontinuous Galerkin

We multiply (1) by an arbitrary $\varphi_h^n \in S_h^n$, integrate over an element $K \in \mathscr{T}_h$ and apply Green's theorem. By summing over all $K \in \mathscr{T}_h$ and rearranging, we get

$$\frac{d}{dt} \int_\Omega u(t)\, \varphi_h^n \, dx + \sum_{\Gamma \in \mathscr{F}_h} \int_\Gamma \mathbf{f}(u) \cdot \mathbf{n}\, [\varphi_h^n]\, dS - \sum_{K \in \mathscr{T}_h} \int_K \mathbf{f}(u) \cdot \nabla \varphi_h^n \, dx = 0. \tag{2}$$

The boundary convective terms will be treated similarly as in the finite volume method, i.e. with the aid of a numerical flux $H(u, v, \mathbf{n})$:

$$\int_\Gamma \mathbf{f}(u) \cdot \mathbf{n}\, [\varphi_h^n]\, dS \approx \int_\Gamma H(u^{(L)}, u^{(R)}, \mathbf{n})\, [\varphi_h^n]\, dS. \tag{3}$$

We assume that H is *Lipschitz continuous, consistent* and *conservative*, cf. [3].

Finally, we define the *convective form* $b_h(\cdot, \cdot)$ defined for $v, \varphi \in H^1(\Omega, \mathscr{T}_h)$:

$$b_h(v, \varphi) = \sum_{\Gamma \in \mathscr{F}_h} \int_\Gamma H(v^{(L)}, v^{(R)}, \mathbf{n})\, [\varphi]\, dS - \sum_{K \in \mathscr{T}_h} \int_K \mathbf{f}(v) \cdot \nabla \varphi \, dx.$$

Definition 1 (Standard DG scheme) *We seek* $u_h : [0,T] \to S_h^n$ *such that*

$$\frac{d}{dt}\big(u_h(t), \varphi_h^n\big) + b_h\big(u_h(t), \varphi_h^n\big) = 0, \quad \forall \varphi_h^n \in S_h^n, \forall t \in (0,T). \tag{4}$$

We note that if we take $n = 0$, i.e. $u_h : (0,T) \to S_h^0$, then from the definition of b_h, we see that the DG scheme (4) is equivalent to the standard FV method.

3 Reconstructed Discontinuous Galerkin

For $v \in L^2(\Omega)$, we denote by $\Pi_h^n v$ the $L^2(\Omega)$-projection of v on S_h^n:

$$\Pi_h^n v \in S_h^n, \quad (\Pi_h^n v - v, \varphi_h^n) = 0, \qquad \forall\, \varphi_h^n \in S_h^n. \tag{5}$$

Obviously, if $K \in \mathscr{T}_h$, then the function $(\Pi_h^n v)|_K$ is the $L^2(K)$-projection of $v|_K$ on $P^n(K)$. The basis of the method lies in the observation that (2) can be viewed as an equation for the evolution of $\Pi_h^n u(t)$, where u is the exact solution of (1). In other words, due to (5), $\Pi_h^n u(t) \in S_h^n$ satisfies the following equation for all $\varphi_h^n \in S_h^n$:

$$\frac{d}{dt}\int_\Omega \Pi_h^n u(t)\, \varphi_h^n \, dx + \sum_{\Gamma \in \mathscr{F}_h} \int_\Gamma \mathbf{f}(u) \cdot \mathbf{n}\, [\varphi_h^n]\, dS - \sum_{K \in \mathscr{T}_h} \int_K \mathbf{f}(u) \cdot \nabla \varphi_h^n \, dx = 0. \tag{6}$$

Now, let $N > n$ be an integer. We assume that there exists a piecewise polynomial function $U_h^N(t) \in S_h^N$, which is an approximation of $u(t)$ of order $N+1$, i.e.

$$U_h^N(x,t) = u(x,t) + O(h^{N+1}), \quad \forall x \in \Omega, \forall t \in [0,T]. \tag{7}$$

This is possible, if u is sufficiently regular in space, e.g. $u(t) \in W^{N+1,\infty}(\Omega)$, cf.[1]. Now we incorporate the approximation $U_h^N(t)$ into (6): the exact solution u satisfies

$$\frac{d}{dt}\big(\Pi_h^n u(t), \varphi_h^n\big) + b_h\big(U_h^N(t), \varphi_h^n\big) = E(\varphi_h^n, t), \quad \forall \varphi_h^n \in S_h^n, \forall t \in (0,T), \tag{8}$$

where $E(\varphi_h^n, t)$ is an error term defined as

$$E(\varphi_h^n, t) = b_h\big(U_h^N(t), \varphi_h^n\big) - b_h\big(u(t), \varphi_h^n\big). \tag{9}$$

Lemma 1 *The following estimate holds for all* $t \in [0,T]$*:*

$$E(\varphi_h^n, t) = O(h^N) \|\varphi_h^n\|_{L^2(\Omega)}. \tag{10}$$

Proof: Due to the consistency and Lipschitz continuity of H, we have on $\Gamma \in \mathscr{F}_h$

$$\mathbf{f}(u) \cdot \mathbf{n} - H(U_h^{N,(L)}, U_h^{N,(R)}, \mathbf{n}) = H(u,u,\mathbf{n}) - H(U_h^{N,(L)}, U_h^{N,(R)}, \mathbf{n}) = O(h^{N+1}).$$

Furthermore, due to the Lipschitz-continuity of $\mathbf{f}$, we have on element $K \in \mathscr{T}_h$

$$\mathbf{f}(u) - \mathbf{f}(U_h^N) = O(h^{N+1}).$$

Estimate (10) follows from these results and the application of the *inverse* and *multiplicative trace inequalities*, cf. [3]. □

It remains to construct a sufficiently accurate approximation $U_h^N(t) \in S_h^N$ to $u(t)$, such that (7) is satisfied. This leads to the following problem.

Definition 2 (Reconstruction problem) *Let $v : \Omega \to \mathbb{R}$ be sufficiently regular. Given $\Pi_h^n v \in S_h^n$, find $v_h^N \in S_h^N$ such that $v - v_h^N = O(h^{N+1})$ in Ω. We define the corresponding reconstruction operator $R : S_h^n \to S_h^N$ by $R\Pi_h^n v := v_h^N$.*

By setting $U_h^N(t) := R\Pi_h^n u(t)$ in (8), we obtain the following equation for the $L^2(\Omega)$-projections of the exact solution u onto the space S_h^n:

$$\frac{d}{dt}\left(\Pi_h^n u(t), \varphi_h^n\right) + b_h\left(R\Pi_h^n u(t), \varphi_h^n\right) = O(h^N)\|\varphi_h^n\|_{L^2(\Omega)}, \quad \forall \varphi_h^n \in S_h^n. \tag{11}$$

By neglecting the right-hand side and approximating $u_h^n(t) \approx \Pi_h^n u(t)$, we arrive at the following definition of the *reconstructed discontinuous Galerkin* (RDG) scheme.

Definition 3 (Reconstructed DG scheme) *We seek $u_h^n : [0,T] \to S_h^n$ such that*

$$\frac{d}{dt}\left(u_h^n(t), \varphi_h^n\right) + b_h\left(Ru_h^n(t), \varphi_h^n\right) = 0, \quad \forall \varphi_h^n \in S_h^n, \forall t \in (0,T). \tag{12}$$

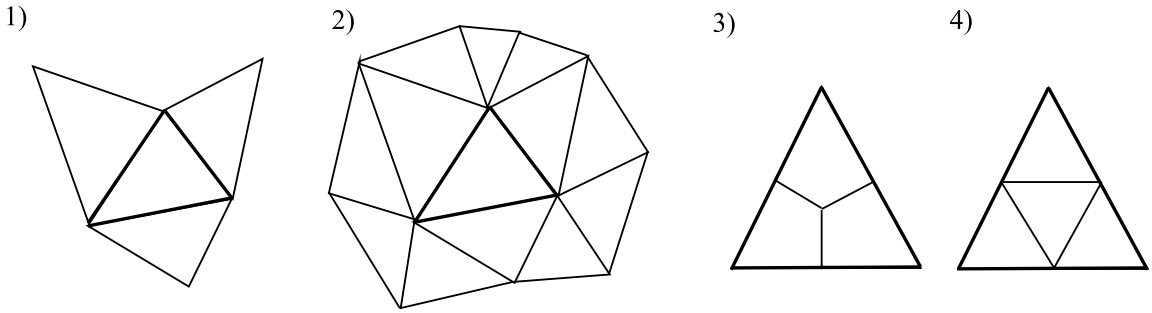
There are several points worth mentioning.

- Equation (11) indicates that the RDG scheme is formally Nth order in space.
- The derivation of the RDG scheme follows the methodology of higher order FV and SV schemes, cf. [7]. The basis of these schemes is an equation for the evolution of averages of the exact solution on individual elements (i.e. an equation for $\Pi_h^0 u(t)$). Equation (11) is a direct generalization for the case of higher order $L^2(\Omega)$-projections $\Pi_h^n u(t)$, $n \geq 0$.
- Both $u_h^n(t)$ and φ_h^n lie in S_h^n. Only $Ru_h^n(t)$, lies in the higher dimensional space S_h^N. Despite this fact, equation (11) indicates that we may expect $u - Ru_h^n = O(h^{N+1})$, although $u - u_h^n = O(h^{n+1})$.
- Numerical quadrature must be employed to evaluate surface and volume integrals in (12). Since test functions are in S_h^n, as compared to S_h^N in the corresponding Nth order standard DG scheme, we may use lower order (i.e. more efficient) quadrature formulae as compared to the standard DG method.
- In practice, we must also discretize (12) with respect to time. As in the case of higher order FV methods, we use an explicit time stepping method. The upper limit on stable time steps, given by a CFL-like condition, is more restrictive with growing N. However, in the RDG scheme, stability properties are essentially inherited from the lower order scheme, therefore a larger time step is possible as compared to the corresponding Nth order standard DG scheme.

3.1 Explicit Time Discretization

For simplicity, we formulate the forward Euler method, which is only the first order accurate, however in Section 5, higher order Adams-Bashforth methods are used.

Let us construct a partition $0 = t_0 < t_1 < t_2 \ldots$ of the time interval $[0,T]$ and define the time step $\tau_k = t_{k+1} - t_k$. We use the approximation $u_h^{n,k} \approx u_h^n(t_k)$, where $u_h^{n,k} \in S_h^n$. The forward Euler scheme is given by:



Source: Own

Fig. 1. *1) FV stencil for linear reconstruction, 2) FV stencil for quadratic reconstruction, 3) Control volumes in a spectral volume for linear reconstruction, 4) Analogy to the SV approach for DG methods - partition of triangle into control volumes, e.g. cubic reconstruction from linear data*

Definition 4 (Explicit RDG scheme) *We seek* $u_h^{n,k} \in S_h^n,\ k = 0,1,\ldots$ *such that*

$$\left(\frac{u_h^{n,k+1} - u_h^{n,k}}{\tau_k}, \varphi_h^n\right) + b_h\left(Ru_h^{n,k}, \varphi_h^n\right) = 0, \quad \forall \varphi_h^n \in S_h^n,\ k = 0,1,\ldots, \tag{13}$$

where $u_h^{n,0} = u_{h,0}$ *is an* S_h^n *approximation of the initial condition* u^0.

3.2 Construction of the Reconstruction Operator

In analogy to the construction of reconstruction operators in higher order FV schemes, we propose two approaches.

3.2.1 ”Standard” Approach

In the *standard approach*, a stencil (a group of neighboring elements and the element under consideration) is used to build an Nth degree polynomial approximation to u on the element under consideration ([4, 6]). In the FV method, the von Neumann neighborhood of an element is used as a stencil to obtain a piecewise linear reconstruction, cf. Figure 1, 1). However, for higher order reconstructions, the size of the stencil increases dramatically, cf. Figure 1, 2), causing higher degrees than quadratic to be very time consuming. In the case of the RDG scheme, we need not increase the stencil size to obtain higher order accuracy, it suffices to take the von Neumann neighborhood and increase the order of the underlying DG scheme.

In analogy to the FV method, the reconstruction operator R is constructed on each stencil independently and satisfies that $R\Pi_h^n$ is in some sense *polynomial preserving*. Specifically, for each element K and its corresponding stencil S, we require that for all $p \in P^N(S)$

$$\left(\left(R\Pi_h^n\right)\big|_S p\right)\Big|_K = p\big|_K. \tag{14}$$

This requirement allows us to study approximation properties of R using the Bramble–Hilbert technique as in the standard finite element method, [1]. The disadvantage of this approach is that for unstructured meshes, the coefficients of the reconstruction operator must be stored for each individual stencil.

In the FV method, different conditions on R than (14) are often used, e.g. continuous or discrete least squares. Special care must be taken in the vicinity of steep gradients and discontinuities, where the Gibbs phenomenon may occur. In this case different strategies are employed, e.g. limiting, ENO and WENO schemes, TVD etc. The generalization of these concepts to the RDG method is left for future work.

3.2.2 Spectral Volume Approach

In the *spectral volume approach*, we start with a partition of Ω into so-called *spectral volumes* S, for example triangles in 2D. The triangulation $\mathscr{T}_h$ is formed by subdividing each spectral volume S into sub-cells K, called *control volumes*, cf. [7]. In the FV method, the order of accuracy of the reconstruction determines the number of control volumes to be generated in each spectral volume. For example, for a linear reconstruction on a triangle, the triangle is divided into three control volumes, Figure 1, 3). Again, in the RDG scheme, we may use only the smallest available partition into control volumes, and increase the accuracy by increasing the order of the underlying scheme, cf. Figure 1, 4).

The reconstruction operator is constructed on each spectral volume independently such that it is in some sense polynomial preserving, i.e. for each spectral volume S, we require that for all $p \in P^N(S)$

$$\left(R\Pi_h^n\right)\big|_S p = p. \tag{15}$$

The advantage of this approach is that all spectral volumes are affine equivalent, we construct the reconstruction operator R only on one reference spectral volume.

4 Relation between RDG and Standard DG

The only difference between the DG scheme (4) and RDG scheme (12) is the presence of the reconstruction operator R in the first variable of $b_h(\cdot,\cdot)$. While the error analysis of (4) is well understood (at least for convection-diffusion problems [3]), the analysis of (12) or (13) poses a new challenge. The problem lies in the fact that we cannot test (12) with $\varphi_h^n := Ru_h^{n,k}$ or something similar, since $Ru_h^{n,k} \notin S_h^n$. Therefore, we need to establish a relation between (12) and the Nth order DG method, instead of only the nth order DG method.

Definition 5 (Auxiliary problem) *We seek $\tilde{u}_h^{N,k} \in S_h^N$ such that*

$$\left(\frac{\tilde{u}_h^{N,k+1} - \tilde{u}_h^{N,k}}{\tau_k}, \varphi_h^N\right) + b_h\left(R\Pi_h^n \tilde{u}_h^{N,k}, \varphi_h^N\right) = 0, \quad \forall \varphi_h^N \in S_h^N,\ k = 0,1,\ldots, \tag{16}$$

where $\tilde{u}_h^{N,0}$ is an S_h^N approximation of the initial condition u^0.

Lemma 2 *Let $u_h^{n,0} = \Pi_h^n \tilde{u}_h^{N,0}$. Then $u_h^{n,k} \in S_h^n$, the solution of (13) and the solution $\tilde{u}_h^{N,k} \in S_h^N$ of (16) satisfy*

$$u_h^{n,k} = \Pi_h^n \tilde{u}_h^{N,k}, \quad \forall k = 0,1,\cdots. \tag{17}$$

Proof: We prove (17) by induction:

$k = 1$: Since $u_h^{n,0} = \Pi_h^n \tilde{u}_h^{N,0}$, we have for all $\varphi_h^n \in S_h^n$

$$\begin{aligned}(\Pi_h^n \tilde{u}_h^{N,1}, \varphi_h^n) &= (\tilde{u}_h^{N,1}, \varphi_h^n) = (\tilde{u}_h^{N,0}, \varphi_h^n) - \tau_k b_h\left(R\Pi_h^n \tilde{u}_h^{N,0}, \varphi_h^n\right) \\ &= (u_h^{n,0}, \varphi_h^n) - \tau_k b_h\left(Ru_h^{n,0}, \varphi_h^n\right) = (u_h^{n,1}, \varphi_h^n),\end{aligned}$$

hence $(\Pi_h^n \tilde{u}_h^{N,1} - u_h^{n,1}, \varphi_h^n) = 0$ for all $\varphi_h^n \in S_h^n$. Therefore $\Pi_h^n \tilde{u}_h^{N,1} = u_h^{n,1}$.

$k > 1$: Assume (17) holds for some $k > 1$. Then for all $\varphi_h^n \in S_h^n$

$$\begin{aligned}\left(\Pi_h^n \tilde{u}_h^{N,k+1}, \varphi_h^n\right) &= \left(\tilde{u}_h^{N,k+1}, \varphi_h^n\right) = \left(\tilde{u}_h^{N,k}, \varphi_h^n\right) - \tau_k b_h\left(R\Pi_h^n \tilde{u}_h^{N,k}, \varphi_h^n\right) \\ &= \left(u_h^{n,k}, \varphi_h^n\right) - \tau_k b_h\left(R u_h^{n,k}, \varphi_h^n\right) = \left(u_h^{n,k+1}, \varphi_h^n\right),\end{aligned}$$

therefore $\Pi_h^n \tilde{u}_h^{N,k+1} = u_h^{n,k+1}$. This completes the induction step $k \to k+1$. □

As a corollary, error estimates for the auxiliary problem imply error estimates for the RDG scheme (12). Problem (16) is basically the standard Nth order DG scheme with the operator $R\Pi_h^n$ in the first variable of $b_h(\cdot,\cdot)$. Therefore, sufficient knowledge of the properties of $R\Pi_h^n$ (which is polynomial preserving) and standard DG method error estimates would imply the estimates for the RDG scheme.

5 Numerical Experiments

We present numerical experiments for the periodic advection of a 1D sine wave on uniform meshes. We use the P^1 RDG scheme with P^5 reconstruction and the P^2 RDG scheme with P^8 reconstruction. The reconstruction operators described in Section 3.2.1 are used. Experimental orders of accuracy α in various norms on meshes with N elements are given in Tables 1 and 2. Here $e_h = u - Ru_h^n$ at t corresponding to ten periods. The increase in accuracy due to reconstruction is clearly visible.

Tab. 1. *1D advection of sine wave, P^1 RDG scheme with P^5 reconstruction*

N	$\|\|e_h\|\|_{L^\infty(\Omega)}$	α	$\|\|e_h\|\|_{L^2(\Omega)}$	α	$\|e_h\|_{H^1(\Omega,\mathscr{T}_h)}$	α
4	5.82E-03	–	3.49E-03	–	3.65E-02	–
8	7.53E-05	6.27	4.43E-05	6,30	1.06E-03	5,11
16	9.07E-07	6.38	5.95E-07	6,22	3.58E-05	4,89
32	1.82E-08	5.64	8.70E-09	6,10	1.16E-06	4,95
64	3.41E-10	5.74	1.33E-10	6,03	3.67E-08	4,98

Source: Own

Tab. 2. *1D advection of sine wave, P^2 RDG scheme with P^8 reconstruction*

N	$\|\|e_h\|\|_{L^\infty(\Omega)}$	α	$\|\|e_h\|\|_{L^2(\Omega)}$	α	$\|e_h\|_{H^1(\Omega,\mathscr{T}_h)}$	α
4	2.90E-03	–	1.85E-03	–	1.63E-02	–
8	7.75E-06	8.55	3.56E-06	9.02	1.03E-04	7.30
16	2.10E-08	8.53	6.64E-09	9.07	4.34E-07	7.89
32	7.21E-11	8.18	4.02E-11	7.37	1.76E-09	7.94

Source: Own

Conclusion

We have presented a possible generalization of higher order reconstruction operators as used in the FV method to the DG method. Two constructions of the reconstruction operator R are presented, the first analogous to the standard FV case (already treated in [2]) and the construction analogous to the SV method. The resulting scheme has many advantages over standard DG, FV and SV schemes:

- To increase the order of the scheme, the reconstruction stencil need not be enlarged, we may simply increase the order of the underlying DG scheme.
- Test functions are from the lower order space, hence more efficient quadratures may be used than in the corresponding higher order DG scheme.
- Since the RDG scheme is basically a lower order DG scheme with higher order reconstruction, the CFL condition is less restrictive than for the corresponding higher order DG scheme.

Acknowledgments

This work was supported by the Grant No. P201/11/P414 of the Czech Science Foundation. The author is a (junior) researcher (or member) of the University center for mathematical modelling, applied analysis and computational mathematics (Math MAC).

Literature

[1] CIARLET, P.G.: *The Finite Elements Method for Elliptic Problems*. North-Holland, Amsterdam, New York, Oxford, 1979.

[2] DUMBSER, M.; BALSARA, D.; TORO, E.F.; MUNZ, C.D.: A unified framework for the construction of one-step finite-volume and discontinuous Galerkin schemes. *J. Comput. Phys.* 227 (2008), pp. 8209–8253.

[3] FEISTAUER, M.; KUČERA, V.: Analysis of the DGFEM for nonlinear convection-diffusion problems. *Electronic Transactions on Numerical Analysis*, Vol. 32, No.1, (2008), pp. 33–48.

[4] KRÖNER, D.: *Numerical Schemes for Conservation Laws*. Wiley und Teubner, 1996.

[5] KUČERA, V.: Higher-Order Reconstruction: From Finite Volumes to Discontinuous Galerkin. *Proc. of FVCA 6, Finite Volumes for Complex Applications VI Problems and Perspectives*, Springer Berlin Heidelberg, pp. 613–621, 2011.

[6] LEVEQUE, R.J.: *Finite Volume Methods for Hyperbolic Problems*. Cambridge University Press, Cambridge, 2002.

[7] WANG, Z. J.: Spectral (Finite) Volume Method for Conservation Laws on Unstructured Grids. Basic Formulation. *J. Comput. Phys.* 178 (2002), pp. 210–251.

RNDr. Václav Kučera, Ph.D.

Aplikace rekonstrukčních operátorů v nespojité Galerkinově metodě

Tento článek představuje přehled základních konceptů nutných k zapojení rekonstrukčních operátorů, známých z metody konečných objemů (FV) a spektrálních objemů (SV) vyššího řádu, do nespojité Galerkinovy (DG) metody. Takovýto operátor konstruuje aproximace vyšších řádů z DG metody nižšího řádu, čímž zvyšuje řád konvergence, zároveň vede k efektivnějšímu numerickému schématu než příslušné DG schéma vyššího řádu samo o sobě. Probereme teoretické i implementační aspekty a numerické experimenty.

Die Anwendung von Rekonstruktionsoperatoren in der kontinuierlichen Galerkin-Methode

Dieser Artikel bietet einen Überblick über die grundlegenden, zu Einbeziehung der Rekonstruktionsoperatoren notwendigen Konzepte, die aus der Methode des finiten Inhalts (FV) und der spektralen Inhalte (SV) höherer Ordnung bekannt sind, in die nicht kontinuierliche Galerkin-Methode (DG). Ein solcher Operator konstruiert eine Annäherung hherer Ordnungen aus der DG-Methode niederer Ordnung, wodurch er die Ordnung der Konvergenz erhöht. Gleichzeitig führt er zu einem effektiveren numerischen Schema als das zugehörige DG-Schema höherer Ordnung selbst. Wir behandeln die theoretischen und implementären Aspekte und die numerischen Experimente.

Zastosowanie operatorów rekonstrukcyjnych w nieciągłej metodzie Galerkina

W niniejszym artykule przedstawiono podstawowe koncepcje niezbędne do włączenia operatorów rekonstrukcyjnych, znanych z metody objętości skończonych (FV) oraz objętości spektralnych (SV) wyższego rzędu, do nieciągłej metody Garlekina (DG). Taki operator konstruuje aproksymacje wyższego rzędu z metody DG niższego rzędu, co zwiększa poziom konwergencji i jednocześnie prowadzi do bardziej efektywnego schematu numerycznego w porównaniu ze schematem DG wyższego rzędu. Omówiono aspekty teoretyczne i implementacyjne oraz eksperymenty numeryczne.

NUMERICAL SOLUTION OF REACTION-DIFFUSION EQUATIONS

Jan Lamač

Technical University of Liberec
Faculty of Science, Humanities and Education
Department of Mathematics and Didactics of Mathematics
Studentská 2, 461 17, Liberec, Czech Republic

*Charles University in Prague
Faculty of Mathematics and Physics
Department of Numerical Mathematics
Sokolovská 83, Praha 8, 186 75, Czech Republic

**Czech Technical University in Prague
Faculty of Nuclear Sciences and Physical Engineering
Břehová 7, 115 19 Praha, Czech Republic
jan.lamac@centrum.cz

Abstract

The subject of the presented paper is a mathematical analysis and numerical solution of the system of nonlinear nonstationary reaction-diffusion equations. Firstly, using the invariant region technique, the proof of both the existence and uniqueness of the solution and problem data continuous dependence is carried out. After time discretization of the problem the Galerkin finite elements method is applied and a priori error estimates of the method are derived. A suitable mesh adaptivity is discussed as well. The method is finally implemented and tested on several examples.

Keywords: Invariant region; reaction; diffusion; finite elements method; adaptivity.

Introduction

Reaction-diffusion equations provide an efficient mathematical model for description of changes in concentration of one or more substances under an influence of two basic physical (and chemical) processes - diffusion and reaction. Using these equations one can also describe several mechanisms that occure in nature (see [5]). One of them is the evolution of bacteria populations.

Let us denote $u_i = u_i(x,t)$, $i = 1,2,\ldots,N$, concentrations of N types of bacteria at time t and place x, then the mathematical model of the evolution of bacteria populations is described by the system of nonlinear nonstationary reaction-diffusion equations (see [6],[7],[8])

$$\frac{\partial u_i}{\partial t} = D_i \Delta u_i + \left(r_i - \sum_{j=1}^{N} a_{ij} u_j \right) u_i, \qquad \text{in } Q_T, \quad i = 1,2,\ldots,N, \tag{1}$$

where $Q_T = \Omega \times (0,T)$ and $D_i, r_i > 0, a_{ij} \geq 0$ are given constants. For each function u_i, $i = 1,2,\ldots,N$, we also prescribe initial and boundary conditions

$$u_i(x,0) = u_{0i}(x) \qquad \text{in } \Omega, \tag{2}$$

$$\frac{\partial u_i}{\partial \boldsymbol{n}} = g_i \qquad \text{on } \Gamma_N \times (0,T), \tag{3}$$

$$u_i = u_{Di} \qquad \text{on } \Gamma_D \times (0,T), \tag{4}$$

where $\boldsymbol{u}_0 : \Omega \to \mathbb{R}^N$, $\boldsymbol{g}$: $\Gamma_N \times (0,T) \to \mathbb{R}^N$ and $\boldsymbol{u}_D : \Gamma_D \times (0,T) \to \mathbb{R}^N$ are given functions and Γ_N, Γ_D are parts of the boundary $\partial\Omega$ such that $\overline{\Gamma_N \cup \Gamma_D} = \partial\Omega$ and $\Gamma_N \cap \Gamma_D = \emptyset$.

1 Weak Formulation

Let us denote $I = (0,T)$ and suppose that for each $i = 1,2,\ldots,N$, there exist functions $\widetilde{u}_{Di} \in L^2(I,H^1(\Omega)) \cap L^\infty(I,L^2(\Omega))$ with $\widetilde{u}'_{Di} \in L^2(I,H^{-1}(\Omega))$ such that for almost all $t \in I$ holds $\widetilde{u}_{Di}(t)|_{\Gamma_D} = u_{Di}(t)$ in the sense of traces. Further, let $g_i \in L^2(I,L^2(\Gamma_N))$ and $u_{0i} \in L^2(\Omega)$ for each $i = 1,2,\ldots,N$ and in order to exclude trivial solutions let us also suppose that for each $i = 1,2,\ldots,N$, one has $\|u_{0i}\|_2 \geq m_i > 0$.

If we then define $V = \{v \in H^1(\Omega) |\, v = 0 \text{ on } \Gamma_D \text{ in the sense of traces}\}$, we call $\boldsymbol{u} \in L^\infty(I,L^2(\Omega))^N \cap L^2(I,H^1(\Omega))^N$ the weak solution of the problem (1) provided that for each $i = 1,2,\ldots,N$, holds

- $\frac{\mathrm{d}u_i}{\mathrm{d}t} \in L^2(I,V^*)$,
- $u_i(0) = u_{0i}$,
- $u_i(t) - \widetilde{u}_{Di}(t) \in V$ for almost all $t \in I$
- and for each $v \in V$ and almost all $t \in I$ holds
  $\left(\frac{\mathrm{d}u_i}{\mathrm{d}t}(t), v\right) + D_i(\nabla u_i(t), \nabla v) = \left((r_i - \sum_{j=1}^{N} a_{ij}u_j(t))u_i(t), v\right) + D_i(g_i, v)_{\Gamma_N}$.

This formulation can be easily derived using integration and Green's theorem.

2 Existence and Uniqueness

The existence and uniqueness of the solution of the problem (1) follow from the boundedness of this solution. The boundedness of the solution of the problem (1) can be proved by construction of the *invariant region* (see [4] or [10]).

Definition 1. *The invariant region of the problem (1) is every closed set $\Sigma \subset \mathbb{R}^N$ with the property*

$$\boldsymbol{u}_0, \boldsymbol{u}_D \in \Sigma \text{ for all } t \in [0,T],\ x \in \Omega \Rightarrow \boldsymbol{u}(x,t) \in \Sigma \text{ for all } t \in [0,T],\ x \in \Omega. \tag{5}$$

Let us denote $f_i(\boldsymbol{z}) = \left(r_i - \sum_{j=1}^{N} a_{ij}z_j\right) z_i$. If one wants to certify that Σ is the invariant region of the problem (1), it is sufficient (see [10]) to prove that for each $\boldsymbol{z} \in \partial\Sigma$ the vector $\boldsymbol{f}(\boldsymbol{z})$ points inside the set Σ. Considering $a_{ii} > 0$ for each $i = 1,2,\ldots,N$, we can define $\Sigma = \times_{i=1}^{N}[0,M_i]$, where $M_i > \max\{\|u_{0i}\|_\infty, \|u_{Di}\|_\infty, r_i/a_{ii}\}$. Consequently, for every $\boldsymbol{z} \in \partial\Sigma$ there exists $i \in \{1,2,\ldots,N\}$ such that $z_i = 0$ or $z_i = M_i$ and thus we can denote $\boldsymbol{z}^{iL}, \boldsymbol{z}^{iR} \in \partial\Sigma$ such that $z_i^{iL} = 0$ and $z_i^{iR} = M_i$. Finally, it is obvious, that vectors $\boldsymbol{f}(\boldsymbol{z}^{iL}), \boldsymbol{f}(\boldsymbol{z}^{iR})$ points inside the set Σ if and only if $f_i(\boldsymbol{z}^{iL}) > 0$ and $f_i(\boldsymbol{z}^{iR}) < 0$. In the second case one has

$$f_i(\boldsymbol{z}^{iR}) = \left(r_i - \sum_{\substack{j=1 \\ j\neq i}}^{N} a_{ij}z_j^{iR} - a_{ii}M_i\right) M_i < -M_i \sum_{\substack{j=1 \\ j\neq i}}^{N} a_{ij}x_j^{iR} \leq 0, \tag{6}$$

while in the first case holds

$$f_i(\boldsymbol{z}^{iL}) = (r_i - \sum_{\substack{j=1 \\ j\neq i}}^{N} a_{ij}z_j^{iL} - a_{ii}.0) \cdot 0 = 0. \tag{7}$$

Thus the vector $\boldsymbol{f}(\boldsymbol{z}^{iL})$ is tangent to $\partial\Sigma$ and does not points inside the set Σ. Nevertheless, this drawback can be removed by showing that the problem (1) is *f-stable*.

Definition 2. *We call the problem (1) f-stable, if there exists a sequence of functions* $\{\boldsymbol{f}_m\}_{k=1}^{\infty}$ *such that* $\|\boldsymbol{f}-\boldsymbol{f}_m\|_{\mathscr{C}^1(\Sigma)^N} \overset{m\to\infty}{\longrightarrow} 0$ *implies* $\|\boldsymbol{u}-\boldsymbol{u}_m\|_{\mathscr{C}(I,L^2(\Omega))^N} \overset{m\to\infty}{\longrightarrow} 0$*, where* $\boldsymbol{u}$ *is the solution of the problem (1) with the function* $\boldsymbol{f}$ *on the right-hand side and* $\boldsymbol{u}_m$ *is the solution of the problem (1) with the function* $\boldsymbol{f}_m$ *on the right-hand side.*

Remark 1. *The f-stability of the problem (1) can be shown using the sequence* $\boldsymbol{f}_m$ *satisfying* $f_{m,i}(\boldsymbol{z}) = f_i(\boldsymbol{z}) + \frac{\|u_{0i}\|_2 - z_i}{m}$.

Theorem 1. *(Existence, uniqueness and data continuous dependence) Let the functions* $\boldsymbol{u}_0, \widetilde{\boldsymbol{u}}_D$ *and* $\boldsymbol{g}$ *satisfy assumptions from the previous section. If the problem (1) is f-stable, then there exists a unique solution* $\boldsymbol{u}$ *of the problem (1) and positive constants* $B_i = B_i(T)$, $i = 1,2,3$, *independent from* $\boldsymbol{u}$*, such that :*

$$\sup_{t\in[0,T]} \|\boldsymbol{u}(t)\|_2^2 \;\leq\; B_1\, \mathscr{K}(\boldsymbol{u}_0, \widetilde{\boldsymbol{u}}_D, \boldsymbol{g}), \tag{8}$$

$$\int_0^T \|\nabla \boldsymbol{u}(t)\|_2^2 \,\mathrm{d}t \;\leq\; B_2\, \mathscr{K}(\boldsymbol{u}_0, \widetilde{\boldsymbol{u}}_D, \boldsymbol{g}), \tag{9}$$

$$\int_0^T \|(\boldsymbol{u})'(t)\|_{-1,2}^2 \,\mathrm{d}t \;\leq\; B_3\, \mathscr{K}(\boldsymbol{u}_0, \widetilde{\boldsymbol{u}}_D, \boldsymbol{g}), \tag{10}$$

where

$$\mathscr{K}(\boldsymbol{u}_0, \widetilde{\boldsymbol{u}}_D, \boldsymbol{g}) = \|\boldsymbol{u}_0\|_2^2 + \sup_{t\in[0,T]} \|\widetilde{\boldsymbol{u}}_D(t)\|_2^2 + \int_0^T \|\widetilde{\boldsymbol{u}}_D'(t)\|_{-1,2}^2 + \|\widetilde{\boldsymbol{u}}_D(t)\|_{1,2}^2 + \|\boldsymbol{g}(t)\|_{2,\Gamma_N}^2 \,\mathrm{d}t. \tag{11}$$

Proof. Sketch of the proof: Since $\boldsymbol{u}$ is bounded, $\boldsymbol{f}(\boldsymbol{u})$ is Lipschitz-continuous and the existence and uniqueness come from the Banach fix-point theorem and Gronwall's lemma respectively. Estimates (8)-(10) can be derived from the weak formulation using Young's inequality. □

Previous theorem also implies that for sufficiently smooth data ($\mathscr{K}(\boldsymbol{u}_0, \widetilde{\boldsymbol{u}}_D, \boldsymbol{g}) < \infty$), the solution of the problem (1) belongs to $\mathscr{C}((0,T), L^2(\Omega))$.

3 Numerical Solution

Let us choose $p \in \mathbb{N}$, denote $\tau = T/p$ and define a partition of the interval $[0,T]$: $t_k = k\tau$, for $k = 0,1,\ldots,p$. Further, let us denote $u_i^k(x) = u_i(x,t_k)$, $g_i^k(x) = g_i(x,t_k)$ and $u_{Di}^k(x) = u_{Di}(x,t_k)$, for $i = 1,2,\ldots,N$ and $k = 1,2,\ldots,p$. We also suppose $u_i^0(x) = u_{0,i}(x)$ for $i = 1,2,\ldots,N$ and approximate the time derivative using the backward difference, i.e. $u_i'(x,t_{k+1}) \approx \frac{u_i^{k+1}(x) - u_i^k(x)}{\tau}$ for $k = 0,1,\ldots,p-1$.

Then the time-discretization (see [9]) of the problem (1) reads: For each $k = 0,1,\ldots,p-1$ find a function $\boldsymbol{u}^{k+1} : \Omega \to \mathbb{R}^N$ such that for each $i = 1,2,\ldots,N$, holds

$$\frac{u_i^{k+1} - u_i^k}{\tau} \;=\; D_i \Delta u_i^{k+1} + \left(r_i - \sum_{j=1}^N a_{ij} u_j^k\right) u_i^{k+1} \qquad \text{in } \Omega, \tag{12}$$

$$\frac{\partial u_i^{k+1}}{\partial \boldsymbol{n}}(x) \;=\; g_i^{k+1}(x) \;=\; g_i(x,t_{k+1}) \qquad \text{on } \Gamma_N, \tag{13}$$

$$u_i^{k+1}(x) \;=\; u_{Di}^{k+1}(x) \;=\; u_{Di}(x,t_{k+1}) \qquad \text{on } \Gamma_D, \tag{14}$$

where we used linearization $u_j^k u_i^{k+1} \approx u_j^{k+1} u_i^{k+1}$. Consequently, the corresponding weak formulation of the problem (12)-(14) reads: For each $k = 0, 1, \ldots, p-1$ and $i = 1, 2, \ldots, N$, find $u_i^{k+1} \in H^1(\Omega)$ such that

- $u_i^{k+1} - \widetilde{u}_{Di}^{k+1} \in V$
- and for all $v \in V$ holds
  $\left(\frac{1}{\tau} - r_i\right)(u_i^{k+1}, v) + D_i(\nabla u_i^{k+1}, \nabla v) + \sum_{j=1}^{N} a_{ij}(u_j^k u_i^{k+1}, v) = \frac{1}{\tau}(u_i^k, v) + D_i(g_i^{k+1}, v)_{\Gamma_N}$.

3.1 Finite Elements Method

Let us define the triangulation $\mathscr{T}_h$ of the domain Ω consisting of a finite number of triangular elements K_i and satisfying $\overline{K_i} \cap \overline{K_j} \in \{\emptyset$, a common vertex, edge or face$\}$ for each $K_j, K_i \in \mathscr{T}_h$ (see [1]). Further, let us define a finite-dimensional space $X_h = \{v_h \in C(\overline{\Omega}) |\ v_h|_K \in P_1(K)\ \forall K \in \mathscr{T}_h\} \subset H^1(\Omega)$ and define also $V_h = X_h \cap V$.

If we denote $\{\varphi_h^m\}_{m=1}^{N_h}$ the basis of the space V_h and $u_{hDi}^k \in X_h$ the X_h-interpolation of $\widetilde{u}_{Di}^k$, we can define the solution of (12)-(14) obtained by the finite elements method as $u_{hi}^k = u_{hDi}^k + \sum_{m=1}^{N_h} c_{i,m}^k \varphi_h^m$, where $c_{i,m}^k \in \mathbb{R}$, $i = 1, 2, \ldots, N$. The function u_{hi}^k then satisfies

- $u_{hi}^{k+1} - u_{hDi}^{k+1} \in V_h$
- and for all $s = 1, 2, \ldots, N_h$ holds
  $\sum_{m=1}^{N_h} b_{i,sm}^k c_{i,m}^{k+1} = f_{i,s}^k$, where

$$b_{i,sm}^k = \left(\frac{1}{\tau} - r_i\right)(\varphi_m, \varphi_s) + D_i(\nabla\varphi_m, \nabla\varphi_s) + \sum_{j=1}^{N} a_{ij}(u_j^k \varphi_m, \varphi_s) \qquad \text{and} \qquad (15)$$

$$\begin{aligned} f_{i,s}^k = {} & \frac{1}{\tau}(u_{hi}^k, \varphi_s) + D_i(g_i^{k+1}, \varphi_s)_{\Gamma_N} - \\ & - \left(\frac{1}{\tau} - r_i\right)(u_{hDi}^{k+1}, \varphi_s) - D_i(\nabla u_{hDi}^{k+1}, \nabla\varphi_s) - \sum_{j=1}^{N} a_{ij}(u_{hj}^k u_{hDi}^{k+1}, \varphi_s). \end{aligned} \qquad (16)$$

Thus one can compute a vector $\boldsymbol{c}_i^{k+1} = (c_{i,s}^{k+1})_{s=1}^{N_h}$ by solving the system of linear equations $\mathbb{B}_i^k \boldsymbol{c}_i^{k+1} = \boldsymbol{f}_i^k$ with a matrix $\mathbb{B}_i^k = (b_{i,sm}^k)_{s,m=1}^{N_h}$ and a right-hand side $\boldsymbol{f}_i^k = (f_{i,s}^k)_{s=1}^{N_h}$.

3.2 Mesh Adaptivity

When using some advanced techniques (e.g. [2]) for mesh-adaptation, we have to combine the properties of all N functions and thus it is necessary to construct one function that is a "combination" of them. This function should change its value rapidly wherever any of N functions changes its value rapidly. One can construct such a function for instance in a following way: At each node P_j of the triangulation $\mathscr{T}_h$ let us order all values $u_{hi}^k(P_j)$, $i = 1, 2, \ldots, N$, in such a way that $u_{hi_m}^k(P_j) \geq u_{hi_{m+1}}^k(P_j)$ for $m = 1, 2, \ldots, N-1$. Then the value of the compound function ψ_h^k at P_j is defined by

$$\psi_h^k(P_j) = \sum_{m=1}^{N} (-1)^{m+1} u_{hi_m}^k(P_j). \qquad (17)$$

However, one has to also choose a suitable projection between a new and the old mesh. This projection can reduce an order of convergence or even result in a nonconsistency of the method.

4 Error Analysis

While solving partial differential equations numerically, one has to estimate the error of applied numerical method. The required estimates are provided by the following theorem (see [3]).

Theorem 2. *Let* $\tau \in \left(0, \frac{1}{\rho+\|\mathbb{A}\|_\infty M}\right)$, *where* $\rho = \max_i r_i$, $M = \max_i M_i$ *and* $\|\mathbb{A}\|_\infty = \max_i \sum_{j=1}^N a_{ij}$. *Further, let* $u_i, u_i' \in L^\infty(0,T,H^{s+1}(\Omega))$, $u_i'' \in L^\infty(0,T,L^2(\Omega))$ *and* $g_i \in L^\infty(0,T,H^{s+1}(\Gamma_N))$ *for each* $i = 1,2,\ldots,N$, *and any* $s \geq 0$. *Then there exist constants* $C_L, C_H > 0$ *independent from* τ, h *and* $\beta = \max_i D_i$, *such that*

$$\|\boldsymbol{e}_h\|^2_{\mathscr{S},\infty} = \max_{0\leq k\leq p} \|\boldsymbol{e}_h^k\|^2_{\mathscr{S}} = \max_{\substack{0\leq k\leq p\\ 0\leq i\leq N}} \|e_{hi}^k\|_2^2 \leq C_L\left(\tau^2 + \beta h^{2s} + h^{2(s+1)} + \beta^2 h^{2(s+1)}\right),$$

$$|\boldsymbol{e}_h|^2_{\mathscr{S},2} = \max_{0\leq i\leq N}\left(\tau D_i \sum_{k=0}^{p} |e_{hi}^k|_1^2\right) \leq C_H\left(\tau^2 + \beta h^{2s} + h^{2(s+1)} + \beta^2 h^{2(s+1)}\right).$$

The Theorem rests upon division of the error into two parts. The estimate of the discretization error comes from the approximative properties of the finite element spaces. The error of the numerical method is estimated using Young's inequality and boundedness of $\boldsymbol{u}_h$ and time derivatives of $\boldsymbol{u}$.

5 Numerical Results

5.1 Example 1 - Neumann Boundary Conditions

In $\Omega = [0,1]^2$ we solve a problem (12)-(14) with $N = 3$, $\Gamma_D = \emptyset$ and coefficients $D_i = 10^{-4}$ and $r_i = 1$, for $i = 1,2$ a 3. In addition $a_{11} = a_{22} = a_{33} = 1$, $a_{12} = a_{23} = a_{31} = 2$ and $a_{21} = a_{13} = a_{32} = 7$. Initial conditions are chosen in such a way that $u_{0i} = \chi_{\mathscr{M}_i}$, for $i = 1,2,3$, where

$$\mathscr{M}_1 = \{(x,y) \in \Omega |\ (x-0.23)^2 + (y-0.2)^2 \leq (1/4)^2\}, \quad (18)$$

$$\mathscr{M}_2 = \{(x,y) \in \Omega |\ (x-0.67)^2 + (y-0.75)^2 \leq (1/5)^2\}, \quad (19)$$

$$\mathscr{M}_3 = \{(x,y) \in \Omega |\ (x-0.75)^2 + (y-0.5)^2 \leq (3/10)^2\}. \quad (20)$$

The invariant region of this problem is a block $[0,M_1] \times [0,M_2] \times [0,M_3]$, where $M_i = \max\{\|u_{0i}\|_\infty, 1\} = 1$ for $i = 1,2,3$. Further, we choose $T = 100$ with $p = 200$ and thus $\tau_k = 0.5$. Neumann boundary conditions are homogeneous, i.e. $g_i^k = 0$, for $i = 1,2,3$ and $k = 0,1,\ldots,p$. The interpolation of the initial condition and a solution computed at $T = 100$ are depicted on Figures 1 and 2.

5.2 Example 2 - Dirichlet Boundary Conditions

In $\Omega = [0,1]^2$ we solve a problem (12)-(14) with $N = 3$, $\Gamma_N = \emptyset$ and coefficients $D_i = 10^{-4}$ and $r_i = 1$, for $i = 1,2,3$. In addition $a_{11} = a_{22} = a_{33} = 1$, $a_{12} = a_{23} = a_{31} = 2$ and $a_{21} = a_{13} = a_{32} = 7$. Initial conditions are chosen in such a way that $u_{0i} = \chi_{\mathscr{M}_i}$, for $i = 1,2,3$, where

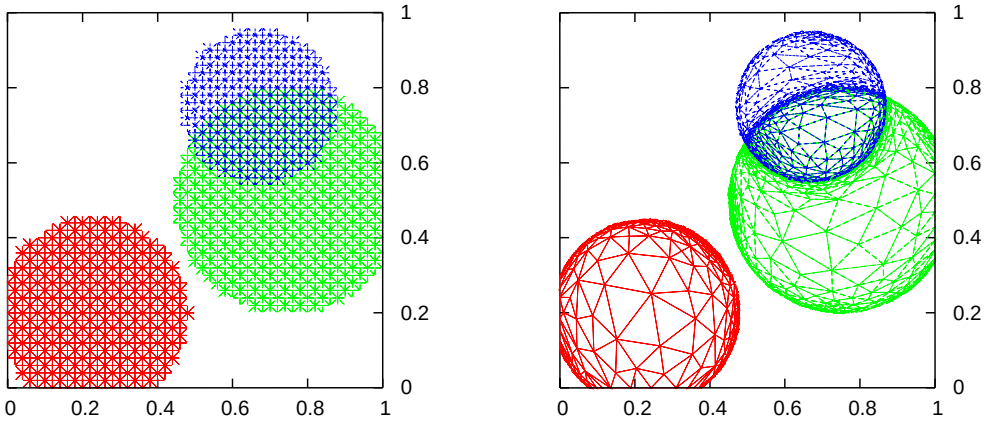
$$\mathscr{M}_1 = \{(x,y) \in \Omega |\ 9(x-0.4)^2 + 25(y-0.65)^2 \leq 1\}, \quad (21)$$

$$\mathscr{M}_2 = \{(x,y) \in \Omega |\ (x-0.67)^2 + (y-0.65)^2 \leq (1/5)^2\}, \quad (22)$$

$$\mathscr{M}_3 = \{(x,y) \in \Omega |\ (x-0.6)^2 + (y-0.4)^2 \leq (3/10)^2\}. \quad (23)$$

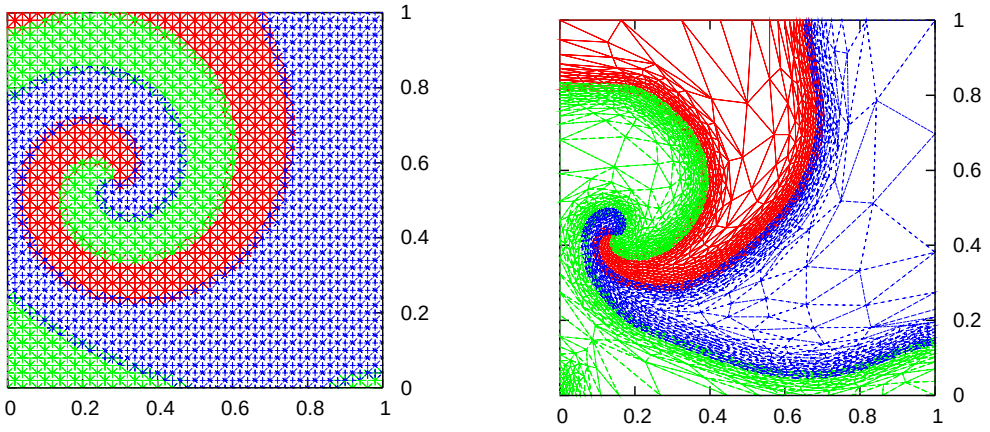
We choose $T = 50$ with $p = 100$ and thus $\tau_k = 0.5$. Dirichlet boundary conditions are $u_{Di}^k(x,y) = y$, for all $i = 1,2,3$ and $k = 0,1,\ldots,p$.

Since $\|u_{Di}^k\|_\infty = 1$, the invariant region of this problem is the same as the previous one. The interpolation of the initial condition and a solution computed at $T = 50$ are depicted on Figures 3 and 4.



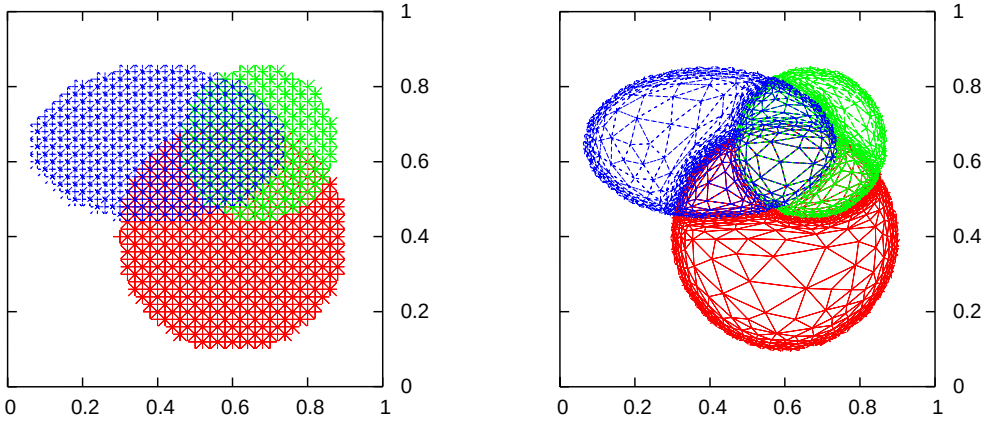
Source: Own

Fig. 1. *The interpolation of the initial condition of Example 1 on an isotropic (left, 5000 elements, 2601 nodes) and adapted mesh (right, 2724 elements, 1301 nodes)*



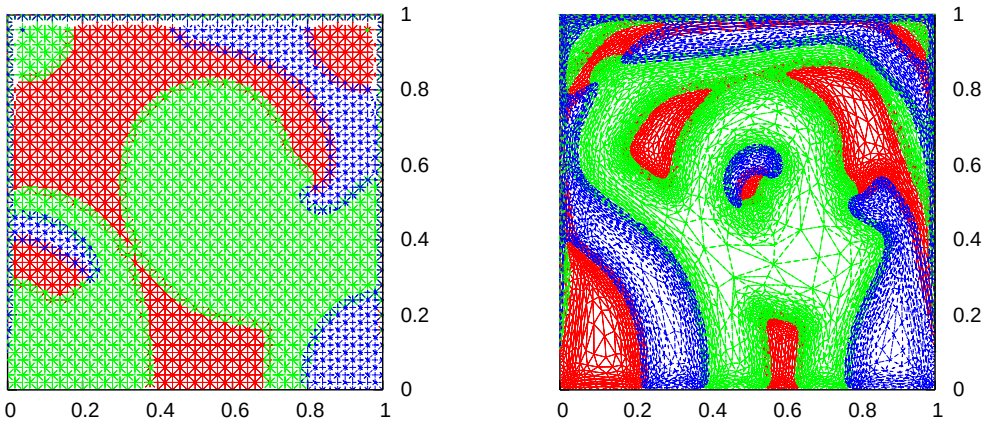
Source: Own

Fig. 2. *The solution of Example 1 computed at* $T = 100$ *on an isotropic (left, 5000 elements, 2601 nodes) and adapted mesh (right, 3694 elements, 1899 nodes)*



Source: Own

Fig. 3. *The interpolation of the initial condition of Example 2 on an isotropic (left, 5000 elements, 2601 nodes) and adapted mesh (right, 1760 elements, 3501 nodes)*



Source: Own

Fig. 4. *The solution of Example 2 computed at $T = 50$ on an isotropic (left, 5000 elements, 2601 nodes) and adapted mesh (right, 5021 elements, 9830 nodes)*

Conclusion

We have successfully derived and tested a numerical method for solving reaction-diffusion equations. We have also proven that for sufficiently smooth and suitable data the solution of the equations is continuous and belongs to the space $\mathscr{C}((0,T),L^2(\Omega))$. For numerical solution we have used the Galerkin finite elements method and derived a priori error estimates on isotropic meshes. We have used adaptively refined meshes as well, however, the error estimates are no longer valid on them. In order to obtain error estimates on adapted meshes a better projection should be used.

Acknowledgments

This work was supported by the European Social Fund project for constitution and improvement of a team for demanding technical computations on parallel computers at Technical University of Liberec, reg. num. CZ.1.07/2.3.00/09.0155, the research programme num. MSM6840770010, by the research center of Ministry of Education, Youth and Sports num. LC06052 and by the Grant Agency of the Czech Republic, grant No. P201/11/1304.

Literature

[1] CIARLET, P. G.: *The finite element method for elliptic problems*. North-Holland Publishing Company, Amsterdam-New York-Oxford, 1978.

[2] DOLEJŠÍ, V.: ANGENER, version 3.1 - Fortran source package able to generate triangulations [online]. Available from WWW: <http://www.karlin.mff.cuni.cz/∼dolejsi/angen/angen3.1.htm>

[3] DOLEJŠÍ, V.; FEISTAUER, M.: Analysis of semiimplicit DGFEM for nonlinear convection-diffusion problems. *Preprint Series of the School of Mathematics*, Charles University Prague, No. MATH-KNM-2004/1, 2004.

[4] ESLEROVÁ, L.: *Numerická analýza dynamiky reakčně-difuzních rovnic*. Master's Thesis, Department of mathematics FJFI ČVUT, 1996.

[5] HOLODNIOK, M. et al.: *Metody analýzy nelineárních dynamických modelů*. Academia, Praha, 1986.

[6] LAMAČ, J.: *Numerické metody pro řešení reakčně-difuzních rovnic*. Rešeršní práce, Department of mathematics FJFI ČVUT, Praha, 2009.

[7] LAMAČ, J.: *Numerické metody pro řešení reakčně-difuzních rovnic*. Výzkumný úkol, Department of mathematics FJFI ČVUT, Praha, 2010.

[8] LAMAČ, J.: *Numerické metody pro řešení reakčně-difuzních rovnic*. Master's Thesis, Department of mathematics FJFI ČVUT, Praha, 2012.

[9] REKTORYS, K.: *The method of discretization in time and partial differential equations*. SNTL-Publishers, Praha, 1985.

[10] SMOLLER, J.: *Shock Waves and Reaction-Diffusion Equations*. Springer Verlag, New York - Heidelberg - Berlin, 1983.

Ing. Mgr. Jan Lamač

Numerické řešení reakčně-difuzních rovnic

Obsahem předložené práce je matematická analýza a numerické řešení soustavy nelineárních nestacionárních difuzně-reakčních rovnic. Nejprve je s využitím invariantního regionu proveden důkaz existence a jednoznačnosti řešení a spojité závislosti na datech úlohy. Po časové diskretizaci problému je aplikována Galerkinova metoda konečných prvků a odvozeny apriorní odhady chyby numerické metody. Rovněž je diskutována vhodná adaptace použitých triangulací. Na závěr je celá metoda implementována a otestována na různých příkladech.

Numerische Lösung reaktionsdiffuser Gleichungen

Der Inhalt der vorgelegten Arbeit besteht in der mathematischen Analyse und der numerischen Lösung des Systems nichtlinearer nicht stationärer reaktionsdiffuser Gleichungen. Zunächst wird unter Verwendung einer invarianten Region der Beweis der Existenz und der Eindeutigkeit der Lösung und der kontinuierlichen Abhängigkeit von den Daten der Aufgabe erbracht. Nach einer zeitlichen Diskretisierung des Problems wird die Galerkin-Methode der finiten Elemente angewandt und es werden A-priori-Schätzungen des Fehlers der numerischen Methode abgeleitet. Ebenfalls wird eine geeignete Adaption der gebrauchten Triangulationen diskutiert. Zum Abschluss wird die gesamte Methode implementiert und an verschiedenen Beispielen getestet.

Numeryczne rozwiązywanie rŕwnań reakcyjno-dyfuzyjnych

Przedmiotem niniejszego opracowania jest analiza matematyczna oraz numeryczne rozwiązanie układu nieliniowych równań niestacjonarnych dyfuzyjno-reakcyjnych. W pierwszej kolejności przy wykorzystaniu niezmiennego regionu przeprowadzono dowód na istnienie i jednoznaczność rozwiązania oraz ciągłą zależność od danych zadania. Po czasowej dyskretyzacji problemu zastosowano metodę Galerkina elementów skończonych oraz oszacowano a priori błąd metody numerycznej. Omówiono także odpowiednią adaptację zastosowanych triangulacji. W zakończeniu przedstawiono zastosowanie całej metody i sprawdzono ją na różnych przykładach.

NUMERICAL EVALUATION OF RHEOLOGICAL EXPERIMENT

Petr Salač
***Ivo Matoušek**

Technical University of Liberec
Faculty of Science, Humanities and Education
Studentská 2, 461 17, Liberec, Czech Republic
petr.salac@tul.cz

*Technical University of Liberec
Faculty of Mechanical Engineering
Studentská 2, 461 17, Liberec, Czech Republic
ivo.matousek@tul.cz

Abstract

A viscoelastic simply supported rotationally symmetric body, fixed on a base, is considered. The body is loaded by a flat plunger, which moves in the direction of the *z* axis by a constant velocity *v*. In this work the reaction force is computed. This allows us to compare numerical results with data from rheological experiment (see [6], [7]). The variational formulation of the problem is derived and transformed to cylindrical coordinates. Some results of numerical calculations are presented.

Keywords: Viscoelasticity; axisymmetric hyperbolic problems; dimensional reduction.

Introduction

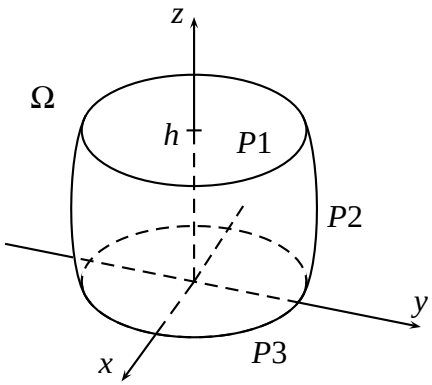
Mathematical modeling of technological processes has been already regarded as a powerful tool for an optimization of technological processes also in glass industry. The fundamental issue of virtual modelling of silica glass forming is an accuracy of numerical outputs. Critical factor is not only definition of boundary conditions, mainly thermal ones, but also specification of material properties. Number of methods for an identification of rheological properties exists [1]. The disadvantage of the most of published models is independent description of the both stages - the stage with the dominant influence of an elastic component of deformation and that one with a dominant viscous flow [8]. One of the most effective methods is isothermal compression method which is based on the evaluation of the force response on compression loading of cylindrical samples [4], [6], [7].

The advantage of this method is its relatively simplicity and possibility to evaluate both elastic and viscous properties of glass melt simultaneously during one experiment. However the critical issue of this method is an accuracy of evaluation of experimental outputs. Several methods were suggested.

In this contribution we introduce the variational formulation of the rheological experiment model, which includes viscoelastic deformations. This problem will be used as a state problem in formulations of various identification problems for various model parameters, that are planed as next step of our research. Nevertheless recent numerical results are roughly in conformity with results of experiments (see [6]).

1 Formulation of the Problem

Forming of glass is quite complicated process which contains both elastic and plastic responses to strain from stress. This is the main reason of using a viscoelastic model to describe relationship between stress and strain.



Source: Own

Fig. 1. *Scheme of glass sample*

We consider a viscoelastic isotropic homogeneous cylindrical body Ω symmetrical according to z axis with bases formed by two parallel circles with radii R, and altitude h. We consider the body which is fixed on both bases and which is free on its surrounding surface. We denote P_1 upper, resp. P_3 bottom, base and P_2 surrounding surface of the body. The body is deformed by flat plunger moving by a constant velocity v in the direction of the z axis placed on the upper base. To represent changes of the shape of the body we define a deformation tensor by the formula

$$\varepsilon_{ij}(\boldsymbol{x},t) = \frac{1}{2}\left(\frac{\partial u_i(\boldsymbol{x},t)}{\partial x_j} + \frac{\partial u_j(\boldsymbol{x},t)}{\partial x_i}\right) . \tag{1}$$

The stress is represented by symmetrical stress tensor σ. We consider stress strain relation given by viscoelastic generalized Hook's law in the form

$$\sigma_{ij}(\boldsymbol{x},t) = \delta_{ij}\int_{-\infty}^{t}\lambda(t-\tau)\frac{\partial\varepsilon_{kk}(\boldsymbol{x},\tau)}{\partial\tau}\,d\tau + 2\int_{-\infty}^{t}\mu(t-\tau)\frac{\partial\varepsilon_{ij}(\boldsymbol{x},\tau)}{\partial\tau}\,d\tau\,, \tag{2}$$

where $\lambda(t)$ and $\mu(t)$ denote relaxation functions describing glass properties in pressing, resp. in shear and δ_{ij} is the Kroneker symbol.
The balance of a linear momentum for the dynamic problem has the form

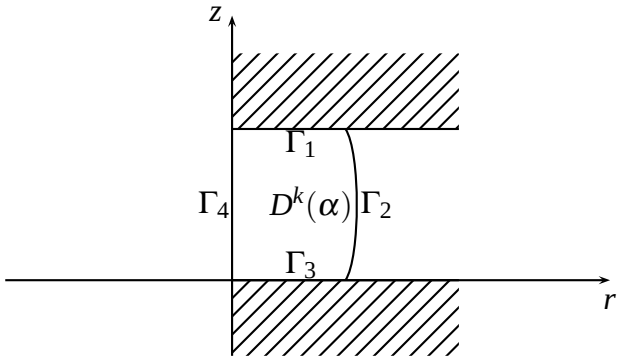
$$\frac{\partial\sigma_{ij}(\boldsymbol{x},t)}{\partial x_j} + \mathscr{F}_i(\boldsymbol{x},t) = \rho\frac{\partial^2 u_i(\boldsymbol{x},t)}{\partial t^2} \quad i=1,2,3\,, \tag{3}$$

where $\boldsymbol{\mathscr{F}}(\boldsymbol{x},t)$ represents the body forces and ρ the density of glass.
We use the time discretization method:
Let the time interval $[0,T]$ be divided into the subintervals $[t_{k-1},t_k]$, for $k=1,2,\ldots,p$, then (3) has at each time level the form

$$\frac{\partial\sigma_{ij}^k(\boldsymbol{x})}{\partial x_j} + F_i^k(\boldsymbol{x}) = \rho\frac{z_i^k(\boldsymbol{x}) - 2z_i^{k-1}(\boldsymbol{x}) + z_i^{k-2}(\boldsymbol{x})}{h^2} \qquad i=1,2,3, \quad k=2,\ldots,p, \tag{4}$$

where $z_i^k(\boldsymbol{x}) = u(\boldsymbol{x},t_k)$, $\sigma_{ij}^k(\boldsymbol{x}) = \sigma_{ij}(\boldsymbol{x},t_k)$, $F_i^k(\boldsymbol{x}) = \mathscr{F}_i(\boldsymbol{x},t_k)$ and $h = \frac{T}{p}$.
According to the fact that the body, support and load have rotational symmetry we transform

the problem to cylindrical coordinates and apply dimensional reduction of an angle to get two-dimensional problem.
We are going to solve the problem in the region $D^k(\alpha)$ dependent on time the level t_k bounded by the axis r (part Γ_3), the axis z (part Γ_4), the straight line $z = h - tv$ (part Γ_1) and the free boundary described by the function $\alpha(z)$ of variable z (part Γ_2).



Source: Own

Fig. 2. *Domain after the dimensional reduction*

According to the fact that the problem has rotational symmetry we assume that a displacement vector component in the direction ϑ is zero ($z^k_\vartheta(\boldsymbol{x}) = 0$), similarly $\frac{\partial z^k_r}{\partial \vartheta} = 0$, and $\frac{\partial z^k_z}{\partial \vartheta} = 0$. We denote the physical components of the displacement vector by two functions, e.g.

$$\begin{aligned} z^k_r(\boldsymbol{x}) &= u^k, \\ z^k_z(\boldsymbol{x}) &= w^k, \\ (z^k_\vartheta(\boldsymbol{x}) &= 0). \end{aligned}$$

The relationship between the displacement vector and the strain tensor is in the form

$$\varepsilon^k_{rr} = \frac{\partial u^k}{\partial r}\,, \tag{5}$$

$$\varepsilon^k_{zz} = \frac{\partial w^k}{\partial z}\,, \tag{6}$$

$$\varepsilon^k_{\vartheta\vartheta} = \frac{u^k}{r}\,, \tag{7}$$

$$\varepsilon^k_{rz} = \frac{1}{2}\left(\frac{\partial u^k}{\partial z} + \frac{\partial w^k}{\partial r}\right), \tag{8}$$

$$(\varepsilon^k_{r\vartheta} = 0,\ \ \varepsilon^k_{z\vartheta} = 0)\,. \tag{9}$$

The components of the stress tensor $\boldsymbol{\sigma}$ have the form

$$\sigma^k_{rr} = \frac{1}{h}\int_{-\infty}^{t} \lambda(t-\tau)(e^k(\boldsymbol{x}) - e^{k-1}(\boldsymbol{x})) + 2\mu(t-\tau)(\varepsilon^k_{rr}(\boldsymbol{x}) - \varepsilon^{k-1}_{rr}(\boldsymbol{x}))\,d\tau\,, \tag{10}$$

$$\sigma^k_{zz} = \frac{1}{h}\int_{-\infty}^{t} \lambda(t-\tau)(e^k(\boldsymbol{x}) - e^{k-1}(\boldsymbol{x})) + 2\mu(t-\tau)(\varepsilon^k_{zz}(\boldsymbol{x}) - \varepsilon^{k-1}_{zz}(\boldsymbol{x}))\,d\tau\,, \tag{11}$$

$$\sigma^k_{\vartheta\vartheta} = \frac{1}{h}\int_{-\infty}^{t} \lambda(t-\tau)(e^k(\boldsymbol{x}) - e^{k-1}(\boldsymbol{x})) + 2\mu(t-\tau)(\varepsilon^k_{\vartheta\vartheta}(\boldsymbol{x}) - \varepsilon^{k-1}_{\vartheta\vartheta}(\boldsymbol{x}))\,d\tau\,, \tag{12}$$

$$\sigma_{rz}^k = \frac{2}{h}\int_{-\infty}^{t} \mu(t-\tau)(\varepsilon_{rz}^k(\boldsymbol{x}) - \varepsilon_{rz}^{k-1}(\boldsymbol{x}))\, d\tau\,, \tag{13}$$

$$(\sigma_{r\vartheta}^k = 0, \quad \sigma_{z\vartheta}^k = 0)\,, \tag{14}$$

where

$$e = \varepsilon_{rr} + \varepsilon_{zz} + \varepsilon_{\vartheta\vartheta}\,. \tag{15}$$

The bilinear form representing mechanical work of inner forces has the form

$$\begin{aligned} A(\boldsymbol{u},\boldsymbol{\varphi}) \;=\; & -\frac{1}{h}\int_{D^k(\alpha)}\int_{-\infty}^{t} \Bigg((\lambda(t-\tau)+2\mu(t-\tau)) \left[\frac{\partial u^k(\boldsymbol{x})}{\partial r}\frac{\partial \varphi_1(\boldsymbol{x})}{\partial r} r + \right. \\ & \left. + \frac{\partial w^k(\boldsymbol{x})}{\partial z}\frac{\partial \varphi_2(\boldsymbol{x})}{\partial z} r \;+\; u^k(\boldsymbol{x})\varphi_1(\boldsymbol{x})\frac{1}{r} \right] \;+ \\ & + \mu(t-\tau) \left[\frac{\partial w^k(\boldsymbol{x})}{\partial r}\frac{\partial \varphi_2(\boldsymbol{x})}{\partial r} r \;+ \frac{\partial w^k(\boldsymbol{x})}{\partial r}\frac{\partial \varphi_1(\boldsymbol{x})}{\partial z} r + \right. \\ & \left. + \;\frac{\partial u^k(\boldsymbol{x})}{\partial z}\frac{\partial \varphi_2(\boldsymbol{x})}{\partial r} r \;+\; \frac{\partial u^k(\boldsymbol{x})}{\partial z}\frac{\partial \varphi_1(\boldsymbol{x})}{\partial z} r \right] \Bigg)\, d\tau\, d\boldsymbol{x} \;+ \\ & + \frac{\rho}{h^2}\int_{D^k(\alpha)} \Big(u^k(\boldsymbol{x})\varphi_1(\boldsymbol{x}) r \;+ w^k(\boldsymbol{x})\varphi_2(\boldsymbol{x}) r \Big)\, d\boldsymbol{x}\,. \end{aligned} \tag{16}$$

Linear functional representing mechanical work of outward forces has the form

$$\begin{aligned} \langle \boldsymbol{F}(t),\boldsymbol{\varphi}\rangle \;=\; & \int_{D^k(\alpha)} \Big([\, F_1^k(\boldsymbol{x})\varphi_1(\boldsymbol{x}) r + F_2^k(\boldsymbol{x})\varphi_2(\boldsymbol{x}) r\,] - \\ & -\frac{1}{h}\int_{-\infty}^{t} \Bigg((\lambda(t-\tau)+2\mu(t-\tau)) \left[\frac{\partial u^{k-1}(\boldsymbol{x})}{\partial r}\frac{\partial \varphi_1(\boldsymbol{x})}{\partial r} r + \right. \\ & \left. + \frac{\partial w^{k-1}(\boldsymbol{x})}{\partial z}\frac{\partial \varphi_2(\boldsymbol{x})}{\partial z} r \;+\; u^{k-1}(\boldsymbol{x})\varphi_1(\boldsymbol{x})\frac{1}{r} \right] \;+ \\ & + \mu(t-\tau) \left[\frac{\partial w^{k-1}(\boldsymbol{x})}{\partial r}\frac{\partial \varphi_2(\boldsymbol{x})}{\partial r} r \;+ \frac{\partial w^{k-1}(\boldsymbol{x})}{\partial r}\frac{\partial \varphi_1(\boldsymbol{x})}{\partial z} r + \right. \\ & \left. + \;\frac{\partial u^{k-1}(\boldsymbol{x})}{\partial z}\frac{\partial \varphi_2(\boldsymbol{x})}{\partial r} r \;+\; \frac{\partial u^{k-1}(\boldsymbol{x})}{\partial z}\frac{\partial \varphi_1(\boldsymbol{x})}{\partial z} r \right] \Bigg)\, d\tau \;- \\ & - \frac{\rho}{h^2} \Big[\Big(u^{k-2}(\boldsymbol{x})\varphi_1(\boldsymbol{x}) r \;+ w^{k-2}(\boldsymbol{x})\varphi_2(\boldsymbol{x}) r \Big) \;- \\ & \quad - 2\left(u^{k-1}(\boldsymbol{x})\varphi_1(\boldsymbol{x}) r \;+ w^{k-1}(\boldsymbol{x})\varphi_2(\boldsymbol{x}) r \right) \Big] \Big)\, d\boldsymbol{x}\,. \end{aligned} \tag{17}$$

Boundary conditions:
The flat plunger moves in the direction of the z axis by the constant velocity v acting on part of boundary Γ_1, i.e.

$$\left.\begin{array}{l} u = 0 \\ w = v(t-t_k) \end{array}\right\} \;\text{ on } \quad \Gamma_1\,. \tag{18}$$

The part of boundary Γ_2 represents the so called free boundary which is deformed by the influence of the inner forces and is not able to catch any force (tangent or normal)

$$\left.\begin{array}{l} \sigma_{rr} = 0 \\ \sigma_{zz} = 0 \end{array}\right\} \;\text{ on } \quad \Gamma_2\,. \tag{19}$$

The part of boundary Γ_3 is fixed, i.e.

$$\left.\begin{array}{l} u=0 \\ w=0 \end{array}\right\} \text{ on } \quad \Gamma_3 \,. \tag{20}$$

The part of boundary Γ_4 is formed by the axis of symmetry and has properties of contact with solid support (without friction), i.e.

$$\left.\begin{array}{l} u=0 \\ \frac{\partial w}{\partial r}=0 \end{array}\right\} \text{ on } \quad \Gamma_4 \,. \tag{21}$$

We define the space $W^{1,2,r}(D^k(\alpha))$ with the norm

$$\|u\|_{1,2,r} \;=\; \left(\int_{D^k(\alpha)} \left[\left(\frac{\partial u}{\partial r} \right)^2 + \left(\frac{\partial u}{\partial z} \right)^2 + u^2 \right] r \, d\boldsymbol{x} \right)^{\frac{1}{2}} . \tag{22}$$

We define the space of functions with the finite energy as the weighted Sobolev space $\mathscr{H}(D^k(\alpha))$

$$\mathscr{H}(D^k(\alpha)) = \{ \hat{\boldsymbol{u}} \equiv (u,w) \in W^{1,2,r}(D^k(\alpha)) \times W^{1,2,r}(D^k(\alpha)) \} \,. \tag{23}$$

We denote

$$\mathscr{V}_1 = \{ u \in C^\infty(D^k(\alpha)) \mid \operatorname{supp} u \cap \Gamma_4 = \emptyset,\; u=0 \;\text{ on }\; \Gamma_1 \cup \Gamma_3 \} \,. \tag{24}$$

Let V_1 be the closure of the set $\mathscr{V}_1$ in the space $W^{1,2,r}(D^k(\alpha))$.
Further we denote

$$\mathscr{V}_2 = \{ u \in C^\infty(D^k(\alpha)) \mid u=0 \;\text{ on }\; \Gamma_1 \cup \Gamma_3 \} \,. \tag{25}$$

Let V_2 be the closure of the set $\mathscr{V}_2$ in the space $W^{1,2,r}(D^k(\alpha))$.
We denote by

$$\boldsymbol{H} = \{ \hat{\boldsymbol{u}} \equiv (u,w) \in V_1 \times V_2 \} \tag{26}$$

the space of test functions (i.e. such functions with finite energy which satisfy stable boundary conditions).
We use the principle of virtual displacement to get a variational formulation of the problem:
Let $\hat{u}_0 \in \mathscr{H}(D^k(\alpha))$ be given, which specifies the displacement on the boundary Γ_1 by its traces. We are looking for $\hat{u} \in \mathscr{H}(D^k(\alpha))$ such that

$$\hat{\boldsymbol{u}} \;-\; \hat{\boldsymbol{u}}_0 \;\in\; \boldsymbol{H} \,, \tag{27}$$

$$A(\hat{\boldsymbol{u}}, \hat{\boldsymbol{\varphi}}) \;=\; \langle \boldsymbol{F}(t), \hat{\boldsymbol{\varphi}} \rangle \qquad \forall \hat{\boldsymbol{\varphi}} \in \boldsymbol{H},\;\; k = 2,3,\ldots,p \,. \tag{28}$$

Theorem. The problem (27) - (28) has the unique solution.

Proof: The proof based on the Lax-Milgram theorem is too long and technical to be published.

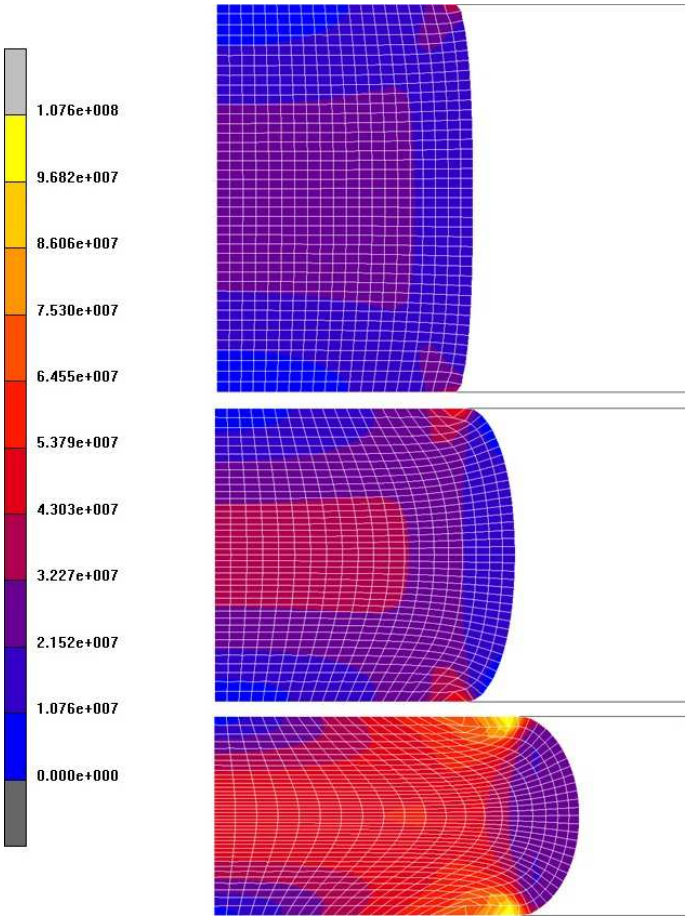
2 Numerical Experiment

The numerical model, describing the course of experimental measurements of rheological properties of melt glass, was created.

The principle of experiment is based on the evaluation of the force (viscoelastic) response of isothermal cylindrical molten glass sample, which is compressed at a constant velocity. Numerical simulation was realized in the commercial FEM (Finite Elements Method) code MSC MARC.

The initial sample sizes were 20,3 mm (diameter) - 18,45 mm (height), the velocities of compression were taken from the range 0,5 - 40 mm/s. The Maxwell model was used for description of material behavior of FLOAT melted glass.
Viscosity of the shaped glass was defined according to the experiment, i.e. $\eta = 10^{7,52}$ [Pa.s]. The modulus of elasticity was selected from the range $E_1 = 2,5.10^8$ $2,5.10^9$ [Pa], molten glass was assumed to be incompressible substance, i.e. The Poisson constant $\nu = 0,5$.

Sticking conditions were presumed between glass and metal punch contact surfaces.



Source: Own

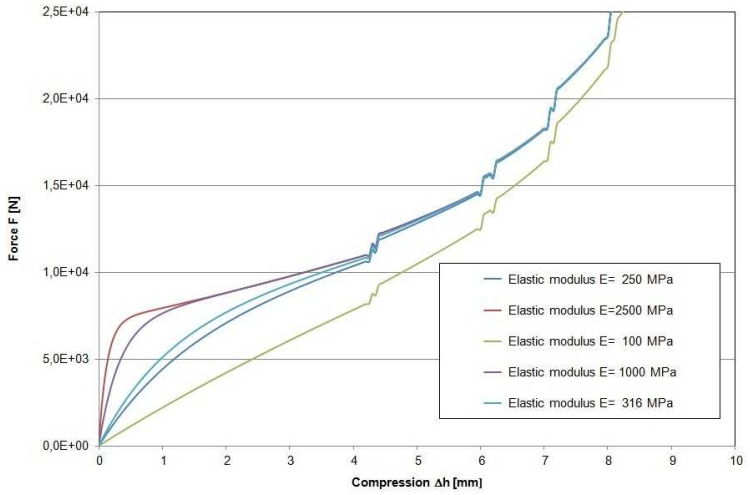
Fig. 3. *Distribution of the stress fields in the form of equivalent Cauchy stress for compression 2, 6 and 10 mm*

The course of distribution of the stress fields in the form of the equivalent Cauchy stress for 3 different stages (compression 2, 6 and 10 mm) are presented in Fig. 3 (for velocity $v = 4$ [mm/s]).

The courses of the force response for different elastic moduli are shown in Fig. 4. From the figure it results that the elastic modulus influences only the first stage of the experiment, second one is only controlled by viscous flow.

Conclusion

In the contribution the model for evaluation of description of viscoelastic force response to compression loading was suggested. Integration of the viscoelastic model of the Maxwell type



Source: Own

Fig. 4. *Course of computed load forces*

to the mathematical model allowed fair description of the force response according to character of the realized experiments [6]. The shape course of deformed glass sample in the experiment was visibly similar to the computed one. Development of the measured load force showed the similar tendency but values became more different from the computed ones during the experiment.

Acknowledgments

The paper was supported by the ESF Project No. CZ.1.07/2.3.00/09.0155 “Constitution and improvement of a team for the demanding technical computations on parallel computers at TU of Liberec”.

Literature

[1] DUFFRENE, L.; GY, R.; MASNIK, J. E.: Temperature dependence of the high-frequency viscoelastic behaviour of a soda-lime-silica glass. *J. Amer. Ceram. Soc.* 81, 1998, no. 5, pp. 1278-1284.

[2] GRÄFE, W.: *Time-dependent Mechanical Properties of Solids.* Trans Tech Publications Inc, 2008, ISBN 978-0-87849-476-7, ISSN 1422-3597.

[3] NEČAS, J., HLAVÁČEK, I.: *Úvod do matematické teorie pružných a pružně plastických těles.* SNTL, Praha, 1983.

[4] HESSEKEMPER, H.; BRÜCKNER R.: Load-dependent flow behaviour of silicate glass melts. *Glastechnische Berichte*, 1988, vol. 61, no. 11, pp. 312-322.

[5] CHRISTENSEN, R. M.: *Theory of viscoelasticity.* Academic Press, NewYork, 2003, ISBN 0-486-42880-X.

[6] MATOUŠEK, I.; VÍTOVÁ, M.: Rheological behaviour of silica glass melts. *Sklář a keramik*, vol. 59, 2009, č. 10-12, pp. 217-223. ISSN 0037-637X.

[7] MATOUŠEK, I.; VÍTOVÁ, M.: Rheological Response of Glass Melts. *Ceramics – Silikáty*, vol. 53, 2009, č. 1, pp. 59-62. ISSN 0862-5468.

[8] YU-CHUNG, T.; CHINGHUA H.; JUNG-CHUNG H.: Glass material model for the forming stage of the glass molding process. *J. Mater. Proc. Technol.*, 2008, vol. 201, no. 1-3, pp. 751-754.

NUMERICKÉ HODNOCENÍ REOLOGICKÉHO EXPERIMENTU

Uvažujeme viskoelastické, prostě podepřené, rotačně symetrické těleso pevně spojené s podkladem. Těleso je zatěžováno plochou lisovací čelistí, která se pohybuje ve směru osy *z* konstantní rychlostí *v*. V předloženém příspěvku je počítána silová odezva. To nám umožní porovnávat numerické výsledky s reálnými daty z reologických experimentů. Je odvozena variační formulace úlohy a transformována do válcových souřadnic. Dále jsou prezentovány numerické výsledky.

NUMERISCHE BEWERTUNG EINES RHEOLOGISCHEN EXPERIMENTS

Wir betrachten einen viskoelastischen, einfach unterstützten, drehsymmetrischen Körper, der fest mit dem Untergrund verbunden ist. Der Körper wird mit der Fläche eines Presskiefers beschwert, die sich in Richtung der Achse aus der konstanten Geschwindigkeit v bewegt. Im vorliegenden Beitrag wird das Kraftecho berechnet. Dies ermöglicht uns einen Vergleich der numerischen Ergebnisse mit den realen Daten aus den rheologischen Experimenten. Daraus wird eine Variantenformulierung der Aufgabe abgeleitet und in Walzenkoordinaten transformiert. Weiter werden numerische Ergebnisse präsentiert.

NUMERYCZNA OCENA EKSPERYMENTU REOLOGICZNEGO

W artykule rozważane jest wiskoelastyczne, prosto podparty, rotacyjnie symetryczne ciało na stałe połączony z podłożem. Na ciało oddziałuje powierzchnia szczęk prasujących, która porusza się w kierunku osi *z* stałą prędkością *v*. W opracowaniu obliczana jest reakcja siłowa. Umożliwia to porównanie wyników numerycznych z realnymi danymi z eksperymentów reologicznych. Określona jest zmienna formuła zadania, która jest transformowana do współrzędnych cylindrycznych. Następnie zaprezentowano wyniki numeryczne.

MULTIPLICATION BY WAVELET MATRIX – EFFICIENT IMPLEMENTATION

Martina Šimůnková

Technical University of Liberec
Faculty of Science, Humanities and Education
Studentská 2, 461 17, Liberec, Czech Republic
martina.simunkova@tul.cz

Abstract

Stiffness matrix of the Dirichlet problem $-(au')' = f$ with a homogeneous boundary value condition in a spline wavelet basis has $O(n \log n)$ non-zero elements [4]. We show that for a constant function a it is just $O(n)$ and moreover we show that it can be stored in $O(1)$ elements. This leads to a linear-time algorithm for multiplication by the wavelet matrix.

Keywords: Dirichlet problem; efficient implementation; Galerkin method; spline wavelets.

Introduction

Our aim is to solve the Dirichlet boundary value problem

$$\begin{aligned} -\Delta u &= f \quad \text{on} \quad \Omega = (0,1)^d \\ u &= 0 \quad \text{on} \quad \partial\Omega \end{aligned}$$

in the higher dimension d by the Galerkin method in a wavelet basis which is a tensor product of wavelet bases in one dimension. The Galerkin method leads to a system of linear algebraic equations with a matrix constructed from matrices (d_{ij}), (g_{ij}), where

$$\begin{aligned} d_{ij} &= \int_0^1 \varphi_i'(x)\varphi_j'(x)\,\mathrm{d}x, \\ g_{ij} &= \int_0^1 \varphi_i(x)\varphi_j(x)\,\mathrm{d}x \end{aligned} \tag{1}$$

and φ_i are basis functions. We need an efficient storage of matrices (d_{ij}), (g_{ij}) and as we solve the system by an iterative method, we also need an efficient implementation of multiplication of matrices by a vector.

In this paper we describe what a wavelet basis is, its construction and then we show that the stiffness matrix (d_{ij}) for the one-dimensional Poisson equation has $O(n)$ non-zero elements, and moreover it can be stored in $O(1)$ space. In the last section we show a numerical experiment on one-dimensional Poisson equation.

1 Spline Wavelet Basis

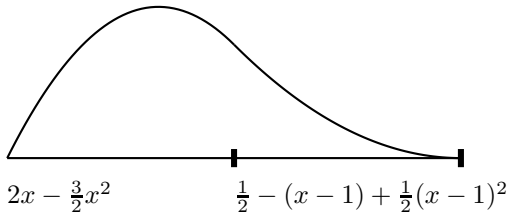
In this section we first describe the spaces V_n of scaling functions based on quadratic splines. Then we deal with the wavelet basis of V_n. First we formulate desired properties for wavelets and explain their consequences, then we show some constructions of wavelet bases.

1.1 Scaling Functions

Spline wavelet basis consists of scaling and wavelet functions. We use two following functions to construct them:

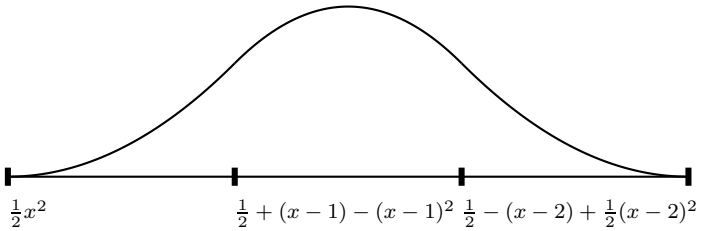
$$\varphi_{bd}(x) = \begin{cases} 2x - \frac{3}{2}x^2 & \text{for } x \in [0,1] \\ \frac{1}{2} - x + \frac{1}{2}x^2 & \text{for } x \in [1,2] \\ 0 & \text{otherwise} \end{cases}$$

$$\varphi_{in}(x) = \begin{cases} \frac{1}{2}x^2 & \text{for } x \in [0,1] \\ \frac{1}{2} + x - x^2 & \text{for } x \in [1,2] \\ \frac{1}{2} - x + \frac{1}{2}x^2 & \text{for } x \in [2,3] \\ 0 & \text{otherwise} \end{cases}$$



Source: Own

Fig. 1. *Boundary scaling function*



Source: Own

Fig. 2. *Inner scaling function*

Both are $\mathscr{C}^1$ piecewise quadratic functions.

Definition 1 *Let $\ell \in \mathbb{N}$. We call functions*

$$\varphi_{\ell,0}(x) = \varphi_{bd}(2^\ell x) \tag{2}$$

$$\varphi_{\ell,i}(x) = \varphi_{in}(2^\ell x - i + 1) \quad \textit{for } i = 1, \ldots, (2^\ell - 2) \tag{3}$$

$$\varphi_{\ell,2^\ell-1}(x) = \varphi_{bd}(2^\ell(1-x)) \tag{4}$$

scaling functions at the level ℓ.

Note that from (3) it follows that the values of integrals

$$\left.\begin{array}{rcl} d_{\ell,i,j} & = & \int_0^1 \varphi'_{\ell,i}\varphi'_{\ell,j}\,\mathrm{d}x \\ g_{\ell,i,j} & = & \int_0^1 \varphi_{\ell,i}\varphi_{\ell,j}\,\mathrm{d}x \end{array}\right\} \quad \text{for } i,j = 1,\ldots,(2^\ell - 2) \tag{5}$$

depend just on the difference $(i-j)$ and the value ℓ. Furthermore, for two different levels ℓ_1, ℓ_2 it holds

$$\begin{array}{rcl} d_{\ell_1,i,j} & = & 2^{\ell_1-\ell_2} d_{\ell_2,i,j} \\ g_{\ell_1,i,j} & = & 2^{\ell_2-\ell_1} g_{\ell_2,i,j} \end{array} \tag{6}$$

Similarly from (2), (4) it follows that

$$\left.\begin{array}{rcl} \int_0^1 \varphi'_{\ell_1,0}\varphi'_{\ell_1,i}\,\mathrm{d}x & = & 2^{\ell_1-\ell_2}\int_0^1 \varphi'_{\ell_2,0}\varphi'_{\ell_2,i}\,\mathrm{d}x \\ \int_0^1 \varphi'_{\ell_1,2^\ell-1}\varphi'_{\ell_1,i}\,\mathrm{d}x & = & 2^{\ell_1-\ell_2}\int_0^1 \varphi'_{\ell_2,2^\ell-1}\varphi'_{\ell_2,i}\,\mathrm{d}x \\ \int_0^1 \varphi_{\ell_1,0}\varphi_{\ell_1,i}\,\mathrm{d}x & = & 2^{\ell_2-\ell_1}\int_0^1 \varphi_{\ell_2,0}\varphi_{\ell_2,i}\,\mathrm{d}x \\ \int_0^1 \varphi_{\ell_1,2^\ell-1}\varphi_{\ell_1,i}\,\mathrm{d}x & = & 2^{\ell_2-\ell_1}\int_0^1 \varphi_{\ell_2,2^\ell-1}\varphi_{\ell_2,i}\,\mathrm{d}x \end{array}\right\} \quad \text{for } i = 1,\ldots,(2^\ell-2) \tag{7}$$

1.2 Wavelets

Let us denote

$$\mathscr{S}_\ell = \{\varphi_{\ell,i} : i = 0\ldots 2^\ell - 1\}$$

and

$$V_\ell = \operatorname{span}\mathscr{S}_\ell \qquad \text{for } \ell \in \mathbb{N}.$$

It holds that

$$V_\ell \subset V_{\ell+1} \qquad \text{for } \ell \in \mathbb{N}.$$

In the next section we describe some constructions of sets $\mathscr{W}_\ell \subset V_{\ell+1}$ which consist of functions $\psi_{\ell,0},\ldots\psi_{\ell,2^\ell-1}$. We require the following properties:

1. The set $\mathscr{S}_\ell \cup \mathscr{W}_\ell$ forms a basis of $V_{\ell+1}$.
2. Functions $\psi \in \mathscr{W}_\ell$ have vanishing moments of order 0, 1 and 2. It means that

$$\int_{\operatorname{supp}(\psi)} x^m \psi(x)\,\mathrm{d}x = 0 \qquad \text{for } m = 0,1,2.$$

We use property 1 iteratively to construct the basis of V_ℓ

$$\mathscr{S}_{l_0} \cup \bigcup_{i=l_0}^{l} \mathscr{W}_i \tag{8}$$

for an appropriate ℓ_0.

The property of vanishing moments has two very important consequences. The first one is more sparse approximation of the solution. The second one concerns matrices (1) – they are less sparse in the basis (8) than in the basis $\mathscr{S}_\ell$, but vanishing moments imply a lot of zero entries of matrices (1) even if supports of functions in the base overlap – more precisely

$$\int_{[0,1]} \psi_i\psi_j = \int_{[0,1]} \psi'_i\psi'_j = 0$$

whenever ψ_i is a quadratic function on the support of ψ_j. In the general case we can write ψ_j as a sum

$$\psi_j(x) = ax^2 + bx + c + \sum_k d_k (\max\{(x - x_k), 0\})^2.$$

We can then express integrals as sums

$$\begin{aligned} \int_{[0,1]} \psi_i \psi_j &= \sum_k d_k \int_{x_k}^1 \psi_i(x)(x - x_k)^2 \,\mathrm{d}x, \\ \int_{[0,1]} \psi_i' \psi_j' &= \sum_k 2 d_k \int_{x_k}^1 \psi_i'(x)(x - x_k) \,\mathrm{d}x. \end{aligned} \tag{9}$$

We use these formulas in cases when levels ℓ_i, ℓ_j of functions ψ_i, ψ_j respectively differ by at least two – then the sum consists of at most two terms.

1.3 Construction of Wavelets

We need wavelets satisfying properties from the previous section. Several such constructions are known: inner bi-orthogonal wavelets [3], boundary bi-orthogonal wavelets [1] and three types of short wavelets [2].

We show construction of one type of short wavelets:

$$\psi_{\ell,0} = -\tfrac{5}{2}\varphi_{\ell+1,0} + \tfrac{47}{12}\varphi_{\ell+1,1} - \tfrac{13}{4}\varphi_{\ell+1,2} + \varphi_{\ell+1,3}, \tag{10}$$

$$\psi_{\ell,i} = -\tfrac{1}{4}\varphi_{\ell+1,2i-1} + \tfrac{3}{4}\varphi_{\ell+1,2i} - \tfrac{3}{4}\varphi_{\ell+1,2i+1} + \tfrac{1}{4}\varphi_{\ell+1,2i+2} \quad \text{for } i = 1, \ldots, 2^\ell - 2, \tag{11}$$

$$\psi_{\ell,2^\ell-1} = -\varphi_{\ell+1,2^{\ell+1}-4} + \tfrac{13}{4}\varphi_{\ell+1,2^{\ell+1}-3} - \tfrac{47}{12}\varphi_{\ell+1,2^{\ell+1}-2} + \tfrac{5}{2}\varphi_{\ell+1,2^{\ell+1}-1}. \tag{12}$$

In general case wavelets of the ℓ-th level are linear combinations of scaling functions of the $(\ell+1)$-th level. There are a few boundary wavelets on both edges – as in (10), (12). We denote their number at each edge by n_{bw}.

Left-edge boundary wavelets are

$$\psi_{\ell,i} = \sum_{j=0}^{n_1} a_{i,j} \varphi_{\ell+1,j}, \quad i = 0, \ldots, (n_{bw} - 1). \tag{13}$$

The upper bound n_1 in the sum corresponds to the support of boundary wavelets

$$\operatorname{supp} \psi_{\ell,i} = [0, (n_1 + 2)2^{-\ell-1}] \quad \text{for } i < n_{bw}.$$

We will measure the support in the units of $2^{-\ell-1}$ and denote $l_{bw} = n_1 + 2$.

Right-edge boundary wavelets are

$$\psi_{\ell,2^\ell-1-i} = \sum_{j=0}^{n_1} a_{i,j} \varphi_{\ell+1,2^{\ell+1}-1-j}, \quad i = 0, \ldots, (n_{bw} - 1)$$

with the same coefficients a_{ij} and the same upper bound l_{bw} of the length.

Inner wavelets $\psi_{\ell,i}$ for

$$i = n_{bw}, \ldots, 2^\ell - n_{bw} - 1$$

are constructed by

$$\psi_{\ell,i} = \sum_{j=0}^{n_2} a_j \varphi_{\ell+1,2(i-n_{bw})+1+j}. \tag{14}$$

It holds

$$\operatorname{supp} \psi_{\ell,i} = [2(i-n_{bw})2^{-\ell-1}, (2(i-n_{bw})+n_2+2)2^{-\ell-1}]. \tag{15}$$

We denote $l_{iw} = n_2 + 2$ the length of the support of inner wavelets in the unit $2^{-\ell-1}$.

All constructions are such that the support of the last inner wavelet contains the point $x = 1$. From here it follows

$$(2(2^{\ell} - 2n_{bw} - 1) + n_2 + 2)2^{-\ell-1} = 1$$

and as $l_{iw} = n_2 + 2$ it follows that

$$(2(2^{\ell} - 2n_{bw} - 1) + l_{iw})2^{-\ell-1} = 1$$

which gives by a straightforward calculation

$$l_{iw} = 2(2n_{bw} + 1). \tag{16}$$

Another straightforward calcualation gives that the center of the support of $\psi_{\ell,i}$ is at the point

$$x = \left(i + \tfrac{1}{2}\right) 2^{-\ell}. \tag{17}$$

In the next section we will use the value $l_w = \max\{l_{bw}, l_{iw}\}$ and the maximal number of discontinuities n_d which bases function can have. As two different discontinuities are distant at least $2^{-\ell-1}$ it holds $n_d \leq l_w - 1$.

Due to the construction of wavelets similar property to (5), (6), (7) it holds:

$$\begin{aligned} \int_0^1 \varphi'_{\ell_1,i} \varphi'_{\ell_1,j} \,\mathrm{d}x &= 2^{\ell_1-\ell_2} \int_0^1 \varphi'_{\ell_2,i} \varphi'_{\ell_2,j} \,\mathrm{d}x \\ \int_0^1 \varphi_{\ell_1,i} \varphi_{\ell_1,j} \,\mathrm{d}x &= 2^{\ell_2-\ell_1} \int_0^1 \varphi_{\ell_2,i} \varphi_{\ell_2,j} \,\mathrm{d}x \end{aligned} \tag{18}$$

and also

$$\begin{aligned} \int_0^1 \varphi'_{\ell,i} \varphi'_{\ell,j} \,\mathrm{d}x &= \int_0^1 \varphi'_{\ell,i+k} \varphi'_{\ell,j+k} \,\mathrm{d}x \\ \int_0^1 \varphi_{\ell,i} \varphi_{\ell,j} \,\mathrm{d}x &= \int_0^1 \varphi_{\ell,i+k} \varphi_{\ell,j+k} \,\mathrm{d}x \end{aligned} \tag{19}$$

whenever all involved wavelets are inner ones.

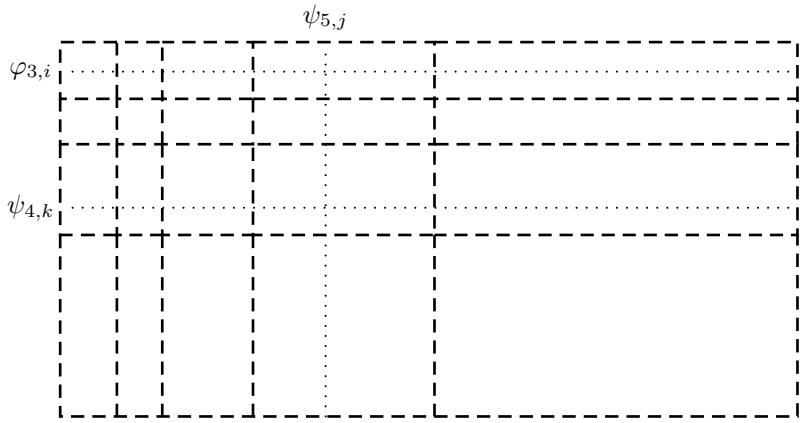
2 Wavelet Matrix

In this section, we first show that matrices (1) in a wavelet basis can be stored in a constant space. Next we show in the Theorem that they have at most linear non-zero elements with respect to the order N of the matrices.

2.1 Storage of Matrices

We store matrices (1) in blocks as in Figure 3. Let us denote $B(\ell)$ the block corresponding to scaling functions and to wavelets at the ℓ-th level. Blocks $B(4)$ and $B(5)$ as well as the top-left block corresponding to scaling functions are stored as a whole. Elements of blocks $B(\ell)$, $\ell = 6, \ldots$ are computed by the formulas (9). We store in an array both the discontinuities of scaling functions and integrals

$$\int_{x_k} \psi(x)(x - x_k)^m \,\mathrm{d}x \qquad m = \begin{cases} 1 \text{ for the matrix } g \\ 2 \text{ for the matrix } d \end{cases}$$



Source: Own

Fig. 3. *Structure of a wavelet matrix*

Now we describe how to pair discontinuities and integrals for inner wavelets (rows corresponding to boundary wavelets are zero due to choice of ℓ_0 in (8) – it is chosen in such a way that the support of the left-edge boundary wavelet does not overlap the support of the right-edge boundary wavelet). We assign the central point of its support (17)

$$x = \left(i + \tfrac{1}{2}\right) 2^{-\ell}$$

to the wavelet $\psi_{\ell,i}$ and from it we measure points x_k of discontinuities of scaling functions. They are positioned at

$$1 \cdot 2^{-3}, 2 \cdot 2^{-3}, \ldots, 6 \cdot 2^{-3}, 7 \cdot 2^{-3}, \tag{20}$$

so they can be inside the support of $\psi_{\ell,i}$ at the points

$$\ldots, (i-1)2^{-\ell}, i2^{-\ell}, (i+1)2^{-\ell}, (i+2)2^{-\ell}, \ldots \tag{21}$$

Inner wavelets are $l_w 2^{-\ell-1}$ long, so we have $n := l_w/2 - 1 = 2n_{nb}$ points (16)

$$(i-n+1)2^{-\ell}, \ldots i2^{-\ell} \tag{22}$$

on the left of x and n points

$$(i+1)2^{-\ell}, \ldots (i+n)2^{-\ell} \tag{23}$$

on the right of x. In total we have $2n = l_w - 1$ points and we will index them by an index $j = 0, \ldots, l_w - 2$. Given discontinuity at $k2^{-3}$ we have an equation

$$(i-n+1+j)2^{-\ell} = k2^{-3},$$

and so

$$i = k2^{\ell-3} + n - 1 - j.$$

As the value of the integral depends just on j and ℓ and the dependence on ℓ is given by (9), we have 8 scaling functions $\times$ 7 discontinuities $\times$ $(l_w - 2)$ integrals and the cycle of the length $56(l_w - 2)$ to construct any block $B(6),\ldots$.

Let us denote by $B(\ell_1, \ell_2)$ blocks of wavelet functions of the levels ℓ_1 and ℓ_2. In case that ℓ_1 and ℓ_2 differ by at least two we use similar idea as described above for the blocks of scaling functions. Other blocks are stored in a compressed way, separately boundary wavelets and inner wavelets and to reconstruct the whole block we use (18), (19).

2.2 Number of Non-Zero Elements

Matrices (1) in a wavelet basis have a block structure. The first block in a row and in a column corresponds to scaling functions of the third level, the second one to wavelets of the third level, then the other levels follow (Figure 3).

In the following theorem, we show that matrices (1) in a wavelet basis have at most linear number of non-zero elements with respect to the order N of the matrices.

Theorem. Let $n \in \mathbb{N}$, $n \geq 3$,

$$\mathscr{B}_n = \{\varphi_{3,i} : i = 0, \ldots, 7\} \cup \bigcup_{\ell=4}^{n} \{\psi_{\ell,i} : i = 0, \ldots, < 2^{\ell-1} - 1\},$$

let N be $|\mathscr{B}_n| = 2^n$, l_w be the maximum of length of basis functions in $2^{-\ell}$ units, n_d be a maximal number of discontinuities of basis functions. Then the matrices (1) have at most $(2n_d l_w - 2n_d + 2l_w - 1)N$ non-zero elements.

PROOF.

- *Square diagonal blocks:* the integral is zero for basis functions with disjoint supports. Supports of neighbour functions are shifted by 2^{-l+1} – see (15) – so every function meets in its support at most $(l_w - 1)$ functions and every row will contain at most $(l_w - 1)$ non-zero elements. In total the N rows in diagonal blocks contain at most $(l_w - 1)N$ non-zero elements.

- *Blocks above the diagonal:* every basis function has at most n_d points of discontinuity and every discontinuity is met by at most $(l_w/2 - 1)$ wavelets of a given finer level. So every row in every block contains at most $n_d(l_w/2 - 1)$ non-zero elements. We will calculate them by blocks in the same column from right to left and get at most $n_d(l_w/2-1)(N/2+N/4+N/8+\cdots+8) < n_d(l_w/2-1)N$ non-zero elements.

- *In blocks under diagonal* there is the same amount of non-zero elements as over diagonal as matrices are symmetric.

We get

$$(l_w - 1)N + 2n_d(l_w/2 - 1)N = (n_d l_w - 2n_d + l_w - 1)N$$

in total. □

3 Numerical Experiment

We tested our implementation on a one-dimensional Poisson equation on the interval $[0, 1]$ with the solution

$$u(x) = (1 - x)(1 - e^{-50x}). \tag{24}$$

The columns of the table contain:

- Level of wavelet basis L.
- Matrix order N.
- Number of iterations (we use conjugate gradient method) #CG.
- Total time of conjugate gradient method in seconds.
- Time per cycle in seconds.
- L_2 norm of error of a solution.

We run it on a processor with 2.4 GHz frequency.

Note that the very slowly increasing number of iterations shows that wavelet basis is well-conditioned.

Tab. 1. *Results of numerical experiment*

L	N	#CG	time of CG(s)	time per cycle	L_2 norm
12	32 768	47	0.46	0.0098	2.1×10^{-12}
13	65 536	47	0.92	0.020	2.6×10^{-13}
14	131 072	49	1.9	0.040	3.3×10^{-14}
15	262 144	49	4.0	0.082	4.1×10^{-15}
16	524 288	50	8.3	0.17	5.2×10^{-16}
17	1 048 576	51	17	0.34	6.5×10^{-17}
18	2 097 152	51	35	0.69	8.1×10^{-18}

Source: Own

Conclusion

We presented an efficient implementation of multiplication by a wavelet matrix. Our next aim is to use it to solve a multi-dimensional Poisson equation on hypercube with a basis which is a tensor product of presented basis.

Acknowledgments

This work was supported by the ESF Project No. CZ.1.07/2.3.00/09.0155 "Constitution and improvement of a team for the demanding technical computations on parallel computers at TU of Liberec".

Literature

[1] ČERNÁ, D.; FINĚK, V.: Construction of optimally conditioned cubic spline wavelets on the interval. In: *Advances in Computational Mathematics* 34(2): pp. 219-252, 2011. ISSN 1019-7168.

[2] ČERNÁ, D.; FINĚK, V.: The construction of well-conditioned wavelet basis based on quadratic B-splines. To appear In: *ICNAAM (T. E. Simos, G. Psihoyios, Ch. Tsitouras eds.)*, American Institute of Physics, New York, 2012.

[3] COHEN, A.; DAUBECHIES, I.; FEAVEAU, J.-C.: Biorthogonal bases of compactly supported wavelets. In: *Commun. Pure Apply. Math.* 45: pp. 485-560, 1992. ISSN 1097-0312.

[4] URBAN, K.: *Wavelet Methods for Elliptic Partial Differential Equations*. Oxford University Press, 2009. ISBN 978-0-19-852605-6.

RNDr. Martina Šimůnková, Ph.D.

NÁSOBENÍ WAVELETOVOU MATICÍ – EFEKTIVNÍ IMPLEMENTACE

Matice tuhosti Dirichletovy okrajové úlohy $-(au')' = f$ v bázi splajnových waveletů má dle K. Urbana $O(n \log n)$ nenulových prvků. Ukážeme, že pro konstantní funkci a jich je ve skutečnosti $O(n)$ a popíšeme algoritmus, který počítá násobení vektoru touto maticí v $O(n)$ operacích. V implementaci používáme wavelety na bázi kvadratických splajnů.

DIE MULTIPLIKATION DER WAVELET-MATRIX – EINE EFFEKTIVE IMPLEMENTIERUNG

Die Zähigkeitsmatrix der Dirichlet-Randaufgabe $-(au')' = f$ auf der Basis der Spline-Wavelets hat nach K. Urban $O(n \log n)$ Nicht-Null-Elemente. Wir zeigen, dass es für die konstante Funktion a in Wirklichkeit $O(n)$ gibt, und wir beschreiben einen Algorithmus, der die Multiplikation des Vektors durch diese Matrix in $O(n)$-Operationen berechnet. In der Implementierung benutzen wir Wavelets auf der Basis quadratischer Splines.

MNOŻENIE MACIERZĄ FALKOWĄ – EFEKTYWNE WDRAŻANIE

Macierz sztywności Dirichleta krańcowego zadania $-(au')' = f$ w bazie falek splajnowych ma wg K. Urbana $O(n \log n)$ elementów niezerowych. Pokazano, że dla funkcji stałej a jest ich w rzeczywistości $O(n)$ oraz opisano algorytm, który oblicza mnożenie wektora poprzez tę macierz w $O(n)$ operacjach. Wdrażane są falki na bazie splajnów drugiego stopnia.

Optimal Error Estimates for Nonstationary Singularly Perturbed Problems for Low Discretization Orders

Miloslav Vlasák
***Hans–Göerg Roos**

Charles University in Prague
Faculty of Mathematics and Physics
Department of Numerical Mathematics
Sokolovská 83, Praha 8, 186 75, Czech Republic
vlasak@karlin.mff.cuni.cz

*Technical University of Dresden
Institute of Numerical Mathematics
Helmholzstrasse 10, 01069 Dresden, Germany
hans-goerg.roos@tu-dresden.de

Abstract

We consider an unsteady 1D singularly perturbed convection–diffusion problem. We discretize such a problem by the linear finite element method (FEM) on a Shishkin mesh and by a discontinuous Galerkin method in time. We present optimal a priori error estimates for low order time discretizations.

Keywords: Convection–diffusion; Shishkin mesh; time discontinuous Galerkin method.

Introduction

We focus ourselves on the analysis of the solution of an unsteady linear singularly perturbed convection–diffusion equation. This type of equation can be considered as a simplified model problem to many important problems, especially to Navier–Stokes equations, which describe compressible flow.

We discretize our equation in space by the simple conforming linear finite element method on a Shishkin mesh and then we discretize the resulting system of ordinary differential equations by the discontinuous Galerkin method.

The space discretization of such a problem is a difficult task and it stimulated development of many stabilization methods (e.g. a streamline upwind Petrov–Galerkin (SUPG) method, local projection stabilizations) and layer–adapting techniques (e.g. Shishkin meshes, Bakhvalov meshes). For the complete overview see [6].

Considering the space discretization on Shishkin meshes, we will follow the theory for stationary singularly perturbed problems based on the solution decomposition, which enables us to derive a priori error estimates independent of the diffusion parameter even with respect to the norms (seminorms) of the exact solution, which can be also highly dependent on the diffusion parameter. For the details see [6].

The discontinuous Galerkin (DG) method is a very popular approach for solving ordinary differential equations arising from the space discretization of parabolic problems, which is based on the piecewise polynomial approximation in time. Among important advantages we should mention unconditional stability for arbitrary order, which allows us to solve stiff problems efficiently, and good smoothing property, which enables us to work with inexact or rough data. We should also mention that the DG method is suitable for changes in our computational

domain and in computational spaces, which allows us to exploit adaptivity during the computational process. For introduction to the DG time discretization see e.g. [8].

In [1] and [4] the authors study the DG method in time and the local projection stabilization respectively the DG method in space on standard meshes for singularly perturbed problems. The error estimates in these papers contain norms of the exact solutions which go to infinity if the diffusion parameter goes to 0.

There are only few papers dealing with finite elements in space on the special meshes combined with any discretization in time. We should mention [3] and [5], where the backward difference formula (BDF) time discretizations and the θ–scheme are used and a priori error estimates are derived. In [5] the authors also study the DG time discretization and derive suboptimal error estimates.

In contrast to the results in [5], we present a sketch of the proof of optimal error estimates for the DG time discretization for lower convergence orders in $L^\infty(L^2)$ norm.

The main difficulty in proving optimal error estimates for the DG time discretization combined with a space discretization on a Shishkin mesh is the fact that we cannot employ a standard technique of the proof, which is based on the construction of a suitable projection, which enables us to eliminate a discrete time derivative in the error equation, see e.g. [7]. This technique enforces us to do the upper bound of the projection error contained in stationary terms, which depends on a higher time derivative of the exact solution in H^1 seminorm, which depends on the diffusion parameter.

1 Continuous Problem

Let us consider the 1D parabolic singularly perturbed problem

$$\begin{aligned} \frac{\partial u}{\partial t}(x,t) - \varepsilon \frac{\partial^2 u}{\partial x^2}(x,t) + b\frac{\partial u}{\partial x}(x,t) + cu(x,t) &= f(x), \quad \forall x \in (0,1), t \in (0,T), \\ u(0,t) = u(1,t) &= 0, \quad \forall t \in (0,T), \\ u(x,0) &= u^0(x), \quad \forall x \in (0,1), \end{aligned} \tag{1}$$

where functions $f \in L^2(0,1)$, $u^0 \in L^2(0,1)$, $0 < \varepsilon << 1$ and functions b and c are sufficiently smooth with $b(x) > \beta > 0$. By substitution in time variable we can achieve

$$c - \frac{1}{2}\frac{\partial b}{\partial x}(x) \geq c_0 > 0. \tag{2}$$

Let us define the bilinear form

$$a(u,v) = \int_0^1 \varepsilon \frac{\partial u}{\partial x}\frac{\partial v}{\partial x} + \left(b\frac{\partial u}{\partial x} + cu \right) v dx. \tag{3}$$

To simplify the text we will use the following notation. $(.,.)$ and $\|.\|$ are $L^2(0,1)$ scalar product and norm, $|.|_1$ and $\|.\|_1$ are $H^1(0,1)$ seminorm and norm.

It is possible to show that the solution has in general the boundary layer at $x = 1$. When we assume sufficiently compatible data, we can avoid the existence of interior layers, which enables us to concentrate on the boundary layer only, see [6]. Moreover, it is possible to prove

$$\left| \frac{\partial^{k+j} u(x,t)}{\partial^k x \partial^j t} \right| \leq C \left(1 + \frac{1}{\varepsilon^k e^{\beta(1-x)/\varepsilon}} \right), \quad \forall k, j \geq 0. \tag{4}$$

This result shows dependence of space derivatives on ε, which complicates deriving of standard a priori error estimates.

1.1 Discretization

We want to discretize the problem (1) by the standard finite element method on Shishkin meshes in space. This technique allows us to derive a priori error estimates that are independent of ε. We define the parameter

$$\sigma = \frac{5}{2}\frac{\varepsilon}{\beta}\log(N), \tag{5}$$

where N is the number of mesh points. We assume our mesh points are equidistantly distributed in intervals $[0, 1-\sigma]$ and $[1-\sigma, 1]$, with the same number of mesh points in both intervals. Let us define the conforming linear finite element space V_N on our mesh.

To discretize the problem (1) in time we assume the time partition $0 = t_0 < t_1 < \ldots < t_r = T$ with time intervals $I_m = (t_{m-1}, t_m)$, time steps $\tau_m = |I_m| = t_m - t_{m-1}$ and $\tau = \max_{m=1,\ldots,r} \tau_m$. We denote the function values at the nodes as $v^m = v(t_m)$. To be able to use the Galerkin type of discretization we denote the space of piecewise polynomial functions

$$V_N^\tau = \{v \in L^2(0,T,V_N) : v|_{I_m} = \sum_{j=0}^{q} v_{j,m} t^j,\ v_{j,m} \in V_N\}. \tag{6}$$

For the functions from such a space we need to define the values at the nodes of time partition

$$v_\pm^m = v(t_m \pm) = \lim_{t \to t_m \pm} v(t) \tag{7}$$

and the jumps

$$\{v\}_m = v_+^m - v_-^m. \tag{8}$$

Definition 1 *We say that the function $U \in V_N^\tau$ is the approximate solution to the problem (1) if*

$$\begin{aligned} \int_{I_m} (U', v) + a(U, v)dt + (\{U\}_{m-1}, v_+^{m-1}) &= \int_{I_m} (f, v)dt, \\ &\quad \forall v \in V_N^\tau,\ \forall m = 1, \ldots, r \\ (U_-^0, v) &= (u^0, v) \quad \forall v \in V_N. \end{aligned} \tag{9}$$

2 Error Analysis

We define the weighted norm

$$\|v\|_\varepsilon^2 = \varepsilon |v|_1^2 + \|v\|^2, \quad \forall v \in H^1(0,1). \tag{10}$$

It is possible to show that

$$a(v,v) \geq \min(c_0, 1)\|v\|_\varepsilon^2 \geq 0. \tag{11}$$

2.1 Stationary Problem

In this part we want to go through some well known results for the singularly perturbed problems (for the details see [6]). Let us assume a related stationary problem

$$a(u,v) = (f^*, v), \quad \forall v \in H_0^1(0,1), \tag{12}$$

with some $f^* : L^2(0,1)$, and the corresponding discrete finite element problem on the layer–adapted mesh. Let us define the Ritz projection $R : H_0^1(0,1) \to V_N$ satisfying

$$a(u - Ru, v) = 0, \quad \forall v \in V_N. \tag{13}$$

It is possible to prove following estimates:

$$\|u - Ru\|_\varepsilon \leq CN^{-1}\log(N), \tag{14}$$
$$\|u - Ru\| \leq C(N^{-1}\log(N))^2, \tag{15}$$

with C independent of ε. For the proof see e.g. [6].

2.2 Radau Quadrature

Let us define the Radau quadrature on each interval I_m

$$Q[f] = \sum_{i=0}^{q} w_i f(t_{m,i}), \tag{16}$$

where $t_{m,i}$ are Radau quadrature nodes in I_m with $t_{m,0} = t_m$. Such a quadrature has the algebraic order $2q$ and the coefficients of the quadrature satisfy $0 \leq w_i \leq \tau_m$.

It is possible to express our method (9) by

$$Q[(U', v)] + Q[a(U, v)] + (\{U\}_{m-1}, v_+^{m-1}) = Q[(f, v)], \quad \forall v \in V_N^\tau. \tag{17}$$

Since the equation for the continuous solution (1) is defined at every point $t \in I_m$, we can see that

$$Q[(u', v)] + Q[a(u, v)] + (\{u\}_{m-1}, v_+^{m-1}) = Q[(f, v)], \quad \forall v \in V_N^\tau. \tag{18}$$

2.3 Projections

We define the space

$$V^\tau = \{v \in L^2(0,T,H_0^1(0,1)) : v|_{I_m} = \sum_{j=0}^{q} v_{j,m} t^j,\ v_{j,m} \in H_0^1(0,1)\}. \tag{19}$$

We define the time projection $P : C([0,T],H_0^1(0,1)) \to V^\tau$, such that

$$Pu(t) = \sum_{i=0}^{q} \ell_i(t) u(t_{m,i}), \tag{20}$$

where ℓ_i is a Lagrange interpolation basis function for the quadrature node $t_{m,i}$. Since

$$RPu(t) = R\sum_{i=0}^{q} \ell_i(t) u(t_{m,i}) = \sum_{i=0}^{q} \ell_i(t) Ru(t_{m,i}) = PRu(t), \tag{21}$$

we can see that projections P and R commute. We define the space–time projection $\pi = PR$: $C(0,T,H_0^1(0,1)) \to V_N^\tau$.

Now, we present some basic approximation properties of our projections P and π.

Lemma 1 *Let u be the exact solution of (1). Then*

$$\sup_{I_m} \|Pu - u\| \quad \leq \quad C\tau^{q+1}, \tag{22}$$

where the constant C does not depend on τ.

The proof can be made by standard arguments. It is an analogy to e.g. [2, Theorem 3.1.5] in Bochner spaces.

Lemma 2 *Let u be the exact solution of (1). Then*

$$\sup_{I_m} \|\pi u - u\| \quad \leq \quad C(\tau^{q+1} + (N^{-1}\log(N))^2), \tag{23}$$

where the constant C does not depend on τ *or N*.

The proof of the lemma follows directly from estimates of the projection P and R and from the triangle inequality.

2.4 Main Result

We are ready to present the main result.

Theorem 1 *Let u be an exact solution of (1) and* $U \in V_N^\tau$ *be its discrete approximation given by (9) with* $q = 0, 1$. *Then*

$$\max_{m=1,\ldots,r} \sup_{I_m} \|U - u\| \leq C\left((N^{-1}\log(N))^2 + \tau^{q+1}\right). \tag{24}$$

To prove the theorem we divide the error $U - u$ into projection part $\eta = \pi u - u$ and $\xi = U - \pi u \in V_N^\tau$. Then we subtract the equation for the exact solution (18) from the equation for the discrete solution (17) and we obtain

$$\begin{aligned} &\int_{I_m} (\xi', v) + a(\xi, v)dt + (\{\xi\}_{m-1}, v_+^{m-1}) \\ &= -Q[(\eta', v)] - (\{\eta\}_{m-1}, v_+^{m-1}) - Q[a(\eta, v)]. \end{aligned} \tag{25}$$

Since $Pu(t_{m,i}) = u(t_{m,i})$, we get

$$Q[a(\eta, v)] = \sum_{i=0}^{q} w_i a(Ru(t_{m,i}) - u(t_{m,i}), v) = 0 \tag{26}$$

we need to estimate the rest of the right–hand side only.

Lemma 3 *Let u be an exact solution of (1). Then*

$$Q[(\eta', v)] + (\{\eta\}_{m-1}, v_+^{m-1}) \leq \tau_m C\left(\tau^{q+1} + (N^{-1}\log(N))^2\right) \sup_{I_m} \|v\|, \quad \forall v \in V_N^\tau. \tag{27}$$

The proof of the lemma is rather long and technical and will be published in detail in forthcoming paper.

We can estimate the right–hand side of (25) by Lemma 3. Then we obtain by setting $v = 2\xi$

$$\begin{aligned} &\|\xi_-^m\|^2 - \|\xi_-^{m-1}\|^2 + \|\{\xi\}_{m-1}\|^2 + 2\min(c_0,1)\int_{I_m} \|\xi\|_\varepsilon^2 dt \\ &\leq \tau_m C\left(\tau^{q+1} + (N^{-1}\log(N))^2\right) \sup_{I_m} \|\xi\|. \end{aligned} \tag{28}$$

We need to deal with the term at the right–hand side. For the case $q = 0$ we know that $\|\xi\|$ is constant with respect to time and we can exchange the last expression by the term $\tau_m C(\tau^{q+1} + (N^{-1}\log(N))^2)\|\xi_-^m\|$. Then it is sufficient to employ Young's inequality and the discrete Gronwall lemma to finish the proof of the theorem. The case $q = 1$ is still quite simple. For $q = 1$ the term $\|\xi\|$ is linear with respect to time and so we can find its supremum at one of the end points of the interval I_m. If the supremum occurs at the point t_m, we can follow the same idea as in the case $q = 0$. If the supremum occurs at the point t_{m-1}, we can divide our term

$$\|\xi_+^{m-1}\| = \|\xi_+^{m-1} - \xi_-^{m-1} + \xi_-^{m-1}\| \leq \|\{\xi\}_{m-1}\| + \|\xi_-^{m-1}\|. \tag{29}$$

Then we need again to employ carefully Young's inequality and the discrete Gronwall lemma to finish the proof of the theorem.

Conclusion

For simplicity, we assume the main result with $q = 0, 1$ only. Nevertheless, the result holds true even for arbitrary $q \geq 0$. Then the proof of the theorem will be more complicated and we will not consider such cases. The result holds also true for consistent stabilization methods and general layer–adapted meshes with a slightly different term describing convergence behavior with respect to space. It is also not important to restrict ourselves to 1D. We can simply extend the results from [3] discussing multidimensional case. The fully general result with complete detailed proofs will be publishes in forthcoming paper.

Acknowledgments

The first author is a junior researcher of the University centre for mathematical modelling, applied analysis and computational mathematics (Math MAC).

Literature

[1] AHMED, N.; MATTHIES, G.; TOBISKA, L.; XIE, H.: Discontinuous Galerkin time stepping with local projection stabilization for transient convection-diffusion-reaction problems. *Comput. Methods Appl. Mech. Eng.*, 200(21-22): pp. 1747-1756, 2011.

[2] CIARLET, P. G.: *The finite element methods for elliptic problems.* Repr., unabridged republ. of the orig. 1978. Classics in Applied Mathematics. 40. Philadelphia, PA: SIAM. xxiv, 530 p. , 2002.

[3] DOLEJŠÍ, V.; ROOS, H.-G.: BDF-FEM for parabolic singularly perturbed problems with exponential layers on layer-adapted meshes in space. *Neural Parallel Sci. Comput.*, 18(2): pp. 221-235, 2010.

[4] FEISTAUER, M.; HÁJEK, J.; ŠVADLENKA, K.: Space-time discontinuos Galerkin method for solving nonstationary convection-diffusion-reaction problems. *Appl. Math., Praha*, 52(3): pp. 197-233, 2007.

[5] KALAND, L.; ROOS, H.-G.: Parabolic singularly perturbed problems with exponential layers: robust discretizations using finite elements in space on Shishkin meshes. *Int. J. Numer. Anal. Model.*, 7(3): pp. 593-606, 2010.

[6] ROOS, H.-G.; STYNES, M.; TOBISKA, L.: *Robust numerical methods for singularly perturbed differential equations. Convection-diffusion-reaction and flow problems. 2nd ed.* Springer Series in Computational Mathematics 24. Berlin: Springer. xiv, 604 p., 2008. ISBN 978-3-540-34466-7/hbk.

[7] SCHÖTZAU, D.: *hp-DGFEM for parabolic evolution problems. Application to diffusion and viscous incompressible flow*. Ph.D. thesis, ETH Zürich, 1999.

[8] THOMÉE, V.: *Galerkin finite element methods for parabolic problems.* 2nd revised and expanded ed. Berlin: Springer. xii, 370 p., 2006. ISBN 3-540-33121-2/hbk.

Optimální odhady chyb nestacionárních singulárně perturbovaných problémů pro diskretizace nízkého řádu

Uvažujeme nestacionární jednodimenzionální singuálrně perturbovaný problém. Diskretizujeme tento problém v prostoru pomocí metody konečných prvků na Shishkinových sítích a v čase pomocí nespojité Galerkinovy metody. Ukážeme optimální apriorní odhady chyb pro časové diskretizace nízkého řádu.

Optimale Schätzungen nichtstationärer singulär gestörtes Probleme für die Diskretierung niederer Ordnung

Wir betrachten ein nichtstationäres eindimensionales singulär gestörtes Problem. Wir diskretieren dieses Problem im Raum mit Hilfe der Methode der finiten Elemente auf den Shishkin'schen Netzen und in der Zeit mit Hilfe der nichtkontinuierlichen Galerkin-Methode. Wir zeigen die optiamlen A-priori-Schätzungen von Fehlern für die zeitliche Diskretierung niederer Ordnung.

Optymalne szacowanie błędów niestacjonarnych osobliwie zaburzonych problemów dla dyskretyzacji niskiego rzędu

W artykule opisano niestacjonarny jednowymiarowy osobliwie zaburzony problem. Problem ten jest dyskretyzowany w przestrzeni przy pomocy metody elementów skończonych na siecіach Shishkina oraz w czasie przy pomocy nieciągłej metody Galerkina. Pokazano optymalne a priori szacunki błędu dla czasowej dyskretyzacji niskiego rzędu.

On Two Flexible Methods of 2-Dimensional Regression Analysis

Petr Volf

Technical University of Liberec
Faculty of Science, Humanities and Education
Studentská 2, 461 17, Liberec, Czech Republic

*Institute of Information Theory and Automation
Academy of Sciences of the Czech Republic
Prague 8, Czech Republic
petr.volf@tul.cz

Abstract

The paper deals with the problem of non-parametric statistical modeling of 2-dimensional surfaces from observed data, i.e. the regression analysis. In general, the model is constructed from a set of basal functions, as are the splines, gaussians and others. However, such modeling means to estimate a large set of parameters (locations of functional units and parameters of their combination). We shall present two approaches allowing reduction of the number of needed parameters. Namely, a well known method of projection pursuit, and the less known method of Gordon surface. Further, we shall analyze possible serious consequences of sparse data to precision of model and uncertainty of prediction. Methods will be illustrated in artificial examples.

Keywords: Statistics; regression analysis; splines; projection pursuit; Gordon surface; prediction error.

Introduction

Though the main concern of the paper is two-dimensional regression analysis, we shall start, in Part 1, with a 1-dimensional case. We think that it is the best way how to show the way of modeling curves from functional units and the problems connected with such an approach. The use of localized units (as B-splines or gaussians) is convenient when we wish to describe a non-regularly varying function (a signal, for instance). However, as we shall see, the use of combination of localized units requires also sufficiently 'localized' (i.e. dense) measurements, to avoid unexpected non-precision of model performance.

This problem will be illustrated first in the case of a 1-dimensional curve, variance of prediction will be computed, and its relationship with measurement design shown. Then, in Part 2, we shall devote to the 2-dimensional regression models. We shall recall some approaches how the number of involved parameters can be reduced. Then, after brief overview of the projection pursuit method, we concentrate, in Part 3, to the model construction via so called Gordon surface. Even here, we shall analyze the relation between data design and model (and prediction) uncertainty.

In Part 1 we shall employ also the MCMC (Markov chain Monte Carlo) procedures in the framework of the Bayes statistical inference. The reason is that MCMC generates a representation of posterior distribution of estimated model. With its aid it is possible to visualize the variability of estimates. Simultaneously, estimated distribution of predicted values is obtained. More details on the MCMC methods can be found elsewhere, for instance in [5] and also in [8].

1 Nonparametric Regression

Let us consider first a pair of one-dimensional random variables X (input variable, predictor) and Y (output variable, response) and a general response model defined by a density $f(y;r(x))$ of conditional probability of Y for given $X = x$, where $r(x)$ is a smooth non-parametrized response function. This definition involves, as a special case, the standard regression model $Y = r(X) + \varepsilon$, where ε is a random Gauss noise with zero mean and an unknown variance σ^2.

There are essentially two different ways how to estimate unknown function r. The first consists in the local (e.g. kernel) smoothing. The other approach, studied here, employs the approximation of $r(x)$ by a combination of functions from some functional basis. For instance radial basis functions (gaussians), polynomial splines, goniometric functions or wavelets are popular choices. Hence, the model of response function has the form

$$r_M(x) = \boldsymbol{\alpha}' \boldsymbol{B}(x;\boldsymbol{\beta}) = \sum_{j=1}^{M} \alpha_j B_j(x;\boldsymbol{\beta}), \tag{1}$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_M)'$ is a vector of linear parameters, B_j are basis functions and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_M)'$ is a vector of parameters of the basis functions (e. g. knots of splines, centers and scales of radial functions). While the estimates of $\boldsymbol{\alpha}$ can be obtained directly from linear regression context, estimation of $\boldsymbol{\beta}$ is a difficult optimization problem. As a solution to the nonlinear problem for coefficients $\boldsymbol{\beta}$ as well as to optimal choice of M, it is possible to use the Bayes methodology in combination with the Markov chain Monte Carlo (MCMC) algorithms. In this framework, the parameter $\boldsymbol{\beta}$ is considered to be a multi-dimensional random vector, with a prior distribution satisfying certain constraint. Simultaneously, M is also regarded as a random variable, with some prior on $\{0, 1, 2, \ldots, M_{max}\}$.

1.1 Modeling via B–splines

Polynomial splines are constructed from piece-wise polynomials which are joined together in the 'knots'. At these points, continuity conditions are fulfilled. We mostly deal with the cubic splines which have continuous two first derivatives. There are several variants of functional bases creating the spline, we prefer the B–splines as they are localized. Let us consider an interval $[a,b]$ and a set of M different inner knots, $\beta_0 = a < \beta_1 < \ \ldots\ \beta_M < \beta_{M+1} = b$, let us add six other 'dummy' knots $\beta_{-j} = a - j(\beta_1 - a),\ \beta_{M+1+j} = b + j(b - \beta_m),\ j = 1,2,3$. One way how to define the B–spline function, following for instance [9], employs divided differences:

$$B_j(x,\boldsymbol{\beta}) = \sum_{k=j-2}^{j+2} \left\{ (x-\beta_k)_+^3 \Big/ \prod_{s=j-2, s\neq k}^{j+2} (\beta_k - \beta_s) \right\},$$

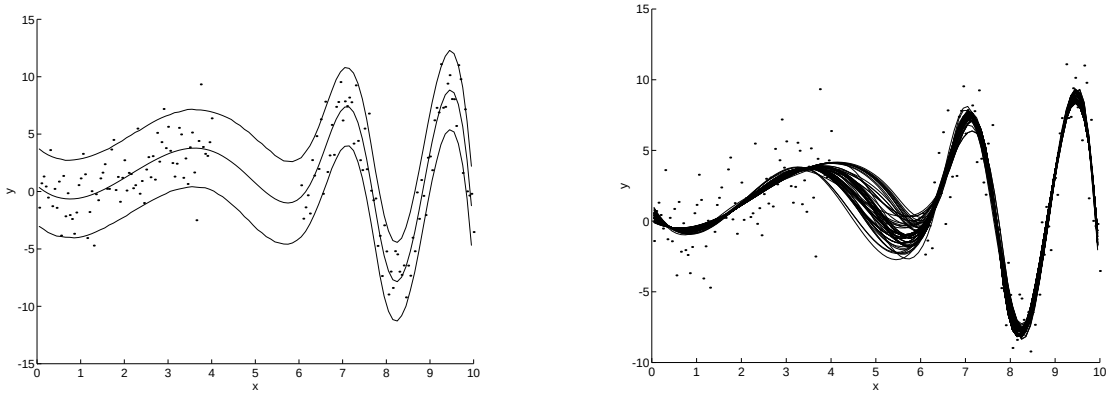
for $j = -1, 0, 1,\ \ldots\ M, M+1, M+2$. Here, function $(u)_+$ means $u \cdot 1[u > 0]$. Thus, M inner knots define $M+4$ basal cubic units. Each unit B_j is zero outside the interval $(\beta_{j-2}, \beta_{j+2})$. The interference of two neighboring units depends on position of the knots, a change of one knot has effect on several nearest units only (e. g. in the case of cubic B–splines, five units have to be updated when one knot is changed). It means that in model (1) a change of one β_k results in updating of only five $\alpha_j,\ j = k-2, k-1, k, k+1, k+2$. This leads to the reduction of necessary computations.

1.2 Optimal Number of Units

It is expected that the model with more units decreases residual variance (or increases the likelihood). Therefore we should examine whether the addition of one unit from corresponding

functional basis improves the fit of the model 'sufficiently'. In a non–Bayes setting, this is often measured by a penalty criterion, e.g. Akaike's AIC, Schwarz's BIC, GCV (see also Friedman [4]). Similar is the criterion $\hat{\sigma}^2_M exp(\frac{M}{N^\gamma})$, where γ is a number from $(0.5,1)$, $\hat{\sigma}^2_M$ is the estimate of residual variance σ^2, M denotes the number of used units, N is the extent of data sample.

Equivalently, we can obtain the penalty as a part of the acceptance probability in the MCMC algorithm. Let us assume that the prior for variable M is specified in such a way that the proportion $Q_0(M^*)/Q_0(M) < 1$ if $M^* > M$. For instance, if prior is proportional to $\exp(-M/N^\gamma)$. Such a choice decreases the chance to accept a model with higher number of units, if the gain of that model (measured by likelihood ratio) is low. The addition of new units can be complementary controlled by a rule guaranteeing a reasonable minimal distance between them and by prescribing maximal number of units.



Source: Own

Fig. 1. *Left: Data, estimated regression curve (central) and $\pm 2\hat{\sigma}$ bands. Right: Variability of MCMC generated posterior representation $r^{(m)}(x)$*

Example 1. 160 uniformly distributed points x_i were sampled in (0,4) ∪ (6,10), and output values $y_i = r(x_i) + \varepsilon_i$ were then generated, where

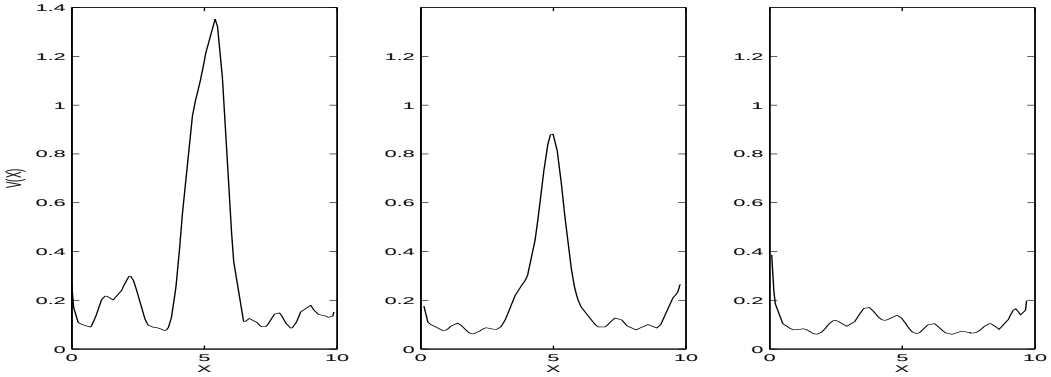
$$r(x) = x \sin\left(\left(\frac{x}{0.25}\right)^2\right)$$

and ε_i were independent identically distributed Gauss random variables with mean zero and variance $\sigma^2 = 4$. For estimation of function $r(x)$, the cubic B-splines were used. As regards the prior for their knots, we used uniform distribution on the set $\{0 < \beta_1 < \beta_2 < \ \ldots \ < \beta_M < 10\}$. 100 loops of the Markov chain generation were performed. One loop updated sequentially all components of $\boldsymbol{\beta}$, with possible change of M. It means that it contains between 20 – 50 iterations of model, depending on actual number M.

Only the final result after each loop was registered as a new member of the chain, $r^{(m)}$. The chain obtained in such a manner has a rather low autocorrelation. The average of this sequence of functions, after skipping first $s = 20$ of them,

$$\hat{r}(x) = \frac{1}{S-s} \sum_{m=s+1}^{S} r^{(m)}(x), \tag{2}$$

serves then as the final estimate of $r(x)$, empirical variance or the quantiles of the set $\{r^{(m)}(x), m = s+1, ..., S\}$ yield information about the uncertainty of the estimate at given x. In the case of Gauss random noise, unknown parameter σ is estimated from the averaged squares of residuals of the preceding iteration. The procedure started from 7 initial units defined by 3 inner knots located equidistantly inside (0,10). It converged to final 13 units, with final estimate $\hat{\sigma}^2 = 4.18$. Figure 1, left plot, shows the points $\{x_i, y_i\}$, the estimate $\hat{r}(x)$ and $\hat{r}(x) \pm 2\hat{\sigma}$ intervals connected to two bands. However, the right plot shows the variability of last 80 members of chain $r^{(m)}(x)$. It can be taken as a representation of posterior distribution of $r(x)$, quite sufficient for illustration of certain important features. Namely, it is seen how the variability of estimate increases in the region with sparse data. It also means that there increases uncertainty of prediction of true values of $r(x)$ (compare also discussion in [2], Ch. 10). A vertical cut at a given x represents Bayes prediction distribution of corresponding $r(x)$.



Source: Own

Fig. 2. *Evaluation of function $V(x)$, in the case without data in (4,6) (left), with one measurement added to $x = 5$ (center), from data distributed uniformly through whole (0,10) (right)*

1.3 Variance of Prediction

In the present part we shall recall some well known results quantifying the variance of prediction in linear regression model, see for instance [1], and adapt them to our case. For simplicity, let us assume that the functional units, i. e. their number and inner parameters, are fixed. Hence, we solve the linear regression case

$$y_i = \boldsymbol{B}^T(x_i) \cdot \boldsymbol{\alpha} + \varepsilon_i, \quad i = 1, \ldots, n,$$

where ε_i are the i.i.d. normal random variables $\mathcal{N}(0, \sigma^2)$, $\boldsymbol{B}(x_i) = (B_1(x_i), \ldots, B_M(x_i))^T$ are functional units (e. g. B-splines) evaluated at data-points $\boldsymbol{x} = (x_1, \ldots, x_n)^T$. Denote $\boldsymbol{B}$ the $n \times M$ matrix with rows $\boldsymbol{B}^T(x_i)$, $\boldsymbol{A} = (\boldsymbol{B}^T \cdot \boldsymbol{B})^{-1}$, $\boldsymbol{y} = (y_1, \ldots, y_n)^T$. Then the least squares method yields the estimate

$$\hat{\boldsymbol{\alpha}} = \boldsymbol{A} \cdot \boldsymbol{B}^T \cdot \boldsymbol{y}, \quad \hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha} \sim \mathcal{N}(\boldsymbol{O}, \sigma^2 \cdot \boldsymbol{A}),$$

where $\boldsymbol{O}$ is the null vector. Further, at a selected point z the prediction of $y(z)$ is $\hat{y}(z) = \boldsymbol{B}^T(z) \cdot \hat{\boldsymbol{\alpha}}$. Its expectation is $r(z)$, while its variance equals

$$\mathrm{var}\big(\hat{y}(z)\big) = \boldsymbol{B}^T(z) \cdot \boldsymbol{A} \cdot \boldsymbol{B}(z) \cdot \sigma^2.$$

We see that it depends both on data ($\boldsymbol{A} = \boldsymbol{A}(\boldsymbol{x})$) from which the model was estimated, and on position of prediction point z.

On this basis we can study interaction of data design and prediction error, particularly the influence of additional measurements to model precision. For illustration, let us return to Example 1. Let the model be constructed from B-splines defined by 7 inner knots placed equidistantly inside $(0,10)$, so that the model contains $M = 11$ B-spline functions. Figure 2 shows function $V(z) = \boldsymbol{B}^T(z)\boldsymbol{A}\boldsymbol{B}(z)$ in 3 cases. The left plot corresponds to data shown in Figure 1, i. e. without measurements in interval $(4,6)$. The central plot corresponds to the situation when one measurement was added to $x = 5$. Finally, the right-hand plot shows function $V(z)$ computed from data distributed uniformly over whole interval $(0,10)$, without any gap.

2 Multi-Dimensional Case

Now we assume that the input variable $\boldsymbol{X} = (X_1, \ldots, X_p)$ is a p-component vector. In general, the multivariate regression modeling has to deal with functions of interactions of several predictors. Such a function is as a rule modeled by a tensor product of one-dimensional units. The problem with multivariate units is not only that their number grows (exponentially) with dimension, but that there also grows a number of units (and parameters) which are influenced by updating of one component of 'inner' parameter (e. g. of one knot). For instance, a function $r(\boldsymbol{x})$ in R^2 can be modeled as

$$r(\boldsymbol{x}) = \alpha_{00} + \sum_{j=1}^{M_1} \alpha_{j0} B_j(x_1, \boldsymbol{\beta}) + \sum_{k=1}^{M_2} \alpha_{0k} C_k(x_2, \boldsymbol{\gamma}) + \sum\sum \alpha_{jk} B_j(x_1, \boldsymbol{\beta}) C_k(x_2, \boldsymbol{\gamma}). \qquad (3)$$

Such a function contains $M_1 + M_2$ inner parameters β_j, γ_k, but $1 + M_1 + M_2 + M_1 M_2$ 'linear' parameters α_{jk}. The high number of parameters leads naturally to the high time needed for (iterative, as a rule) computations. That is why there are attempts to reduce the dimensionality of 'decision space' of the model construction. Especially the methods based on the idea of decision tree are successful. The most known one is the CART (Classification and Regression Tree), giving a histogram-like result, and its modification giving a continuous function, the MARS (Multi-dimensional Adaptive Regression Splines) [4].

In the simplest scenario the multi-dimensional model has an additive form. The response function $r(\boldsymbol{x})$ is then a sum of p functions of one variable. In our context

$$r(\boldsymbol{x}) = \sum_{k=1}^{p} \sum_{j=1}^{M_k} \alpha_{jk} B_{jk}(x_k, \boldsymbol{\beta}_k) \;=\; \sum_{k=1}^{p} \boldsymbol{\alpha}_k' \boldsymbol{B}_k(x_k, \boldsymbol{\beta}_k),$$

vectors $\boldsymbol{\beta}_k$ and values M_k, $k = 1, \ldots, p$, are optimized iteratively. Most of algorithms innovate sequentially one component after another. Naturally, flexibility of such a model is rather limited.

2.1 Projection Pursuit

This is one of the approaches how to model the multi-dimensional interactions of input variables [7]. The PP estimator has the following form

$$r^*(\boldsymbol{x}) = \sum_{j=1}^{K} s_j(\boldsymbol{\beta}'_j \boldsymbol{x}),$$

i.e. it is the sum of 1-d functions of linear combinations (rotations) of covariates. Now, the objective is to find an optimal K and optimal p-dimensional vectors $\boldsymbol{\beta}_j$. The problem of estimation of non-parametrized functions s_j is then solved in the framework of additive model. The space

of $\boldsymbol{\beta}$'s is limited to such that $|\boldsymbol{\beta}_j| = 1$. Notice also that the rotation in R^p is fully described by an angle having $p-1$ components, each with values in $[0,\pi]$. So that the dimension of nonlinear parameter is actually $p-1$. For instance, in R^2 the angles of rotations can be ordered to a sequence $0 \le \gamma_1 < \gamma_2 < ... < \gamma_K < \pi$, so that the process of solution is quite similar to that dealing with optimal location of knots in R^1. Thus, the problem of construction of 2-dimensional model is changed to two nested 1-dimensional problems.

On the other hand, it is well known that the projection pursuit is highly sensitive, that the dependence on $\boldsymbol{\beta}$'s is rather non-smooth. The method is implemented to several popular statistical software packets (S-plus, R).

2.2 Variance of Prediction in PP

We shall study now the error or prediction in a similar manner as in part 2.4. Again, let us assume that the angles of rotations γ_k, $k = 1,\dots,K$, are selected, and that also functional units $B_{jk}(z)$ are already fixed, $j = 1,2,\dots,M_k$ for k-th projection. Thus, we shall deal with the model

$$y(\boldsymbol{x}) = \sum_{k=1}^{K} \sum_{j=1}^{M_k} \alpha_{jk} B_{jk}(z_k),$$

where we denoted $z_k = \gamma_k \star \boldsymbol{x} = \cos\gamma_k \cdot x_1 + \sin\gamma_k \cdot x_2$ the projection of point $\boldsymbol{x} = (x_1, x_2)$ to direction γ_k. Thus, we again deal with the linear regression scheme

$$y_i = \boldsymbol{B}^T(\boldsymbol{x}_i) \cdot \boldsymbol{\alpha} + \varepsilon_i,$$

where $\boldsymbol{B}(\boldsymbol{x}_i) = \{B_{jk}(z_{ki}),\ j = 1,\dots,M_k,\ k = 1,\dots,K\}$, $z_{ki} = \gamma_k \star \boldsymbol{x}_i$, $\boldsymbol{\alpha} = \{\alpha_{jk}\}$, its dimension is $M = \sum_{k=1}^{K} M_k$, and $(y_i, \boldsymbol{x}_i)$, $i = 1,\dots,n$, are measurements. Formally the case is the same as the case discussed in 2.3. A numerical illustration is a part of Example 2 in the next section.

3 A Model of Gordon Surface

The Gordon surface [6] is one of constructions of smooth surfaces used mostly in engineering applications, some others from this set are for instance Extrusion Surface, Ruled Surface or Coons Patch. We shall adapt it here to the needs of 2-dimensional statistical regression analysis.

Let us consider a surface S given by a smooth function $y(\boldsymbol{x})$, $\boldsymbol{x} \in X \subset R_2$. We assume that S is given in a form of the Gordon surface, namely: Consider a set of K lines L_k $(k = 1,\dots,K)$ crossing the domain X. Cuts of surface S above lines L_k are smooth functions $y_k(z)$, $z \in L_k$. For each point $\boldsymbol{x} \in X$, let $\boldsymbol{x}(k)$ be its projection to L_k. Further, let $R_k(\boldsymbol{x})$ be a weight of point $\boldsymbol{x}$ w.r. to line L_k. We assume that $R_\ell(\boldsymbol{x}) = 1$ if $\boldsymbol{x} \in L_\ell$ and for points lying between lines L_k the weight is given by a convenient (sufficiently smooth) kernel function, e. g. by a Gauss kernel, with property $\sum_{k=1}^{K} R_k(\boldsymbol{x}) = 1$. Then let us define

$$r(\boldsymbol{x}) = \sum_{k=1}^{K} R_k(\boldsymbol{x}) \cdot r_k(\boldsymbol{x}(k)) \tag{4}$$

as a surface value at $\boldsymbol{x}$. Regarding the functions $r_k(z)$, it is assumed that they are linear combinations of 1-dimensional functional units (e. g. B-splines or radial-basic functions) $\varphi_{km}(z)$

$$r_k(z) = \sum_{m=1}^{M_k} w_{km}\, \varphi_{km}(z) = \boldsymbol{w}'_k \cdot \boldsymbol{\varphi}_k(z).$$

Hence, each such function contains M_k unknown parameters w_{km} and localized units $\varphi_{km}(z)$. The units, as well as blending (mixing) weights $R_k(\boldsymbol{x})$, should be selected by an analyst. It is seen that the model has a similar feature as the projection pursuit, instead of beams given by rotations the core is given by a net of lines, and the projection of data points is weighted. The simplest choice of lines L_k are equidistant vertical (or horizontal) lines crossing X.

3.1 Statistical Model and Estimation

Let us now imagine a set of measurements $(y_i, \boldsymbol{x}_i)$, $i = 1, \ldots, n$ of the surface S at points $\boldsymbol{x}_i$ with errors ε_i (i. e. a regression model)

$$y_i = r(\boldsymbol{x}_i) + \varepsilon_i,$$

ε_i are centered independent Gauss variables with $\operatorname{var}\varepsilon_i = \sigma^2$. The objective of statistical analysis is to estimate unknown parameters w_{km}, compute the variance of estimates and also of prediction $r(\boldsymbol{x})$ at a point $\boldsymbol{x}$, based on these estimates. We shall use the following approach: First, functions r_k will be estimated. Then $\widehat{r}_k(z)$ will be propagated to whole X by blending given in expression (4).

Again, denote by $\boldsymbol{x}_i(k)$ projections of points $\boldsymbol{x}_i$ to lines L_k. Let P_{ki} be again the weight of $\boldsymbol{x}_i$ w.r. to line L_k. Further, denote $\boldsymbol{y} = (y_1, \ldots, y_n)'$, for each k denote Φ_k a matrix $n \times M_k$ with elements $\varphi_{mk}(\boldsymbol{x}_i(k))$, P_k the $\operatorname{diag}\{P_{ki}\}$ matrix $n \times n$ and G_k $\operatorname{diag}\left\{(P_{ki})^{-1/2}\right\}$ matrix. Let us eventually omit indices i such that $P_{ki} = 0$. A natural estimator of parameters $\boldsymbol{w}_k$ is then the result of weighted least squares method,

$$\widehat{\boldsymbol{w}}_k = \left(\Phi_k' P_k \Phi_k\right)^{-1} \Phi_k' P_k \boldsymbol{y},$$

corresponding to a regression model $\boldsymbol{y} = \Phi_k \boldsymbol{w}_k + G_k \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)'$. It follows that $\widehat{\boldsymbol{w}}_k = \boldsymbol{w}_k + \boldsymbol{A}_k \cdot \Phi_k' P_k G_k \boldsymbol{\varepsilon}$, with $\boldsymbol{A}_k = \left(\Phi_k' P_k \Phi_k\right)^{-1}$, therefore $\mathrm{E}\,\widehat{\boldsymbol{w}}_k = \boldsymbol{w}_k$ and

$$\operatorname{var}(\widehat{\boldsymbol{w}}_k) = \boldsymbol{A}_k \cdot \sigma^2.$$

In general, we can consider one set of weights, $R_k(\boldsymbol{x})$, for blending functions y_k, and a different set, $P_k(\boldsymbol{x})$, as weights used in the estimation procedure. Naturally, it is possible to set $P_k(\boldsymbol{x}) \equiv R_k(\boldsymbol{x})$. As regards the large sample (asymptotic) behavior of the procedure, it is natural to adopt the approach common in statistical smoothing methods. Namely, as n increases, the width of kernels is controlled by a decreasing window-width, while simultaneously the number of lines, K, should increase, both with a proper rate. In such a way, the consistency of estimation procedure can be guaranteed.

3.2 Prediction

Now, let $\boldsymbol{x}$ be an arbitrary point from X, different from all $\boldsymbol{x}_i$. Predicted value suggested by (4), where parameters $\boldsymbol{w}_k$ are substituted by their estimates, is given as

$$\widehat{r}(\boldsymbol{x}) = \sum_{k=1}^{K} R_k(\boldsymbol{x}) \cdot \widehat{\boldsymbol{w}}_k' \boldsymbol{\varphi}_k(\boldsymbol{x}(k)),$$

while its 'true' value (which we do not know) would be $y(\boldsymbol{x}) = r(\boldsymbol{x}) + \varepsilon_x$, where $r(\boldsymbol{x})$ is given by (4), ε_x is the same random variable like all ε_i, independent on them. We are interested in the difference

$$\widehat{r}(\boldsymbol{x}) - r(\boldsymbol{x}) = \sum_{k=1}^{K} R_k(\boldsymbol{x}) \left(\widehat{\boldsymbol{w}}_k - \boldsymbol{x}_k\right)' \boldsymbol{\varphi}_k(\boldsymbol{x}(k)).$$

Let us denote by $\boldsymbol{w} = (\boldsymbol{w}_1', \boldsymbol{w}_2', \ldots, \boldsymbol{w}_K')'$ the vector of length $M \times 1$, where $M = \sum_{k=1}^{K} M_k$ containing all parameters, similarly for $\widehat{\boldsymbol{w}}$. Further,

$$\boldsymbol{b}(\boldsymbol{x}) = \left(R_1(\boldsymbol{x})\,\boldsymbol{\varphi}_1(\boldsymbol{x})', R_2(\boldsymbol{x})\,\boldsymbol{\varphi}_2(\boldsymbol{x})', \ldots, R_K(\boldsymbol{x})\,\boldsymbol{\varphi}_K(\boldsymbol{x})'\right)'$$

is also $M \times 1$ vector, depending on $\boldsymbol{x}$. Then, we obtain that

$$\widehat{r}(\boldsymbol{x}) - r(\boldsymbol{x}) = \boldsymbol{b}'(\boldsymbol{x})\,(\widehat{\boldsymbol{w}} - \boldsymbol{w}).$$

Its expectation is zero, while its variance equals

$$\mathrm{var}\,(\widehat{r}(\boldsymbol{x}) - r(\boldsymbol{x})) = \boldsymbol{b}'(\boldsymbol{x}) \cdot \mathscr{V} \cdot \boldsymbol{b}(\boldsymbol{x}),$$

where $\mathscr{V}$ is the $M \times M$ covariance matrix $\mathrm{cov}(\widehat{\boldsymbol{w}} - \boldsymbol{w})$. It is, naturally, symmetric, positive definite almost surely, and composed from blocks $V_{k\ell}$, $k, \ell = 1, \ldots, K$, of dimension $M_k \times M_k$:

$$\begin{aligned} V_{k\ell} &= \mathrm{E}\left\{(\widehat{\boldsymbol{w}}_k - \boldsymbol{w}_k) \cdot (\widehat{\boldsymbol{w}}_\ell - \boldsymbol{w}_\ell)'\right\} = \\ &= \mathrm{E}\left\{\boldsymbol{A}_k \Phi_k' P_k G_k\, \boldsymbol{\varepsilon}\,\boldsymbol{\varepsilon}'\, G_\ell P_\ell \Phi_\ell \boldsymbol{A}_\ell\right\} = \boldsymbol{A}_k \Phi_k' P_k G_k G_\ell P_\ell \Phi_\ell \boldsymbol{A}_\ell \cdot \sigma^2. \end{aligned}$$

"Diagonal" blocks are then $V_{kk} = \boldsymbol{A}_k \cdot \sigma^2$. Finally, the difference $\widehat{r}(\boldsymbol{x}) - y(\boldsymbol{x})$ has the expectation zero, too, and variance $\boldsymbol{b}'(\boldsymbol{x})\,\mathscr{V}\,\boldsymbol{b}(\boldsymbol{x}) + \sigma^2$.

As the matrix $\mathscr{V}$ depends on the design of observed points $\boldsymbol{x}_i$, the variability of prediction at a point $\boldsymbol{x}$ depends on information in its neighborhood. We thus, in the following example, see the same phenomenon as in Example 1, i.e. a growing uncertainty of model in sparse data regions.

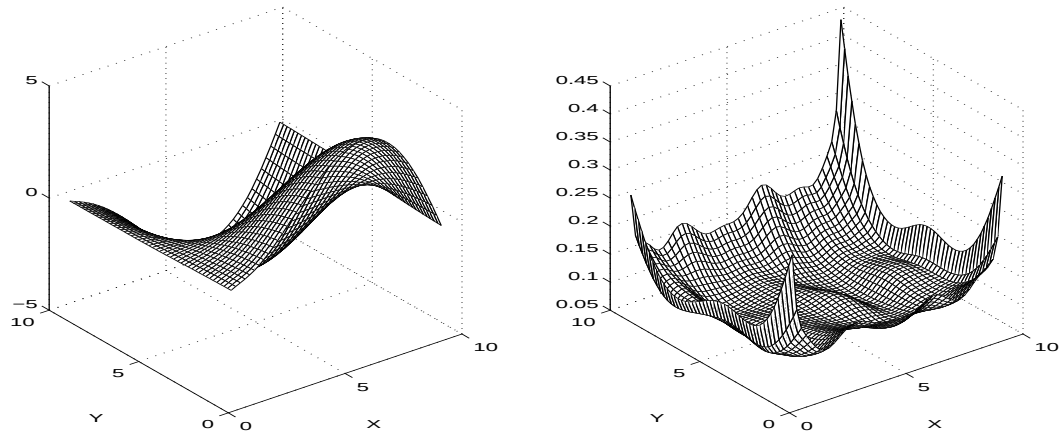
The model considered here contains $M = \sum_{k=1}^{K} M_k$ parameters. In a standard regression setting the number of observations should be $n > M$ to guarantee a reliable estimation. Notice that in the Gordon surface model $n > \max_k M_k$ suffices.

3.3 Analysis of Prediction Variance

As the direct computation of variance of prediction is here rather complicated (also due the presence of weighting functions) we shall prefer "empirical" illustration utilizing the following example.

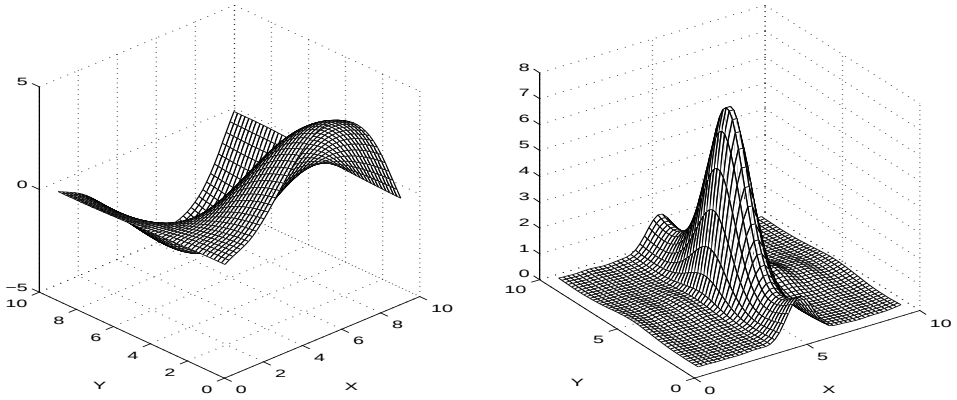
Example 2. The data of size $n = 200$ were generated from the function $r(\boldsymbol{x}) = x_1 \cdot sin(x_1/3) + sin(x_2/2 + 1)$ and additional standard Gauss random noise. For estimation, we fixed all 'nonlinear' components of the Gordon surface, namely: Number of lines, K, was fixed to 7, they were located horizontally, equidistantly in $(0, 10)$. To fit 1-dimensional curves along these lines, the cubic B-splines were used. For them, 5 equidistant knots were placed along each line. The same Gauss kernels were used both for weighting and blending.

In order to evaluate variability of estimates, the analysis was repeated $N = 200$ times. Figure 3 shows the mean and variance from 200 surfaces, estimated from full data covering the whole region $(0, 10) \times (0, 10)$. Then, Figure 4 displays the mean and variance of 200 surfaces estimated from data with missing values in $(4, 6) \times (4, 6)$ square. There is no significant differences between the means, they correspond, more-less, to the 'true' surface $r(\boldsymbol{x})$. However, the large variance in Figure 4 is the warning that one cannot rely just on estimated trend, that the inference has to be accompanied by the analysis of variance (and confidence, concerning the parameters). Notice also that the peak of variance has the vertical direction (in $X \times Y$ plane), i. e. orthogonal to lines L_k. It is caused by mixing of curves constructed along the lines, this mixing propagates the peak in the x direction.



Source: Own

Fig. 3. *The mean (left) and variance (right) estimated from the data covering uniformly the whole region* $(0,10) \times (0,10)$



Source: Own

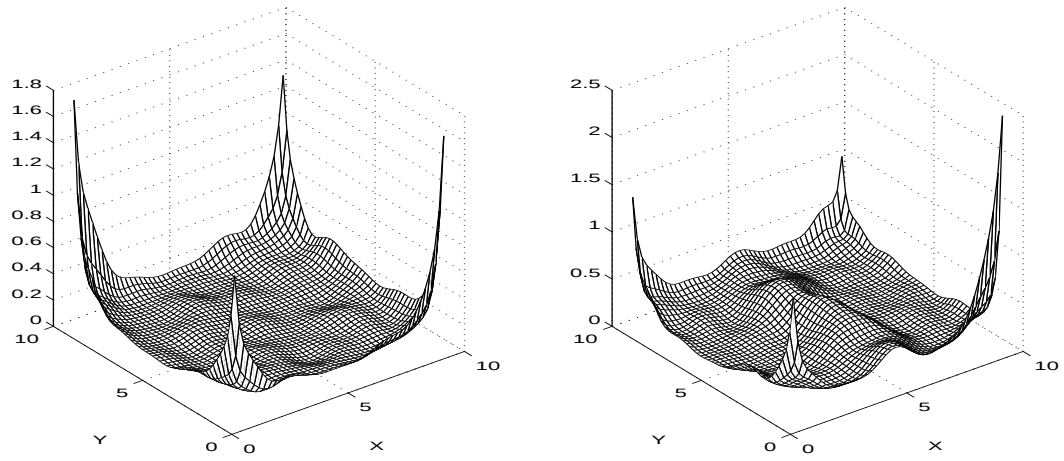
Fig. 4. *The mean (left) and variance (right) estimated from the data with missing values in region* $(4,6) \times (4,6)$

3.4 Comparison with Projection Pursuit

The same data as in Example 2 were repeatedly (with $N = 200$ repetitions) analyzed with the projection pursuit method. Again, the structure of model was fixed, we selected just four angles equidistantly inside $(0,\pi)$, each projection was fitted by B-splines with 5 inner knots located equidistantly between minimum and maximum value of projected $\boldsymbol{x}$-s. Such a selection was quite satisfactory, resulting surface fitted well (not worse than the Gordon surface shown on preceding figures). Figure 5 shows estimated variance of results (i.e. from 200 repetitions), left plot corresponds to full data, the right plot again to data with a gap in the region $(4,6) \times (4,6)$. The variance in the second case is still significant, however much lower compared to results of Gordon model application. It is probably caused by the structure of core lines (parallel lines in Gordon construction, a rosette in projection pursuit).

Conclusion

The objective of the paper was to examine several ways how to construct statistical model of 2-dimensional surfaces and to study advantages and problems of presented methods, from



Source: Own

Fig. 5. *Estimated variance of projection pursuit regression, from full data (left), from the data with missing values in region* $(4,6) \times (4,6)$ *(right)*

the point of their accuracy, flexibility, and also computational effort. The main attention was focused on the method of Gordon surface construction, and to the study of relationship between the design of data (input variables), selected set of localized functional units, and resulting variance of model estimate and prediction.

Acknowledgments

This work was supported by the Grant No. 209/10/2045 of the GA CR.

Literature

[1] ANDĚL, J.: *Foundantions of Mathematical Statistics (in Czech: Základy matematické statistiky)*. Matfyzpress, Praha, 2005.

[2] BISHOP, C.: *Neural Networks for Pattern Recognition*. Cambridge Univ. Press, Cambridge, 1992.

[3] DE BOOR, C.: *A Practical Guide to Splines*. Springer Verlag, Berlin, 1978.

[4] FRIEDMAN, J.H.: Multivariate adaptive regression splines, with Discussion and Rejoinder. *Annals Statist.* 19, 1991, pp. 1–141.

[5] GAMERMAN, D.: *Markov Chain Monte Carlo*. Chapman and Hall, New York, 1997.

[6] GORDON, W.J.: Spline-blended surface interpolation through curve networks. *Journal of Mathematics and Mechanics* 18, 1969, pp. 931–952.

[7] HUBER, P.J.: Projection pursuit. *Annals Statist.* 13, 1985, pp. 435–475.

[8] VOLF, P.: MCMC methods of randomized optimization and data analysis. In: *Proceedings of the ICPM 2007*, TU Liberec, 2007, pp. 123-130.

[9] WOLD, S.: Spline functions in data analysis. *Technometrics* 16, 1974, pp. 1–11.

doc. Petr Volf, CSc.

O DVOU FLEXIBILNÍCH METODÁCH PRO DVOUROZMĚRNOU REGRESNÍ ANALÝZU

Práce je věnována neparametrickému statistickému modelování dvourozměrných ploch na základě pozorovaných dat, tj. dvourozměrné regresní analýze. Model je konstruován jako kombinace jednoduchých bázových funkcí, jako jsou spliny, gaussiány apod. Nevýhodou takového přístupu je potřeba odhadnout velké množství parametrů. Budou ukázány dva přístupy, které počet potřebných parametrů redukují. A to metoda zvaná projection pursuit, která je statistikům dobře známá, a pak méně známá metoda konstrukce tzv. Gordonovy plochy. Zároveň také ukážeme, jak nepřítomnost dat v určité oblasti zvyšuje variabilitu modelu a tedy i neurčitost predikce. Metody budou předvedeny na umělých příkladech.

ÜBER ZWEI FLEXIBLE METHODEN FÜR EINE ZWEIDIMENSIONALE REGRESSANALYSE

Diese Arbeit befasst sich mit einer nichtparametrischen statistischen Modellierung zweidimensionaler Flächen auf Grundlage beobachteter Daten, d. h. mit der zweidimensionalen Regressanalyse. Das Modell ist konstruiert als Kombination einfacher Basenfunktionen wie z. B. Splines, Gaussian'sche Funktionen u. ä. Als Nachteil eines solchen Ansatzes erweist sicht die Notwendigkeit, eine große Menge Parameter zu schätzen. Es werden zwei Ansätze gezeigt, welche die Anzahl der erforderlichen Parameter reduziert, und zwar projection pursuit, welche den Statistikern wohl bekannt ist, und dann die weniger bekannte Konstruktionsmethode der Gordon-Flächen. Gleichzeitig zeigen wir, wie die Abwesenheit von Daten auf einem bestimmten Gebiet die Variabilität des Modells und damit auch die Unbestimmtheit der Prädiktion erhöht. Die Methoden werden an künstlichen Beispielen vorgeführt.

O DWU ELASTYCZNYCH METODACH DO DWUWYMIAROWEJ ANALIZY REGRESJI

Artykuł poświęcony jest nieparametrycznemu statystycznemu modelowaniu powierzchni dwuwymiarowych na podstawie obserwowanych danych, tj. dwuwymiarowej analizie regresji. Model skonstruowano jako połączenie prostych funkcji bazowych, takich jak splajny, gaussiany itp. Wadą takiego podejścia jest konieczność oszacowania dużej liczby parametrów. Pokazano dwa podejścia ograniczające liczbę niezbędnych parametrów. To metoda nazywana projection pursuit, która jest dobrze znana statystykom oraz mniej znana metoda konstrukcji tzw. powierzchni Gordona. Jednocześnie pokazano, jak brak danych w pewnym obszarze zwiększa zmienność modelu, czyli także niepewność prognozowania. Metody zaprezentowano na sztucznych przykładach.

LIST OF AUTHORS

Name	E-Mail
Mgr. Daniela Bímová, Ph.D.	daniela.bimova@tul.cz
RNDr. Daniela Bittnerová, CSc.	daniela.bittnerova@cz
RNDr. Dana Černá, Ph.D.	dana.cerna@tul.cz
Prof. Dr. hab. Andrzej Cegielski	a.cegielski@wmie.uz.zgora.pl
prof. RNDr. Vít Dolejší, Ph.D. DSc.	dolejsi@karlin.mff.cuni.cz
prof. Alexandre Ern, Ph.D.	ern@cermics.enpc.fr
RNDr. Václav Finěk, Ph.D.	vaclav.finek@tul.cz
Mahmoud Gad, M.Sc.	Mahmoud_Attya_Or@yahoo.com
Prof. Dr. rer. nat. habil. Christian Grossmann	christian.grossmann@tu-dresden.de
doc. Ing. Milan Hokr, Ph.D.	milan.hokr@tul.cz
RNDr. Jiří Hozman, Ph.D.	jiri.hozman@tul.cz
RNDr. Václav Kučera, Ph.D.	vaclav.kucera@email.cz
Ing. Mgr. Jan Lamač	jan.lamac@centrum.cz
RNDr. Ctirad Matonoha, Ph.D.	matonoha@cs.cas.cz
Ing. Ivo Matoušek, Ph.D.	ivo.matousek@tul.cz
Ing. Josef Novák, Ph.D.	josef.novak@tul.cz
Zdena Ondračková	zdenka.ondrackova@tul.cz
Gerta Plačková, prom. mat.	gerta.plackova@tul.cz
Prof. Dr. rer. nat. habil. Hans–Göerg Roos	hans-goerg.roos@tu-dresden.de
RNDr. Petr Salač, CSc.	petr.salac@tul.cz
RNDr. Martina Šimůnková, Ph.D.	martina.simunkova@tul.cz
RNDr. Miloslav Vlasák, Ph.D.	vlasak@karlin.mff.cuni.cz
Ing. Martin Vohralík, Ph.D.	vohralik@ann.jussieu.fr
doc. Petr Volf, CSc.	petr.volf@tul.cz

LIST OF REVIEWERS

Name	Work Location
Aneja Ritu, Prof.	Geogia State University
Andrášová Hana, PaedDr., Ph.D.	Jihočeská univerzita v Českých Budějovicích
Anchor John R., Dr.	University of Huddersfield
Antoch Jaromír, Prof., RNDr., CSc.	Matematicko-fyzikální fakulta UK v Praze
Bachmann Pavel, Ing., Ph.D.	Univerzita Hradec Králové
Baraniecka Anna, Dr.	Uniwersytet Ekonomiczny we Wroclawiu
Barči Tomáš, PhDr., Ing., Ph.D.	EGAP, a.s., Praha
Barták Miroslav, PhDr., Ph.D.	Univerzita J. E. Purkyně v Ústí nad Labem
Bejrová Martina, Ing., Ph.D.	ŠKODA AUTO, a.s.
Betáková Lucie, doc., PhDr., MA, Ph.D.	Jihočeská univerzita v Českých Budějovicích
Biniek Kamila (Jeleńska), Mgr	Karkonoska państwowa szkoła wyższa w Jeleniej Górze
Blechová Beáta, Ing., Ph.D.	Slezská univerzita v Opavě
Blin Jutta, Prof. Dr. phil.	Hochschule Zittau/Görlitz
Borozdina, Olga, Dr.	Univerzita St. Petersburg
Brauweiler Jana, Dr. rer. pol.	Internationales Hochschulinstitut Zittau
Budaj Pavol, Ing., Ph.D.	Katolícka univerzita v Ružomberku
Bureš Vladimír, doc., Ing., Ph.D.	Univerzita Hradec Králové
Busch-Lauer Ines Andrea, Prof., Dr.	Fachhochschule Zwickau
Čech Jaroslav, Prof., Ing., CSc.	Vysoké učení technické v Brně
Černá Dana, RNDr., Ph.D.	Technická univerzita v Liberci
Čiháková Silvia Aguilar, Ing., Ph.D.	Technická univerzita v Liberci
Daněk Ladislav, doc., Ing., CSc.	Vysoké učení technické v Brně
Delakowitz Bernd, Prof., Dr. rer. nat	Hochschule Zittau/Görlitz
Dipayan Das, Prof., Ph.D.	IIT Delhi
Domosławski Zbigniew, Prof., Dr. hab. N. Med.	Karkonoska państwowa szkoła wyższa w Jeleniej Górze
Doucek Petr, Prof., Ing., CSc.	Vysoká škola ekonomická v Praze
Dynybyl Vojtěch, Prof., Ing., CSc.	ČVUT Praha
Eck Vladimír, doc., Ing, CSc.	ČVUT Praha
Felixová Kateřina, Ing., Ph.D.	Univerzita J. E. Purkyně v Ústí nad Labem
Ficek Jaromír, Ing., Ph.D.	VÚTS, a.s., Liberec

Name	Work Location
Fielko Eva, Ing., Ph.D.	Metropolitní univerzita v Praze
Finěk Václav, RNDr., Ph.D.	Technická univerzita v Liberci
Fliegel Vítězslav, doc., Ing., CSc.	Technická univerzita v Liberci
Gerstlberger Wolfgang, Univ.-Prof., Dr. rer. pol. habil.	University of Southern Denmark
Griebel Bernd, Prof., Dr. phil.	Hochschule Zittau/Görlitz
Grossmann Christian, Prof., Dr. rer. nat. habil.	Technical University Dresden
Hájek Ladislav, Prof., Ing., CSc.	Univerzita Hradec Králové
Harland Peter E., Prof., Dr.	Internationales Hochschulinstitut Zittau
Herzig Ingo, M.A., PhDr.	Technická univerzita v Liberci
Hes Aleš, doc., Ing., CSc.	Česká zemědělská univerzita v Praze
Heßberg Silke, Prof., Dr.-Ing.	Westsächsische Hochschule Zwickau
Hinke Jana, Ing., Ph.D.	Západočeská univerzita v Plzni
Hlavatý Ivo, doc., Ing., Ph.D.	Technická univerzita Ostrava
Hokr Milan, doc., Ing., Ph.D.	Technická univerzita v Liberci
Homišin Jaroslav, Prof., Ing., CSc.	Technická univerzita v Košiciach
Holá Jana, Ing., Ph.D.	Univerzita Pardubice
Holík Martin, Mgr. Ph.D.	OSVČ, Praha
Honců Jan, Prof., Ing., CSc.	Technická univerzita v Liberci
Hortel Milan, Ing., DrSc.	Akademie věd ČR, Praha
Hozman Jiří, RNDr., Ph.D.	Technická univerzita v Liberci
Hruš Miroslav, Ing., CSc.	Odborný poradce, Liberec
Hyžík Jaroslav, Prof., Ing., CSc.	Technická univerzita v Liberci
Chocholoušková Hana, Mgr.	Státní archiv Liberec
Ircingová Jarmila, Ing., Ph.D.	Západočeská univerzita v Plzni
Jáčová Helena, PhDr., Ing., Ph.D.	Technická univerzita v Liberci
Jihlavec Jan, Mgr., DiS.	Technická univerzita v Liberci
Jílková Jiřina, Prof. Ing., CSc.	Univerzita J. E. Purkyně v Ústí nad Labem
Jirčíková Eva, Ing., Ph.D.	Univerzita Tomáše Bati ve Zlíně
Jirman Pavel, Ing.	Technická univerzita v Liberci
Kala Tomáš, Ing., DrSc., DBA.	Univerzita Hradec Králové
Kasper Tomáš, Doc., PhDr., Ph.D.	Technická univerzita v Liberci
Kellner Jan, Ing., Ph.D.	KPMG Česká republika, s.r.o.

Name	Work Location
Klápšťová Květoslava, Mgr., Ph.D.	Technická univerzita v Liberci
Klíma Radek, Ing.	Cadence Innovation s.r.o., Liberec
Knápková Adriana, Ing., Ph.D.	Univerzita Tomáše Bati ve Zlíně
Koláčková Ludmila, Mgr.	Univerzita obrany Brno
Kovárník Jaroslav, Ing., Ph.D.	Univerzita Hradec Králové
Kretschmar Gerlinde, Prof., Dr.-Ing.	Hochschule Zittau/Görlitz
Kučera Václav, RNDr., Ph.D.	Univerzita Karlova v Praze
Kurek Robert, Dr.	Uniwersytet Ekonomiczny we Wroclawiu
Ładysz Jerzy, Dr.	Uniwersytet Ekonomiczny we Wroclawiu
Lachout Martin, PhDr., Ph.D.	Metropolitní univerzita Praha
Landorová Anděla, Prof., Ing., CSc.	Vysoká škola obchodní v Praze
Lässig, Jörg, Prof., Dr.	Hochschule Zittau/Görlitz
Lizák Pavol, doc., Ing., Ph.D.	Trenčianska univerzita Alexandra Dubčeka
Lori Willfried, Prof., Dr.-Ing.	Westsächsische Hochschule Zwickau
Marouśková Marie, Prof., PhDr., CSc.	Univerzita J. E. Purkyně v Ústí nad Labem
Matonoha Ctirad, RNDr., Ph.D.	Akademie věd ČR, v.v.i. Praha
Mlýnek Jaroslav, doc., RNDr., CSc.	Technická univerzita v Liberci
Modrlák Osvald, doc., Ing., CSc.	Technická univerzita v Liberci
Mohelská Hana, doc., Ing., Ph.D.	Univerzita Hradec Králové
Mošová Vratislava, RNDr., CSc.	Moravská vysoká škola, Olomouc, o.p.s.
Mráz Jan, Ing., Ph.D.	EGAP, a.s., Praha
Mrkvička Tomáš, doc. RNDr., Ph.D.	Jihočeská univerzita v Českých Budějovicích
Müller Hardy, Prof., Dr.	Westsächsische Hochschule Zwickau
Müller Miloš, Ing., Ph.D.	LENAM s.r.o., Liberec
Najzar Karel, doc. RNDr., CSc.	Univerzita Karlova v Praze
Nehls Uwe, Prof., Dr.- Ing.	FH Oldenburg
Neuhoff Antje, M.A.	Technische Universität Dresden
Norková, Alena, Mgr.	Univerzita J. E. Purkyně v Ústí nad Labem
Nouza Jan, Prof. Ing. CSc.	Technická univerzita v Liberci
Nováková Kateřina, Ing.	Česká školní inspekce Liberec
Opa Miroslav, Ing., Ph.D.	Demoautoplast, s.r.o., Čelákovice
Ortová Martina, Ing., Ph.D.	Technická univerzita v Liberci
Papáček Štěpán, Ing., Ph.D.	Jihočeská univerzita v Českých Budějovicích

Name	Work Location
Paseková Marie, doc., Ing., Ph.D.	Univerzita Tomáše Bati ve Zlíně
Pavelka Tomáš, Ing., Ph.D.	Vysoká škola ekonomická v Praze
Pawłowski Maciej, Dr., Inż.	Politechnika Wrocławska
Pełczyńska Marzena, M.D., Ph.D.	Karkonoska państwowa szkoła wyższa w Jeleniej Górze
Pełka Marcin, Mgr.	Uniwersytet Ekonomiczny we Wrocławiu
Pešková Radka, Ing., Ph.D.	VŠEM Praha
Picek Jan, prof., RNDr., CSc.	Technická univerzita v Liberci
Piotrowski Przemyslaw, Mgr.	Karkonoska państwowa szkoła wyższa w Jeleniej Górze
Pištěk Luděk, Ing.	TOS Varnsdorf
Procházka Martin, Ing.	Okresní hospodářská komora Liberec
Radzik Tadeusz, Prof., Dr.	Karkonoska państwowa szkoła wyższa w Jeleniej Górze
Rahmanová Šárka, Ing., MBA, Ph.D.	Aareal Capital Corporation
Richter Ernst, Dr.-Ing.	Hochschule Zittau/Görlitz
Rozkovec Jiří, Mgr.	Technická univerzita v Liberci
Seidler Christine, Dr.	Internationales Hochschulinstitut Zittau
Schmidt Fritz Jochen, Prof., Dr.-Ing. habil.	Hochschule Zittau/Görlitz
Schöne Karin, M.A.	Technische Universität Dresden
Schönherr Jürgen, Prof., Dr.-Ing.	Hochschule Zittau/Görlitz
Skála Marek, Mgr. Ing., Ph.D.	Technická univerzita v Liberci
Skrbek Jan, doc. Dr., Ing.	Technická univerzita v Liberci
Sovová Ilona, Mgr.	Technická univerzita v Liberci
Stößel Bernd, Prof., Dr.-Ing.	Hochschule Zittau/Görlitz
Strahl Danuta, Prof., Dr. hab.	Uniwersytet Ekonomiczny we Wroclawiu
Svoboda Milan, PhDr., Ph.D.	Technická univerzita v Liberci
Szargot Maciej, Prof., Dr. hab.	Uniwersytet Humanistyczno przyrodniczny, Piotrków Trybunalski
Ševčík Ladislav, Prof., Ing., CSc.	Technická univerzita v Liberci
Štěpánek Libor, PhDr., Mgr., Ph.D.	Masarykova univerzita v Brně
Tettenborn Oliver, M.A.	Internationales Hochschulinstitut Zittau
Theilig Holger, Prof., Dr.-Ing. habil.	Hochschule Zittau/Görlitz
Trešl Jiří, Doc., Ing., CSc.	Vysoká škola ekonomická v Praze

Name	Work Location
Tvrdoň Michal, Mgr., Ing., Ph.D.	Slezská univerzita v Opavě
Urbánek Václav, doc., Ing., CSc.	Vysoká škola ekonomická v Praze
Vacek Jiří, Doc., Ing., CSc.	Technická univerzita v Liberci
Vašutová Jaroslava, doc., PaedDr., Ph.D.	Univerzita Karlova v Praze
Vítek Leoš, Doc., Ing., Ph.D.	Vysoká škola ekonomická v Praze
Vlasák Miloslav, RNDr., Ph.D.	Univerzita Karlova v Praze
Vlčková Kateřina, Mgr. et Mgr., Ph.D.	Masarykova univerzita Brno
Vogt Matthias-Theodor, Prof., Dr.	Hochschule Zittau/Görlitz
Vomáčková Helena, doc., Ing., CSc.	Univerzita J. E. Purkyně v Ústí nad Labem
Walter Johann Heinrich, Prof., Dr.-Ing., Dipl.-Math.	HS für Technik und Wirtschaft Dresden
Wierick Dieter, Prof. (em.), Dr.-Ing.	Hochschule Zittau/Görlitz
Will Markus, Dipl.-Ing. (FH)	Hochschule Zittau/Görlitz
Winnicki Tomasz, Prof. zw. Dr. hab. Inż.	Karkonoska państwowa szkoła wyższa w Jeleniej Górze
Winzeler Marius, Dr. Des., lic. phil.	Städtische Museen Zittau
Woldt Claudia, Dr.	Technische Universität Dresden
Worlitz Frank, Prof., Dr.-Ing.	Hochschule Zittau/Görlitz
Zaremba – Warnke Sabina, Dr.	Uniwersytet Ekonomiczny we Wroclawiu
Zelenka Jaroslav, Mgr.	Technická univerzita v Liberci
Žďánská Vladimíra, MUDr.	Privátní lékař, Liberec
Žižka Miroslav, doc., Ing., Ph.D.	Technická univerzita v Liberci

GUIDELINES FOR CONTRIBUTORS

Guidelines for contributors are written in the form of a model article, which is available as a Word document at http://acc-ern.tul.cz/images/journal/ACC_Journal_Template.docx .

EDITORIAL BOARD

Název časopisu (*Journal Title*)	ACC JOURNAL
Ročník (*vol./year/issue*)	XVIII (4/2012/Issue D)
Autor (*Author*)	kolektiv autorů (*composite authors*)
Vydavatel (*Published by*)	Technická univerzita v Liberci Studentská 2, Liberec 1, 461 17 IČO 46747885, DIČ CZ 46 747 885
Schváleno rektorátem TU v Liberci dne	20. 11. 2012, č. j. RE 102/12
Vyšlo (*Published*)	21.12. 2012
Počet stran (*Number of pages*)	173
Vydání (*Edition*)	první (*first*)
Evidenční číslo periodického tisku (*Registry reference number of periodical print*)	MK ČR E 18815
Tištěná verze ISSN (*ISSN printed version*)	1803-9782
Počet výtisků (*Number of copies*)	62 ks (*pieces*)
Adresa redakce (*Address of the editorial office*)	Technická univerzita v Liberci Akademické koordinační středisko v Euroregionu Nisa (ACC) Studentská 2, Liberec 1 461 17, Česká republika Tel. +420 485 352 318, Fax +420 485 352 229 e-mail: acc-journal@tul.cz http://acc-ern.tul.cz
Tiskne (*Print*)	Vysokoškolský podnik, s.r.o. Hálkova 6, Liberec 1 460 01, Česká republika

Upozornění pro čtenáře

Příspěvky v časopise jsou recenzovány a prošly jazykovou redakcí.

Readers' notice

Contributions in the journal have been reviewed and edited.

Předplatné

Objednávky předplatného přijímá redakce. Cena předplatného za rok je 900,- Kč mimo balné a poštovné. Starší čísla lze objednat do vyčerpání zásob (cena 200,- Kč za kus).

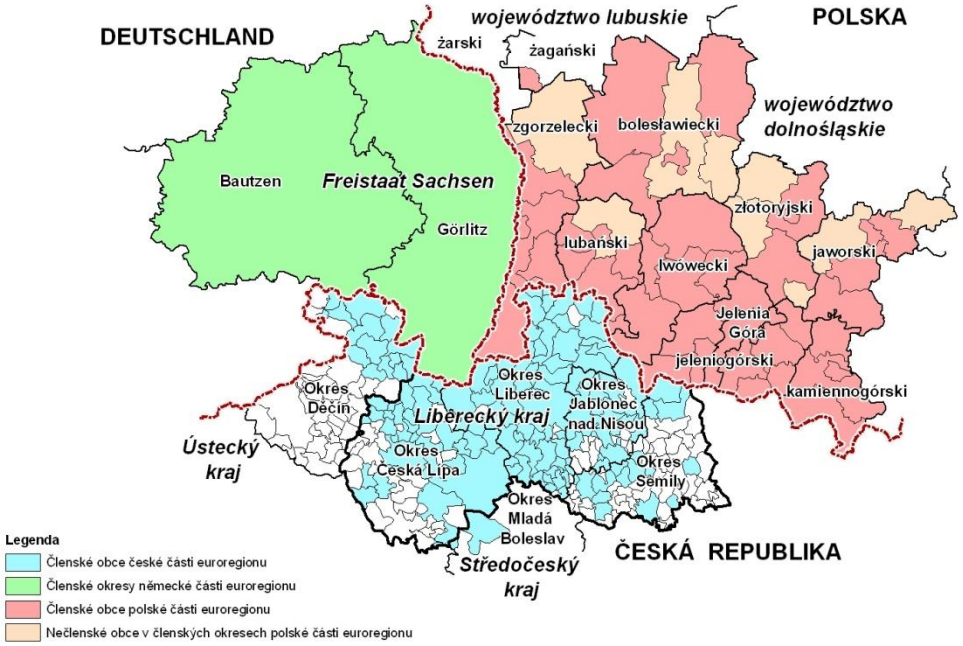
Subscription

Subscription orders must be sent to the editorial office. The price is 40 € a year excluding postage and packaging. It is possible to order older issues only until present supplies are exhausted (8 € an issue).

Časopis ACC JOURNAL vychází obvykle dvakrát ročně (červen, prosinec).

Two issues of ACC JOURNAL are published every year (June, December).

Euroregion Neisse-Nisa-Nysa k 31. 12. 2009



ISSN 1803-9782