International Conference

Programs and Algorithms of Numerical Mathematics 13

Photo courtesy of C. Fonville

in honor of Ivo Babuška's 80th birthday

under the auspices of Prof. Václav Pačes,
the President of the Academy of Sciences
of the Czech Republic

Mathematical Institute, Academy of Sciences,
Žitná 25, Prague, Czech Republic
May 28–31, 2006

PROCEEDINGS

Edited by
J. Chleboun, K. Segeth, T. Vejchodský

MATEMATICKÝ ÚSTAV AV ČR

Mathematical Institute
Academy of Sciences of the Czech Republic
Prague 2006

ISBN 80-85823-54-3
Matematický ústav AV ČR
Praha 2006

Contents

Preface

This book contains most papers presented at the international conference Programs and Algorithms of Numerical Mathematics (PANM) held in Prague, Czech Republic, May 28–31, 2006, in honor of Ivo Babuška's 80th birthday. It is the thirteenth volume in the series of the PANM proceedings.

The conference was organized by the Mathematical Institute of the Academy of Sciences of the Czech Republic (ASCR) and continued the previous PANM seminars (conferences) held in Alšovice, Bratříkov, Janov nad Nisou, Kořenov, Lázně Libverda, and Dolní Maxov in the period 1983–2004. The objective of this series of seminars has been to provide a forum for presentation and discussion of advanced topics, new approaches, and applications of computational methods; moreover, the participation of PhD students and young scientists has been encouraged.

The conference was honored by the presence of Ivo Babuška, who spent three days and four nights in Prague during his 2006 Europe Tour focused on conferences organized in Prague, Zurich, and London to celebrate his birthday, and who used to be a leading scientist in the Institute in the 1960s.

This year, the conference left its traditional Jizera Mountains locations, and was held on the Prague premises of the Institute, to make the participation of distinguished foreign guests easier or even possible. To further emphasize the significance of the conference, it was organized under the auspices of Professor Václav Pačes, president of the ASCR, who, moreover, presented the Honorary Medal "De scientia et humanitate optime meritis" (the ASCR highest distinction) to Ivo Babuška during a closed meeting in the course of the conference.

It would not have been possible to organize the conference without the financial support of the Czech Science Foundation, the Grant Agency of the ASCR, and the ASCR (project no. 201/04/1503, project no. A1019201, and Institutional Research Plan no. AV0Z10190503, respectively). The Organizing Committee included Jan Chleboun, Michal Křížek, Petr Přikryl, Karel Segeth, Alena Šolcová, and Tomáš Vejchodský.

More than 80 participants from the field took part in the conference. Although most of them came from Czech universities and the institutes of the ASCR, foreign scholars were also present (Canada, China, Finland, Germany, the Netherlands, United Kingdom, and the U.S.A.). They witnessed the presentation of the Honorary Medal of the Ministry of Education, Youth, and Sports to Ivo Babuška on the third day of the conference.

As regards the technical aspects of this book of proceedings, all the papers have been reproduced directly from materials submitted by the authors, but an attempt has been made to use a unified layout for each paper. We are indebted to Mrs. Hana Bílková for her effort in preparing the manuscripts for print and to Dr. Karel Horák and Mr. Ladislav Capanda for their technical help with printing the book.

The editors and organizers also wish to thank all scientists who peer reviewed the submitted manuscripts. In many cases, their comments and recommendations led to substantial improvements of the manuscripts.

Besides this PANM volume, a special issue of Applications of Mathematics (no. 3, 2007) consisting of selected papers presented at the conference will be published.

J. Chleboun, K. Segeth, T. Vejchodský

NUMERICAL MODELING OF FLOW AND POLLUTION DISPERSION OVER REAL TOPOGRAPHY[*]

Luděk Beneš, Karel Kozel, Ivo Sládek

1. Introduction

The Atmospheric Boundary Layer (ABL) is the lowest part of the atmosphere. Its thickness usually ranges from several hundred meters to approximately two kilometers. The air pollution resulting from rapid industrialization has become a serious environmental problem mainly in the North Bohemia region. In this contribution, the influence of several types of obstacles on dustiness of coal depot in open coal mine was numerically modeled.

2. Mathematical models

In our computations, the flow in ABL is assumed to be viscous, steady, incompressible, turbulent and indifferently stratified. Two different mathematical and numerical methods have been used for numerical simulations.

- **The full RANS model**

The first model is based on Reynolds Averaged Navier–Stokes equations. The governing equations are considered in the conservative, non-dimensional, and vector form:

$$W_t + F_x + G_y + H_z = (KR)_x + (KS)_y + (KT)_z + f_{\mathrm{v}}, \qquad (1)$$

where $F = (u,\ u^2{+}p,\ uv,\ uw,\ uC\,)^T$, $G = (v,\ vu,\ v^2{+}p,\ vw,\ vC\,)^T$, $H = (w,\ wu,\ wv,\ w^2+p,\ wC\,)^T$, $R = (0,\ u_x,\ v_x,\ w_x,\ C_x/\sigma_C)^T$, $S = (0,\ u_y,\ v_y,\ w_y,\ C_y/\sigma_C)^T$, $T = (0,\ u_z,\ v_z,\ w_z,\ C_z/\sigma_C)^T$. $W = (p/\beta^2,\ u,\ v,\ w,\ C)^T$ stands for the vector of unknown variables the pressure, three velocity components $V = (u, v, w)^T$, and the concentration of passive pollutant, respectively. Further f_{v} denotes the volume force, σ_C is the turbulent Prandtl's number, β artificial compressibility coefficient and finally K represents the turbulent diffusion coefficient, see equation (5). The artificial compressibility method is used for the numerical solution of this model.

- **Boussinesq equations**

The NS equations are simplified by the so called Boussinesq approximation. The instantenous values of the density, pressure and potential temperature can be decomposed into two parts: the large synoptic scale part denoted by subscript $_0$ and

[*] The financial support for the presented project is partly provided by the Research Plan MSM No. 6840770003 and GAČR 205/06/0727.

its perturbation denoted by $''$. Then the governing equations for the neutrally stratified flow can be rewritten in the following form

$$(\rho_0 u)_x + (\rho_0 v)_y + (\rho_0 w)_z = 0\,, \tag{2}$$

$$V_t + uV_x + vV_y + wV_z = -\frac{\nabla p''}{\rho_0} + \frac{1}{\rho_0}\Big\{[\rho_0 K V_x]_x + [\rho_0 K V_y]_y + [\rho_0 K V_z]_z\Big\} + f_{\mathrm{v}}\,. \tag{3}$$

The transport equations for the passive pollutant C is

$$C_t + uC_x + vC_y + wC_z = \left[\left(\frac{K}{\sigma_C}C_x\right)_x + \left(\frac{K}{\sigma_C}C_y\right)_y + \left(\frac{K}{\sigma_C}C_z\right)_z\right]. \tag{4}$$

2.1. Turbulence model

Closure of both systems of governing equations (1) and (2)–(4) is achieved by a simple algebraic turbulence model designed for ABL flow. The model is based on the Bousinesq hypothesis. The diffusion coefficient K has the following form in the dimensional case

$$K = \nu + \nu_T, \qquad \nu_T = l^2\sqrt{(u_z)^2 + (v_z)^2}, \tag{5}$$

where ν_T and ν are the turbulent and laminar viscosities. The mixing length l is according to Blackadar computed from

$$l = \frac{\kappa(z+z_0)}{1+\kappa(z+z_0)/l_\infty}, \qquad l_\infty = \frac{27\,|V_g|\,10^{-5}}{\lambda}, \tag{6}$$

where κ is the von Karman constant, λ denotes the Coriolis parameter, z_0 the roughness length, l_∞ denotes the mixing length for $z \to \infty$ and V_g is the geostrophic wind velocity at the upper boundary of the domain.

3. Numerical methods

We have solved the governing systems of equations with stationary boundary conditions and we suppose that we obtain the expected steady-state solution for $t \to \infty$. Structured non-orthogonal grids made of hexahedral (in 3D case) and quadrilateral (in 2D case) control cells are used.

3.1. Finite volume method

The finite volume method (cell-centered type) together with the 3–stage explicit Runge–Kutta time integration scheme have been applied to solve equation (1). For discretization of viscous fluxes, a second octohedral mesh was used.

The numerical method is theoretically second order accurate in space and time on orthogonal grids. In addition, it must be stabilized by the artificial viscosity term of fourth order to remove spurious oscillations in the flow-field due to sharp gradients of computed quantities and also due to the central differences used for the space discretization of convective terms.

3.2. Finite difference method

A semi-implicit finite difference scheme has been used for the model (2)–(4). The special combination of different nonsymmetric space discretizations at time level n and $n+1$ leads to the numerical scheme that is centered and second order both in space and time. In order to improve the convergence of this method for large Reynolds numbers the artificial viscosity terms either of the fourth or the second order are added. To discretize the governing system (2)–(4) we have constructed a non-orthogonal structured boundary–terrain fitted mesh.

3.3. Boundary conditions

Both models use the following boundary conditions.

- Inlet: $u = U_0(z/L)^\alpha$, $v = w = C = 0$, where L is vertical length of the domain and α is a power law exponent (we usually set $\alpha = 2/9$).
- Outlet: $u_x = v_x = w_x = C_x = 0$.
- Wall: the no-slip condition for the velocity components, $\partial C/\partial n = 0$.
- Top: $u = U_0$, $v = 0$, $\partial w/\partial z = \partial C/\partial z = 0$.
- Sides: periodic or non–periodic.

3.4. Validation of models

The first model (1) has been validated through the ERCOFTAC's test-case of fully developed channel flow over 2D polynomial-shaped hill mounted on a flat plate. The Almeida's experimental and the ERCOFTAC's $k-\varepsilon$ reference numerical data have been used for the comparison, see [5].

The second model (2)–(4) has been validated on the experimental and reference numerical data obtained by G.H. Kim [6]. Boundary layer type of flow over the sinusoidal 2D-single-hills of different shapes has been tested [1].

The results from both validation studies has shown very good agreement with the target data.

4. Numerical results

This practical problem is related to the flow over a surface coal field located in the open coal mine in the North Bohemia. This numerical study is a continuation of the project we have been solving since 2001 in cooperation with Brown Coal Research Institute in Most. The major task was to design a safety obstacle close to a coal depot in order to decrease the level of pollutant concentrations in the down stream region which is inhabited. Several types of obstacles as solid wall, protective tree line, forest block and shelter belt were tested.

The influence of the forest blocks and the protective walls on the dustiness of the coal depot has been studied on the real topography of the coal depot.

The model of a real 3D relief was created on the basis of the topographic data obtained by the Brown Coal Research Institute in Most. The whole topography has been divided into two parts. The computational Domain 1, (see Figs. 1,2) is 800 m long, 480 m wide and the upper side is at 1000 m. The coal depot has dimensions

$80 \times 20\,m$ and it is situated at the origin. For better resolution of the flow field close to the depot, the second Domain 2 $400 \times 240\,m$ was imbedded (see Fig. 2). The data obtained on Domain 1 were used as the boundary and initial conditions for the computations on Domain 2.

Both domains have been discretized using $100 \times 60 \times 40$ mesh cells, so the horizontal resolution is $8\,m$ on Domain 1 and $4\,m$ on Domain 2. Both grids are significantly thickened close to the ground with $\Delta_{z_{\min}} \approx 0.6\,m$. Two variants were computed in 3D: basic (without protective obstacles) and with two forest blocks situated before and behind the coal depot.

The **solid wall** was simulated by the column of a few cells. All the velocity components have been set to zero in all of these cells. For the **forest block**, the force vector $\vec{f_v}$ includes the specific aerodynamic force corresponding to the drag induced by the vegetation, i.e.

$$\vec{f_v} = (-r_h|V|u, -r_h|V|v, -r_h|V|w)^T. \tag{7}$$

Here the $r_h(z)$ denotes the total resistance parameter. The vertical profile of this parameter has been set-up in the following way:

$$r_h(z) = \begin{cases} rz/(0.75h) & \text{for} \quad 0 \le z/h \le 0.75, \\ r(1 - z/h)/(1 - 0.75) & \text{for} \quad 0.75 \le z/h \le 1.0, \end{cases} \tag{8}$$

where the drag coefficient value r is given a priori.

The other parameters are: mean free stream velocity $U = 10\,m/s$, roughness parameter $z_0 = 0.1\,m$ and power law exponent $2/9$ are used for the inlet velocity profile (Domain 1). The forest blocks are $10\,m$ high with the drag coefficient $r = 0.19$. The wall is $5\,m$ high.

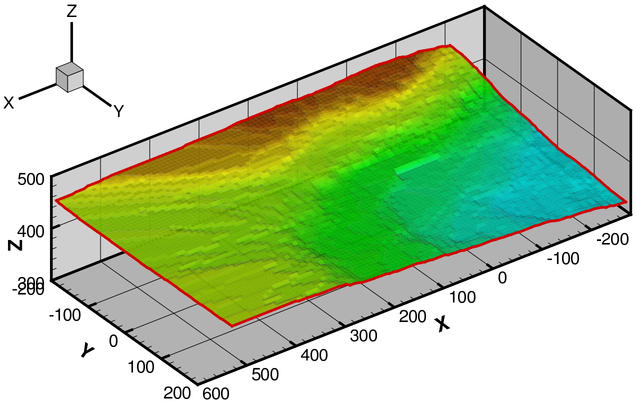


Fig. 1: *Topography of the mine–Domain 1.*

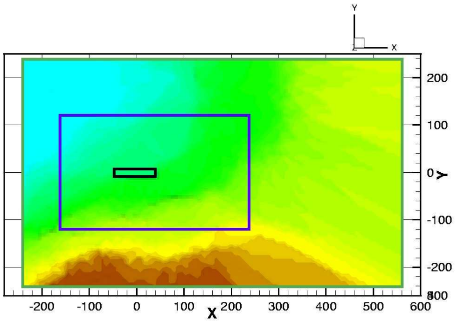


Fig. 2: *Computational Domain 1, Domain 2 (larger rectangle), coal depot (smaller rectangle).*

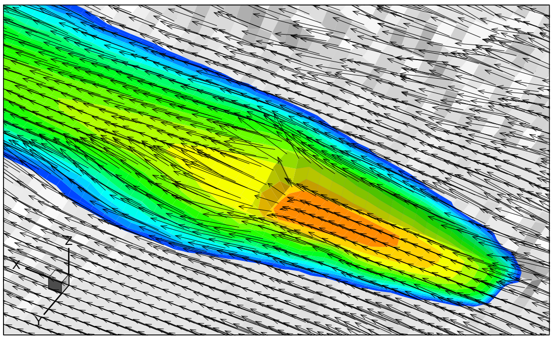


Fig. 3: *Velocity vectors close to the coal depot – basic situation. Colored by the concentration.*

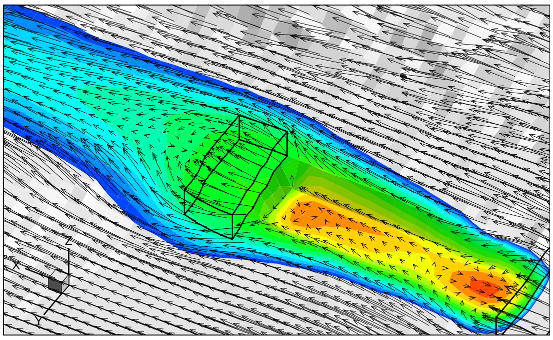


Fig. 4: *Velocity vectors close to the coal depot – situation with two forests. Colored by the concentration.*

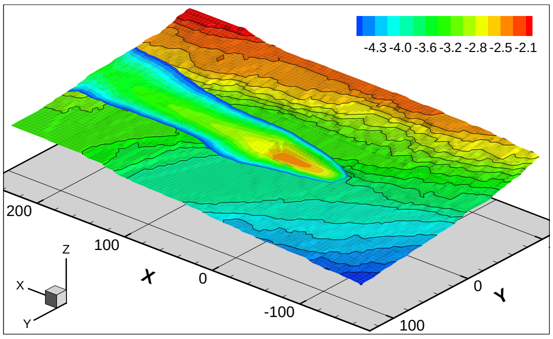


Fig. 5: *Concentration of the pollution in the logarithmic scale completed by altitude – basic situation.*

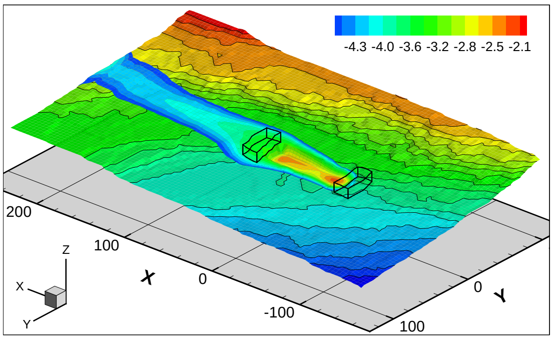


Fig. 6: *Concentration of the pollution in the logarithmic scale completed by altitude – situation with two forests.*

In Figs. 3 and 4 and Figs. 5 and 6 we can see the comparison of the flow field and the pollution dispersion in two different cases – basic and with two forests before and behind the coal depot. From these figures one can see considerable reduction of the dustiness in the second case. It is due to the significant deceleration of the flow behind the forest on the area of coal depot.

The majority of variants has been tested in 2D only. From Domain 1 the 2D middle cut ($y = 0$) was chosen. This cut was discretized by 800 × 40 cells (horizontal resolution $1\,m$), vertical distribution is the same as in 3D. Also the other computational parameters are the same as in 3D.

The seven different positions and combinations of walls and forests were computed in 2D case: basic – without protective obstacles (zak), with the forest block before (a) behind (b) and on both sides (ab) of the depot, and with the wall before (fa) behind (fb) and on both sides (fab).

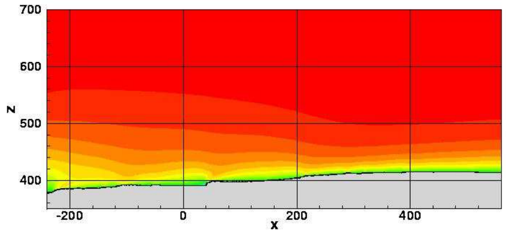


Fig. 7: *Basic variant – velocity component u.*

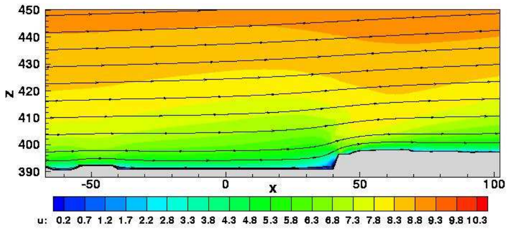


Fig. 8: *Basic var. – velocity component u with streamlines close to the coal depot.*

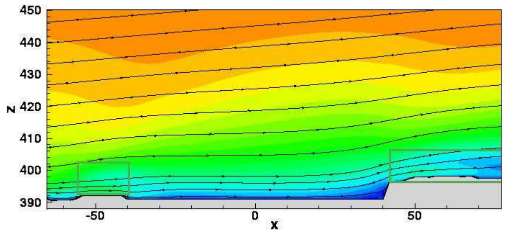


Fig. 9: *Variant ab – velocity component u with streamlines close to the c.d.*

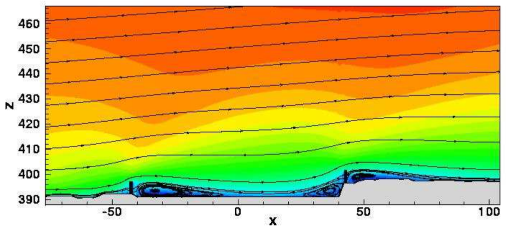


Fig. 10: *Variant fab – velocity component u with streamlines close to the c.d.*

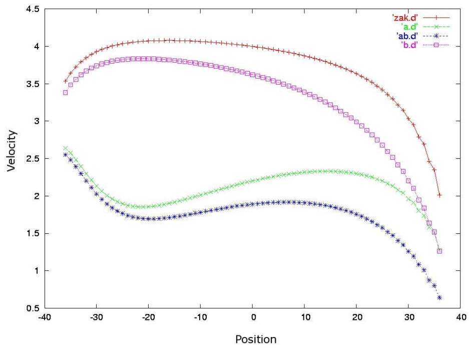


Fig. 11: *The longitudinal distribution of near-ground velocity in case of forest.*

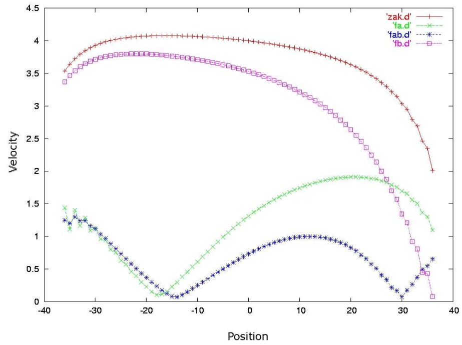


Fig. 12: *The longitudinal distribution of near-ground velocity in case of walls.*

Figs. 7–10 show the comparison of the basic variant with two different cases in 2D: with forest on both sides (ab) of the coal depot and also with a wall on both sides (fab). In Fig. 10 large recirculation zones behind the walls are shown. In contrast, the flow going through the forest is decelerated smoothly without recirculation, Fig. 9.

In our model, the source intensity is proportional to the vertical velocity gradient, and the mesh is uniform on the coal depot. Therefore the local source intensity is proportional to the ground velocity.

Fig. 11 and Fig. 12 shows the longitudinal distribution of near ground velocity for an obstacle of type forest block (left) and walls (right).

References

[1] L. Beneš, T. Bodnár, Ph. Fraunié, K. Kozel: *Numerical modelling of pollution dispersion in complex terrain.* In: G. Latini, C.A. Brebbia (eds), Air Pollution IX., Southampton, WIT Press, 2001, 85–94.

[2] T. Bodnár, Ph. Fraunié, K. Kozel, I. Sládek: *Numerical simulation of complex atmospheric boundary layer problems.* In: J.M. Redondo (ed.), ERCOFTAC Bulletin No. 60, March 2004, 5–12.

[3] T. Bodnár, I. Sládek, E. Gulíková: *Numerical simulation of wind flow in the vicinity of forest block.* In: S.N. Atluri (ed.), Advances in Computational & Experimental Engineering & Sciences. Forsyth, Tech Science Press, 2004, 554–559.

[4] E. Gulíková, T. Bodnár, V. Píša: *Improvement of numerical models for solution of dust air pollution.* In: J. Příhoda, K. Kozel (eds.), Colloquium FLUID DYNAMICS 2006, Prague, IT ASCR, 2005, 63–66.

[5] G.P. Almeida, D.F.G. Durao, M.V. Heitor: *Wake flows behind two dimensional model hills.* Exp. Thermal and Fluid Science **7**, 1992, 87–101.

[6] G.H. Kim, M.Ch. Lee, C.H. Lim, H.N. Kyong: *An experimental and numerical study on the flow over two-dimensional hills.* Journal of Wind Eng. and Industrial Aerodynamics **66**, 1997, 17–33.

NUMERICAL ANALYSIS OF MATHEMATICAL MODEL OF HEAT AND MOISTURE TRANSPORT IN CONCRETE AT HIGH TEMPERATURES[*]

Michal Beneš, Petr Mayer

Abstract

In this paper, we present a nonlinear mathematical model for numerical analysis of the behaviour of concrete subject to transient heating according to the standard ISO fire curve. This example allows us to analyse and better understand physical phenomena taking place in heated concrete (thermal spalling).

1. Balance equations of mathematical model

The behaviour of concrete at high temperature is dependent on its composite structure, on the physical and chemical composition of the cement paste, which is a highly porous, hygroscopic material. In the whole temperature range, the gas phase is a mixture of dry air and water vapour. Therefore, the moist concrete is modelled as a multiphase material.

The global multiphase system is treated within the framework of averaging theories starting from microscopic level and applying mass, area and volume averaging operators to the local form of governing equations.

The mathematical model consists of the following balance equations for the α-phase, in particular w, resp. g, resp. ga, resp. gw denotes the liquid phase, resp. the gas phase, resp. the dry air, resp. water vapour,

$$\begin{aligned} \frac{D^\alpha}{Dt}(\eta^\alpha\rho^\alpha) + (\eta^\alpha\rho^\alpha)\nabla\cdot\mathbf{v}^\alpha &= \hat{e}^\alpha_\beta, \quad \alpha,\beta = w,\, gw,\, ga, \quad \alpha\neq\beta, & (1)\\ (\rho C)\frac{\partial T}{\partial t} + (\rho C\mathbf{v})\nabla\cdot T - \nabla\cdot(\lambda\nabla T) &= -\dot{m}_{phase}\Delta h_{phase} + \dot{m}_{dehydr}\Delta h_{dehydr}, & (2)\end{aligned}$$

where η^α is the volume fraction of phase α,

$$\eta_w = \phi S_w, \quad \eta_g = \phi S_g, \quad S_w + S_g = 1,$$

where S_w, resp. S_g denotes the degree of water saturation, resp. the degree of gas saturation, ϕ is the porosity, ρ^α and $\mathbf{v}^\alpha$ denote the averaged density and mass-averaged velocity of the α-phase. The mass source term $\hat{e}^\alpha_\beta$ on the right-hand side represents

[*] This research has been supported by Ministry of Education, Youth and Sports of the Czech Republic, No. 1M6840770001 within the frame of research centre CIDEAS and VZ 03 CEZ MSM 6840770003.

exchange of mass with interfaces separating individual phases (phase changes), as well as the terms on the right hand side of (2) represent the energy required for evaporation of liquid water and the energy required for release of bound water by dehydration. The convection term $(\rho C\mathbf{v})\nabla \cdot T$ in equation (2) is ignored provided that the transfer of energy by convection is included in the empirical relationship for the thermal conductivity $\lambda = \lambda(T)$.

2. Boundary and initial conditions

The model consisted of a rectangular section of the concrete wall, 0.1 m thickness, exposed to transient heating from one side according to the standard ISO FIRE curve

$$T_\infty(t) = T_{ISO-FIRE}(t) = 345\log(8t+1) + 293.15, \quad [t] = \text{min}. \tag{3}$$

In the case of heat transfer through the boundary at normal temperatures, the boundary conditions correspond to the Newton's law of cooling (Neuman's conditions)

$$-(\rho_w \mathbf{v}_l \Delta h_{phase} - \lambda \nabla T).\mathbf{n} = 0, \tag{4}$$

$$-(\rho_{gw}\mathbf{v}_g + \rho_w \mathbf{v}_l + \rho_g \mathbf{v}^d_{gw}).\mathbf{n} = 0. \tag{5}$$

On the part of the boundary, where the high temperature is analyse, the radiative boundary conditions

$$\begin{aligned}(\rho_w \mathbf{v}_l \Delta h_{phase} - \lambda \nabla T).\mathbf{n} &= \alpha_c(T - T_\infty) + e\sigma(T^4 - T^4_\infty), & (6)\\ (\rho_{gw}\mathbf{v}_g + \rho_w \mathbf{v}_l + \rho_g \mathbf{v}^d_{gw}).\mathbf{n} &= \beta_c(\rho_{gw} - \rho_{gw\infty}), & (7)\end{aligned}$$

are of importance, where the terms on the right hand side of (6) represent the heat energy dissipated by convection and radiation to the surrounding medium, and the term on the ride hand side of (7) is the substance dissipated into the surrounding medium.

The initial conditions for concrete were set as follows: the uniform temperature $T = 293.15$ K, the uniform gas pressure 101325 Pa and the uniform capillary pressure 97.3 MPa, according to ρ_{gw} and ρ_{ga}.

3. Thermodynamic approach, constitutive relationships and material data

Dry air, water vapour and their mixture are assumed to behave as perfect gases, therefore Dalton's law and the Clapeyron equation are assumed as state equations. Water vapour pressure, p_{gw} is obtained from the Kelvin equation. As the constitutive equations for fluid phases (capillary water, gas phase) the multiphase Darcy's law has been applied.

Mathematical model of multiphase flow and heat transfer in concrete contains a several parameters and coefficients describing the properties of concrete and fluids: porosity $\phi = \phi(T)$, saturation $S = S(p_c)$, solid phase density $\rho = \rho(T)$, absolute permeability $\mathbf{K} = \mathbf{K}(p_g, T)$, relative permeability of gas phase $K_{rg} = K_{rg}(p_c, T)$, relative permeability of liquid phase $K_{rw} = K_{rw}(p_c, T)$, gas-phase dynamic viscosity $\mu_g = \mu_g(p_g, p_c, T)$, liquid phase dynamic viscosity $\mu_w = \mu_w(T)$, thermal capacity of the system $\rho C_p(T)$, thermal conductivity of the system $\lambda(T)$, enthalpy of vaporisation $\Delta h_{vap}(T)$. Formulas are given in detail in [3].

The relationship between capillary pressure and saturation in multiphase flow problems demonstrates memory effects and hysteresis. Differential equations with hysteresis have been the subject of studies, from the mathematical point of view, since 1960s. Hassanizadeh and Gray employ conservation laws for mass, momentum and energy, and the second law of thermodynamics in order to develop constitutive equations which describe two-phase flow in a porous medium (see [1], [2]). The following combination of terms contributes to the entropy production Λ:

$$T\Lambda = \ldots - \phi \dot{S}^w (p^g - p^w - p^c) + \ldots \geq 0.$$

For a linear theory, Hassanizadeh and Gray have suggested the relationship

$$p^g - p^w - p^c(S) + \tau(S)\dot{S} = 0,$$

where p^w, resp. p^g, designate the water, resp. the gas, pressure.

Under equilibrium condition without dynamic effects in the capillary relation the following definition of the capillary pressure can be used

$$p^c = p^g - p^w, \quad S_w = S_w(p^c). \tag{8}$$

In some particular cases, it is possible to use relation (8) even if the material system demonstrates hysteresis. For instance, in slow processes with monotonically decreasing (or increasing) saturation. Equation (8) is usually determined from experiments. In the literature several approximations of the relationship (8) have been suggested. In the present model the relationship

$$S_w(p_c) = S_r^w + (S_s^w - S_r^w)\left[1 + \left(\frac{p_c}{p_b^c}\right)^n\right]^{-m} \tag{9}$$

is employed. In (9) p_b^c denotes the air entry value, also referred to as bubbling pressure, which can be viewed as a characteristic pressure that has to be reached before the air actually enters the pores; m and n are empirical constants to fit the curves to experimental data.

A further step of this research is the influence of the dynamic or non-equilibrium effects and hysteresis, e.g. $\dot{S} \neq 0$, to hydro-thermal behavior of rapidly heated concrete.

4. Numerical algorithm

The space discretization of the energy conservation equation (2) is carried out by means of the finite element method ($h = 0.001$ m), we obtain the finite element model in the form

$$\mathbf{C}(\mathbf{T})\dot{\mathbf{T}} - \mathbf{K}(\mathbf{T})\mathbf{T} = \mathbf{f}(\mathbf{T}, \rho_{gw}, \rho_{ga}). \tag{10}$$

Time discretization of (10) is accomplished through an implicit difference scheme compared with $\mathbf{T}$ ($\Delta t = 1$ s)

$$[\mathbf{C}(\mathbf{T}_{n+1}) + \Delta t\mathbf{K}(\mathbf{T}_{n+1})]\,\mathbf{T}_{n+1} = \mathbf{C}(\mathbf{T}_{n+1})\mathbf{T}_n + \mathbf{f}(\mathbf{T}_{n+1}, \rho_{gw(n)}, \rho_{ga(n)}). \tag{11}$$

The Newton-Raphson method is applied to the nonlinear system (11) in the following iteration procedure: Let us denote

$$\Phi(\mathbf{T}^{(l)}_{n+1}) = \left[C_{ij}(\mathbf{T}^{(l)}_{n+1}) + \Delta t K_{ij}(\mathbf{T}^{(l)}_{n+1})\right]\mathbf{T}_{n+1} - C_{ij}(\mathbf{T}^{(l)}_{n+1})\mathbf{T}_n + f_i(\mathbf{T}^{(l)}_{n+1}, \rho_{gw(n)}, \rho_{ga(n)}),$$

then the solution at the end of the $(l+1)$st iteration is then given by

$$\mathbf{T}^{(l+1)}_{n+1} = \mathbf{T}^{(l)}_{n+1} - \mathbf{J}^{-1}_{\Phi}(\mathbf{T}^{(l)}_{n+1})\Phi(\mathbf{T}^{(l)}_{n+1}), \tag{12}$$

where $\mathbf{J}_\Phi$ is the (three-diagonal) Jacobi matrix.

Now we modify the dry air conservation equation and the water species conservation equation to the form

$$\begin{aligned}
\phi\frac{\partial}{\partial t}\left[(1-S)\rho_{ga}\right] + (1-S)\rho_{ga}\frac{\partial\phi_{hydr}}{\partial t} + \nabla.(\rho_{ga}\mathbf{v}_g) + \nabla.(\rho_g\mathbf{v}^d_{ga}) &= 0, && (13)\\
\phi\frac{\partial}{\partial t}\left[(1-S)\rho_{gw}\right] + (1-S)\rho_{gw}\frac{\partial\phi_{hydr}}{\partial t} + \nabla.(\rho_{gw}\mathbf{v}_g) + \nabla.(\rho_g\mathbf{v}^d_{ga}) &= \\
-\phi\frac{\partial}{\partial t}(S\rho_w) - S\rho_w\frac{\partial\phi_{hydr}}{\partial t} - \nabla.(\rho_w\mathbf{v}_l) - \frac{\partial}{\partial t}(\Delta m_{hydr}) & && (14)
\end{aligned}$$

with regard to Dalton's law and Clapeyron equations of state of perfect gases $\rho_g = \rho_{gw} + \rho_{ga}$ to the form

$$\phi\frac{\partial}{\partial t}\left[(1-S)\rho_{ga}\right] + (1-S)\rho_{ga}\frac{\partial\phi_{hydr}}{\partial t} + \nabla.(\phi(1-S)\rho_{ga}\mathbf{v}_g) + \nabla.(\phi(1-S)\rho_g\mathbf{v}^d_{ga}) = 0, \tag{15}$$

$$\begin{aligned}
\phi\frac{\partial}{\partial t}\left[(1-S)\rho_g + S\rho_w\right] + \left[(1-S)\rho_g + S\rho_w\right]\frac{\partial\phi_{hydr}}{\partial t} + \\
+\nabla.(\phi\left[(1-S)\rho_g + S\rho_w\right]\mathbf{v}_g) + \nabla.(\phi S\rho_w(\mathbf{v}_w - \mathbf{v}_g)) = -\frac{\partial}{\partial t}(\Delta m_{hydr}). && (16)
\end{aligned}$$

Now we introduce the substitution $\mathbf{X} = (1-S)\rho_g + S\rho_w$, $\mathbf{Y} = (1-S)\rho_{ga}$ to (15) and (16)

$$\begin{aligned}
\phi\frac{\partial\mathbf{Y}}{\partial t} + \mathbf{Y}\frac{\partial\phi_{hydr}}{\partial t} + \nabla.(\phi\mathbf{Y}\mathbf{v}_g) + \nabla.(\phi(1-S)\rho_g\mathbf{v}^d_{ga}) &= 0, && (17)\\
\phi\frac{\partial\mathbf{X}}{\partial t} + \mathbf{X}\frac{\partial\phi_{hydr}}{\partial t} + \nabla.(\phi\mathbf{X}\mathbf{v}_g) + \nabla.(\phi S\rho_w(\mathbf{v}_w - \mathbf{v}_g)) &= -\frac{\partial}{\partial t}(\Delta m_{hydr}). && (18)
\end{aligned}$$

After discretization of the latter equaitons we get

$$X^j_i.\left[\phi^j_i - \Delta t_1\frac{A_h}{\rho_s}\frac{(T^n_i - T^{n-1}_i)}{\Delta t} + \Delta t_1\frac{\phi^j_i(v_g)^{j-1}_i}{h}\right] =$$

$$= \phi_i^j X_i^{j-1} + \Delta t_1 \frac{X_{i-1}^j (v_g)_{i-1}^{j-1} \phi_{i-1}^j}{h} - \Delta t_1 A_h \frac{T_i^n - T_i^{n-1}}{\Delta t} -$$

$$-\Delta t_1 \frac{\phi_i^j S_i^{j-1} (\rho_w)_i^j (v_w - v_g)_i^{j-1} - \phi_{i-1}^j S_{i-1}^{j-1} (\rho_w)_{i-1}^j (v_w - v_g)_{i-1}^{j-1}}{h}, \tag{19}$$

$$Y_i^j . \left[\phi_i^j - \Delta t_1 \frac{A_h}{\rho_s} \frac{(T_i^n - T_i^{n-1})}{\Delta t} + \Delta t_1 \frac{\phi_i^j (v_g)_i^{j-1}}{h} \right] = \phi_i^j Y_i^{j-1} + \Delta t_1 \frac{Y_{i-1}^j (v_g)_{i-1}^{j-1} \phi_{i-1}^j}{h} +$$

$$+\Delta t_1 \frac{\phi_i^j (1 - S_i^{j-1}) (\rho_g)_i^j (v_{ga}^d)_i^{j-1} - \phi_{i-1}^j (1 - S_{i-1}^{j-1}) (\rho_g)_{i-1}^j (v_{ga}^d)_{i-1}^{j-1}}{h}, \tag{20}$$

where

$$\begin{aligned} X_i^j &= (1 - S_i^j)(\rho_g)_i^j + S_i^j (\rho_w)_i^j, \\ Y_i^j &= (1 - S_i^j)(\rho_{ga})_i^j. \end{aligned}$$

Since $\rho_g = \rho_{gw} + \rho_{ga}$, then

$$X_i^j = (1 - S_i^j)(\rho_{gw})_i^j + S_i^j (\rho_w)_i^j + Y_i^j. \tag{21}$$

Let us denote $\mathcal{F}((\rho_{gw})_i^j) = (1 - S_i^j)(\rho_{gw})_i^j + S_i^j (\rho_w)_i^j - X_i^j + Y_i^j$, where

$$\left[S((p_c)_i^j) \right]_i^j = \left[1 + \left(\frac{(p_c)_i^j}{p_b^c} \right)^n \right]^{-m}, \qquad \left[p_c((\rho_{gw})_i^j) \right]_i^j = -(\rho_w)_i^j \frac{R T_i^j}{M_w} \ln \left[\frac{T_i^j R}{(p^{gws})_i^j} (\rho_{gw})_i^j \right].$$

For given X_i^j, Y_i^j from (19) and (20), we find a solution $(\rho_{gw})_i^j$ of the nonlinear equation (21) written in the form

$$\mathcal{F}((\rho_{gw})_i^j) = 0, \tag{22}$$

with Newton's iteration procedure; the solution at the rst iteration is given by

$$\left\{ (\rho_{gw})_i^j \right\}^r = \left\{ (\rho_{gw})_i^j \right\}^{r-1} - \frac{\mathcal{F}' \left(\left\{ (\rho_{gw})_i^j \right\}^{r-1} \right)}{\mathcal{F} \left(\left\{ (\rho_{gw})_i^j \right\}^{r-1} \right)}, \tag{23}$$

where

$$\mathcal{F}' \left(\left\{ (\rho_{gw})_i^j \right\}^{r-1} \right) = 1 - S \left(p_c \left((\rho_{gw})_i^j \right) \right) + \frac{\partial S}{\partial p_c} . p_c' \left((\rho_{gw})_i^j \right) ((\rho_w)_i^j - (\rho_{gw})_i^j),$$

$$\frac{\partial S}{\partial p_c} (p_c) = -\frac{mn}{p_b^c} \left(\frac{p_c}{p_b^c} \right)^{n-1} \left[1 + \left(\frac{p_c}{p_b^c} \right)^n \right]^{-m-1}, \qquad p_c' (\rho_{gw}) = -\frac{T R \rho_w}{M_w \rho_{gw}}.$$

From boundary conditions (4) and (5) we get

$$\rho_{gw} (\mathbf{v}_g + \mathbf{v}_{gw}^d - \beta_c) + \rho_{ga} \mathbf{v}_{gw}^d = -\rho_w \mathbf{v}_l - \beta_c \rho_{gw\infty}, \tag{24}$$

$$p_{atm} = p_g = p_{ga} + p_{gw} = \frac{TR}{M_a}\rho_{ga} + \frac{TR}{M_w}\rho_{gw}, \tag{25}$$

where $\rho_{gw}(0)$ and $\rho_{ga}(0)$ are the solutions of (24) and (25) and finally

$$\begin{aligned} X_0^j &= (1 - S_0^j)(\rho_{gw}(0) + \rho_{ga}(0)) + S_0^j(\rho_w)_0^j, & (26)\\ Y_0^j &= (1 - S_0^j)\rho_{ga}(0). & (27)\end{aligned}$$

Analogously, for the boundary conditions (6) and (7), we get

$$\begin{aligned} X_l^j &= (1 - S_l^j)(\rho_{gw}(l) + \rho_{ga}(l)) + S_l^j(\rho_w)_l^j, & (28)\\ Y_l^j &= (1 - S_l^j)\rho_{ga}(l). & (29)\end{aligned}$$

5. Numerical results

Numerical algorithm was implemented in the model by coding in FORTRAN. Following figures show developments of temperature, saturation and water vapour pressure. An increase of temperature and capillary pressure and corresponding decrease of saturation are observed only in the confined layer in the range 50 mm from the heated surface. The swift evaporation of water inside the wall implies the rapid desaturation in the zone of increased vapour pressure. Analysis of these results allows for better understanding of hygro-thermal behaviour of concrete elements near the heated boundary.

6. Thermal spalling

In 1996 the fire with temperatures up to 700 °C occurred in the transport Tunnel connecting England and France, as in the similar case in 1999 in St. Gotthard tunnel,

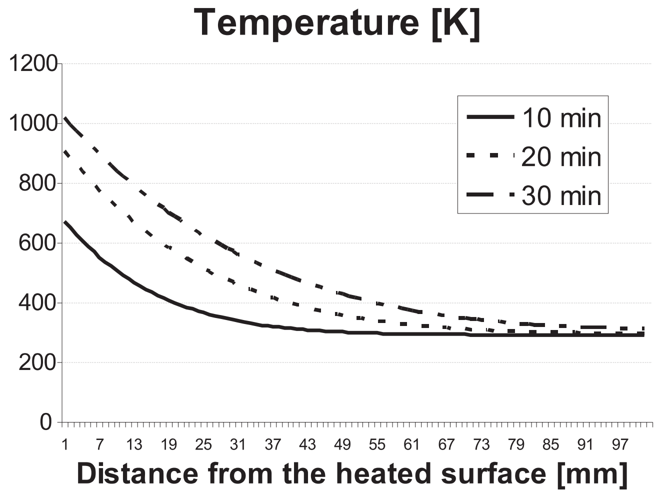


Fig. 1: *Temperature distributions.*

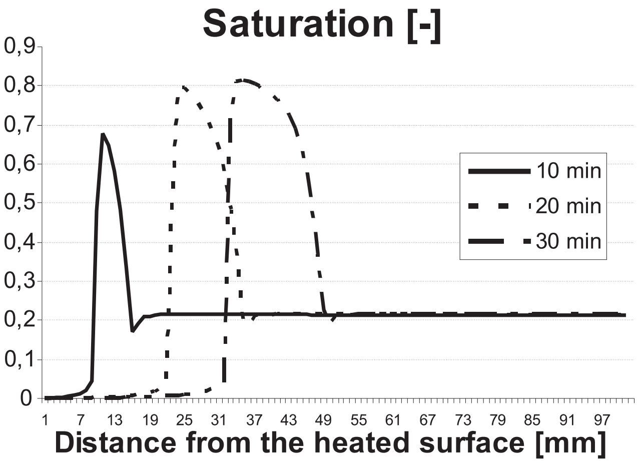


Fig. 2: *Saturation distributions.*

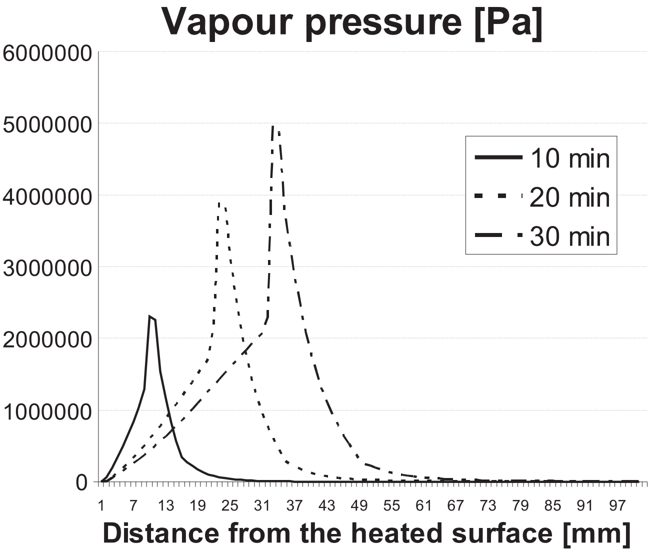
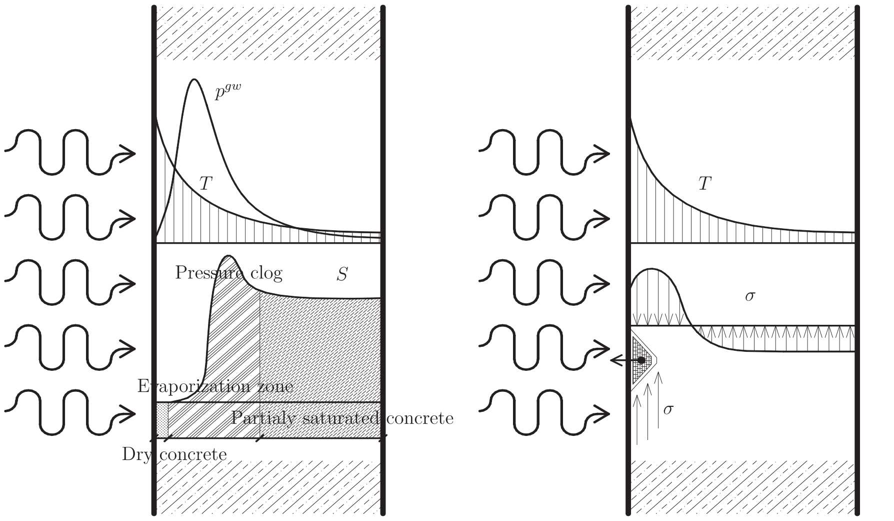


Fig. 3: *Vapour pressure distributions.*

the fire destroyed the concrete structure by thermal spalling over a length of a few hundred meters.

An interesting phenomenon, very specific for heated concrete, is the so-called thermal spalling, which can sometimes be explosive. Its physical causes are still not fully understood. Two main phenomena are generally considered to explain this transient thermal behavior of High Performance Concrete (see [4], [5], [6]):

- generation of internal vapor pressures, which exceed the local tensile strength of the material,
- thermo-mechanical stresses associated with thermal gradients increased by the local consumption of energy associated with vaporization and dehydration.



Spalling Due to Vapour Pressure

Spalling Due to Restrain Thermal Dilatation

Fig. 4: *Thermal Spalling Hypothesis.*

References

[1] A.Y. Beliaev, S.M. Hassanizadeh: *A theoretical model of hysteresis and dynamic effects in the capillary relation for two-phase flow in porous media.* Transport in Porous Media **43**, 2001, 487–510.

[2] S.M. Hassanizadeh, W.G. Gray: *Mechanics and thermodynamics of multiphase flow in porous media including interphase boundaries.* Adv. Water Resour. **13**, 1990, 169–186.

[3] D. Gawin, C.-E. Majorana, B.-A. Schrefler: *Numerical analysis of hygro-thermal behaviour and damage of concrete at high temperature.* Mech. Cohes.-Frict. Mater. **4**, 1999, 37–74.

[4] D. Gawin, F. Pesavento, B.-A. Schrefler: *Towards prediction of the thermal spalling risk through a multi-phase porous media model of concrete.* Comput. Methods Appl. Mech. Engrg. **195**, 2006, 5707–5729.

[5] B.-A. Schrefler: *Multiphase flow in deforming porous material.* Int. J. Numer. Meth. Eng. **60**, 2004, 27–50.

[6] F.-J. Ulm, O. Coussy, Z.-P. Bažant: *The "Chunnel" fire. I: chemoplastic softening in rapidly heated concrete.* Journal of Engineering Mechanics **125**, 3, 1999, 283–289.

HIERARCHICAL FEM: STRENGTHENED CBS INEQUALITIES, ERROR ESTIMATES AND ITERATIVE SOLVERS[*]

Radim Blaheta

Abstract

This paper describes natural decomposition of hierarchical finite element spaces, discusses a characterization of this decomposition via strengthened CBS inequality and uses this decomposition for development of hierarchical error estimates and iterative solution methods.

1. Introduction

A subsequent refinement of a finite element grid provides a sequence of nested grids and hierarchy of nested finite element spaces as well as a natural hierarchical decomposition of these spaces. This decomposition can be characterized by the constant from the corresponding Cauchy–Bunyakowski–Schwarz (CBS) inequality. In Section 2, we summarize some older and recent results concerning this constant. The CBS analysis is exploited in Section 3 for investigation of the so called hierarchical error estimates. We shall show that such estimates are robust with respect to coefficient jumps and anisotropy as well as to the element shape. Hierarchical error estimates can be used for both global and local error assessment. Local estimates can be used for local refinement and construction of hierarchy of locally refined spaces. In Section 4, we outline the hierarchical decomposition in this case. Note that this decomposition can be used for defining various iterative solution methods and preconditioners.

2. FE hierarchy and natural decomposition

Let us consider a model boundary value problem in $\Omega \subset R^d$ $(d = 2, 3)$,

$$\text{find} \quad u \in V \ : \ a(u,v) = b(v) \qquad \forall v \in V, \tag{1}$$

where $V = H_0^1(\Omega)$, $b(v) = \int\limits_{\Omega} fv dx$ for $f \in L_2(\Omega)$ and

$$a(u,v) = \int\limits_{\Omega} \sum_{ij}^{d} k_{ij} \frac{\partial u}{\partial x_i} \frac{\partial v}{\partial x_j} \, dx \, .$$

Above $K = (k_{ij})$ is a symmetric and positive definite matrix of coefficients.

[*]This work was supported by grant No. 1ET400300415 of the Academy of Sciences of the Czech Republic.

We also consider a coarse triangular or tetrahedral finite element grid $\mathcal{T}_H$ of Ω and a fine grid $\mathcal{T}_h$, which arises by a refinement of the coarse elements. By $\mathcal{N}_H$ and $\mathcal{N}_h$, we denote the set of nodes corresponding to $\mathcal{T}_H$ and $\mathcal{T}_h$, respectively. Naturally, $\mathcal{N}_h = \mathcal{N}_H \cup \mathcal{N}_H^+$, where $\mathcal{N}_H^+$ is the complement of $\mathcal{N}_H$ in $\mathcal{N}_h$.

Now, we can introduce the finite element spaces U_H and U_h ($U_H \subset U_h$) of functions which are continuous and linear on the elements of the triangulation $\mathcal{T}_H$, and $\mathcal{T}_h$, respectively. We shall also speak about a *hierarchy of triangulations and finite element spaces.*

Let $\{\phi_i^H\}$ and $\{\phi_i^h\}$ be the standard nodal finite element bases of U_H and U_h , i.e. $\phi_i^H(x_j) = \delta_{ij}$ for all $x_j \in \mathcal{N}_H$, $\phi_i^h(x_j) = \delta_{ij}$ for all $x_j \in \mathcal{N}_h$. Then we can also introduce a *hierarchical basis* $\{\bar{\phi}_i^h\}$ in U_h ,

$$\bar{\phi}_i^h = \begin{cases} \phi_i^h & \text{if} \quad x_i \in \mathcal{N}_H^+ , \\ \phi_i^H & \text{if} \quad x_i \in \mathcal{N}_H . \end{cases}$$

It gives a *natural hierarchical decomposition* of the space U_h,

$$U_h = U_H \oplus U_H^+ , \tag{2}$$

where $U_H^+ = \text{span}\ \{\phi_i^h,\ x_i \in \mathcal{N}_H^+\}$.

The decomposition (2) is characterized by the angle between the subspaces or the strengthened CBS inequality with the constant $\gamma = \cos(U_H, U_H^+)$, which is defined as follows:

$$\begin{aligned} \gamma &= \cos(U_H, U_H^+) \\ &= \sup \left\{ \frac{\mid a(u,v) \mid}{\sqrt{a(u,u)}\sqrt{a(v,v)}} : u \in U_H,\ a(u,u) \neq 0,\ v \in U_H^+,\ a(v,v) \neq 0 \right\} . \end{aligned} \tag{3}$$

If $\mathcal{T}_h$ arises from $\mathcal{T}_H$ by a regular division of the coarse grid triangles into m^2 congruent triangles in 2D or a regular division (given by the affine mapping to a reference rectangular tetrahedra) of the coarse grid tetrahedra into m^3 tetrahedra (see Fig. 1) and if the coefficients $K = (k_{ij})$ are constant on the coarse grid elements then for general anisotropic coefficients and arbitrary shape of the coarse grid elements, we get

$$\gamma \le \sqrt{\frac{m^2-1}{m^2}} \qquad \text{and} \qquad \gamma \le \sqrt{\frac{(m^2-1)(m^2+2)}{m^2(m^2+1)}}$$

for 2D and 3D case, respectively. See [1], [4] and the references given there for more details.

Note that in special cases we get smaller values of γ. For example, $\gamma \le \sqrt{3/8}$ for isotropic coefficients and equilateral triangles [8], $\gamma \le \sqrt{1/2}$ for isotropy and rectangular finite elements [8], [1] or $\gamma \le \sqrt{3/4}$ for orthotropy $k_{ij} = k_i \delta_{ij}$ and rectangular tetrahedra [4] .

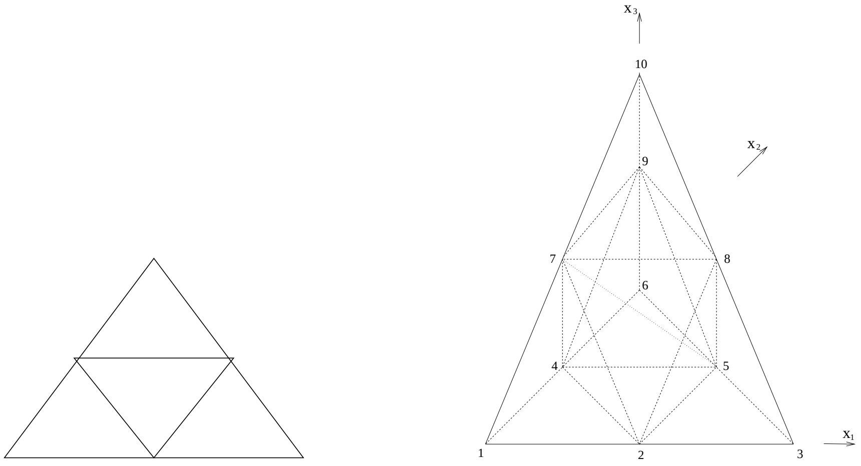


Fig. 1: *Decompositions in 2D and 3D with multiplicity* $m = 2$.

3. Hierarchical error estimates

Hierarchical error estimates were introduced in papers by R.E. Bank, see [3]. The aim is to estimate the error $e_H = u - u_H$, where $u_H \in V_H$ is the finite element approximation of the exact solution $u \in V$ of the considered boundary value problem (1), $V_H = U_H \cap V$.

Let us also introduce the spaces $V_h = U_h \cap V$ and $V_H^+ = U_H^+ \cap V,\ V_h = V_H \oplus V_H^+$ and let u_h be the finite element approximation of u in V_h, i.e.

$$u_h \in V_h : \, a(u - u_h, z) = 0 \quad \forall z \in V_h \, . \tag{4}$$

Lemma 1 *Let there is a positive constant* $\beta < 1$ *such that*

$$\| \, u - u_h \, \|_a \leq \beta \, \| \, u - u_H \, \|_a, \tag{5}$$

where $\| \, v \, \|_a = \sqrt{a(v,\, v)}$. *Then*

$$\frac{1}{1+\beta} \, \| \, u_H - u_h \, \|_a \leq \| \, u - u_H \, \|_a \leq \frac{1}{1-\beta} \, \| \, u_H - u_h \, \|_a \, . \tag{6}$$

Proof see e.g. [3].

The assumption (5) is crucial and need not be fulfilled in any case, see e.g. [6] for a counterexample. If this assumption holds, then

$$\eta = \| \, u_H - u_h \, \|_a \tag{7}$$

is the two-level a posteriori error estimate.

For practical use, the computation of η is too expensive. The hierarchical decomposition $V_h = V_H \oplus V_H^+$ then suggest to use an approximation w_h to u_h,

$$w_h \in V_H^+ : \, a(u - u_H - w_h, z) = 0 \quad \forall z \in V_H^+ \, . \tag{8}$$

Lemma 2 *Let the saturation assumption (5) holds and* $\eta_H = \| w_h \|_a$. *Then*

$$\frac{1}{(1+\beta)(1+\gamma)}\eta_H \leq \| u - u_H \|_a \leq \frac{1}{(1-\beta)(1-\gamma)}\eta_H, \tag{9}$$

where $\gamma = \cos(V_H, V_H^+)$.

Proof see e.g. [3].

Note that η_H is called the hierarchical error estimate.

Let us now consider algebraic formulation of the fine grid finite element approximation in the hierarchical basis. We get

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$$

where u_1 and u_2 correspond to $\mathcal{N}_H^+ \setminus \partial\Omega$ and $\mathcal{N}_H \setminus \partial\Omega$, respectively, and

$$\begin{aligned} A_{11} &= [a(\phi_j^h, \phi_i^h) : x_i, x_j \in \mathcal{N}_H^+ \setminus \partial\Omega], \\ A_{12} &= [a(\phi_j^H, \phi_i^h) : x_i \in \mathcal{N}_H^+ \setminus \partial\Omega, \ x_j \in \mathcal{N}_H \setminus \partial\Omega], \text{ etc}. \end{aligned}$$

Then

$$\begin{aligned} u_H \text{ is represented by } w_2 : A_{22}w_2 &= b_2 \\ w_h \text{ is represented by } w_1 : A_{11}w_1 &= b_1 - A_{12}u_2 \end{aligned}$$

and $\eta_1 = \| w_h \|_a = \sqrt{\langle A_{11}w_1, w_1 \rangle} = \| w_1 \|_A$.

The computation of w_1 can be still too expensive and we can be interested in a possible simplification, e.g. by approximation $\bar{A}_{11} \sim A_{11}$ such that

$$\bar{w}_1 : \bar{A}_{11}\bar{w}_1 = b_1 - A_{12}u_2$$

can be computed in a number of operations proportional to the number of elements in $\mathcal{N}_H^+$ (i.e. $O(\#\mathcal{N}_H^+)$ operations) and provide a good approximation to w_1.

The simplest case is to replace A_{11} by its diagonal, but then the relation between $\| \bar{w}_1 \|_A$ and $\| w_1 \|_A$ depends on anisotropy and/or shape of the elements.

For 2D case, another approximation can be constructed as in the paper [2]. It gives nice bounds independent on the discretization size and both anisotropy and element shape,

$$(1 - \sqrt{\frac{7}{15}}) \| \bar{w}_1 \|_A \leq \| w_1 \|_A \leq (1 + \sqrt{\frac{7}{15}}) \| \bar{w}_1 \|_A .$$

Moreover, $\bar{w}_1$ can be computed in $O(\#\mathcal{N}_H^+)$ operations.

4. Locally refined hierarchy

The hierarchical error estimators discussed in the previous section are global, but their value can be computed from contributions of macroelements corresponding to coarse grid elements to $\| w_h \|_a$. These local contributions or another local estimators can be used for determination of these coarse grid elements, which should be refined. After their refining, we can either work with special hanging nodes or make another refinement of the surrounding elements by their bisection, see Fig. 2.

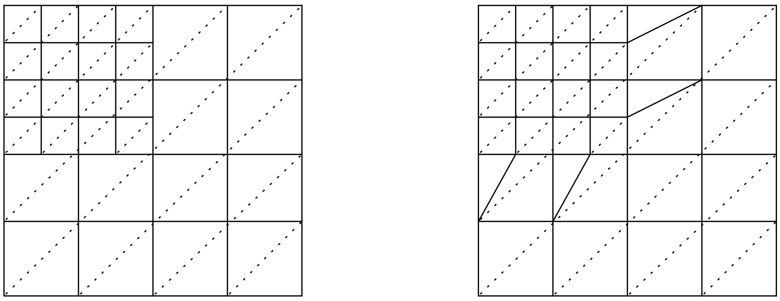


Fig. 2: *Local refinement with hanging nodes (left) and bisection (right).*

Again we get spaces U_H and U_h and the natural decomposition $U_h = U_H \oplus U_H^+$. The constant $\gamma = \cos(U_H, U_H^+)$, is then important for special iterative solution methods like FAC or BEPS, see [7], [5] and the references therein.

Theorem 1

- *In the case of local refinement with hanging nodes, γ remains the same as in the case of global refinement.*
- *In the case of local refinement with bisection, we obtain the same constant γ only in special cases (e.g. orthotropic problems $k_{ij} = k_i \delta_{ij}$ and refinement like on Fig.2 right). Generally, γ is not further robust with respect to anisotropy or the element shape.*

The proof of the first statement can be found in [7], the second statement will be discussed in a forthcoming paper.

5. Conclusions

The paper shows the hierarchical finite element method with hierarchical error estimates, which are robust with respect to coefficients jumps between coarse elements and both physical and numerical anisotropy. The finite element problems on locally refined grids can be solved by iterative methods, see [7] and [5]. The convergence of these methods can be again estimated with the aid of the strengthened CBS constant.

Acknowledgement

The author is grateful to an anonymous referee for careful reading and useful comments.

References

[1] O. Axelsson, R. Blaheta: *Two simple derivations of universal bounds for the C.B.S. inequality constant.* Appl. Math. **49**, 2004, 57–72.

[2] O. Axelsson, A. Padiy: *On the additive version of the algebraic multilevel iteration method for anisotropic elliptic problems.* SIAM J. Sci. Comp. **20**, 1999, 1807–1830.

[3] R.E. Bank: *Hierarchical bases and the finite element method.* Acta Numerica **5**, 1996, 1–43.

[4] R. Blaheta: *Nested tetrahedral grids and strengthened C.B.S. inequality.* Numer. Linear Algebra Appl. **10**, 2003, 619–637.

[5] R. Blaheta, P. Byczanski, R. Kohut: *Composite grid finite element method: Implementation and iterative solution with inexact subproblems.* Appl. Math. **47**, 2002, 83–100.

[6] S.C. Brenner, C. Carstensen: *Finite element methods.* In: E. Stein, R. de Borst and T.J.R. Hughes (eds), Encyclopedia of Computational Mechanics, J. Wiley, 2004, 73–118.

[7] J. Mandel, S. McCormick: *Iterative solution of elliptic equations with refinement: the two-level case.* In: T. Chan, R. Glowinski, G.A. Meurant, J. Periaux and O. Widlund (eds), Domain Decomposition Methods for PDEs II, SIAM, Philadelphia 1989, 81–92.

[8] J.F. Maitre, F. Musy: *The contraction number of a class of two-level methods, an exact evaluation for some finite element subspaces and model problems.* In: W. Hackbusch, U. Trottenberg (eds), Multigrid Methods, Lecture Notes in Math. **960**, Springer-Verlag, Berlin 1982, 535–544.

ACCURACY INVESTIGATION OF A STABILIZED FEM FOR SOLVING FLOWS OF INCOMPRESSIBLE FLUID*

Pavel Burda, Jaroslav Novotný, Jakub Šístek

Abstract

In computer fluid dynamics, employing stabilization to the finite element method is a commonly accepted way to improve the applicability of this method to high Reynolds numbers. Although the accompanying loss of accuracy is often referred, the question of quantifying this defect is still open. On the other hand, practitioners call for measuring the error and accuracy. In the paper, we present a novel approach for quantifying the difference caused by stabilization.

Dedicated to Professor Ivo Babuška on the occasion of his 80th birthday.

1. Introduction

The finite element method equipped with stabilization has proven to be a powerful tool for solving flows of incompressible fluids with high Reynolds numbers. But applying stabilization can lead to a change of the approximate solution in a serious way, as was discussed in [2].

The aim of our present research is to quantify the difference and find a way to predict it. Application of a posteriori error estimates seems to be a promising way to face these tasks.

Several numerical examples are presented to show the effect of stabilization and to investigate the accuracy.

2. Mathematical model

The considered mathematical model is the system of Navier-Stokes equations in two space dimensions (1) accompanied by the continuity equation (2). The aim is to search the vector of velocity $\mathbf{u}(\mathbf{x},t) = (u_1(\mathbf{x},t), u_2(\mathbf{x},t)) \in [\mathcal{C}^2(\overline{\Omega})]^2$ and pressure $p(\mathbf{x},t) \in \mathcal{C}^1(\overline{\Omega})/\mathbb{R}$ such that

$$\frac{\partial \mathbf{u}}{\partial t} + (\mathbf{u}\cdot\nabla)\mathbf{u} - \nu\Delta\mathbf{u} + \nabla p = \mathbf{f} \quad \text{in } \Omega \times [0,T]\,, \tag{1}$$

$$\nabla \cdot \mathbf{u} = 0 \quad \text{in } \Omega \times [0,T]\,, \tag{2}$$

where ν denotes kinematic viscosity and $\mathbf{f}(\mathbf{x},t)$ stands for intensity of volume force.

*This research has been supported by grant No. 106/05/2731 of the Grant Agency of the Czech Republic.

Boundary conditions (3)–(4) are allowed. For time dependent problems, initial condition (5) is considered.

$$\begin{aligned} \mathbf{u} &= \mathbf{g} \ \text{ on } \Gamma_g \times [0,T] && (3)\\ -\nu(\nabla\mathbf{u})\mathbf{n} + p\mathbf{n} &= \mathbf{0} \ \text{ on } \Gamma_h \times [0,T] && (4)\\ \mathbf{u} &= \mathbf{u}_0 \ \text{ in } \Omega,\ t=0 && (5) \end{aligned}$$

For the solution by the finite element method, we consider the weak formulation of the problem (1)–(2). We introduce function spaces based on Sobolev spaces

$$\begin{aligned} V_g &= \Big\{\mathbf{v} = (v_1, v_2) \mid \mathbf{v} \in [H^1(\Omega)]^2; \mathbf{Tr}\ v_i = g_i, i = 1,2, \text{ on } \Gamma_g\Big\},\\ V &= \Big\{\mathbf{v} = (v_1, v_2) \mid \mathbf{v} \in [H^1(\Omega)]^2; \mathbf{Tr}\ v_i = 0, i = 1,2, \text{ on } \Gamma_g\Big\}. \end{aligned}$$

Now, we seek velocity $\mathbf{u}(\mathbf{x},t) = (u_1(\mathbf{x},t), u_2(\mathbf{x},t)) \in V_g$ such that $\mathbf{u} - \mathbf{u}_g \in V$ and pressure $p(\mathbf{x},t) \in L_2(\Omega)/\mathbb{R}$ for $t \in [0,T]$ satisfying

$$\int_\Omega \frac{\partial \mathbf{u}}{\partial t} \cdot \mathbf{v}\mathrm{d}\Omega + \int_\Omega (\mathbf{u}\cdot\nabla)\mathbf{u}\cdot\mathbf{v}\mathrm{d}\Omega + \nu\int_\Omega \nabla\mathbf{u} : \nabla\mathbf{v}\mathrm{d}\Omega - \int_\Omega p\nabla\cdot\mathbf{v}\mathrm{d}\Omega = \int_\Omega \mathbf{f}\cdot\mathbf{v}\mathrm{d}\Omega \quad (6)$$

$$\int_\Omega \psi\nabla\cdot\mathbf{u}\mathrm{d}\Omega = 0 \quad (7)$$

for any $\mathbf{v} \in V$ and $\psi \in L_2(\Omega)$. Operation ":" used in (6) is defined as

$$\nabla\mathbf{u} : \nabla\mathbf{v} \ = \ \frac{\partial u_x}{\partial x}\frac{\partial v_x}{\partial x} + \frac{\partial u_x}{\partial y}\frac{\partial v_x}{\partial y} + \frac{\partial u_y}{\partial x}\frac{\partial v_y}{\partial x} + \frac{\partial u_y}{\partial y}\frac{\partial v_y}{\partial y}\,. \quad (8)$$

3. Approximation of the problem by FEM

We use Hood-Taylor finite elements, which lead to the following function spaces

$$\begin{aligned} V_{gh} &= \Big\{\mathbf{v}_h = (v_{h_1}, v_{h_2}) \in [\mathcal{C}(\overline{\Omega})]^2;\ v_{h_i}\mid_K \in R_2(\overline{K}),\ i=1,2,\ \mathbf{v}_h = \mathbf{g} \text{ in nodes on } \Gamma_g\Big\}\\ Q_h &= \Big\{\psi_h \in \mathcal{C}(\overline{\Omega});\ \psi_h\mid_K \in R_1(\overline{K})\Big\}\\ V_h &= \Big\{\mathbf{v}_h = (v_{h_1}, v_{h_2}) \in [\mathcal{C}(\overline{\Omega})]^2;\ v_{h_i}\mid_K \in R_2(\overline{K}),\ i=1,2,\ \mathbf{v}_h = \mathbf{0} \text{ in nodes on } \Gamma_g\Big\} \end{aligned}$$

where V_{gh} is the space for approximation of velocities, Q_h for pressure and test functions for the continuity equation, and V_h for test functions for momentum equations. Here

$$R_m(\overline{K}) = \begin{cases} P_m(\overline{K}), & \text{if } K \text{ is a triangle}\\ Q_m(\overline{K}), & \text{if } K \text{ is a quadrilateral} \end{cases}$$

and P_m, Q_m have the usual meaning. Among all the advantages of these elements, we consider it to be rather important, that they lead to functions satisfying Babuška-Brezzi (*inf-sup*) stability condition (9).

$$\exists C_B > 0, const.\ \forall\psi_h \in Q_h\ \exists\mathbf{v}_h \in V_h \quad (\psi_h, \nabla\cdot\mathbf{v}_h)_0 \geq C_B\|\psi_h\|_0\|\mathbf{v}_h\|_1 \quad (9)$$

4. SemiGLS stabilization technique

In [4], semiGLS stabilization technique was derived as a modification of Galerkin Least Squares method, proposed by Hughes, Franca, and Hulbert [3]. We search the approximate velocity $\mathbf{u}_h \in V_{gh}$ and pressure $p_h \in Q_h$ satisfying in Ω

$$B_{sGLS}(\mathbf{u}_h, p_h; \mathbf{v}_h, \psi_h) = L_{sGLS}(\mathbf{v}_h, \psi_h), \quad \forall \mathbf{v}_h \in V_h, \ \ \forall \psi_h \in Q_h \,, \tag{10}$$

where

$$\begin{aligned} B_{sGLS}(\mathbf{u}_h, p_h; \mathbf{v}_h, \psi_h) &\equiv \int_\Omega \frac{\partial \mathbf{u}_h}{\partial t} \cdot \mathbf{v}_h \mathrm{d}\Omega + \int_\Omega (\mathbf{u}_h \cdot \nabla)\mathbf{u}_h \cdot \mathbf{v}_h \mathrm{d}\Omega \\ + \ \ \nu \int_\Omega \nabla \mathbf{u}_h : \nabla \mathbf{v}_h \mathrm{d}\Omega &- \int_\Omega p_h \nabla \cdot \mathbf{v}_h \mathrm{d}\Omega + \int_\Omega \psi_h \nabla \cdot \mathbf{u}_h \mathrm{d}\Omega + \\ + \ \ \sum_{K=1}^{N} \int_K \left[\frac{\partial \mathbf{u}_h}{\partial t} + (\mathbf{u}_h \cdot \nabla)\mathbf{u}_h - \nu \Delta \mathbf{u}_h + \nabla p_h \right] &\cdot \tau \left[(\mathbf{u}_h \cdot \nabla)\mathbf{v}_h - \nu \Delta \mathbf{v}_h + \nabla \psi_h \right] \mathrm{d}\Omega, \end{aligned}$$

$$L_{sGLS}(\mathbf{v}_h, \psi_h) \ \ \equiv \ \ \int_\Omega \mathbf{f} \cdot \mathbf{v}_h \mathrm{d}\Omega + \sum_{K=1}^{N} \int_K \mathbf{f} \cdot \tau \left[(\mathbf{u}_h \cdot \nabla)\mathbf{v}_h - \nu \Delta \mathbf{v}_h + \nabla \psi_h \right] \mathrm{d}\Omega \,.$$

Here τ denotes stabilization parameter. The way to determine it is mentioned in [2]. Index *sGLS* is an abbreviation of semiGLS.

5. Evaluating of the accuracy

A straightforward way to evaluate the effect of stabilization is to compute the difference between solution with and without stabilization. This method was proposed in [2] accompanied by numerical examples and is applicable in the range of Reynolds numbers, where we can solve the problem both with and without stabilization. Such difference represents "pure distortion" caused by stabilization.

To get the idea about achieved accuracy of our solution, it is also suitable to apply a posteriori error estimates. We use following estimate derived for Hood-Taylor elements

$$\mathcal{U}^2(u_1 - u_{1h}, u_2 - u_{2h}, p - p_h) \le \mathcal{E}^2(u_{1h}, u_{2h}, p_h) \,, \tag{11}$$

where the terms represent

$$\mathcal{U}^2(u_1 - u_{1h}, u_2 - u_{2h}, p - p_h) = \|(e_{u_1}, e_{u_2})\|^2_{1,K} + \|e_p\|^2_{0,K} \,,$$

$$\begin{aligned} \mathcal{E}^2(u_{1h}, u_{2h}, p_h) \ \ &= \ \ C \Big[h_K^2 \int_K \big(r_1^2(u_{1h}, u_{2h}, p_h) + r_2^2(u_{1h}, u_{2h}, p_h) \big) \mathrm{d}\Omega \\ &+ \ \ \int_K r_3^2(u_{1h}, u_{2h}, p_h) \mathrm{d}\Omega \Big] \,, \end{aligned}$$

and $r_1(u_{1h}, u_{2h}, p_h)$, $r_2(u_{1h}, u_{2h}, p_h)$, and $r_3(u_{1h}, u_{2h}, p_h)$ stand for residuals of the system (1)–(2); (u_1, u_2, p) denotes an exact solution, (u_{1h}, u_{2h}, p_h) an approximate solution computed by FEM, and $(e_{u_1}, e_{u_2}, e_p) = (u_1 - u_{1h}, u_2 - u_{2h}, p - p_h)$ an error of approximate solution. Constant C is determined from a numerical experiment described in [1], as well as details on the a posteriori estimates.

Such approach is applicable for any Reynolds number, for which we can find solution by the stabilized method and estimates the whole difference between a stabilized solution and an exact one.

6. Results of numerical experiments

To demonstrate the approach using a posteriori error estimates, we present results for a problem of a lid driven cavity and a channel with a sudden extension of diameter. Both problems are steady and the results for measuring distortion caused by the stabilization can be found in [2].

In Figures 1 – 2, we can observe the effect of stabilization on streamlines inside cavity for three levels of mesh fineness. A posteriori error estimates in the cavity are presented in Figures 3 – 5. They represent the relative error in percents. We can observe, that while the regions of higher error are decreasing for the Newton method without stabilization when refining the mesh, they remain almost independent of refinement for the stabilized method.

Geometry of the channel is described in Figure 6. Streamlines in the channel for Reynolds number 1,000 are presented in Figure 7, and in Figure 8, there are a posteriori error estimates to compare the differences.

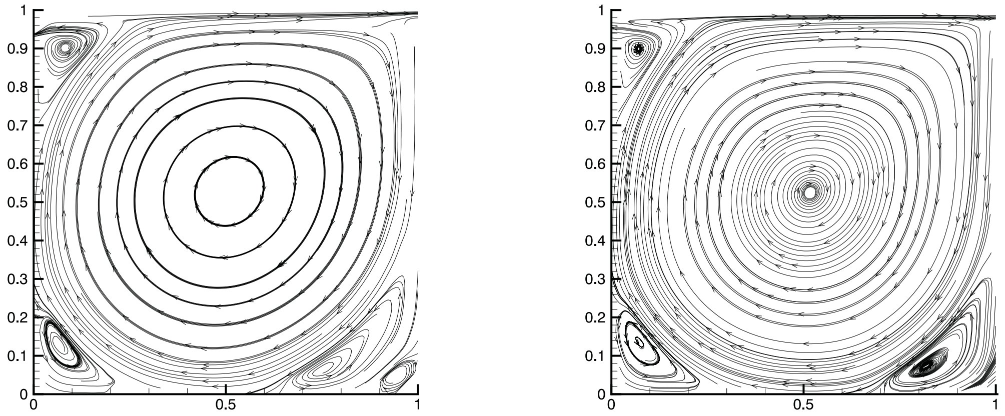


Fig. 1: *Streamlines, Re = 10,000, mesh 32×32 without stabilization (left) and by semiGLS (right).*

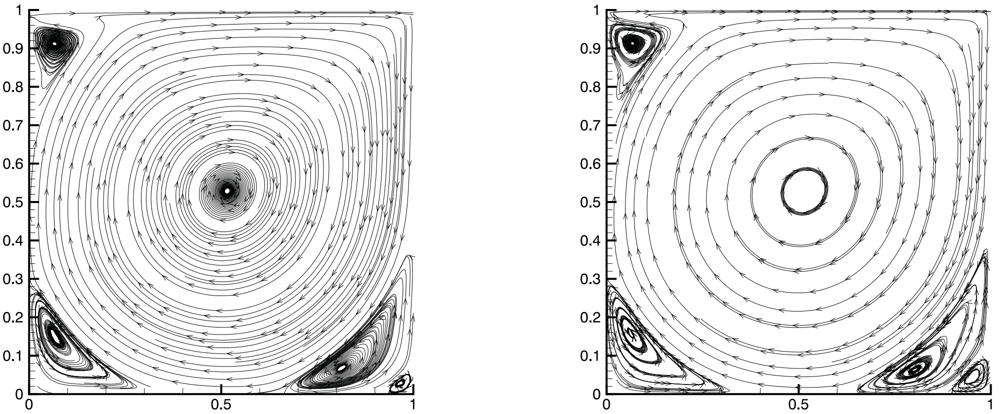


Fig. 2: *Streamlines by semiGLS, Re = 10,000, mesh 64×64 (left) and 128×128 (right).*

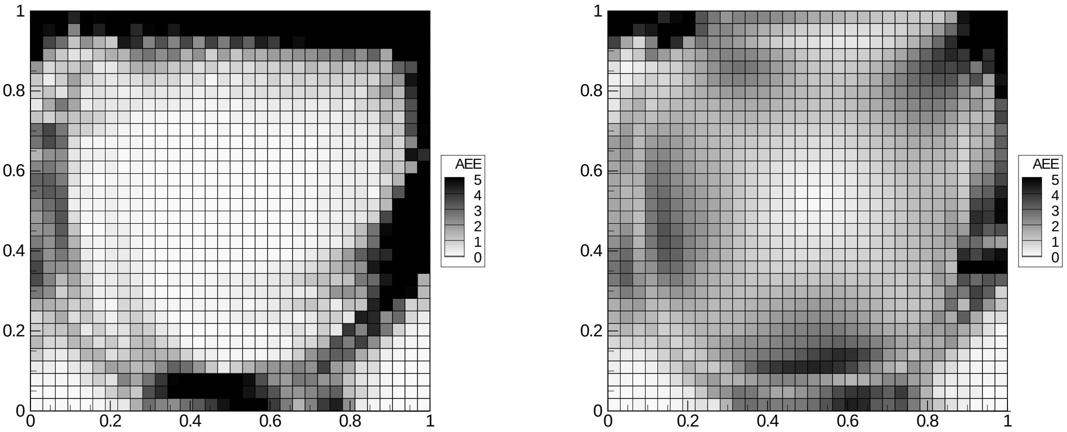


Fig. 3: *A posteriori errors on elements, Re = 10,000, mesh 32×32 without stabilization (left) and by semiGLS (right).*

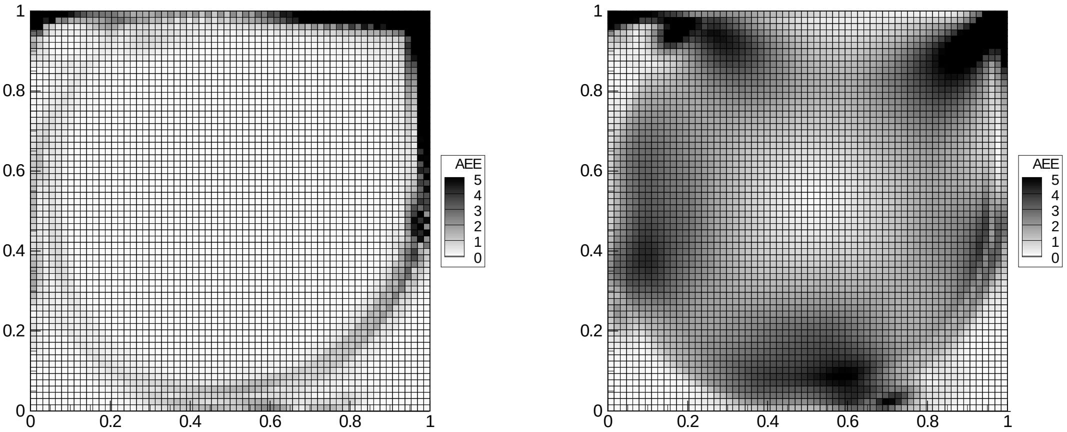


Fig. 4: *A posteriori errors on elements, Re = 10,000, mesh 64×64 without stabilization (left) and by semiGLS (right).*

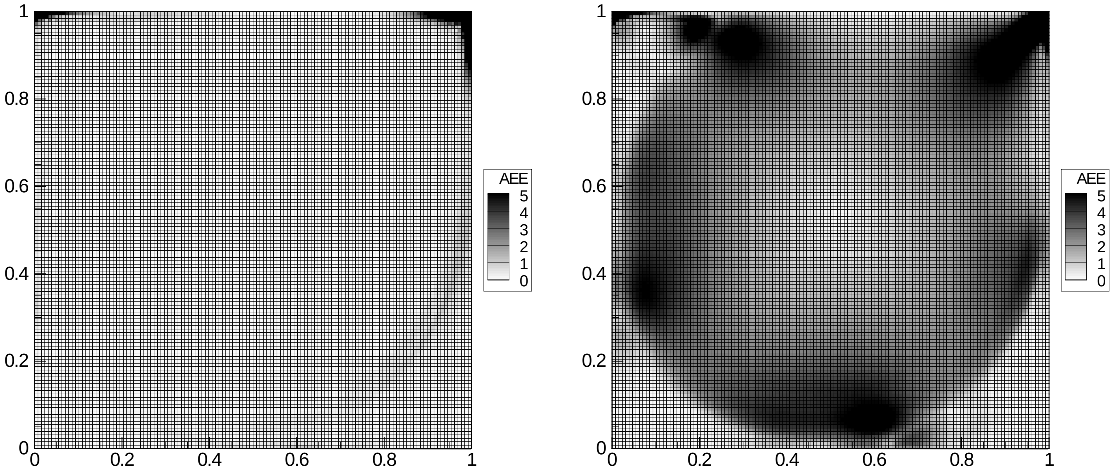


Fig. 5: *A posteriori errors on elements, Re = 10,000, mesh 128×128 without stabilization (left) and by semiGLS (right).*

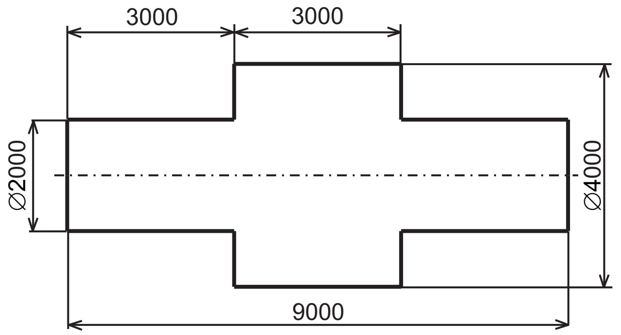


Fig. 6: *Geometry of the channel (dimensions in milimeters).*

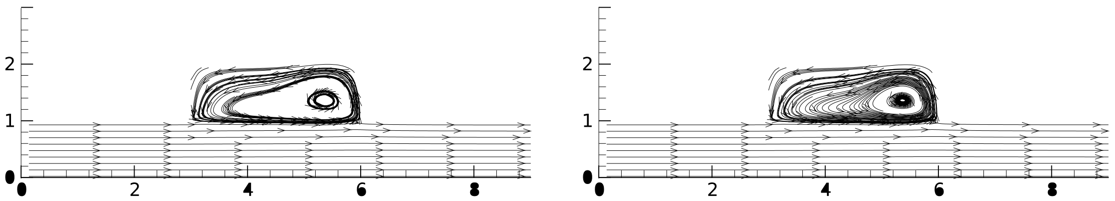


Fig. 7: *Streamlines in the channel by the Newton method without stabilization (left) and by the semiGLS algorithm (right), Re = 1,000.*

7. Conclusion

We have developed a stabilized method and tested it on various problems, where it provided promising results. This means, that we were able to reach markably higher Reynolds numbers using this method than using method without stabilization.

The cost for using stabilization is a loss of accuracy. This loss is hard to predict, but we are able to quantify it and estimate it a posteriori. We have presented two

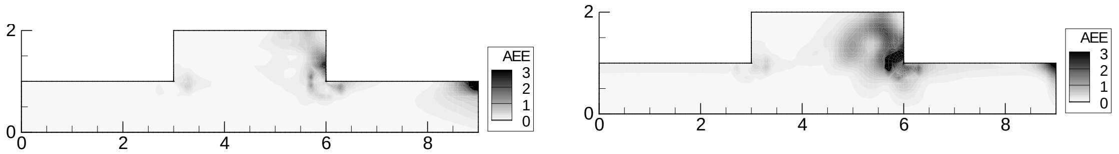


Fig. 8: *A posteriori error estimates in the channel by the Newton method without stabilization (left) and by the semiGLS algorithm (right), Re = 1,000.*

approaches for such evaluation, based on comparing approximate solutions with and wihout stabilization and on a posteriori error estimation.

As the main ideas resulting from the research we could mention, that for reaching higher Reynolds numbers, stabilization should be efficiently combined with mesh refinement, because both of these factors improve the stability of the method. We have shown, that residual stabilization is not as innocent in practice as available proofs of convergence claim, and people, who use stabilized methods, should be aware of this fact and always take care of the final accuracy of their computations.

References

[1] P. Burda, J. Novotný, B. Sousedík: *A posteriori error estimates applied to flow in the channel with corners.* Mathematics and Computers in Simulation **61**, 2003, 375–383.

[2] P. Burda, J. Novotný, J. Šístek: *On a modification of GLS stabilized FEM for solving incompressible viscous flows.* Internat. J. Numer. Methods Fluids **51**, 2006, 1001–1016.

[3] T.J.R. Hughes, L.P. Franca, G.M. Hulbert: *A new finite element formulation for computational fluid dynamics: VIII. The Galerkin/least-squares method for advective-diffusive equations.* Comput. Methods Appl. Mech. Engrg. **73**, 1989, 173–189.

[4] J. Šístek: *Stabilization of finite element method for solving incompressible viscous flows* (diploma thesis). Praha, ČVUT 2004.

ON A TRAFFIC PROBLEM[*]

Lubor Buřič, Vladimír Janovský

Abstract

We consider a macroscopic follow-the-leader model of a road traffic. The novelty is that we incorporate the possibility to overtake a slower car. We introduce two ways to simulate overtaking. One is based on swapping initial conditions after the overtaking occurs. Second approach is to formulate the problem as a Filippov system with discontinuous right-hand sides.

1. Introduction

A massive traffic is the phenomenon of our civilization. The mathematical modeling of traffic flows has a long tradition, see e.g. [1] for a recent review. We will consider a class of *macroscopic follow-the-leader models*, see e.g. [2]: Consider the system

$$\frac{\mathrm{d}x_i}{\mathrm{d}t} = y_i\,, \qquad \frac{\mathrm{d}y_i}{\mathrm{d}t} = V(x_{i+1}-x_i) - y_i\,, \quad x_{N+1} = x_1 + L, \tag{1}$$

$i = 1,\dots,N$. It models N cars on a circular road of the length L. The pairs (x_i, y_i) are interpreted as the position $x_i \equiv \mathrm{mod}(x_i, L)$ and the velocity y_i of the car number i. The acceleration $\mathrm{d}y_i/\mathrm{d}t$ of each car depends on the difference between the car velocity y_i and the *optimal velocity function* $V = V(x_{i+1}-x_i)$. In particular, we will consider the hyperbolic optimal velocity function $r \mapsto V(r)$ defined as

$$V(r) = V^{\max}\frac{\tanh\left(a(r-1)\right) + \tanh(a)}{1+\tanh(a)}\,, \tag{2}$$

where $V^{\max}$ and a are positive constants. The choice of V imposes a driving law and we assume that this law is the same for all N drivers. The difference

$$h_i \equiv x_{i+1} - x_i\,, \quad i = 1,\dots,N\,, \tag{3}$$

is called *headway* (of the i-th car). Note that we can also formulate the model (1) in the state space of headway and velocity components

$$\frac{\mathrm{d}h_i}{\mathrm{d}t} = y_{i+1} - y_i\,, \qquad \frac{\mathrm{d}y_i}{\mathrm{d}t} = V(h_i) - y_i\,, \qquad i = 1,\dots,N. \tag{4}$$

[*] The research of the first author was supported by the grant MSM 6046137306 of the Ministry of Education, Youth and Sports, Czech Republic. The second author was supported by the Grant Agency of the Czech Republic (grant No. 201/06/0356) and also by the research project MSM 0021620839 of The Ministry of Education, Youth and Sports, Czech Republic.

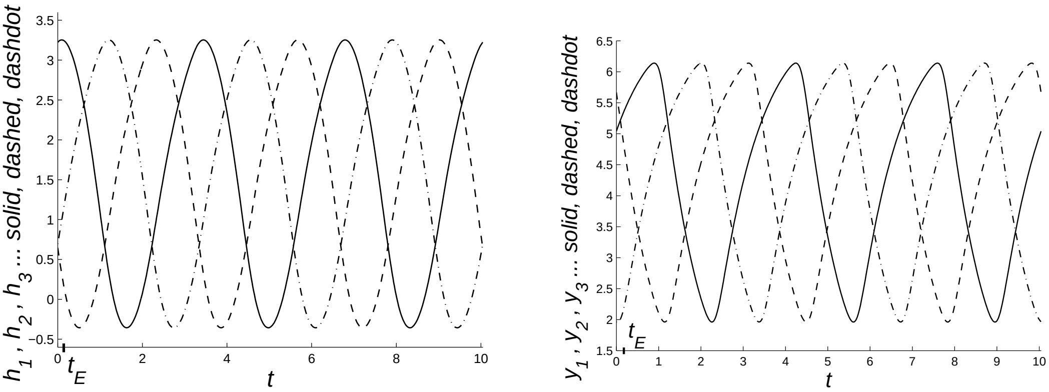


Fig. 1: *Velocity vs. time, headway vs. time: negative headway is non physical.*

Given an initial condition $[x^0, y^0] \in \mathbb{R}^N \times \mathbb{R}^N$, the system (1) defines a flow on $\mathbb{R}^N \times \mathbb{R}^N$

$$[x^0, y^0] \mapsto [x(t), y(t)] \equiv \Psi(t, [x^0, y^0])\,, \quad t \in \mathbb{R}\,. \tag{5}$$

Without loss of generality, we may order x^0 as

$$s \leq x_1^0 \leq x_2^0 \leq \cdots \leq x_{N-1}^0 \leq x_N^0 \leq L + s\,,$$

where $s \in \mathbb{R}$ is an arbitrary phase shift. It is easy to check that there exists a family of *quasi-stationary solutions*, see e.g. [2]. For example, in case $N = 3$ let $x^0 = [s; s + L/3; s + 2L/3]$, $y^0 = [c; c; c]$, $c \equiv V(L/3)$ where $s \in \mathbb{R}$ is an arbitrary phase shift. Then the flow (5) is given by $x(t) = [s + ct; s + L/3 + ct; s + 2L/3 + ct]$, $y(t) = [c; c; c]$ for all t. Therefore, velocity and headway components are constant.

These solutions were observed both stable and unstable. The stability exchange is due to the Hopf bifurcation, see [3]: In certain parameter regions, quasi-stationary solutions co-exist with periodic solutions to (1).

Fig. 1 shows the periodic solution for $N = 3$ cars and the parameter setting $L = 4.56281$, $V^{\max} = 7$, $a = 2$. The periodicity concerns the velocity and headway components. In [4], the authors noted that the solutions to (1) which yield the negative headway are problematic to interprete physically. They called them *non physical solutions*. For example, the trajectory on Fig. 1 becomes non physical since $t_E = 0.2074$. Observe that

$$h_2(t_E) \equiv x_2(t_E) - x_3(t_E) = 0\,, \quad y_2(t_E) > y_3(t_E)\,. \tag{6}$$

The natural interpretation is that the car No 2 is about to **overtake** the car No 3.

The authors of [4] tried to generalize the model (1) in such a way that the periodic solutions become physical for a larger parameter regions. We will follow a different idea. We are going to simulate the overtaking. The resulting model is a piecewise smooth dynamical system composed by pieces of (1).

2. Overtaking

The idea is as follows: On the left Fig. 2, three consecutive trajectories due to the flow (5) are sketched. The headway of the k-th car, namely $h_k(t) = x_{k+1}(t) - x_k(t)$, becomes negative for $t > t_E$. Note that we can compute the time t_E for which $h_k(t_E) = 0$ within a prescribed precision in MATLAB environment (see `odeset`, `Event location`). We define a new initial condition at $t = t_E$ by naturally swapping $[x_k(t_E), y_k(t_E)]$ and $[x_{k+1}(t_E), y_{k+1}(t_E)]$. The resulting trajectories, see Fig. 2 on the right, have discontinuous first derivatives (the solid and dashed lines). Note that x_k on the right Fig. 2 corresponds to position of the k-th car only for $t \leq t_E$. For $t > t_E$, x_k is position of the $(k + 1)$-st car. Overtaking algorithm solves the problem in two runs, simulation with swapping of initial conditions and postprocessing to produce final trajectories of cars. In the postprocessing stage, we assemble pieces of the final trajectories on Fig. 3 from lines obtained on Fig. 2. They have continuous derivatives and discontinuous headway components. The velocities are continuous.

Let us illustrate performance of the algorithm. We consider $N = 14$, $V^{\max} = 34$ and $a = 2$; the same data as in [3], Figure 9. The steady state at $L = 15$ is known

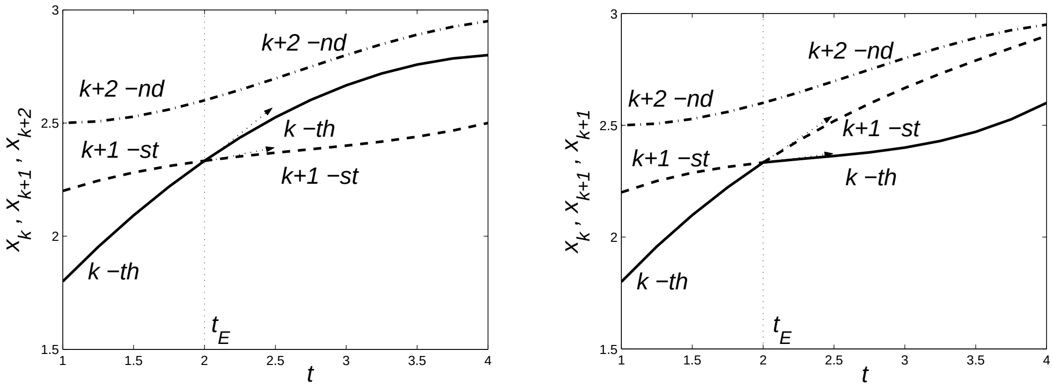


Fig. 2: *On the left: A sketch of three trajectories of the flow (5). On the right: The trajectories after imposing the swap of the initial condition at* $t = t_E$.

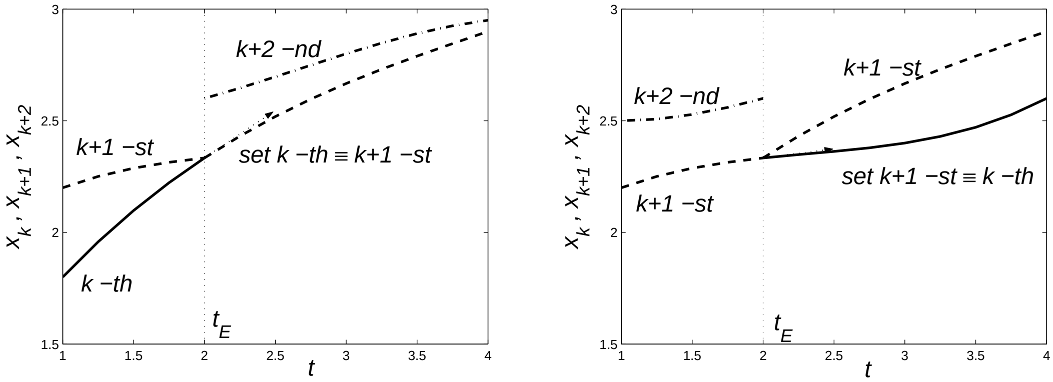


Fig. 3: *On the left: Trajectory of the k-th car. On the right: Trajectory of the* $k + 1$*-st car. The relevant headway components are discontinuous at* t_E.

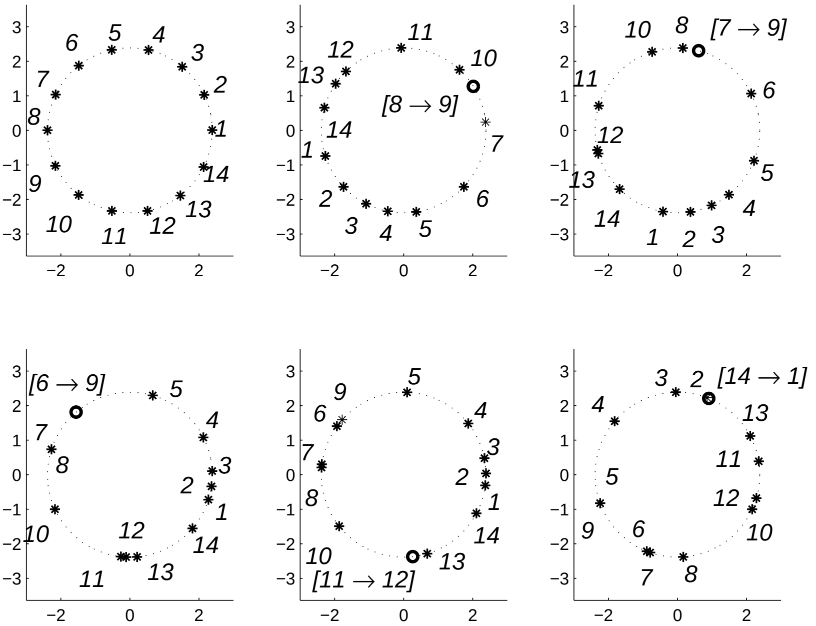


Fig. 4: *A slightly perturbed steady state at* $t = 0$*, top left. Sequence of overtaking (Events): No* 8 *overtakes No* 9*, symbolically* $[8 \longrightarrow 9]$ *at time* $t_E(1) = 1.9136$*,* $[7 \longrightarrow 9]$ *at* $t_E(2) = 2.0426$*,* $[6 \longrightarrow 9]$ *at* $t_E(3) = 2.2294$*,* $[11 \longrightarrow 12]$ *at* $t_E(4) = 2.2546$*,* $[14 \longrightarrow 1]$ *at* $t_E(5) = 2.4605$*.*

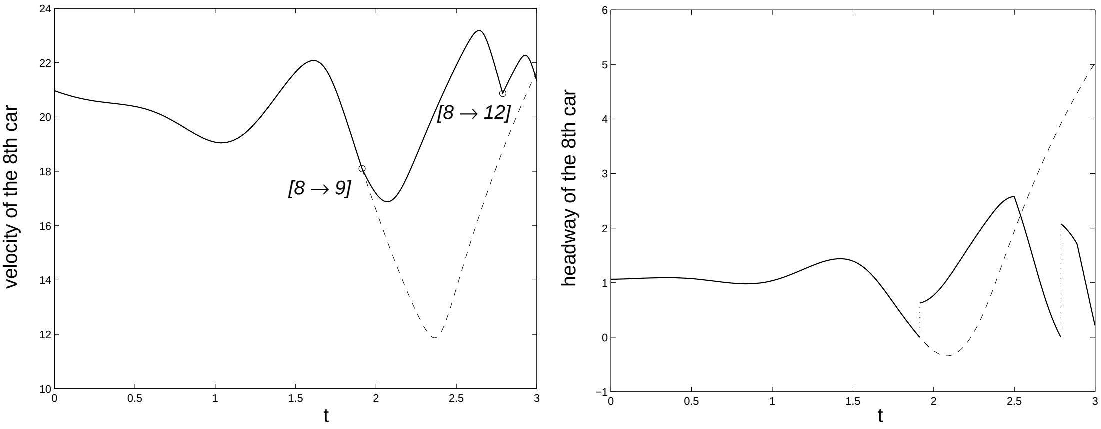


Fig. 5: *Velocity and headway of the* 8*-th car vs time. No* 8 *overtakes No* 9 *and No* 12*. Dashed: The model without overtaking, i.e.* $y_8(t)$*. Since* $t = 1.9136$*, dashed solution becomes non physical.*

to be unstable. Perturbing this steady state slightly, we let the above algorithm work till the time $t = 3$. There were indicated 18 swaps on the track. Five of them are shown on Fig. 4. As an example, we describe the trajectory of the 8-th car for $0 \leq t \leq 3$, see Fig. 5, giving a comparison with the "smooth" model (1).

3. Long time behaviour

Given an initial condition $[x^0, y^0] \in \mathbb{R}^N \times \mathbb{R}^N$ and a time instant $t \geq 0$, let the above algorithm return the actual positions and velocities $[x(t), y(t)] \in \mathbb{R}^N \times \mathbb{R}^N$ of all N cars on the track. We formally define

$$[x^0, y^0] \mapsto [x(t), y(t)] \equiv \Pi(t, [x^0, y^0]), \quad t \geq 0. \tag{7}$$

The aim is to investigate asymptotic properties of the overtaking model as $t \to \infty$. We report on invariant objects we observed. For instance, in the case $N = 3$, one can observe phase-shifted reflectionally symmetric oscillations similar to those predicted for a ring of coupled oscillators, see [5], Chapter XVIII, § 4: Let $N = 3$, $V^{\max} = 7$, $a = 2$ and $L = 3.6998$. Let us set $x^0 = [0.1504; 2.6756; 3.5599]$, $y^0 = [4.2668; 5.1647; 2.9087]$. Due to (7), the velocity $y(t)$ is periodic, see Fig. 6. Its period T can be computed numerically. The cars No 1 and No 2 oscillate out-of-phase with the period $T = 4.8525$. The 3-rd car oscillates twice as rapidly as the other two. The corresponding headway components $h(t)$ oscillate similarly, see Fig. 7. Consider $N = 3$ for simplicity. We will show that Overtaking Model, in the state space of headway and velocity components, can be formulated as a *Filippov system*, see [6].

4. Formulation via a Filippov system

Let us consider $N = 3$. In this case we have only two possible configurations of the cars on the road, see Fig. 8. In the first configuration, cars are running ordered "123" along the circuit in the anticlockwise direction, whereas in the second configuration, the cars are ordered "132". It should be noted that the car numbering is fixed during the computation. The configuration of the cars changes when any car overtakes the other one.

Let us define new variables

$$h_{ij} = x_j - x_i, \quad i \neq j, \tag{8}$$

which describe a gap between the car No i and the car No j. It is clear that h_{ji} can be computed from the relation

$$h_{ji} = L - h_{ij}, \quad i \neq j, \tag{9}$$

which reflects the fact that we consider a closed road. Therefore we can use h_{12}, h_{23} and h_{31} as state variables, only. Remaining gaps h_{13}, h_{21} and h_{32} can be computed from the equation (9).

We will redefine the optimal velocity function as follows. We use the function (2) on the interval $[0, L]$ only, and repeat function values with period L, see Fig. 9 for example. Driving law is independent on whether the car ahead is lap down or lap forward. We denote this new periodic discontinuous optimal velocity function as $\tilde{V}$.

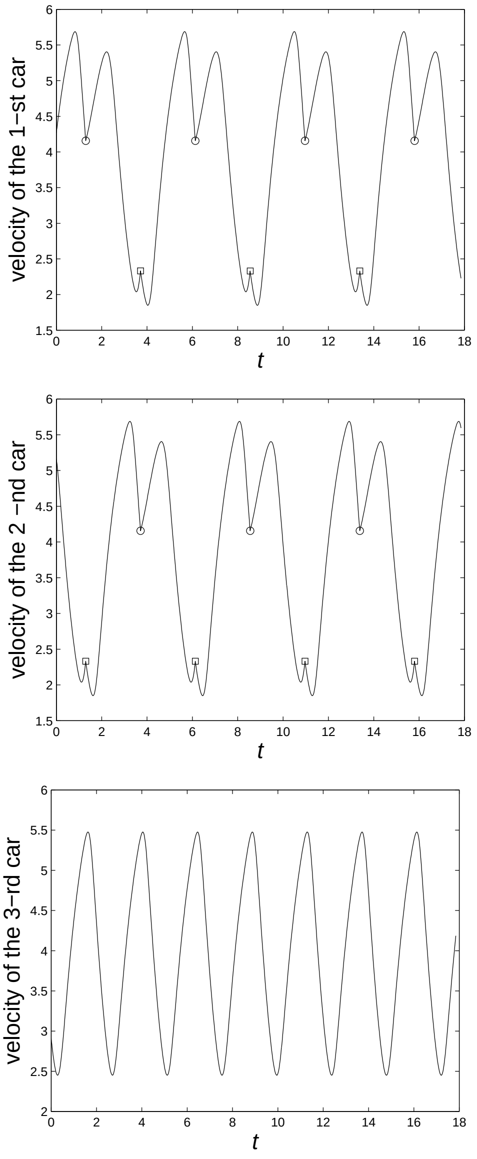


Fig. 6: *The velocity waveforms of period* $T = 4.8363$.

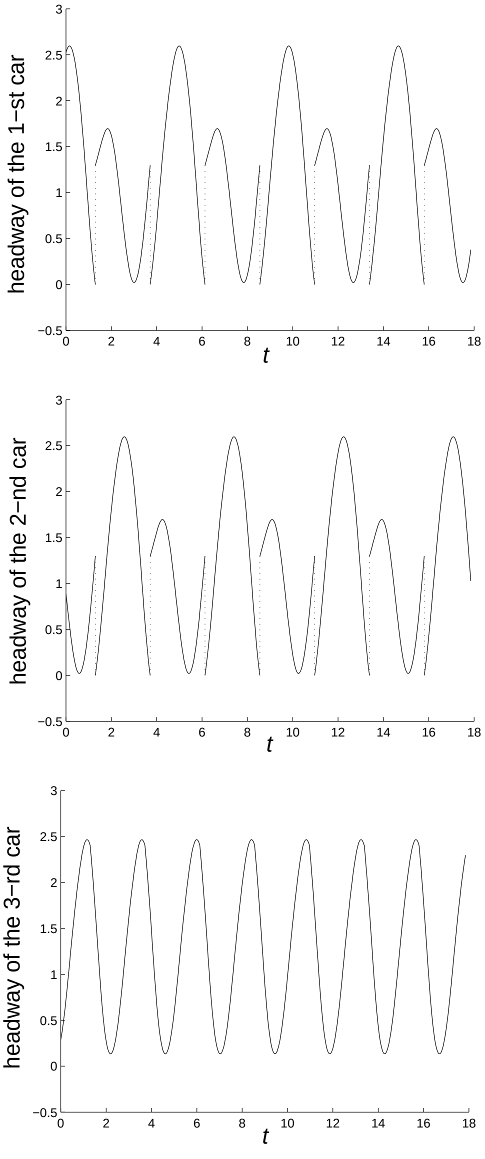


Fig. 7: *Discontinuous headway components.*

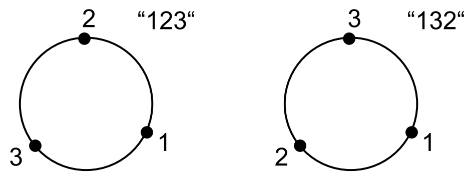


Fig. 8: *Two possible configurations of the cars on the road.*

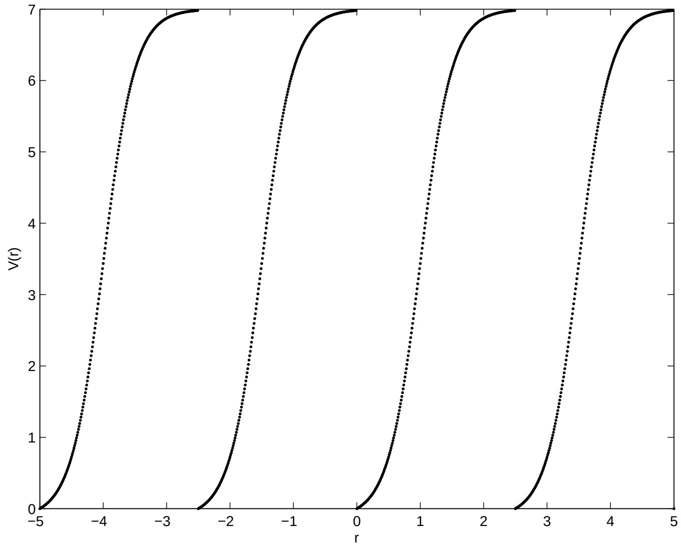


Fig. 9: *Discontinuous optimal velocity function $\tilde{V}$; $V^{\max} = 7$, $a = 2$, $L = 2.5$.*

If the system is in the configuration “123” it is described by the following system of differential equations

$$\begin{aligned}
\frac{\mathrm{d}h_{12}}{\mathrm{d}t} &= y_2 - y_1\,, & \frac{\mathrm{d}y_1}{\mathrm{d}t} &= \tilde{V}(h_{12}) - y_1\,,\\
\frac{\mathrm{d}h_{23}}{\mathrm{d}t} &= y_3 - y_2\,, & \frac{\mathrm{d}y_2}{\mathrm{d}t} &= \tilde{V}(h_{23}) - y_2\,,\\
\frac{\mathrm{d}h_{31}}{\mathrm{d}t} &= y_1 - y_3\,, & \frac{\mathrm{d}y_3}{\mathrm{d}t} &= \tilde{V}(h_{31}) - y_3\,.
\end{aligned} \tag{10}$$

After overtaking occurs, the configuration of the cars changes to “132” and then the system (10) changes to the following one

$$\begin{aligned}
\frac{\mathrm{d}h_{12}}{\mathrm{d}t} &= y_2 - y_1\,, & \frac{\mathrm{d}y_1}{\mathrm{d}t} &= \tilde{V}(h_{13}) - y_1 = \tilde{V}(L - h_{31}) - y_1\,,\\
\frac{\mathrm{d}h_{23}}{\mathrm{d}t} &= y_3 - y_2\,, & \frac{\mathrm{d}y_2}{\mathrm{d}t} &= \tilde{V}(h_{21}) - y_2 = \tilde{V}(L - h_{12}) - y_1\,,\\
\frac{\mathrm{d}h_{31}}{\mathrm{d}t} &= y_1 - y_3\,, & \frac{\mathrm{d}y_3}{\mathrm{d}t} &= \tilde{V}(h_{32}) - y_3 = \tilde{V}(L - h_{23}) - y_1\,.
\end{aligned} \tag{11}$$

Finally, if

$$h_{ij} = kL\,, \quad k \in \mathbb{Z}\,, \tag{12}$$

for some i, j then the i-th car and the j-th car are involved in overtaking. More precisely, if h_{ij} increases when it crosses the boundary (12), then the i-th is overtaken by the j-th one. On the other hand, if h_{ij} decreases when it crosses the boundary (12), then the i-th car overtakes the j-th one. During the computation, we swap systems (10) and (11) after each overtaking. Since the function $\tilde{V}$ is discontinuous and right hand sides of systems (10) and (11) are different, the system given by equations (10), (11) and (12) is a Filippov system, see [6].

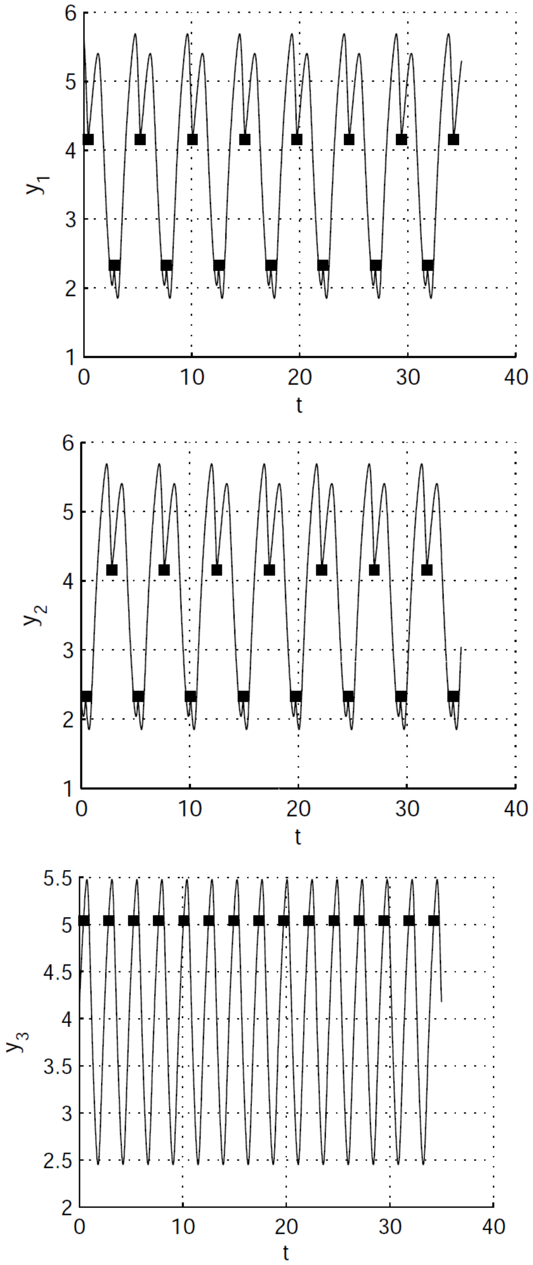


Fig. 10: *The velocity components of the solution of Filippov system* (10),(11),(12).

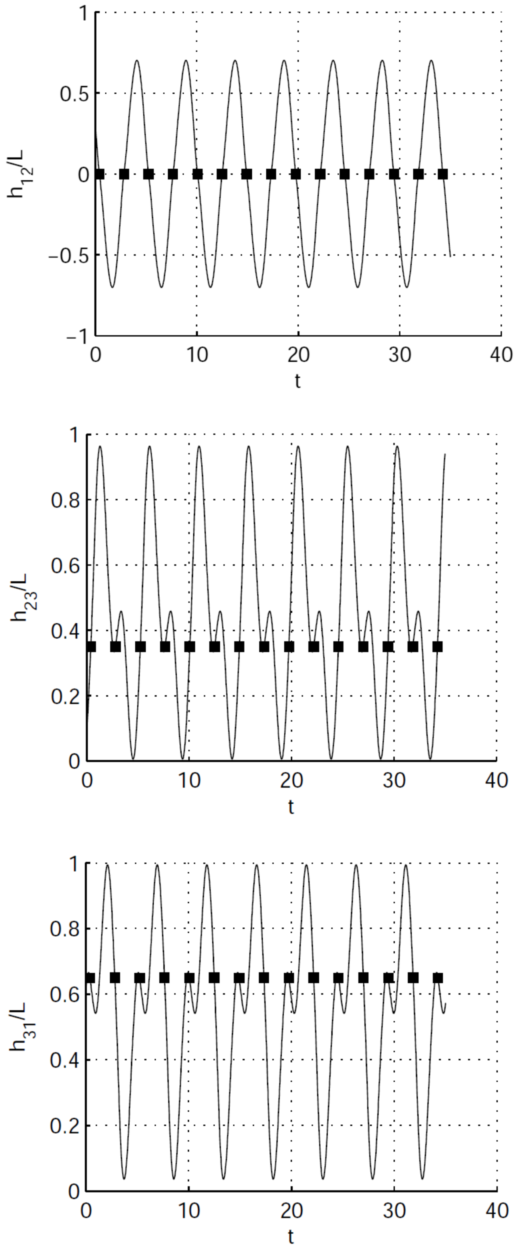


Fig. 11: *The gap components of the solution of Filippov system* (10),(11),(12).

5. Comparison and comments

In this section, we provide a numerical solution of the discontinuous model. The problem was solved in MATLAB by `ode15s` procedure with the event location to detect overtaking. Special attention is paid to the comparison of the results given by the overtaking algorithm described in the section 2 and 3 and results obtained from the discontinuous model. We have fixed values of parameters $V^{\max} = 7$ and $a = 2$. Experiments were started from the "123" configuration.

The numerical solution was obtained for the length of the track $L = 3.6998$, with the initial condition

$$[h_{12}(0), h_{23}(0), h_{31}(0)] = [1.1396, 0.3138, 2.2464]\,, \tag{13a}$$

$$[y_1(0), y_2(0), y_3(0)] = [5.6485, 2.2919, 4.0906]\,. \tag{13b}$$

Results are plotted on the Fig. 10 and 11. On each figure, the overtaking events are marked by the full square.

The velocity components of the solution of both models are similar, compare Fig. 10 and Fig. 6. This shows that both approaches results in the same behaviour of the cars on the track.

Headway components of the solutions are not similar. The model (4) is identical to the "123" configuration of the discontinuous model, i.e. the system (10). Thus, $h_{12} = h_1$, $h_{23} = h_2$ and $h_{31} = h_3$ until no overtaking occurs in the system. After overtaking, since $h_1 = h_{13}$, h_{12} is not equal to h_1 (until next overtaking occurs), etc. Therefore, the solution curves on Fig. 11 correspond to that ones on Fig. 7 only partially. Since the function $h_{12}(t)$ is the only one crossing the boundary represented by the equation (12), the cars No 1 and No 2 overtake each other alternatively and the car No 3 is not involved in any overtaking.

Let us note that gaps $h_{ij}(t)$ are continuous functions, but headway components $h_i(t)$ are discontinuous, see Fig. 11 and 7. The velocities y_i are continuous but they do not have continuous derivatives, see Fig. 10 and 6.

References

[1] D. Helbing: *Traffic and related self-driven many-partical systems.* Rev. Modern Phys. **73**, 2001, 1067–1141.

[2] M. Bando, K. Hasebe, A. Nakayama, A. Shibata, Y. Sugiyama: *Dynamical model of traffic congestion and numerical simulation.* Phys. Rev. E **51**, 1995, 1035–1042.

[3] I. Gasser, G. Sirito, B. Werner: *Bifurcation analysis of a class of 'car following' traffic models.* Physica D **197**, 2004, 222–241.

[4] I. Gasser, T. Seidel, G. Sirito, B. Werner: *Bifurcation analysis of a class of 'car following' traffic models II: variable reaction times and aggressive drivers.* Transport Theory and Statistical Physics **35**, 2006, to appear.

[5] M. Golubitsky, I. Stewart, D.G. Schaeffer: *Singularities and groups in bifurcation theory, Volume II.* New York, Springer Verlag 1988.

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

A FICTITIOUS DOMAIN APPROACH TO THE NUMERICAL SOLUTION OF ELLIPTIC BOUNDARY VALUE PROBLEMS DEFINED IN STOCHASTIC DOMAINS[*]

Claudio Canuto, Tomáš Kozubek

1. Introduction

In [2], we present an efficient method for the numerical solution of elliptic PDEs in domains depending on random variables. The key feature is the combination of a fictitious domain approach and a polynomial chaos expansion. The PDE is solved in a larger, fixed domain (the fictitious domain), with the original boundary condition enforced via a Lagrange multiplier acting on a random manifold inside the new domain. A (generalized) Wiener expansion is invoked to convert such a stochastic problem into a deterministic one, depending on an extra set of real variables (the stochastic variables). Discretization is accomplished by standard mixed finite elements in the physical variables and a Galerkin projection method with numerical integration (which coincides with a collocation scheme) in the stochastic variables. A stability and convergence analysis of the method, as well as numerical results, are provided in [2]. The convergence is "spectral" in the polynomial chaos order, in any subdomain which does not contain the random boundaries.

2. Setting of the problem

Let $(\mathit{\Omega}, F, P)$ be a complete probability space, where $\mathit{\Omega}$ is the set of outcomes, F is the σ-algebra of events and P is the probability measure. For any $\omega \in \mathit{\Omega}$, let $D(\omega) \subset \mathbb{R}^2$ be a bounded domain depending on ω; its boundary $\mathit{\Gamma}(\omega) := \partial D(\omega)$ is assumed to be polygonal or of class $C^{1,1}$, i.e., the boundary is locally represented by functions, whose first derivatives are Lipschitz continuous. We suppose that all domains are contained with their boundaries in a domain $\hat{D} \subset \mathbb{R}^2$, which will serve as the fictitious domain in the fictitious domain formulation (see Figure 1).

For the sake of simplicity, we will be concerned with the following model boundary value problem in $D(\omega)$: Find $u : \overline{D(\omega)} \times \Omega \to \mathbb{R}$ such that almost surely (a.s.) in Ω we have

$$\begin{cases} -\triangle u(\,\cdot\,,\omega) &= f \ \text{ in } D(\omega), \\ u(\,\cdot\,,\omega) &= 0 \ \text{ on } \mathit{\Gamma}(\omega), \end{cases} \qquad \big(\mathcal{P}(\omega)\big)$$

[*]This research was supported by the European project Breaking Complexity, n. HPRN-CT-2002-00286 and by the grants IAA1075402 and 1ET400300415 of the Grant Agency of the Czech Academy of Sciences.

where f is a given function in $L^2(\hat{D})$. The case of Neumann or mixed boundary conditions or of random coefficients and data (independent of the random variables describing the domain) could be handled at no extra difficulty.

Solving the discrete problem $\big(\mathcal{P}(\omega)\big)$ for any $\omega \in \Omega$ using, e.g., the finite element method, means that by varying ω we have to: (i) remesh the new domain $D(\omega)$; (ii) assemble the new stiffness matrix and the right hand side vector; (iii) solve the new system of linear equations. Thus the efficiency of solving the discrete problems is crucial. Hereafter, we will explore a fictitious domain method with nonfitted meshes as a possible way to increase efficiency: indeed, this approach avoids completely step (i) and partially step (ii), since the stiffness matrix remains the same for any admissible domain.

3. The fictitious domain (FD) formulation

In this section, we will consider problem $\big(\mathcal{P}(\omega^*)\big)$ for a given event $\omega^* \in \Omega$; we will simplify our notation by setting $D := D(\omega^*)$, $\Gamma := \Gamma(\omega^*)$ and $u = u(\,\cdot\,, \omega^*)$.

Let $\hat{D}$ be the fictitious domain containing $\overline{D}$. The corresponding fictitious domain formulation reads as follows:

$$\begin{cases} \text{Find } (\hat{u}, \lambda) \in V \times M \text{ such that} \\ \int_{\hat{D}} \nabla \hat{u} \cdot \nabla v \, d\mathbf{x} + \langle \lambda, \tau v \rangle_\Gamma = \int_{\hat{D}} f v \, d\mathbf{x}, \quad \forall v \in V, \\ \langle \mu, \tau \hat{u} \rangle_\Gamma = 0, \quad \forall \mu \in M, \end{cases} \tag{$\hat{\mathcal{P}}$}$$

where the symbol $\langle .,. \rangle$ denotes the duality pairing between $M := H^{-1/2}(\Gamma)$ and $H^{1/2}(\Gamma)$, $\tau : H^1_0(\hat{D}) \to H^{1/2}(\Gamma)$ stands for the trace mapping and V is a closed subspace of $H^1(\hat{D})$. Typical choices for V are: $H^1(\hat{D})$, $H^1_0(\hat{D})$, or $H^1_P(\hat{D}) = \{v \mid \quad v \in H^1(\hat{D}),\ v \text{ is periodic on } \partial\hat{D}\}$ if $\hat{D}$ is a cartesian product of intervals.

The reason for introducing the space of the Lagrange multipliers M is to fulfil the requirement that $\hat{u}_{|_D}$ solves $\big(\mathcal{P}(\omega^*)\big)$.

The well-posedness of this problem for any $f \in L^2(\hat{D})$ follows from classical results on abstract saddle-point problems (see [1]). Hence the saddle-point problem $\big(\hat{\mathcal{P}}\big)$ has a unique solution $(\hat{u}, \lambda) \in V \times M$. In addition, $\hat{u}_{|_D} = u$ and $\lambda = \left[\frac{\partial u}{\partial n}\right]$, the jump of the normal derivative of u across Γ.

3.1. Discretization of the FD formulation

Problem $\big(\hat{\mathcal{P}}\big)$ will be approximated by using the mixed finite element method (see [1]). For this purpose the spaces V and M are replaced by suitable finite dimensional subspaces V_h and M_H. More specifically, V_h contains all *continuous piecewise bilinear* functions $\hat{v}_h$ constructed over a uniform rectangulation of $\hat{D}$ and satisfying boundary condition on $\partial\hat{D}$ dependent on the choice of V. Further, M_H contains all *piecewise constant* functions μ_H constructed over a partition of ∂D. For more details we refer to [3].

The resulting algebraic formulation is

$$\begin{pmatrix} \mathbb{A} & \mathbb{B}^T \\ \mathbb{B} & \mathbb{O} \end{pmatrix} \begin{pmatrix} \mathbf{u} \\ \boldsymbol{\lambda} \end{pmatrix} = \begin{pmatrix} \mathbf{f} \\ \mathbf{0} \end{pmatrix}, \tag{$\mathbf{P}$}$$

where $\mathbb{A}$ is *the stiffness matrix*, $\mathbb{B}$ is the matrix coupling the primal variable $\mathbf{u}$ and the Lagrange multiplier $\boldsymbol{\lambda}$, which are the vectors of the nodal values of $\hat{u}_h$ (approximation of $\hat{u}$ from V_h) and of the constant values of λ_H (approximation of λ from M_H), respectively, and $\mathbf{f}$ *is the load vector*.

To solve $\big(\mathbf{P}\big)$, we use the first equation to eliminate the vector $\mathbf{u} = \mathbb{A}^{-1}(-\mathbb{B}^T\boldsymbol{\lambda}+\mathbf{f})$ from the second one, and we solve the resulting system for $\boldsymbol{\lambda}$, $\mathbb{B}\mathbb{A}^{-1}\mathbb{B}^T\boldsymbol{\lambda} = \mathbb{B}\mathbb{A}^{-1}\mathbf{f}$, by a *conjugate gradient method*. The size of $\mathbb{B}\mathbb{A}^{-1}\mathbb{B}^T$ is much smaller than the size of $\mathbb{A}$. Generally, we do not need any preconditioning but we are able to construct preconditioners to the Schur complement based on the pseudoinverse and multigrid techniques. The multiplication by $\mathbb{A}^{-1}$ can be realized efficiently, e.g., by Choleski factorization with symmetric approximate minimum degree reordering, multigrid approach, domain decomposition method or by using fast solvers based on *the Fourier Analysis* and *the cyclic reduction*.

4. The stochastic FD formulation

We go back to the stochastic setting. The FD formulation $\big(\hat{\mathcal{P}}\big)$ suggests the following stochastic FD formulation: Find $\hat{u}(\,\cdot\,,\omega) \in H_0^1(\hat{D})$ and $\lambda(\,\cdot\,,\omega) \in M(\omega) := H^{-1/2}(\mathit{\Gamma}(\omega))$ such that, a.s. in $\mathit{\Omega}$,

$$\begin{cases} \int_{\hat{D}} \nabla\hat{u}(\,\cdot\,,\omega)\cdot\nabla v\, d\mathbf{x} + \langle\lambda(\,\cdot\,,\omega),\tau v\rangle_{\mathit{\Gamma}(\omega)} = \int_{\hat{D}} f v\, d\mathbf{x}, & \forall v \in H_0^1(\hat{D}), \\ \langle\mu,\tau\hat{u}(\,\cdot\,,\omega)\rangle_{\mathit{\Gamma}(\omega)} = 0, & \forall\mu \in M(\omega). \end{cases} \tag{$\hat{\mathcal{P}}(\omega)$}$$

We assume that, a.s., $\mathit{\Gamma}(\omega)$ is obtained from a reference $C^{1,1}$ or polygonal boundary $\mathit{\Gamma}_0$ as the image of a piecewise smooth invertible mapping $\gamma_0(\omega)$. More precisely, we assume that $\mathit{\Gamma}(\omega) = \gamma_0(\omega)(\mathit{\Gamma}_0)$, where $\gamma_0(\omega)$ belongs to $C^{1,p}(\mathit{\Gamma}_0)$ (the space of all continuous and piecewise continuously differentiable mappings $\gamma : \mathit{\Gamma}_0 \to \mathbb{R}^2$) and its inverse $\gamma_0(\omega)^{-1}$ exists and belongs to $C^{1,p}(\mathit{\Gamma}(\omega))$. The function $\gamma_0 : \Omega \to C^{1,p}(\mathit{\Gamma}_0)$ is assumed to be a random variable belonging to $L^\infty(\Omega, dP; C^{1,p}(\mathit{\Gamma}_0))$, i.e., γ_0 is a jointly measurable function on the Borel sets of $\mathit{\Gamma}_0 \times \Omega$ for which there exists a constant $g_0 > 0$ such that $\|\gamma_0(\omega)\|_{C^{1,p}(\mathit{\Gamma}_0)} \le g_0$ a.s. in Ω; the same occurs for the inverse mapping, i.e., $\|\gamma_0(\omega)^{-1}\|_{C^{1,p}(\mathit{\Gamma}(\omega))} \le g_0$ a.s. in Ω.

Let $\mathbb{E}[X] = \int_\Omega X(\omega)\, dP(\omega)$ be the expected value of a real-valued random variable X. Let $L^2(\Omega, dP) = \{X : \Omega \to \mathbb{R} \,|\, X$ is a random variable such that $\mathbb{E}[X^2] < +\infty\}$ be the space of second order random variables over the probability space $(\mathit{\Omega}, F, P)$. We denote by $L^2(\Omega, dP;\ H_0^1(\hat{D}))$ the space of the random variables $v : \mathit{\Omega} \to H_0^1(\hat{D})$ (i.e., $v : \hat{D}\times\Omega \to \mathbb{R}$ is jointly measurable and $v(\,\cdot\,,\omega) \in H_0^1(\hat{D})$ a.s. in Ω) with finite second order moment $\mathbb{E}\Big[\|v\|^2_{H_0^1(\hat{D})}\Big] = \int_{\hat{D}} \mathbb{E}[|\nabla v|^2]\, d\mathbf{x} < +\infty$. The definition of the space $L^2(\Omega, dP; H^{-1/2}(\mathit{\Gamma}_0))$ is similar. Finally, the space $L^2(\Omega, dP; H^{-1/2}(\mathit{\Gamma}))$ is defined as follows: $\mu \in L^2(\Omega, dP; H^{-1/2}(\mathit{\Gamma}))$ means that $\mu_0 \in L^2(\Omega, dP; H^{-1/2}(\mathit{\Gamma}_0))$, where $\mu_0(\omega) \in H^{-1/2}(\mathit{\Gamma}_0)$ is defined a.s. in Ω by the conditions $\langle\mu_0, v_0\rangle_{\mathit{\Gamma}_0} = \langle\mu, v_0\circ\gamma_0^{-1}\rangle_{\mathit{\Gamma}(\omega)}$ for all $v_0 \in H^{1/2}(\mathit{\Gamma}_0)$.

With such notation at hand, the stochastic FD formulation given at the beginning of the section can be made precise as follows: Find $\hat{u} \in L^2(\Omega, dP; H_0^1(\hat{D}))$ and $\lambda \in L^2(\Omega, dP; H^{-1/2}(\mathit{\Gamma}))$ such that

$$\begin{cases} \mathbb{E}\left[\int_{\hat{D}} \nabla \hat{u} \cdot \nabla v \, d\mathbf{x}\right] + \mathbb{E}\left[\langle \lambda, \tau v\rangle_{\mathit{\Gamma}}\right] = \mathbb{E}\left[\int_{\hat{D}} f v \, d\mathbf{x}\right], \quad \forall v \in L^2(\Omega, dP; H_0^1(\hat{D})), \\ \mathbb{E}\left[\langle \mu, \tau \hat{u}\rangle_{\mathit{\Gamma}}\right] = 0, \quad \forall \mu \in L^2(\Omega, dP; H^{-1/2}(\mathit{\Gamma})). \end{cases} \qquad (\hat{\mathcal{P}}^S)$$

Our next step will be to transform this stochastic problem into a purely deterministic one. This will be accomplished by expanding the random variables into polynomial chaos.

5. (Wiener) polynomial chaos

This section is devoted to recalling some basic facts about polynomial chaos (see, e.g., [4]), as well as to setting the notation.

Let $Y_1(\omega), \ldots, Y_k(\omega), \ldots$ be a sequence of independent standard Gaussian random variables with zero mean and unit variance, i.e., $\mathbb{E}[Y_k] = 0$, $\mathbb{E}[Y_k Y_\ell] = \delta_{k\ell}$ for all $k, \ell \geq 1$. On the other hand, given a real variable y, let $\{H_n(y)\}_{n\geq 0}$ be the sequence of Hermite polynomials on the real line, satisfying

$$\frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} H_n(y) H_m(y) \, \mathrm{e}^{-y^2/2} dy = \delta_{nm}, \qquad n, m \geq 0,$$

where δ_{nm} is the Kronecker symbol. Next, denote by $\mathbf{y} = (y_k)_{k\geq 1} \in \mathbb{R}^{\mathbb{N}_0}$ any infinite sequence of real variables, and by $\boldsymbol{\nu} = (\nu_k)_{k\geq 1} \in \mathbb{N}^{\mathbb{N}_0}$ any infinite sequence of integers which is “finite”, i.e., such that $\nu_k > 0$ only for a finite number of indices; let $|\boldsymbol{\nu}| = \sum_{k\geq 1} \nu_k$. Define the multidimensional Hermite polynomials of order $|\boldsymbol{\nu}|$ as $H_{\boldsymbol{\nu}}(\mathbf{y}) = \prod_{k=1}^{\infty} H_{\nu_k}(y_k)$; note that the definition is meaningful since $H_0(y) \equiv 1$, hence, $H_{\boldsymbol{\nu}}(\mathbf{y})$ actually depends only on a finite number of components of $\mathbf{y}$. These polynomials are mutually orthonormal, in the following sense:

$$(H_{\boldsymbol{\nu}}, H_{\boldsymbol{\mu}}) := \prod_{k=1}^{\infty} \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} H_{\nu_k}(y_k) H_{\mu_k}(y_k) \, \mathrm{e}^{-y_k^2/2} dy_k = \delta_{\boldsymbol{\nu}\boldsymbol{\mu}}, \quad \forall \boldsymbol{\nu}, \boldsymbol{\mu}.$$

Setting $\mathbf{Y}(\omega) := (Y_k(\omega))_{k\geq 1}$ for all $\omega \in \mathit{\Omega}$, the random variables $\mathcal{H}_{\boldsymbol{\nu}} : \omega \mapsto H_{\boldsymbol{\nu}}(\mathbf{Y}(\omega))$ are independent and with unit variance, since $\mathbb{E}[\mathcal{H}_{\boldsymbol{\nu}} \mathcal{H}_{\boldsymbol{\mu}}] = (H_{\boldsymbol{\nu}}, H_{\boldsymbol{\mu}}) = \delta_{\boldsymbol{\nu}\boldsymbol{\mu}}$, $\forall \boldsymbol{\nu}, \boldsymbol{\mu}$. They form the so-called *Wiener chaos* (sometimes termed *homogeneous chaos* or *Hermite chaos*). The Cameron-Martin theorem states that the family $\{\mathcal{H}_{\boldsymbol{\nu}}\}$ so defined forms an orthonormal basis of the space $L^2(\Omega, dP)$ of the second order random variables over a Gaussian space. The precise result is as follows.

Theorem 5.1 *Let* $\Phi \in L^2(\Omega, dP)$ *and let* $\Phi_{\boldsymbol{\nu}} = \mathbb{E}[\Phi \mathcal{H}_{\boldsymbol{\nu}}]$ *for any finite* $\boldsymbol{\nu}$. *Then,* $\Phi = \sum_{\boldsymbol{\nu} \text{ finite}} \Phi_{\boldsymbol{\nu}} \mathcal{H}_{\boldsymbol{\nu}}$ *in* $L^2(\Omega, dP)$.

This means, for instance, that we have $\mathbb{E}\left[\left(\Phi - \sum_{|\boldsymbol{\nu}|\leq N} \Phi_{\boldsymbol{\nu}} \mathcal{H}_{\boldsymbol{\nu}}\right)^2\right] \to 0$ as $N \to \infty$.

The Cameron-Martin theorem states that $\Phi(\omega)=\varphi(\mathbf{Y}(\omega))$, where $\varphi:\mathbb{R}^{\mathbb{N}_0}\to\mathbb{R}$ is formally defined as $\varphi(\mathbf{y})=\sum_{\boldsymbol{\nu}\text{ finite}}\Phi_{\boldsymbol{\nu}}H_{\boldsymbol{\nu}}(\mathbf{y})$. In many situations of interest, Φ will be possible to express using a finite number of random variables $Y_k(\omega)$, say using $\mathbf{Y}_K(\omega):=(Y_1(\omega),\dots,Y_K(\omega))$; then, $\Phi(\omega)=\varphi(\mathbf{Y}_K(\omega))$ with $\varphi:\mathbb{R}^K\to\mathbb{R}$ defined as $\varphi(\mathbf{y})=\sum_{\boldsymbol{\nu}\in\mathbb{N}^K}\Phi_{\boldsymbol{\nu}}H_{\boldsymbol{\nu}}(\mathbf{y})$ for $\mathbf{y}\in\mathbb{R}^K$ and satisfying

$$\frac{1}{(\sqrt{2\pi})^K}\int_{\mathbb{R}^K}\varphi^2(\mathbf{y})\,\mathrm{e}^{-\mathbf{y}^T\mathbf{y}/2}d\mathbf{y}<+\infty.$$

Thus, for our variable Φ, the condition $\Phi\in L^2(\Omega,dP)$ is equivalent to $\varphi\in L^2_\varrho(\mathbb{R}^K)$, where the weight function ϱ is defined as $\varrho(\mathbf{y})=\frac{1}{(\sqrt{2\pi})^K}\mathrm{e}^{-\mathbf{y}^T\mathbf{y}/2}$. The variable $\mathbf{y}$ will be termed the *stochastic* variable, whereas the spatial variables $\mathbf{x}$ and s will be referred to as the *deterministic* variables.

So far, we have focussed on Gaussian random variables. Similar representations can be given for second order random variables over other probabilistic spaces admitting a density function. The system of orthonormal polynomials which gives rise to a *generalized polynomial chaos*, similar to the Wiener chaos, is determined by the density function; for instance, the uniform density obviously leads to the Legendre polynomials. We refer to [4] for more details.

In general terms, a second order random variable Φ depending on a finite number K of mutually independent real random variables $Y_1(\omega),\dots,Y_K(\omega)$ with zero mean and unit variance with respect to a density function ρ, can be represented as

$$\Phi(\omega)=\varphi(\mathbf{Y}_K(\omega)),\qquad \mathbf{Y}_K(\omega):=(Y_1(\omega),\dots,Y_K(\omega)),\tag{1}$$

where $\varphi=\varphi(\mathbf{y})$ satisfies $\varphi\in L^2_\varrho(\mathbf{I})$: here, $\mathbf{I}=I^K$, where I is the interval of the real line on which ρ is defined, and $\varrho(\mathbf{y})=\prod_{k=1}^K\rho(y_k)$. Since $L^2_\varrho(\mathbf{I})=\bigotimes_{k=1}^K L^2_\rho(I)$, a natural orthonormal basis $\{\psi_{\boldsymbol{\nu}}\}_{\boldsymbol{\nu}\in\mathbb{N}^K}$ in this space is provided by the tensor product of a one-dimensional family of orthonormal functions $\{\psi_n\}_{n\in\mathbb{N}}$ in $L^2_\rho(I)$; we assume that these functions are algebraic polynomials, as it occurs in the most relevant situations.

6. The deterministic formulation of $\big(\hat{\mathcal{P}}^S\big)$

We go back to the stochastic formulation $\big(\hat{\mathcal{P}}^S\big)$. We assume that the boundary $\Gamma(\omega)$ of $D(\omega)$ depends on ω via K mutually independent real random variables $Y_1(\omega),\dots,Y_K(\omega)$ with zero mean and unit variance with respect to a density function ρ defined on some interval $I\subseteq\mathbb{R}$. Let $\mathbf{Y}_K(\omega)$ and ϱ be defined as above. Since we assumed in Section 4 that $\Gamma(\omega)=\gamma_0(\omega)(\Gamma_0)$, equation (1) easily yields $\gamma_0(\omega)=\gamma_0^*(\mathbf{Y}_K(\omega))$, where $\gamma_0^*=\gamma_0^*(\mathbf{y})$ is a family of $C^{1,p}(\Gamma_0)$-mappings defined in $\mathbf{I}=I^K$, with inverses $\gamma_0^*(\mathbf{y})^{-1}$ in $C^{1,p}(\Gamma^*(\mathbf{y}))$. Thus, $\Gamma^*(\mathbf{y})=\gamma_0^*(\mathbf{y})(\Gamma_0)$ is a parametrization of the set of the admissible boundaries of the stochastic domains $D(\omega)$.

Since $\hat{u}$ and λ depend on ω only through $\Gamma(\omega)$, the Doob-Dynkin lemma assures that this dependence takes place via $\mathbf{Y}_K(\omega)$, i.e., we have $\hat{u}(\,\cdot\,,\omega)=\hat{u}^*(\,\cdot\,,\mathbf{Y}_K(\omega))$

and $\lambda(\,\cdot\,,\omega) = \lambda^*(\,\cdot\,,\mathbf{Y}_K(\omega))$, where $\hat{u}^*(\,\cdot\,,\mathbf{y}) \in H_0^1(\hat{D})$ and $\lambda^*(\,\cdot\,,\mathbf{y}) \in H^{-1/2}(\mathit{\Gamma}^*(\mathbf{y}))$, a.e. in $\mathbf{I}$. Condition $\hat{u} \in L^2(\Omega, dP; H_0^1(\hat{D}))$ is then equivalent to $\hat{u}^* \in L^2_\varrho(\mathbf{I}; H_0^1(\hat{D}))$; similarly, $\lambda \in L^2(\Omega, dP; H^{-1/2}(\mathit{\Gamma}))$ is equivalent to $\lambda^* \in L^2_\varrho(\mathbf{I}; H^{-1/2}(\mathit{\Gamma}^*))$ (with obvious meaning of the notation).

We now recall the formula $\mathbb{E}\,[\Phi] = \int_{\mathbf{I}} \varphi(\mathbf{y})\varrho(\mathbf{y})\,d\mathbf{y}$ which holds for all random variables $\Phi(\omega) = \varphi(\mathbf{Y}_K(\omega))$ with $\varphi \in L^1_\varrho(\mathbf{I})$. By applying this formula several times, we transform the stochastic problem $\big(\hat{\mathcal{P}}^S\big)$ into the following deterministic problem: Find $\hat{u}^* \in L^2_\varrho(\mathbf{I}; H_0^1(\hat{D}))$ and $\lambda^* \in L^2_\varrho(\mathbf{I}; H^{-1/2}(\mathit{\Gamma}^*))$ such that

$$\left\{\begin{array}{l} \int_{\mathbf{I}}\int_{\hat{D}} \nabla\hat{u}^* \cdot \nabla v^*\,d\mathbf{x}\,\varrho(\mathbf{y})\,d\mathbf{y} + \int_{\mathbf{I}} \langle\lambda^*, \tau v^*\rangle_{\mathit{\Gamma}^*(\mathbf{y})}\varrho(\mathbf{y})\,d\mathbf{y} = \int_{\mathbf{I}}\int_{\hat{D}} f v^*\,d\mathbf{x}\,\varrho(\mathbf{y})\,d\mathbf{y}, \\ \hfill \forall v^* \in L^2_\varrho(\mathbf{I}; H_0^1(\hat{D})), \qquad \big(\hat{\mathcal{P}}^D\big) \\ \int_{\mathbf{I}} \langle\mu^*, \tau\hat{u}^*\rangle_{\mathit{\Gamma}^*(\mathbf{y})}\varrho(\mathbf{y})\,d\mathbf{y} = 0, \quad \forall\mu^* \in L^2_\varrho(\mathbf{I}; H^{-1/2}(\mathit{\Gamma}^*)). \end{array}\right.$$

7. Discretization of the deterministic formulation

Discretization is accomplished by standard mixed finite elements in the physical variables as in Section 3 and a Galerkin projection method with numerical integration (which coincides with a collocation scheme) in the stochastic variables. Thus instead of solving very large algebraic saddle-point system resulting from the discretization of $\big(\hat{\mathcal{P}}^D\big)$, we will solve n deterministic problems $\big(\hat{\mathcal{P}}\big)$ for n different configurations of the stochastic domain $D(\mathbf{y})$, where n is the number of Gauss (collocation) points $\mathbf{y_q}$. We can simply parallelize all computations. For more details see [2], where a stability and convergence analysis of the method have been presented. We showed that, in any subdomain that does not contain the random boundaries, the convergence is “spectral” in the polynomial chaos order.

8. Numerical examples

In this section, we illustrate the efficiency of our approach on a model example with nonhomogeneous Dirichlet boundary condition for which we do not know an analytic solution. Therefore basic Monte Carlo (MC) simulation without using any special optimization technique is used to validate the result.

Example 1. Let $\hat{D} := (0,1) \times (0,1)$ be the fictitious domain. Let $\mathbf{y} = (y_1, y_2)$ be a stochastic vector variable, associated with two independent normal distributions $Y_k \sim N[\overline{y}, \sigma]$, $k = 1,2$, with $\overline{y} = (a+b)/2$, $\sigma = (b-a)/8$; the density function $\rho(y_k)$ is truncated from $\mathbb{R}$ to the interval $I = [a,b]$, $a = 0.25$ and $b = 0.35$. In a polar coordinate system centered at $\mathbf{x}_0 = (0.5, 0.5)$, consider the control points C_k, $k = 0,\dots,15$, whose angles are $\varphi_k = k\pi/8$ and whose radii are constant, $r_k = 0.3$, except for $k = 5$ and $k = 6$: for these control points, the radii are given by the variables y_1 and y_2, respectively (see Figure 1). The boundary $\mathit{\Gamma}(\mathbf{y})$ is obtained by connecting the control points via a piecewise Bèzier curve of the second order, identified by the Bèzier triples (M_k, C_{k+1}, M_{k+1}), with $M_k = (C_k + C_{k+1})/2$ for $k = 0,\dots,15$ and

$C_{16} = C_0,\ M_{16} = M_0$. All possible configurations of the stochastic domain $D(\mathbf{y})$ are obtained by moving the control nodes C_5 and C_6 along the depicted lines.

We consider the problem

$$\Big\{\ -\triangle u(\mathbf{x},\mathbf{y}) = 60 \ \text{ in } D(\mathbf{y}), \quad u(\mathbf{x},\mathbf{y}) = g \ \text{ on } \mathit{\Gamma}(\mathbf{y}), \qquad (\mathcal{P}(\mathbf{y}))$$

where $g(\varphi) = 0$, $\varphi \in [-\pi, 0]$ and $g(\varphi) = 1 - \cos(2\varphi)$, $\varphi \in (0, \pi)$.

Figures 2 and 3 provide comparisons between the results produced by basic Monte Carlo (MC) simulation, for different numbers of trials N, and second order Polynomial Chaos (PC) results, obtained by solving 9 independent deterministic problems. The results are depicted along the line $L = \{(x_1, \frac{1}{2})|\ x_1 \in [0, 1]\}$. The two vertical dot and dash lines bound the domain $D(\mathbf{y})$ which is fixed in this cross-section for all $\mathbf{y} \in I^2$. While the Monte Carlo approximation of the mean value is good already for moderate numbers of trials, an acceptable approximation of the variance is obtained only with a number of trials in the order of several hundreds.

For more examples and deeper understanding we refer to [2].

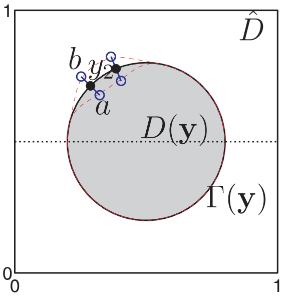


Fig. 1: *Geometry of $D(\mathbf{y})$.*

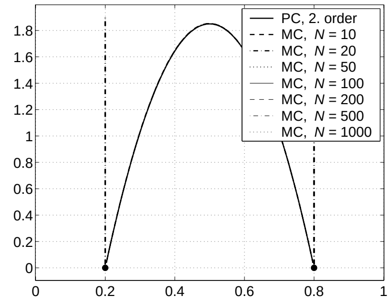


Fig. 2: *PC vs MC, h fixed: mean.*

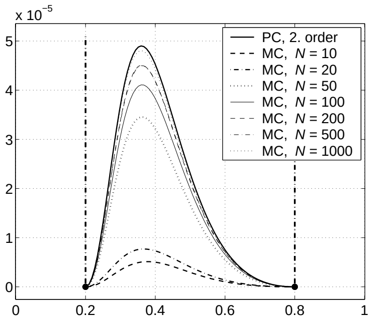


Fig. 3: *PC vs MC, h fixed: variance.*

References

[1] F. Brezzi, M. Fortin: *Mixed and hybrid finite element methods.* Springer-Verlag, New York, 1991.

[2] C. Canuto, T. Kozubek: *A fictitious domain approach to the numerical solution of PDEs in stochastic domains.* Rapporto Interno N. **12**, Politecnico di Torino, 2006, 1–27 (will be published in Numerische Mathematik, 2006).

[3] J. Haslinger, R.A.E. Mäkinen: *Introduction to shape optimization, theory, approximation, and computation.* SIAM, Philadelphia, 2003.

[4] D. Xiu, G.E. Karniadakis: *The Wiener-Askey polynomial chaos for stochastic differential equations.* SIAM J. Sci. Comput. **24**, 2002, 619–644.

ON A SANDIA STRUCTURAL MECHANICS CHALLENGE PROBLEM[*]

Jan Chleboun

1. Introduction

A structural mechanics prediction problem was proposed by Ivo Babuška, Fabio Nobile, and Raul Tempone as one of the uncertain input data problems specially designed to challenge the participants of Validation Challenge Workshop, Sandia National Laboratories, Albuquerque, NM, USA, May 21-23, 2006; see [1].

The prediction problem concerns the structure sketched in Figure 1 (left), the coordinates of the joints are given in meters. The rods are joined by pins (zero moment connections) at the junctions and hinges. The horizontal beam (number 4 in Figure 1 (left)) is loaded by a uniform force. The vertical displacement of P, the midpoint of the horizontal beam, is denoted by δ_P and exaggerated. Since the force acts downward, δ_P is negative; we refer to [1], where $w(P_m) \equiv \delta_P$, for details.

The prediction problem is posed as follows: What is the probability that $\delta_P \geq -3$ mm? Or, in a broader sense: How can we assess the occurrence of the $\delta_P \geq -3$ mm phenomenon?

The difficulty of the problem lies in limited information about material parameter E, the modulus of elasticity (Young modulus) of the truss structure members. The material of the bars and the beam is represented by the Young modulus that is assumed to be a homogeneous random field. The modulus and its probabilistic properties are not known and have to be inferred and characterized from available data. In [1], three embedded sets of data are presented. In this analysis, we confine ourselves to the first, most limited dataset.

It comprises: a vector $E_0^{\mathrm{v}} = (13.26, 10.86, 14.77, 10.94, 11.05)$ of five local values of E in GPa, see the top five values in the third column of [1, Table 6]; a vector $E_{20}^{\mathrm{v}} = (11.65, 11.21, 11.45, 10.89, 11.67)$ of five averaged values of E (in GPa) inferred from the elongation of sample rods 20 cm long (calibration experiments; cross-section area $A = 4$ mm^2, force $F = 1200$ N), see the top five values in the second column of [1, Table 6] for the elongations; a vector $E_{80}^{\mathrm{v}} = (11.94, 11.65)$ of two averaged values of E (in GPa) inferred from the elongation of sample rods 80 cm

[*]This research was supported by the Czech Science Foundation, grant 201/04/1503, and by the Academy of Sciences of the Czech Republic, Institutional Research Plan No. AV0Z10190503. The author thanks Ivo Babuška, Fabio Nobile, and Raul Tempone for fruitful discussions about the challenge problem, and Ivan Saxl for valuable comments on an earlier version of the manuscript.

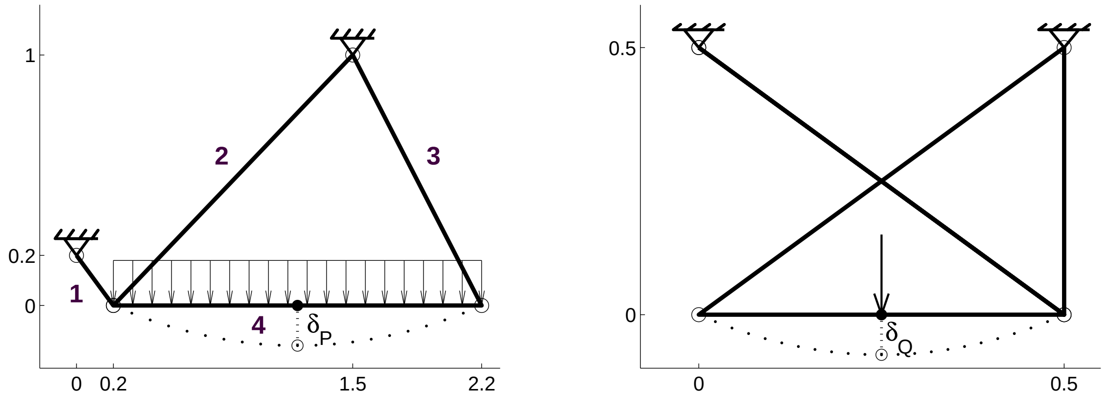


Fig. 1: *Prediction problem (left). Accreditation experiment (right).*

long (validation experiments; $A = 4$ mm^2, $F = 1200$ N), see the top two values in the second column of [1, Table 7]; and δ_Q, a particular displacement observed in a “similar”, point-loaded structure in an *accreditation experiment* [1], see Figure 1 (right) and the first value of $w(P)$ (correctly $w(Q)$) in [1, Table 8]. Vectors E_0^{v}, E_{20}^{v}, and E_{80}^{v} stem from sampling random variables E_0, E_{20}, and E_{80}, respectively.

The geometry of the structures as well as of the individual bars and beams is exactly known. The structures are statically determined, therefore the load-to-displacement mapping can be expressed by an explicit formula, see [1] for details.

2. Analysis

The probability distribution of the Young modulus value is unknown. The number of measurements is not sufficient to allow for strong results of a statistical analysis; the estimates of probability related parameters would be poor. Nevertheless, a stochastic-based approach will be used to tackle the uncertainty problem.

Let us identify the longitudinal axis of each rod with a local coordinate system axis in such a way that the left end of the rod coincides with the origin. Along each rod, the Young modulus is supposed to be a stationary random field, $E(x)$, where $x \in [0, L]$ and L is the length of the rod. Some features of this field are assumed to be independent of x. For example, $\mathbf{E}(E(x))$, the expected value of the Young modulus at point x, is assumed to be constant and independent of x and of a particular choice of the rod; similarly for higher statistical moments. This is why we can identify $E(x)$ with an x-independent random variable E_0 in certain analyses.

We have to assume that $E(x_1)$ and $E(x_2)$ are not mutually independent, especially if x_1 is “close” to x_2. In other words, $E(x_1)$ and $E(x_2)$ are correlated. However, the formula representing the model of correlation is not known. We will *assume* a formula dependent on one parameter, called correlation length, that has to be determined from the available data.

The method of treating the prediction problem can be summarized as follows:

- Choose respective intervals I_0 and I_{20} that exceed the range of the measured values E_0^{v} and E_{20}^{v}. These intervals represent the assumed range of random variables E_0 and E_{20}.
- Assume a probability distribution of E_0 and E_{20} (uniform or normal).
- For $1/E$, assume a covariance function with an unknown correlation length L_{corr}.
- By using the assumptions, calculate the correlation length L_{corr}.
- By knowing L_{corr}, infer I_{80}, an interval representing the range of the random variable E_{80}, and check it against E_{80}^{v}. It is assumed that E_{80} retains the probability distribution of E_0 and E_{20} (uniform or normal).
- By knowing L_{corr}, infer an interval representing the range of the random variable δ_Q and check it against the value of δ_Q coming from the accreditation experiment.
- By knowing L_{corr}, infer an interval representing the range of δ_P and check it against the -3 mm limit given in the prediction problem. Try to make a conclusion.

Inevitably, expert opinion is required to make realistic assumptions needed in the above-listed steps.

The intervals I_0 and I_{20} are constructed to have a common center. They are interpreted as either the respective intervals in which both E_0 and E_{20} are uniformly distributed (i.e., E_0 and E_{20} do not exceed I_0 and I_{20}, respectively), or the intervals covering 95% of normally distributed values E_0 and E_{20} (i.e., the probability that these random quantities leave their intervals is 0.05). The normal distribution assumption can be challenged because normally distributed values of E would allow for a negative Young moduli (with a low probability), which is physically impossible.

We have a double reason for using the normal distribution. First, we wish to compare the results obtained for the uniform probability distribution with the results obtained for a non-uniform distribution. Second, we wish to minimize the use of numerical methods, which is possible for the above-mentioned distributions. Moreover, the normal distribution assumption may still be adequate for understanding the dominant behavior of the rods and structures.

Let us recall that E_0 is a random field of local values of E, that is, a field identical to $E(x)$ except for the localization at a particular x.

We assume that the covariance function of $1/E$ is related to the variance of $1/E$ in a particular way mediated through an L_{corr}-dependent function (cf. [2, Example 1]):

$$\operatorname{cov}\bigl[1/E(x_1), 1/E(x_2)\bigr] = \operatorname{var}\bigl(1/E_0\bigr)\exp\bigl(-|x_1 - x_2|/L_{\mathrm{corr}}\bigr). \tag{1}$$

If a bar of length L and cross-section area A is axially loaded by a force F, then for δ_L, its elongation, holds

$$\delta_L = \frac{F}{A}\int_0^L \frac{1}{E(x)}\,dx. \tag{2}$$

Since $E(x)$ is a random variable, δ_L is a random variable, too.

By (2) used in $\mathrm{var}(\delta_L) = \mathbf{E}\big[\delta_L^2\big] - (\mathbf{E}[\delta_L])^2$ and by (1), we infer

$$\begin{aligned}\mathrm{var}(\delta_L) &= \frac{F^2}{A^2}\int_0^L\int_0^L \Big\{\mathbf{E}\big[1/E(x_1), 1/E(x_2)\big] - \big(\mathbf{E}\big[1/E_0\big]\big)^2\Big\}\, dx_1\, dx_2 \\ &= \frac{F^2}{A^2}\int_0^L\int_0^L \mathrm{cov}\big[1/E(x_1), 1/E(x_2)\big]\, dx_1\, dx_2 \\ &= \frac{F^2}{A^2}\,\mathrm{var}\big(1/E_0\big)\int_0^L\int_0^L \exp\Big(-\frac{|x_1-x_2|}{L_{\mathrm{corr}}}\Big)\, dx_1\, dx_2. \qquad (3)\end{aligned}$$

If we define E_L as the *effective* modulus of elasticity inferred from the prolongation of the bar of length L via the equality $\delta_L = FL/(AE_L)$, then E_L is also a random variable and its variance can be calculated as $\mathrm{var}(\delta_L) = \mathrm{var}(1/E_L)F^2L^2/A^2$. By this equality combined with (3), we eliminate $\mathrm{var}(\delta_L)$ and obtain

$$\frac{\mathrm{var}(1/E_L)}{\mathrm{var}(1/E_0)} = \frac{1}{L^2}\int_0^L\int_0^L \exp\left(-\frac{|x_1-x_2|}{L_{\mathrm{corr}}}\right) dx_1\, dx_2. \qquad (4)$$

To solve (4), we evaluate $\mathrm{var}(1/E_0)$ stemming from the assumed probability distribution of E_0 in the interval I_0. We evaluate $\mathrm{var}(1/E_L)$ for $L = 20$ cm in a similar way by using assumptions about E_{20} and I_{20}. After exact integration (done analytically by `Maple`), the right-hand side of (4) becomes

$$2z + 2z^2(\exp(-1/z) - 1), \text{ where } z = L_{\mathrm{corr}}/L, \qquad (5)$$

an explicit function of L_{corr} and L. By using (5) and by fixing $L = 20$ cm, we can numerically solve (4) for L_{corr}.

As soon as $\mathrm{var}(1/E_0)$ is inferred from the assumptions and L_{corr} is known from (4), we can use (4) to directly calculate $\mathrm{var}(1/E_L)$ for $L = 80$ cm and the other bar lengths appearing in the truss structures. We assume that $\mathrm{var}(1/E_{80})$ corresponds to either a uniform or normal distribution of E_{80}. Under these assumptions, we can infer I_{80} as either the entire range of a uniformly distributed random variable E_{80} or the 95% confidence interval of normally distributed random variable E_{80}, and check whether or not the validation data lie in I_{80}.

To obtain the vertical displacements δ_Q and δ_P, the axial elongation of the rods has to be combined with the bending of the transversally loaded beams, see [1] for details. The bending is expressed by the Green function technique. As a consequence, to compute the corresponding variance of the vertical displacements δ_Q and δ_P, integrals such as

$$\int_0^L\int_0^L \phi(x_1)\psi(x_2)\exp\Big(-\frac{|x_1-x_2|}{L_{\mathrm{corr}}}\Big)\, dx_1\, dx_2 \qquad (6)$$

have to be evaluated. In the most complex setting of (6), ϕ and ψ are continuous piece-wise quadratic (in the accreditation experiment) or cubic (in the prediction problem) functions. Again, `Maple` is able to analytically integrate expression (6). In a similar (but simpler) way, the respective means of δ_Q and δ_P can be calculated through the knowledge of $\int_0^L \phi(x_1)\psi(x_2)\,dx_1\,dx_2$ and $\mathbf{E}[1/E_0]$.

Since we have made various assumptions, it will be useful to parametrize at least some of them to make the model partially parameter-dependent. By playing with the values of the parameters and by analyzing the model response, we hope to get at least some insight into the impact of uncertain inputs on the prediction problem truss behavior.

Let us define E_{m} as the mean of all the measured values of the Young modulus, that is E_0^{v}, E_{20}^{v}, and E_{80}^{v} taken together, twelve values in total. Let us introduce three fundamental parameters: E_{coef}, E_0^{ratio}, and I_{20}-to-I_0 ratio. The first parameter stands for a multiplicative coefficient that is used to control the centers of intervals I_0, I_{20}, and I_{80} that are defined as coincident with $E_{\mathrm{coef}}E_{\mathrm{m}}$. The second parameter, E_0^{ratio}, is related to the distance between E_0^{v} (the measured values) and the ends of the interval I_0 that covers E_0^{v}. In detail, if I_{c} is the complement of I_0 in the set of real numbers, then $E_{\mathrm{ratio}} = \mathrm{dist}(I_{\mathrm{c}}, E_0^{\mathrm{v}})/l_0$, where l_0 is the difference between the maximum and the minimum of the measured values E_0^{v}. Finally, the I_{20}-to-I_0 ratio is simply the ratio of the length of I_{20} to the length of I_0.

Let us comment on Figure 2. The uppermost graph depicts the measured values E_0^{v}, E_{20}^{v}, and E_{80}^{v} (marked by ×) as well as the respective intervals I_0, I_{20}, and I_{80} they are embedded in. Unlike I_0 and I_{20}, which are assumed, I_{80} is calculated from the

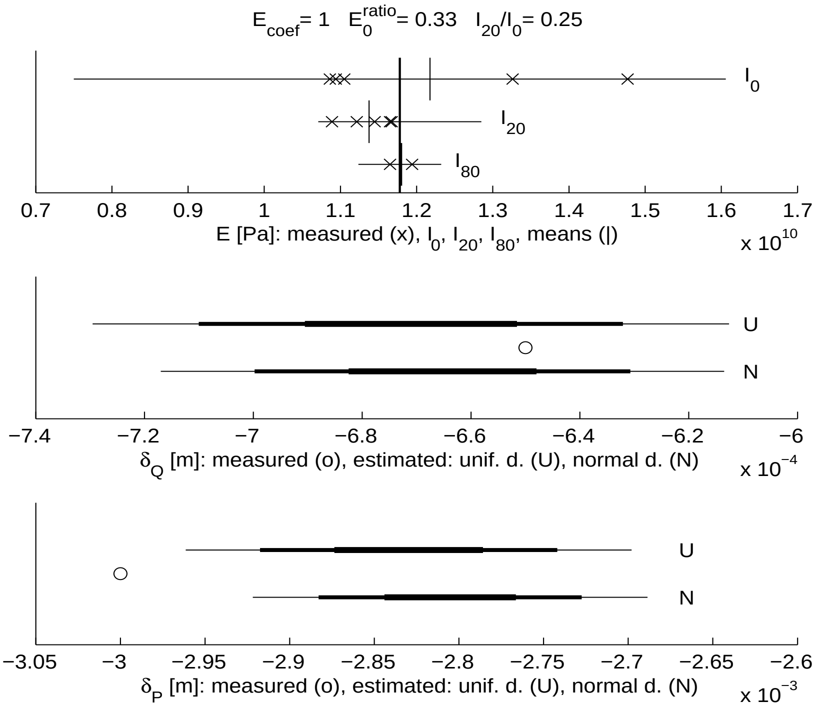


Fig. 2: *Model outputs for fixed parameters.*

assumptions and (4). Two intervals I_{80} should be depicted; one determined by the uniform distribution assumption, the other determined by the normal distribution assumption. Since, however, they almost coincide, only one line appears in the graph. The means of the measured values are marked by short vertical lines. The long vertical line marks the average value of the Young modulus we *assume* and calculate with, i.e., $E_{\mathrm{coef}}E_{\mathrm{m}}$. Note that E^{v}_{20} comprises five values but two of them almost coincide.

The middle graph shows the measured δ_Q marked by a small circle, and the estimated intervals for δ_Q constructed from the mean of δ_Q and the standard deviation of δ_Q inferred via the method outlined on the previous pages. The width of the lines marks the intervals determined by the mean and the first three multiples of (plus and minus) the standard deviation. Intervals stemming from the uniform (U, upper line) and normal (N, lower line) distribution of E_0 are drawn.

In Figure 2, the last graph is a parallel to the graph described in the previous paragraph, this time for δ_P, however. We see that the mean of δ_P is greater than the given acceptable limit of δ_P (-3 mm, marked by $\circ$). Indeed, it is more than three standard deviation values "on the safe side" even in the case of uniformly distributed Young modulus values.

Figure 3 shows what happens if we let the I_{20}-to-I_0 ratio change. In other words, we fix the interval encompassing the measured local Young moduli and we let the interval I_{20} get larger and larger. As a consequence, the inferred interval I_{80} becomes larger too, and the possible ranges of δ_Q and δ_P also increase. In Figure 3, the two thin lines depict E^{ratio}_{20} and E^{ratio}_{80}. These quantities have the meaning similar to that of E^{ratio}_{0}, but are defined by means of the pairs I_{20}, E^{v}_{20}, and I_{80}, E^{v}_{80}. The larger the ratio, the larger the distance between the measured values of the Young modulus and the ends on the respective intervals I_{20} and I_{80}. The two assumptions on the random variable distribution lead to two graphs of E^{ratio}_{80}. Since they almost coincide, only one line is depicted in Figure 3.

The "accreditation" dash and dash-dotted lines show the ratio of the distance between the measured δ_Q value and the calculated average of δ_Q to the standard deviation of δ_Q. Again, two assumed distributions of E are considered (u, uniform; n, normal).

The two thick lines that graph negative values depict the ratio of the difference between the limit displacement of -3 mm and the calculated average of δ_P to the standard deviation of δ_P. The uniform distribution assumption leads to a worse separation from the limit displacement than the normal distribution assumption. We observe that if the I_{20}-to-I_0 ratio increases, the distance between the set limit (-3 mm) and the calculated average decreases, that is, the probability that $\delta_P \leq -3$ mm increases.

The $*$ and $\Box$ symbols stem from the Chebyshev inequality [3, Section 33.10, inequality (3)], that is, they mark an upper bound on the probability that $\delta_P \leq -3$ mm; the estimates are multiplied by 10 in Figure 3. The values depend on the assumed distributions of E (u, uniform; n, normal).

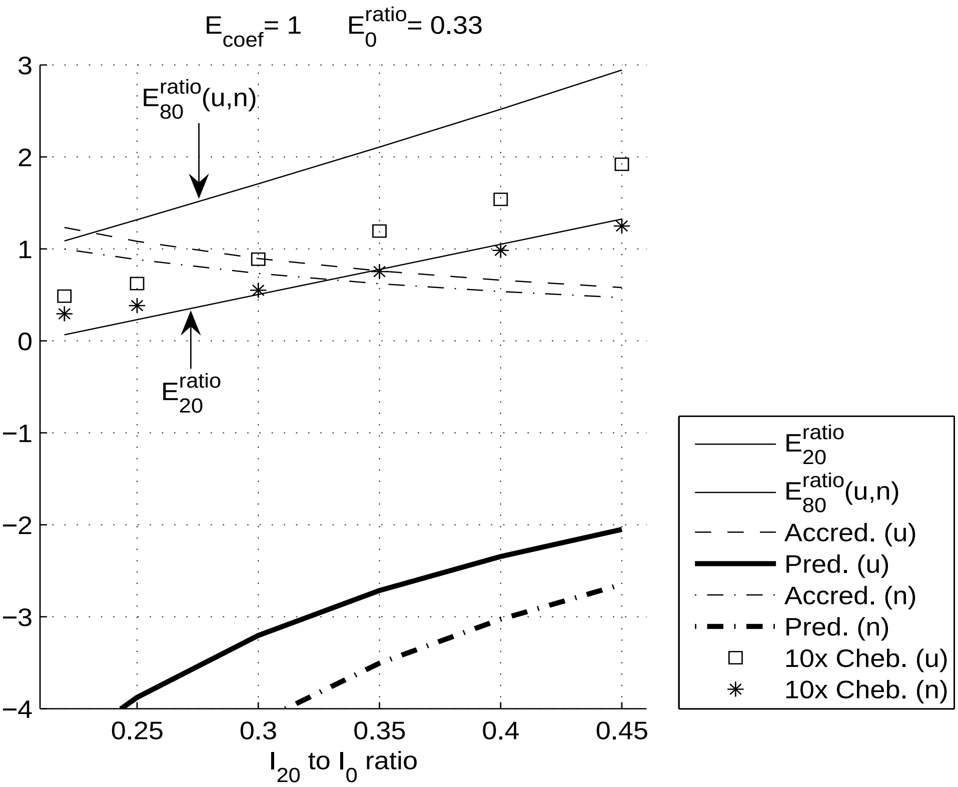


Fig. 3: *Model outputs for variable parameters.*

The graphs in Figure 2 and Figure 3 might indicate that the probability of reaching or exceeding the limit displacement in the prediction problem is sufficiently low even if we allow for rather large intervals to cover E_0^{v} and E_{20}^{v}. However, the graphs corresponding to perturbed values of E_{coef} (not displayed here) reveal a substantial sensitivity of outputs to the assumed average of the Young modulus. Its decrease ($E_{\mathrm{coef}} = 0.98$, for instance) seems to be acceptable from the view of the measured data, but brings the predicted average displacement closer to the limit (less than three standard deviations). To make a more definite conclusion on the prediction problem solution, more data from measurements would be needed.

References

[1] I. Babuška, F. Nobile, R. Tempone: *Model validation challenge problem: static frame problem.*
`http://www.esc.sandia.gov/VCWwebsite/MechanicsProblemDescrip.pdf`

[2] S. Rahman, B.N. Rao: *An element-free Galerkin method for probabilistic mechanics and reliability.* Int. J. Solids Struct. **38**, 2001, 9313–9330.

[3] K. Rektorys et al.: *Survey of applicable mathematics II.* Prometheus, Praha, 2000 (in Czech).

A SECOND ORDER UNCONDITIONALLY POSITIVE SPACE-TIME RESIDUAL DISTRIBUTION METHOD FOR SOLVING COMPRESSIBLE FLOWS ON MOVING MESHES*

Jiří Dobeš, Herman Deconinck

Abstract

A space-time formulation for unsteady inviscid compressible flow computations in 2D moving geometries is presented. The governing equations in Arbitrary Lagrangian-Eulerian formulation (ALE) are discretized on two layers of space-time finite elements connecting levels n, $n + 1/2$ and $n + 1$. The solution is approximated with linear variation in space (P1 triangle) combined with linear variation in time. The space-time residual from the lower layer of elements is distributed to the nodes at level $n+1/2$ with a limited variant of a positive first order scheme, ensuring monotonicity and second order of accuracy in smooth flow under a time-step restriction for the timestep of the first layer. The space-time residual from the upper layer of the elements is distributed to both levels $n + 1/2$ and $n + 1$, with a similar scheme, giving monotonicity without any time-step restriction. The two-layer scheme allows a time marching procedure thanks to initial value condition imposed on the first layer of elements. The scheme is positive and second order accurate in space and time for arbitrary meshes and it satisfies the Geometric Conservation Law condition (GCL) by construction.

Example calculations are shown for the Euler equations of inviscid gas dynamics, including the 1D problem of gas compression under a moving piston and transonic flow around an oscillating NACA0012 airfoil.

1. Introduction

Residual Distribution (RD) schemes have reached a certain level of maturity for the simulation of steady flow problems. The RD approach allows to construct second order methods on a compact stencil, which are positive at the same time. They are used as state of the art methods to solve complex steady problems e.g. 3D turbulent Navier-Stokes equations or Magneto-Hydro-Dynamic equations [5, 2, 1].

In [9] it has been noted that for an unsteady computation a mass matrix coupling space and time discretizations has to be taken into the account, otherwise the *spatial* accuracy is lost. This matrix is not a M-matrix, hence if inverted, the positivity of the spatial discretization is compromised.

In [10, 1, 3] an alternative approach for unsteady problems has been proposed, based on space-time RD schemes for a bilinear space-time element approximation. In particular, in [10, 1] a first order scheme corresponding to the N scheme with Crank-Nicholson time integration has been shown to be positive under a time step restriction. This restriction can be overcome by adding one more time layer [3].

*This research was partially supported by Research plan of MSMT no. 6840770003.

An extension of the conditionally positive, one layer method for moving meshes was presented in [6]. In this paper we extend the two layer method for computations on moving meshes. Because the underlying scheme can be written as the modification of the spatial N scheme with Crank-Nicholson time integration, we use the Arbitrary Lagrangian-Eulerian formulation of the RD method [11].

2. ALE formulation

We define the ALE mapping which for each $t \in I$ associates a point $\vec{Y}$ of reference configuration Ω_0 to a point $\vec{x}$ on the current domain configuration Ω_t, $\mathcal{A}_t : \Omega_0 \subset \mathbb{R}^d \to \Omega_t \subset \mathbb{R}^d$, $\vec{x}(\vec{Y}, t) = \mathcal{A}_t(\vec{Y})$. The ALE mapping $\mathcal{A}_t$ is chosen sufficiently smooth and invertible with nonzero determinant of Jacobian $J_{\mathcal{A}_t}$. A domain velocity $\vec{w}(\vec{x}, t)$ is defined as the time derivative of $\vec{x}$ for constant $\vec{Y}$. We start from the conservative ALE formulation of the Euler equations in d spatial dimensions

$$\frac{1}{J_{\mathcal{A}_t}} \frac{\partial J_{\mathcal{A}_t} \mathbf{u}}{\partial t}\bigg|_{\vec{Y}} + \nabla_x \cdot [\vec{\mathbf{f}}(\mathbf{u}) - \mathbf{u}\vec{w}] = 0\,, \tag{1}$$

where $\mathbf{u} = (\rho, \rho v_i, E)^T$ is the vector of conserved variables and $\vec{\mathbf{f}}(\mathbf{u})$ the well known vector of flux functions. The system is closed with the equation for a perfect gas. The problem is equipped with an appropriate set of initial and boundary conditions. The following equality, called geometrical conservation law, will be used later

$$\nabla_x \cdot \vec{w} = \frac{1}{J_{\mathcal{A}_t}} \frac{\partial J_{\mathcal{A}_t}}{\partial t}\bigg|_{\vec{Y}}. \tag{2}$$

The RD schemes operate on the quasi-linear form of the equation, which can be obtained with $\nabla_x \cdot (\mathbf{u}\vec{w}) = \vec{w} \cdot \nabla_x \mathbf{u} + \mathbf{u}\nabla_x \cdot \vec{w}$ and identity (2)

$$\frac{1}{J_{\mathcal{A}_t}} \frac{\partial J_{\mathcal{A}_t} \mathbf{u}}{\partial t}\bigg|_{\vec{Y}} + \left(\frac{\partial \vec{\mathbf{f}}}{\partial \mathbf{u}} - I\vec{w}\right) \cdot \nabla_x \mathbf{u} - \frac{\mathbf{u}}{J_{\mathcal{A}_t}} \frac{\partial J_{\mathcal{A}_t}}{\partial t}\bigg|_{\vec{Y}} = 0. \tag{3}$$

3. Numerical scheme

The problem is solved on mesh $\mathcal{T}^h$ consisting of simplex elements $\{E\}$. The unknowns are stored in the vertices of the mesh. A straightforward application of the N scheme [6] with Crank-Nicholson time integrator operating between layers n and $n + 1/2$ (i.e. lower layer of the elements) to the problem (3) gives

$$\frac{S_i^{n+1/2} u_i^{n+1/2} - S_i^n u_i^n}{\Delta t^{\text{lower}}} + \sum_{E \in \mathcal{D}_i} \frac{1}{2} \left[\left(k_i^+ (u_i - u_{\text{in}})\right)^{n+1/2} + \left(k_i^+ (u_i - u_{\text{in}})\right)^n \right]^E -$$

$$- \frac{u_i^{n+1/2} + u_i^n}{2} \frac{S_i^{n+1/2} - S_i^n}{\Delta t^{\text{lower}}} = 0,\ k_i = \overline{\left(\frac{\partial \vec{\mathbf{f}}}{\partial \mathbf{u}} - I\vec{w}\right)} \cdot \frac{\vec{n}_i}{d},\ u_{\text{in}} = -(\sum_{i \in E} k_i^+)^{-1} \sum_{j \in E} k_j^- u_j, \tag{4}$$

where S_i is the area of median dual cell around node i, $\mathcal{D}_i$ denote all the elements sharing node i, u_i^n is the solution in node i at time level n, Δt is the time-step, $\vec{n}_i$ is the normal to the face opposite to the node i scaled by its surface and k_i^+ is the positive part of the upwind matrix k_i in the sense of its eigen-decomposition. Note that the Jacobian includes the mesh velocity. The Jacobian and mesh velocity $\vec{w}$ are taken in an averaged state, such that the resulting method is conservative [11, 4]. Note, that the method presented here is different from [11] in the treatment of the source term, what allows us to show the positivity of the scheme for scalar problems under the time-step restriction

$$\Delta t^{\text{lower}} \leq \frac{\mu(E^{n+1/2}) + \mu(E^n)}{k_i^{+,E}(d+1)}, \qquad \forall i, E \in \mathcal{T}^h, \tag{5}$$

where $\mu(E)$ is the volume of element E. This method can be interpreted as a space-time method, distributing *space-time* nodal contribution $\phi_i^{E^{\text{ST}}}$ from the lower layer of the elements (element between levels n and $n+1/2$) to the nodes at level $n+1/2$

$$u_i^{n+1/2,m+1} = u_i^{n+1/2,m} - \alpha_i \sum_{E \in \mathcal{D}_i} \phi_i^{E^{\text{ST}},\text{lower,n+1/2}}, \tag{6}$$

where α_i is relaxation parameter given by the explicit stability constraint.

Although the method is implicit, it suffers from the time-step restriction (5). As a cure, we add a second layer of elements with similar scheme, operating between levels $n+1/2$ and $n+1$. This scheme distributes portions of the space-time residual of the upper layer as follows:

- To the nodes at $n+1$:

$$\phi_i^{E^{\text{ST}},\text{upper,n+1}} = \mu(E_i^{n+1})u_i^{n+1} - \mu(E_i^{n+1/2})u_i^{n+1/2} +$$
$$+\Delta t^{\text{upper}} \sum_{E \in \mathcal{D}_i} \frac{1}{2}\left(k_i^+(u_i - u_{\text{in}})\right)^{n+1,E} - \frac{u_i^{n+1} + u_i^{n+1/2}}{2}\Big(\mu(E_i^{n+1}) - \mu(E_i^{n+1/2})\Big). \tag{7}$$

- To the nodes at $n+1/2$:

$$\phi_i^{E^{\text{ST}},\text{upper,n+1/2}} = \Delta t^{\text{upper}} \sum_{E \in \mathcal{D}_i} \frac{1}{2}\left(k_i^+(u_i - u_{\text{in}})\right)^{n+1/2,E}. \tag{8}$$

Relaxation procedure (6) has then the form

$$u_i^{n+1/2,m+1} = u_i^{n+1/2,m} - \alpha_i \sum_{E \in \mathcal{D}_i} \left(\phi_i^{E^{\text{ST}},\text{lower,n+1/2}} + \phi_i^{E^{\text{ST}},\text{upper,n+1/2}}\right) \tag{9}$$

$$u_i^{n+1,m+1} = u_i^{n+1,m} - \alpha_i \sum_{E \in \mathcal{D}_i} \phi_i^{E^{\text{ST}},\text{upper,n+1}} \tag{10}$$

and the scheme is formally unconditionally stable with arbitrary Δt^{upper}.

The space-time nodal contribution can be seen as a space-time residual distributed with (implicitly defined) distribution coefficient

$$\phi_i^{E^{\rm ST}} = \beta_i \phi^{E^{\rm ST}}, \qquad \sum_{i \in E} \beta_i = 1. \tag{11}$$

The scheme described above is at most first order accurate. As it was proven in [1], a condition for second order of accuracy is the uniform boundedness of the distribution coefficients β_i. One of the possibilities to modify the distribution coefficients is [1]

$$\beta_i^{\rm mod} = \frac{\beta_i^+}{\sum_{j \in E} \beta_j^+}. \tag{12}$$

This modification preserves the sign of the distribution coefficients and ensures its uniform boundedness, hence the method becomes second order accurate, while keeping its positivity. In the case of the Euler equations the modification of the distribution coefficients is performed on *simple waves* given by the projection of the residual to the Jacobian eigenvectors [1].

4. Numerical results

The first test case is motivated by an internal aerodynamics problem, namely flow in a piston engine. A gas at rest is enclosed between two opposite walls in the chamber. One of the walls slowly starts to move, compressing the gas inside the chamber. This problem can be solved by the method of characteristics [12] until the head of the pressure wave reflects from the end wall or a shock is created[1]. We have used a rectangular domain of size 5 × 1 with initial conditions $u_0 = 0$, $\rho_0 = 1.4$ and $p_0 = 1$. The piston starts to accelerate with derivative of acceleration $\dddot{x} = 0.2$. The numerical solution is plotted at time $t = 4$, when the piston has reached $x = 2.13\bar{3}$. The mesh consist of 372 nodes and 674 triangular elements with 30 nodes along the cylinder wall and 6 nodes along the end wall. Comparison is made with a finite volume method using a linear least square reconstruction, Barth's limiter, three point backward differentiation scheme on moving meshes [7] (Fig. 1). The solution given by the RD scheme perfectly follows the analytical solution, while the FV scheme gives bigger differences.

The next problem involves a piston instantaneously accelerated to a uniform speed. From the Rankine-Hugoniot jump conditions we can compute the solution analytically.[2] The comparison is shown in Fig. 2 at $t = 2$. Note the perfectly monotone shock capturing. Both the FV and RD schemes give comparable results. Note also the entropy layer in the vicinity of the piston, which is present both for RD and FV methods. Its source has to be still investigated.

[1] The analytical solution is avaliable on an email request. Email: `Jiri.Dobes@fs.cvut.cz`.

[2] Piston velocity is 0.8, flow velocity is $u_L = 0.8$, $u_R = 0$, density is $\rho_L = 2.8191$, $\rho_R = 1.4$ and pressure is $p_L = 2.78$, $p_R = 1$. Shock speed is 0.79461.

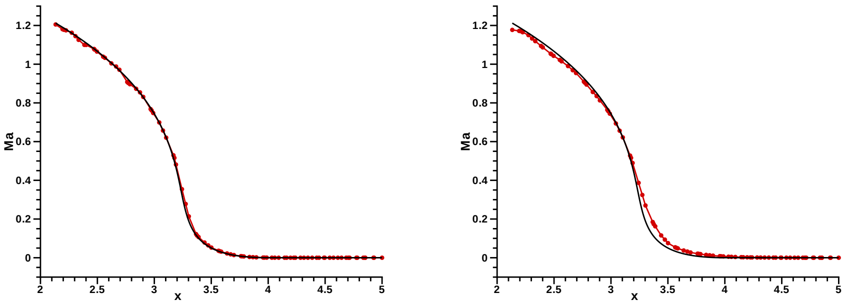


Fig. 1: *Smooth compression of the gas, Mach number cut. Left: present scheme. Right: FV scheme.*

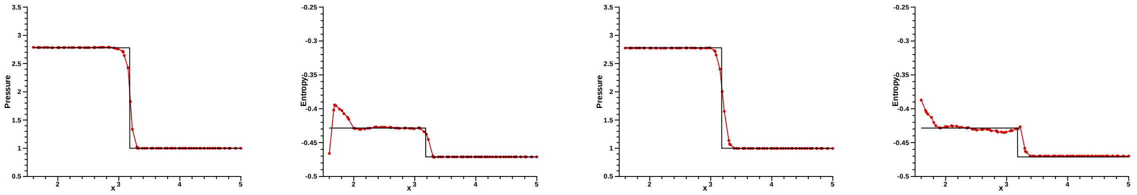


Fig. 2: *Compression of the gas with a shock. Pressure and entropy cut. Left half: present scheme. Right half: FV scheme.*

Finally, a fully 2D test involves a NACA 0012 airfoil which is sinusoidally pitching around its a quarter chord (test case AGARD CT 5[8]). The free stream Mach number is 0.755 and the mean angle of incidence is 0.016° . The airfoil performs a sinusoidal pitching motion with an amplitude of 2.51°

$$\alpha = 2.51 \sin(2kt) + 0.016, \tag{13}$$

where k is the reduced frequency of oscillation with respect to the half chord

$$k = \frac{\omega c}{2u_\infty} = 0.0814, \tag{14}$$

where c is the chord, u_∞ is the free-stream velocity and ω the frequency.

The problem was solved on an unstructured mesh consisting of 5711 nodes and 11153 elements with 206 nodes around the airfoil. The free stream boundary was located 20 chords away from the airfoil. The solution at time $t = 115$ is plotted in Fig. 3. The FV solution is plotted by a dotted line, while the RD solution is plotted as the continuous lines. The FV solution is more dissipative, as one can notice above the profile, where the RD isolines are more crisp and running straight into the shock. Interesting is a comparison of the lift coefficient depending on the angle of incidence. On the zoom, one can notice a higher peak of the lift given by the RD method than by the FV method, which points to the higher accuracy.

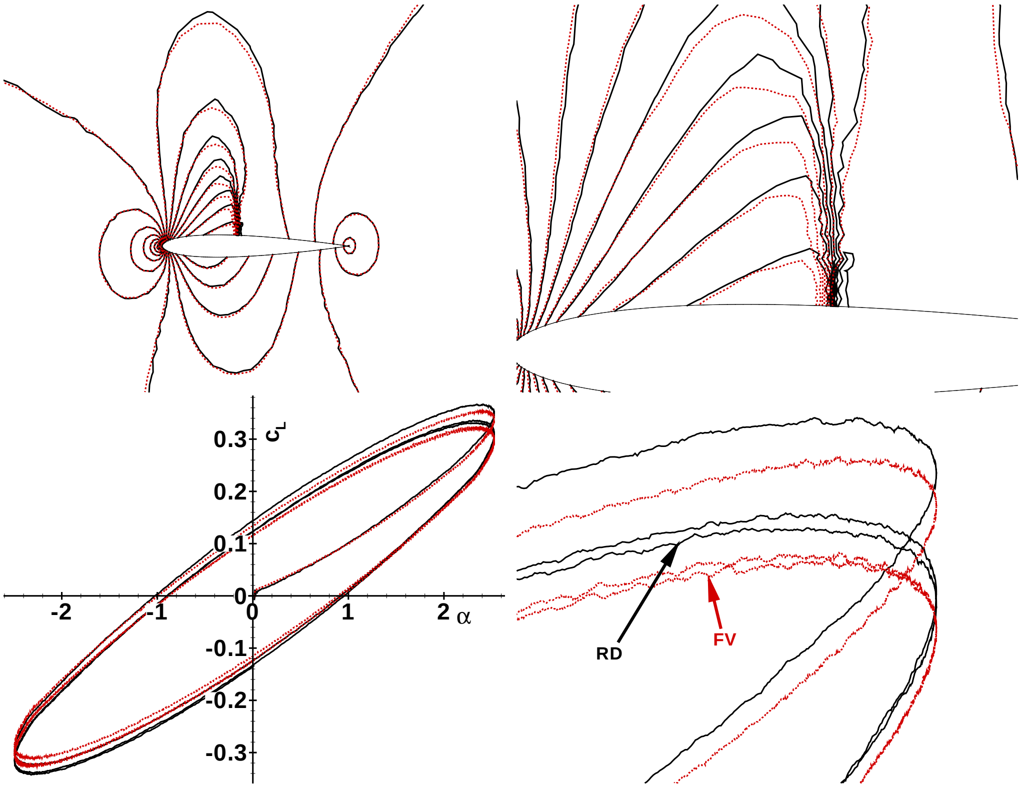


Fig. 3: *Flow past oscillating NACA 0012 airfoil. Top: isolines of the pressure. Bottom: dependence of the lift on the angle. RD method – continuous line, FV method dotted line. CFL* = 5.

5. Conclusions

The two layer N-modified space-time multidimensional upwind residual distribution scheme of [1] was extended for computations on moving meshes. The scheme is unconditionally positive and second order accurate on moving meshes. The method was tested on a 1D piston problem (solved in 2D settings), where we have shown excellent agreement with the analytical solution. The method was then applied to the problem of a transonic flow around an oscillating NACA 0012 airfoil, showing the more accurate and less dissipative behavior of RD scheme with respect to the state of the art FV scheme.

References

[1] R. Abgrall and M. Mezine: *Construction of second order accurate monotone and stable residual distribution schemes for unsteady flow problems.* Journal of Computational Physics **188**, 2003, 16–55.

[2] A. Csík, H. Deconinck, and S. Poedts: *Monotone residual distribution schemes for the ideal 2D magnetohydrodynamic equations on unstructured grids.* AIAA Journal **39**, 8, August 2001, 1532–1541.

[3] Á. Csík, M. Ricchiuto, and H. Deconinck: *Space-time residual distribution schemes for hyperbolic conservation laws over linear and bilinear elements.* 33rd Computational Fluid dynamics Course, Von Karman Institute for Fluid Dynamics, 2003.

[4] H. Deconinck, P. L. Roe, and R. Struijs: *A multidimensional generalization of Roe's flux difference splitter for the Euler equations.* Computers and Fluids **22**, 1993, 215–222.

[5] H. Deconinck, K. Sermeus, and R. Abgrall: *Status of multidimensional upwind residual distribution schemes and applications in aeronautics.* AAIA Paper 2000-2328, AIAA, 2000.

[6] J. Dobeš and H. Deconinck: *A second order space-time residual distribution method for solving compressible flow on moving meshes.* AIAA Paper 2005-0493, AIAA, 2005. Presented on 43rd AIAA Aerospace Sciences Meeting and Exhibit 10–13 January 2005, Reno, Nevada.

[7] B. Koobus and C. Farhat: *Second-order time-accurate and geometrically conservative implicit schemes for flow computations on unstructured dynamic meshes.* Comput. Methods Appl. Mech. Engrg. **170**, 1–2, 1999, 103–129.

[8] R. H. Landon: *NACA 0012. oscillatory and transient pitching, compendium of unsteady aerodynamic measurements.* Technical Report AGARD-R-702, AGARD, 1982.

[9] J. Maerz: *Improving time accuracy for residual distribution schemes.* Project Report 1996-17, Von Karman Institute for Fluid Dynamics, Belgium, Chausée do Waterloo 72, B-1640 Rhode Saint Genèse, Belgium, June 1996.

[10] M. Mezine and R. Abgrall: *Upwind multidimensional residual schemes for steady and unsteady flows.* In: ICCFD2, International Conference on Computational Fluid Dynamics 2, Sydney, Australia, 15–19 July 2002, 165–170.

[11] C. Michler, H. D. Sterck, and H. Deconinck: *An arbitrary Lagrangian Eulerian formulation for residual distribution schemes on moving grids.* Computers and Fluids **32**, 1, 2003, 59–71.

[12] M. J. Zucrow and J. D. Hoffman: *Gas dynamics.* John Wiley and Sons, Inc., 1976.

SCALABLE ALGORITHMS FOR CONTACT PROBLEMS WITH GEOMETRICAL AND MATERIAL NON-LINEARITIES[*]

Jiří Dobiáš, Svatopluk Pták, Zdeněk Dostál, Vít Vondrák

1. Introduction

Contact modelling is still a challenging problem of non-linear computational mechanics. The complexity of such problems is related to the a priori unknown contact interface and contact tractions. Their evaluations have to be part of the solution. In addition, the solution across the contact interface is non-smooth.

FETI (Finite Element Tearing and Interconnecting) method [1] belongs to the class of non-overlapping spatial domain decomposition method. Its key concept stems from the idea that the spatial sub-domains, into which the domain is partitioned, are 'glued' by Lagrange multipliers. After eliminating the primal variables, which are displacements, the original problem is reduced to a small, relatively well conditioned, typically equality constrained quadratic programming problem that is solved iteratively. The CPU time that is necessary for both the elimination and iterations can be reduced nearly proportionally to the number of the processors, so that the algorithm exhibits parallel scalability. Observing that the equality constraints may be used to define so called 'natural coarse grid', Farhat, Mandel and Roux modified the basic FETI algorithm so that they were able to prove its numerical scalability, i.e. asymptotically linear complexity.

If the FETI method is applied to the contact problems, the same methodology can be used to prescribe conditions of non-penetration between bodies.

After brief theoretical introduction, this paper is concerned with demonstration of scalability of a new variant of the FETI domain decomposition method, called TFETI (Total FETI) method, and application of the classic FETI method to the solution to contact problems with other non-linearities.

2. Theoretical background

Let us consider a contact problem between two solid deformable bodies. This is basically the boundary value problem known from the solid mechanics. The problem is depicted in Figure 1. Two bodies are denoted by $(\Omega_1, \Omega_2) \subset \mathbf{R}^n, n = 2$ or $n = 3$ where n stands for number of spatial dimensions. Γ stands for boundaries of the bodies that are sub-divided into three disjoint parts. The Dirichlet and Neumann boundary conditions are prescribed on the parts Γ^u and Γ^f, respectively. The third

[*]This work was supported by grant No. 101/05/0423 of the Grant Agency of the Czech Republic.

type of the boundary condition, Γ^c, is defined along the contact interface. The mathematical description of the problem is given by the governing equations expressing equilibrium conditions of the system, along with the boundary conditions.

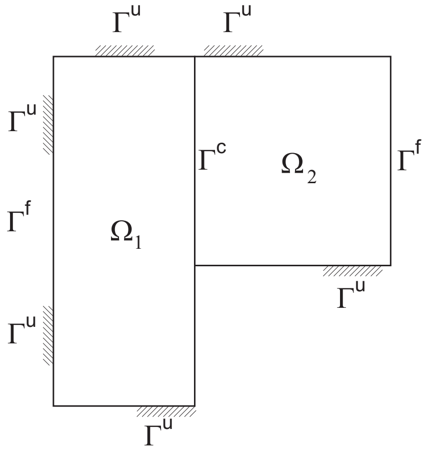


Fig. 1: *Contact problem.*

The result of application of the classic FETI method to the system of bodies from Figure 1 is depicted in Figure 2. The sub-domain Ω_1 is decomposed into two sub-domains with fictitious interface between them.

The fundamental idea of the FETI method is that the compatibility between sub-domains is ensured by means of the Lagrange multipliers or forces in this context. In Figure 2, λ^E denotes the forces along the fictitious interface and λ^I stands for the forces generated by contact.

The original FETI method assumes that Dirichlet boundary conditions are inherited from the original problem, which is shown in Figure 2. This fact implies that the defect of the stiffness matrices of individual sub-domains may vary from zero, for the sub-domains with enough Dirichlet conditions, to the maximum (6 for 3D solid mechanics problems and 3 for 2D ones) in the case of sub-domains exhibiting some rigid body modes. General solution to such systems requires computation of a generalised inverse and a basis of the null spaces of the underlying singular matrices. The problem is that the magnitudes of the defects are difficult to obtain because this computation is disposed to the round off errors [2].

To circumvent the problem, Dostál came up with a novel solution [3]. His idea was to release all prescribed Dirichlet boundary conditions and enforce them by the Lagrange multipliers as it is shown in Figure 3. The effect of the procedure on the stiffness matrices of the sub-domains is that their defects are the same and its magnitude is known beforehand.

The mathematical description of the FETI method can be found, e.g., in [4] and the TFETI method in [3].

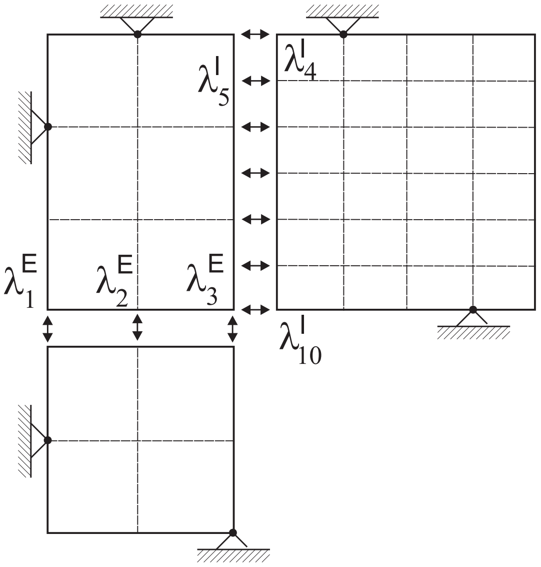


Fig. 2: *FETI method.*

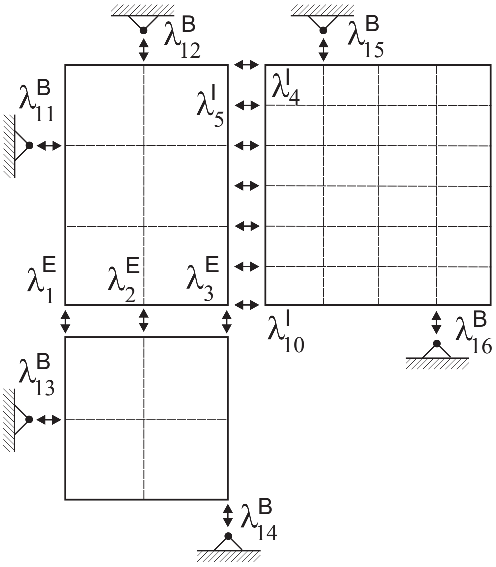


Fig. 3: *TFETI method.*

Application of the FETI and TFETI methods to the contact problems converts the original problem to the quadratic programming one with simple bounds and equality constraints. This problem is further transformed by Semi-Monotonic Augmented Lagrangians with Bound and Equality constraints (SMALBE) method to the sequence of simply bounded quadratic programming problems. These auxiliary problems may be solved efficiently by the Modified Proportioning with Reduced Gradient Projection (MPRGP) method which is described in more details in [5]. It was proved in [6] that application of combination of both these methods to solution to contact problems benefit the numerical and parallel scalability.

We extended the FETI and TFETI method to problems with the geometric and material non-linearities. The above mentioned approach is directly applicable to solution to the contact problems, but with other conditions linear, i.e. for linear elasticity with small displacements and rotations, and frictionless contact. Any additional non-linearity necessitates employment of the nested iteration strategy, where the outer loop is concerned with the material and geometric non-linear effects, contact geometry update, and equilibrium iterations.

3. Numerical experiments

We shall show results of three sets of numerical experiments we carried out. The first one documents numerical scalability of the FETI and TFETI methods. The second case is concerned with contact problem of two cylinders, and the third one with contact problem of the pin in hole with small clearance.

Numerical experiments in the second and third cases were carried out with our general purpose finite element package PMD [7].

3.1. Poisson's problem

Consider a Poisson's problem $\triangle u = 1 \quad \text{in} \quad \Omega$, where $\Omega = (0,1) \times (0,1)$. Dirichlet boundary conditions are prescribed along one edge of the domain, and Neumann conditions along remaining edges. This scalar boundary value problem can be interpreted as the deformation perpendicular to the domain for a thin membrane under lateral pressure, while the physical meaning of the right hand side is the applied pressure divided by membrane tension per unit length. We used bilinear quadrilateral elements for discretisation of the problem.

We carried out a series of computations with changing decomposition parameter H and discretisation parameter h. The results are summarised in Table 1.

H	h	prim.	dual FETI	dual TFETI	CG steps FETI	CG steps TFETI
1/2	1/4	36	11	17	7	4
1/4	1/8	144	63	75	12	5
1/8	1/16	576	287	311	13	7
1/16	1/32	2304	1215	1263	15	11
1/2	1/8	100	19	29	9	9
1/4	1/16	400	111	131	16	12
1/8	1/32	1600	511	551	18	16
1/16	1/64	6400	2175	2255	20	21
1/2	1/16	324	35	53	14	9
1/4	1/32	1296	207	243	22	14
1/8	1/64	5184	959	1031	24	20
1/16	1/128	20736	4095	4239	23	23

Tab. 1: *Scalability of FETI and TFETI.*

The table also shows numbers of primal variables and numbers of dual variables for both FETI and TFETI. We observe from the last two columns that performances of FETI and TFETI are close and that both algorithms exhibit the numerical scalability as can be seen from number of the conjugate gradient (CG) steps.

Figure 4 shows the case corresponding to the first line in Table 1, i.e. $H = 1/2$ and $h = 1/4$. There are four sub-domains there, each with nine primal variables so that the total number is 36. The FETI dual variables are explicitly depicted. The number of the TFETI dual variables is obtained as the sum of the FETI dual variables and the Dirichlet boundary conditions, which are indicated by triangles.

3.2. Contact problem of two cylinders

Consider contact of two cylinders with parallel axes. The diameter of the upper cylinder $R_u = 1$ m and of the lower one $R_l = \infty$. In spite of the fact that it is a 2D problem, it is modelled with 3D continuum trilinear elements with two layers

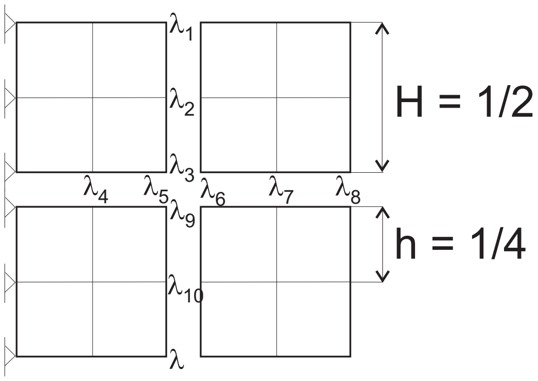


Fig. 4: *Decomposition and discretisation of the domain.*

of them along the axis of symmetry of the upper cylinder. Nevertheless, it is clear that number of layers is irrelevant. The boundary conditions are imposed in such a way that from the physical viewpoint it is the plane strain problem. The model consists of 8904 elements and 12765 nodes. The upper cylinder is loaded by 40 MN/m along the upper line of the upper cylinder.

Figure 5 shows solution to linearly elastic and linearly geometric problem in terms of the deformed mesh. The material properties are as follows: Young's modulus $E = 2.0 \times 10^{11}$ Pa and Poisson's ratio $\nu = 0.3$.

The second problem was computed on the same mesh with the same loading, but we considered linearly–elastic–perfectly–plastic material with yield stress $\sigma_Y =$ 800 MPa. We also considered the geometric non-linearity, i.e. large displacements and finite rotations. The deformed mesh is depicted in Figure 6.

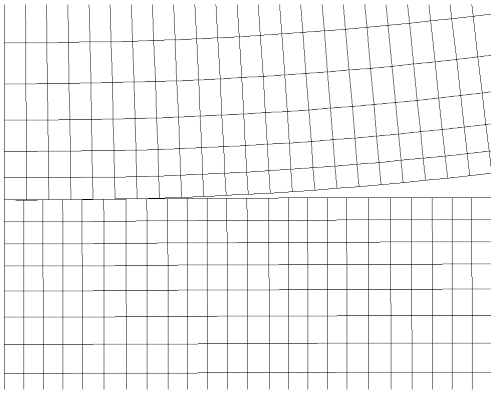


Fig. 5: *Deformed mesh, linear problem.*

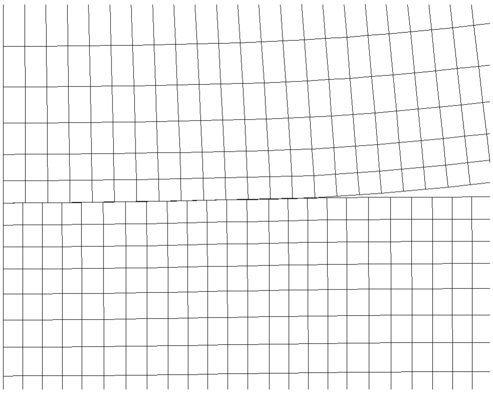


Fig. 6: *Deformed mesh, non-linear problem.*

3.3. Pin-in-hole contact problem

Consider problem of a circular pin in circular hole with small clearance. The radius of the hole is 1 m and the pin has its radius by 1% smaller. Again, the 2D problem is modelled with 3D trilinear elements. The model consists of 15844 elements and 28828 nodes. The pin is loaded along its centre line by 133 MN/m. The geometric non-linearity was considered. The material properties are the same as in the previous case.

Figure 7 shows von Mises stress distribution on the deformed mesh.

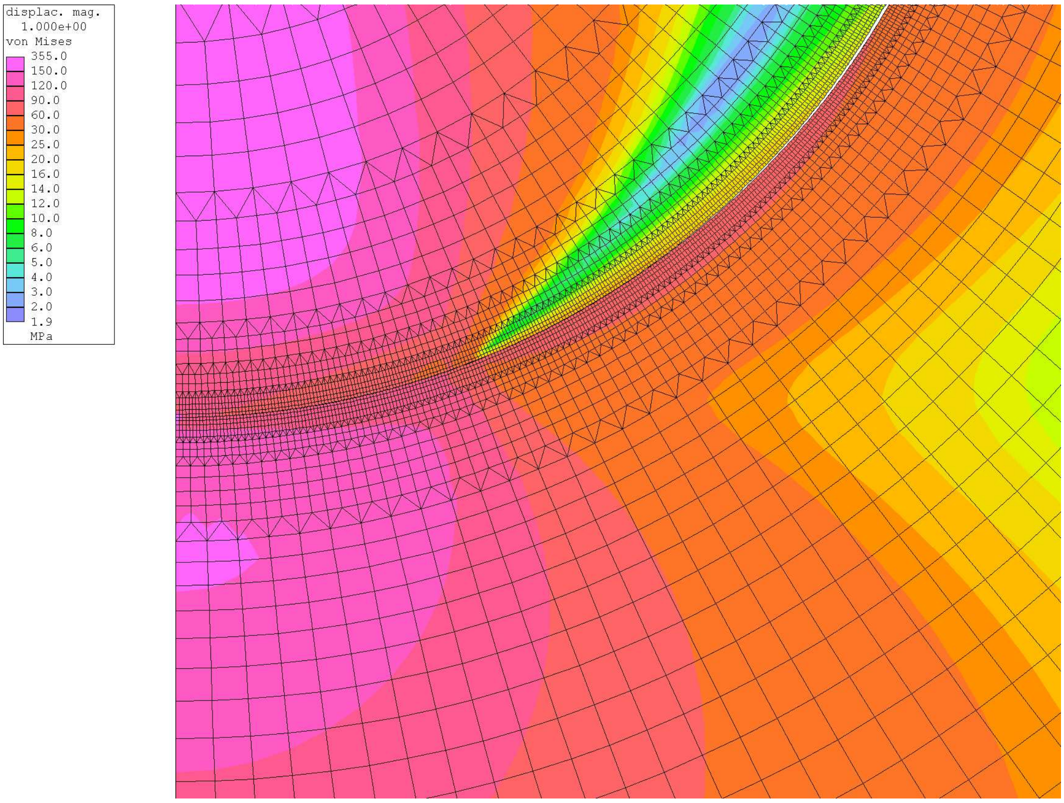


Fig. 7: *Deformed mesh, non-linear problem, von Mises stress.*

4. Conclusion

A new variant of the original FETI domain decomposition method was presented. It is called TFETI and its basic idea, in comparison with FETI, consists in replacement of Dirichlet boundary conditions by Lagrange multipliers or forces in this context. It is of great importance from the computational point of view, because the defect of stiffness matrices of all sub-domains is the same and its magnitude is known beforehand. Numerical experiments show that algorithm stemming from TFETI exhibits the numerical scalability. We also show results of solution to contact problems by the FETI method.

References

[1] Ch. Farhat, F.-X. Roux: *A method of finite element tearing and interconnecting and its parallel solution algorithm.* Int. J. Numer. Methods Engng. **32**, 1991, 1205–1227.

[2] Ch. Farhat, M. Géradin: *On the general solution by a direct method of a large-scale singular system of linear equations: application to the analysis of floating structures.* Int. J. Numer. Methods Engng. **41**, 1998, 675–696.

[3] Z. Dostál, D. Horák, R. Kučera: *Total FETI - an easier implementable variant of the FETI method for numerical solution of elliptic PDE.* Will be published in Communications in Numerical Methods in Engineering.

[4] Z. Dostál, D. Horák, R. Kučera, V. Vondrák, J. Haslinger, J. Dobiáš, S. Pták: *FETI based algorithms for contact problems: scalability, large displacements and 3D Coulomb friction.* Comp. Meth. Appl. Mech. Eng. **194**, 2005, 395–409.

[5] Z. Dostál, J. Schöberl: *Minimizing quadratic functions over non-negative cone with the rate of convergence and finite termination.* Comput. Optim. Appl. **30**, 2005, 23–43.

[6] Z. Dostál: *Inexact semi-monotonic augmented Lagrangians with optimal feasibility convergence for convex bound and equality constrained quadratic programming.* SIAM J. on Num. Anal. **43**, 2005, 96–115.

[7] PMD: *Manuals on* `http://www.it.cas.cz/manual/pmd`.

AN EFFICIENT IMPLEMENTATION OF THE SEMI-IMPLICIT DISCONTINUOUS GALERKIN METHOD FOR COMPRESSIBLE FLOW SIMULATION*

Vít Dolejší

Abstract

We deal with a numerical simulation of the inviscid compressible flow with the aid of the combination of the discontinuous Galerkin method (DGM) and backward difference formulae. We recall the mentioned numerical scheme and discuss implementation aspects of DGM, particularly a choice of basis functions and numerical quadratures for integrations. An illustrative numerical example is presented.

1. Introduction

Our aim is to develop a sufficiently robust, accurate and efficient numerical scheme for a simulation of compressible flows. Among several types of numerical schemes the discontinuous Galerkin method (DGM) seems to be a promising technique, see e.g., [2], [3], [5], [8], [9]. DGM is based on a piecewise polynomial but discontinuous approximation and represents a generalization of the finite element and finite volume methods. Although authors mostly claim that DGM is very suitable for the compressible flow simulation they admit one disadvantage: a high computational cost which prevents DGM from practical applications. Therefore an efficient implementation exhibits a challenging task.

In this paper we recall the semi-implicit numerical scheme proposed in [4],which is based on a combination of DGM for the space semi-discretization and the backward difference formula for the time discretization (Section 3.). Then we discuss some implementation aspects with respect to the CPU time, particularly a choice of the basis functions and numerical quadratures for integrations (Section 4.). Finally, one numerical example of an unsteady inviscid compressible flow through the forward facing step is presented for an illustration.

2. Problem formulation

The system of the *Euler equations* describing 2D inviscid compressible flow can be written in the form

$$\frac{\partial \boldsymbol{w}}{\partial t} + \sum_{s=1}^{2} \frac{\partial \boldsymbol{f}_s(\boldsymbol{w})}{\partial x_s} = 0 \quad \text{in } Q_T = \Omega \times (0,T), \tag{1}$$

*This work is a part of the research project MSM 0021620839 financed by the Ministry of Education of the Czech Republic and was partly supported by the Grant No. 316/2006/B-MAT/MFF of the Grant Agency of Charles University Prague.

where $\Omega \subset I\!R^2$ is a bounded polygonal domain occupied by a gas, $T > 0$ is the length of a time interval, $\boldsymbol{w} = (w_1, \dots, w_4)^{\mathrm{T}} = (\rho,\, \rho v_1,\, \rho v_2,\, e)^{\mathrm{T}}$ is the *state vector* and $\boldsymbol{f}_s(\boldsymbol{w}) = (\rho v_s,\, \rho v_s v_1 + \delta_{s1} p,\, \rho v_s v_2 + \delta_{s2} p,\, (e+p)\, v_s)^{\mathrm{T}},\ s = 1, 2$, are the *inviscid (Euler) fluxes*. We use the following notation: ρ – density, p – pressure, e – total energy, $\boldsymbol{v} = (v_1, v_2)$ – velocity, δ_{sk} – Kronecker symbol, $\gamma > 1$ – Poisson adiabatic constant. The equation of state implies that $p = (\gamma - 1)\,(e - \rho|\boldsymbol{v}|^2/2)$. The system (1) is equipped with a set of initial and boundary conditions, for details see, e.g., [7].

3. Discretization

In [4], we presented the discretization of the Euler equations (1) by the discontinuous Galerkin method (DGM). Therefore we do not derive the numerical scheme again but only present the main relations.

Let $\mathcal{T}_h \equiv \{K_i\}_{i \in I}$ denote a triangulation of the closure $\overline{\Omega}$ of the domain Ω into a finite number of closed elements (triangles or quadrilaterals) $K_i,\ i \in I$ with mutually disjoint interiors. Let $\partial K_i \equiv \cup_{j \in S(i)} \Gamma_{ij}\ \forall K_i \in \mathcal{T}_h$, where $S(i),\ i \in I$ are suitable index sets, Γ_{ij} is either a common face between neighbouring elements K_i and K_j or a boundary face (i.e. $\Gamma_{ij} \subset \partial\Omega$). Moreover, $\boldsymbol{n}_{ij} = ((n_{ij})_1, (n_{ij})_2)$ is the unit outer normal to ∂K_i on the face Γ_{ij}.

The approximate solution of (1) is sought in the space of discontinuous piecewise polynomial functions $\boldsymbol{S}_h$ defined by

$$\boldsymbol{S}_h \equiv [S_h]^4, \quad S_h \equiv S^{p,-1}(\Omega, \mathcal{T}_h) \equiv \{v; v|_K \in P^p(K)\ \forall\, K \in \mathcal{T}_h\}, \tag{2}$$

where $P^p(K)$ denotes the space of all polynomials on K of degree at most $p \geq 0$, p is an integer. For $\boldsymbol{w}_h, \boldsymbol{\varphi}_h \in \boldsymbol{S}_h$ we introduce the forms

$$\begin{aligned}
(\boldsymbol{w}_h, \boldsymbol{\varphi}_h) &= \int_\Omega \boldsymbol{w}_h(\boldsymbol{x}) \cdot \boldsymbol{\varphi}_h(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}, \qquad (3)\\
\boldsymbol{b}_h(\boldsymbol{w}_h, \boldsymbol{\varphi}_h) &= -\sum_{K \in \mathcal{T}_h} \int_K \sum_{s=1}^{2} \boldsymbol{f}_s(\boldsymbol{w}_h) \cdot \frac{\partial \boldsymbol{\varphi}_h}{\partial x_s}\, \mathrm{d}\boldsymbol{x}\\
&\quad + \sum_{K_i \in \mathcal{T}_h} \sum_{j \in S(i)} \int_{\Gamma_{ij}} \boldsymbol{H}(\boldsymbol{w}_h|_{\Gamma_{ij}}, \boldsymbol{w}_h|_{\Gamma_{ji}}, \boldsymbol{n}_{ij}) \cdot \boldsymbol{\varphi}_h \mathrm{d}S,
\end{aligned}$$

where $\boldsymbol{H}$ is a *numerical flux*, $\boldsymbol{w}(t)|_{\Gamma_{ij}}$ and $\boldsymbol{w}(t)|_{\Gamma_{ji}}$ are the values of $\boldsymbol{w}$ on Γ_{ij} considered from the interior and the exterior of K_i, respectively, and at time t. The values of $\boldsymbol{w}(t)|_{\Gamma_{ji}}$ for $\Gamma_{ij} \subset \partial\Omega$ are given by the boundary conditions, for details, see [7]. Then we define the *semidiscrete problem*:

Definition 1: Function $\boldsymbol{w}_h$ is a *semidiscrete solution* of the problem (1), if

$$\begin{aligned}
&\text{a)} \quad \boldsymbol{w}_h \in C^1([0, T]; \boldsymbol{S}_h), \qquad (4)\\
&\text{b)} \quad \left(\frac{\partial \boldsymbol{w}_h(t)}{\partial t}, \boldsymbol{\varphi}_h\right) + \boldsymbol{b}_h(\boldsymbol{w}_h(t), \boldsymbol{\varphi}_h) = 0 \quad \forall\, \boldsymbol{\varphi}_h \in \boldsymbol{S}_h\ \forall\, t \in (0, T),\\
&\text{c)} \quad \boldsymbol{w}_h(0) = \boldsymbol{w}_h^0,
\end{aligned}$$

	approximation		
#dof	P_1	P_2	P_3
FEM	n	$2.5n$	$6n$
DGM	$6n$	$12n$	$20n$

Tab. 1: *Comparison of degree of freedom of DGM and FEM for a triangular grid having n vertices.*

where $\boldsymbol{w}_h^0 \in \boldsymbol{S}_h$ denotes the initial condition. Here $C^1([0,T]; \boldsymbol{S}_h)$ is the space of continuously differentiable mappings of the interval $[0,T]$ into $\boldsymbol{S}_h$.

The problem (4), a) – c) exhibits a system of ordinary differential equations for $\boldsymbol{w}_h(t)$ which has to be discretized by a suitable ODE method. In [4] we introduced the semi-implicit discretization of (4), a) – c), where the form $\boldsymbol{b}_h(\cdot,\cdot)$ was linearized and then the linear terms were treated implicitly by a multi-step backward difference formula and the nonlinear terms were approximated by a suitable higher order explicit extrapolation. Then the full space-time discretization leads to a system of linear algebraic equations at each time level, the numerical scheme is practically unconditionally stable and has a high order of accuracy with respect to the time coordinate.

Since for the purposes of this paper an exact form of the time discretization is not important we write the full space-time discretization schematically by

$$\left(\boldsymbol{M} + \tau_k \boldsymbol{C}(\boldsymbol{w}_h^k)\right) \boldsymbol{w}_h^{k+1} = \boldsymbol{g}(\boldsymbol{w}_h^k), \quad k = 0, 1, \ldots,$$

where $\boldsymbol{w}_h^k \in \boldsymbol{S}_h,\ k = 0, 1, \ldots$ represents an approximation of the solution at $t = t_k$, $\boldsymbol{M}$ is the mass matrix (6), $\boldsymbol{C}(\cdot)$ is a matrix representing the form $\boldsymbol{b}_h(\cdot,\cdot)$, $\boldsymbol{g}(\cdot)$ is a right-hand-side and $\tau_k \equiv t_{k+1} - t_k$ is a time step. For more details see [4].

4. Implementation aspects

Although DGM exhibits a very promising approach for a simulation of compressible flows, its main disadvantage is a higher number of degrees of freedom in comparison with the classical finite element method (FEM) which leads to a higher requirement on CPU time. Table 1 compares the degrees of freedom of DGM and FEM on a triangular grid with n vertices (than the number of triangles $\approx 2n$) for piecewise linear, quadratic and cubic approximations. We observe several times higher number of degrees of freedom for DGM than FEM. Therefore a very efficient implementation is a natural requirement for an industrial use of DGM. We discuss two items: choice of basis functions and numerical quadratures. Other aspects (e.g. linear solver, preconditioning,...) are a subject of the future research.

4.1. Choice of basis

For a numerical simulation of compressible flows it is suitable to use meshes consisting of triangles and quadrilaterals since numerical experiments show that quadrilaterals are better for a resolution of effects within boundary layers around solid walls whereas triangles are more suitable for capturing of discontinuities (e.g., shock waves) with a general direction. An use of the Lagrangian basis known from FEM is not suitable for a combination of triangles and quadrilaterals. Since we have a discontinuous approximation we can employ a local basis on each element independently. A natural choice is an use of the Taylor basis on element $K_i \in \mathcal{T}_h$ in the form

$$\{\psi_j^{K_i}\}_{j=1}^{dof_{K_i}} \equiv \{(x_1 - x_1^{K_i})^{n_x}(x_2 - x_2^{K_i})^{n_y},\ n_x, n_y \geq 0,\ n_x + n_y \leq p\}, \tag{5}$$

where p is the degree of the polynomial approximation on K_i, $dof_{K_i} = (p+1)(p+2)/2$ is the number of degree of freedom on K_i and $(x_1^{K_i}, x_2^{K_i})$ is the barycentre of K_i.

However, numerical experiments show that the Taylor basis (5) is not suitable for a computations, since the *mass matrix* defined by

$$\boldsymbol{M} \equiv \left\{m_{(K_i,n_i),(K_j,n_j)}\right\}_{n_i=1,\dots,dof_{K_i},K_i\in\mathcal{T}_h}^{n_j=1,\dots,dof_{K_j},K_j\in\mathcal{T}_h}, \quad m_{(K_i,n_i),(K_j,n_j)} \equiv \int_\Omega \psi_{n_i}^{K_i}\,\psi_{n_j}^{K_j}\,\mathrm{d}x \tag{6}$$

has elements with very different magnitudes which causes a slow convergence of the linear algebraic problem. In order to save some CPU-time it is possible to use an approach [1] where basis (5) is replaced by the following one

$$\{\tilde\psi_j^{K_i}\}_{j=1}^{dof_{K_i}},\ \tilde\psi_j^{K_i} \equiv \frac{\psi_j^{K_i}}{\|\psi_j^{K_i}\|_{L^2(\Omega)}}, \quad j = 1,\dots,dof_{K_i},\ K_i \in \mathcal{T}_h. \tag{7}$$

Based on numerical experiments we observed that the choice of the basis (7) saves the computational time approximately 50% in comparison with the basis (5).

We extended the idea from [1] in such a way that not only "normalization" but the full orthonormalization of the basis (5) is carried out. So that we employ the basis

$$\{\bar\psi_j^{K_i}\}_{j=1}^{dof_{K_i}}, \quad \text{such that}\ \ (\bar\psi_j^{K_i}, \bar\psi_l^{K_i})_{L^2(\Omega)} = \delta_{jl}, \quad j,l = 1,\dots,dof_{K_i},\ K_i \in \mathcal{T}_h. \tag{8}$$

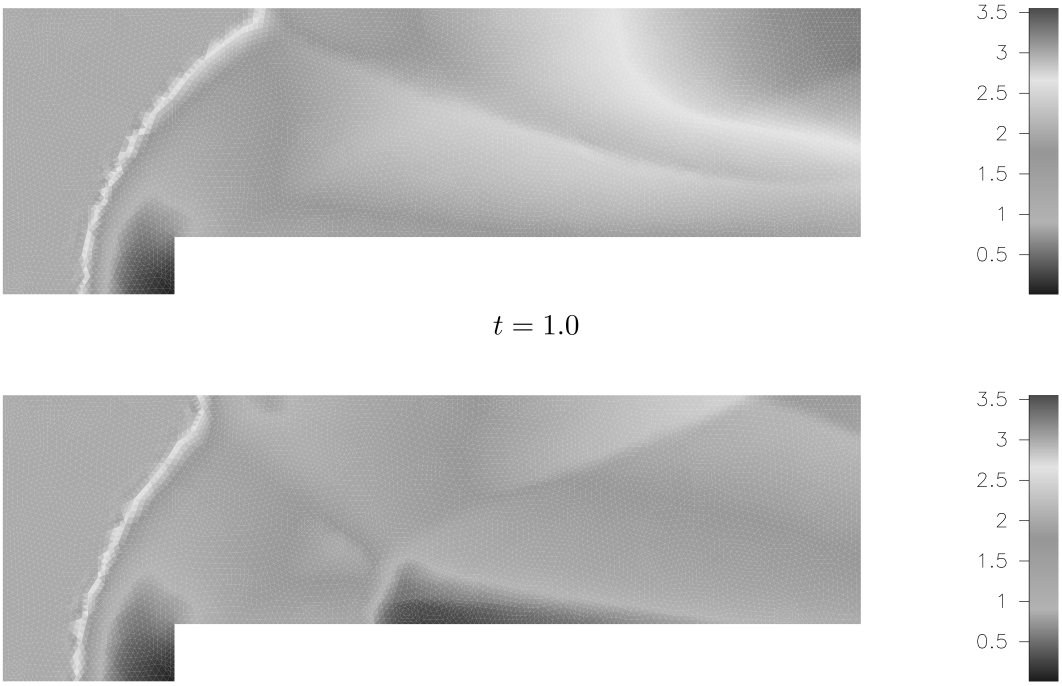
The orthonormalization is carried out by the Grant-Schmidt ortogonalization process. Although it is a known fact, that this algorithm is ill-conditioned we do not observed any problem with the stability of the Grant-Schmidt ortogonalization. It is caused by the fact that the dimension of the finite element space on each element (dof_{K_i}) is small and moreover if the basis is not (exactly) orthogonal it does not mind. We observe that the choice of the basis (8) saves the computational time approximately 90% in comparison with the basis (5).

4.2. Numerical integration

The integrals in (3) have to be evaluated with the aid of suitable numerical quadratures. An use of a numerical quadrature with a low order of accuracy can cause a loss of accuracy and on the other hand a numerical quadrature with a higher number of integration nodes requires longer CPU time. Therefore an use of some

type of integral	integ. rule	#nodes	order
edge	Gauss	$2p$	$4p-1$
quadrilateral	2D Gauss	$(2p)^2$	$4p-1$
triangle	Dunavant		$3p-1$

Tab. 2: *List of the used quadrature rules with the orders of accuracy and the number of integration nodes, p is the degree of polynomial approximation.*



$t = 1.0$

$t = 3.0$

Fig. 1: *Forward facing step, P_3 approximation, Mach number distributions.*

"optimal" numerical quadratures is necessary in order to balance the CPU-costs and the accuracy. Based on numerical experiments the classical Gauss quadrature formulas were employed for edge integrals. Concerning the volume integrals we used the 2D version of the Gauss formulas for quadrilateral elements and the Dunavant rules [6] for triangular elements. Table 2 shows the used quadrature rules with the orders of accuracy and the number of integration nodes.

In order to obtain a high efficiency of the implementation, the values of the test functions in integration nodes are evaluated a priori, so that we do not use any mappings of reference elements to physical ones. Therefore the evaluation of integrals in (3) exhibits a simple multiplicative multiplications of real arrays. This was the reason why we use the programming language Fortran 95 which is optimalized for arrays operations.

5. Numerical example

We consider a flow through the well-known forward facing step proposed in [10] with a constant initial condition given by $\rho = 1.4$, $\boldsymbol{v} = (3, 0)$, $p = 1$. Figure 1 shows Mach number distributions obtained by P_3 approximation on a grid having 1 033 triangles at $t = 0.1$ and $t = 0.3$.

References

[1] F. Bassi: Private communication, Charles University Prague, 2005.

[2] F. Bassi, S. Rebay: *High-order accurate discontinuous finite element solution of the 2D Euler equations.* J. Comput. Phys. **138**, 1997, 251–285.

[3] V. Dolejší: *On the discontinuous Galerkin method for the numerical solution of the Navier–Stokes equations.* Int. J. Numer. Methods Fluids **45**, 2004, 1083–1106.

[4] V. Dolejší: *Higher order semi-implicit discontinuous Galerkin finite element schemes for compressible flow simulation.* In: Software and Algorithms of Numerical Mathematics, 2005, (submitted).

[5] V. Dolejší, M. Feistauer: *Semi-implicit discontinuous Galerkin finite element method for the numerical solution of inviscid compressible flow.* J. Comput. Phys. **198**, 727–746.

[6] D.A. Dunavant: *High degree efficient symmetrical gaussian quadrature rules for the triangle.* Int. J. Numer. Methods Eng. **21**, 1985, 1129–1148.

[7] M. Feistauer, J. Felcman, I. Straškraba: *Mathematical and computational methods for compressible flow.* Oxford University Press 2003.

[8] R. Hartmann, P. Houston: *Adaptive discontinuous Galerkin finite element methods for the compressible Euler equations.* J. Comput. Phys. **183**, 2002, 508–532.

[9] J.J.W. van der Vegt, H. van der Ven: *Space-time discontinuous Galerkin finite element method with dynamic grid motion for inviscid compressible flows. I: General formulation.* J. Comput. Phys. **182**, 2002, 546–585.

[10] P. Woodward, P. Colella: *The numerical simulation of two-dimensional fluid flow with strong shocks.* J. of Comput. Phys. **54**, 1984, 115–173.

NUMERICAL SIMULATION OF INTERACTION OF FLUIDS AND SOLID BODIES[*]

Lenka Dubcová, Miloslav Feistauer, Petr Sváček

1. Introduction

In this work we focus on the numerical simulation of an aeroelastic problem. We consider two–dimensional viscous incompressible flow around an airfoil with two degrees of freedom. It means that the airfoil can oscillate in the vertical direction and rotate around an elastic axis.

The mathematical model of flow is represented by the Navier–Stokes equations and the continuity equation. The initial condition and mixed boundary conditions are added to this system. The numerical simulation consists of the finite element solution of the Navier–Stokes equations coupled with the system of the ordinary differential equations, which describes the airfoil motion.

Since the computational domain is time dependent and the grid is moving, we use the Arbitrary-Lagrangian-Eulerian (ALE) formulation of the Navier–Stokes equations [7]. High Reynolds numbers (10^5–10^6) require the application of a turbulent model.

2. Formulation of the problem

We assume that $(0,T)$ is a time interval and by Ω_t we denote a computational domain occupied by the fluid at time t. The boundary $\partial\Omega_t$ consists of disjoint parts $\Gamma_D, \Gamma_O, \Gamma_{W_t}$, where Γ_D represents the inlet and inpermeable fixed walls, Γ_O the outlet and Γ_{W_t} is the boundary of the airfoil at time t. The fluid flow is characterised by the velocity $\boldsymbol{u} = \boldsymbol{u}(\boldsymbol{x},t) = (u_1(\boldsymbol{x},t), u_2(\boldsymbol{x},t))$ and the kinematic pressure $p = p(\boldsymbol{x},t)$. By ρ we denote the fluid density. The ALE method is based on the ALE mapping of the reference domain $\Omega_{ref} = \Omega_0$ onto the current domain Ω_t:

$$\boldsymbol{A}_t : \Omega_{ref} \mapsto \Omega_t, \quad \boldsymbol{X} \mapsto \boldsymbol{x}(\boldsymbol{X},t) = \boldsymbol{A}_t(\boldsymbol{X})\,. \tag{1}$$

By $\boldsymbol{w}$ we denote the domain velocity: $\boldsymbol{w} = \frac{\partial}{\partial t}\,\boldsymbol{x}(\boldsymbol{X},t)$. In the domain Ω_t we consider the Navier–Stokes system written in the following ALE form

[*] No. 201/05/0005 of the Grant Agency of the Czech Republic. The research of M. Feistauer was partly supported by the research project MSM 0021620839 financed by the Ministry of Education of the Czech Republic.

$$\frac{D^A}{Dt}\boldsymbol{u} + [(\boldsymbol{u}-\boldsymbol{w})\cdot\nabla]\,\boldsymbol{u} + \nabla p - \nu\Delta\boldsymbol{u} = 0 \quad \text{in} \quad \Omega_t, \tag{2}$$

$$\operatorname{div}\boldsymbol{u} = 0 \quad \text{in} \quad \Omega_t, \tag{3}$$

equipped with the initial condition

$$\boldsymbol{u}(\boldsymbol{x},0) = \boldsymbol{u}_0, \quad \boldsymbol{x}\in\Omega_0, \tag{4}$$

and the boundary conditions

$$\begin{aligned} &\text{a) } \boldsymbol{u}|_{\Gamma_D} = \boldsymbol{u}_D, \qquad \text{b) } \boldsymbol{u}|_{\Gamma_{W_t}} = \tilde{\boldsymbol{u}}_\Gamma = \boldsymbol{w}|_{\Gamma_{W_t}}, \\ &\text{c) } -(p-p_{ref})\,\boldsymbol{n} + \nu\frac{\partial\boldsymbol{u}}{\partial\boldsymbol{n}} = 0 \quad \text{on} \quad \Gamma_O. \end{aligned} \tag{5}$$

The vertical displacement H and rotation α of the airfoil are described by the system [7]

$$\begin{aligned} m\ddot{H} + S_\alpha\ddot{\alpha}\cos\alpha + k_{HH}H + d_{HH}\dot{H} - S_\alpha\dot{\alpha}^2\sin\alpha &= -L(t), \\ S_\alpha\ddot{H}\cos\alpha + I_\alpha\ddot{\alpha} + k_{\alpha\alpha}\alpha + d_{\alpha\alpha}\dot{\alpha} &= M(t), \end{aligned} \tag{6}$$

where m denotes the mass of the airfoil, S_α, I_α are the static moment and the inertia moment around the elastic axis, k_{HH}, $k_{\alpha\alpha}$ denote the bending stiffness and the torsional stiffness, d_{HH}, $d_{\alpha\alpha}$ are the structural dampings. The aerodynamic lift force $L(t)$ and the aerodynamic torsional moment $M(t)$ are define by the relations

$$L = -\int_{\Gamma_{Wt}}\sum_{j=1}^{2}\tau_{2j}n_j dS, \quad M = -\int_{\Gamma_{Wt}}\sum_{i,j=1}^{2}\tau_{ij}n_j r_i^{\text{ort}} dS, \tag{7}$$

$$\tau_{ij} = \rho\left[-p\delta_{ij} + \nu\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right)\right], \quad r_1^{\text{ort}} = -(x_2 - x_{EO2}),\ r_2^{\text{ort}} = x_1 - x_{EO1}.$$

These relations determine the interaction between the moving fluid and the airfoil.

3. Discrete problem

Time discretization. We consider a partition $0 = t_0 < t_1 < \cdots < T$, $t_k = k\tau$. On each time level we approximate the solution $\boldsymbol{u}(t_n) \approx \boldsymbol{u}^n$ and $p(t_n) \approx p^n$ and use the second order two step scheme to approximate the ALE derivative. The unknown functions $\boldsymbol{u}^{n+1} : \Omega_{t_{n+1}} \mapsto I\!R^2$ and p^{n+1}: $\Omega_{t_{n+1}} \mapsto I\!R$ satisfy the system

$$\frac{3\boldsymbol{u}^{n+1} - 4\hat{\boldsymbol{u}}^n + \hat{\boldsymbol{u}}^{n-1}}{2\tau} + \Big((\boldsymbol{u}^{n+1} - \boldsymbol{w}^{n+1})\cdot\nabla\Big)\,\boldsymbol{u}^{n+1} + \nabla p^{n+1} - \nu\Delta\boldsymbol{u}^{n+1} = 0,$$

$$\operatorname{div}\boldsymbol{u}^{n+1} = 0, \tag{8}$$

and the boundary conditions (5). The function $\hat{\boldsymbol{u}}^j$ denotes the velocity at time t_j transformed to the domain $\Omega_{t_{n+1}}$.

Space discretization. System (8) is discretized by the finite element metod, based on the weak formulation of our problem: on each time level we want to find the weak solution $U = (\boldsymbol{u}, p) = (\boldsymbol{u}^{n+1}, p^{n+1}) \in W \times Q$, which satisfies

$$a(U, U, V) = f(V), \quad \text{for all}\, V = (\boldsymbol{v}, q) \in X \times Q, \tag{9}$$

and $\boldsymbol{u}$ fulfills the boundary conditions (5), a)–b). Here

$$\begin{aligned}
& W = (H^1(\Omega))^2, \quad X = \{\boldsymbol{v} \in W; \boldsymbol{v}|_{\Gamma_D \cup \Gamma_{W_t}} = 0\}, \quad Q = L^2(\Omega), \qquad (10)\\
& a(U^*, U, V) = \frac{3}{2\tau}(\boldsymbol{u}, \boldsymbol{v})_\Omega + \nu\,(\nabla \boldsymbol{u}, \nabla \boldsymbol{v})_\Omega + (((\boldsymbol{u}^* - \boldsymbol{w}^{n+1}) \cdot \nabla)\boldsymbol{u}, \boldsymbol{v})_\Omega \\
& \quad - (p, \nabla \cdot \boldsymbol{v})_\Omega + (\nabla \cdot \boldsymbol{u}, q)_\Omega, \\
& f(V) = \frac{1}{2\tau}(4\hat{\boldsymbol{u}}^n - \hat{\boldsymbol{u}}^{n-1}, \boldsymbol{v})_\Omega - \int_{\Gamma_O} p_{\text{ref}} \boldsymbol{v} \cdot \boldsymbol{n}\, dS, \\
& U = (\boldsymbol{u}, p), \ V = (\boldsymbol{v}, q), \ U^* = (\boldsymbol{u}^*, p).
\end{aligned}$$

(The symbol $(\cdot,\cdot)$ denotes the $L^2(\Omega)$-scalar product.) In order to apply the finite element method, we approximate the spaces W, X, Q by finite dimensional subspaces W_h, X_h, Q_h, which are defined on a triangulation $\mathcal{T}_h$, and we want to find the approximate solution $U_h = (\boldsymbol{u}_h, p_h) \in W_h \times Q_h$ such that

$$a(U_h, U_h, V_h) = f(V_h) \quad \forall V_h \in X_h \times Q_h, \tag{11}$$

and $\boldsymbol{u}_h$ satisfies an approximation of conditions (5), a)–b). In our computations we use

$$Q_h = \{q \in Q \cap C(\bar{\Omega});\, q|_K \in P^1(K), \forall K \in \mathcal{T}_h\},$$
$$W_h = \{\boldsymbol{v} \in W \cap (C(\bar{\Omega}))^2;\, \boldsymbol{v}|_K \in (P^2(K))^2, \forall K \in \mathcal{T}_h\}, \qquad X_h = W_h \cap X.$$

The couple (X_h, Q_h) satisfies the Babuška-Brezzi condition. Because the Reynolds numbers are high, we use a suitable stabilization of the FEM. Here we apply the approach proposed by Lube in [4]. (For more details, see [7].) The solution of the nonlinear discrete problem is realized by the Oseen iterations.

4. Modelling of turbulence

The flow with a sufficiently small Reynolds number Re is laminar, but if Re increases, the flow loses its stability and becomes turbulent. We apply the algebraic turbulent model [5], which is based on the Reynolds averaging leading to the Reynolds averaged Navier-Stokes equations

$$\text{div}\, \overline{\boldsymbol{u}} = 0, \tag{12}$$

$$\frac{\partial \overline{u_i}}{\partial t} + (\overline{\boldsymbol{u}} \cdot \nabla)\, \overline{u_i} + \frac{\partial \overline{p}}{\partial x_i} - \nu \Delta \overline{u_i} - \sum_{j=1}^{2} \frac{\partial R_{ji}}{\partial x_j} = 0, \quad i = 1, 2, \tag{13}$$

for averaged quantities $\overline{\boldsymbol{u}}, \overline{p} + p'$. The components $R_{ji} = -\overline{u_i' u_j'}, \quad i,j = 1,2$, of the Reynolds stress tensor are expressed by Boussinesq's hypothesis in the form

$$R_{ij} = \nu_T \Big(\frac{\partial \overline{u_i}}{\partial x_j} + \frac{\partial \overline{u_j}}{\partial x_i}\Big), \tag{14}$$

(see, e.g. [3]). Here ν_T is called the turbulent viscosity. It depends on the coordinates, velocity and other variables. To compute ν_T we use two algebraic models designed by Baldwin-Lomax and Rostand [5].

System (12), (13) and (14) is again rewritten in the ALE form and discretized similarly as in Section 3 with the only difference in the definition of the form $a(U^*, U, V)$. The details are contained in [2].

5. Numerical results

5.1. Flow along a flat plate

In order to validate the proposed technique, we compare our numerical results of the simulation of flow along a flat plate with the theory of turbulent flow [6], using the Baldwin-Lomax model and the Rostand model. Let us define the function Y^+ and u^+

$$Y^+(Y) = \frac{u_\tau Y}{\nu}, \qquad\qquad u^+ = \frac{\mathcal{U}_\infty}{u_\tau},$$

where Y is the distance from the plate, $\mathcal{U}_\infty$ is the far field velocity and u_τ is the wall-shear velocity.

Figure 1 (left) shows the comparison of the numerical results with theory. Figure 1 (right) shows the comparison of theoretical dependence of the friction coefficient C_f on the local Reynolds number $Re_x = U_\infty x_1/\nu$ with our computations. The agreement of the computation with theory is very good.

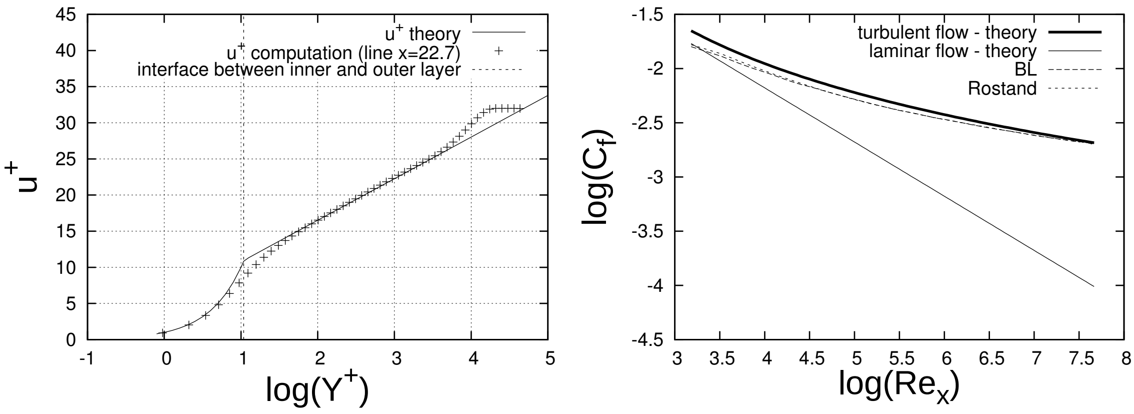


Fig. 1: *The function u^+ in dependence on Y^+ (left); The friction coefficient (right).*

5.2. Flow along the airfoil NACA 0012

Now let us consider flow past the airfoil NACA 0012, which oscillates around the elastic axis (25% of the length of the airfoil) with prescribed frequence $f = 30\,\text{Hz}$ and total amplitude $\alpha^* = 5°$. We compute the pressure coefficient

$$C_p = \frac{P}{\frac{1}{2}\,\rho\,\mathcal{U}_\infty^2},$$

and evaluate $C_{p_{\text{mean}}}$, the time mean value of $C_p(t)$ and the so-called real and imaginary components of the amplitudes C_p' and C_p'' from the relation

$$C_p(t) = C_{p_{\text{mean}}} + C_p' \sin(\omega t) + C_p'' \cos(\omega t).$$

In Figures 2 and 3, there is the comparison of the numerical results with experiments [1]. Although the algebraic model of turbulence is very simple, it gives good results.

Finally, the coupled problem of flow induced airfoil vibrations is solved using the finite element method for the flow problem, combined with the Runge-Kutta method

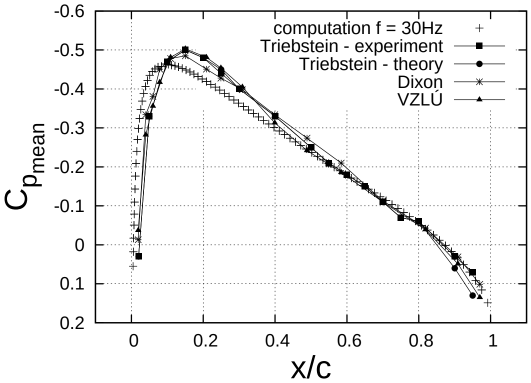


Fig. 2: *The mean value of $C_p(t)$.*

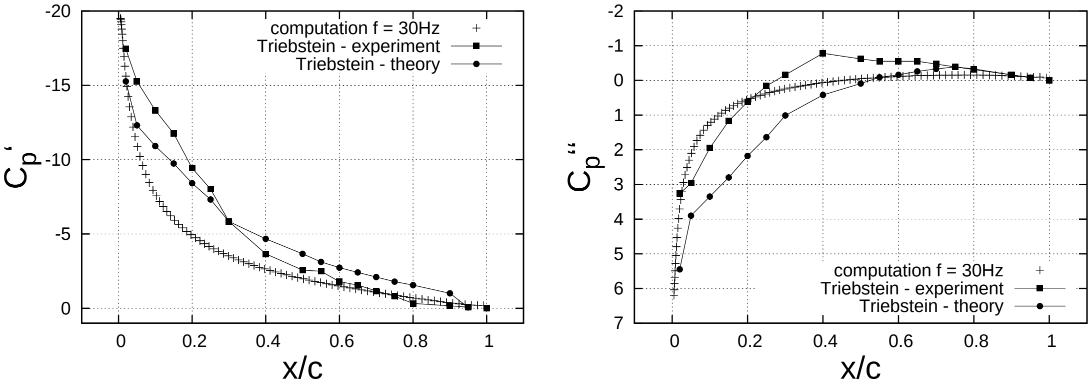


Fig. 3: *The real and imaginary components of the amplitudes.*

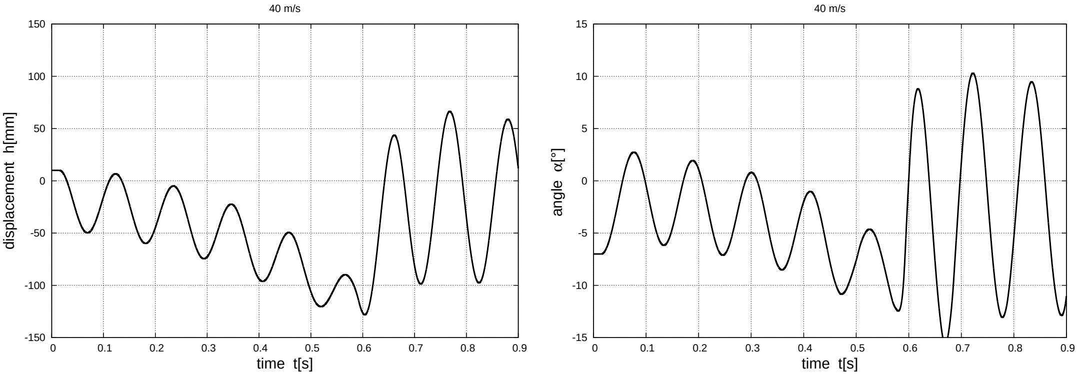


Fig. 4: *Flow induced airfoil vibrations for* $\mathcal{U}_\infty = 40\,m/s$

for system (6) transformed to a first order system. In Figure 4, the displacement H and rotation angle α are plotted in dependence on time for the far field velocity $\mathcal{U}_\infty = 40\,\mathrm{m/s}$. In this case the vibrations are not damped and we get the regime called flutter.

References

[1] J. Benetka: *Measurement of an oscillating airfoil in slotted measurement spaces with various heights.* Tech. Rep. Z-2610/81, Aeronautical Research and Test Institute, Prague, Letňany, 1981 (in Czech).

[2] L. Dubcová: *Numerical simulation of interaction of fluids and solid bodies.* Master Degree Thesis, Charles University Prague, Faculty of Mathematics and Physics, Prague, 2006 (in Czech).

[3] V. John: *Large eddy simulation of turbulent incompressible flows.* Springer, Berlin, 2004.

[4] G. Lube: *Stabilized Galerkin finite element methods for convection dominated and incompressible flow problems.* In: J.K. Kowalski, A. Wakulicz (eds), Numerical Analysis and Mathematical Modelling, Banach Cent. Publ. **29**, (1994), 85–104.

[5] J. Příhoda: *Algebraic models of turbulence and their application to the solution of the averaged Navier-Stokes equations.* Research report Z-1153/90, Institute of Thermomechanics, Czech Academy of Science, Prague, 1990 (in Czech).

[6] H. Schlichting, K. Gersten: *Boundary layer theory.* 8th edition, Springer, Berlin, 2000.

[7] P. Sváček, M. Feistauer, J. Horáček: *Numerical simulation of flow induced airfoil vibrations with large amplitudes.* J. Fluids Structures (to appear).

FINITE VOLUME WLSQR SCHEME AND ITS APPLICATIONS TO TRANSONIC FLOWS[*]

Jiří Fürst

Abstract

This article describes the development of a high order numerical method for the solution of compressible transonic flows. The discretisation in space is based on the standard finite volume method of Godunov's type. A higher order of accuracy is achieved by a piecewise polynomial interpolation similar to the ENO or weighted ENO methods (see e.g. [8]).

1. Introduction

The weighted least square reconstruction (WLSQR) of pointwise data at the cell faces from given cell averages is developed with the aim to simplify the implementation of the standard ENO procedure especially for the case of unstructured meshes. The reconstruction procedure uses single stencil and computes an interpolation polynomial by minimizing the weighted interpolation error over the cells in this stencil.

The complete finite volume scheme equipped with the piecewise linear reconstruction was successfully used for the solution many transonic flow problems (see e.g. [4, 5]). This article presents the basic analytical results as well as some new numerical experiments with the WLSQR scheme especially for the case of inviscid 3D flows and turbulent flows in 2D. The WLSQR reconstruction has been used for the conservative variables as well as for the model of turbulence.

The flow is described by the set of the Euler or the Navier-Stokes equations in conservative form

$$W_t + F(W)_x + G(W)_y = F^v(W)_x + G^v(W)_y + S(W), \tag{1}$$

where $W = [\rho, \rho u, \rho v, e]^T$ is the vector of conservative variables, $F(W)$ and $G(W)$ are the inviscid fluxes, $F^v(W)$ and $G^v(W)$ are the viscous fluxes ($F^v = G^v = 0$ for the case of the Euler equations) and $S(W)$ is a source term, for more details see [2].

The equations equipped with proper boundary conditions are solved numerically using an unstructured mesh and a finite volume scheme with all unknowns located at cell centers. The fluxes through the cell interfaces are approximated by the Gauss quadrature with the physical fluxes replaced by the numerical ones

$$\int_{C_i \cap C_j} (F(W), G(W)) \cdot d\vec{S} \approx \sum_{q=1}^{J} \omega_q F^{AUSMPW+}(W_{ijq}^L, W_{ijq}^R, \vec{S_{ijq}}). \tag{2}$$

[*]This work was supported by grant No. 201/05/0005 of the Grant Agency of the Czech Republic and by the Research Plan MSM No. 6840770010.

Here W^L_{ijq} and W^R_{ijq} denotes the values of the vector of unknowns interpolated to the Gauss point q of the interface $C_i \cap C_j$ from the left cell or from the right cell, respectively. $F^{AUSMPW+}$ denotes the numerical flux described in [9] and ω_q are the weights of the Gauss quadrature. The resulting finite volume scheme for inviscid case can be then written in semi-discrete form

$$|C_i|\frac{dW_i}{dt} = -\sum_{j\in\mathcal{N}_i}\sum_{q=1}^{J}\omega_q F^{AUSMPW+}(W^L_{ijq}, W^R_{ijq}, \vec{S_{ijq}}). \tag{3}$$

Here C_i is the i-th cell, $W_i = \int_{C_i} W(\vec{x},t)d\vec{x}$, and $\mathcal{N}_i = \{j : \dim(C_i \cap C_j) = 1\}$.

The basic first order scheme can be obtained by setting $J = 1$, $W^L_{ijq} = W_i$, and $W^R_{ijq} = W_j$.

2. The WLSQR interpolation

However the basic first order scheme posses very good mathematical properties, it is well known, that it is very diffusive. Therefore a use of higher order schemes is preferred, especially for the viscous flow calculations. The higher order scheme can be constructed within this framework simply by improving the interpolation of W^L and W^R. There exist several methods for the construction of a stable interpolation, the most known are the limited least squares of Barth [1], the ENO/WENO schemes [8], or the TVD schemes [7].

The use of limiters as in the TVD or the Barth's schemes usually cut the order of accuracy near extrema and may also hamper the convergence to a steady state. On the other hand, the implementation of ENO/WENO schemes is relatively complicated for unstructured meshes. Therefore a novel reconstruction procedure was introduced in [5]. Denote by ϕ a component of W. Then the interpolation polynomial $P_i(\vec{x};\phi)$ for the cell C_i is constructed by minimizing the weighted interpolation error [1]

$$\mathrm{err} := \sum_{j\in\mathcal{M}_i}\left[w_{ij}\left(\int_{C_j}\tilde{P}(\vec{x};\phi)\,d\vec{x} - |C_j|\phi_j\right)\right]^2 \tag{4}$$

with respect to the conservativity constraint

$$\int_{C_j} P_i(\vec{x};\phi)\,d\vec{x} = |C_j|\phi_j. \tag{5}$$

The weights w_{ij} are chosen in such a way, that the magnitude of w is big whenever the solution is smooth and w is close to zero when the solution is discontinuous, see formula (6). The single stencil $\mathcal{M}_i$ is selected according to the order of the polynomial P.

[1] Herefrom comes the name of the method - the Weighted Least Square Reconstruction.

2.1. The second order scheme

The formally second order scheme can be obtained by using linear polynomials P_i. For this case, the choice of $\mathcal{M}_i := \mathcal{M}_i^1 = \{j : C_i \cap C_j \neq \emptyset\}$ (i.e. cells touching C_i at least by a vertex) has been tested together with the weights

$$w_{ij} = \sqrt{\frac{h^{-r}}{\left|\frac{\phi_i - \phi_j}{h}\right|^p + h^q}},\ j \in \mathcal{M}_i, \tag{6}$$

with h being the distance between cell centers of C_i and C_j and $p = 4$, $q = -3$, and $r = 3$. The analysis of simplified cases has been carried out in [4] showing a stability of WLSQR interpolation for special discontinuous data.

2.2. The third order scheme

This approach can be extended to a scheme which has formally third order of accuracy by using quadratic polynomials P_i. It is also necessary to enlarge the stencil to $\mathcal{M}_i := \mathcal{M}_i^2 = \mathcal{M}_i^1 \cup \{j : C_j \cap \mathcal{M}_i^1 \neq \emptyset\}$ (i.e. the stencil is extended by the cells touching $\mathcal{M}_i^1$). Although there are no analytical results for quadratic reconstruction, the same definition of w_{ij}, $j \in \mathcal{M}_i^2$ has been used successfully.

2.3. Analysis of weights in WLSQR interpolation

The complete analysis of this three-parametric family of weights is very difficult task, therefore we investigate here only effects of p and q. The value of r was kept constant $r = 3$ in this work.

In [3] the theoretical analysis of 1D piecewise linear reconstruction using regular mesh has been developed with the following results:

Lemma 2.1 *Assume a sufficiently smooth function $u(x)$ having cell averages u_i and weights $w \neq 0$. Then the piecewise linear WLSQR interpolation polynomial approximates $u(x)$ with second order of accuracy, i.e.*

$$P(x; u) = u(x) + \mathcal{O}(h^2). \tag{7}$$

In the case of discontinuous data the total variation of the interpolant for $u(x)$ defined as $u(x) = 1$ for $x < x_{shock}$ and $u(x) = 0$ for $x \geq x_{shock}$ has been analyzed and the following TV-estimate has been proven

$$TV(P(x; u)) \leq TV(u) + 6h^{1+q/p}. \tag{8}$$

Several numerical experiments for piecewise linear WLSQR method in [3] have shown, that the choices $p, q, r = 4, -2, 3$ or $4, -3, 3$ are appropriate at least for inviscid transonic flows in test channel. Therefore we chose here $p, q, r = 4, -2, 3$ also for the piecewise quadratic WLSQR method.

2.4. Numerical experiments with the WLSQR scheme

The numerical analysis of the order of accuracy of an upwind scheme with WLSQR interpolation has been done in [3] for the case of linear advection in 2D and for the non-linear Burgers equation in 2D. The numerical experiments proved, that the order of accuracy corresponds well to the order of the reconstruction for the case of smooth data i.e. the scheme without reconstruction has order of accuracy almost 1, the scheme with piecewise linear reconstruction almost 2, and finally the scheme with quadratic reconstruction almost 3. On the other hand, the order of accuracy drops to one as soon as there are moving discontinuities.

3. Applications in turbomachinery

The above mentioned numerical method has been applied to the solution of transonic flows in 2D turbine cascades. The compressible viscous flow is described by the set of the Euler equations or the Favre averaged Navier-Stokes equations (RANS) coupled with the TNT $k-\omega$ model of turbulence (see [10]). The turbulent transonic flow through a 2D turbine cascade was solved using a hybrid mesh with quadrilaterals around the profile, in the mixing region behind the outlet edge and at the outlet part of boundary. The remaining part of the domain was filled up with triangles. The total number of elements was 24087 with $y_1^+ < 1$ (here y_1^+ is the size of the first cell near the wall in normal direction in wall coordinates, see [11]).

Figure 1 shows the isolines of the Mach number the detail of isolines of entropy

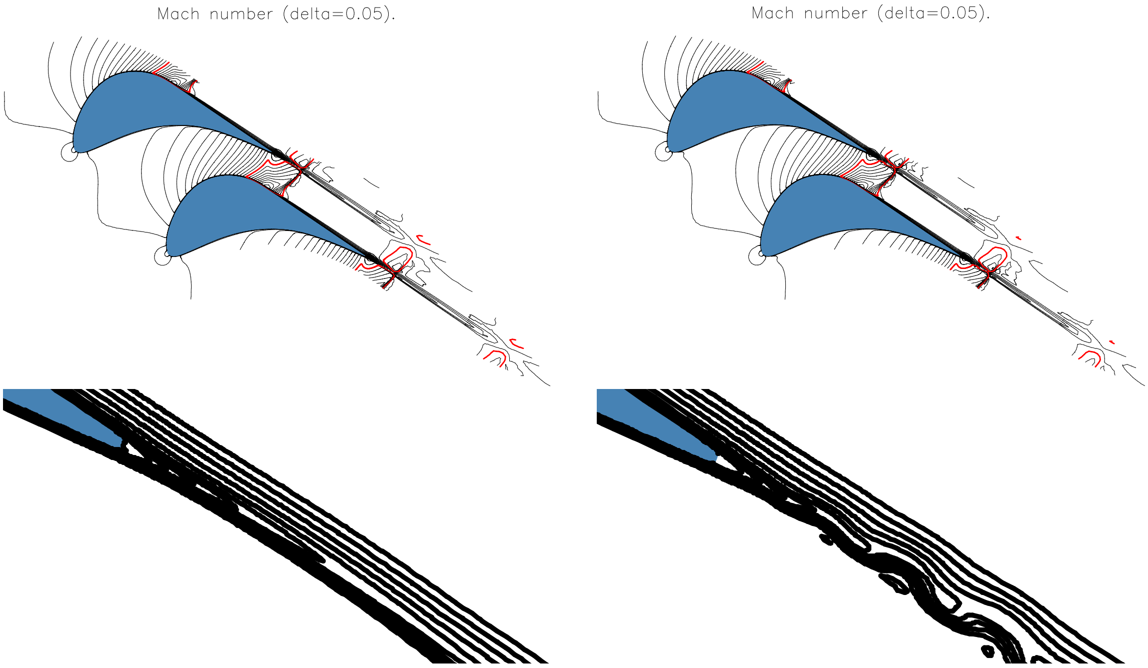


Fig. 1: *Isolines of Mach number (above) and entropy (below) in 2D turbine cascade, second (left) and third (right) order solution.*

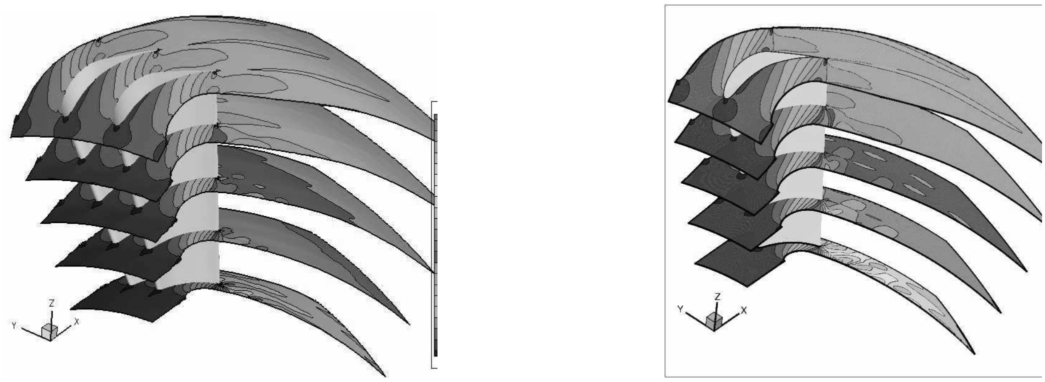


Fig. 2: *Isolines of Mach number for inviscid flow through a 3D turbine stator, WLSQR method on the left (coarser mesh), TVD MC scheme on the right (finer mesh).*

near the outlet edge obtained with the help of the second and the third order method for the flow characterized by the outlet Mach number $M_{2i} = 0.906$ and Reynolds number $Re = 848000$. The isolines of entropy document clearly the difference between those two results - the second order scheme gives stationary solution whereas the wake is unsteady for the third order solution.

Last example is the inviscid transonic flow through 3D turbine cascade. We assume that the flow is periodic from blade to blade and therefore it is possible to solve the flow field just in one period. The domain is discretized using a structured mesh with hexahedral cells. The inflow and outflow conditions depend on the radius. Figure 2 compares the distribution of Mach number obtained with the piecewise linear WLSQR method with AUSM flux using a structured mesh with $100 \times 20 \times 20$ cell. It can be seen, that the solution is comparable to the reference solution obtained with TVD MacCormack scheme with finer mesh having $200 \times 40 \times 40$ cells. Similar results were also obtained by J. Halama [6] using cell vertex Ni's scheme with Jameson's artificial viscosity.

4. Conclusion

The article describes briefly the weighted least-square reconstruction procedure. The proposed WLSQR reconstruction posses good stability even for the case of transonic turbulent flows and is easily extensible to 3D case as well as to third order of accuracy. The difference between second and third order scheme was demonstrated for the case of 2D flows through a turbine. The third order scheme uses less numerical dissipation and produces an unsteady solution in this case.

References

[1] T.J. Barth, D.C. Jesperson: *The design and application of upwind schemes on unstructured meshes.* AIAA Paper 89–0366, AIAA, Jan 1989.

[2] M. Feistauer, J. Felcman, I. Straškraba: *Mathematical and computational methods for compressible flow.* Numerical Mathematics and Scientific Computation. Oxford University Press, 2003.

[3] J. Fürst: *Numerical solution of inviscid and viscous flows using modern schemes and quadrilateral or triangular mesh.* In: M. Beneš, (ed.), Proceedings of the Czech-Japanese Seminar in Applied Mathematics, Czech Technical University in Prague, August 2004.

[4] J. Fürst: *A finite volume scheme with weighted least square reconstruction.* In: S. Raghay F. Benkhaldoun, D. Ouazar, (eds), Finite Volumes for Complex Applications IV, Hermes Science, 2005, 345–354.

[5] J. Fürst, K. Kozel: *Second and third order weighted ENO scheme on unstructured meshes.* In: F. Benkhaldoun and R. Vilsmeier, (eds), Finite Volumes for Complex Applications. Problems and Perspectives, Hermes, July 2002.

[6] J. Halama, T. Arts, J. Fořt: *Numerical solution of steady and unsteady transonic flow in turbine cascades and stages.* Computers and Fluids **33**, 2004, 729–740.

[7] A. Harten: *High resolution schemes for hyperbolic conservation laws.* Journal of Computational Physics **49**, 1983, 357–393.

[8] C. Hu, C.-W. Shu: *Weighted essentially non-oscillatory schemes on triangular meshes.* Journal of Computational Physics **150**, 1999, 97–127.

[9] K.H. Kim, C. Kim, O.-H. Rho: *Methods for the accurate computations of hypersonic flows i. AUSMPW+ scheme.* Journal of Computational Physics **174**, 2001, 38–80.

[10] J.C. Kok: *Resolving the dependence on free stream values for the k-omega turbulence model.* Technical Report NLR-TP-99295, NLR, 1999.

[11] D.C. Wilcox: *Turbulence modeling for CFD.* DCW Industries, Inc., second edition edition, 1998.

TWO-SIDED A POSTERIORI ESTIMATES OF GLOBAL AND LOCAL ERRORS FOR LINEAR ELLIPTIC TYPE BOUNDARY VALUE PROBLEMS*

Antti Hannukainen, Sergey Korotov

Abstract

The paper is devoted to the problem of reliable control of accuracy of approximate solutions obtained in computer simulations. This task is strongly related to the so-called a posteriori error estimates, giving computable bounds for computational errors and detecting zones in the solution domain, where such errors are too large and certain mesh refinements should be performed. Mathematical model described by a linear elliptic (reaction-diffusion) equation with mixed boundary conditions is considered. We derive in a simple way two-sided (upper and lower) easily computable estimates for global (in terms of the energy norm) and local (in terms of linear functionals with local supports) control of the computational error, which is understood as the difference between the exact solution of the model and the approximation. Such two-sided estimates are completely independent of the numerical technique used to obtain approximations and can be made as close to the true errors as resources of a concrete computer used for computations allow.

Keywords: a posteriori error estimation, error control in energy norm, error control in terms of linear functionals, reaction-diffusion equation, mixed boundary conditions.

MSC: 65N15, 65N30

1. Introduction

Many physical and mechanical phenomena can be described by means of mathematical models presenting boundary value problems of elliptic type [7, 15]. Various numerical techniques (the finite difference method, the finite element method (FEM), the finite volume method etc.) are well developed for finding approximate solutions for such problems, see, e.g., [6]. However, in order to be practically meaningful, computer simulations always require an accuracy verification of computed approximations. Such a verification is the main purpose of a posteriori error estimation methods.

In the present paper, we recall two different ways of measuring the computational error, which is understood as the difference $u - \bar{u}$ between the exact solution u and approximation $\bar{u}$, in the global (energy) norm and in terms of linear bounded functionals. These two ways of measurement (and also of control – via a posteriori error

*The first author was supported by the project no. 211512 from the Academy of Finland. The second author was supported by the Academy Research Fellowship no. 208628 from the Academy of Finland.

estimation procedures) of the error are very natural and commonly used nowadays in both mathematical and engineering communities. The global error estimation (see [1, 2, 3, 4, 12, 13, 16, 18, 19, 25, 26] and references therein) normally gives a general presentation on the quality of approximation and a stopping criterion to terminate the calculations. However, practitioners are often interested not only in the value of the overall error, but also in errors over certain critical (and usually local) parts of the solution domain (for example, in fracture mechanics – see [23, 24] and references therein). This reason initiated another trend in a posteriori error estimation which is based on the concept of control of the computational error locally. One common way to perform such a control is to introduce a suitable linear functional ℓ related to subdomain of interest and to construct a posteriori computable estimate for $\ell(u - \bar{u})$, see [4, 5, 8, 11, 14, 20].

It is worth to mention here that most of estimates proposed so far strongly rely on the fact that the computed solutions are true finite element (FE) approximations which, in fact, rarely happens in real computations, e.g., due to quadrature rules, forcibly stopped iterative processes, various round-off errors, or even possible bugs in FE codes.

In this work, on the base of a model elliptic problem with mixed (Dirichlet/Neumann) boundary conditions, we present two relatively simple technologies for obtaining *computable guaranteed two-sided (upper and lower) a posteriori error estimates* needed for reliable control in both global (in the energy norm) and local (in terms of linear functionals) ways. The estimates derived are valid for *any conforming approximations* independently of numerical methods used to obtain them, and can be made *arbitrarily close* to the true errors. In real-life calculations this closeness only depends on resources of a concrete computer used. Some variant of the present paper was published as a preprint [9] in February 2006 (see also [10]).

2. Formulation of problem

For standard definitions of functional spaces and finite element terminology used in the paper we refer to [6].

2.1. Model problem

We introduce the model elliptic problem which consists of the governing equation (1) and mixed (Dirichlet/Neumann) boundary conditions (2)–(3): Find a function u such that

$$-\mathrm{div}(A\nabla u) + cu = f \quad \text{in } \Omega, \tag{1}$$

$$u = u_0 \quad \text{on } \Gamma_D, \tag{2}$$

$$\nu^T \cdot A\nabla u = g \quad \text{on } \Gamma_N, \tag{3}$$

where Ω is a bounded domain in $\mathbf{R}^d$ with a Lipschitz continuous boundary $\partial\Omega$, such that $\overline{\partial\Omega} = \overline{\Gamma}_D \cup \overline{\Gamma}_N$, $\mathrm{meas}_{d-1}\,\Gamma_D > 0$ and ν is the outward normal to the boundary.

It is common practice to pose problem (1)–(3) in the so-called weak form: Find $u \in u_0 + H^1_{\Gamma_D}(\Omega)$ such that

$$\int_\Omega A\nabla u \cdot \nabla w \, dx + \int_\Omega cuw \, dx = \int_\Omega fw \, dx + \int_{\Gamma_N} gw \, ds \qquad \forall w \in H^1_{\Gamma_D}(\Omega), \tag{4}$$

where

$$H^1_{\Gamma_D}(\Omega) := \{v \in H^1(\Omega) \mid v = 0 \text{ on } \Gamma_D\}.$$

For the purposes of the weak formulation, we assume, that $f \in L_2(\Omega)$, $u_0 \in H^1(\Omega)$, $g \in L_2(\Gamma_N)$, $c \in L_\infty(\Omega)$, the coefficient matrix A is symmetric, with entries $a_{ij} \in L_\infty(\Omega)$, $i, j = 1, \ldots, d$, and is such that

$$C_2|\xi|^2 \geq A(x)\xi \cdot \xi \geq C_1|\xi|^2 \qquad \forall \xi \in \mathbf{R}^d \quad \text{for a.e. } x \in \Omega. \tag{5}$$

In addition, the coefficient c is assumed to be either zero or bounded away from zero by a positive constant c_0, i.e. $c \equiv 0$ in $\Omega \setminus \overline{\Omega^c}$, where

$$\Omega^c := \text{supp } c = \{x \in \Omega \mid c(x) \geq c_0 > 0\}. \tag{6}$$

If we define bilinear form $a(\cdot, \cdot)$ and linear form $F(\cdot)$ as follows

$$a(v, w) := \int_\Omega A\nabla v \cdot \nabla w \, dx + \int_\Omega cvw \, dx, \quad v, w \in H^1(\Omega),$$
$$F(w) := \int_\Omega fw \, dx + \int_{\Gamma_N} gw \, ds, \quad w \in H^1(\Omega),$$

then weak formulation (4) can be written in a short form: Find $u = u_0 + u^*$, where $u^* \in H^1_{\Gamma_D}(\Omega)$, such that $a(u, w) = F(w) \quad \forall w \in H^1_{\Gamma_D}(\Omega)$.

Remark 2.1 *The weak solution defined by* (4) *exists and is unique in view of well-known Lax-Milgram lemma (see, e.g.,* [6]*).*

The so-called *energy functional* J of problem (4) is defined as follows

$$J(w) := \frac{1}{2}a(w, w) - \bar{F}(w), \qquad w \in H^1(\Omega), \tag{7}$$

where $\bar{F}(w) := F(w) - a(u_0, w)$, and the corresponding *energy norm* is defined as $\sqrt{a(\cdot, \cdot)}$.

Remark 2.2 *It is well-known that problem* (4) *(namely, finding the function u^*) is equivalent to the problem of finding the minimizer (which is equal to u^*) of the energy functional* (7) *over the space $H^1_{\Gamma_D}(\Omega)$.*

2.2. Types of error control

Let $\bar{u} = u_0 + \bar{u}^*$ be *any function* from $u_0 + H^1_{\Gamma_D}(\Omega)$ (e.g., computed by some numerical method) considered as an approximation of u. It is a natural practice to measure the overall accuracy of the approximation $\bar{u}$ in terms of the above-defined energy norm. Thus, our first goal is to construct reliable and easily computable two-sided estimates for controlling the following value

$$a(u - \bar{u}, u - \bar{u}) = \int_\Omega A\nabla(u - \bar{u}) \cdot \nabla(u - \bar{u})\, dx + \int_\Omega c(u - \bar{u})^2\, dx. \tag{8}$$

The second type of error control considered in the paper is two-sided estimation of the value of the difference $u - \bar{u}$ in terms of a linear bounded functional ℓ

$$\ell(u - \bar{u}). \tag{9}$$

Remark 2.3 *It is clear that existence of an estimate for* (9) *also allows to estimate the value $\ell(u)$ (often called quantity of interest or goal-oriented quantity* [1]*). Really, $\ell(u) = \ell(u - \bar{u}) + \ell(\bar{u})$ where $\ell(\bar{u})$ is computable and $\ell(u - \bar{u})$ is estimated. The value of $\ell(u)$ can be sometimes more important to know than the solution u itself (see* [23, 24]*).*

Remark 2.4 *If the functional ℓ in* (9) *is defined as some integral over small subdomain (or line) in $\overline{\Omega}$, then reliable two-sided estimation of $\ell(u - \bar{u})$ helps to control the behaviour of the error $u - \bar{u}$ locally in that subdomain (or over the line). For example, one can be interested in estimation of $\ell(u - \bar{u}) = \int_S \varphi(u - \bar{u})\, dx$ with S be a subdomain in Ω or a line in Γ_N (where the solution is also unknown), see* [11] *for more details and numerical results in this respect.*

2.3. Inequalities and constants

In what follows we shall need the Friedrichs inequality

$$\|w\|_{0,\Omega} \leq C_{\Omega,\Gamma_D} \|\nabla w\|_{0,\Omega} \quad \forall w \in H^1_{\Gamma_D}(\Omega), \tag{10}$$

and the inequality in the trace theorem

$$\|w\|_{0,\partial\Omega} \leq C_{\partial\Omega} \|w\|_{1,\Omega} \quad \forall w \in H^1(\Omega), \tag{11}$$

where C_{Ω,Γ_D} and $C_{\partial\Omega}$ are positive constants, depending only on Ω, Γ_D, and $\partial\Omega$. The above used denotation $\|\cdot\|_{0,\Omega}$ and $\|\cdot\|_{1,\Omega}$ stand for the standard norms in $L_2(\Omega)$ and $H^1(\Omega)$, respectively. The symbol $\|\cdot\|_{0,\partial\Omega}$ means the norm in $L_2(\partial\Omega)$. Proofs of inequalities (10) and (11) can be found, e.g., in [17].

3. Two-sided estimates of error in energy norm

In this section we shall employ the denotation χ_S for a characteristic function of set S, i.e., $\chi_S(x) = 1$ if $x \in S$, and $\chi_S(x) = 0$ if $x \notin S$. We also define $|||\mathbf{y}|||_\Omega := \sqrt{\int_\Omega A\mathbf{y} \cdot \mathbf{y}\, dx}$ for $\mathbf{y} \in L_2(\Omega, \mathbf{R}^d)$.

3.1. Upper estimate

Proposition 3.1 *For the error in the energy norm* (8) *we have the following upper estimate*

$$a(u-\bar{u},u-\bar{u}) \leq \left\| \frac{1}{\sqrt{c}}(f+\operatorname{div}\mathbf{y}^*-c\bar{u}) \right\|^2_{0,\Omega^c} + \\ + (1+\alpha)\|\!|A^{-1}\mathbf{y}^*-\nabla\bar{u}|\!\|^2_\Omega + \left(1+\frac{1}{\alpha}\right)(1+\beta)\frac{C^2_{\Omega,\Gamma_D}}{C_1}\|f+\operatorname{div}\mathbf{y}^*\|^2_{0,\Omega\setminus\overline{\Omega}^c} \\ + \left(1+\frac{1}{\alpha}\right)\left(1+\frac{1}{\beta}\right)C^2_{\Omega,\partial\Omega}\|g-\nu^T\cdot\mathbf{y}^*\|^2_{0,\Gamma_N}, \quad (12)$$

where α and β are arbitrary positive real numbers, $\mathbf{y}^*$ is any function from

$$H_N(\Omega,\operatorname{div}) := \left\{\mathbf{y}\in L_2(\Omega,\mathbf{R}^d) \mid \operatorname{div}\mathbf{y}\in L_2(\Omega),\ \nu^T\cdot\mathbf{y}\in L_2(\Gamma_N)\right\},$$

and $C_{\Omega,\partial\Omega} := C_{\partial\Omega}\sqrt{1+C^2_{\Omega,\Gamma_D}}/\sqrt{C_1}$.

Proof: First of all, we notice that it actually holds, cf. (6),

$$a(u-\bar{u},u-\bar{u}) = \|\!|\nabla(u-\bar{u})|\!\|^2_\Omega + \|\sqrt{c}(u-\bar{u})\|^2_{0,\Omega^c}. \quad (13)$$

Further, using the fact that $u-\bar{u}\in H^1_{\Gamma_D}(\Omega)$ and identity (4) we observe that

$$a(u-\bar{u},u-\bar{u}) = \int_\Omega f(u-\bar{u})dx + \int_{\Gamma_N} g(u-\bar{u})\,ds - \int_\Omega A\nabla\bar{u}\cdot\nabla\,(u-\bar{u})\,dx \\ - \int_\Omega c\bar{u}(u-\bar{u})\,dx = \int_\Omega (f-c\bar{u})(u-\bar{u})\ dx + \int_{\Gamma_N} g(u-\bar{u})\,ds \\ - \int_\Omega (A\nabla\bar{u}-\mathbf{y}^*)\cdot\nabla(u-\bar{u})\ dx - \int_\Omega \mathbf{y}^*\cdot\nabla(u-\bar{u})\ dx, \quad (14)$$

where $\mathbf{y}^*$ is any function from the space $H_N(\Omega,\operatorname{div})$ defined in the formulation of the theorem. Applying the Green's formula to the last term in above gives

$$\int_\Omega \mathbf{y}^*\cdot\nabla(u-\bar{u})\ dx = \int_{\Gamma_N}(\nu^T\cdot\mathbf{y}^*)(u-\bar{u})\,ds - \int_\Omega \operatorname{div}\mathbf{y}^*(u-\bar{u})\ dx.$$

Using this identity and equation (14) we obtain

$$a(u-\bar{u},u-\bar{u}) = \int_\Omega A(A^{-1}\mathbf{y}^*-\nabla\bar{u})\cdot\nabla(u-\bar{u})\ dx + \int_\Omega (f+\operatorname{div}\mathbf{y}^*-c\bar{u})(u-\bar{u})\ dx \\ + \int_{\Gamma_N}(g-\nu^T\cdot\mathbf{y}^*)(u-\bar{u})\,ds. \quad (15)$$

Now, we proceed by estimating the three terms in the right-hand side (RHS) of equality (15). The first term can be estimated by the Cauchy-Schwarz inequality as

$$\int_\Omega A(A^{-1}\mathbf{y}^* - \nabla\bar{u}) \cdot \nabla(u-\bar{u})\; dx \le |\!|\!|A^{-1}\mathbf{y}^* - \nabla\bar{u}|\!|\!|_\Omega\, |\!|\!|\nabla(u-\bar{u})|\!|\!|_\Omega. \qquad (16)$$

The second term in the RHS of equality (15) can be estimated using Friedrichs inequality (10), ellipticity condition (5), denotation (6), and a simple inequality $a\,b \le \frac{1}{2}a^2 + \frac{1}{2}b^2$ as follows

$$\begin{aligned}
&\int_\Omega (f + \operatorname{div}\mathbf{y}^* - c\bar{u})(u-\bar{u})\; dx \\
= \;& \int_{\Omega^c} \frac{1}{\sqrt{c}}(f + \operatorname{div}\mathbf{y}^* - c\bar{u})\,\sqrt{c}(u-\bar{u})\; dx + \int_\Omega \chi_{\Omega\setminus\overline{\Omega}^c}(f + \operatorname{div}\mathbf{y}^* - c\bar{u})\,(u-\bar{u})\; dx \\
\le \;& \|\sqrt{c}(u-\bar{u})\|_{0,\Omega^c} \Big\| \frac{1}{\sqrt{c}}(f + \operatorname{div}\mathbf{y}^* - c\bar{u})\Big\|_{0,\Omega^c} \\
& \qquad\qquad + \|\chi_{\Omega\setminus\overline{\Omega}^c}(f + \operatorname{div}\mathbf{y}^* - c\bar{u})\|_{0,\Omega}\, \|u-\bar{u}\|_{0,\Omega} \\
\le \;& \frac{1}{2}\|\sqrt{c}(u-\bar{u})\|^2_{0,\Omega^c} + \frac{1}{2}\Big\| \frac{1}{\sqrt{c}}(f + \operatorname{div}\mathbf{y}^* - c\bar{u})\Big\|^2_{0,\Omega^c} \qquad (17) \\
& \qquad\qquad + \frac{C_{\Omega,\Gamma_D}}{\sqrt{C_1}}\|f + \operatorname{div}\mathbf{y}^* - c\bar{u}\|_{0,\Omega\setminus\overline{\Omega}^c} |\!|\!|\nabla(u-\bar{u})|\!|\!|_\Omega.
\end{aligned}$$

Finally, the third term can be estimated using inequalities (10) and (11) and the ellipticity condition (5) as

$$\begin{aligned}
&\int_{\Gamma_N} (g - \nu^T \cdot \mathbf{y}^*)(u-\bar{u})\, ds \le \|g - \nu^T \cdot \mathbf{y}^*\|_{0,\Gamma_N} \|u-\bar{u}\|_{0,\Gamma_N} \\
&\quad \le \; C_{\partial\Omega}\|g - \nu^T \cdot \mathbf{y}^*\|_{0,\Gamma_N}\|u-\bar{u}\|_{1,\Omega} \le C_{\Omega,\partial\Omega}\|g - \nu^T \cdot \mathbf{y}^*\|_{0,\Gamma_N} |\!|\!|\nabla(u-\bar{u})|\!|\!|_\Omega. \quad (18)
\end{aligned}$$

Using (16), (17), and (18) to estimate the terms on the RHS of (15), we obtain

$$\begin{aligned}
& a(u-\bar{u}, u-\bar{u}) \\
& \le \frac{1}{2}\Big(|\!|\!|A^{-1}\mathbf{y}^* - \nabla\bar{u})|\!|\!|_\Omega + C_{\Omega,\partial\Omega}\|g - \nu^T \cdot \mathbf{y}^*\|_{0,\Gamma_N} + \frac{C_{\Omega,\Gamma_D}}{\sqrt{C_1}}\|f + \operatorname{div}\mathbf{y}^* - c\bar{u}\|_{0,\Omega\setminus\overline{\Omega}^c}\Big)^2 \\
& \qquad + \frac{1}{2}|\!|\!|\nabla(u-\bar{u})|\!|\!|^2_\Omega + \frac{1}{2}\|\sqrt{c}(u-\bar{u})\|^2_{0,\Omega^c} + \frac{1}{2}\Big\|\frac{1}{\sqrt{c}}(f + \operatorname{div}\mathbf{y}^* - c\bar{u})\Big\|^2_{0,\Omega^c}. \quad (19)
\end{aligned}$$

Using now (13) and the final inequality (19), multiplying by two and regrouping, we immediately get for the error in the energy norm that

$$\begin{aligned}
& a(u-\bar{u}, u-\bar{u}) = |\!|\!|\nabla(u-\bar{u})|\!|\!|^2_\Omega + \|\sqrt{c}(u-\bar{u})\|^2_{0,\Omega^c} \le \Big\|\frac{1}{\sqrt{c}}(f + \operatorname{div}\mathbf{y}^* - c\bar{u})\Big\|^2_{0,\Omega^c} \\
& + \Big(|\!|\!|A^{-1}\mathbf{y}^* - \nabla\bar{u}|\!|\!|_\Omega + \frac{C_{\Omega,\Gamma_D}}{\sqrt{C_1}}\|f + \operatorname{div}\mathbf{y}^*\|_{0,\Omega\setminus\overline{\Omega}^c} + C_{\Omega,\partial\Omega}\|g - \nu^T \cdot \mathbf{y}^*\|_{0,\Gamma_N}\Big)^2. \quad (20)
\end{aligned}$$

Finally, using two times the inequality $(a+b)^2 \le (1+\lambda)a^2 + (1+\frac{1}{\lambda})b^2$, valid for any $\lambda > 0$, for the terms in the round brackets in (20), we get estimate (12).

□

3.2. Lower estimate

Proposition 3.2 *For the error in the energy norm* (8) *we have the following lower bound*

$$a(u-\bar{u}, u-\bar{u}) \geq 2(J(\bar{u}^*) - J(w)), \tag{21}$$

where w is any function from $H^1_{\Gamma_D}(\Omega)$ and the functional J is defined in (7)*.*

Proof: First, we prove that

$$a(u-\bar{u}, u-\bar{u}) = 2(J(\bar{u}^*) - J(u^*)). \tag{22}$$

Really, we have

$$\begin{aligned}
2(J(\bar{u}^*) - J(u^*)) &= a(\bar{u}^*, \bar{u}^*) - 2\bar{F}(\bar{u}^*) - a(u^*, u^*) + 2\bar{F}(u^*) \\
= a(\bar{u}^*, \bar{u}^*) - a(u^*, u^*) &+ 2\bar{F}(u^* - \bar{u}^*) = a(\bar{u}^*, \bar{u}^*) - a(u^*, u^*) + 2a(u^*, u^* - \bar{u}^*) \\
&= a(\bar{u}^*, \bar{u}^*) + a(u^*, u^*) - 2a(u^*, \bar{u}^*) = a(u-\bar{u}, u-\bar{u}).
\end{aligned}$$

Since u^* minimizes the energy functional, we have $J(u^*) \leq J(w) \quad \forall w \in H^1_{\Gamma_D}(\Omega)$, which proves (21). □

Remark 3.1 *The estimate* (21) *has a practical meaning only if w satisfies $J(w) \leq J(\bar{u}^*)$. For example, if $\bar{u}^*$ comes from a FE-solution obtained using mesh S_h, suitable w can be constructed, e.g., by solving the weak problem* (4) *on a hierarcially refined mesh S_τ.*

3.3. Comments on two-sided estimates (12) and (21)

- In order to derive the upper (12) and the lower (21) estimates, we did not specify the function $\bar{u}$ to be a finite element approximation (or computed by some another numerical method). In fact, it is simply any function from the set $u_0 + H^1_{\Gamma_D}(\Omega)$.

- The upper estimate (12) cannot be improved. Really, if one takes $\mathbf{y}^* = A\nabla u$, which obviously belongs to $H_N(\Omega, \mathrm{div})$, then the last two terms in the right-hand side of (12) vanish. Further, taking $\alpha = 0$, we finally observe that the inequality (12) holds as equality. To prove that the lower estimate (21) cannot be improved either, we should, obviously, take $w = u^* \in H^1_{\Gamma_D}(\Omega)$ and use (22).

- The upper estimate (12) contains only two global constants, C_{Ω,Γ_D} and $C_{\partial\Omega}$, which do not depend on the computational process. They have to be computed (or accurately estimated from above) only once when the problem is posed.

- In many works, devoted to a posteriori error estimation, one usually takes $c \equiv 0$. In this case $a(u-\bar{u}, u-\bar{u}) = \|\nabla(u-\bar{u})\|^2_\Omega$, the set $\Omega^c = \emptyset$, and the estimate (12) takes a simpler form

$$a(u-\bar{u},u-\bar{u}) \leq (1+\alpha)|\!|\!|A^{-1}\mathbf{y}^* - \nabla\bar{u}|\!|\!|^2_\Omega + \Big(1+\frac{1}{\alpha}\Big)(1+\beta)\frac{C^2_{\Omega,\Gamma_D}}{C_1}\|f+\operatorname{div}\mathbf{y}^*\|^2_{0,\Omega}$$
$$+\Big(1+\frac{1}{\alpha}\Big)\Big(1+\frac{1}{\beta}\Big)C^2_{\Omega,\partial\Omega}\|g-\nu^T\cdot\mathbf{y}^*\|^2_{0,\Gamma_N}. \quad (23)$$

- For the pure Dirichlet boundary condition, the third term in RHS of (23) vanishes, and, since the estimate is valid for any positive β, we can take it to be zero. Then, we get the estimate

$$a(u-\bar{u},u-\bar{u}) \leq (1+\alpha)|\!|\!|A^{-1}\mathbf{y}^* - \nabla\bar{u}|\!|\!|^2_\Omega + \Big(1+\frac{1}{\alpha}\Big)\frac{C^2_{\Omega,\Gamma_D}}{C_1}\|f+\operatorname{div}\mathbf{y}^*\|^2_{0,\Omega}. \quad (24)$$

- The upper estimate (24) was first obtained in [19] using complicated tools of the duality theory, and later it was also obtained in [21] for the Poisson equation, using the Helmholz decomposition of $L_2(\Omega,\mathbf{R}^d)$. The estimate (23) is derived in [22] using the duality theory again. Our approach of derivation of the estimates is different from those used in the above mentioned works and is simplier.

- In the case of pure Dirichlet conditions, only the constant C_{Ω,Γ_D} has to be computed or estimated from above.

- In the case of pure Dirichlet condition and if $c \geq c_0 > 0$ in Ω, we need not estimate any constants at all.

In what follows we shall use the following denotations for the upper and lower bounds of the error in the energy norm (8)

$$M^\oplus(\bar{u},\mathbf{y}^*,\alpha,\beta) = \Big\|\frac{1}{\sqrt{c}}(f+\operatorname{div}\mathbf{y}^* - c\bar{u})\Big\|^2_{0,\Omega^c} + (1+\alpha)|\!|\!|A^{-1}\mathbf{y}^* - \nabla\bar{u}|\!|\!|^2_\Omega$$
$$+\Big(1+\frac{1}{\alpha}\Big)(1+\beta)\frac{C^2_{\Omega,\Gamma_D}}{C_1}\|f+\operatorname{div}\mathbf{y}^*\|^2_{0,\Omega\setminus\overline{\Omega}^c} + \Big(1+\frac{1}{\alpha}\Big)\Big(1+\frac{1}{\beta}\Big)C^2_{\Omega,\partial\Omega}\|g-\nu^T\cdot\mathbf{y}^*\|^2_{0,\Gamma_N},$$

and

$$M^\ominus(\bar{u},w) = 2(J(\bar{u}) - J(w)).$$

Sometimes we shall use only a short denotation $M^\oplus$ or $M^\ominus$ for the corresponding bounds if it does not lead to misunderstanding.

4. Two-sided estimates for local errors

Two-sided estimates for controlling the error $u-\bar{u}$ in terms of linear functional (9) are essentially based on the usage of an auxiliary (often called *adjoint*) problem formulated below.

Adjoint problem: Find $v \in H^1_{\Gamma_D}(\Omega)$ such that

$$\int_\Omega A\nabla v \cdot \nabla w \, dx + \int_\Omega cvw \, dx = \ell(w) \quad \forall w \in H^1_{\Gamma_D}(\Omega).$$

The adjoint problem can be rewritten in a shorter form similarly to the main problem (4): Find $v \in H^1_{\Gamma_D}(\Omega)$ such that $a(v,w) = \ell(w) \quad \forall w \in H^1_{\Gamma_D}(\Omega)$. In particular this means, that the bilinear forms of the main and adjoint problems coincide.

The adjoint problem is uniquely solvable due to the assumption that ℓ is a linear bounded functional. However, the exact solution v of it is usually very hard (or even impossible) to find in analytical form and, thus, we only have some approximation for v, which we denote by the symbol $\bar{v}$ in what follows, assuming again only that $\bar{v} \in H^1_{\Gamma_D}(\Omega)$.

Proposition 4.1 *(cf.* [8]*) The following error decomposition holds*

$$\ell(u-\bar{u}) = E_0(\bar{u},\bar{v}) + E_1(u-\bar{u}, v-\bar{v}),$$

where

$$E_0(\bar{u},\bar{v}) = \int_\Omega f\bar{v}\, dx + \int_{\Gamma_N} g\bar{v}\, ds - \int_\Omega A\nabla\bar{v}\cdot\nabla\bar{u}\, dx - \int_\Omega c\bar{v}\bar{u}\, dx, \tag{25}$$

$$E_1(u-\bar{u}, v-\bar{v}) = \int_\Omega A\nabla(u-\bar{u})\cdot\nabla(v-\bar{v})\, dx + \int_\Omega c(u-\bar{u})(v-\bar{v})\, dx.$$

The first term E_0 is, obviously, directly computable once we have $\bar{u}$ and $\bar{v}$ computed, but the term E_1 contains unknown gradients ∇u and ∇v. In order to estimate it, we notice first that $E_1(u-\bar{u}, v-\bar{v}) \equiv a(u-\bar{u}, v-\bar{v})$. Further, the following relation obviously holds for any positive α:

$$\begin{aligned} 2E_1(u-\bar{u}, v-\bar{v}) = a\Big(\alpha(u-\bar{u}) + \frac{1}{\alpha}(v-\bar{v}), \alpha(u-\bar{u}) + \frac{1}{\alpha}(v-\bar{v})\Big) \\ - \alpha^2 a(u-\bar{u}, u-\bar{u}) - \frac{1}{\alpha^2} a(v-\bar{v}, v-\bar{v}). \end{aligned} \tag{26}$$

The last two terms in the above identity present the errors in the energy norm for main and adjoint problems. Thus, we can immediately use the two-sided estimates from Section 3, written in somewhat simplified form:

$$M^\ominus \le a(u-\bar{u}, u-\bar{u}) \le M^\oplus, \quad M^\ominus_{ad} \le a(v-\bar{v}, v-\bar{v}) \le M^\oplus_{ad},$$

where subindex "*ad*" means that the corresponding estimate is obtained for the adjoint problem.

As far it concerns the first term in the right-hand side of (26), we observe that

$$a\Big(\alpha(u-\bar{u})+\frac{1}{\alpha}(v-\bar{v}),\alpha(u-\bar{u})+\frac{1}{\alpha}(v-\bar{v})\Big)=$$
$$=a\Big(\big(\alpha u+\frac{1}{\alpha}v\big)-\big(\alpha\bar{u}+\frac{1}{\alpha}\bar{v}\big),\big(\alpha u+\frac{1}{\alpha}v\big)-\big(\alpha\bar{u}+\frac{1}{\alpha}\bar{v}\big)\Big).$$

The function $\alpha u+\frac{1}{\alpha}v$ can be perceived as the solution of the following problem (called as the *mixed problem* in what follows): Find $u_\alpha\in u_0+H^1_{\Gamma_D}(\Omega)$ such that

$$\int_\Omega A\nabla u_\alpha\cdot\nabla w\,dx+\int_\Omega cu_\alpha w\,dx=\alpha F(w)+\frac{1}{\alpha}\ell(w)\quad\forall w\in H^1_{\Gamma_D}(\Omega),$$

which is uniquely solvable due to the fact that $\alpha F(w)+\frac{1}{\alpha}\ell(w)$ is, obviously, also linear bounded functional.

The function $\alpha\bar{u}+\frac{1}{\alpha}\bar{v}\in H^1_{\Gamma_D}(\Omega)$ can be considered as an approximation of u_α, and we can again apply the techniques of Section 3 in order to obtain the following two-sided estimates (writen again in simplified form)

$$M^\ominus_{mix}\le a\Big(\alpha(u-\bar{u})+\frac{1}{\alpha}(v-\bar{v}),\alpha(u-\bar{u})+\frac{1}{\alpha}(v-\bar{v})\Big)\le M^\oplus_{mix},$$

where subindex "*mix*" means that the estimates are obtained for the mixed problem.

Further, we immediately observe that

$$\frac{1}{2}\Big(M^\ominus_{mix}-\alpha^2M^\oplus-\frac{1}{\alpha^2}M^\oplus_{ad}\Big)\le E_1(u-\bar{u},v-\bar{v}),$$

and

$$E_1(u-\bar{u},v-\bar{v})\le\frac{1}{2}\Big(M^\oplus_{mix}-\alpha^2M^\ominus-\frac{1}{\alpha^2}M^\ominus_{ad}\Big).$$

The above considerations can be summarized as follows.

Proposition 4.2 *For the error in terms of linear functional $\ell(u-\bar{u})$ we have the following upper estimate*

$$\ell(u-\bar{u})\le E_0(\bar{u},\bar{v})+\frac{1}{2}\Big(M^\oplus_{mix}-\alpha^2M^\ominus-\frac{1}{\alpha^2}M^\ominus_{ad}\Big),$$

and the following lower estimate

$$\ell(u-\bar{u})\ge E_0(\bar{u},\bar{v})+\frac{1}{2}\Big(M^\ominus_{mix}-\alpha^2M^\oplus-\frac{1}{\alpha^2}M^\oplus_{ad}\Big),$$

where the directly computable term $E_0(\bar{u},\bar{v})$ is defined in (25).

Remark 4.1 *For practical realisations of the above technologies, see e.g.* [8, 9, 21, 22].

References

[1] M. Ainsworth and J. T. Oden: *A posteriori error estimation in finite element analysis.* John Wiley & Sons, Inc., 2000.

[2] I. Babuška and W. C. Rheinbold: *Error estimates for adaptive finite element computations.* SIAM J. Numer. Anal., **15**, 1978, 36–754.

[3] I. Babuška and T. Strouboulis: *The finite element method and its reliability.* Oxford University Press Inc., New York, 2001.

[4] W. Bangerth and R. Rannacher: *Adaptive finite element methods for differential equations.* Lectures in Mathematics ETH Zürich, Birkhäuser Verlag, Basel, 2003.

[5] R. Becker and R. Rannacher: *A feed-back approach to error control in finite element methods: Basic approach and examples.* East-West J. Numer. Math., **4**, 1996, 237–264.

[6] Ph. G. Ciarlet: *The finite element method for elliptic problems.* Studies in Mathematics and its Applications, **4**, North-Holland Publishing Co., Amsterdam-New York-Oxford, 1978.

[7] I. Faragó and J. Karátson: *Numerical solution of nonlinear elliptic problems via preconditioning operators: theory and applications.* Advances in Computation: Theory and Practice, **11**, Nova Science Publishers, Inc., Hauppauge, NY, 2002.

[8] A. Hannukainen and S. Korotov: *Techniques for a posteriori error estimation in terms of linear functionals for elliptic type boundary value problems.* Far East J. Appl. Math. **21**, 2005, 289–304.

[9] A. Hannukainen and S. Korotov: *Computational technologies for reliable control of global and local errors for linear elliptic type boundary value problems.* Preprint A494, Helsinki University of Techology (February 2006) (submitted).

[10] S. Korotov: *A posteriori error estimation for linear elliptic problems with mixed boundary conditions.* Preprint A495, Helsinki University of Techology, March 2006.

[11] S. Korotov: *A posteriori error estimation of goal-oriented quantities for elliptic type BVPs.* J. Comput. Appl. Math. **191**, 2, 2006, 216–227.

[12] S. Korotov: *Two-sided a posteriori error estimates for linear elliptic problems with mixed boundary conditions.* To appear in Appl. Math.

[13] S. Korotov and D. Kuzmin: *A new approach to a posteriori error estimation for convection-diffusion problems.* I. Getting started. Technical Report **335**, University of Dortmund, 2006. Submitted to SIAM J. Numer. Anal.

[14] S. Korotov, P. Neittaanmäki and S. Repin: *A posteriori error estimation of goal-oriented quantities by the superconvergence patch recovery.* J. Numer. Math. **11**, 2003, 33–59.

[15] M. Křížek and P. Neittaanmäki: *Finite element approximation of variational problems and applications.* Pitman Monographs and Surveys in Pure and Applied Mathematics, 50. Longman Scientific & Technical, Harlow; copublished in USA with John Wiley & Sons, Inc., New York, 1990.

[16] C. Lovadina and R. Stenberg: *Energy norm a posteriori error estimates for mixed finite element methods.* Math. Comp. **75**, 2006, 1659–1674.

[17] J. Nečas: *Les Méthodes Directes en Théorie des Équations Elliptiques.* Academia, Prague, 1967.

[18] P. Neittaanmäki and S. Repin: *Reliable methods for computer simulation.* Error Control and A Posteriori Estimates, Studies in Mathematics and its Applications, **33**, Elsevier Science B.V., Amsterdam, 2004.

[19] S. Repin: *A posteriori error estimation for nonlinear variational problems by duality theory.* Zap. Nauchn. Sem. S.-Peterburg. Otdel. Mat. Inst. Steklov. (POMI) **243**, 1997, 201–214.

[20] S. Repin: *Two-sided estimates of deviation from exact solutions of uniformly elliptic equations.* Amer. Math. Soc. Transl. **209**, 2003, 43–171.

[21] S. Repin, S. Sauter and A. Smolianski: *A posteriori error estimation for the Dirichlet problem with account of the error in the approximation of boundary conditions.* Computing **70**, 2003, 205–233.

[22] S. Repin, S. Sauter and A. Smolianski: *A posteriori error estimation for the Poisson equation with mixed Dirichlet/Neumann boundary conditions.* J. Comput. Appl. Math. **164/165**, 2004, 601–612.

[23] M. Rüter, S. Korotov and Ch. Steenbock: *Goal-oriented error estimates based on different FE-solution spaces for the primal and the dual problem with application to linear elastic fracture mechanics.* Comput. Mech. (in press).

[24] M. Rüter and E. Stein: *Goal-oriented a posteriori error estimates in linear fracture mechanics.* Comput. Methods Appl. Mech. Engrg. **195**, 2006, 251–278.

[25] T. Vejchodský: *Guaranteed and locally computable a posteriori error estimate.* IMA J. Numer. Anal. (in press).

[26] R. Verfürth: *A review of a posteriori error estimation and adaptive mesh-refinement techniques.* Wiley-Teubner, 1996.

BENCHMARK CALCULATIONS OF THE VARIABLE-DENSITY FLOW IN POROUS MEDIA[*]

Milan Hokr

1. Introduction

Variable-density (or density-driven, density-dependent) porous media flow problem is a coupled problem of water flow and solute transport: the water velocity as a result of the flow problem is a parameter in the solute transport problem (standard case) and the solution density as a parameter in the flow problem is dependent on concentration, a result of the transport problem (specific for variable-density flow) [1].

Several standard benchmark problems are used for tests of numerical schemes and simulation codes [2, 1]; they are mostly derived from real-world problems of seawater intrusion and salt deposits. We propose a new benchmark problem, with a configuration derived from a case-study of groundwater flow and contaminant transport in the former uranium leaching site Stráž pod Ralskem in the north of the Czech Republic. The improvement is in parametrization of the intensity of the density coupling, allowing to study the efficiency of numerical schemes in dependence on physical parameters and also to find the limits for using simpler numerical schemes for the variable-density flow problem.

2. Governing equations

The groundwater porous media flow with the Boussinesque approximation [2] is governed by the Darcy's law and the mass-balance (continuity) equation

$$\boldsymbol{u} = (\boldsymbol{K}(\nabla h + \varrho_r \nabla z)), \qquad \kappa \frac{\partial h}{\partial t} - \nabla \cdot \boldsymbol{u} = q\,, \tag{1}$$

where h is the pressure head, ϱ_r is the relative solution density (with respect to the fresh water density), $\boldsymbol{u}$ is the Darcy velocity, q is the source/sink rate, $\boldsymbol{K}$ is the hydraulic conductivity tensor, κ is the storativity coefficient, and z is the vertical coordinate. The solute transport is governed by the advection-diffusion equation

$$\frac{\partial\,(nc)}{\partial t} + \nabla \cdot (\boldsymbol{u}c) - \nabla \cdot (n\boldsymbol{D}\nabla c) = qc_0\,, \tag{2}$$

[*] This work was supported with the subvention from the Grant Agency of the Czech Republic, project code 102/05/P284.

where c is the solute concentration, c_0 is a concentration in the source/sink, $\boldsymbol{D}$ is the hydrodynamic dispersion tensor [2], and n is the porosity.

The flow and transport equations are coupled through the Darcy velocity $\boldsymbol{u} = \boldsymbol{K}(\nabla h + \varrho_r \nabla z)$ and through the relative density, which is a function of the concentration, in the simplest case $\varrho_r(c) = 1 + c/\varrho_0$, where ϱ_0 is the fresh water density.

3. Numerical schemes

We use two schemes (MHFEM and CVFEM) denoted by the name of the method used for the flow problem. In both schemes, the advective transport problem is solved by principally same upwind finite volumes (the only difference is the position of the control volume – primal or dual mesh, see below). The hydrodynamic dispersion term is not evaluated in neither of the schemes. The main motivation for the choice of these numerical methods is the consistent discrete representation of velocity in both the flow and transport schemes, preserving the local mass balance. Both the methods use a discretization with trilateral prisms, which allow to use simpler mesh topology with the horizontal triangulation and the prisms ordered to layers and columns.

We use the computer codes (different for each method) developed before for general groundwater problems, with the variable-density term recently added. The codes have been successfully tested in several model and real-world problems.

3.1. Mixed-hybrid finite-element scheme

The MHFEM scheme is based on the weak formulation of the system of equations (1) on a system of elements $e \in \mathcal{E}_h$ with the additional constraint condition of mass balance expressed by Lagrange multipliers [5]. Thus, there are three unknown functions approximated with the following discrete spaces: the pressure head h by piecewise constant functions (in elements), the Lagrange multipliers (physically "pressure on inter-element interfaces") by piecewise constant functions on sides, and the velocity $\boldsymbol{u}$ by piecewise linear vector functions (lowest-order Raviart-Thomas space). The exact formulation for the specific case of trilateral prismatic elements is given in [5].

The approximation of the variable-density term results directly in the right-hand side of the weak formulation of the first equation of (1), i.e.

$$\sum_{e\in\mathcal{E}_h} \{ (\boldsymbol{A}\boldsymbol{u}^e, \boldsymbol{v}^e)_{0,e} - (p^e, \nabla \cdot \boldsymbol{v}^e)_{0,e} + \langle \lambda^e, \boldsymbol{\nu}^e \cdot \boldsymbol{v}^e \rangle_{\partial e \cap \Gamma_h} \} =$$

$$\sum_{e\in\mathcal{E}_h} \{ \langle p_D, \boldsymbol{\nu}^e \cdot \boldsymbol{v}^e \rangle_{\partial e \cap \partial\Omega_D} + \langle \varrho_r z, \boldsymbol{v}^e \cdot \boldsymbol{\nu}^e \rangle_{\partial e} - (\varrho_r z, \nabla \cdot \boldsymbol{v}^e)_{0,e} \}, \qquad (3)$$

where $\boldsymbol{A} = \boldsymbol{K}^{-1}$, $\boldsymbol{v}$ are test functions from the same Raviart-Thomas space as $\boldsymbol{u}$, $\boldsymbol{\nu}$ is the outward normal vector, $(\cdot,\cdot)_{0,e}$ and $\langle \cdot,\cdot \rangle_{\partial e}$ are the L_2 scalar products on the element volume and the element boundary respectively, $\partial\Omega_D$ is the Dirichlet boundary, and p_D is the boundary value of p. In the discrete form, the last two terms on the right-hand side are evaluated as a difference between the z coordinates of the mass centre of the element and the mass centre of the particular side. The time discretisation is implicit Euler, but in the calculations below we use a sequence of steady states with variable parameters, which corresponds to a very large value of the storativity κ.

The finite volume scheme for the transport problem is described in [3]; the cells are geometrically identical with the elements of MHFEM flow problem solution, we use the cell-centred approximation, the upwind weighting of the advective flux, and the explicit time discretisation. The MHFEM discrete unknowns of the velocity approximation are the fluxes through element sides, conservative with respect to the element volumes, which are directly the input value for the discrete advection term.

3.2. Control-volume finite-element scheme

The CVFEM scheme is based on a combination of two ideas: understanding the basic piecewise linear finite element solution with the triangular mesh as a finite volume scheme on the dual mesh (control volumes around the mesh nodes) and combining the FE scheme for 2D horizontal triangulation with the finite differences for the vertical discretization. This technique including the variable-density term in a mass-balance form is derived in [4].

The weak formulation, semidiscrete in the vertical direction, for a layer k is

$$(K^{xy}\nabla_{xy}h_k, \nabla_{xy}\phi_k)_{\Omega_k} - \left(\frac{1}{\Delta z_k}\left[K^z_{k+\frac{1}{2}}\frac{h_{k+1}-h_k}{\Delta z_{k+\frac{1}{2}}} - K^z_{k-\frac{1}{2}}\frac{h_k-h_{k-1}}{\Delta z_{k-\frac{1}{2}}}\right], \phi_k\right)_{\Omega_k} = (q_k, \phi_k)_{\Omega_k}, \tag{4}$$

where K^{xy} and K^z are components of $\boldsymbol{K}$ in the x, y directions and z direction respectively, ∇_{xy} is the ∇ operator in the xy direction, $\Delta z_{k+\frac{1}{2}}$ is the vertical discretisation step between the layers k and $k+1$, $(\cdot,\cdot)_{\Omega_k}$ is the L_2 scalar product in the layer k (horizontal projection of problem domain Ω), ϕ_k is a piecewise linear test function.

The pressures and the concentrations are evaluated in the mesh nodes, the velocity is represented as fluxes along mesh edges, the flux between nodes i and j is $u_{ij} = \mathbb{A}_{ij}(h_i - h_j)$, where $\mathbb{A}$ is the global stifness matrix, and h_i, h_j are the nodal values of pressure head.

3.3. Variable-density coupling

The model uses the explicit time stepping, i.e. in each time step, the flow problem is solved with the density distribution from the previous time step and then the transport problem is solved with the updated velocity field. This approach requires a small time step. The benchmark below is sensitive to change of the coupling time step in the beginning of the time interval, but the sufficient time step is still 10 times larger than the stability condition given by the upwind scheme for the solute transport. In the calculations, the time step is 40 days for the transport scheme and 360 days for the coupling iterations.

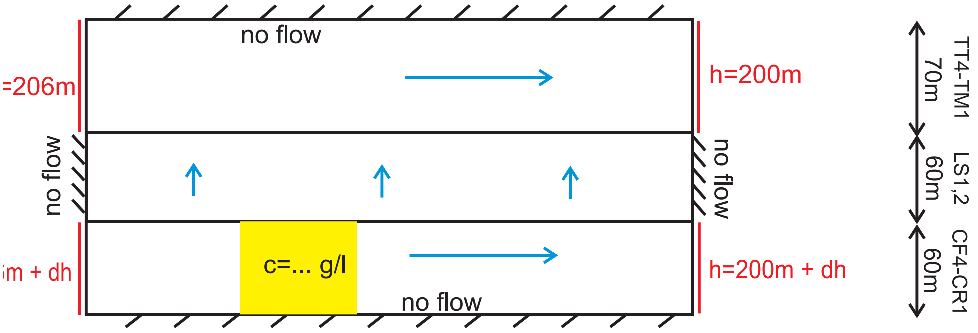


Fig. 1: *Configuration of the benchmark problem, position of boundary and initial conditions.*

Layer code	K_x, K_y	K_z	n	dz	$c_{ini}^{(10)}$	$c_{ini}^{(30)}$	$c_{ini}^{(50)}$
	m/day	m/day	1	m	g/l	g/l	g/l
TT4 – TT1	6 – 10	6 – 10	0.07	10–15	0	0	0
TM2 – TM1	0.4	0.1	0.07	12.5	0	0	0
LS2 – LS1	1e-4	4e-4	0.05	30	0	0	0
CF4 – CF3	0.5	0.25	0.08	12.5	10	25	40
CF2	0.05	0.025	0.04	7.5	10	20	30
CF1	0.5	0.25	0.08	7.5	10	25	40
CR2 – CR1	2 – 4	2	0.1	8–12	10	30	50

Tab. 1: *Discretization and material parameters in the benchmark: horizontal and vertical conductivity, porosity, layer thickness, and three variants of initial concentration. Some lines represent multiple layers with slightly variable parameters.*

4. Benchmark structure

4.1. Discretization and material parameters

The benchmark problem is built as geometrically simple domain representing the most of the character of the real groundwater system in Stráž pod Ralskem. The domain is 2000 m long (left–right), 190 m high and 40 m wide (front–back), discretized with prisms coupled in hexahedrons, each of the size 40 × 40 × dz (the thickness varies). The vertical discretisation is by 14 layers with thickness dz according to the real geological structure (Fig. 1, Tab. 1). We use the codes originated from the rock names: "T" the top permeable part (aquifer), "L" the semi-isolator, "C" the bottom part (aquifer).

4.2. Boundary conditions

The boundary conditions are Dirichlet (prescribed pressure head h) and homogeneous Neumann (zero flux $\boldsymbol{u} \cdot \boldsymbol{\nu} = 0$) for the flow problem (Fig. 1). The pressure head difference between the bottom and the top part is a parameter dh, representing the intensity of the hydraulic force in comparison with the gravity force on denser liquid (larger dh means less density-dependent coupling), $dh = 1\,\mathrm{m}$, $3\,\mathrm{m}$, and $10\,\mathrm{m}$. For the solute advection problem, zero Dirichlet at the inflow boundary is prescribed (fresh water $c = 0$), no boundary condition is prescribed at the outflow boundary, and the position of zero flux boundaries is the same as for the flow problem.

4.3. Initial conditions

The initial distribution of head and velocity (flow problem) is given by the boundary conditions above (constant-density steady state). As the initial distribution of concentration (transport problem), we use a simple representation of a contamination plum in the bottom aquifer, with zero concentration elsewhere (Fig. 1).

The contamination plum is defined by constant concentration for each layer, with horizontal dimension (length) 280 m and position 200 m from the left, with vertical inhomogeneity given by field measurements. We use three variants (referred by the most bottom value) in Tab.1. They are the second parameter of density-coupling (the higher is the concentration, the more is the density influence).

5. Results

We observe the behaviour of the system in the time interval of 200 years. During this interval the contamination in the most permeable layers leaves the domain, but the slowly moving contamination in the less permeable layers moves to the central and the right part of the domain and the transfer upwards is well visible (Fig. 2).

The objective of the numerical benchmark study is to compare two different approximations (equations coupled/uncoupled), two different numerical schemes, and mesh refinement. The results are expressed by integral values of concentration over each layer of the discretization (total mass in a layer). This technique is kept from previous use of the benchmark for hydrogeological parametric studies.

5.1. Basic study of parameter influence

Table 2 compares the total transfer to the top aquifer for the combinations of the three values of the piezometric head difference dh and the three variants of initial contamination, calculated with MHFEM scheme. For each combination, we also compare the variable-density and the constant-density model formulation.

For the head difference $dh = 1\,\mathrm{m}$, the hydraulic force is small and the gravity force and the density-driven process dominate, so much that the mass transfer upwards partly decreases with rising concentration. For the head difference $dh = 3\,\mathrm{m}$ and $dh = 10\,\mathrm{m}$, the hydraulic force becomes more significant but the density effect keeps important. The smallest influence and the weakest coupling is as expected for $dh = 10\,\mathrm{m}$ and $c = 10\,\mathrm{g/l}$. The basic analysis in Tab. 2 documents the necessity of the variable-density model and a good sensitivity on the density approximation required for variable-density benchmarks.

Initial	$dh = 1$		$dh = 3$		$dh = 10$	
conc.	var.dens.	const.dens.	var.dens.	const.dens.	var.dens.	const.dens.
g/l	ton	ton	ton	ton	ton	ton
10	0.165	0.568	5.07	11.897	67.883	100.643
30	0.105	1.419	6.181	29.578	116.717	251.165
50	0.117	2.269	7.058	47.26	156.037	401.688

Tab. 2: *Evaluation of the parameter influence and comparison of the variable-density versus the constant-density approximation, by means of a single value of the total mass transfer to the upper aquifer (subdomain).*

	$dh = 1$ $c = 50$			$dh = 3$ $c = 30$			$dh = 10$ $c = 10$		
	orig	ref1	ref2	orig	ref1	ref2	orig	ref1	ref2
top	0.828	0.028	1E-04	6.645	1.091	0.184	75.55	66.72	48.11
isolator	33.46	14.13	3.579	62.04	31.01	13.58	105.2	138.7	88.78
bottom	167.1	141.2	57.82	120.4	125.1	56.9	24.58	52.45	54.23

Tab. 3: *Study of the mesh refinement in z direction, results expressed by three values of the total mass in the bottom, middle, and top part of the domain.*

5.2. Mesh refinement

We narrow the study to the following three combinations representing the weakest, medium and the strongest density coupling respectively: (a) $dh = 10\,\mathrm{m}$, $c = 10\,\mathrm{g/l}$, (b) $dh = 3\,\mathrm{m}$, $c = 30\,\mathrm{g/l}$, and (c) $dh = 1\,\mathrm{m}$, $c = 50\,\mathrm{g/l}$. The mesh is refined in the z direction, i.e. each layer in Tab. 1 is divided into two equal.

The results of CVFEM calculation[1] expressed as mass sums in each of the three parts are in Tab. 3. The density influence is similar in all the original and the refined meshes, but there is no visible convergence. Generally, finer mesh lead to smaller transfer to upper layers, which can be caused by smaller numerical diffusion. On the other hand, the overall trend visualised by concentration field is similar for all discretizations (Fig. 2). The difficulty for comparing the MHFEM and CVFEM schemes is in the different position of unknowns with respect to the material parameters in the layers. As examples of secondary importance, the three corresponding values in Tabs. 2 and 3 are less different than with respect to the mesh refinement.

6. Conclusion

The results confirm the great enough sensitivity of the defined benchmark on the variable-density coupling. Moreover, the chosen parameters well cover the interval between the weak and strong coupling.

[1] The refinements for MHFEM were not evaluated, because the code uses external solver of the system of linear algebraic equations, which leads to very slow calculation in the iterations. We currently work on a more efficient implementation.

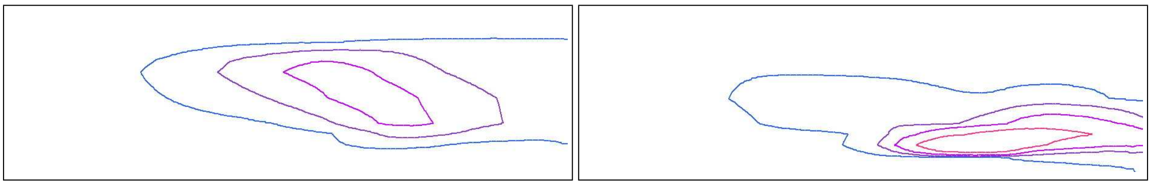


Fig. 2: *Isolines of concentration in the final time 200 years for the smallest ($dh = 10\,m$, $c = 10\,g/l$, left) and the largest ($dh = 1\,m$, $c = 50\,g/l$, right) density influence. The isoline values are (from outside) 0.1, 0.5, 1, and 2 g/l.*

On the other hand, the problem configuration and the used schemes do not allow to obtain mesh independent results. The reason can be that the influence of inhomogeneity inside the three subdomains and the changes of the numerical diffusion related to the mesh refinements amplify each other. The use of integral values also complicates the interpretation: in the bottom subdomain, there is a strong influence by escape of the mass from the domain (different in each layer of the mesh) and in the top subdomain, the value is inappropriately sensitive to the numerical approximation because it is a very small fraction of the original mass (large error relative to the local value, but smaller relative to the maximum or average value in the domain), e.g. in the case of the top layer value for $dh = 1\,\mathrm{m}$ and $c = 50\,\mathrm{g/l}$).

Here the solutions and evaluation criteria sufficient for the hydrogeological studies are not enough accurate for more exact statements on the numerical properties. We assume that an identical configuration without the internal material inhomogeneity and finer meshes in both the vertical and the horizontal directions, planned for future work, would give a better understanding of the solution behaviour.

References

[1] H.J.G. Diersch, O. Kolditz: *Variable-density flow and transport in porous media: approaches and challenges.* Adv. in Water Res. **25**, 2002, 899–944.

[2] E.O. Holzbecher: *Modeling density-driven flow in porous media: Basics, numerics, software.* Springer-Verlag Berlin and Heidelberg 1998.

[3] M. Hokr, J. Maryška, and J. Šembera: *Modelling of transport with nonequilibrium effects in dualporosity media.* In: Chen, Glowinski, Li (eds.), Current Trends in Scientific Computing, Amer. Math. Soc., 2003, 175–182.

[4] M. Hokr and V. Wasserbauer: *Velocity approximation in finite-element method for density-driven porous media flow.* In: *Sborník 3. Matematický workshop s mezinárodní účastí*, FAST VUT Brno, 2004, 49–50, full paper on CD.

[5] J. Maryška, M. Rozložník, and M. Tůma: *Mixed-hybrid finite-element approximation of the potential fluid flow problem.* J. Comput. Appl. Math. **63**, 1995, 383–392.

REMARK ON COMPUTING THE ANALYTIC SVD*

Dáša Janovská, Vladimír Janovský

Abstract

A new technique for computing Analytic SVD is proposed. The idea is to follow branches for just one selected singular value and the corresponding left/right singular vector.

1. Introduction

A singular value decomposition (SVD) of a real matrix $A \in \mathbb{R}^{m\times n}$, $m \geq n$, is a factorization $A = U\Sigma V^T$, where $U \in \mathbb{R}^{m\times m}$ and $V \in \mathbb{R}^{n\times n}$ are orthogonal matrices and $\Sigma = \mathrm{diag}(s_1, \ldots, s_n) \in \mathbb{R}^{m\times n}$. The values s_i, $i = 1, \ldots, n$, are called singular values. They may be defined to be nonnegative and to be arranged in nonincreasing order.

Let A depend smoothly on a parameter $t \in \mathbb{R}$, $t \in [a, b]$. The aim is to construct a path of SVD's

$$A(t) = U(t)\Sigma(t)V(t)^T\,, \tag{1}$$

where $U(t)$, $\Sigma(t)$ and $V(t)$ depend smoothly on $t \in [a, b]$. If A is a real analytic matrix function on $[a, b]$, then there exists *Analytic Singular Value Decomposition* (ASVD), see [1]: There exists a factorization (1) that *interpolates* classical SVD defined at $t = a$, i.e.

- the factors $U(t)$, $V(t)$, and $\Sigma(t)$ are real analytic on $[a, b]$;
- for each $t \in [a, b]$, both $U(t) \in \mathbb{R}^{m\times m}$ and $V(t) \in \mathbb{R}^{n\times n}$ are orthogonal matrices and $\Sigma(t) = \mathrm{diag}(s_1(t), \ldots, s_n(t)) \in \mathbb{R}^{m\times n}$ is a diagonal matrix;
- at $t = a$, the matrices $U(a)$, $\Sigma(a)$ and $V(a)$ are the factors of the classical SVD of the matrix $A(a)$.

Diagonal entries $s_i(t) \in \mathbb{R}$ of $\Sigma(t)$ are called *singular values*. Due to the requirement of smoothness, singular values may be negative and also their ordering may be arbitrary. Under certain assumptions, ASVD may be uniquely determined by the factors at $t = a$. For a theoretical background, see [9]. As far as the computation is concerned, an incremental technique is proposed in [1]: Given a point on the path,

*The research of both authors was partially supported by the Grant Agency of the Czech Republic (grant No. 201/06/0356). The first and the second author acknowledge financial support by the research projects MSM 6046137306 and MSM 0021620839, respectively, of The Ministry of Education, Youth and Sports, Czech Republic.

one computes a classical SVD for a neighboring parameter value. Next, one computes permutation matrices which link the classical SVD to the next point on the path. The procedure is approximative with a local error of order $O(h^2)$, where h is the step size.

An alternative technique for computing ASVD is presented in [12]: A non-autonomous vector field $H : \mathbb{R} \times \mathbb{R}^N \to \mathbb{R}^N$ of a huge dimension $N = n + n^2 + m^2$ can be constructed in such a way that the solution of the initial value problem for the system $x' = H(t, x)$ is linked to the path of ASVD. Moreover, [12] contributes to the analysis of *non-generic points*, see [1], of the ASVD path. These points could be, in fact, interpreted as singularities of the vector field H. In [11], both approaches are compared.

A continuation algorithm for computing ASVD is presented in [7]. It follows a path of a few *selected* singular values and left/right singular vectors. It is aimed to treat large sparse matrices. The continuation algorithm is of a predictor-corrector type. The relevant predictor is based on Euler method hence on an ODE solver. In this respect, there is a link to [12]. Nevertheless, the Newton-type corrector guarantees the solution with a prescribed precision.

The continuation may get stuck at the points, where a nonsimple singular value $s_i(t)$ turns up for a particular parameter t and index i. In [1, 12], such points are called non-generic points of the path. They are related to the branching of singular value paths. The code in [7] incorporates extrapolation strategies in order to "jump over" such a point.

In the present contribution, we will review the continuation proposed in [7], see Section 2. We suggest and investigate the idea to continue just **one** singular value and the corresponding left/right singular vector. Finally, we report on numerical experiments.

2. Preliminaries

Let us recall the notion of a singular value of a matrix $A \in \mathbb{R}^{m\times n}$, $m \geq n$:

Definition 2.1 *We say that $s \in \mathbb{R}$ is a singular value of the matrix A if there exist $u \in \mathbb{R}^m$ and $v \in \mathbb{R}^n$ such that*

$$Av - su = 0\,, \quad A^T u - sv = 0\,, \quad \|u\| = \|v\| = 1\,. \tag{2}$$

The vectors v and u are called the right and the left singular vectors of the matrix A.

Note that s is defined up to its sign: if the triplet (s, u, v) satisfies (2) then at least three more triplets

$$(s, -u, -v)\,, \quad (-s, -u, v)\,, \quad (-s, u, -v)\,,$$

can be interpreted as singular values, left and right singular vectors of A.

Definition 2.2 *For a given $s \in \mathbb{R}$, let us set*

$$\mathcal{M}(s) \equiv \begin{pmatrix} -sI_m & A \\ A^T & -sI_n \end{pmatrix},$$

where $I_m \in \mathbb{R}^{m\times m}$ and $I_n \in \mathbb{R}^{n\times n}$ are identities.

Definition 2.3 *We say that $s \in \mathbb{R}$ is a simple singular value of a matrix A if there exist $u \in \mathbb{R}^m$ and $v \in \mathbb{R}^n$ such that*

$$(s, u, v)\,, \quad (s, -u, -v)\,, \quad (-s, -u, v)\,, \quad (-s, u, -v)$$

are the only solutions to (2)*. A singular value s which is not a simple singular value is called nonsimple singular value.*

Remark 2.1 *Let $s \neq 0$.*

1. *s is a simple singular value of A if and only if* $\dim \operatorname{Ker} \mathcal{M}(s) = 1$*.*
2. *s is a simple singular value of A if and only if s^2 is a simple eigenvalue of A^TA. In particular, $v \in \mathbb{R}^n$ and $u \in \mathbb{R}^m$,*

$$A^TAv = s^2 v\,, \quad \|v\| = 1\,, \quad u = \frac{1}{s}Av\,,$$

are the relevant right and left singular vectors of A.

Remark 2.2 *$s = 0$ is a simple singular value of A if and only if $m = n$ and* $\dim \operatorname{Ker} A = 1$*.*

Remark 2.3 *Let s_i, s_j, $s_i \neq s_j$, be simple singular values of A. Then $s_i \neq \pm s_j$.*

Let us recall the idea of [7]: The branches of *selected* singular values and corresponding left/right singular vectors $s_i(t)$, $U_i(t) \in \mathbb{R}^m$, $V_i(t) \in \mathbb{R}^n$ are considered i.e.,

$$A(t)V_i(t) = s_i(t)U_i(t)\,, \quad A(t)^TU_i(t) = s_i(t)V_i(t)\,, \tag{3}$$

$$U_i(t)^TU_i(t) = V_i(t)^TV_i(t) = 1 \tag{4}$$

for $t \in [a, b]$. The natural orthogonality conditions $U_i(t)^TU_j(t) = V_i(t)^TV_j(t) = 0$, $i \neq j$, $t \in [a, b]$, are added. Given p, $p \leq n$, the selected singular values $S(t) = (s_1(t), \ldots, s_p(t)) \in \mathbb{R}^p$, and the corresponding left/right singular vectors $U(t) = [U_1(t), \ldots, U_p(t)] \in \mathbb{R}^{m\times p}$, $V(t) = [V_1(t), \ldots, V_p(t)] \in \mathbb{R}^{n\times p}$ are followed as $t \in [a, b]$.

In the operator setting, let

$$F : \mathbb{R} \times \mathbb{R}^p \times \mathbb{R}^{m\times p} \times \mathbb{R}^{n\times p} \to \mathbb{R}^{m\times p} \times \mathbb{R}^{n\times p} \times \mathbb{R}^{p\times p} \times \mathbb{R}^{p\times p} \tag{5}$$

be defined as

$$F(t,X) \equiv \left(A(t)V - U\Sigma, A^T(t)U - V\Sigma, U^TU - I, V^TV - I\right) \;, \tag{6}$$

where $X \equiv (S,U,V) \in \mathbb{R}^p \times \mathbb{R}^{m\times p} \times \mathbb{R}^{n\times p}$, $\Sigma = \operatorname{diag}(S)$ and $I \in \mathbb{R}^{p\times p}$ is the identity. Under certain assumptions, the set of *overdetermined nonlinear equations*

$$F(t,X) = 0 \tag{7}$$

implicitly defines a curve in $\mathbb{R} \times \mathbb{R}^N$, where $\mathbb{R}^N$, $N = p(1+n+m)$, and $\mathbb{R}^p \times \mathbb{R}^{m\times p} \times \mathbb{R}^{n\times p}$ are isomorphic. The image of F, namely $\mathbb{R}^{m\times p} \times \mathbb{R}^{n\times p} \times \mathbb{R}^{p\times p} \times \mathbb{R}^{p\times p}$, and $\mathbb{R}^M$, $M = p(m+n+2p)$, are isomorphic.

The curve (7) can be parameterized by t, i.e. $t \mapsto X(t) = (S(t), U(t), V(t))$ so that $F(t, X(t)) = 0$ as $t \in [a,b]$. Given a solution $X(t)$ at $t = a$, the curve is initialized. For this purpose, we may select p singular values and left/right singular vectors computed via the classical SVD of the matrix $A(a)$, see e.g. [4].

In [7], the *tangent continuation*, see [2], Algorithm 4.25, p. 107, is applied. It is a predictor-corrector algorithm with an adaptive stepsize control. Let us note that the sparsity of $A(t)$ as $t \in [a,b]$ can be exploited.

3. Continuation of a single singular value

In this section, we will consider the idea of pathfollowing of **one** singular value and the corresponding left/right singular vector. We will expect the path to be locally a branch $s_i(t)$, $U_i(t) \in \mathbb{R}^m$, $V_i(t) \in \mathbb{R}^n$ satisfying conditions (3)&(4) for $t \in [a,b]$.

We consider the i-th branch, $1 \le i \le m$, namely, the branch which is initialized by $s_i(a)$, $U_i(a) \in \mathbb{R}^m$, $V_i(a) \in \mathbb{R}^n$ computed by the classical SVD, see [4]. Note that the SVD algorithm orders all singular values in descending order $s_1(a) \ge \ldots \ge s_i(a) \ge \ldots \ge s_m(a) \ge 0$. We assume that $s_i(a)$ is **simple**. For the analysis of this assumption, see Remark 2.1 and Remark 2.2.

Remark 3.1 *Let $s \neq 0$.*

1. *If $\mathcal{M}(s) \begin{pmatrix} u \\ v \end{pmatrix} = 0$ then $u^Tu = v^Tv$.*
2. *If in addition $\mathcal{M}(s) \begin{pmatrix} \tilde{u} \\ \tilde{v} \end{pmatrix} = 0$ then $u^T\tilde{u} = v^T\tilde{v}$.*
3. *s is a singular value of A if and only if $\dim \operatorname{Ker} \mathcal{M}(s) \ge 1$.*

For $\mathcal{M}(s)$, see Definition 2.2.

Note that if $s_i(t) \neq 0$ then due to Remark 3.1 one of the scaling conditions (4) is **redundant**. It motivates the following

Definition 3.1 *Consider a mapping*

$$f : \mathbb{R} \times \mathbb{R}^{1+m+n} \longrightarrow \mathbb{R}^{1+m+n} \,,$$

$$t \in \mathbb{R} \,, \quad x = (s, u, v) \in \mathbb{R}^1 \times \mathbb{R}^m \times \mathbb{R}^n \longmapsto f(t, x) \in \mathbb{R}^{1+m+n} \,,$$

where

$$f(t, x) \equiv \left(\begin{array}{l} -su + A(t)v \\ A^T(t)u - sv \\ v^T v - 1 \end{array} \right) . \tag{8}$$

As an alternative to (8) we will also use

$$f(t, x) \equiv \left(\begin{array}{l} -su + A(t)v \\ A^T(t)u - sv \\ u^T u + v^T v - 2 \end{array} \right) \tag{9}$$

with an equivalent scaling.

The equation

$$f(t, x) = 0 \,, \quad x = (s, u, v) \,, \tag{10}$$

may locally define a branch $x(t) = (s(t), u(t), v(t)) \in \mathbb{R}^{1+m+n}$ of singular values $s(t)$ and left/right singular vectors $u(t)$ and $v(t)$. The branch is initialized at t^0 that plays the role of $t(a)$. It is assumed that there exists $x^0 \in \mathbb{R}^{1+m+n}$ such that $f(t^0, x^0) = 0$. The initial condition $x^0 = (s^0, u^0, v^0) \in \mathbb{R}^{1+m+n}$ plays the role of already computed SVD-factors $s_i(a) \in \mathbb{R}^1$, $U_i(a) \in \mathbb{R}^m$ and $V_i(a) \in \mathbb{R}^n$.

We solve (10) on an open interval $\mathcal{J}$ of parameters t such that $t^0 \in \mathcal{J}$.

Theorem 3.1 *Let* $(t^0, x^0) \in \mathcal{J} \times \mathbb{R}^{1+m+n}$*,* $x^0 = (s^0, u^0, v^0)$ *be a root of* $f(t^0, x^0) = 0$*. Assume that* $s^0 \neq 0$ *is a simple singular value of* $A(t^0)$*.*

Then there exists an open subinterval $\mathcal{I} \subset \mathcal{J}$ *containing* t^0 *and a unique function* $t \in \mathcal{I} \longmapsto x(t) \in \mathbb{R}^{1+m+n}$ *such that* $f(t, x(t)) = 0$ *for all* $t \in \mathcal{I}$ *and that* $x(t^0) = x^0$*. Moreover, if* $A \in C^k(\mathcal{I}, \mathbb{R}^{m \times n})$*,* $k \geq 1$*, then* $x \in C^k(\mathcal{I}, \mathbb{R}^{1+m+n})$*. If* $A \in C^\omega(\mathcal{I}, \mathbb{R}^{m \times n})$ *then* $x \in C^\omega(\mathcal{I}, \mathbb{R}^{1+m+n})$*.*

Proof Note that the assumptions yield that the partial differential $f_x(t^0, x^0) \in \mathbb{R}^{1+m+n} \times \mathbb{R}^{1+m+n}$ at (t^0, x^0) is a regular matrix.

Assuming $A \in C^k(\mathcal{I}, \mathbb{R}^{m \times n})$, $k \geq 1$, the statement is a consequence of Implicit Function Theorem, see e.g. [6]. In case that $A \in C^\omega(\mathcal{I}, \mathbb{R}^{m \times n})$, i.e. A is real analytic, again Implicit Function Theorem holds, see [10]. ♢

In case that $s^0 = 0$ is a simple singular value of $A(t^0)$, see Remark 2.2, the analysis is much more complicated. In the present paper we prefer to announce the result as a conjecture:

Conjecture 3.1 *Let $(t^0, x^0) \in \mathcal{J} \times \mathbb{R}^{1+m+n}$, $x^0 = (s^0, u^0, v^0)$ be a root of $f(t^0, x^0) = 0$. Assume that $s^0 = 0$ is a simple singular value of $A(t^0)$ i.e. $m = n$ and $\dim \operatorname{Ker} A(t^0) = 1$. Let $(u^0)^T A'(t^0) v^0 \neq 0$.*

Then there exists an open subinterval $\mathcal{I} \subset \mathcal{J}$ containing t^0 and a unique function $t \in \mathcal{I} \longmapsto x(t) \in \mathbb{R}^{1+2n}$ such that $f(t, x(t)) = 0$ for all $t \in \mathcal{I}$ and $x(t^0) = x^0$. Moreover, if $A \in C^k(\mathcal{I}, \mathbb{R}^{n \times n})$, $k \geq 1$, then $x \in C^k(\mathcal{I}, \mathbb{R}^{1+2n})$. If $A \in C^\omega(\mathcal{I}, \mathbb{R}^{n \times n})$ then $x \in C^\omega(\mathcal{I}, \mathbb{R}^{1+2n})$.

Let us compare:

Remark 3.2 *Consider the defining equation (7) for $p = 1$. It represents an* **overdetermined** *system for $(t, X) \in \mathbb{R} \times \mathbb{R}^{1+m+n}$. In [7], the condition (7) is meant in the least-squares sense. The compatibility of the solution set to (7), see [2] p. 93 for the notion, has been checked a posteriori. On the other hand, the formulation via (10) suggests that the solution set (t, x) to (10) is under certain assumption an implicitly defined curve in $(t, x) \in \mathbb{R} \times \mathbb{R}^{1+m+n}$.*

The practical advantage of (10) is that we can use the ready-made packages for continuation of an implicitly defined curves. In particular, we implemented a Matlab toolbox MATCONT, [3].

In Conclusions to [7], we admitted that the continuation of a bunch of p selected singular values and the relevant left/right singular vectors may get stuck. Note that the same phenomena was reported as the alternative methods are concerned, see [1, 12, 11]. In Introduction we noted that the continuation problems are related to nonsimple singular values on the path (see Definition 2.3). In [1, 12], these points are called *non-generic.*

Pathfollowing of the solution set of (10) via MATCONT is very robust. It does not usually get stuck. On the other hand, one has to be careful when interpreting the results. In principle, the minimal stepsize `MinStepsize` should be sufficiently small.

In [8], the non-generic points of the path are considered. The claim is that a non-generic point does not persist a sufficiently small perturbation of $A(t)$. In other words, given an $A(t)$ on a finite interval $a \leq t \leq b$ then, “usually”, the set of non-generic points on the path is empty.

4. Numerical experiments

We consider the same problem as in [7] namely, the homotopy

$$A(t) = t\,A2 + (1 - t)\,A1\,, \quad t \in [0, 1]\,, \tag{11}$$

where the matrices

$$A1 \equiv \texttt{well1033.mtx}\,, \quad A2 \equiv \texttt{illc1033.mtx}$$

are taken over from `http://math.nist.gov/MatrixMarket/`. Note that $A1, A2 \in \mathbb{R}^{1033\times 320}$ are sparse while $A1$ and $A2$ are well and ill-conditioned. The aim is to continue

- 10 smallest singular values, left/right singular vectors of $A(t)$,
- 10 largest singular values, left/right singular vectors of $A(t)$.

The continuation is initialized at $t = 0$: The initial decomposition of $A1$ was computed via `SVDS`, see MATLAB Function Reference.

The results of continuation are resumed on Figure 1 and Figure 2. The branches are depicted in turns by solid and dash curves. This should underline that the branches do not cross each other. The computation complies with Theorem 3.1.

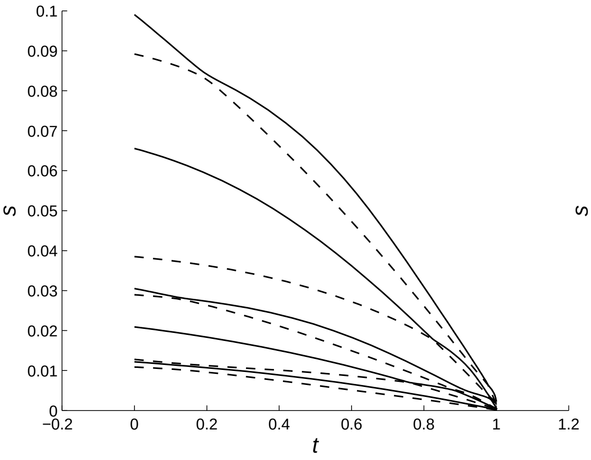


Fig. 1: *Ten smallest singular values s versus parameter t.*

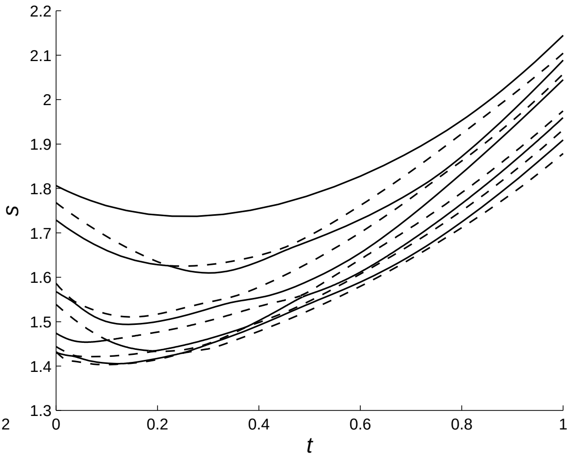


Fig. 2: *Ten largest singular values s versus parameter t.*

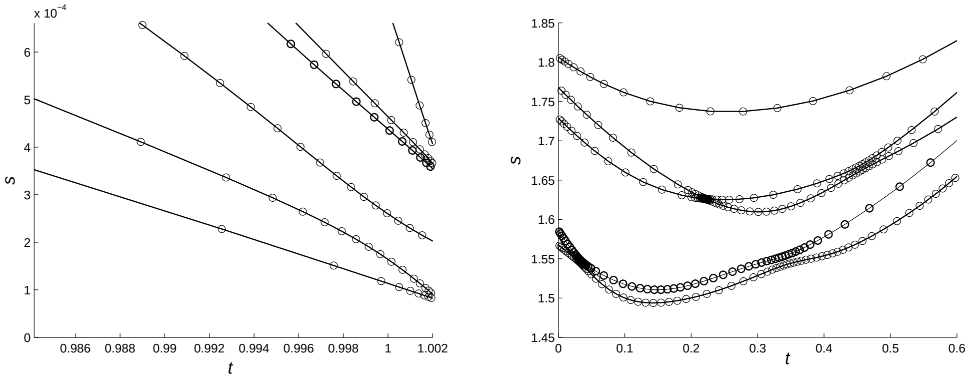


Fig. 3: *Zooms: Ten smallest s vs. t. Ten largest s vs. t.*

The zooms of the branches are shown on Figure 3. Each curve is computed as a sequence of isolated points marked by circles. The adaptive stepsize control refines the stepsize individually for each branch.

Note that the branches reported in [7] are not computed correctly. They cross each other occasionally: In the case of a stagnation, the continuation algorithm tries to jump over a prospective non-generic point on the path. A simple extrapolation strategy is used to continue. The branching scenario often suggests to follow a wrong branch. The message is that the branching is not generic.

In [7], the stepsize is always changed simultaneously for all p selected singular values. Treating each branch individually, see Figure 3, is much more efficient.

As the second example, we consider another homotopy

$$A(t) = t\,A3 + (1-t)\,A4\,, \quad t \in [-3, 10]\,, \tag{12}$$

where the matrices

$$A3 \equiv \texttt{cavity01.mtx}\,, \quad A4 \equiv \texttt{cavity02.mtx}$$

are taken over from `http://math.nist.gov/MatrixMarket/`. $A3, A4 \in \mathbb{R}^{317\times 317}$ are sparse square matrices.

The aim is to continue the smallest singular value and the relevant left/right singular vector over the interval $[-3, 10]$. The plot of the smallest singular value vs. t is shown on Figure 4. Note that $s(t)$ changes sign. It illustrates Conjecture 3.1: If the sign change occurs at t^0, the condition $(u^0)^T A'(t^0) v^0 \neq 0$ means that $s(t)$ crosses zero at t^0 "transversally", i.e. $s'(t^0) \neq 0$. It complies with the situation on Figure 4.

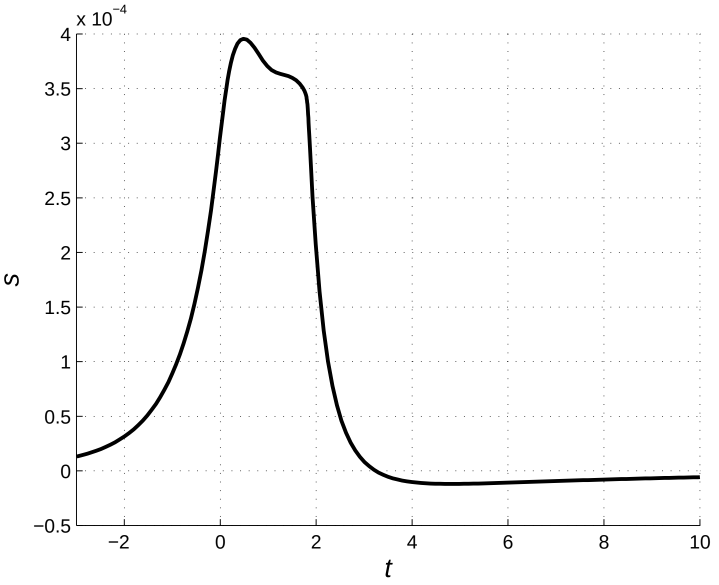


Fig. 4: *The smallest singular value s versus parameter t.*

5. Conclusions

In order to perform the Analytic SVD, we suggested to compute **separate** branches of singular values and the relevant left/right singular vectors. We can use any standard software for the pathfollowing of an implicitly defined curve. It seems, see [8], that the branches do not intersect generically. In other words, the branching scenario which concerns non-generic points, see [1, 12], does not persist sufficiently small perturbations of $A(t)$. So far, the claim is not rigorously proved. Nevertheless, the numerical experience supports the claim.

6. Appendix

We shall comment on Remark 3.1, Remark 2.1 and Remark 2.2. In particular, Remark 3.1 is based on Lemma 6.1 and Lemma 6.2, Remark 2.1 follows from Lemma 6.3 and Remark 2.2 is due to Lemma 6.4. Let us prove the Lemmas.

Lemma 6.1 *Let* $s \neq 0$, $\mathcal{M}(s)\begin{pmatrix} u \\ v \end{pmatrix} = 0$. *Then* $u^T u = v^T v$.

Proof By definition, we assume

$$-su + Av = 0\,, \quad A^T u - sv = 0\,.$$

Multiplying the first equation by u^T from the left and the second equation by v^T from the left, we get

$$u^T u = -\frac{1}{s} u^T Av\,, \quad v^T v = -\frac{1}{s} v^T A^T u\,.$$

Note that $v^T A^T u = (Av)^T u = u^T Av$. Therefore, $u^T u - v^T v = -\frac{1}{s}(u^T Av - u^T Av) = 0$. ♢

Lemma 6.2 *Let* $s \neq 0$, $\mathcal{M}(s)\begin{pmatrix} u \\ v \end{pmatrix} = 0$, $\mathcal{M}(s)\begin{pmatrix} \tilde{u} \\ \tilde{v} \end{pmatrix} = 0$. *Then* $u^T \tilde{u} = v^T \tilde{v}$.

Proof We assume

$$\begin{aligned} -su + Av &= 0\,, \quad A^T u - sv = 0\,, \\ -s\tilde{u} + A\tilde{v} &= 0\,, \quad A^T \tilde{u} - s\tilde{v} = 0\,. \end{aligned}$$

Therefore,

$$\begin{aligned} \tilde{u}^T(-su + Av) &= 0\,, \quad \tilde{v}^T(A^T u - sv) = 0\,, \\ u^T(-s\tilde{u} + A\tilde{v}) &= 0\,, \quad v^T(A^T \tilde{u} - s\tilde{v}) = 0\,. \end{aligned}$$

Since $s \neq 0$,

$$\tilde{u}^T u = -\frac{1}{s}\tilde{u}^T Av\,, \quad \tilde{v}^T u = -\frac{1}{s}\tilde{v}^T A^T u = -\frac{1}{s}(A\tilde{v})^T u$$

and

$$u^T \tilde{u} = -\frac{1}{s} u^T A \tilde{v} \,, \quad v^T \tilde{v} = -\frac{1}{s} v^T A^T \tilde{u} = -\frac{1}{s} (Av)^T \tilde{u} \,.$$

We conclude that

$$\tilde{u}^T u - v^T \tilde{v} = -\frac{1}{s} \tilde{u}^T Av + \frac{1}{s} (Av)^T \tilde{u} \,.$$

Since $v^T \tilde{v} = \tilde{v}^T v$ and $(Av)^T \tilde{u} = \tilde{u}^T Av$,

$$\tilde{u}^T u - \tilde{v}^T v = 0 \,.$$

$\diamondsuit$

Lemma 6.3 *A triplet $s \neq 0$, $u \in \mathbb{R}^m$ and $v \in \mathbb{R}^n$ satisfies (2) if and only if*

$$A^T Av = s^2 v \,, \quad u = \frac{1}{s} Av \,, \quad \|v\| = 1 \,, \quad s \neq 0 \,. \tag{13}$$

Proof Let $s \neq 0$, u and v satisfy (2). From the first equation in (2), $0 = Av - su = 0$, we conclude that $0 = A^T(Av - su) = A^T Av - sA^T s = A^T Av - s^2 v$ since $A^T u = sv$. Moreover, $su = Av$, i.e. $u = \dfrac{1}{s}\, Av$.

Let $s \neq 0$, u and v satisfy (13). Then $A^T u - su = A^T(\frac{1}{s} Av) - sv = \frac{1}{s} A^T Av - sv = sv - sv = 0$ and $Av - su = Av - s(\frac{1}{s} Av) = Av - Av = 0$. Finally, $u^T u = u^T(\frac{1}{s} Av) = \frac{1}{s} u^T Av = \frac{1}{s} (A^T u)^T v = \frac{1}{s}\, sv^T v = 1$. $\diamondsuit$

Note that a nonzero simple singular value s can be identified with a nonzero simple eigenvalue s^2 of the matrix $A^T A$, see Lemma 6.3.

Lemma 6.4 *$s = 0$ is a simple singular value of A if and only if $m = n$ and $\dim \mathrm{Ker} A = 1$.*

Proof Let $m = n$, $\dim \mathrm{Ker} A = 1$. As a consequence, $\dim \mathrm{Ker} A^T = 1$. Then there exist $u \in \mathbb{R}^m$ and $v \in \mathbb{R}^n$ such that

$$Av = 0 \,, \quad A^T u = 0 \,, \quad \|u\| = \|v\| = 1 \,, \tag{14}$$

i.e. $(s = 0, u, v)$ satisfies (2). Clearly, $(s = 0, u, v)$ and $(s = 0, -u, -v)$ and $(s = 0, -u, v)$ and $(s = 0, u, -v)$ are the only possibilities to solve (2).

If $m > n$ then $\dim \mathrm{Ker} A^T \geq 2$ and hence (14) has infinitely many solutions. If $\dim \mathrm{Ker} A \geq 2$, one can also find infinitely many solutions to (14). $\diamondsuit$

References

[1] A. Bunse-Gerstner, R. Byers, V. Mehrmann, N.K. Nichols: *Numerical computation of an analytic singular value decomposition of a matrix valued function.* Numer. Math. **60**, 1991, 1–39.

[2] P. Deuflhart, A. Hohmann: *Numerical analysis in modern scientific computing. An introduction.* New York, Springer Verlag 2003.

[3] A. Dhooge, W. Govaerts, Yu.A. Kuzetsov: *MATCONT: A Matlab package for numerical bifurcation analysis of ODEs.* ACM Transactions on Mathematical Software **31**, 2003, 141–164.

[4] G.H. Golub, C.F. van Loan: *Matrix computations.* 3rd ed. Baltimore, The Johns Hopkins University Press 1996.

[5] W. Govaerts: *Numerical methods for bifurcations of dynamical equilibria.* Philadelphia, SIAM 2000.

[6] S.N. Chow, J.K. Hale: *Methods of bifurcation theory.* New York, Springer-Verlag 1982.

[7] V. Janovský, D. Janovská, K. Tanabe: *Computing the analytic singular value decomposition via a pathfollowing.* In: Proceedings of ENUMATH 2005, Springer Verlag, New York, 2006, 911–918.

[8] D. Janovská, V. Janovský: *On non-generic points of the analytic SVD.* In: T.E. Simos, G. Psihoyios, Ch. Tsitouras (Eds.), International Conference on Numerical Analysis and Applied Mathematics 2006., WILEY-VCH Verlag, Weinheim, 2006, 162–165.

[9] T. Kato: *Perturbation theory for linear operators, second ed.* New York, Springer Verlag 1976.

[10] S. Krantz, H. Parks: *A primer of real analytic functions.* New York, Birkhäuser 2002.

[11] V. Mehrmann, W. Rath: *Numerical methods for the computation of analytic singular value decompositions: electronic transactions on numerical analysis* **1**, 1993, 72–88.

[12] K. Wright: *Differential equations for the analytic singular value decomposion of a matrix.* Numer. Math. **63**, 1992, 283–295.

A COMPUTATIONAL COMPARISON OF METHODS DIMINISHING SPURIOUS OSCILLATIONS IN FINITE ELEMENT SOLUTIONS OF CONVECTION–DIFFUSION EQUATIONS[*]

Volker John, Petr Knobloch

Abstract

This paper presents a review and a computational comparison of various stabilization techniques developed to diminish spurious oscillations in finite element solutions of scalar stationary convection–diffusion equations. All these methods are defined by enriching the popular SUPG discretization by additional stabilization terms. Although some of the methods can substantially enhance the quality of the discrete solutions in comparison to the SUPG method, any of the methods can fail in very simple situations and hence none of the methods can be regarded as reliable. We also present results obtained using the improved Mizukami–Hughes method which is often superior to techniques based on the SUPG method.

1. Introduction

During the past three decades, much effort has been devoted to the numerical solution of the scalar convection–diffusion equation

$$-\varepsilon\,\Delta u + \boldsymbol{b}\cdot\nabla u = f \quad \text{in } \Omega, \qquad\qquad u = u_b \quad \text{on } \partial\Omega. \tag{1}$$

Here $\Omega \subset \mathbb{R}^2$ is a bounded domain with a polygonal boundary $\partial\Omega$, $\varepsilon > 0$ is the constant diffusivity, $\boldsymbol{b} \in W^{1,\infty}(\Omega)^2$ is a given convective field, $f \in L^2(\Omega)$ is an outer force, and $u_b \in H^{1/2}(\partial\Omega)$ represents the Dirichlet boundary condition. In our numerical tests also less regular boundary conditions are considered.

Problem (1) describes the stationary distribution of a physical quantity u (e.g., temperature or concentration) determined by two basic physical mechanisms, namely the convection and diffusion. The broad interest in solving problem (1) is also caused by the fact that it is a simple model problem for convection–diffusion effects which appear in many more complicated problems arising in applications, e.g., in convection–dominated incompressible fluid flow problems which are described by the Navier–Stokes equations. Despite the apparent simplicity of problem (1), its numerical solution is by no means easy when convection is strongly dominant (i.e., when $\varepsilon \ll |\boldsymbol{b}|$). In this case, the solution of (1) typically possesses interior and boundary layers, which often leads to unwanted spurious (nonphysical) oscillations in the numerical solution.

[*]The research of Petr Knobloch is a part of the project MSM 0021620839 financed by MSMT and it was partly supported by the Grant Agency of Charles University in Prague under the grant No. 316/2006/B–MAT/MFF.

In this paper, we concentrate on the solution of (1) using the finite element method. The simplest finite element discretization of (1) is the classical Galerkin formulation which, in simple settings, is equivalent to a central finite difference discretization. Thus, it is not surprising that, in the convection dominated regime, the Galerkin solution is usually globally polluted by spurious oscillations and hence the Galerkin discretization is inappropriate.

To enhance the stability and accuracy of the Galerkin discretization of (1) in the convection dominated case, various stabilization strategies have been developed. The most popular stabilization technique within the framework of finite element discretizations of (1) is the streamline upwind/Petrov–Galerkin (SUPG) discretization proposed by Brooks and Hughes [2], see Section 2. It can be observed that the solutions obtained with the SUPG method possess often spurious oscillations in the vicinity of layers.

To diminish the oscillations of SUPG solutions, a large class of finite element methods has been constructed by adding yet additional stabilization terms to the SUPG discretization of (1). Usually, these terms depend on the element residuals of the discrete solution and therefore the resulting methods are consistent and hence higher–order accurate. We shall discuss such stabilization methods in Section 3. The stabilization terms introduce additional artificial diffusion and often depend on the unknown discrete solution in a nonlinear way. It is believed that, for a proper amount of artificial diffusion, we obtain a discrete solution which represents a good approximation of the solution of (1) and does not contain any spurious oscillations. Therefore, the design of suitable formulas specifying the artificial diffusion introduced by the stabilization terms was a subject of an extensive research during the past two decades.

The main aim of this paper is to present a computational comparison of the above–mentioned stabilization techniques by means of two standard test problems whose solutions possess characteristic features of solutions of (1). In addition, we shall introduce a new simple model problem of the type of (1) for which none of the above–mentioned stabilization methods gives a satisfactory discrete solution. This indicates the necessity to seek other ways of approximating the solution to the convection–diffusion equation (1). We also present results obtained using the improved Mizukami–Hughes method which is often superior to techniques based on the SUPG method and which gives good approximations to the solutions of all three test problems considered in this paper. In the whole paper we confine ourselves to conforming piecewise linear triangular finite elements.

The plan of the paper is as follows. In the next section, we formulate two discretizations of the problem (1): the Galerkin discretization and the SUPG method. In Section 3, we present a review of various additional stabilization terms added to the SUPG discretization to diminish spurious oscillations at layers. Also, we mention the improved Mizukami–Hughes method. Then the results of our numerical tests are reported in Section 4. Finally, the paper is closed by Section 5 containing our conclusions.

Throughout the paper, we use the standard notations $L^p(\Omega)$, $W^{k,p}(\Omega)$, $H^k(\Omega) = W^{k,2}(\Omega)$, etc. for the usual function spaces. The norm and seminorm in the Sobolev space $H^k(\Omega)$ will be denoted by $\|\cdot\|_{k,\Omega}$ and $|\cdot|_{k,\Omega}$, respectively. The inner product in the space $L^2(\Omega)$ or $L^2(\Omega)^2$ will be denoted by $(\cdot,\cdot)$. For a vector $\boldsymbol{a} \in \mathbb{R}^2$, we denote by $|\boldsymbol{a}|$ its Euclidean norm.

2. The Galerkin discretization of (1) and the SUPG method

To define a finite element discretization of (1), we introduce a triangulation $\mathcal{T}_h$ of the domain Ω consisting of a finite number of open triangular elements K possessing the usual compatibility properties. Using this triangulation, we define the finite element space

$$\mathrm{V}_h = \{v \in H_0^1(\Omega)\,;\; v|_K \in P_1(K) \quad \forall\; K \in \mathcal{T}_h\}\,,$$

where $P_1(K)$ is the space of linear functions on K. Further, we introduce a piecewise linear function $\widetilde{u}_{bh} \in H^1(\Omega)$ such that $\widetilde{u}_{bh}|_{\partial\Omega}$ approximates the boundary condition u_b. Then the usual Galerkin finite element discretization of the convection–diffusion equation (1) reads:

Find $u_h \in H^1(\Omega)$ such that $u_h - \widetilde{u}_{bh} \in \mathrm{V}_h$ and

$$a(u_h, v_h) = (f, v_h) \qquad \forall\; v_h \in \mathrm{V}_h\,,$$

where

$$a(u, v) = \varepsilon\,(\nabla u, \nabla v) + (\boldsymbol{b}\cdot\nabla u, v)\,.$$

Since the Galerkin method lacks stability if convection dominates diffusion, Brooks and Hughes [2] proposed to enrich it by a residual–based stabilization term yielding the streamline upwind/Petrov–Galerkin (SUPG) method:

Find $u_h \in H^1(\Omega)$ such that $u_h - \widetilde{u}_{bh} \in \mathrm{V}_h$ and

$$a(u_h, v_h) + (R(u_h), \tau\, \boldsymbol{b}\cdot\nabla v_h) = (f, v_h) \qquad \forall\; v_h \in \mathrm{V}_h\,, \tag{2}$$

where $\tau \in L^\infty(\Omega)$ is a nonnegative stabilization parameter and

$$R(u_h) = \boldsymbol{b}\cdot\nabla u_h - f$$

is the residual (note that $\Delta u_h = 0$ on any element of the triangulation).

A delicate problem is the choice of the stabilization parameter τ in (2). Theoretical investigations of the SUPG method (see, e.g., Roos *et al.* [20]) provide certain bounds for τ for which the SUPG method is stable and leads to (quasi–)optimal convergence of the discrete solution u_h. However, it has been reported many times that the choice of τ inside these bounds may dramatically influence the accuracy of the discrete solution. Therefore, over the last two decades, much research has also been devoted to the choice of τ and various strategies for the computation of τ have been proposed, see, e.g., the review in the recent paper by John and

Knobloch [14]. Let us stress that the definition of τ mostly relies on heuristic arguments and the 'best' way of choosing τ for general convection–diffusion problems is not known. Here we define τ on any element $K \in \mathcal{T}_h$ by the formula

$$\tau|_K \equiv \tau_K = \frac{h_K}{2\,|\boldsymbol{b}|}\,\xi(\mathrm{Pe}_K) \qquad \text{with} \qquad \mathrm{Pe}_K = \frac{|\boldsymbol{b}|\,h_K}{2\,\varepsilon}\,, \tag{3}$$

where h_K is the diameter of K in the direction of the convection $\boldsymbol{b}$, Pe_K is the local Péclet number and ξ is the so–called upwind function defined by $\xi(\alpha) = \coth\alpha - 1/\alpha$. If $\boldsymbol{b}|_K$ is not constant, then the parameters h_K, Pe_K and τ_K are generally functions of the points $x \in K$. The formula (3) is a generalization of an analogous one–dimensional formula which guarantees that, for the one–dimensional case of (1) with constant data, the SUPG solution with continuous piecewise linear finite elements on a uniform division of an interval Ω is nodally exact, c.f. Christie *et al.* [9].

3. A short review of stabilization methods based on the SUPG method

The SUPG method produces to a great extent accurate and oscillation–free solutions but it does not preclude spurious oscillations (overshooting and undershooting) localized in narrow regions along sharp layers. Although these nonphysical oscillations are usually small in magnitude, they are not permissible in many applications. An example are chemically reacting flows where it is essential to guarantee that the concentrations of all species are nonnegative. The small spurious oscillations may also deteriorate the solution of nonlinear problems, e.g., in two–equations turbulence models.

The oscillations along sharp layers are caused by the fact that the SUPG method is neither monotone nor monotonicity preserving (in contrast with the continuous problem (1)). Therefore, various terms introducing artificial crosswind diffusion in the neighborhood of layers have been proposed to be added to the SUPG formulation in order to obtain a method which is monotone, at least in some model cases, or which at least reduces the local oscillations. This procedure is referred to as discontinuity capturing or shock capturing. A detailed review of such methods was recently published by John and Knobloch [14].

Usually, the additional artificial diffusion is introduced by adding either the term

$$(\widetilde{\varepsilon}^{iso}\,\nabla u_h, \nabla v_h) \tag{4}$$

or the term

$$(\widetilde{\varepsilon}^{cd}\,D\,\nabla u_h, \nabla v_h) \qquad \text{with} \qquad D = \begin{cases} I - \dfrac{\boldsymbol{b}\otimes\boldsymbol{b}}{|\boldsymbol{b}|^2} & \text{if } \boldsymbol{b} \neq \boldsymbol{0}, \\ 0 & \text{if } \boldsymbol{b} = \boldsymbol{0} \end{cases} \tag{5}$$

to the left–hand side of (2). The former term introduces isotropic artificial diffusion whereas the latter one adds the artificial diffusion in the crosswind direction only (note that D is the projection onto the line orthogonal to $\boldsymbol{b}$, I being the identity tensor). A basic problem of all these methods is to find the proper amount of artificial diffusion which leads to sufficiently small nonphysical oscillations (requiring that artificial diffusion is not 'too small') and to a sufficiently high accuracy (requiring that artificial diffusion is not 'too large'). The derivation of formulas for $\widetilde{\varepsilon}^{iso}$ and $\widetilde{\varepsilon}^{cd}$ is typically based either on a convergence analysis or on investigations of the discrete maximum principle or (very often) on heuristic arguments. Usually, the parameter $\widetilde{\varepsilon}^{iso}$ or $\widetilde{\varepsilon}^{cd}$ depends on the unknown discrete solution u_h and hence the resulting method is nonlinear.

Many formulas for $\widetilde{\varepsilon}^{iso}$ rely on replacing the convection $\boldsymbol{b}$ in the SUPG weighting function by another upwind direction. This approach is used, e.g., in the methods of Hughes *et al.* [13], Tezduyar and Park [21], Galeão and do Carmo [12], do Carmo and Galeão [8] and in the modifications of these methods mentioned below. Let us mention at least the idea of Galeão and do Carmo [12]. They introduced an approximate streamline direction $\boldsymbol{b}_h$ for which the discrete solution u_h elementwise satisfies the equation (1) with $\boldsymbol{b}$ replaced by $\boldsymbol{b}_h$. Minimizing the difference between $\boldsymbol{b}$ and $\boldsymbol{b}_h$, they found that $\boldsymbol{b}_h = \boldsymbol{b} - \boldsymbol{z}_h$ with

$$\boldsymbol{z}_h = \frac{R(u_h)\,\nabla u_h}{|\nabla u_h|^2}\,.$$

(Here and in the following it is understood that, if $\nabla u_h = \boldsymbol{0}$ in the denominator, the respective expression is replaced by zero.) Finally, they replaced the function $\tau\,\boldsymbol{b}$ in (2) by $\tau\,\boldsymbol{b} + \sigma\,\boldsymbol{z}_h$ with a nonnegative parameter σ. That leads to the discretization (2) with the additional term (4) on the left–hand side, where

$$\widetilde{\varepsilon}^{iso} = \sigma\,\frac{|R(u_h)|^2}{|\nabla u_h|^2}\,. \tag{6}$$

Based on ideas of Hughes *et al.* [13], Galeão and do Carmo [12] defined the parameter σ by

$$\sigma = \max\{0, \tau(\boldsymbol{z}_h) - \tau(\boldsymbol{b})\}\,. \tag{7}$$

The notation of the type $\tau(\boldsymbol{b}^\star)$ denotes a value computed using the formula (3) with $\boldsymbol{b}$ replaced by some function $\boldsymbol{b}^\star$. Note that $\boldsymbol{b}^\star$ influences the value of $\tau_K(\boldsymbol{b}^\star)$ also through the definition of h_K.

Do Carmo and Galeão [8] proposed to simplify (7) to

$$\sigma = \tau(\boldsymbol{b})\,\max\left\{0, \frac{|\boldsymbol{b}|}{|\boldsymbol{z}_h|} - 1\right\}\,, \tag{8}$$

which assures that the term (4) is added only if the above–introduced vector $\boldsymbol{b}_h$ satisfies the natural requirement $\boldsymbol{b}\cdot\boldsymbol{b}_h > 0$. It may also be advantageous to set

$$\sigma = \tau(\boldsymbol{b})\,\max\left\{0, \frac{|\boldsymbol{b}|}{|\boldsymbol{z}_h|} - \zeta_h\right\} \quad \text{with} \quad \zeta_h = \max\left\{1, \frac{\boldsymbol{b}\cdot\nabla u_h}{R(u_h)}\right\}\,, \tag{9}$$

which was proposed by Almeida and Silva [1]. Further variants of this approach were developed by do Carmo and Galeão [8] and do Carmo and Alvarez [6] who proposed techniques which should suppress the addition of the artificial diffusion in regions where the solution of (1) is smooth. A finer tuning of the stabilization parameters was introduced by do Carmo and Alvarez [7]. Let us also mention that, motivated by assumptions of a rather general error analysis, Knopp *et al.* [18] suggested to replace (6), on any element $K \in \mathcal{T}_h$, by

$$\widetilde{\varepsilon}^{iso}|_K = \sigma\,|Q_K(u_h)|^2 \qquad \text{with} \qquad Q_K(u_h) = \frac{\|R(u_h)\|_{0,K}}{S_K + \|u_h\|_{1,K}}\,, \tag{10}$$

where S_K are appropriate positive constants. The stabilization term (4) was also used by Johnson [15] who proposed to set

$$\widetilde{\varepsilon}^{iso}|_K = \max\{0, \alpha\,[\mathrm{diam}(K)]^\nu\,|R(u_h)| - \varepsilon\} \qquad \forall\, K \in \mathcal{T}_h \tag{11}$$

with some constants α and $\nu \in (3/2, 2)$. He suggested to take $\nu \sim 2$.

The crosswind artificial diffusion term (5) was first considered by Johnson *et al.* [16]. A straightforward generalization of their approach leads to

$$\widetilde{\varepsilon}^{cd}|_K = \max\{0, |\boldsymbol{b}|\,h_K^{3/2} - \varepsilon\} \qquad \forall\, K \in \mathcal{T}_h\,. \tag{12}$$

The value $h_K^{3/2}$ was motivated by a careful analysis of the numerical crosswind spread in the discrete problem, i.e., of the maximal distance in which the right–hand side f significantly influences the discrete solution. The resulting method is linear but non–consistent and hence it is restricted to finite elements of first order of accuracy.

Codina [10] proposed to define $\widetilde{\varepsilon}^{cd}$, for any $K \in \mathcal{T}_h$, by

$$\widetilde{\varepsilon}^{cd}|_K = \frac{1}{2}\max\left\{0, C - \frac{2\,\varepsilon\,|\nabla u_h|}{\mathrm{diam}(K)\,|\boldsymbol{b}\cdot\nabla u_h|}\right\}\mathrm{diam}(K)\,\frac{|R(u_h)|}{|\nabla u_h|}\,, \tag{13}$$

where C is a suitable constant, and he recommended to set $C \approx 0.7$ for (bi)linear finite elements. To improve the properties of the resulting method for $f \neq 0$, John and Knobloch [14] replaced (13) by

$$\widetilde{\varepsilon}^{cd}|_K = \frac{1}{2}\max\left\{0, C - \frac{2\,\varepsilon\,|\nabla u_h|}{\mathrm{diam}(K)\,|R(u_h)|}\right\}\mathrm{diam}(K)\,\frac{|R(u_h)|}{|\nabla u_h|}\,. \tag{14}$$

If $f = 0$, it is equivalent to the original method (13). Finally, Knopp *et al.* [18] proposed to use (5) with $\widetilde{\varepsilon}^{cd}$ defined, for any $K \in \mathcal{T}_h$, by

$$\widetilde{\varepsilon}^{cd}|_K = \frac{1}{2}\max\left\{0, C - \frac{2\,\varepsilon}{Q_K(u_h)\,\mathrm{diam}(K)}\right\}\mathrm{diam}(K)\,Q_K(u_h)\,, \tag{15}$$

which leads to a method having properties convenient for theoretical investigations. The definition of $Q_K(u_h)$ is the same as in (10).

It is also possible to add both isotropic and crosswind artificial diffusion terms to the left–hand side of (2) as proposed by Codina and Soto [11]. They set

$$\widetilde{\varepsilon}^{iso} = \max\{0, \varepsilon_{\rm dc} - \tau(\boldsymbol{b})\,|\boldsymbol{b}|^2\}\,, \qquad \widetilde{\varepsilon}^{cd} = \varepsilon_{\rm dc} - \widetilde{\varepsilon}^{iso}\,,$$

where $\varepsilon_{\rm dc}$ is defined by the formula (14). However, for the test examples considered in the present paper, this method gave very similar results as (5) with $\widetilde{\varepsilon}^{cd}$ given by (14) and hence we shall not consider it in the following.

For triangulations consisting of weakly acute triangles, Burman and Ern [3] proposed to use (5) with $\widetilde{\varepsilon}^{cd}$ defined, on any $K \in \mathcal{T}_h$, by

$$\widetilde{\varepsilon}^{cd}|_K = \frac{\tau(\boldsymbol{b})\,|\boldsymbol{b}|^2\,|R(u_h)|}{|\boldsymbol{b}|\,|\nabla u_h| + |R(u_h)|}\,\frac{|\boldsymbol{b}|\,|\nabla u_h| + |R(u_h)| + \tan\alpha_K\,|\boldsymbol{b}|\,|D\,\nabla u_h|}{|R(u_h)| + \tan\alpha_K\,|\boldsymbol{b}|\,|D\,\nabla u_h|} \tag{16}$$

($\widetilde{\varepsilon}^{cd} = 0$ if one of the denominators vanishes). The parameter α_K is equal to $\pi/2 - \beta_K$ where β_K is the largest angle of K. If $\beta_K = \pi/2$, it is recommended to set $\alpha_K = \pi/6$. To improve the convergence of the nonlinear iterations, we replace $|R(u_h)|$ by $|R(u_h)|_{\rm reg}$ with $|x|_{\rm reg} \equiv x\,\tanh(x/2)$. Based on numerical experiments, John and Knobloch [14] simplified (16) to

$$\widetilde{\varepsilon}^{cd}|_K = \frac{\tau(\boldsymbol{b})\,|\boldsymbol{b}|^2\,|R(u_h)|}{|\boldsymbol{b}|\,|\nabla u_h| + |R(u_h)|}\,. \tag{17}$$

In this case, no regularization of the absolute values is applied.

Another stabilization strategy for linear simplicial finite elements was introduced by Burman and Hansbo [5]. The term to be added to the left–hand side of (2) is defined by

$$\sum_{K\in\mathcal{T}_h} \int_{\partial K} \Psi_K(u_h)\,\mathrm{sign}(\frac{\partial u_h}{\partial \boldsymbol{t}_{\partial K}})\,\frac{\partial v_h}{\partial \boldsymbol{t}_{\partial K}}\,\mathrm{d}\sigma\,, \tag{18}$$

where $\boldsymbol{t}_{\partial K}$ is a tangent vector to the boundary ∂K of K,

$$\Psi_K(u_h) = \mathrm{diam}(K)\,(C_1\,\varepsilon + C_2\,\mathrm{diam}(K))\,\max_{E\subset\partial K}|\,[|\boldsymbol{n}_E\cdot\nabla u_h|]_E\,|\,, \tag{19}$$

$\boldsymbol{n}_E$ are normal vectors to edges E of K, $[|v|]_E$ denotes the jump of a function v across the edge E and C_1, C_2 are appropriate constants. Further, Burman and Ern [4] proposed to use (18) with $\Psi_K(u_h)$ defined by

$$\Psi_K(u_h)|_E = C\,|\boldsymbol{b}|\,[\mathrm{diam}(K)]^2\,|\,[|\nabla u_h|]_E\,| \qquad \forall\ E\subset\partial K \tag{20}$$

or by

$$\Psi_K(u_h) = C\,|R(u_h)|\,, \tag{21}$$

where C is a suitable constant.

Finally, let us mention the improved Mizukami–Hughes method, originally introduced by Mizukami and Hughes [19] and recently improved by Knobloch [17]. It is a method of another type than the methods presented in this section since its derivation does not start from the SUPG discretization. However, like the SUPG method, it is a Petrov–Galerkin method. The weighting functions generally depend on the unknown discrete solution and hence the method is nonlinear. The advantage of the Mizukami–Hughes method is that the discrete solution always satisfies the discrete maximum principle and is usually rather accurate. Drawbacks of the method are that it is defined for conforming linear triangular finite elements only and that it is not clear how to generalize the Mizukami–Hughes method to more complicated convection–diffusion problems than presented in this paper.

4. Numerical results

In this section, we shall present numerical results obtained using the methods from Sections 2 and 3 for the following three test problems:

Example 1. Solution with parabolic and exponential boundary layers. We consider the convection–diffusion equation (1) in $\Omega = (0,1)^2$ with $\varepsilon = 10^{-8}$, $\boldsymbol{b} = (1,0)^T$, $f = 1$ and $u_b = 0$. The solution $u(x,y)$ of this problem, see Figure 1, possesses an exponential boundary layer at $x = 1$ and parabolic boundary layers at $y = 0$ and $y = 1$. In the interior grid points, the solution $u(x,y)$ is very close to x. This example has been used, e.g., in [19].

Example 2. Solution with interior layer and exponential boundary layers. We consider the convection–diffusion equation (1) in $\Omega = (0,1)^2$ with $\varepsilon = 10^{-8}$, $\boldsymbol{b} = (\cos(-\pi/3), \sin(-\pi/3))^T$, $f = 0$ and

$$u_b(x,y) = \begin{cases} 0 & \text{for } x = 1 \text{ or } y \leq 0.7, \\ 1 & \text{else.} \end{cases}$$

The solution, see Figure 3, possesses an interior layer in the direction of the convection starting at $(0, 0.7)$. On the boundary $x = 1$ and on the right part of the boundary $y = 0$, exponential layers are developed. This example has been used, e.g., in [13].

Example 3. Solution with two interior layers. We consider the convection–diffusion equation (1) in $\Omega = (0,1)^2$ with $\varepsilon = 10^{-8}$, $\boldsymbol{b} = (1,0)^T$, $u_b = 0$ and

$$f(x,y) = \begin{cases} 16\,(1 - 2\,x) & \text{for } (x,y) \in [0.25, 0.75]^2, \\ 0 & \text{else.} \end{cases}$$

The solution, see Figure 5, possesses two interior layers layer at $(0.25, 0.75) \times \{0.25\}$ and $(0.25, 0.75) \times \{0.75\}$. In $(0.25, 0.75)^2$, the solution $u(x,y)$ is very close to the quadratic function $(4\,x - 1)(3 - 4\,x)$. This example has not been published before.

All the numerical results discussed in this section were computed on uniform triangulations $\mathcal{T}_h$ of Ω of the type depicted in Fig. 7, which consist of $2(N \times N)$ equal right–angled isosceles triangles ($N = 5$ in Fig. 7). We used either $N = 20$ or $N = 64$. Figs. 2, 4 and 6 show the SUPG solutions for Examples 1, 2 and 3, respectively. Although the formula (3) for the stabilization parameter τ can be regarded as optimal in all three cases, we observe significant spurious oscillations in layer regions.

Denoting

$$\Omega_1 = \{(x,y) \in \Omega\,;\ x \le 0.5,\, y \ge 0.1\}\,, \qquad \Omega_2 = \{(x,y) \in \Omega\,;\ x \ge 0.7\}\,,$$

we introduce the following measures of oscillations in the discrete solutions u_h of Examples 1 and 2:

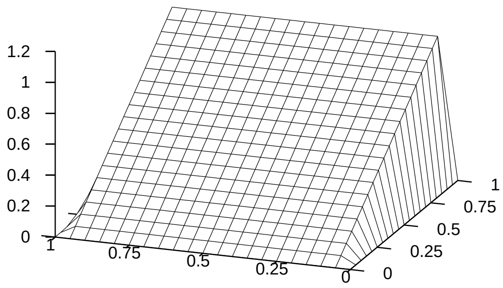


Fig. 1: *Example 1, solution u.*

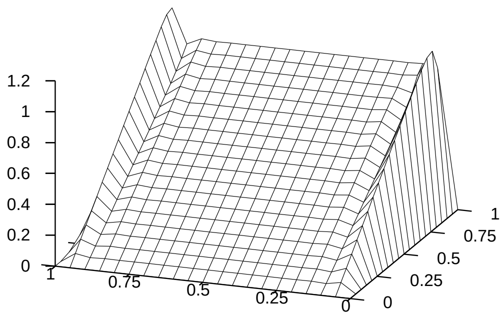


Fig. 2: *Example 1, SUPG, $N = 20$.*

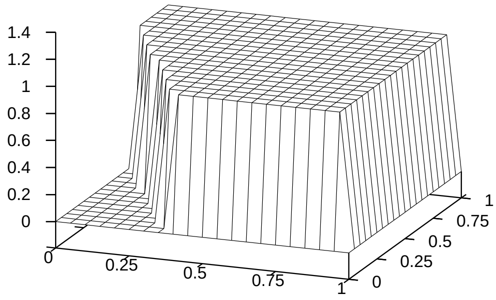


Fig. 3: *Example 2, solution u.*

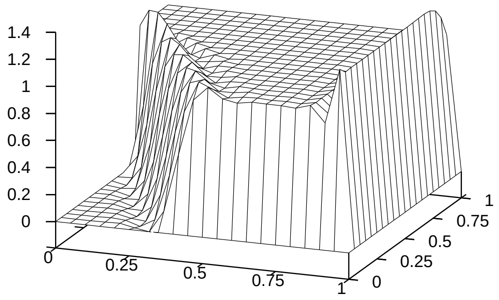


Fig. 4: *Example 2, SUPG, $N = 20$.*

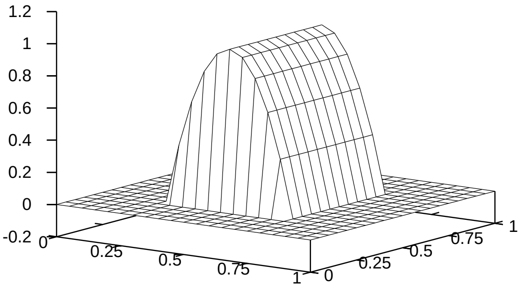


Fig. 5: *Example 3, solution u.*

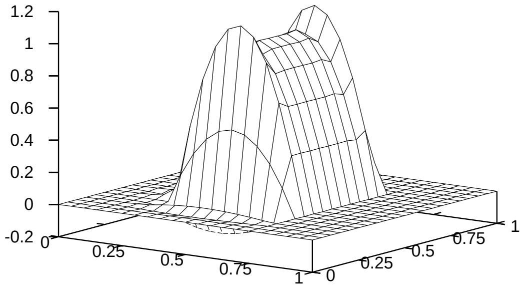


Fig. 6: *Example 3, SUPG, $N = 20$.*

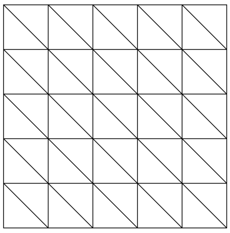


Fig. 7: *Considered type of triangulations.*

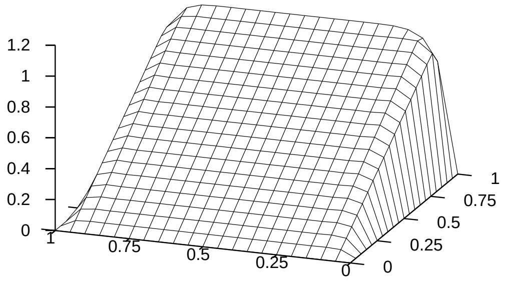


Fig. 8: *Example 1, Cmod,C* $=0.6$*,N* $=20$.

$$
\begin{aligned}
osc_{\mathrm{para}} &:= 10 \max_{y\in[0,1]} \{u_h(0.5,y) - u_h(0.5,0.5)\}\,,\\
osc_{\mathrm{int}} &:= \left(\sum_{(x,y)\in\Omega_1} (\min\{0,u_h(x,y)\})^2 + (\max\{0,u_h(x,y)-1\})^2 \right)^{1/2}\,,\\
osc_{\mathrm{exp}} &:= \left(\sum_{(x,y)\in\Omega_2} (\max\{0,u_h(x,y)-1\})^2 \right)^{1/2}\,.
\end{aligned}
$$

The measure osc_{para} characterizes the oscillations of u_h in the parabolic boundary layer regions of Example 1 whereas osc_{int} and osc_{exp} measure the oscillations of u_h in the interior and exponential layer regions of Example 2. The summations are performed over the nodes (x, y) of the mesh. Fig. 9 shows the values of these measures for most of the methods discussed in the previous section and we see that there are significant differences between the size of the oscillations. For the six best methods the results are also shown in Fig. 10 which suggests that the best methods are the improved Mizukami–Hughes method and the crosswind artificial diffusion method defined by (5) with (12).

However, a suppression of oscillations does not imply that the respective discrete solution u_h is a good approximation of u since the layers can be smeared considerably. Therefore, we also define the following measures:

$$
smear_{\mathrm{para}} := \max_{y\in[1/N,1-1/N]} \{u_h(0.5,0.5) - u_h(0.5,y)\}\,, \qquad smear_{\mathrm{int}} := x_2 - x_1\,,
$$

$$
smear_{\mathrm{exp}} := \frac{1}{10} \left(\sum_{(x,y)\in\Omega_2} (\min\{0,u_h(x,y)-1\})^2 \right)^{1/2}\,.
$$

The measure $smear_{\mathrm{para}}$ characterizes the smearing of the parabolic boundary layer in Example 1 whereas $smear_{\mathrm{int}}$ and $smear_{\mathrm{exp}}$ measure the smearing of the interior and exponential layers in Example 2. In the definition of $smear_{\mathrm{int}}$, the value x_1 is the x–coordinate of the first point on the cut line $(x, 0.25)$ with $u_h(x_1, 0.25) \geq 0.1$ and x_2

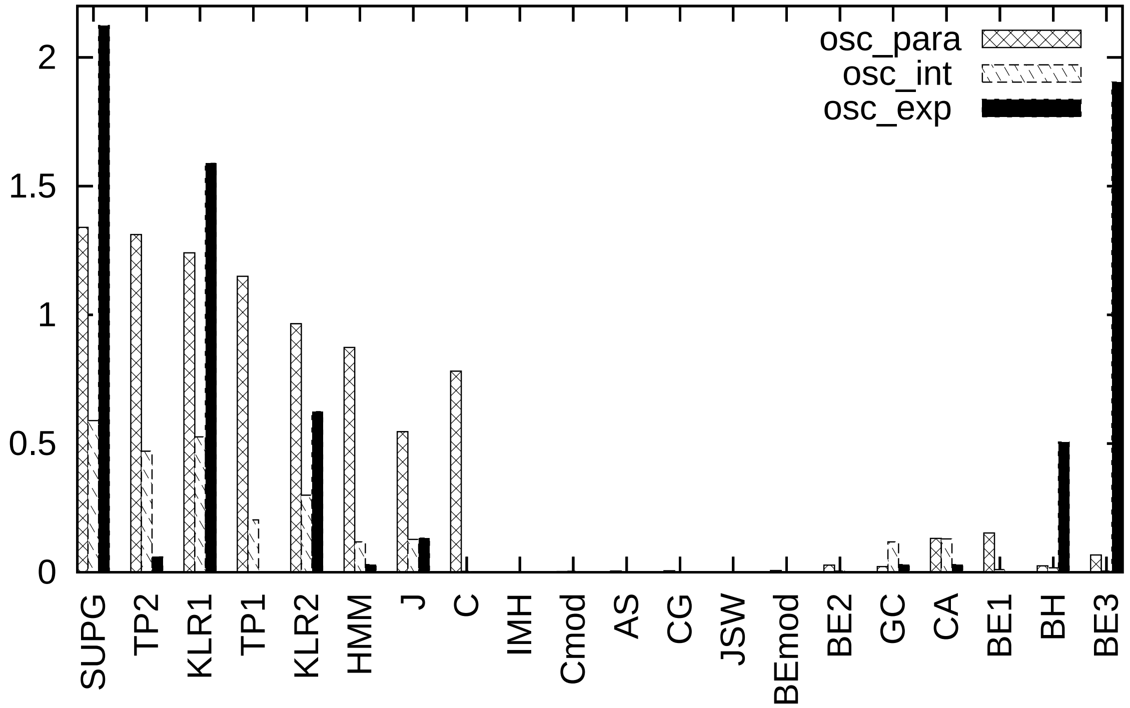


Fig. 9: *Measures of oscillations in discrete solutions of Examples 1 and 2 for* $N = 64$ *and various discretizations. Methods adding isotropic artificial diffusion (4): TP1, TP2 - [21], KLR1 - (10), (7) with* $S_K = 1$*, HMM - [13], J - (11) with* $\alpha = 0.3$ *and* $\nu = 2$*, AS - (6), (9), CG - (6), (8), GC - (6), (7), CA - [6]. Methods adding crosswind artificial diffusion (5): KLR2 - (15) with* $C = 0.6$ *and* $S_K = 1$*, C - (13) with* $C = 0.6$*, Cmod - (14) with* $C = 0.6$*, JSW - (12), BEmod - (17), BE1 - (16). Edge stabilizations (18): BE2 - (21) with* $C = 5 \cdot 10^{-5}$*, BH - (19) with* $C_1 = 0.5$ *and* $C_2 = 0.01$*, BE3 - (20) with* $C = 0.05$*. IMH - improved Mizukami–Hughes method [17].*

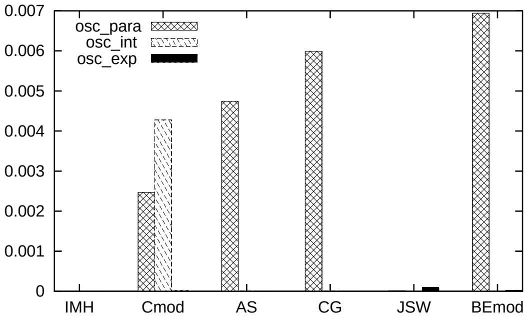


Fig. 10: *Measures of oscillations in discrete solutions of Examples 1 and 2 for IMH, Cmod, AS, CG, JSW and BEmod.*

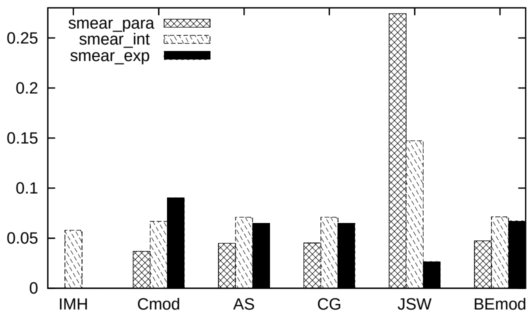


Fig. 11: *Measures of smearing in discrete solutions of Examples 1 and 2 for IMH, Cmod, AS, CG, JSW and BEmod.*

is the x–coordinate of the first point with $u_h(x_2, 0.25) \geq 0.9$. The summation is again performed over the nodes (x, y) of the mesh. The results in Fig. 11 show that the method JSW leads to a considerable smearing of the layers and that the improved Mizukami–Hughes methods does not smear boundary layers for Examples 1 and 2. The remaining four methods (Cmod, AS, CG and BEmod) seem to be comparable.

The above results and many other numerical tests we performed indicate that the best methods are the improved Mizukami–Hughes method [17], the isotropic artificial diffusion methods by do Carmo and Galeão [8] given by (4), (6), (8) and by Almeida and Silva [1] given by (4), (6), (9) and the modified crosswind artificial diffusion methods by Codina [10] given by (5), (14) and by Burman and Ern [3] given by (5), (17). For Example 1, the improved Mizukami–Hughes method gives a nodally exact discrete solution and the remaining four methods give comparable discrete solutions, one of which is depicted in Fig. 8. For Example 2, the discrete solutions obtained using the improved Mizukami–Hughes method and the method by do Carmo and Galeão [8] (denoted by CG) are depicted in Figs. 12 and 13, respectively. For the methods Cmod, AS and BEmod, the discrete solutions are similar as in Fig. 13. Thus, we see that the methods IMH, Cmod, AS, CG and BEmod are able to substantially improve the quality of the discrete solution in comparison to the SUPG method.

Unfortunately, this is not always the case. We observed that often also the methods Cmod, AS, CG and BEmod may produce results with spurious oscillations. This may also happen for Example 2 if a triangulation similar as in Fig. 7 but with different numbers of vertices in x– and y–directions is used. But also for a triangulation of the type from Fig. 7, the methods Cmod, AS, CG and BEmod may give a wrong solution. This is the case for Example 3 as Figs. 14 and 15 show. We see that the oscillations along the interior layers disappeared but the discrete solution is not correct in a region where it should vanish. We observed this phenomenon for all the SUPG based methods discussed in Section 3. Fig. 16 shows an approximation of u obtained using the improved Mizukami–Hughes method which is much better than for the other four methods, however not perfect. Moreover, in contrast with these methods, the IMH solution improves if the mesh is refined.

Let us demonstrate that the phenomenon shown in Figs. 14 and 15 has to be

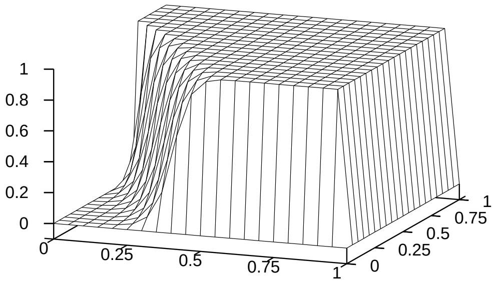


Fig. 12: *Example 2, IMH,* $N = 20$.

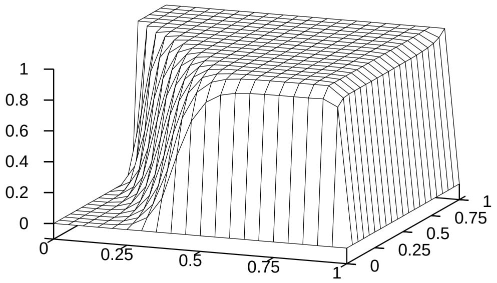


Fig. 13: *Example 2, CG,* $N = 20$.

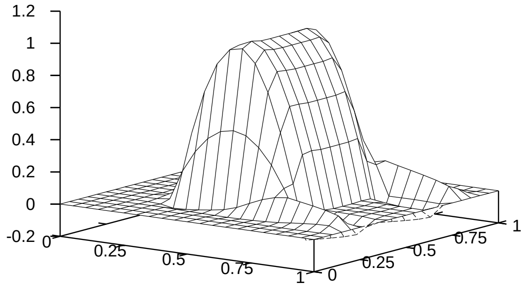


Fig. 14: *Example 3, CG, $N = 20$.*

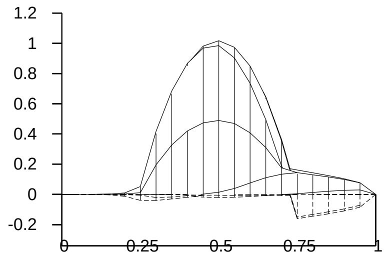


Fig. 15: *As in Fig. 14 (other view).*

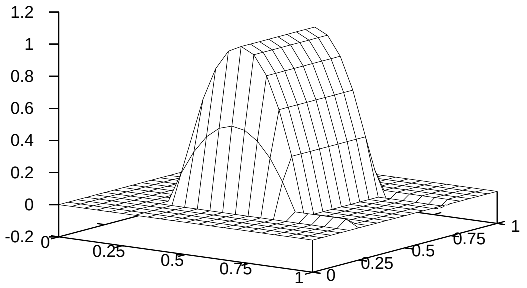


Fig. 16: *Example 3, IMH, $N = 20$.*

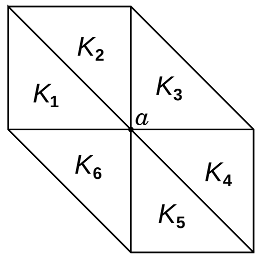


Fig. 17: *Support of a basis function.*

expected if the discrete solution should suppress the spurious oscillations present in the SUPG solution. Thus, let us assume that a solution of the discrete problem obtained by adding the term (4) or (5) to the left–hand side of (2) does not contain spurious oscillations and does not smear the inner layers significantly. Then, on any vertical mesh line intersecting the interval $(0.5, 0.75)$ on the x–axis, we may find a vertex $a \in (0.5, 0.75) \times (0, 0.25)$ surrounded by elements $K_1, \dots, K_6$ as depicted in Fig. 17 such that $\nabla u_h \approx \mathbf{0}$ on $K_4 \cup K_5 \cup K_6$ but $\partial u_h/\partial y$ is positive and nonnegligible on K_2. The elements $K_1, \dots, K_6$ make up the support of the standard piecewise linear basis function equal 1 at a. Using this basis function as v_h in the discrete problem and denoting by $\widetilde{\varepsilon}$ the artificial diffusion in (4) or (5), it is easy to show that

$$\frac{\partial u_h}{\partial x}\Big|_{K_2} \approx \frac{3}{h}\,\frac{\partial u_h}{\partial y}\Big|_{K_2} (2\,\varepsilon + \widetilde{\varepsilon}|_{K_2} + \widetilde{\varepsilon}|_{K_3})\,,$$

where h denotes the length of a leg of K_2. This means that, on some elements in the region $(0.5, 0.75) \times (0, 0.25)$, the discrete solution has to grow in the x–direction, which is exactly what is observed in Figs. 14 and 15. A deeper explanation of this phenomenon will be a subject of our future research.

5. Conclusions

In this paper we presented a review and a computational comparison of various stabilization techniques based on the SUPG method which have been developed to diminish spurious oscillations in finite element solutions of scalar stationary convection–diffusion equations. We identified the best methods and demonstrated that they are able to substantially enhance the quality of the discrete solutions in comparison to the SUPG method. However, we have also shown that these methods can fail for very simple test problems. An alternative to these methods seems to be the improved Mizukami–Hughes method which gives good results in all the cases considered.

References

[1] R.C. Almeida, R.S. Silva: *A stable Petrov–Galerkin method for convection–dominated problems.* Comput. Methods Appl. Mech. Engrg. **140**, 1997, 291–304.

[2] A.N. Brooks, T.J.R. Hughes: *Streamline upwind/Petrov–Galerkin formulations for convection dominated flows with particular emphasis on the incompressible Navier–Stokes equations.* Comput. Methods Appl. Mech. Engrg. **32**, 1982, 199–259.

[3] E. Burman, A. Ern: *Nonlinear diffusion and discrete maximum principle for stabilized Galerkin approximations of the convection–diffusion–reaction equation.* Comput. Methods Appl. Mech. Engrg. **191**, 2002, 3833–3855.

[4] E. Burman, A. Ern: *Stabilized Galerkin approximation of convection–diffusion–reaction equations: Discrete maximum principle and convergence.* Math. Comput. **74**, 2005, 1637–1652.

[5] E. Burman, P. Hansbo: *Edge stabilization for Galerkin approximations of convection–diffusion–reaction problems.* Comput. Methods Appl. Mech. Engrg. **193**, 2004, 1437–1453.

[6] E.G.D. do Carmo, G.B. Alvarez: *A new stabilized finite element formulation for scalar convection–diffusion problems: The streamline and approximate upwind/Petrov–Galerkin method.* Comput. Methods Appl. Mech. Engrg. **192**, 2003, 3379–3396.

[7] E.G.D. do Carmo, G.B. Alvarez: *A new upwind function in stabilized finite element formulations, using linear and quadratic elements for scalar convection–diffusion problems.* Comput. Methods Appl. Mech. Engrg. **193**, 2004, 2383–2402.

[8] E.G.D. do Carmo, A.C. Galeão: *Feedback Petrov–Galerkin methods for convection–dominated problems.* Comput. Methods Appl. Mech. Engrg. **88**, 1991, 1–16.

[9] I. Christie, D.F. Griffiths, A.R. Mitchell, O.C. Zienkiewicz: *Finite element methods for second order differential equations with significant first derivatives.* Int. J. Numer. Methods Eng. **10**, 1976, 1389–1396.

[10] R. Codina: *A discontinuity–capturing crosswind–dissipation for the finite element solution of the convection–diffusion equation.* Comput. Methods Appl. Mech. Engrg. **110**, 1993, 325–342.

[11] R. Codina, O. Soto: *Finite element implementation of two–equation and algebraic stress turbulence models for steady incompressible flows.* Int. J. Numer. Meth. Fluids **30**, 1999, 309–333.

[12] A.C. Galeão, E.G.D. do Carmo: *A consistent approximate upwind Petrov–Galerkin method for convection–dominated problems.* Comput. Methods Appl. Mech. Engrg. **68**, 1988, 83–95.

[13] T.J.R. Hughes, M. Mallet, A. Mizukami: *A new finite element formulation for computational fluid dynamics: II. Beyond SUPG.* Comput. Methods Appl. Mech. Engrg. **54**, 1986, 341–355.

[14] V. John, P. Knobloch: *A comparison of spurious oscillations at layers diminishing (SOLD) methods for convection–diffusion equations: Part I.* Preprint Nr. 156, FR 6.1 – Mathematik, Universität des Saarlandes, Saarbrücken, 2005.

[15] C. Johnson: *Adaptive finite element methods for diffusion and convection problems.* Comput. Methods Appl. Mech. Engrg. **82**, 1990, 301–322.

[16] C. Johnson, A.H. Schatz, L.B. Wahlbin: *Crosswind smear and pointwise errors in streamline diffusion finite element methods.* Math. Comput. **49**, 1987, 25–38.

[17] P. Knobloch: *Improvements of the Mizukami–Hughes method for convection–diffusion equations.* Comput. Methods Appl. Mech. Engrg. **196**, 2006, 579–594.

[18] T. Knopp, G. Lube, G. Rapin: *Stabilized finite element methods with shock capturing for advection–diffusion problems.* Comput. Methods Appl. Mech. Engrg. **191**, 2002, 2997–3013.

[19] A. Mizukami, T.J.R. Hughes: *A Petrov–Galerkin finite element method for convection–dominated flows: An accurate upwinding technique for satisfying the maximum principle.* Comput. Methods Appl. Mech. Engrg. **50**, 1985, 181–193.

[20] H.–G. Roos, M. Stynes, L. Tobiska: *Numerical methods for singularly perturbed differential equations. Convection–Diffusion and Flow problems.* Springer–Verlag, Berlin, 1996.

[21] T.E. Tezduyar, Y.J. Park: *Discontinuity–capturing finite element formulations for nonlinear convection–diffusion–reaction equations.* Comput. Methods Appl. Mech. Engrg. **59**, 1986, 307–325.

NUMERICAL SOLUTION OF NEWTONIAN FLOW IN BYPASS AND NON-NEWTONIAN FLOW IN BRANCHING CHANNELS*

R. Keslerová, K. Kozel, V. Prokop

Abstract

This paper deals with the numerical solution of Newtonian and non-Newtonian flows. The flows are supposed to be laminar, viscous, incompressible and steady. The model used for non-Newtonian fluids is a variant of the power-law. Governing equations in this model are incompressible Navier-Stokes equations. For numerical solution we could use artificial compressibility method with three stage Runge-Kutta method and finite volume method in cell centered formulation for discretization of space derivatives. The following cases of flows are solved: flow through a bypass connected to main channel in 2D and 3D and non-Newtonian flow through branching channels in 2D. Some 2D and 3D results that could have an application in the area of biomedicine are presented.

1. Mathematical model

The motivation for numerical solution of the fluid flow of Newtonian and non-Newtonian fluids arises in many applications, e.g. in the biomedicine, food industry, chemistry, glaciology etc. Many common fluids are non-Newtonian: paints, solutions of various polymers, food products. The main points of non-Newtonian behaviour are the ability of the fluid to shear thin or shear thicken in shear flows, the presence of non-zero normal stress differences in shear flows, the ability of the fluid to yield stress, the ability of the fluid to exhibit relaxation, the ability of the fluid to creep, see [1]. The solution of flows in branching channels and channels with bypass is important for modelling of blood flow in arteries. The study of blood flow in large and medium arteries is a very complex task because of the heterogeneous nature of the problem and the extreme complexity of blood and arterial wall dynamics. Although blood is actually a non-Newtonian suspension of cells in plasma, it is reasonable to model it as a Newtonian fluid in vessels greater than approximately 0.5 mm in diameter [2]. The occurring shear rates are in a range where non-Newtonian effects are only in minor significance to the flow parameters. This type of flow could be described by conservation laws of mass and momentum (Navier-Stokes equations), where the influence of exterior forces and heat exchange is not taken into account. In this case the model of a vessel is a tube with rigid walls. The pulsatile character of blood flow is not considered as well as the elasticity of arterial walls.

[*] This work was sponsored by Research Plan MSM 6840770010 and GA AS CR No. A 100190505.

First, we consider the Newtonian fluids. The system of 2D Navier-Stokes equations for Newtonian fluids in dimensionless conservative form has the form:

$$\tilde{R}W_t + F_x + G_y = \frac{\tilde{R}}{\mathrm{Re}}\Delta W, \qquad \tilde{R} = \mathrm{diag}\|0,1,1\|. \tag{1}$$

where the Reynolds number defined as $\mathrm{Re} = dw^*/\nu$ in 2D and $\mathrm{Re} = d_h w^*/\nu$ in 3D is an important parameter of the flow. Quantity w^* is a characteristic velocity (the speed of upstream flows), $\nu = \eta/\rho$ is the kinematic viscosity, d is a length scale (the width of the channel), $d_h = 4S/O$ is the hydraulic diameter, S is the area section of the duct and O is the wetted perimeter. In equation (1), $W = (p, u, v)^T$ is the vector of solution, $\tilde{R} = \mathrm{diag}\|0,1,1\|$, and $F = (u, u^2 + p, uv)^T, G = (v, uv, v^2 + p)^T$ denote inviscid fluxes, (u, v) is the dimensionless velocity vector ($u = u^*/q_\infty$, $v = v^*/q_\infty$), p denotes the dimensionless pressure ($p = p^*/\rho q_\infty^2$), t is the dimensionless time ($t = t^* q_\infty/l$), and q_∞ is defined as a velocity of incoming flow ($q_\infty = u^*$).

In the case of non-Newtonian fluids the power-law fluids are considered. The dominant difference from the Newtonian behaviour is shear thinning or shear thickening. From variety of power-law fluids we choose the simplest one:

$$\tau(\mathbf{e}) = 2\nu_0 |\mathbf{e}|^r \mathbf{e}, \tag{2}$$

where τ is the stress tensor, $\mathbf{e} = (e_{ij}),\ i,j = 1,2$, is the strain tensor with components $e_{11} = u_x,\ e_{12} = e_{21} = (v_x + u_y)/2,\ e_{22} = v_y,\ |\mathbf{e}|$ denotes the Euclidean norm of the tensor, ν_0 is a positive constant related to the limit of generalized viscosity $\mu_g(\kappa)$ when $\kappa \to 0$, r is a constant of the model. The model captures the shear thinning fluid if $r \in (-1, 0)$, shear thickening fluid if $r > 0$, and $r = 0$ corresponds to the Newtonian fluid. For the non-Newtonian fluids the system of 2D Navier-Stokes equations and the continuity equation in two dimensional case written in the dimensionless conservative form reads

$$\tilde{R}W_t + F_x + G_y = \frac{\tilde{R}}{\mathrm{Re}}(R_x + S_y) \tag{3}$$

where $R = (0, g_{11}, g_{21})^T,\ S = (0, g_{12}, g_{22})^T,\ g_{ij} = 2|\mathbf{e}|^r e_{ij},\ i,j = 1,2$, with components of e_{ij} defined above. The terms on the right-hand side can be expanded as follows

$$\begin{aligned} (g_{11})_x + (g_{12})_y &= 2|\mathbf{e}|^r_x u_x + |\mathbf{e}|^r_y (u_y + v_x) + |\mathbf{e}|^r \Delta u, \\ (g_{21})_x + (g_{22})_y &= |\mathbf{e}|^r_x (u_y + v_x) + 2|\mathbf{e}|^r_y v_y + |\mathbf{e}|^r \Delta v. \end{aligned} \tag{4}$$

Let us stress that subindices $_x$ and $_y$ denote partial derivatives with respect to x and y and that Δ stands for the 2D Laplacian. At the inlet the Dirichlet boundary condition for velocity vector $(u, v)^T$ is prescribed, at the outlet the pressure value is given. On the wall the zero Dirichlet boundary conditions for the components of velocity are used.

2. Numerical model

For further solution of the system of equations (1), the artificial compressibility method is used. The continuity equation is completed with the term p_t/a^2, where

$a^2 > 0$. The pressure satisfies the artificial equation of state: $p = \rho/\delta$, in which ρ is the artificial density, δ is the artificial compressibility, that is connected to the artificial speed of sound by relation $a = \delta^{-\frac{1}{2}}$, see [3]. Then system of governing equations has the form

$$W_t + F_x + G_y = \frac{\tilde{R}}{\mathrm{Re}} \left(R_x + S_y \right), \tag{5}$$

where $W = (p/a^2, u, v)^T$. System of equations (5) is solved by a three stage Runge-Kutta method with given steady boundary conditions. At the inlet an extrapolation of the pressure is used. At the outlet the value of the pressure is prescribed by $p = p_2$, where p_2 is the dimensionless value of the pressure, that is higher then the initial value of the pressure at the inlet to ensure pressure gradient. On the walls there are non-permeability and no-slip conditions. The multistage Runge-Kutta method is stabilized by the artificial viscosity term (Jameson's type, see [4]):

$$\begin{aligned} W_{i,j}^{n} &= W_{i,j}^{(0)} \\ W_{i,j}^{(r)} &= W_{i,j}^{(0)} - \alpha_r \Delta t \overline{R} W_{i,j}^{(r-1)}, \quad r = 1, \ldots, m, \\ W_{i,j}^{n+1} &= W_{i,j}^{(m)}, \quad m = 3, \end{aligned}$$

where W_{ij}^n denotes an approximation of W at grid point (x_i, x_j) and at a time $t = t_n$, $\Delta t = t_n - t_{n-1}$ is the time step, and

$$\overline{R} W_{i,j}^{(r-1)} = \tilde{R} W_{i,j}^{(r-1)} - D W_{i,j}^{n}.$$

The coefficients are $\alpha_1 = 0.5, \alpha_2 = 0.5, \alpha_3 = 1.0$ and the term DW_{ij}^n is described below. The numerical method is of the second order in time and space. The form of residual $\tilde{R}W_{i,j}^n$ depends on the method used for the space discretization, which is in this case the finite volume method in the cell centered formulation:

$$\tilde{R} W_{i,j} = \frac{1}{\mu_{ij}} \sum_{k=1}^{4} \big[\big(F_k^i - \frac{1}{Re} F_k^v \big) \Delta y_k - \big(G_k^i - \frac{1}{Re} G_k^v \big) \Delta x_k \big], \tag{6}$$

where $F^i = F, G^i = G$ are inviscid fluxes and $F^v = (0, u_x, v_x)^T, G^v = (0, u_y, v_y)^T$ are viscous fluxes, the index k corresponds to the side of a finite volume. The artificial viscosity term $DW_{i,j}^n$ depends in this case on the second derivatives of the pressure and is used to improve stability of the solution. The dissipative artificial viscosity term is constructed as follows:

$$
\begin{aligned}
DW &= D_xW + D_yW,\\
D_xW &= d_{i+\frac{1}{2},j} - d_{i-\frac{1}{2},j},\\
D_yW &= d_{i,j+\frac{1}{2}} - d_{i,j-\frac{1}{2}},\\
d_{i+\frac{1}{2},j} &= \frac{h_{i+\frac{1}{2},j}}{\Delta t}\epsilon^{(2)}_{i+\frac{1}{2},j}(W_{i+1,j} - W_{i,j}),\\
\nu_{i,j} &= \frac{|p_{i+1,j} - 2p_{i,j} + p_{i-1,j}|}{|p_{i+1,j}| + 2p_{i,j} + |p_{i-1,j}|},\\
\epsilon^{(2)}_{i+\frac{1}{2},j} &= \kappa^{(2)}\max(\nu_{i+1,j}, \nu_{i,j}),
\end{aligned}
$$

where $\kappa^{(2)}$ has to be chosen in order to achieve convergence of the method.

3. Numerical results

In this section we present steady numerical results obtained for the steady flow with the aid of the above methods. First, numerical results for channels with one entrance and two exit parts are presented. Figures 2 and 4 show the fluid velocity distribution for Reynolds number 1500 for non-Newtonian fluid. In Figures 3 and 5 the numerical results for the two-dimensional case (Newtonian fluids) and the convergence of the residuals of the vector $W = (p, u, v)^T$ are shown. The symbol q stands for the velocity magnitude, i.e. $q = \sqrt{u^2 + v^2}$. The other figures represent 3D flow for $\mathrm{Re} = 500$.

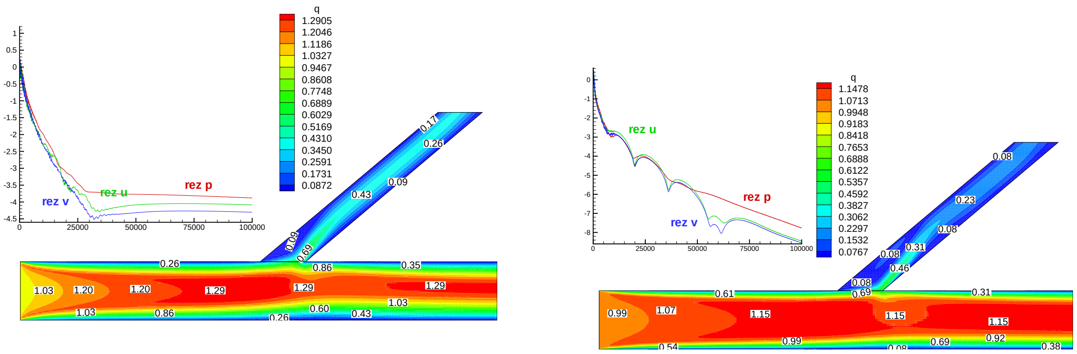


Fig. 1, 2: *Velocity magnitude distribution, non-Newtonian and Newtonian fluids,* $\mathrm{Re} = 1500$.

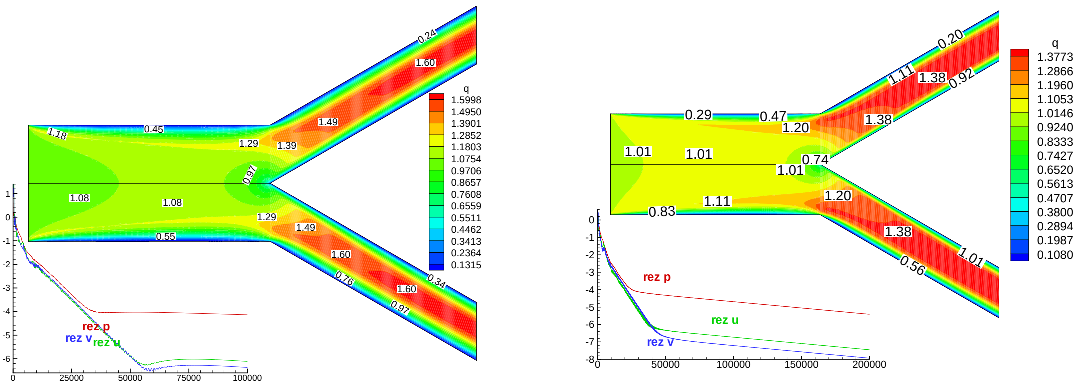


Fig. 3, 4: *Velocity magnitude distribution, non-Newtonian and Newtonian fluids,* $\mathrm{Re} = 1500$.

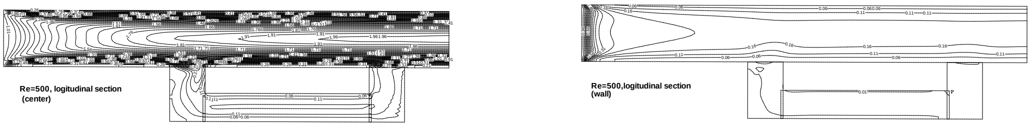


Fig. 5, 6: *Isolines of velocity in angular bypass for* $\mathrm{Re} = 500$, *3D case in the central plane xy, 3D case in the xy plane near the wall.*

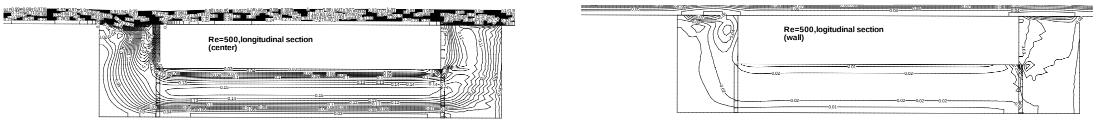


Fig. 7, 8: *Isolines of velocity in angular bypass for* $\mathrm{Re} = 500$, *3D case in the central plane xy, 3D case in the xy plane near the wall, details of regions.*

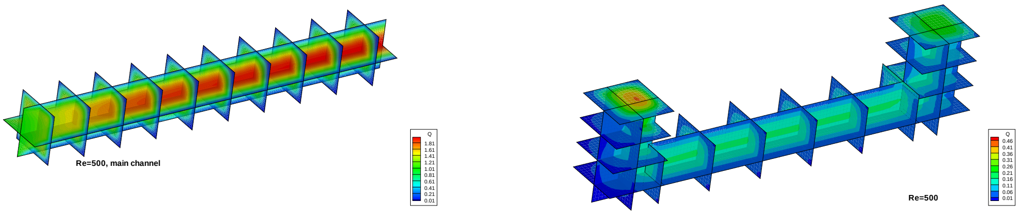


Fig. 9, 10: *The figure shows behaviour of flow in angular bypass for* $\mathrm{Re} = 500$ *in the form of isolines of velocity, x-y-z cross section of the main channel and bypass.*

References

[1] K.R. Rajagopal: *Mechanics of non-Newtonian fluids.* In: G.P. Galdi, J. Nečas (eds.), Recent Developments in Theoretical Fluid Dynamics, Pitaman Research Notes in Math. **291**, Longman & Technical, Essex, 1993, 129–162.

[2] T. Taylor, T. Yamaguchi: *Three-dimensional simulation of blood flow in an abdominal aortic aneurysm, steady and unsteady flow cases.* J. Biomech. Engrg. **116**, 1994, 89–97.

[3] A.J. Chorin: *A numerical nethod for solving incompressible viscous flow problems.* Journal of Computational Physics **135**, 1997, 118–125.

[4] A. Jameson, W. Schmidt, E. Turkel: *Numerical solution of the Euler equations by finite volume methods using Runge-Kutta time-stepping schemes.* AIAA 14th Fluid and Plasma Dynamic Conference, June 23–25, 1981, Palo Alto, California, 1981, 1981–1259.

[5] R.J. LeVeque: *Numerical methods for conservation laws.* Birkhäuser Verlag, Basel, Switzerland, 1990.

[6] R. Dvořák, K. Kozel: *Mathematical modeling in aerodynamics.* CTU Prague, 1996.

[7] R. Keslerová, K. Kozel: *Numerical solution of 2D and 3D incompressible laminar flows through a branching channel.* In: Proc. Topical Problems of Fluid Mechanics, Prague, 2005, 55–58.

THE USE OF BASIC ITERATIVE METHODS FOR BOUNDING A SOLUTION OF A SYSTEM OF LINEAR EQUATIONS WITH AN M-MATRIX AND POSITIVE RIGHT-HAND SIDE

Martin Kocurek

Abstract

This article presents a simple method for bounding a solution of a system of linear equations $Ax = b$ with an M-matrix and positive right-hand side [1]. Given a suitable approximation to an exact solution, the bounds are constructed by one step in a basic iterative method.

1. Motivation

When we use iterative methods for solving sets of linear algebraic equations $Ax = b$, we guess an accuracy of the computed solution according to a residual vector. Unfortunately, small norm of the residual vector doesn't imply that we are close to the exact solution. If we could instead construct an upper and lower bound, we could guess an accuracy of the computed solution better.

2. Basic terms and definitions

Definition 2.1 *Let matrices A, B have the same dimension. We say that $A \geq B$ if $a_{ij} \geq b_{ij}$ holds for every i, j. Matrix A is called **nonnegative**, if $A \geq O$, where O is the zero matrix.*

Definition 2.2 *A real square matrix $A = (a_{ij})_{i,j=1}^{n}$ is called **M-matrix**, if*

1. *$a_{ii} > 0$, $i = 1, \dots, n$,*
2. *$a_{ij} \leq 0$ for $i \neq j$, $i, j = 1, \dots, n$,*
3. *exists $A^{-1} \geq 0$.*

Definition 2.3 *Let us split matrix A into two matrices V, W, so that $A = V - W$. If matrix V is nonsingular, then $V - W$ is called **splitting** of matrix A. The splitting of matrix A is called **regular** if V is nonsingular with $V^{-1} \geq 0$ and $W \geq 0$.*

A splitting $Ax = (V - W)x = b$ yields an iterative method

$$x^{(k+1)} = V^{-1}Wx^{(k)} + V^{-1}b,$$

which is convergent if and only if the spectral radius satisfies $\rho(V^{-1}W) < 1$.

As usual, we split matrix A into $D - L - U$, where D is the diagonal of A and L, U are strictly lower and upper triangular parts of A, respectively. The classical iterative methods are obtained by setting

- $V = I$, $W = I - A$ ... Fixed-point iterations
- $V = D$, $W = L + U$ ... Method of Jacobi
- $V = D - L$, $W = U$ ... Method of Gauss-Seidel

From now on we consider matrix A to be an M-matrix and the right-hand side to be positive. The three methods mentioned above can be written as

$$x^{(k+1)} = \mathbf{T}x^{(k)} + \mathbf{d}, \qquad \mathbf{T} := V^{-1}W, \qquad \mathbf{d} := V^{-1}b.$$

Furthermore, for all these methods (for fixed-point iterations $a_{ii} \le 1,\ i = 1, \dots, n$, is required) $V - W$ is a regular splitting and (see [3], Theorem 3.13)

$$\mathbf{T} \ge 0, \quad \mathbf{d} > 0, \quad \rho(\mathbf{T}) < 1.$$

3. Bounds for the solution

Lemma 3.1 *Let x be the exact solution to $Ax = b$. Let us consider an iterative process $x^{(k+1)} = \boldsymbol{T}x^{(k)} + \boldsymbol{d}$ with $\boldsymbol{T} \ge 0$ and $\rho(\boldsymbol{T}) < 1$. If*

$$x^{(l+1)} \ge x^{(l)} \tag{1}$$

for some $l \in \mathbf{N}$, then

$$x \ge x^{(l+2)} \ge x^{(l+1)}. \tag{2}$$

Similarly, if $x^{(l+1)} \le x^{(l)}$ for some $l \in \mathbf{N}$, then

$$x \le x^{(l+2)} \le x^{(l+1)}.$$

Notice that condition (1) is equivalent to $Ax^{(l)} \le b$, see [3]. Proof of this lemma is easy and can be found in [1].

If we get an approximation $x^{(k)}$ and a modificating vector v, we will try to find a vector $y^{(k)} = x^{(k)} + \delta v$ so that this $y^{(k)}$ has property (1),

$$y^{(k+1)} = \mathbf{T}y^{(k)} + \mathbf{d} \ge y^{(k)}. \tag{3}$$

Solving this inequality with variable δ we find a set of acceptable parameters δ^U. In the same way we find δ^L by solving the opposite inequality. Then we set the upper and lower bounds to be in the following form:

$$x^{(k)} + \delta^L v \le x \le x^{(k)} + \delta^U v.$$

Inequalities (3) have the form

$$\delta^L(I - \mathbf{T})v \le r^{(k)} \quad \text{and} \quad \delta^U(I - \mathbf{T})v \ge r^{(k)}, \quad \text{where} \quad r^{(k)} = \mathbf{d} - (I - \mathbf{T})x^{(k)}. \tag{4}$$

Sufficient condition for these inequalities to have a solution is $(I - \mathbf{T})v > 0$, or equivalently

$$r^{(v)} < \mathbf{d}, \tag{5}$$

where $r^{(v)} = \mathbf{d} - (I - \mathbf{T})v$. Thus, $\mathbf{d} - r^{(v)} = (I - \mathbf{T})v$ and inequalities (4) will be

$$\delta^L(\mathbf{d} - r^{(v)}) \leq r^{(k)}, \qquad \delta^U(\mathbf{d} - r^{(v)}) \geq r^{(k)}.$$

Optimal solution, which yields the highest lower bound $x^L = x^{(k)} + \delta^L v$ and the lowest upper bound $x^U = x^{(k)} + \delta^U v$, is (index i denotes i-th component of a vector)

$$\delta^L = \min_{i=1,\ldots,n} \frac{r_i^{(k)}}{\mathbf{d}_i - r_i^{(v)}}, \qquad \delta^U = \max_{i=1,\ldots,n} \frac{r_i^{(k)}}{\mathbf{d}_i - r_i^{(v)}}.$$

Condition $(I - \mathbf{T})v > 0$ holds for any approximation $v = x^{(k)}$, which has its residual vector $r^{(k)} < \mathbf{d}$, see (5). Here it is useful to have a positive right-hand side b (and therefore $\mathbf{d} > 0$). Therefore, if the residual vector of the approximation $x^{(k)}$ is small enough, we may take $v = x^{(k)}$, $r^{(v)} = r^{(k)}$ and the bounds will be

$$x^U = x^{(k)}(1 + \delta^U), \qquad x^L = x^{(k)}(1 + \delta^L),$$

where

$$\delta^L = \min_{i=1,\ldots,n} \frac{r_i^{(k)}}{\mathbf{d}_i - r_i^{(k)}}, \qquad \delta^U = \max_{i=1,\ldots,n} \frac{r_i^{(k)}}{\mathbf{d}_i - r_i^{(k)}}.$$

4. Application to irreducible Markov chains

Let us now consider a system corresponding to an automaton with n states. This automaton changes its state, switches from one state to another, in certain time steps. If a probability of switching to another state depends on the current state only, we call this system *Markov Chain.* If there exists a connection between every two states, we call this Markov chain *irreducible.*

Probability of transition from i-th state to j-th (if the system is in the i-th state) is denoted by p_{ij}. In this manner we construct a *transition probability matrix* P, which is stochastic (row sums are equal to 1).

A useful characteristic of Markov chain is its *mean first passage times matrix*, denoted M. Its elements m_{ij} are average times between leaving i-th state and reaching j-th state (it is useful when j-th state is dangerous and means some kind of failure). It is computed from the following equation, see [4],

$$M = P(M - M_D) + E,$$

where $M_D = \operatorname{diag}\{m_{11}, \ldots, m_{nn}\}$ and $E = (e_{ij})_{i,j=1}^n$, $e_{ij} = 1$, $i, j = 1, \ldots, n$. If we write this equation for each column separately, we get a set of linear algebraic equations

$$[I - P(I - e_i e_i^T)]M_i = e,$$

where M_i denotes the i-th column of M and $e = (1, \dots, 1)^T$. Matrix of this system is a diagonally dominant M-matrix and the method described above can be applied to find bounds for the solution.

If we use the fixed-point iterations, $M_i^{(k+1)} = P(I - e_i e_i^T) M_i^{(k)} + e$, for solving this problem with $x^{(0)} = e$, we get an approximation $x^{(k)}$, which has its residual vector $r^{(k)} < \mathbf{d} = e$ (condition (5)), after k iterations, $k \leq n$ [1]. Usually it is $k \ll n$.

5. Numerical example

We show these bounds in the following example. Let us consider a set of linear equations with the right-hand side e and matrix ([2], p. 55–56)

$$A = \begin{pmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1/3 & -2/3 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & -0.8 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1/3 & 0 & -2/3 & 0 & 0 & 0 \\ 0 & -1/7 & 0 & 0 & 1 & -2/7 & 0 & -4/7 & 0 & 0 \\ 0 & 0 & -0.2 & 0 & 0 & 1 & 0 & 0 & -0.8 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & -1/3 & 0 & 0 & 0 & 1 & -2/3 & 0 \\ 0 & 0 & 0 & 0 & -1/3 & 0 & 0 & 0 & 1 & -2/3 \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

The method of fixed-point iterations with initial vector e is used for solving this system. The first three columns of the following tables show the vectors of the lower bounds x^L, the vector of the exact solution x, and the vectors of the upper bounds x^U. Approximate solutions $x^{(k)}$ used for creating these bounds are presented in the fourth columns and their residual vectors $r^{(k)}$ in the fifth columns. Furthermore, an error factor δ_{err} is computed as an additional criterion of convergence,

$$\delta_{\text{err}} = \min_{i=1,\dots,n} \frac{x_i^L}{x_i^U}. \tag{6}$$

6. Conclusions

Systems of linear algebraic equations with an M-matrix appear in many parts of mathematics. If the right-hand side vector of the given system is positive, we may use this simple method to bound the exact solution with help of basic iterative methods.

The obtained bounds may be used to verify the accuracy of the computed solution. The approximate solutions x^L, x^U computed in Table 1 can be used to restart the iterative process [1].

Lower bnd. x^L	Exact sol. x	Upper bnd. x^U	Approx. $x^{(k)}$	Residual $r^{(k)}$
99.269406	105.000000	106.412140	79.876550	0.236850
98.320974	104.000000	105.395466	79.113400	0.236850
82.846169	87.579104	88.807202	66.661688	0.195356
104.635728	110.710448	112.164585	84.194530	0.247642
102.307712	108.223881	109.669062	82.321305	0.244195
98.700281	104.376119	105.802065	79.418607	0.237525
104.397207	110.453731	111.908902	84.002605	0.249366
103.464330	109.453731	110.908902	83.251972	0.249366
101.479317	107.325373	108.781061	81.654742	0.239748
99.647874	105.376119	106.817840	80.181082	0.237525

Tab. 1: *Solution after $k = 150$ iterations, $\|r^{(k)}\|$= 0.752329, δ_{err}=0.932877.*

Lower bnd. x^L	Exact sol. x	Upper bnd. x^U	Approx. $x^{(k)}$	Residual $r^{(k)}$
104.727984	105.000000	105.062731	103.534808	0.013813
103.730431	104.000000	104.061990	102.548621	0.013813
87.354444	87.579104	87.633660	86.359207	0.011393
110.422097	110.710448	110.775045	109.164048	0.014442
107.943055	108.223881	108.288080	106.713250	0.014241
104.106702	104.376119	104.439464	102.920605	0.013852
110.166244	110.453731	110.518374	108.911110	0.014543
109.169430	109.453731	109.518374	107.925653	0.014543
107.047876	107.325373	107.390039	105.828270	0.013982
105.104214	105.376119	105.440165	103.906753	0.013852

Tab. 2: *Solution after $k = 450$ iterations, $\|r^{(k)}\|$= 0.043876, δ_{err}=0.996814.*

Lower bnd. x^L	Exact sol. x	Upper bnd. x^U	Approx. $x^{(k)}$	Residual $r^{(k)}$
104.999085	105.000000	105.000210	104.995017	0.000047
103.999094	104.000000	104.000208	103.995064	0.000047
87.578349	87.579104	87.579287	87.574955	0.000039
110.709478	110.710448	110.710664	110.705188	0.000049
108.222936	108.223881	108.224096	108.218743	0.000048
104.375213	104.376119	104.376332	104.371169	0.000047
110.452765	110.453731	110.453948	110.448485	0.000049
109.452775	109.453731	109.453948	109.448534	0.000049
107.324440	107.325373	107.325590	107.320281	0.000048
105.375205	105.376119	105.376334	105.371122	0.000047

Tab. 3: *Solution after k=1050 iterations, $\|r^{(k)}\|$= 0.000149, δ_{err}=0.999989.*

Disadvantages of this approach are given by strict conditions that need to be fulfilled. Most restrictive conditions are the positive right-hand side and the need for a modificating vector. The positive right-hand side appears in some problems arising in modelling of Markov chains. The modificating vector is obtained either by computing a sufficient approximation, which is sometimes very difficult, or by using an extremely slow iterative method. On the other hand, having the modificating vector, one matrix-vector multiplication is enough to construct these bounds.

References

[1] M. Kocurek: *Acceleration of iterative methods for computing the mean first passage times matrix.* Diploma Thesis, MFF UK Praha, 2005 (in Czech).

[2] W.J. Stewart: *Introduction to the numerical solution of Markov chains.* Princeton University Press, Princeton, New Jersey, 1994.

[3] R.S.Varga: *Matrix iterative analysis.* Prentice-Hall, Englewood Cliffs, New Jersey, 1962.

[4] Š. Klapka: *Markov models in signalling systems.* Disertation Thesis, MFF UK Praha, 2002 (in Czech).

[5] Y. Saad: *Iterative methods for sparse linear systems.* SIAM, Philadelphia, 1996.

ON THE LONGEST-EDGE BISECTION ALGORITHM*

Aleš Kropáč,   Michal Křížek

There are many methods for refining finite element simplicial partitions in R^d, $d \in \{2, 3, \dots\}$. One of them is the longest-edge bisection algorithm. It is very popular for its simplicity, especially in the three-dimensional space. It chooses the longest edge in a given simplicial partition. Dividing this edge by its midpoint, we can define a locally refined partition by simplices that surround this midpoint. Repeating this process, we obtain a family of nested face-to-face partitions (see Figures 1, 2, and 4). This approach is much simpler (especially for $d > 2$) than the standard local refinement of simplicial partitions that uses red and green subdivisions (see, e.g., [3], [7]). Note that this family is never uniquely defined, since during the refinement process there appear many new edges having the same length due to the bisections. For instance, the last but one bisection in Figure 1 is not uniquely determined.

There is an extensive literature devoted to numerical analysis of the longest-edge bisection algorithm, see [1]–[20]. For instance, Rosenberg and Stenger [16] for $d = 2$ show that angles of triangles do not tend to zero for infinitely many steps of a bisection algorithm. A somewhat stronger result has been achieved by M. Stynes [20] who showed that the repeated bisection process yields only a finite number of similarity-distinct subtriangles. This number is bounded when the discretization parameter h tends to zero. However, Stynes admits the so-called hanging nodes which do not appear in face-to-face partitions considered in this paper.

Without loss of generality we can analyse the longest-edge bisection algorithm only for one simplex from a given initial simplicial partition. In Figures 1 and 2, we observe subsequent partitions of a triangle and a tetrahedron by the longest-edge bisection algorithm.

The worst case from the point of degeneracy happens when the regular simplex is bisected (see [5]). For instance, for the equilateral triangle, the minimal angle is halved. On the other hand, this situation does not occur while bisecting obtuse and right triangles. In the next theorem we prove that in this case the minimal angle does not change.

Theorem 1. *Let α be the smallest angle of a nonacute triangle. Bisecting the longest edge determines two triangles whose all angles are not less than α.*

*This paper was supported by Institutional Research Plan nr. AV0Z 10190503 and Grant nr. 201/04/1503 of the Grant Agency of the Czech Republic.

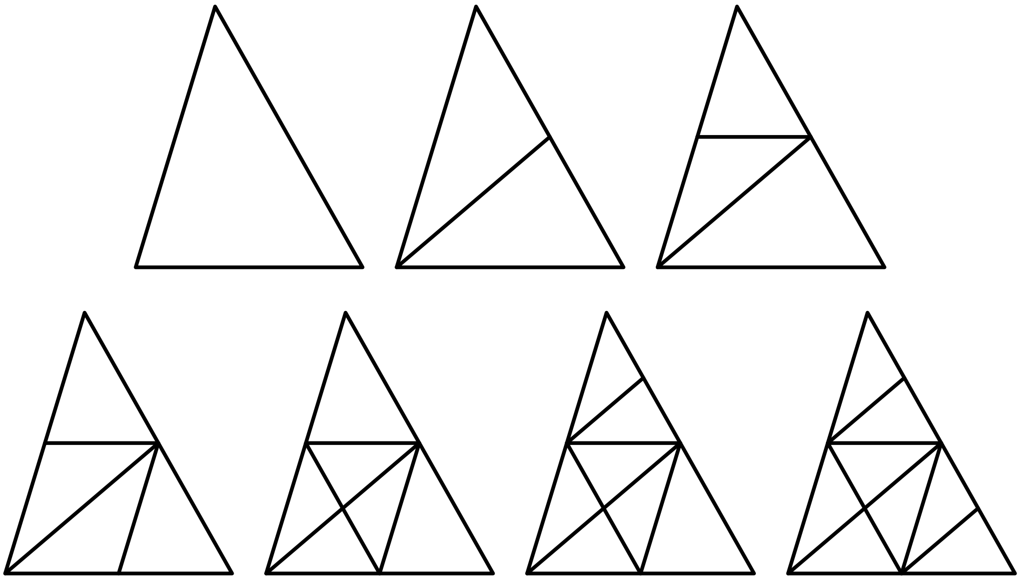


Fig. 1:

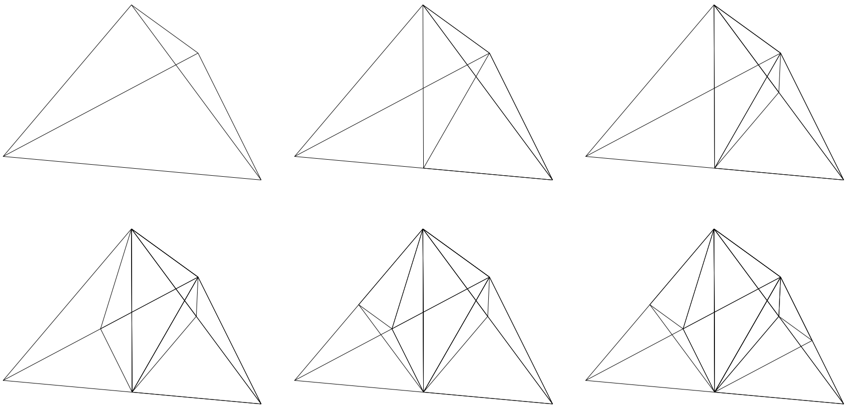


Fig. 2:

Proof. Let a nonacute triangle be given. Denote its angles so that

$$\alpha \leq \beta \leq \frac{\pi}{2} \leq \gamma \tag{1}$$

and let

$$a \leq b \leq c \tag{2}$$

be the associated edges.

Now bisect the triangle by the median t to the longest edge c. Denote the new angles by $\alpha_1, \beta_1, \gamma_1$, and γ_2 as illustrated in Figure 3. We show that all these angles are not less than α.

By the Cosine theorem we see that

$$\begin{aligned} a^2 &= t^2 + \left(\frac{c}{2}\right)^2 - tc\cos\alpha_1, \\ b^2 &= t^2 + \left(\frac{c}{2}\right)^2 - tc\cos\beta_1. \end{aligned}$$

From this and (2) we find that $\cos\alpha_1 \geq \cos\beta_1$. Since $\alpha_1 + \beta_1 = \pi$ and the function cos is decreasing on the whole interval $[0, \pi]$, we have

$$\alpha_1 \leq \frac{\pi}{2} \leq \beta_1. \tag{3}$$

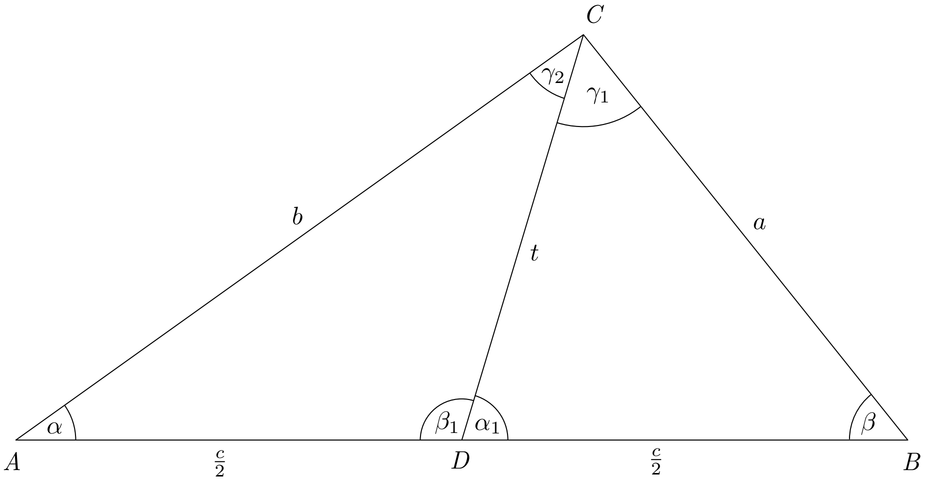


Fig. 3:

Denote vertices of the original triangle ABC as marked in Figure 3. Let D be the midpoint of the segment AB and let C' be such a point that D is the midpoint of the segment CC', i.e., $ACBC'$ is a parallelogram. Using the triangle inequality for the triangle ACC' and relation (2), we get $2t < a + b \leq 2b$, i.e.,

$$t < b.$$

From this and the Sine theorem we obtain

$$\frac{\sin\alpha}{a} = \frac{\sin\beta}{b} < \frac{\sin\beta}{t} = \frac{\sin\alpha_1}{a},$$

which implies that

$$\alpha \leq \alpha_1. \tag{4}$$

Finally, by (1) we know that $\gamma \geq \frac{\pi}{2}$, and therefore, $t \leq \frac{c}{2}$. Using again the Sine theorem, we come to

$$\alpha \leq \gamma_2, \quad \beta \leq \gamma_1. \tag{5}$$

From this, (1), (3), and (4) the lemma follows. □

Remark 1. It is $\gamma_2 \leq \gamma_1$, since by (5) and (1) we have

$$\frac{2\sin\gamma_2}{c} = \frac{\sin\alpha}{t} \leq \frac{\sin\beta}{t} = \frac{2\sin\gamma_1}{c}.$$

Remark 2. From the inequality

$$b \geq \frac{a+b}{2} > \frac{c}{2}$$

we observe that the edge b will be bisected in the next step.

Theorem 2. *Let α_0 be the minimum angle in a given triangulation. Then the longest-edge bisection algorithm yields the following lower bound for any angle α of refined triangles:*

$$\alpha \geq \frac{\alpha_0}{2}.$$

The proof is quite complicated and technical. It is based on some ideas from [16]. We see that for the equilateral triangle the above lower bound $\alpha_0/2$ is attainable. Let us point out that a similar theorem, which guarantees a nondegeneracy in $d = 3$, is still an open problem, even though all triangles on surfaces of all tetrahedra in the partition will be bisected in the same way as for $d = 2$.

Numerical tests. In Figure 4, we observe the initial triangulation and the result of the longest-edge bisection algorithm after 10 and 1000 refining steps.

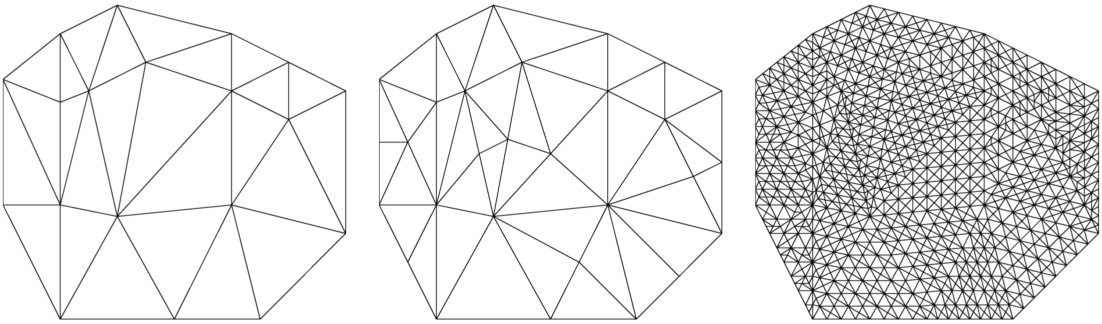


Fig. 4:

To illustrate that repeated bisection process yields only a finite number of similarity-distinct subtriangles, we have chosen the initial triangle with vertices (0,0), (10,0), and (9,3.2). Numerical results in Figure 5 indicate that this number is bounded when $h \to 0$ (cf. [20] for a different approach which produces hanging nodes, in general). In this test we performed 1000 bisections. In Figure 6 we observe values of the maximal and minimal angles from the interval $(0°, 180°)$ during the 1000 bisections. The minimal angle $\approx 18°$ does not change.

The number of nonsimilar subtriangles

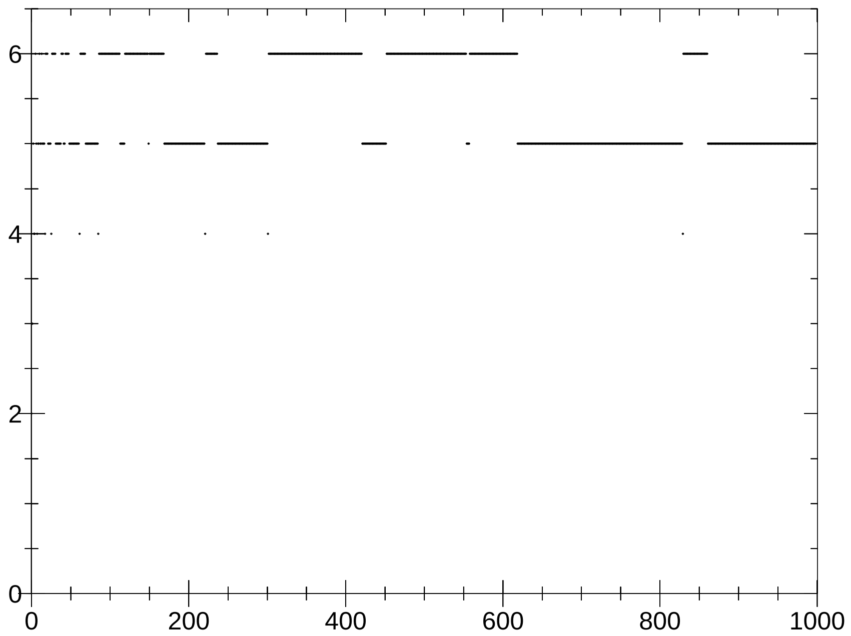


Fig. 5:

Behaviour of the maximal and minimal angles

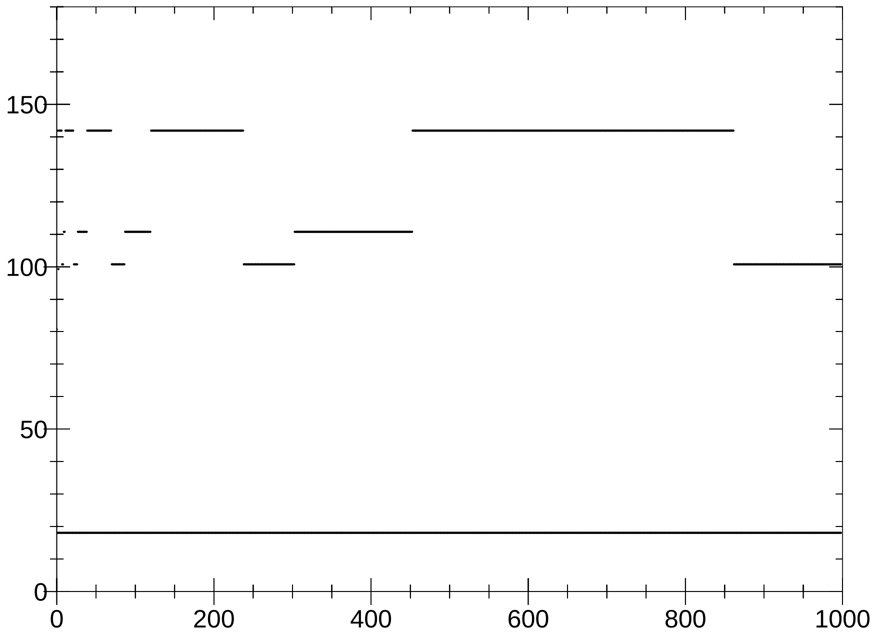


Fig. 6:

References

[1] A. Adler: *On the bisection method for triangles.* Math. Comp. **40**, 1983, 571–574.

[2] D.N. Arnold, A. Mukherjee, L. Pouly: *Locally adapted tetrahedra meshes using bisection.* SIAM J. Sci. Comput. **22**, 2001, 431–448.

[3] E. Bänsch: *Local mesh refinement in 2 and 3 dimensions.* IMPACT Comp. Sci. Engrg. **3**, 1991, 181–191.

[4] A. Eiger, K. Sikorski, F. Stenger: *A bisection method for systems of nonlinear equations.* ACM Trans. Math. Software **10**, 1984, 367–377.

[5] R. Horst: *On generalized bisection of n-simplices.* Math. Comp. **66**, 1997, 691–698.

[6] R.B. Kearfott: *A proof of convergence and an error bound for the method of bisection in R^n.* Math. Comp. **32**, 1978, 1147–1153.

[7] M. Křížek, T. Strouboulis: *How to generate local refinements of unstructured tetrahedral meshes satisfying a regularity ball condition.* Numer. Methods Partial Differential Equations **13**, 1997, 201–214.

[8] A. Liu, B. Joe: *On the shape of tetrahedra from bisection.* Math. Comp. **63**, 1994, 141–154.

[9] A. Liu, B. Joe: *Quality of local refinement of tetrahedral meshes based on bisection.* SIAM. J. Sci. Comput. **16**, 1995, 1269–1291.

[10] J.M. Maubach: *Local bisection refinement.* SIAM J. Sci. Comput. **13**, 1992, 210–227.

[11] Á. Plaza, G.F. Carey: *About local refinement of tetrahedral grids based on bisection.* Proc. 5th Internat. Conf. Meshing Roundtable, 1996, 123–136.

[12] Á. Plaza, M.-C. Rivara: *Mesh refinement based on the 8-tetrahedra longest-edge partition.* Preprint, 1–12.

[13] M.-C. Rivara: *Mesh refinement process based on the generalized bisection of simplices.* SIAM J. Numer. Anal. **21**, 1984, 604–613.

[14] M.-C. Rivara: *New longest-edge bisection algorithm for the refinement and/or improvement of unstructured triangulations.* Internat. J. Numer. Methods Engrg. **40**, 1997, 3313–3324.

[15] M.-C. Rivara, G. Iribarren: *The 4-triangles longest-side partition and linear refinement algorithm.* Math. Comp. **65**, 1996, 1485–1502.

[16] I.G. Rosenberg, F. Stenger: *A lower bound on the angles of triangles constructed by bisection of the longest side.* Math. Comp. **29**, 1975, 390–395.

[17] K. Sikorski: *A three dimensional analogue to the method of bisections for solving nonlinear equations.* Math. Comp. **33**, 1979, 722–738.

[18] Ch. Stamm, S. Eidenbenz, R. Pajarola: *A modified longest edge bisection triangulation.* Preprint, 1–6.

[19] M. Stynes: *On faster convergence of the bisection method for certain triangles.* Math. Comp. **33**, 1979, 717–721.

[20] M. Stynes: *On faster convergence of the bisection method for all triangles.* Math. Comp. **35**, 1990, 1195–1201.

ŠINDEL SEQUENCES AND THE PRAGUE HOROLOGE[*]

Michal Křížek, Alena Šolcová, Lawrence Somer

1. Introduction

The mathematical model of the astronomical clock of Prague was developed by the professor of Prague University, Jan Ondřejův, called *Šindel* (see [2]). The clock was realized by Mikuláš from Kadaň around 1410. The ingenuity of clockmakers of that time can be demonstrated by the following construction.

The astronomical clock of Prague contains a large gear with 24 slots at increasing distances along its circumference (see Figure 1). This arrangement allows for a periodic repetition of 1–24 strokes of the bell each day. There is also a small auxiliary gear whose circumference is divided by 6 slots into segments of arc lengths 1, 2, 3, 4, 3, 2 (see Figure 1). These numbers form a period which repeats after each revolution and their sum is $s = 15$. At the beginning of every hour a catch rises, both gears start to revolve and the bell chimes. The gears stop when the catch simultaneously falls back into the slots on both gears. The bell strikes $1 + 2 + \cdots + 24 = 300$ times every day. Since this number is divisible by $s = 15$, the small gear is always at the same position at the beginning of each day.

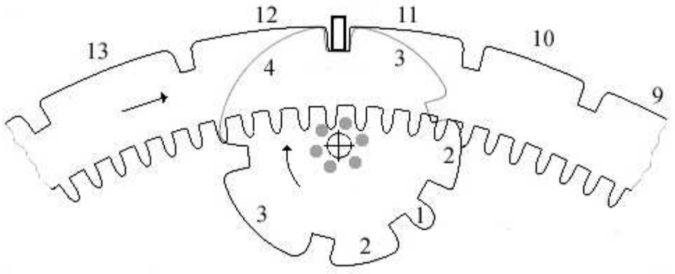


Fig. 1: *The number of bell strokes is denoted by the numbers ..., 9, 10, 11, 12, 13,... along the large gear. The small gear placed behind it is divided by slots into segments of arc lengths 1, 2, 3, 4, 3, 2. The catch is indicated by a small rectangle on the top.*

When the small gear revolves it generates by means of its slots a periodic sequence whose particular sums correspond to the number of strokes of the bell at each hour,

$$1\ 2\ 3\ 4\ \underbrace{3\ 2}_{5}\ \underbrace{1\ 2\ 3}_{6}\ \underbrace{4\ 3}_{7}\ \underbrace{2\ 1\ 2\ 3}_{8}\ \underbrace{4\ 3\ 2}_{9}\ \underbrace{1\ 2\ 3\ 4}_{10}\ \underbrace{3\ 2\ 1\ 2\ 3}_{11}\ \underbrace{4\ 3\ 2\ 1\ 2}_{12}\ \dots \qquad (1)$$

[*] This paper was supported by Institutional Research Plan nr. AV0Z 10190503 and Grant nr. 1P05ME749 of the Ministry of Education of the Czech Republic.

In [4] we showed that we could continue in this way until infinity. However, not all periodic sequences have such a nice summation property. For instance, we immediately find that the period 1, 2, 3, 4, 5, 4, 3, 2 could not be used for such a purpose, since 6 < 4+3. Also the period 1, 2, 3, 2 could not be used, since 2+1 < 4 < 2+1+2.

2. Connections with triangular numbers and periodic sequences

In this section we show how the *triangular numbers*

$$T_k = 1 + 2 + \cdots + k = \frac{k(k+1)}{2}, \quad k = 0, 1, 2, \dots, \tag{2}$$

are related to the astronomical clock. We shall look for all periodic sequences that have a similar property as the sequence 1, 2, 3, 4, 3, 2 in (1), i.e., that could be used in the construction of the small gear. Put $\mathbb{N} = \{1, 2, \dots\}$.

A sequence $\{a_i\}_{i=1}^{\infty}$ is said to be *periodic,* if there exists $p \in \mathbb{N}$ such that

$$\forall\, i \in \mathbb{N} : \quad a_{i+p} = a_i. \tag{3}$$

The finite sequence $a_1, \dots, a_p$ is called a *period* and p is called the *period length.* The smallest p satisfying (3) is called the *minimal period length* and the associated sequence $a_1, \dots, a_p$ is called the *minimal period.*

Definition 1. Let $\{a_i\} \subset \mathbb{N}$ be a periodic sequence. We say that the triangular number T_k for $k \in \mathbb{N}$ is *achievable* by $\{a_i\}$, if there exists a positive integer n such that

$$T_k = \sum_{i=1}^{n} a_i. \tag{4}$$

The periodic sequence $\{a_i\}$ is said to be a *Šindel sequence* if T_k is achievable by $\{a_i\}$ for every $k \in \mathbb{N}$, i.e.,

$$\forall\, k \in \mathbb{N} \quad \exists\, n \in \mathbb{N} : \quad T_k = \sum_{i=1}^{n} a_i. \tag{5}$$

The triangular number T_k on the left-hand side is equal to the sum $1 + \cdots + k$ of hours on the large gear, whereas the sum on the right-hand side expresses the corresponding rotation of the small gear (see Figure 2). For the kth hour, we have

$$k = T_k - T_{k-1} = \sum_{i=m+1}^{n} a_i, \tag{6}$$

where $T_{k-1} = \sum_{i=1}^{m} a_i$. Since $a_i > 0$, the number n depending on k in (5) is unique. From (2) and (4) we also see that $a_1 = 1$ when $\{a_i\}$ is a Šindel sequence.

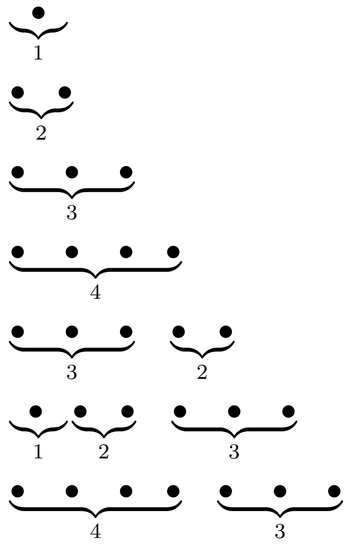


Fig. 2: *The bullets in the kth row indicate the number of strokes at the kth hour (see (6)). The numbers denote lengths of segments on the small gear.*

3. Necessary and sufficient condition for the existence of a Šindel sequence

First we need to define quadratic residues and nonresidues.

Definition 2. Let $n \geq 2$ and a be integers. If the quadratic congruence

$$x^2 \equiv a \pmod{n}$$

has a solution x, then a is called a *quadratic residue modulo n*. Otherwise, a is called a *quadratic nonresidue modulo n*.

Lemma 1. *If f and h are nonnegative integers, then $8f+1$ is a quadratic residue modulo 2^h.*

The proof is a consequence of [5, pp. 105–106]). From now on let

$$s = \sum_{i=1}^{p} a_i \tag{7}$$

denote the sum of the period.

Theorem 1. *A periodic sequence $\{a_i\}$ is a Šindel sequence if and only if for any $n \in \{1, \dots, p\}$ and any $j \in \{1, 2, \dots, a_n - 1\}$ with $a_n \geq 2$ the number*

$$w = 8\Big(\sum_{i=1}^{n} a_i - j\Big) + 1$$

is a quadratic nonresidue modulo s.

Proof. $\Longleftarrow$: Let a periodic sequence $\{a_i\}$ not be a Šindel sequence. According to (5), there exist positive integers ℓ, m, and j such that $a_m \geq 2$, $j \leq a_m - 1$, and

$$T_\ell = \sum_{i=1}^{m} a_i - j. \tag{8}$$

Let $n \in \{1, \dots, p\}$ be such that $n \equiv m \pmod{p}$. Then by (2), (8), (7), and (3),

$$(2\ell+1)^2 = 4\ell^2 + 4\ell + 1 = 8T_\ell + 1 = 8\Big(\sum_{i=1}^{m} a_i - j\Big) + 1 \equiv 8\Big(\sum_{i=1}^{n} a_i - j\Big) + 1 \pmod{s},$$

i.e., $8\big(\sum_{i=1}^{n} a_i - j\big) + 1$ is a square modulo s.

$\Longrightarrow$: Let $\{a_i\}$ be a Šindel sequence with $s = 2^c d$, where $c \geq 0$ and d is odd. Suppose to the contrary that there exist positive integers n, j, and x such that $n \leq p$, $a_n \geq 2$, $j \leq a_n - 1$, $x \leq s$, and

$$w = 8\Big(\sum_{i=1}^{n} a_i - j\Big) + 1 \equiv x^2 \pmod{s}. \tag{9}$$

From Lemma 1 and (9) there exists y such that

$$\begin{aligned} x^2 &\equiv w \pmod{d}, \\ y^2 &\equiv w \pmod{2^{c+3}}. \end{aligned} \tag{10}$$

By the Chinese remainder theorem (see [3, p. 15]) there exists an integer $u \geq 3$ such that $u \equiv x \pmod{d}$ and $u \equiv y \pmod{2^{c+3}}$. Thus, by (10),

$$\begin{aligned} u^2 &\equiv x^2 \equiv w \pmod{d}, \\ u^2 &\equiv y^2 \equiv w \pmod{2^{c+3}}. \end{aligned}$$

Since $\gcd(d, 2^{c+3}) = 1$, we see that

$$u^2 \equiv w \pmod{2^{c+3}d}. \tag{11}$$

Clearly, u is odd, since w is odd. So let $u = 2\ell + 1$, where $\ell \geq 1$. Then, by (11), $u^2 = 4\ell^2 + 4\ell + 1 = w + 2^{c+3}dg$ for some integer g. Hence, since $u \geq 3$, we find by (2), (11), and (9) that

$$T_\ell = \frac{u^2 - 1}{8} = \frac{w - 1}{8} + 2^c dg \equiv \sum_{i=1}^{n} a_i - j \pmod{s}.$$

Thus, there exists a positive integer m such that $m \equiv n \pmod{p}$ and

$$T_\ell = \sum_{i=1}^{m} a_i - j,$$

which contradicts the assumption that $\{a_i\}$ is a Šindel sequence. □

As a byproduct of the proof of Theorem 1, we get the well-known result (see also [1, p. 15] and Figure 3):

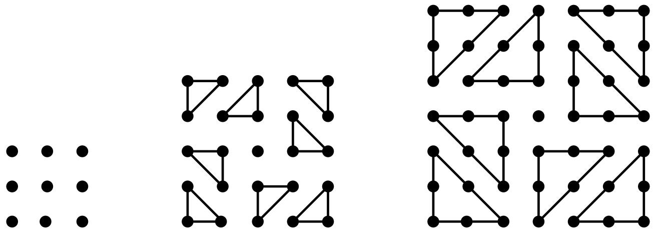


Fig. 3: *The early Pythagoreans knew that if r is a triangular number, then $8r+1$ is a square. This result is mentioned as early as about 100 A.D. in Platonic Questions by the Greek historian Plutarch, see [6, p. 4].*

Corollary 1. *A positive integer r is a triangular number if and only if $8r+1$ is a square.*

Remark 1. In Theorem 1, we require that

$$w = 8\Big(\sum_{i=1}^{n} a_i - j\Big) + 1$$

be a quadratic nonresidue modulo s for various values of n and j when $\{a_i\}$ is a Šindel sequence. A sufficient condition for this to occur is that w be a quadratic nonresidue for some odd prime q dividing s. To see that this condition is not necessary, consider the periodic sequence $\{a_i\}$ given in Example 2 below with $p=11$, $s=25$, and the period 1, 2, 2, 1, 4, 1, 4, 1, 4, 1, 4. Then

$$8\Big(\sum_{i=1}^{5} a_i - 2\Big) + 1 = 65,$$

which is a quadratic nonresidue modulo 25, but is a quadratic residue modulo 5. Note that 5 is the only odd prime dividing $s=25$.

Remark 2. Consider the sequence $\{a_i\}$ with period $1,2,1,1,1,\ldots,1$. Note that

$$w = 8\Big(\sum_{i=1}^{2} a_i - 1\Big) + 1 = 17.$$

By Theorem 1 and the law of quadratic reciprocity one sees that (cf. [3, pp. 23–25]) if s is an odd prime and $s \equiv 1,2,4,8,9,13,15$ or $16 \pmod{17}$, then w is a quadratic residue modulo s and thus, $\{a_i\}$ is not a Šindel sequence. Other patterns of the period of periodic sequences $\{a_i\}$ can be similarly investigated.

4. Construction of the primitive Šindel sequence

Definition 3. A Šindel sequence $\{a'_i\}$ with the minimal period length $p+1$ is said to be *composite,* if there exists a Šindel sequence $\{a_i\}$ and $\ell \in \mathbb{N}$ such that

$$\begin{aligned} a_i &= a'_i, \quad i = 1, \dots, \ell - 1, \\ a_\ell &= a'_\ell + a'_{\ell+1}, \\ a_i &= a'_{i+1}, \quad i = \ell + 1, \dots, p. \end{aligned}$$

The period 1, 2, 3, 2, 2, 3, 2 derived from the period 1, 2, 3, 4, 3, 2 of sequence (1) produces a composite Šindel sequence. In other words, the astronomical clock would also work with the small gear corresponding to this composite Šindel sequence.

Definition 4. A Šindel sequence $\{a_i\}$ is called *primitive* if it is not composite. The sequence 1, 1, 1, ... is called a *trivial* Šindel sequence.

The proof of the next theorem contains an explicit algorithm for finding a primitive Šindel sequence for a given s.

Theorem 2. *Let s be a positive integer. Then there exists a unique primitive Šindel sequence $\{a_i\}$ such that* (7) *holds for one of its not necessarily minimal period lengths p. The primitive Šindel sequence $\{a_i\}$ is trivial if and only if $s = 2^h$ for $h \geq 0$.*

Proof. Let $1 \leq b_1 < b_2 < \cdots < b_t \leq s$ be all the integers such that each $8b_n + 1$ is a square modulo s for $n = 1, \dots, t$. We observe that $b_1 = 1$ and $b_t = s$. Now choose the period as follows: $a_1 = b_1$ and $a_n = b_n - b_{n-1}$ for $n = 2, 3, \dots, t$. Then

$$\forall n \in \{1, 2, \dots, t\} : \quad b_n = \sum_{i=1}^{n} a_i.$$

We claim that $\{a_i\}$ is a Šindel sequence. Note that if $n \in \{1, \dots, t\}$, $a_n \geq 2$, and $j \in \{1, 2, \dots, a_n - 1\}$, then $b_{n-1} < \sum_{i=1}^{n} a_i - j < b_n$. Then $8(\sum_{i=1}^{n} a_i - j) + 1$ is a quadratic nonresidue modulo s, since $8b_1 + 1, \dots, 8b_t + 1$ are all the quadratic residues modulo s. It now follows from Theorem 1 that $\{a_i\}$ is a Šindel sequence.

Moreover, one sees that $\{a_i\}$ is a primitive Šindel sequence having a period length $p = t$ and satisfying (7). It is also clear by construction that $\{a_i\}$ is the unique primitive Šindel sequence satisfying (7) for some period length p.

$\Longleftarrow$: By the above construction of the period, the primitive Šindel sequence corresponding to s is nontrivial if and only if there exists a positive integer $f \leq s$ such that $8f + 1$ is a quadratic nonresidue modulo s. By Lemma 1, $8f + 1$ is always a quadratic residue modulo $s = 2^h$ for $h \geq 0$. Hence, the primitive Šindel sequence corresponding to $s = 2^h$ is the trivial Šindel sequence.

$\Longrightarrow$: Conversely, assume that s has an odd prime divisor q. Let d be a quadratic nonresidue modulo q. Since 8 is invertible modulo q, one sees that if z is the inverse of 8 modulo q and $f \equiv z(d-1) \pmod{q}$, then $8f+1 \equiv d \pmod{q}$. It now follows that the primitive Šindel sequence corresponding to s is nontrivial. □

We have the following immediate corollaries to Theorems 2 and 1:

Corollary 2. *Let $\{a_i\}$ be a periodic sequence with the minimal length p of the period and $s = 2^m$, where m is a nonnegative integer. Then $\{a_i\}$ is a Šindel sequence if and only if $\{a_i\}$ is the trivial Šindel sequence.*

Corollary 3. *A periodic sequence $\{a_i\}$ is a primitive Šindel sequence if and only if for any $n \in \{1, \dots, p\}$ and any $j \in \{1, 2, \dots, a_n - 1\}$ with $a_n \geq 2$ the number*

$$w = 8\Big(\sum_{i=1}^{n} a_i - j\Big) + 1$$

is a quadratic nonresidue modulo s and

$$v = 8\sum_{i=1}^{n} a_i + 1$$

is a quadratic residue modulo s.

Theorem 3. *For any $k \in \mathbb{N}$ there exist $\ell \in \mathbb{N}$ and a Šindel sequence $\{a_i\}$ such that $a_\ell = k$.*

Proof. It was stated in Corollary 1 that for $r \in \mathbb{N}$, $8r+1$ is a square if and only if r is a triangular number. Let $k = T_k - T_{k-1}$ be given (see (6)). Thus it suffices by the proof of Theorem 2 to find a positive integer $s \geq T_k$ such that $8(T_{k-1}+j)+1$ is a quadratic nonresidue modulo s for $j = 1, 2, \dots, k-1$.

For a fixed $j \in \{1, \dots, k-1\}$ let

$$8(T_{k-1}+j)+1 = \prod_{i=1}^{v} p_i^{\alpha_i}$$

be the prime power factorization. Since $8(T_{k-1}+j)+1$ is not a square, some α_i is odd. Without loss of generality, we can assume that α_1 is odd. Let c_1 be a quadratic nonresidue modulo p_1. By the Chinese remainder theorem and Dirichlet's theorem on the infinitude of primes in arithmetic progressions, one can find a prime $q_j \geq T_k$ such that $q_j \equiv 1 \pmod{4}$, $q_j = c_1 \pmod{p_1}$, and $q_j \equiv 1 \pmod{p_i}$ for $i \in \{2, \dots, v\}$. By the law of quadratic reciprocity and the properties of the Jacobi symbol (see [3, p. 24–25]), $8(T_{k-1}+j)+1$ is a quadratic nonresidue modulo q_j. Now simply let s be the product of the distinct q_j's for $j \in \{1, \dots, k-1\}$. □.

5. Numerical examples

We developed a program that generates the primitive Šindel sequence for a given s. It is based on the numerical algorithm presented in the proof of Theorem 2. By this theorem we know that the primitive primitive Šindel sequence is uniquely determined for each positive integer s.

s	Primitive Šindel sequences
1	1
2	1 1
3	1 2
4	1 1 1 1
5	1 2 2
6	1 2 1 2
7	1 2 3 1
8	1 1 1 1 1 1 1 1
9	1 2 3 3
10	1 2 2 1 2 2
11	1 2 1 2 4 1
12	1 2 1 2 1 2 1 2
13	1 1 1 3 2 2 3
14	1 2 3 1 1 2 3 1
15	1 2 3 4 3 2
16	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
17	1 1 1 1 2 4 1 4 2
18	1 2 3 3 1 2 3 3
19	1 1 1 3 1 2 1 5 2 2
20	1 2 2 1 2 2 1 2 2 1 2 2
21	1 2 3 1 3 3 2 6
22	1 2 1 2 4 1 1 2 1 2 4 1
23	1 2 2 1 3 1 3 2 5 1 1 1
24	1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2
25	1 2 2 1 4 1 4 1 4 1 4

Example 1. The period 1, 2, 3, 4, 5, 3, 3, 7, 2, 3, 3, 9 with minimal period length $p = 12$ and $s = 45$ yields a primitive Šindel sequence $\{a_i\}$ with a large value of $a_{12} = 9$ relative to s (see Theorem 3).

Example 2. The next table shows values of all primitive Šindel sequences for $s = 1, \dots, 25$. Anyway, we verified that no primitive Šindel sequence up to $s = 1000$ has such a nice symmetry property as that in (1). From the table we also observe that trivial primitive Šindel sequences appear when $s = 2^h$ for some $h \geq 0$ (see Theorem 2).

References

[1] D.M. Burton: *Elementary number theory*, fourth edition. McGraw-Hill, New York, 1998.

[2] Z. Horský: *The astronomical clock of Prague*. Panorama, Prague, 1988.

[3] M. Křížek, F. Luca, L. Somer: *17 lectures on Fermat numbers: From number theory to geometry*. CMS Books in Mathematics **9**, Springer-Verlag, New York, 2001.

[4] M. Křížek, L. Somer, A. Šolcová: *Jaká matematika se ukrývá v pražském orloji?* Matematika-fyzika-informatika **16**, 2006/2007, 129–137.

[5] I. Niven, H.S. Zuckerman, H.L. Montgomery: *An introduction to the theory of numbers*, fifth edition. John Wiley & Sons, New York, 1991.

[6] J.J. Tattersall: *Elementary number theory in nine chapters*, second edition. Cambridge Univ. Press, 2005.

ON SOLVING NON-SYMMETRIC SADDLE-POINT SYSTEMS ARISING FROM FICTITIOUS DOMAIN APPROACHES*

Radek Kučera, Tomáš Kozubek, Jaroslav Haslinger

1. Introduction

We propose a fast method for finding a pair $(u, \lambda) \in \mathbb{R}^n \times \mathbb{R}^m$ that solves a linear system of algebraic equations called the *(generalized) saddle-point system*:

$$\begin{pmatrix} A & B_1^\top \\ B_2 & 0 \end{pmatrix} \begin{pmatrix} u \\ \lambda \end{pmatrix} = \begin{pmatrix} f \\ g \end{pmatrix}, \tag{1}$$

where the diagonal block A is an $(n \times n)$ matrix, the off-diagonal blocks B_1 and B_2 are $(m \times n)$ matrices with full row-rank and vectors f, g are of order n, m, respectively.

Our contribution is inspired by a class of saddle-point systems arising from fictitious domain formulations of PDEs [3, 4]. Therefore we will be interested especially in systems (1) with n large, A singular and B_1, B_2 sparse. Moreover, we will assume that m is much smaller than n and that the defect l of A, i.e. $l = n - rank\, A$, is much smaller than m.

There are several basic approaches used for solving (1); see e.g. [1]. Due to the structure of our matrices, we pay our attention to the class of methods that are based on the Schur complement reduction. Their key idea consists in eliminating the first unknown u. This leads, in the case of non-singular A, to the reduced system for the second unknown λ. The matrix of this system is the (negative) Schur complement $-S = B_2 A^{-1} B_1^\top$. If this system is solved by an iterative method, we do not need to form S explicitly since only the matrix-vector products with S are needed.

The situation is not so easy if A is singular. In this case, the first unknown u can not be completely eliminated from (1). The Schur complement reduction leads now to another saddle-point system for λ and a new unknown, say α, that corresponds to the null-space of A. Fortunately after applying orthogonal projectors, we obtain an equation only in terms of λ. As our original saddle-point system (1) is non-symmetric, this equation can be solved by a *projected* Krylov method for non-symmetric matrices. In our numerical tests, we will use the projected variant of the BiCGSTAB algorithm.

*Supported by the National Program of Research "Information Society" under project 1ET400300415 and by the grant IAA1075402 of the Grant Agency of the Czech Academy of Sciences.

The presented method generalizes ideas used in the algebraic description of FETI domain decomposition methods [2], in which A is symmetric, positive semidefinite and $B_1 = B_2$.

2. A new variant of the fictitious domain method

Let Ω be a bounded domain in $\mathbb{R}^d$, $d = 2, 3$ with the Lipschitz boundary $\partial\Omega$, which is split into three non-overlapping parts Γ_D, Γ_N and Γ_G (see Figure 1). We will be concerned with the following abstract class of mixed boundary value problems:

$$\left.\begin{aligned} \mathcal{L}u = f &\quad \text{in} \quad \Omega, \\ u = g_D &\quad \text{on} \quad \Gamma_D, \\ \frac{\partial u}{\partial \nu_{\mathcal{L}}} = g_N &\quad \text{on} \quad \Gamma_N, \\ \frac{\partial u}{\partial \nu_{\mathcal{L}}} + \beta u = g_G &\quad \text{on} \quad \Gamma_G, \end{aligned}\right\} \qquad (\mathcal{P})$$

where $\mathcal{L}$ is an elliptic operator of the second order, $f \in L^2(\Omega)$, $g_D \in H^{1/2}(\Gamma_D)$, $g_N \in L^2(\Gamma_N)$, $g_G \in L^2(\Gamma_G)$, β is a constant and $\frac{\partial}{\partial \nu_{\mathcal{L}}}$ denotes the normal derivative on $\partial\Omega$. We assume that $(\mathcal{P})$ has a unique solution u.

Any fictitious domain (FD) formulation of PDEs transforms the original problem defined in a domain Ω to a new one solved in a simple shaped domain $\hat{\Omega}$ (e.g. a box), which contains $\overline{\Omega}$. Its solution will be denoted by $\hat{u}$. The standard boundary Lagrange multiplier FD approach (see [3]) gives rise to a singularity of $\hat{u}$ located on the boundary $\partial\Omega$. This fact can result in an intrinsic error of the computed solution. Therefore we recommend to move this singularity further of $\partial\Omega$, i.e. to enforce the prescribed boundary conditions by new control variables defined not on $\partial\Omega$ but on an

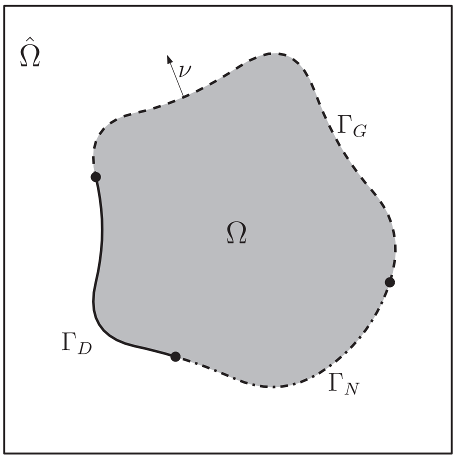


Fig. 1: *Geometry.*

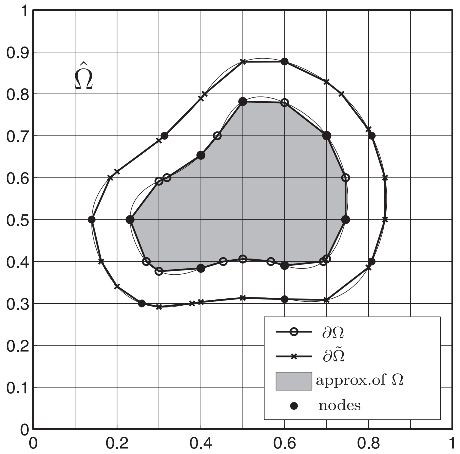


Fig. 2: *Auxiliary boundary $\partial\tilde{\Omega}$.*

auxiliary boundary $\partial\tilde{\Omega} = \tilde{\Gamma}_D \bigcup \tilde{\Gamma}_N \bigcup \tilde{\Gamma}_G$ obtained by shifting the Bezièr approximation of $\partial\Omega = \Gamma_D \bigcup \Gamma_N \bigcup \Gamma_G$ in the outer normal direction with a step δ (see Fig. 2). *This approach improves significantly the error of the computed FD solution and the rates of convergence.*

Let us introduce boundary control variables $\tilde{\lambda}_D \in \tilde{\Lambda}_D := H^{-1/2}(\tilde{\Gamma}_D)$, $\tilde{\lambda}_N \in \tilde{\Lambda}_N := H^{-1/2}(\tilde{\Gamma}_N)$ and $\tilde{\lambda}_G \in \tilde{\Lambda}_G := H^{-1/2}(\tilde{\Gamma}_G)$ defined on $\tilde{\Gamma}_D$, $\tilde{\Gamma}_N$ and $\tilde{\Gamma}_G$, respectively. Instead of $(\mathcal{P})$, we will solve the following problem:

$$\left.\begin{array}{l}
\text{Find } (\hat{u}, \tilde{\lambda}_D, \tilde{\lambda}_N, \tilde{\lambda}_G) \in V \times \tilde{\Lambda}_D \times \tilde{\Lambda}_N \times \tilde{\Lambda}_G \text{ such that}\\
a(\hat{u}, \hat{v}) + \tilde{b}_D(\tilde{\lambda}_D, \tilde{\tau}_D\hat{v}) + \tilde{b}_N(\tilde{\lambda}_N, \tilde{\tau}_N\hat{v}) + \tilde{b}_G(\tilde{\lambda}_G, \tilde{\tau}_G\hat{v}) = (\hat{f}, \hat{v})_{0,\hat{\Omega}} \quad \forall \hat{v} \in V,\\
b_D(\mu_D, \tau_D\hat{u}) = b_D(\mu_D, g_D) \quad \forall \mu_D \in \Lambda_D,\\
b_N(\mu_N, \frac{\partial \hat{u}}{\partial \nu_{\mathcal{L}}}) = b_N(\mu_N, g_N) \quad \forall \mu_N \in \Lambda_N,\\
b_G(\mu_G, \frac{\partial \hat{u}}{\partial \nu_{\mathcal{L}}} + \beta\tau_G\hat{u}) = b_G(\mu_G, g_G) \quad \forall \mu_G \in \Lambda_G,
\end{array}\right\} \qquad (\hat{\mathcal{P}})$$

where $a : V \times V \to \mathbb{R}^1$ is a continuous, coercive bilinear form resulting from the weak formulation of the first equation in $(\mathcal{P})$, $\hat{f}$ is an extension of f from Ω to $\hat{\Omega}$, $\tau_D : V \mapsto H^{1/2}(\Gamma_D)$, $\tau_G : V \mapsto H^{1/2}(\Gamma_G)$, $\tilde{\tau}_D : V \mapsto H^{1/2}(\tilde{\Gamma}_D)$, $\tilde{\tau}_N : V \mapsto H^{1/2}(\tilde{\Gamma}_N)$ and $\tilde{\tau}_G : V \mapsto H^{1/2}(\tilde{\Gamma}_G)$ stand for the trace mappings, respectively, and the bilinear forms b_D, b_N, b_G and $\tilde{b}_D$, $\tilde{b}_N$, $\tilde{b}_G$ denote the corresponding duality pairings. Finally, $\Lambda_D := H^{-1/2}(\Gamma_D)$, $\Lambda_N := H^{1/2}(\Gamma_N)$, $\Lambda_G := H^{1/2}(\Gamma_G)$ and V is a closed subspace of $H^1(\hat{\Omega})$. Typical choices for V are: $H^1(\hat{\Omega})$, $H^1_0(\hat{\Omega})$, or $H^1_P(\hat{\Omega}) = \{v | v \in H^1(\hat{\Omega}),\ v \text{ is periodic on } \partial\hat{\Omega}\}$ if $\hat{\Omega}$ is a cartesian product of intervals.

A discretization of $(\hat{\mathcal{P}})$ based on a mixed finite element method leads to a saddle-point system (1). One can use fairly structured meshes in $\hat{\Omega}$ ensuring favorable properties of the stiffness matrix A. Therefore actions of a generalized inverse $A^\dagger$ (or inverse A^{-1}) are cheaply computable and, in addition, the null-space of A and $A^\top$ can be easily identified [6]. The geometry of $\partial\Omega$ together with the type of boundary conditions are characterized by B_1, B_2, which are highly sparse.

3. Algorithms

Denote $\mathbb{N}(B|\mathbb{V})$ the null-space and $\mathbb{R}(B|\mathbb{V})$ the range-space of an $(m \times n)$ matrix B in a subspace $\mathbb{V} \subset \mathbb{R}^n$. If $\mathbb{V} = \mathbb{R}^n$, we simply write $\mathbb{N}(B) := \mathbb{N}(B|\mathbb{R}^n)$ and $\mathbb{R}(B) := \mathbb{R}(B|\mathbb{R}^n)$. The system (1) has a unique solution iff [5]

$$\mathbb{N}(A) \cap \mathbb{N}(B_2) = \{0\}, \tag{2}$$

$$\mathbb{R}(A|\mathbb{N}(B_2)) \cap \mathbb{R}(B_1^\top) = \{0\}. \tag{3}$$

Suppose that A is singular with the defect $l = \dim \mathbb{N}(A)$, $l \geq 1$ and consider $(n \times l)$ matrices N and M whose columns span the null-space $\mathbb{N}(A)$ and $\mathbb{N}(A^\top)$, respectively. Finally, denote by $A^\dagger$ a generalized inverse to A. In what follows we will consider an arbitrary but fixed selections of $A^\dagger$, N and M.

The *generalized Schur complement* of A in (1) is defined by

$$\mathcal{S} = \begin{pmatrix} -B_2 A^\dagger B_1^\top & B_2 N \\ M^\top B_1^\top & 0 \end{pmatrix}.$$

Notice that $\mathcal{S}$ is invertible provided that (2), (3) are satisfied. The following theorem describes the Schur complement reduction.

Theorem 3.1 [5] *Assume that both B_1, B_2 have full row-ranks and that (2), (3) are satisfied. Then the second component λ of a solution to (1) is the first component of a solution to*

$$\begin{pmatrix} F & G_1^\top \\ G_2 & 0 \end{pmatrix} \begin{pmatrix} \lambda \\ \alpha \end{pmatrix} = \begin{pmatrix} d \\ e \end{pmatrix}, \tag{4}$$

where $F := B_2 A^\dagger B_1^\top, G_1 := -N^\top B_2^\top, G_2 := -M^\top B_1^\top, d := B_2 A^\dagger f - g$ and $e := -M^\top f$. The first component u of a solution to (1) is given by the formulae

$$u = A^\dagger (f - B_1^\top \lambda) + N\alpha.$$

Let us point out that (4) is formally the same saddle-point system as (1), but its size is considerably smaller. We will modify the new system (4) by two orthogonal projectors

$$P_1 := I - G_1^\top (G_1 G_1^\top)^{-1} G_1, \qquad P_2 := I - G_2^\top (G_2 G_2^\top)^{-1} G_2,$$

on $\mathbb{N}(G_1)$, $\mathbb{N}(G_2)$, respectively. Our method is based on the following results.

Lemma 3.1 [5] *The linear operator $P_1 F : \mathbb{N}(G_2) \mapsto \mathbb{N}(G_1)$ is invertible.*

Theorem 3.2 [5] *Let $\lambda_\mathbb{N} \in \mathbb{N}(G_2)$, $\lambda_\mathbb{R} \in \mathbb{R}(G_2^\top)$. Then $\lambda = \lambda_\mathbb{N} + \lambda_\mathbb{R}$ is the first component of a solution to (4) iff*

$$\lambda_\mathbb{R} = G_2^\top (G_2 G_2^\top)^{-1} e$$

and

$$P_1 F \lambda_\mathbb{N} = P_1 (d - F\lambda_\mathbb{R}).$$

The second component α is given by

$$\alpha = (G_1 G_1^\top)^{-1} G_1 (d - F\lambda).$$

Let us summarize the previous results in the algorithm scheme. It turns out to be reasonable to form and store the $(l \times m)$ matrices G_1, G_2 and the $(l \times l)$ matrices $H_1 := (G_1 G_1^\top)^{-1}$, $H_2 := (G_2 G_2^\top)^{-1}$ because l is small. On the other hand, the $(m \times m)$ matrices F, P_1 and P_2 are not assembled explicitly.

```
Algorithm: Projected Schur Complement Method (PSCM)
Step 1.a:  Assemble $G_1 = -N^\top B_2^\top$, $G_2 = -M^\top B_1^\top$, $d = B_2 A^\dagger f - g$ and $e = -M^\top f$.
Step 1.b:  Assemble $H_1 = (G_1 G_1^\top)^{-1}$ and $H_2 = (G_2 G_2^\top)^{-1}$.
Step 1.c:  Assemble $\lambda_{\mathbb{R}} = G_2^\top H_2 e$.
Step 1.d:  Assemble $\tilde{d} = P_1(d - F\lambda_{\mathbb{R}})$.
Step 1.e:  Solve the equation $P_1 F \lambda_{\mathbb{N}} = \tilde{d}$ on $\mathbb{N}(G_2)$.
Step 1.f:  Compute $\lambda = \lambda_{\mathbb{N}} + \lambda_{\mathbb{R}}$.
Step 2:    Compute $\alpha = H_1 G_1 (d - F\lambda)$.
Step 3:    Compute $u = A^\dagger (f - B_1^\top \lambda) + N\alpha$.
```

The heart of the algorithm consists in Step 1.e. Its solution can be computed by a *projected* Krylov subspace method. The projected BiCGSTAB algorithm [5] can be derived from the non-projected one [7] by choosing an initial iterate $\lambda_{\mathbb{N}}^0$ in $\mathbb{N}(G_2)$, projecting the initial residual in $\mathbb{N}(G_2)$ and replacing the operator $P_1 F$ by its projected version $P_2 P_1 F$. Finally, let us point out that convergence of the projected BiCGSTAB algorithm can be accelerated by a reorthogonalization procedure or by a multigrid technique.

4. Numerical experiments

We illustrate the efficiency of the presented method on a model problem $(\mathcal{P})$. Let $\mathcal{L} = -\Delta$, $\Omega = \{(x, y) \in \mathbb{R}^2 | \; (x - 0.5)^2/0.4^2 + (y - 0.5)^2/0.2^2 < 1\}$ and consider the mixed Dirichlet-Neumann boundary conditions with Γ_D and Γ_N corresponding to the upper and lower half-part of the ellipse $\partial\Omega$, respectively. Let us choose the right hand-sides f, g_D and g_N in $(\mathcal{P})$ appropriately to the exact solution $u_{ex}(x, y) = 100\,((x - 0.5)^3 - (y - 0.5)^3)$. In the FD formulation $(\hat{\mathcal{P}})$, we take $\hat{\Omega} \equiv (0, 1) \times (0, 1)$ and $V = H^1_P(\hat{\Omega})$. This space is approximated by piecewise bilinear functions defined on a rectangulation of $\hat{\Omega}$ with a stepsize h. The spaces Λ_D, Λ_N and their tilded counterparts are approximated by piecewise constant functions defined on partitions of polygonal approximations of $\partial\Omega$ and $\partial\tilde{\Omega}$.

In tables below, we report the errors of the approximate solution u_h with respect to the stepsize h in the indicated norms together with the number of primal (n) and control (m) variables, the number of BiCGSTAB iterations and the computational time.

Tables 1 and 2 summarize results obtained by a classical FD method with boundary Lagrange multipliers on $\partial\Omega$. The BiCGSTAB iterations are accelerated by biorthogonalization, when B_2 in (1) is replaced by $(B_2 B_1)^{-1} B_2$.

From Tables 3 and 4 one can see that the errors are significantly smaller, when the auxiliary boundary $\partial\tilde{\Omega}$ (with $\delta = 8h$) is used. Here the BiCGSTAB iterations are accelerated by a multigrid strategy.

Step h	n/m	Iters.	C.time[s]	$\delta_{L^2(\Omega)}$	$\delta_{H^1(\Omega)}$	$\delta_{L^2(\partial\Omega)}$
1/128	16641/40	15	0.188	2.3637e-002	2.1633e+000	9.0989e-002
1/256	66049/70	24	1.36	1.2831e-002	1.4736e+000	4.9341e-002
1/512	263169/124	32	14.24	7.1820e-003	9.9318e-001	2.7571e-002
1/1024	1050625/220	46	93.11	3.9157e-003	7.1732e-001	1.5345e-002

Tab. 1: *Convergence* ***without*** $\partial\tilde{\Omega}$.

Step h	n/m	Iters.	C.time[s]	$\delta_{L^2(\Omega)}$	$\delta_{H^1(\Omega)}$	$\delta_{L^2(\partial\Omega)}$
1/128	16641/40	9	0.11	2.3386e-002	2.1550e+000	8.9462e-002
1/256	66049/70	12	0.735	1.2808e-002	1.4734e+000	4.9238e-002
1/512	263169/124	22	10.03	7.1183e-003	9.9261e-001	2.7336e-002
1/1024	1050625/220	30	60.23	3.8315e-003	7.1694e-001	1.5064e-002

Tab. 2: *Convergence* ***without*** $\partial\tilde{\Omega}$*, biorthogonalization.*

Step h	n/m	Iters.	C.time[s]	$\delta_{L^2(\Omega)}$	$\delta_{H^1(\Omega)}$	$\delta_{L^2(\partial\Omega)}$
1/128	16641/40	25	0.281	5.3431e-004	2.4639e-002	1.8577e-003
1/256	66049/70	39	2.218	1.4133e-004	1.2407e-002	5.7929e-004
1/512	263169/124	99	42.22	4.3848e-005	7.0675e-003	2.2314e-004
1/1024	1050625/220	200	371.5	1.2541e-005	3.6767e-003	6.9726e-005

Tab. 3: *Convergence* ***with*** $\partial\tilde{\Omega}$.

Step h	n/m	Iters.	C.time[s]	$\delta_{L^2(\Omega)}$	$\delta_{H^1(\Omega)}$	$\delta_{L^2(\partial\Omega)}$
1/128	16641/40	16	0.266	7.3218e-004	2.8843e-002	2.3947e-003
1/256	66049/68	20	1.39	1.3533e-004	1.1927e-002	5.0063e-004
1/512	263169/124	33	16.37	3.3349e-005	5.9480e-003	1.4539e-004
1/1024	1050625/220	38	94.25	1.3469e-005	3.7054e-003	5.2209e-005

Tab. 4: *Convergence* ***with*** $\partial\tilde{\Omega}$*, multigrid.*

References

[1] M. Benzi, G.H. Golub, J. Liesen: *Numerical solution of saddle point systems.* Acta Numerica, 2005, 1–137.

[2] C. Farhat, J. Mandel, F.X. Roux: *Optimal convergence properties of the FETI domain decomposition method.* Comput. Methods Appl. Mech. Engrg. **115**, 1994, 365–385.

[3] R. Glowinski, T. Pan, J. Periaux: *A fictitious domain method for Dirichlet problem and applications.* Comput. Meth. Appl. Mech. Engrg. **111**, 1994, 283–303.

[4] J. Haslinger, T. Kozubek, R. Kučera, K. Kunisch, G. Peichl: *Fictitious domain approach for solving boundary value problems with the mixed Dirichlet-Neumann boundary conditions.* In preparation, 2006.

[5] R. Kučera, T. Kozubek, J. Haslinger: *Projected Schur complement method for solving nonsymmetric saddle point systems arising in fictitious domain approach.* Submitted to Appl. Math., 2006.

[6] R. Kučera: *Complexity of an algorithm for solving saddle-point systems with singular blocks arising in wavelet-Galerkin discretizations.* Appl. Math. **50**, 3, 2005, 291–308.

[7] H.A. Van der Vorst: *BiCGSTAB: a fast and smoothly converging variant of BiCG for solution of nonsymmetric linear systems.* SIAM J. Sci. Statist. Comput. **13**, 1992, 631–644.

THE DISCONTINUOUS GALERKIN METHOD FOR LOW-MACH FLOWS[*]

Václav Kučera

1. Introduction

Our goal is to develop a numerical technique allowing the solution of compressible flow with a wide range of the Mach number. This technique is based on the *discontinuous Galerkin finite element method* (DGFEM), which employs piecewise polynomial approximations without any requirement on the continuity on interfaces between neighbouring elements. The DGFEM space semidiscretization is combined with a semi-implicit time discretization (Section 2.) and a special treatment of boundary conditions (Section 3.). In this way we obtain a numerical scheme requiring the solution of only one linear system on each time level. This scheme is successfully tested on flows with Mach numbers as low as 10^{-4}. As for the transonic case it is necessary to avoid the Gibbs phenomenon manifested by spurious overshoots and undershoots in computed quantities near discontinuities and steep gradients. These phenomena do not occur in low Mach number regimes, however in the transonic case they cause instabilities in the semi-implicit solution. Here we present a possibility how to treat this problem (Section 4.). Section 5. presents computational results for small Mach numbers as well as transonic flow.

2. Discretization

We discretize the Euler equations in the conservative form:

$$\begin{aligned} &\frac{\partial \mathbf{w}}{\partial t} + \sum_{s=1}^{2} \frac{\partial \boldsymbol{f}_s(\mathbf{w})}{\partial x_s} = 0 \quad \text{in } \Omega \times (0,T), \\ &\mathbf{w} = (\rho, \rho v_1, \rho v_2, e)^{\mathrm{T}} \in I\!R^4, \\ &\boldsymbol{f}_i(w) = (\rho v_i, \rho v_1 v_i + \delta_{1i} p, \rho v_2 v_i + \delta_{2i} p, (e+p) v_i)^{\mathrm{T}}. \end{aligned} \tag{1}$$

Let $\mathcal{T}_h$ be a partition of $\overline{\Omega}$ into a finite number of triangles with a numbering I. Let $\Gamma_{ij} = \partial K_i \cap \partial K_j$ be a common edge of two triangles. The DGFEM uses the finite element space of discontinuous piecewise polynomial functions.

$$S_h = S^{p,-1}(\Omega, \mathcal{T}_h) = \{v; v|_K \in P_p(K)\ \forall K \in \mathcal{T}_h\}, \tag{2}$$

[*]This work is a part of the research project No. MSM 0021620839 of the Ministry of Education of the Czech Republic. The research is partly supported by the Grant No. 6/2005/R of the Grant Agency of Charles University.

where $P_p(K)$ is the space of all polynomials on K of degree $\leq p$. In the current implementation, P^0, P^1 and P^2 approximations are used along with 5^{th} order Gaussian quadrature rules on elements and edges.

We multiply (1) by a test function $\boldsymbol{\varphi} \in [S_h]^4$ and integrate over $K_i \in \mathcal{T}_h$. With the aid of Green's theorem and summing over all $i \in I$, we obtain

$$\begin{aligned}\frac{d}{dt} \sum_{K_i \in \mathcal{T}_h} \int_{K_i} \mathbf{w} \cdot \boldsymbol{\varphi} \, dx &= \\ &= \underbrace{\sum_{K_i \in \mathcal{T}_h} \int_{K_i} \sum_{s=1}^{2} \boldsymbol{f}_s(\mathbf{w}) \cdot \frac{\partial \boldsymbol{\varphi}}{\partial x_s} \, dx}_{T_1} + \underbrace{\sum_{i \in I} \sum_{j \in S(i)} \int_{\Gamma_{ij}} \mathbf{H}(\mathbf{w}|_{\Gamma_{ij}}, \mathbf{w}|_{\Gamma_{ji}}, \mathbf{n}_{ij}) \cdot \boldsymbol{\varphi} \, dS}_{T_2} . \end{aligned} \tag{3}$$

In the term T_2, we have incorporated an approximation using a numerical flux $\mathbf{H}$, as known from the finite volume method. The approximate solution is defined as $\mathbf{w}_h \in [S_h]^4$ such that (3) holds for all $\boldsymbol{\varphi}_h \in [S_h]^4$.

Scheme (3) represents a system of ordinary differential equations, which we must discretize with respect to time. Explicit time discretization is however undesirable due to a *CFL*-like condition, which limits the time step proportionally to the Mach number. A fully implicit scheme presents us with the task of solving a large nonlinear system on each time level. We therefore use the method presented in [4]. A forward Euler method is used and the nonlinear terms in the scheme are linearized. The resulting systems are solved using block-Jacobi preconditioned GMRES or the UMFPACK direct solver.

The term T_1 in (3) is linearized using homogeneity of the Euler fluxes:

$$T_1 \approx \sum_{i \in I} \int_{K_i} \sum_{s=1}^{2} \frac{D\boldsymbol{f}_s(\mathbf{w}_h^k)}{D\mathbf{w}} \mathbf{w}_h^{k+1} \cdot \frac{\partial \boldsymbol{\varphi}_h}{\partial x_s} \, dx. \tag{4}$$

As for the term T_2, the Vijayasundaram numerical flux is chosen, since it is suitable for linearization. This numerical flux has the form

$$\mathbf{H}_{VS}(\mathbf{w}_L, \mathbf{w}_R, \mathbf{n}) = \mathbb{P}^+ \left(\frac{\mathbf{w}_L + \mathbf{w}_R}{2}, \mathbf{n} \right) \mathbf{w}_L + \mathbb{P}^- \left(\frac{\mathbf{w}_L + \mathbf{w}_R}{2}, \mathbf{n} \right) \mathbf{w}_R. \tag{5}$$

3. Boundary conditions

The choice of appropriate boundary conditions is a delicate problem which plays a key role in the presented algorithm. Boundary conditions are incorporated into the DGFEM, as in the finite volume method, via the choice of $H(\mathbf{w}_L, \mathbf{w}_R, \mathbf{n})$ or $\mathbf{w}_R = \mathbf{w}|_{\Gamma_{ji}}$ for boundary edges. In the case of impermeable walls, we prescribe the *no-stick* condition $\mathbf{v} \cdot \mathbf{n}$. The situation is much more problematic on the inlet and outlet - standard boundary conditions reflect acoustic effects coming from the inside of Ω. This behavior is nonphysical and the reflected interfering density and pressure waves corrupt the solution in the low-Mach number case. To cure this disease new *characteristic based* boundary conditions are derived, which reflect the hyperbolic character of the Euler equations and are transparent to acoustic phenomena. These boundary conditions are a key ingredient in low-Mach calculations.

Using the rotational invariance and homogeneity we write the Euler equations in the nonconservative form

$$\frac{\partial \mathbf{q}}{\partial t} + \mathbb{A}_1(\mathbf{q})\frac{\partial \mathbf{q}}{\partial \tilde{x}_1} = 0, \tag{6}$$

where $\mathbf{q} = \mathbb{Q}(\mathbf{n})\mathbf{w}$ and $\mathbb{Q}(\mathbf{n})$ is a standard 4×4 rotational matrix (see [1]). We linearize this system around the state $\mathbf{q}_i = \mathbb{Q}(\mathbf{n})\mathbf{w}_i$ and obtain a linear system. The goal is to choose the boundary state $\mathbf{q}_j$ in such a way that this initial-boundary problem is well posed, i.e. has a unique solution. This linearized system has a solution which can be written explicitly using the method of characteristics. We shall take some state $\mathbf{q}_j^0 = \mathbb{Q}(\mathbf{n})\mathbf{w}_j^0$. The state $\mathbf{w}_j^0$ is the state vector of the far-field flow. We calculate the eigenvectors $\mathbf{r}_s, s = 1, \ldots, 4$ of the matrix $\mathbb{A}_1(\mathbf{q}_i)$, arrange them as columns in the matrix $\mathbb{T}$ and calculate $\mathbb{T}^{-1}$ (explicit formulae can be found in [1]). We calculate

$$\boldsymbol{\beta} = \mathbb{T}^{-1}\mathbf{q}_i, \quad \boldsymbol{\alpha} = \mathbb{T}^{-1}\mathbf{q}_j^0. \tag{7}$$

Now we calculate the state $\mathbf{q}_j$ according to the presented process:

$$\mathbf{q}_j := \sum_{s=1}^{4} \gamma_s \mathbf{r}_s = \mathbb{T}\boldsymbol{\gamma}, \quad \gamma_s = \begin{cases} \alpha_s, & \lambda_s \geq 0, \\ \beta_s, & \lambda_s < 0 \end{cases} \tag{8}$$

and $\lambda_s,\ s = 1, \ldots, 4$ are eigenvalues of $\mathbb{A}_1(\mathbf{q}_i)$. Finally the sought boundary state is $\mathbf{w}_j = \mathbb{Q}^{-1}(\mathbf{n})\mathbf{q}_j$. Since we have respected the hyperbolic character of the Euler equations, these boundary conditions seem to give a natural choice for the boundary state $\mathbf{w}_j$.

4. Shock capturing

Our approach is based on the discontinuity indicator $g(i)$ proposed in [2] defined by

$$g(i) = \int_{\partial K_i} [\rho_h^k]^2 \, dS \big/ (h_{K_i}|K_i|^{3/4}), \quad K_i \in \mathcal{T}_h. \tag{9}$$

We define a discrete shock indicator on the basis of (9):

$$G(i) = \begin{cases} 0, & g(i) < 1, \\ 1, & g(i) \geq 1. \end{cases}, \quad K_i \in \mathcal{T}_h.$$

To the left-hand side of (3) we add the form $\beta(\boldsymbol{w}_h, \boldsymbol{\varphi}_h)$ defined by

$$\beta(\boldsymbol{w}, \boldsymbol{\varphi}) = C \sum_{i \in I} h_{K_i} G(i) \int_{K_i} \nabla \boldsymbol{w} \cdot \nabla \boldsymbol{\varphi} \, dx, \tag{10}$$

where $C \approx 1$. This artificial term represents a discrete Laplacian with zero Neumann boundary conditions on each element, thus forcing the solution to a piecewise constant function. The stabilization form β is treated implicitly (with $G(i)$ computed from $\boldsymbol{w}_h^k$).

This form limits the order of accuracy on each element lying on a discontinuity. However, it appears that on finely refined grids this is insufficient. Therefore, we propose to augment the left-hand side of (3) by adding the form $J(\boldsymbol{w}_h, \boldsymbol{\varphi}_h)$ defined as

$$J(\boldsymbol{w}, \boldsymbol{\varphi}) = \varepsilon \sum_{i \in I} \sum_{j \in s(i)} \frac{1}{2}\big(G(i) + G(j)\big) \int_{\Gamma_{ij}} [\boldsymbol{w}] \cdot [\boldsymbol{\varphi}] \, dS, \tag{11}$$

where $\varepsilon \approx 1$ and $[u]|_{\Gamma_{ij}} = u_{ij} - u_{ji}$ is the jump on Γ_{ij} of a function $u \in S_h$. In this way we penalize inter-element jumps in the vicinity of the shock wave. This form can be treated implicitly, similarly as $\beta(\boldsymbol{w}, \boldsymbol{\varphi})$.

5. Numerical examples

In this section we present the solution of some test problems in order to demonstrate the accuracy and robustness of the proposed method. In all examples quadratic elements ($r = 2$) were used for obtaining steady state solutions for "$t \to \infty$". The number of time steps necessary to obtain the steady state solution in the following test cases is approximately 100-200.

1) Irrotational flow past a symmetric Joukowski airfoil First we consider flow past a symmetric Joukowski profile with zero angle of attack. Using the complex function method from [3], we can obtain the exact solution of incompressible inviscid irrotational flow for this test case. We assume that the far field Mach number of compressible flow $M_\infty = 0.0001$. Figure 1 shows a detail near the profile of the

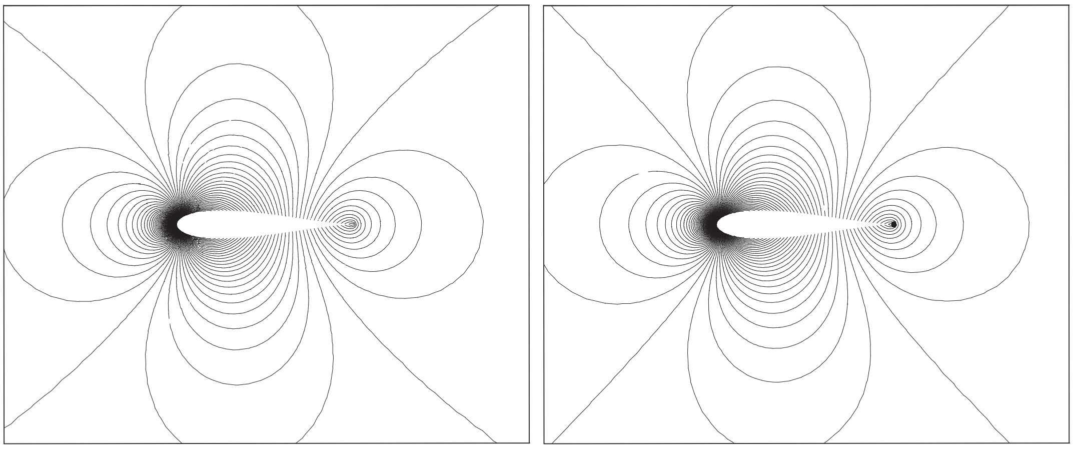


Fig. 1: *Velocity isolines for the approximate solution of compressible flow (left) and for the exact solution of incompressible flow (right).*

velocity isolines for the approximate solution of compressible flow and for the exact solution of incompressible flow, respectively. The mesh was formed by 4103 triangular elements.

2) Irrotational flow past a nonsymmetric Joukowski airfoil The second example deals with a similar problem to the preceding symmetric case.we present flow past a nonsymmetric Joukowski profile with zero angle of attack. Again, using the complex function method we can obtain the exact solution in the case of a nonsymmetric Joukowski profile. The far field Mach number of is again $M_\infty = 0.0001$. Figure 2 shows a detail near the profile of the velocity isolines for the approximate solution of compressible flow and for the exact solution of incompressible flow, respectively. Figure 3 shows a comparison of the velocity distribution along the profile surface for the computed and exact solution. The mesh was formed by 5418 triangular elements.

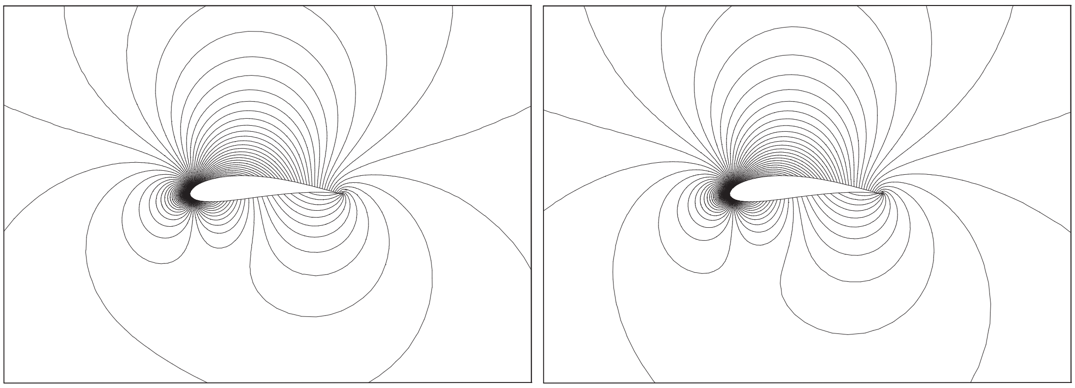


Fig. 2: *Velocity isolines for the approximate solution of compressible flow (left) and for the exact solution of incompressible flow (right).*

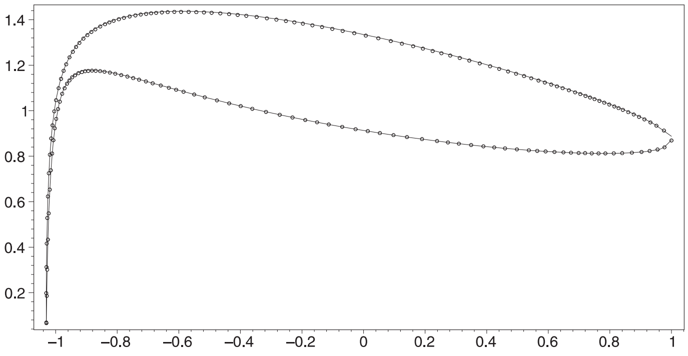


Fig. 3: *Velocity distribution along the profile surface.* ○○○ *– exact solution of incompressible flow,* —– *– approximate solution of compressible flow.*

3) Transonic flow The performance of shock capturing terms from Section 4. is tested on the GAMM channel with a 10% circular bump and the inlet Mach number equal to 0.67. In this case a conspicuous shock wave is developed. Figure 4 shows Mach number isolines and entropy isolines computed by the presented scheme. One can see that this scheme yields the entropy production on the shock wave only, which is correct from the physical point of view. The stabilization parameters in were chosen $\nu_1 = \nu_2 = 0.2$. The mesh was formed by 7753 triangular elements.

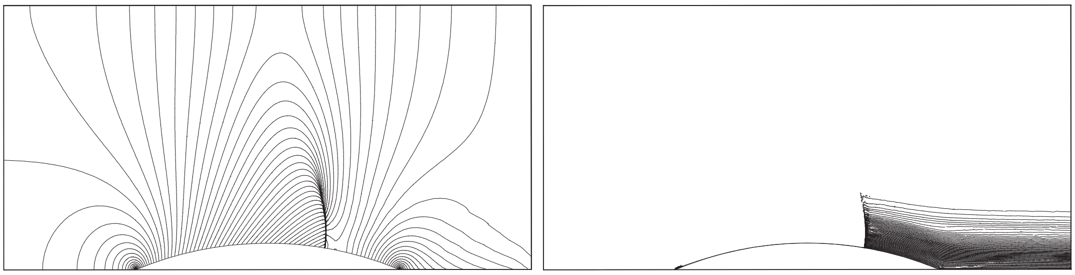


Fig. 4: *GAMM channel transonic flow, Mach number (top) and entropy (bottom) isolines.*

References

[1] Feistauer M., Felcman J., Straškraba I.: *Mathematical and Computational Methods forCompressible Flow*, Oxford University Press, Oxford, 2003.

[2] V. Dolejší, M. Feistauer and C. Schwab. *On some aspects of the discontinuous Galerkin finite element method for conservation laws.* Math. Comput. Simul. **61**, 2003, 333–346.

[3] M. Feistauer. *Mathematical Methods in Fluid Dynamics.* Longman Scientific & Technical, Harlow, 1993.

[4] Dolejší V., Feistauer M.: *A Semi-Implicit Discontinuous Galerkin Finite Element Method for the Numerical Solution of Inviscid Compressible Flows*, The Preprint Series of the School of Mathematics MATH-KNM-2003/1, Charles University, Prague.

ARBITRARY LAGRANGIAN-EULERIAN (ALE) METHODS IN COMPRESSIBLE FLUID DYNAMICS*

Milan Kuchařík, Richard Liska, Pavel Váchal, Mikhail Shashkov

Abstract

The aim of this paper is to present an Arbitrary Lagrangian-Eulerian (ALE [1]) code for simulation of problems in compressible fluid dynamics and plasma physics including heat conduction and laser absorption, in both Cartesian and cylindrical geometries. Various techniques are utilized for mesh adaptation (rezoning), including Winslow smoothing [2], three-step untangling [3] and Reference Jacobian method [4, 5]. For conservative transfer (remapping) of variables onto the rezoned mesh, linear interpolation with a posteriori repairs is used by default. Simulation of high velocity impact, for which pure Lagrangian method fails, proves the usefulness of ALE approach.

1. Introduction

The Arbitrary Lagrangian-Eulerian (ALE) method [1] is a popular tool for simulation of continuum mechanics problems with large shear deformation such as fluid flow and metal forming. Compared to pure Eulerian methods, it is also better suited for moving boundaries and large volume changes of the computational domain, appearing in simulations of laser-plasma interactions and inertial confinement fusion.

The ALE algorithm consists of a classical Lagrangian step in which the mesh moves along with the modeled material, a rezone step in which the mesh is modified to preserve good quality through the computation, and a remapping step in which the solution is conservatively transferred from the old mesh to the new, rezoned one. We present new efficient techniques for the rezoning and remapping stages of the ALE framework and demonstrate some of their properties on a real physical simulation of high velocity impact.

Note that by the ALE method we understand the variation of Lagrangian hydrodynamics which avoids Lagrangian mesh distortion (arising in some problems involving e.g. shear flows) by rezoning and remapping. Another method, unfortunately also called ALE, uses a mesh smoothly moving in a predefined way, typically determined by moving boundaries rather than by fluid motion.

Details on implementation of particular procedures and on the physical background can be found in [6].

*Part of this work was funded by the U.S. Department of Energy under contract W-7405-ENG-36, the DOE Office for Science's ASCR program in Applied Mathematical Sciences and the ASC program. The first three authors have been partly supported by the Czech Science Foundation GAČR 202/03/H162 and Czech Ministry of Education grants MSM 6840770010, MSM 6840770022 and LC528.

2. The Lagrangian step

In pure Lagrangian computation, each mesh cell can be considered as a particle of the fluid, so that the mesh moves along with the simulated problem, with no mass flux between the cells. Euler equations for compressible fluid flow with heat conductivity and laser absorption in Lagrangian coordinates read

$$\frac{1}{\rho}\,\frac{\mathrm{d}\,\rho}{\mathrm{d}\,t} = -\nabla\cdot\vec{v}, \qquad \rho\,\frac{\mathrm{d}\,\vec{v}}{\mathrm{d}\,t} = -\nabla p, \qquad \frac{\mathrm{d}\,\vec{x}}{\mathrm{d}\,t} = \vec{v}, \tag{1a}$$

$$\rho\,\frac{\mathrm{d}\,\varepsilon}{\mathrm{d}\,t} = -p\;\nabla\cdot\vec{v} + \nabla\cdot(\kappa\nabla T) - C_a\nabla\cdot\vec{I}, \tag{1b}$$

where total Lagrangian time derivatives include convective terms: $\frac{\mathrm{d}}{\mathrm{d}\,t} = \frac{\partial}{\partial\,t} + \vec{v}\cdot\nabla$. Scalar quantities (density ρ, pressure p, specific internal energy ε and temperature T) are approximated in mesh cells, while vectors (position $\vec{x}$ and velocity $\vec{v}$) are related to the nodes. To complete the system, one has to supply also the equation of state (EOS). For the ideal polytropic gas, the EOS is $p = (\gamma - 1)\varepsilon\rho$. For other materials, more sophisticated formulas are advised, e.g. the Quotidian EOS [7]. The hyperbolic Lagrangian system is numerically treated by compatible method [8, 9] conserving total energy. Several types of artificial viscosity are incorporated into the difference scheme, such as bulk viscosity, edge viscosity, etc. [6]. Laser absorption is taken into account by the last term in the energy equation (1b).

The system is split into hyperbolic and parabolic parts. The parabolic part

$$\frac{\mathrm{d}\,T}{\mathrm{d}\,t} - \nabla\cdot(\kappa\nabla T) = 0$$

of the energy equation is solved separately by a scheme fully implicit in time, which allows the choice of timestep equal to that of the hyperbolic system. A discretization of operators div and grad by a mimetic method [10] leads to a system with a symmetric and positive definite matrix, which is then solved by conjugate gradient method.

3. Mesh adaptation (rezoning)

During the rezoning process, the quality of strongly deformed parts of the mesh must be improved, so that the computation can continue with desired precision. However, doing more changes than necessary could lead to loss of valuable simulation information gathered so far. If the mesh is really strongly distorted, e.g. containing the “hourglass-shaped” (⋈) quadrilateral cells, one first needs to untangle it, that is to fix all the fully or partly inverted elements. An efficient method to do this is the three-step algorithm [3], combining direct node placement based on geometrical considerations with numerical optimization of a quadratic functional which serves as a local mesh quality indicator. Another option is to prevent evolution of strong deformations (tangling) by regular use of a less expensive rezoning technique, such as the simple Winslow approach [2], where new node positions are given by

$$\vec{x}^{k+1}_{i,j} = \frac{1}{2\,(\alpha^k+\gamma^k)}\,\Big(\alpha^k\,(\vec{x}^k_{i,j+1}+\vec{x}^k_{i,j-1})+\gamma^k\,(\vec{x}^k_{i+1,j}+\vec{x}^k_{i-1,j})- \\ -\frac{1}{2}\,\beta^k\,(\vec{x}^k_{i+1,j+1}-\vec{x}^k_{i-1,j+1}+\vec{x}^k_{i-1,j-1}-\vec{x}^k_{i+1,j-1})\Big) \quad (2)$$

with coefficients $\alpha^k = x^2_\xi + y^2_\xi$, $\beta^k = x_\xi\;x_\eta + y_\xi\;y_\eta$, $\gamma^k = x^2_\eta + y^2_\eta$, where z_ξ, z_η denote finite differences in logical, index coordinates $z_\xi = (z_{i+1,j} - z_{i-1,j})/2, z_\eta = (z_{i,j+1} - z_{i,j-1})/2$. A more sophisticated method is based on the local parametrization and optimization of the Reference Jacobian matrix [4, 5]. First, each node is assigned a virtual reference position $\vec{x}^{(R)}$ by optimization of a local mesh quality estimator in its neighborhood. In particular, in N dimensions, for node V one minimizes the functional

$$Q_V = \sum_{T\in T_V} \|\mathbf{J}_{V,T}\|\cdot\|\mathbf{J^{-1}}_{V,T}\|,$$

which is a sum of condition numbers of the Jacobi mapping matrices

$$\mathbf{J}_{V,T} = [e_{V,1}, e_{V,2}, \ldots, e_{V,N}]$$

given by edges $e_{V,k} = \vec{x}_k - \vec{x}_V$ forming a virtual simplex in the N-dimensional space. The sum is taken over all simplices T sharing node V as a vertex. Then, global optimization is used to find a mesh of good quality, with edges as close as possible to their reference counterparts. This is done by minimization of the functional

$$F_{RJ} = \sum_V \sum_{T\in T_V} \frac{\|\mathbf{J}_{V,T}(x) - \mathbf{J^{(R)}}_{V,T}\|}{\|\mathbf{J^{(R)}}_{V,T}\|},$$

where the sum is taken over all mesh vertices V and the reference Jacobian matrix is defined as

$$\mathbf{J^{(R)}}_{V,T} = \left[e^{(R)}_{V,1}, e^{(R)}_{V,2}, \ldots, e^{(R)}_{V,N}\right], \qquad\qquad e^{(R)}_{V,k} = \vec{x}_k - \vec{x}^{(R)}_V.$$

Both functionals are optimized using the conjugate gradient method, which is well suited for problems with large number of parameters.

The input mesh for this procedure must not contain inverted elements (i.e. simplices with negative volume in the sense of original orientation). Therefore, strongly distorted meshes must be preprocessed by an untangling procedure, e.g. the three-step method [3] mentioned above.

4. Conservative transfer of solution (remapping)

Once the mesh is adapted (rezoned), the discrete values of conserved variables must be transferred (remapped) from the old mesh to this new, rezoned one. This procedure is required to be conservative for mass, each component of momentum, and total energy and must preserve monotonicity (or at least local bounds) for density, velocity and specific internal energy. The remapping should be as accurate as possible. Exact transfer from the old mesh to the new one is required for linear functions. All this is achieved by a method which first interpolates discrete values by a piecewise linear function, then integrates it over swept regions and finally corrects the possibly created overshoots or undershoots by redistribution of these into the neighboring cells (so-called Repair) [11, 12].

Other techniques enforce all imposed requirements already during the remapping process, with no need of a posteriori repair. Many of them combine low-order intercell fluxes (which preserve local bounds by default) with some portion of higher-order (generally unconstrained) fluxes. An example called Flux-Corrected Remapping (FCR) is described in [13].

5. Numerical example

As a practical example, we show a simulation inspired by an experiment performed recently at the Prague Asterix Laser System (PALS) facility: a laser-irradiated aluminum disc ablatively accelerates and strikes a massive aluminum target [14, 15]. Here we focus on the second part, that is on disc impact. The setup is as in Fig. 1(a)

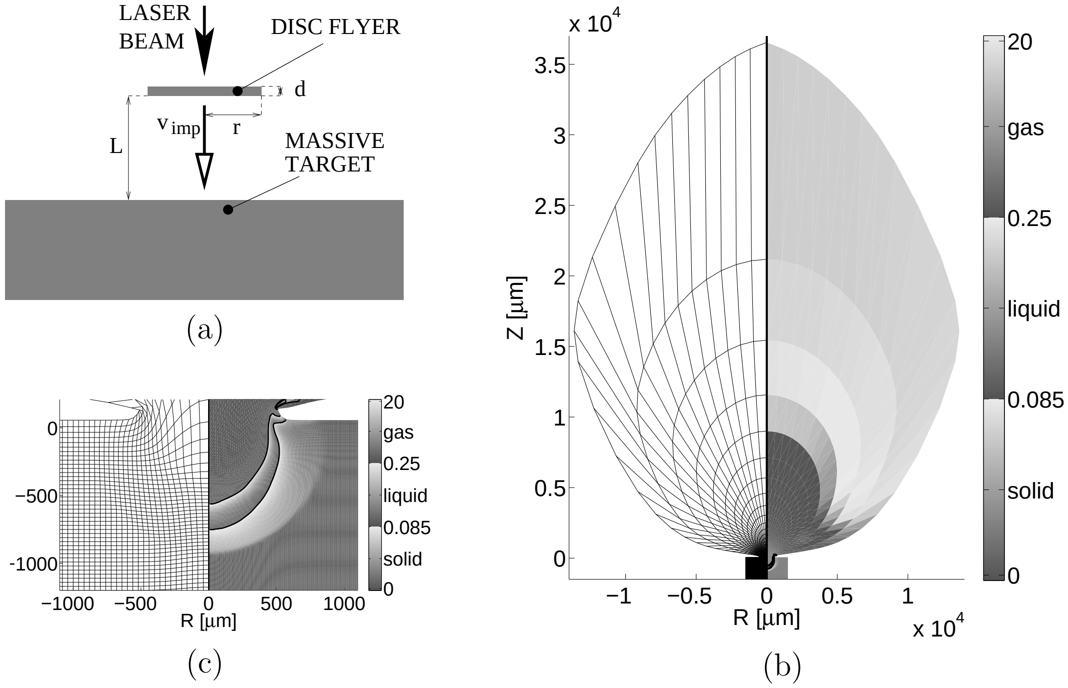


Fig. 1: *Disc impact problem. Experiment setup (a) and temperature at* 80 ns*: whole domain with hot plasma corona (b), detail of crater evolving in the target (c). Only every fourth layer of edges is shown in (c). Solid, liquid and gas phases are shown in separate colormaps.*

with the following parameters: a 400 ps laser pulse with energy 240 J operating in the 3rd harmonic with radius of focal spot on target $r_f = 125\,\mu$m, irradiates a $d = 11\,\mu$m thick disc with radius $r = 150\,\mu$m, located $L = 200\,\mu$m above the target. The disc is ablatively accelerated up to the impact velocity $v_{\mathrm{imp}} = 134\,$km/s and hits the target. Simulation starts at the moment of impact. Pure Lagrangian computation fails very soon (at approximately $t \approx 0.5\,$ns) because of fatal mesh distortion, while the ALE simulation preserves sufficient mesh quality for the computation to continue. In particular, EOS for ideal gas was used, mesh rezoning was performed by Winslow smoothing (2) and remapping by linear interpolation with Repair. The flyer starts to sink into the target, material of both the flyer and the target are compressed, heated and evaporated. Part of the hot material is ionized, ablated and forms an expanding plasma corona, shown at $t = 80\,$ns in Fig. 1(b). Shock wave is propagating into the target, continuing to melt and evaporate its material, see Fig. 1(c), where only every fourth mesh edge in each direction is shown, so that each quadrilateral corresponds to sixteen real cells. Solid, liquid and gas phases are shown by different colormaps in grayscale. In all performed tests, size and shape of the crater approximated the experimental data with reasonable precision.

References

[1] C.W. Hirt, A.A. Amsden, and J.L. Cook: *An arbitrary Lagrangian-Eulerian computing method for all flow speeds.* J. Comp. Phys. **14**, 1974, 227–253. Reprinted in J. Comp. Phys. **135**, 1997, 203–216.

[2] A.M. Winslow: *Equipotential zoning of two-dimensional meshes.* Technical Report UCRL-7312, Lawrence Livermore National Laboratory, 1963.

[3] P. Váchal, R. Garimella, and M. Shashkov: *Untangling of 2D meshes in ALE simulations.* J. Comp. Phys. **196**, 2004, 627–644.

[4] P.M. Knupp, L.G. Margolin, and M.J. Shashkov: *Reference Jacobian optimization-based rezone strategies for Arbitrary Lagrangian-Eulerian methods.* J. Comp. Phys. **176**, 2002, 93–128.

[5] V. Dyadechko, R. Garimella, and M. Shashkov: *Reference Jacobian rezoning strategy for Arbitrary Lagrangian-Eulerian methods on polyhedral grids.* In: Proceedings of the Thirteenth Meshing Roundtable, Williamsburg, VA, September 2004, Sandia National Laboratories, 2004.

[6] M. Kuchařík: *Arbitrary Lagrangian-Eulerian (ALE) methods in plasma physics.* PhD Dissertation. Czech Technical University in Prague, 2006.

[7] R.M. More, K. Warren, D. Young, and G. Zimmerman: *A new quotidian equation of state (QEOS) for hot dense matter.* Physics of Fluids **31**, 1988, 3059–3078.

[8] E.J. Caramana, D.E. Burton, M.J. Shashkov, and P.P. Whalen: *The construction of compatible hydrodynamics algorithms utilizing conservation of total energy.* J. Comp. Phys. **146**, 1998, 227–262.

[9] E.J. Caramana and M.J. Shashkov: *Elimination of artificial grid distortion and hourglass-type motions by means of Lagrangian subzonal masses and pressures.* J. Comp. Phys. **142**, 1998, 521–561.

[10] M. Shashkov and S. Steinberg: *Solving diffusion equations with rough coefficients in rough grids.* J. Comp. Phys. **129**, 1996, 383–405.

[11] M. Kuchařík, M. Shashkov, and B. Wendroff: *An efficient linearity-and-bound-preserving remapping method.* J. Comp. Phys. **188**, 2003, 462–471.

[12] R. Garimella, M. Kuchařík, and M. Shashkov: *An efficient linearity and bound preserving conservative interpolation (remapping) on polyhedral meshes.* Comp. & Fluids, 2006, in press.

[13] P. Váchal and R. Liska: *Sequential flux-corrected remapping for ALE methods.* In: Numerical Mathematics and Advanced Applications. ENUMATH 2005, Springer, 2006, 671–679.

[14] S. Borodziuk, A. Kasperczuk, T. Pisarczyk, K. Rohlena, J. Ullschmied, M. Kalal, J. Limpouch and P. Pisarczyk: *Application of laser simulation method for the analysis of crater formation experiment on PALS laser.* Czechoslovak J. Phys. **53**, 2003, 799–810.

[15] M. Kalal, S. Borodziuk, N.N. Demchenko, S.Yu. Guskov, K. Jungwirth, A. Kasperczuk, V.N. Kondrashov, B. Kralikova, E. Krousky, J. Limpouch, K. Masek, P. Pisarczyk, T. Pisarczyk, M. Pfeifer, K. Rohlena, V.B. Rozanov, J. Skala and J. Ullschmied: *High power laser interaction with single and double layer targets.* In: Proceedings of XXVIII ECLIM, 2004, 249–260.

SOLUTION OF TIME-DEPENDENT CONVECTION-DIFFUSION EQUATIONS WITH THE AID OF HIGHER ORDER ADAPTIVE METHODS WITH RESPECT TO SPACE AND TIME*

Pavel Kůs, Vít Dolejší

1. Introduction

This work deals with the solution of a scalar nonlinear convection–diffusion equation which is a model problem for a numerical simulation of viscous compressible flows. A semi-discretization with respect to the space coordinates, which is carried out with the aid of the discontinuous Galerkin method, yields a system of ordinary differential equations (ODE). Our aim is to develop and implement an efficient adaptive numerical scheme for the solution of this ODE system. We derive two stable multi-step methods of the same order of accuracy and from a difference of both approximate solutions, we estimate a local discretization error with respect to the time. Then we choose the time step in such a way, that local error is approximately equal to a given tolerance. Several numerical simulations were carried out to demonstrate the efficiency of the method.

2. Discontinuous Galerkin method

We consider the following unsteady nonlinear convection–diffusion problem: Find $u : Q_T \to I\!R$ such that

$$\frac{\partial u}{\partial t} + \sum_{s=1}^{d} \frac{\partial f_s(u)}{\partial x_s} = \varepsilon\, \Delta u \ + g \quad \text{in } Q_T, \tag{1}$$

$$u\,\big|_{\partial\Omega\times(0,T)} = u_D, \tag{2}$$

$$u(x,0) = u^0(x), \quad x \in \Omega. \tag{3}$$

Similarly as in the finite element method, we introduce a weak solution u of the problem

$$\frac{\mathrm{d}}{\mathrm{d}t}(u(t),v) + b(u(t),v) + a(u(t),v) = (g(t),v) \qquad \forall v \in H_0^1(\Omega), \tag{4}$$

*This work is a part of the research project MSM 0021620839 financed by the Ministry of Education of the Czech Republic and was partly supported by the Grant No. 316/2006/B-MAT/MFF of the Grant Agency of Charles University Prague.

where $(\cdot,\cdot)$ denotes the L^2-scalar product, $a(\cdot,\cdot)$ is a linear form representing the diffusive term and $b(\cdot,\cdot)$ is a nonlinear form representing the convective term. We also consider appropriate representation of initial and boundary conditions. As in the classical finite element method we use triangulation of domain Ω and a piecewise polynomial discontinuous approximation. More general, even non-convex elements with the hanging nodes are allowed. The approximate solution is sought in a space of piecewise polynomial but discontinuous functions S_h. In order to replace the inter-element continuity, we add some stabilization terms into formulation of a discrete problem. The convective term is approximated with the aid of a numerical flux, known from the finite volume method. We receive the space semidiscretization

$$\left(\frac{\partial u_h(t)}{\partial t}, \varphi_h\right) + b_h(u_h(t), \varphi_h) + a_h(u_h(t), \varphi_h) = 0 \quad \forall\, \varphi_h \in S_h, \tag{5}$$

where $a_h(\cdot,\cdot)$ and $b_h(\cdot,\cdot)$ are the discrete variants of the forms $a(\cdot,\cdot)$ and $b(\cdot,\cdot)$, respectively. For more details, see [1], [2]. The relation (5) represents a system of ordinary differential equations, which must be solved by a suitable method.

3. BDF2 method

The system (5) is stiff, so we have to use an implicit method, such as backward difference formulae (BDF). In contrast to [4] where a combination of explicit and implicit schemes was employed we introduce two implicit schemes of the same order of accuracy. Using this pair of methods, we obtain two solutions and from their difference we estimate the local discretization error.

3.1. Derivation of the method

Now we shall briefly describe derivation of two n-step methods BDF2a and BDF2b for solution of a system of ordinary differential equations with an unknown function $y : (0,T) \to I\!R^m$.

$$\frac{\mathrm{d}y(t)}{\mathrm{d}t} = F(t,y), \qquad y(0) = y^0\,, \tag{6}$$

where $y^0 \in I\!R^m$ and $F : (0,T) \times I\!R^m \to I\!R^m$. Let us denote by $0 = t_0 < t_1 < t_2 < \cdots < t_r = T$ the partition of the interval $(0,T)$, $\tau_k \equiv t_k - t_{k-1},\ k = 1,\dots,r$, $\theta_k = \tau_k/\tau_{k-1},\ k = 1,\dots,r$. Moreover, let y_k denote approximate value of solution $y(t_k),\ k = 0,\dots,r$.

First method is derived from the Taylor formula in t_k. We express values of solution in t_{k-1},…,t_{k-n}. When we neglect higher order terms, we obtain a system of n equations with unknown approximate solutions y_k,…,y_{k-n} and derivatives $y'(t_k)$,…,$y^{(n)}(t_k)$. By eliminating higher order derivatives we obtain method BDF2a :

$$\sum_{i=0}^{n} \alpha_i y_{n-i} = \tau_k F_k \tag{7}$$

The second method can be derived similarly from the Taylor formula in t_{k-1}

$$\sum_{i=0}^{n} \bar{\alpha}_i \bar{y}_{n-i} = \tau_k F_{k-1}. \tag{8}$$

This method is explicit and therefore not suitable for the solution of stiff problems. So we define the method BDF2b as a linear combination of schemes (7) and (8) by

$$\sum_{i=0}^{n} \hat{\alpha}_i \hat{y}_{n-i} = \hat{\gamma}_0 \tau_k F_k + \hat{\gamma}_1 \tau_k F_{k-1}. \tag{9}$$

3.2. Error estimation

From the Taylor formula we also get an estimation of the local discretization error for the BDF2a and BDF2b methods in the form

$$\begin{aligned} e_k \equiv y(t_k) - y_k &\approx f_1(\tau_k, \ldots, \tau_{k-n+1}) y^{(n+1)}(t_k), \\ \hat{e}_k \equiv y(t_k) - \hat{y}_k &\approx f_2(\tau_k, \ldots, \tau_{k-n+1}) y^{(n+1)}(t_{k-1}). \end{aligned} \tag{10}$$

Now let us assume that $y^{n+1}(t_k) \approx y^{n+1}(t_{k-1})$. From (10) we eliminate the term $y^{n+1}(\cdot)$ and after substituton we obtain a computable expression for the local discretization error depending on both approximate solutions only. Therefore we have

$$e_k \approx \delta(y_k - \hat{y}_k), \tag{11}$$

$$\hat{e}_k \approx \hat{\delta}(y_k - \hat{y}_k). \tag{12}$$

We can also combine our two solutions to obtain final solution of a higher order of accuracy by

$$\breve{y}_k = \hat{\delta} y_k - \delta \hat{y}_k, \tag{13}$$

whose order of convergence is equal to $n+1$. In [3], we computed coefficients for $n = 1, 2, 3$ and verified stability of the proposed methods.

4. Full space–time discretization

By a direct application of an implicit method to the semi-discrete problem (5), we obtain a system of nonlinear algebraic equations at each time step, which is expensive to solve. Therefore we use a semi–implicit approach, where the linear terms are treated implicitly, whereas the nonlinear ones explicitly. For the nonlinear terms we employ an explicit higher order extrapolation. Then we obtain the scheme

$$\begin{aligned} \frac{1}{\tau_k} \left(\sum_{l=0}^{n} \alpha_l u_h^{k-l}, v_h \right) + \gamma_0 a_h(u_h^k, v_h) + \gamma_0 b_h \left(\sum_{l=1}^{n} \beta_l u_h^{k-l}, v_h \right) & \\ + \gamma_1 a_h(u_h^{k-1}, v_h) + \gamma_1 b_h(u_h^{k-1}, v_h) = 0 \quad \forall\, v_h \in S_h \, . & \end{aligned} \tag{14}$$

5. Adaptive choice of time step

An important feature of modern numerical algorithms is the adaptivity, i.e., their ability to estimate the local discretization error during execution and adapt a time step in such a way, that the local discretization error is under a given tolerance. Thus, at each time step, we estimate the local discretization error and on the basis of this estimation we choose the next time step. In order to ensure an efficiency of the method the local discretization error at each time step should be approximately equal to the given tolerance TOL. Let us denote by EST the estimate of the local error. Since the order of convergence of the method is equal to $n+1$, we have

$$\mathrm{EST} = C\tau_k^{n+1}. \tag{15}$$

We want to find a time step $\bar{\tau}_k$ such that

$$\mathrm{TOL} = C\bar{\tau}_k^{n+1}. \tag{16}$$

Therefore we define the next time step by

$$\bar{\tau}_k = \tau_k \sqrt[n+1]{\frac{\mathrm{TOL}}{\mathrm{EST}}}. \tag{17}$$

If EST is much larger than TOL, we reject the last time step and compute it again using $\bar{\tau}_k$ instead of τ_k. Otherwise we accept the last time step and compute the next one using $\tau_{k+1} := \bar{\tau}_k$.

6. Numerical results

6.1. Orders of convergence

We investigate the experimental orders of convergence of the presented numerical schemes. We carried out numerical experiments for an ordinary differential equation having the exact solution in the form

$$y = \frac{\mathrm{e}^{\alpha t} - 1}{\mathrm{e}^{\alpha} - 1} \tag{18}$$

on interval $t \in [0,1]$ with $\alpha = 500$. The following table contains the computational errors for the one, two, and three-step BDF and the corresponding orders of convergence.

n	10^{-2}	10^{-3}	10^{-4}	10^{-5}	order
1	1.53×10^{0}	2.07×10^{-2}	2.08×10^{-4}	2.08×10^{-6}	1.96
2	1.04×10^{0}	3.20×10^{-3}	3.45×10^{-6}	3.47×10^{-9}	2.84
3	8.06×10^{-1}	6.31×10^{-4}	7.65×10^{-8}	1.23×10^{-11}	3.64

We observe the order of convergence $n+1$ since we used a combination of two different methods of order n. However, this procedure can not be used in case of the scalar convection-diffusion equation. Not only we do not obtain more accurate solution, but combination of two solutions of order n is even worse. So we have to use one of our two methods of order n.

Further we consider the scalar convection–diffusion equation (1) with the exact solution

$$\bar{u} \ = x(1-x)y(1-y)\frac{e^{\alpha t}-1}{e^{\alpha}-1} \tag{19}$$

on $[0,1]\times[0,1]$ and time interval $t \in [0,1]$. The computational errors and the order of convergence are shown in the following table.

	10^{-1}	5×10^{-2}	10^{-2}	5×10^{-3}	order
$n=1$	3.18×10^{-1}	1.48×10^{-1}	2.74×10^{-2}	1.34×10^{-2}	1.05
$n=2$	1.14×10^{-1}	3.49×10^{-2}	1.42×10^{-3}	2.63×10^{-4}	2.02
$n=3$	8.15×10^{-2}	1.05×10^{-2}	2.58×10^{-4}	9.31×10^{-5}	2.28

We observe, that the numerical order of convergence in this case is approximately n and it corresponds to the expected one. However, for the case $n = 3$, the order is 2.28 only, which is caused by the fact that the solution depends on both time and space discretization and its order of convergence is $O(h^p + \tau^n)$. Hence, if τ^n is so small that h^p has nonnegligible influence then further increase of order of accuracy in time has no effect.

6.2. Efficiency of the adaptive strategy

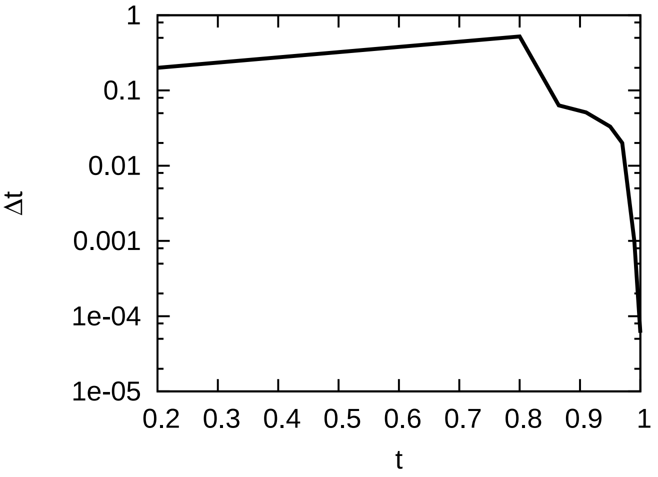
In this section we compare the efficiency of the methods using a constant and adaptive time step. We compared how many time steps are needed to obtain solution with prescribed accuracy.

6.2.1. Ordinary differential equations

First we carried out experiments for the ordinary differential equation with the exact solution (18). The following table shows the numbers of time steps necessary to obtain solution with errors 10^{-2} to 10^{-6}.

		10^{-2}	10^{-3}	10^{-4}	10^{-5}	10^{-6}
constant	$n=1$	1375	4474	14365	45790	143641
	$n=2$	642	1425	3197	6972	15110
	$n=3$	410	855	1586	2879	5222
adaptive	$n=1$	34	81	241	965	2520
	$n=2$	26	36	65	145	266
	$n=3$	24	29	43	70	108

We observe that the adaptive method is more effective. The differential equation is chosen in such a way, that the exact solution is almost constant in the major part of the interval. However, at the end of the interval the solution grows very steeply. So the major part of the interval can be done with few steps, which adaptive method allows. The following figure shows the lengths of time steps with respect to the time.



We observe that the time step is quite long at the beginning of the interval, while at the end it is shortening rapidly.

6.2.2. Scalar convection–diffusion equations

Further we consider the scalar equation (1) with the exact solution in the form (19). The numbers of iterations, which are needed to obtain solution with errors 10^{-1}, 10^{-2} and 10^{-3}, are in the following table, which verifies the efficiency of the adaptive strategy.

		10^{-1}	10^{-2}	10^{-3}
constant	$n = 1$	7	98	> 10000
	$n = 2$	5	27	> 10000
	$n = 3$	4	18	8973
adaptive	$n = 1$	9	29	1335
	$n = 2$	6	11	650
	$n = 3$	5	9	321

References

[1] V. Dolejší and M. Feistauer: *Error estimates of the discontinuous Galerkin method for nonlinear nonstationary convection-diffusion problems.* Numer. Funct. Anal. Optim. **26**, 2005, 2709–2733.

[2] V. Dolejší, M. Feistauer, and V. Sobotíková: *A discontinuous Galerkin method for nonlinear convection–diffusion problems.* Comput. Methods Appl. Mech. Engrg., **19**, 2005, 2709–2733.

[3] P. Kůs: *Solution of convection–diffusion equations with adaptive methods of higher order in space and time.* Master's thesis, Charles University Prague, Faculty of Mathematics and Physics, 2006, (in Czech).

[4] P. Lötstedt, S. Söderberg, A. Ramage, and L. Hemmingsson-Frändén: *Implicit solution of hyperbolic equations with space-time adaptivity.* BIT **42**, 2002, 134–158.

INTERIOR-POINT METHOD FOR LARGE-SCALE l_1 OPTIMIZATION*

Ladislav Lukšan, Ctirad Matonoha, Jan Vlček

Consider the l_1 optimization problem: Minimize function

$$F(x) = \sum_{i=1}^{m} |f_i(x)|, \tag{1}$$

where $f_i : R^n \to R$, $0 \le i \le m$ (m is usually large), are smooth functions depending on a small number of variables. We will assume that these functions are twice continuously differentiable with bounded first and second-order derivatives in a sufficiently large region $\mathcal{D}$.

Minimization of F is equivalent to the sparse nonlinear programming problem with $n + m$ variables $x \in R^n$, $z \in R^m$:

$$\text{minimize} \quad \sum_{i=1}^{m} z_i \quad \text{subject to} \quad -z_i \le f_i(x) \le z_i, \quad 1 \le i \le m. \tag{2}$$

In this contribution, we introduce a trust-region interior-point method for nonconvex nonlinear programming that utilizes a special structure of problem (2). All theoretical results are given without proofs. These proofs can be found in [5].

The constrained problem (2) is replaced by a sequence of unconstrained problems

$$\begin{aligned} \text{minimize} \quad B(x, z; \mu) \;&=\; \sum_{i=1}^{m} z_i - \mu \sum_{i=1}^{m} \log(z_i - f_i(x)) - \mu \sum_{i=1}^{m} \log(z_i + f_i(x)) \\ &=\; \sum_{i=1}^{m} z_i - \mu \sum_{i=1}^{m} \log(z_i^2 - f_i^2(x)), \end{aligned} \tag{3}$$

where $z_i > |f_i(x)|$, $1 \le i \le m$, and $\mu > 0$ (we assume that $\mu \to 0$ monotonically). Barrier function (3) remains unchanged if we replace problem (2) by equivalent problem

$$\text{minimize} \quad \sum_{i=1}^{m} z_i \quad \text{subject to} \quad f_i^2(x) \le z_i^2, \quad 1 \le i \le m. \tag{4}$$

The necessary first-order (KKT) conditions for the solution of (4) have the form

$$\sum_{i=1}^{m} 2w_i f_i(x) \nabla f_i(x) = 0, \quad 2w_i z_i = 1, \quad w_i \ge 0, \quad w_i(z_i^2 - f_i^2(x)) = 0, \quad 1 \le i \le m, \tag{5}$$

*This work was supported by the Grant Agency of the Czech Academy of Sciences, project No. IAA1030405 and the institutional research plan No. AV0Z10300504

where w_i, $1 \le i \le m$, are Lagrange multipliers. Since $z_i = |f_i(x)|$, $1 \le i \le m$, at the solution of (4), we can write (5) in a simpler equivalent form

$$\sum_{i=1}^{m} u_i \nabla f_i(x) = 0, \quad \frac{u_i z_i}{f_i(x)} = 1, \quad z_i^2 - f_i^2(x) = 0, \quad 1 \le i \le m, \tag{6}$$

where $u_i = 2 w_i f_i(x)$ for $1 \le i \le m$.

The special structure of problem (3) allows us to obtain minimizer $z(x;\mu) \in R^m$ of function $B(x,z;\mu)$ for a given $x \in R^n$.

Lemma 1. *Function $B(x,z;\mu)$ (with x fixed) has the unique stationary point, which is its global minimizer. This stationary point is characterized by equations*

$$\frac{2\mu z_i(x;\mu)}{z_i^2(x;\mu) - f_i^2(x)} = 1 \quad \text{or} \quad z_i^2(x;\mu) - f_i^2(x) = 2\mu z_i(x;\mu), \quad 1 \le i \le m, \tag{7}$$

which have solutions

$$z_i(x;\mu) = \mu + \sqrt{\mu^2 + f_i^2(x)}, \quad 1 \le i \le m. \tag{8}$$

Assuming $z = z(x;\mu)$, we denote

$$B(x;\mu) = \sum_{i=1}^{m} z_i(x;\mu) - \mu \sum_{i=1}^{m} \log(z_i^2(x;\mu) - f_i^2(x)) \tag{9}$$

and $u(x;\mu) = u(x, z(x;\mu);\mu)$. In this case, barrier function $B(x;\mu)$ depends only on x.

Lemma 2. *Consider barrier function (9). Then*

$$\nabla B(x;\mu) = g(x;\mu), \tag{10}$$

where $g(x;\mu) = A(x)u(x;\mu) = \sum_{i=1}^{m} \nabla f_i(x) u_i(x;\mu)$ with

$$u_i(x;\mu) = \frac{2\mu f_i(x)}{z_i^2 - f_i^2(x)}, \quad 1 \le i \le m, \tag{11}$$

and

$$\nabla^2 B(x;\mu) = G(x;\mu) + A(x)V(x;\mu)A^T(x), \tag{12}$$

where

$$G(x;\mu) = \sum_{i=1}^{m} u_i(x;\mu) G_i(x) \tag{13}$$

with $G_i(x) = \nabla^2 f_i(x)$, $1 \le i \le m$, and $V(x;\mu) = \operatorname{diag}(v_1(x;\mu), \dots, v_m(x;\mu))$ with

$$v_i(x;\mu) = \frac{2\mu}{z_i^2(x;\mu) + f_i^2(x)}, \quad 1 \le i \le m. \tag{14}$$

Lemma 3. *Let $\nabla^2 B(x;\mu)d = -\nabla B(x;\mu)$. If matrix $G(x;\mu)$ is positive definite, then $d^T g(x;\mu) < 0$ (direction vector d is descent for $B(x;\mu)$).*

Since positive definiteness of matrix $G(x;\mu)$ is not assured, the standard line-search methods cannot be used. For this reason, trust-region methods were developed. These methods use the direction vector obtained as an approximate minimizer of the quadratic subproblem

$$\text{minimize} \quad Q(d) = \frac{1}{2} d^T \nabla^2 B(x;\mu) d + g^T(x;\mu) d \quad \text{subject to} \quad \|d\| \le \Delta, \tag{15}$$

where Δ is the trust region radius. Direction vector d serves for obtaining new point $x^+ \in R^n$. Denoting

$$\rho(d) = \frac{B(x+d;\mu) - B(x;\mu)}{Q(d)}, \tag{16}$$

we set

$$x^+ = x \quad \text{if} \quad \rho(d) \le 0, \quad \text{or} \quad x^+ = x + d \quad \text{if} \quad \rho(d) > 0. \tag{17}$$

Finally, we update the trust region radius in such a way that

$$\begin{aligned} \Delta^+ &= \underline{\beta}\Delta && \text{if} \quad \rho(d) < \underline{\rho}, \\ \Delta^+ &= \Delta && \text{if} \quad \underline{\rho} \le \rho(d) \le \overline{\rho}, \\ \Delta^+ &= \overline{\beta}\Delta && \text{if} \quad \overline{\rho} < \rho(d), \end{aligned} \tag{18}$$

where $0 < \underline{\rho} < \overline{\rho} < 1$ and $0 < \underline{\beta} < 1 < \overline{\beta}$.

Now we are in a position to describe the basic algorithm.

Algorithm 1.

Data: Termination parameter $\underline{\varepsilon} > 0$, minimum value of the barrier parameter $\underline{\mu} > 0$, rate of the barrier parameter decrease $0 < \tau < 1$, trust-region parameters $0 < \underline{\rho} < \overline{\rho} < 1$, trust-region coefficients $0 < \underline{\beta} < 1 < \overline{\beta}$, step bound $\overline{\Delta} > 0$.

Input: Sparsity pattern of matrix A. Initial estimation of vector x.

Step 1: *Initiation.* Choose initial barrier parameter $\mu > 0$ and initial trust-region radius $0 < \Delta \le \overline{\Delta}$. Determine the sparsity pattern of matrix $\nabla^2 B$ from the sparsity pattern of matrix A. Carry out symbolic decomposition of $\nabla^2 B$. Compute values $f_i(x)$, $1 \le i \le m$, and $F(x) = \sum_{1 \le i \le m} |f_i(x)|$. Set $k := 0$ (iteration count).

Step 2: *Termination.* Determine vector $z(x;\mu)$ by (8) and vector $u(x;\mu)$ by (11). Compute matrix $A(x)$ and vector $g(x;\mu) = A(x)u(x;\mu)$. If $\mu \le \underline{\mu}$ and $\|g(x;\mu)\| \le \underline{\varepsilon}$, then terminate the computation. Otherwise set $k := k+1$.

Step 3: *Approximation of the Hessian matrix.* Compute approximation of matrix $G(x;\mu)$ by using differences $A(x+\delta v)u(x;\mu) - g(x;\mu)$ for a suitable set of vectors v (see [1]). Determine Hessian matrix $\nabla^2 B(x;\mu)$ by (12).

Step 4: *Direction determination.* Determine vector d as an approximate solution of trust-region subproblem (15).

Step 5: *Step-length selection.* Determine x^+ by (17) and set $x := x^+$. Compute values $f_i(x)$, $1 \le i \le m$, and $F(x) = \sum_{1 \le i \le m} |f_i(x)|$.

Step 6: *Trust-region update.* Determine new trust-region radius Δ by (18) and set $\Delta := \min(\Delta, \overline{\Delta})$.

Step 7: *Barrier parameter update.* If $\rho(d) \ge \underline{\rho}$ (where $\rho(d)$ is given by (16)), determine a new value of barrier parameter $\mu \ge \underline{\mu}$ (not greater than the current one) by the procedure described below. Go to Step 2.

The use of the maximum step-length $\overline{\Delta}$ has no theoretical significance but is very useful for practical computations. First, the problem functions can sometimes be evaluated only in a relatively small region (if they contain exponentials) so that the maximum step-length is necessary. Secondly, the problem can be very ill-conditioned far from the solution point, thus large steps are unsuitable. Finally, if the problem has more local solutions, a suitably chosen maximum step-length can cause a local solution with a lower value of F to be reached. Therefore, maximum step-length $\overline{\Delta}$ is a parameter, which is most frequently tuned.

Direction vector $d \in R^n$ should be a solution of the quadratic subproblem (15). Usually, an inexact approximate solution suffices. The dog-leg method described in [6], [2], seeks d as a linear combination of the Cauchy step $d_C = -(g^T g / g^T \nabla^2 B g) g$ and the Newton step $d_N = -(\nabla^2 B)^{-1} g$. The Newton step is computed by using either sparse Gill-Murray decomposition [4] or sparse Bunch-Parlett decomposition [3]. The sparse Gill-Murray decomposition has the form $\nabla^2 B + E = LDL^T = R^T R$, where E is a positive semidefinite diagonal matrix (which is equal to zero when $\nabla^2 B$ is positive definite), L is a lower triangular matrix, D is a positive definite diagonal matrix and R is an upper triangular matrix. The sparse Bunch-Parlett decomposition has the form $\nabla^2 B = PLML^T P^T$, where P is a permutation matrix, L is a lower triangular matrix and M is a block-diagonal matrix with 1×1 or 2×2 blocks (which is indefinite when $\nabla^2 B$ is indefinite). The following algorithm is a typical implementation of the dog-leg method.

Algorithm A: Data $\Delta > 0$.

Step 1: If $g^T \nabla^2 B g \le 0$, set $s := -(\Delta / \|g\|) g$ and terminate the computation.

Step 2: Compute the Cauchy step $d_C = -(g^T g / g^T \nabla^2 B g) g$. If $\|d_C\| \ge \Delta$, set $d := (\Delta / \|d_C\|) d_C$ and terminate the computation.

Step 3: Compute the Newton step $d_N = -(\nabla^2 B)^{-1} g$. If $(d_N - d_C)^T d_C \ge 0$ and $\|d_N\| \le \Delta$, set $d := d_N$ and terminate the computation.

Step 4: If $(d_N - d_C)^T d_C \ge 0$ and $\|d_N\| > \Delta$, determine number θ in such a way that $d_C^T d_C / d_C^T d_N \le \theta \le 1$, choose $\alpha > 0$ such that $\|d_C + \alpha(\theta d_N - d_C)\| = \Delta$, set $d := d_C + \alpha(\theta d_N - d_C)$ and terminate the computation.

Step 5: If $(d_N - d_C)^T d_C < 0$, choose $\alpha > 0$ such that $\|d_C + \alpha(d_C - d_N)\| = \Delta$, set $d := d_C + \alpha(d_C - d_N)$ and terminate the computation.

The above algorithm generates direction vectors such that

$$\begin{array}{rcl} \|d\| & \le & \Delta, \\ \|d\| & < & \Delta \quad \Rightarrow \quad \nabla^2 B d = -g, \\ -Q(d) & \ge & \underline{\sigma}\|g\| \min\left(\Delta, \dfrac{\|g\|}{\|\nabla^2 B\|}\right), \end{array}$$

where $0 < \underline{\sigma} < 1$ is a constant. These inequalities imply (see [7]), that a constant $0 < \underline{c} < 1$ exists such that

$$\|d\| \ge \underline{c}\gamma/\overline{B}, \tag{19}$$

where γ is the minimum norm of gradients that have been computed and $\overline{B}$ is an upper bound for $\|\nabla^2 B\|$ (assuming $\mu \ge \underline{\mu} > 0$, we can set $\overline{B} = m(\overline{G} + \overline{g}^2/(2\underline{\mu}))$). Thus

$$B(x+d;\mu) - B(x;\mu) \le \underline{\rho} Q(d) \le -\underline{\rho}\,\underline{\sigma}\,\underline{c}\frac{\gamma^2}{\overline{B}} \quad \text{if} \quad \rho \ge \underline{\rho} \tag{20}$$

by (17) and (19).

Algorithm 1 is sensitive on the way in which the barrier parameter decreases. We have tested various possibilities for the barrier parameter update including simple geometric sequences, which were proved to be unsuitable. Better results were obtained by setting

$$\mu_{k+1} = \mu_k \quad \text{if} \quad \|g_k\|^2 > \tau\mu_k \quad \text{or} \quad \mu_{k+1} = \max(\underline{\mu}, \|g_k\|^2) \quad \text{if} \quad \|g_k\|^2 \le \tau\mu_k, \tag{21}$$

where $0 < \tau < 1$.

In the subsequent considerations, we will assume that $\underline{\varepsilon} = \underline{\mu} = 0$ and all computations are exact.

Lemma 4. *Let Assumption 3 be satisfied. Then values $\{\mu_k\}_1^\infty$, generated by Algorithm 1, form a non-increasing sequence such that $\mu_k \to 0$.*

Lemma 5. *The inequality*

$$B(x_{k+1};\mu_{k+1}) \le B(x_{k+1};\mu_k) - L(\mu_{k+1} - \mu_k) \tag{22}$$

holds with some $L \in R$.

Theorem 1. *Consider sequence $\{x_k\}_1^\infty$, generated by Algorithm 1. Then*

$$\liminf_{k\to\infty} \sum_{i=1}^m u_i(x_k;\mu_k) g_i(x_k) = 0$$

and

$$u_i(x_k;\mu_k) = \frac{f_i(x_k)}{z_i(x_k;\mu_k)}, \quad \lim_{k\to\infty}(z_i^2(x_k;\mu_k) - f_i^2(x_k)) = 0$$

for $1 \le i \le m$.

Remark 1. If we replace (17) by

$$x^+ = x \quad \text{if} \quad \rho(d) < \underline{\rho}, \quad \text{or} \quad x^+ = x + d \quad \text{if} \quad \rho(d) \geq \underline{\rho} \tag{23}$$

in Algorithm 1, then $\lim_{k\to\infty} \|g(x_k;\mu_k)\| = 0$.

Corollary 1. *Let assumptions of Theorem 1 and (23) hold. Then every cluster point $x \in R^n$ of sequence $\{x_k\}_1^\infty$ satisfies KKT conditions (6), where $u \in R^m$ is a cluster point of sequence $\{u(x_k;\mu_k)\}_1^\infty$.*

The efficiency of Algorithm 1 was tested by using extensive collections of test problems. The results are given in [5].

References

[1] T.F. Coleman, J.J. Moré: *Estimation of sparse Hessian matrices and graph coloring problems.* Mathematical Programming **28**, 1984, 243–270.

[2] J.E. Dennis, H.H.W. Mei: *An unconstrained optimization algorithm which uses function and gradient values.* Report No. TR 75-246, 1975.

[3] I.S. Duff, M. Munksgaard, H.B. Nielsen, J.K. Reid: *Direct solution of sets of linear equations whose matrix is sparse and indefinite.* J. Inst. Maths. Applics. **23**, 1979, 235–250.

[4] P.E. Gill, W. Murray: *Newton type methods for unconstrained and linearly constrained optimization.* Mathematical Programming **7**, 1974, 311–350.

[5] L. Lukšan, C. Matonoha, J. Vlček: Trust-region interior-point method for large sparse l_1 optimization. Report V-942, Prague, ICS AS CR, 2005 (to appear in Optimization Methods and Software).

[6] M.J.D. Powell: *A new algorithm for unconstrained optimization.* In: Nonlinear Programming, J.B. Rosen, O.L. Mangasarian, K. Ritter (eds.), Academic Press, London 1970.

[7] M.J.D. Powell: *On the global convergence of trust region algorithms for unconstrained optimization.* Mathematical Programming **29**, 1984, 297–303.

THE APPLICATION OF THE THERMAL BALANCE METHOD FOR COMPUTATION OF WARMING IN ELECTRIC MACHINES[*]

Jaroslav Mlýnek

Abstract

The paper describes the procedure of the thermal balance method implementation for the computation of warming in electrical machines. Our effort will be focused on the temperature distribution in transformer screening under a stationary load. Since the three-dimensional problem is axially symmetric, it will be reduced by means of the cylindrical coordinates to an elliptic partial differential equation of second order with the Newton boundary conditions on a rectangular domain. Results of numerical tests are presented as well.

1. Introduction

Heat energy is being accumulated in an electrical machine during its operation. Thus, the temperature increase in its different parts depends on the accumulated heat energy. The electrical machine operating temperature is an important feature of a proper functioning and lifespan. The highest (and often also the lowest) operating temperature is prescribed for most of machine components.

These requirements could be reached by limiting the ambient temperature, at which the machine works in and by preventing machine parts warming over specified allowable limits. One of the most effective approaches for solving these problems is the description of spreading heat in electrical machines by means of a mathematical model, which is subsequently investigated. At present, mathematical models are often solved by using a variational formulation (see e.g. [3] and [4]). A one-dimensional problem of heat conduction is solved in [5]. This contribution is focused on the computation of warming of a transformer container screening at a stationary load by means of the thermal balance method.

2. Problem definition

Transformer screening is considered in the form of a thin-wall cylinder and the temperature field is supposed to be rotationally symmetric. Therefore, the warming computation problem can be solved in screening cross section on a two-dimensional closed domain Ω ($R_1 \le r \le R_2$, $Z_1 \le z \le Z_2$, see Fig. 1).

The temperature field is described by the elliptic partial differential equation of second order (see [3, p. 221])

$$\lambda_r \left(\frac{\partial^2 \vartheta(r,z)}{\partial r^2} + \frac{1}{r} \frac{\partial \vartheta(r,z)}{\partial r} \right) + \lambda_z \frac{\partial^2 \vartheta(r,z)}{\partial z^2} = -q(r,z) \tag{1}$$

[*] Technical University of Liberec Internal Grant IGS 5130/1 supported this contribution.

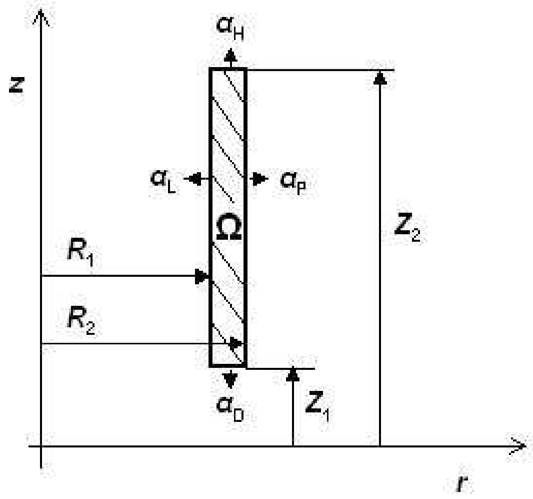


Fig. 1: *Cross section of the transformer screening.*

with the Newton boundary conditions

$$\lambda_r \frac{\partial \vartheta(r,z)}{\partial r} + \alpha_{L,P}\left(\vartheta(r,z) - u(z)\right) = 0 \tag{2}$$

on vertical parts of boundary of Ω and

$$\lambda_z \frac{\partial \vartheta(r,z)}{\partial z} + \alpha_{H,D}\left(\vartheta(r,z) - u(z)\right) = 0 \tag{3}$$

on horizontal parts for appropriate values of r and z. Real values λ_r and λ_z stand for heat conductivities of the material in the r-axis and z-axis directions, respectively; the true solution $\vartheta(r,z)$ denotes screening temperature rise with respect to the surrounding oil temperature. The function $u(z)$ in expressions (2) and (3) allows to respect the variable temperature of oil in the vicinity of screening in the z-axis direction. It is given by the formula $u(z) = Cz$, where C is constant. In expression (1), the function $q(r,z)$ represents the volume density of losses, which is expressed by the following relation:

$$q(r,z) = \delta^2(z)\rho(1 + \alpha_T\vartheta(r,z)), \tag{4}$$

where $\delta(z)$ denotes the density of eddy currents, ρ is the specific resistance of the material used for screening, and α_T is the factor for the dependence of a specific resistance on temperature. In boundary conditions (2) and (3), the constants α_L, α_P, α_H, and α_D stand for the heat transfer coefficients on the left, right, upper, and lower parts of the rectangular domain Ω, respectively.

3. Solving the problem by means of the thermal balance method

Equation (1) can be transformed to a self-adjoint form and after the substitution of the function $q(z)$ from expression (4), the basic equation will be obtained. It describes warming in the cross section Ω of transformer screening:

$$\frac{\partial}{\partial r}\left(\lambda_r r\frac{\partial\vartheta(r,z)}{\partial r}\right)+r\frac{\partial}{\partial z}\left(\lambda_z\frac{\partial\vartheta(r,z)}{\partial z}\right)=-r\delta^2(z)\rho(1+\alpha_T\vartheta(r,z)) \qquad (5)$$

with boundary conditions (2) and (3).

In the domain Ω, a regular rectangular mesh will be constructed with increments

$$h_r=\frac{R_2-R_1}{m} \quad \text{and} \quad h_z=\frac{Z_2-Z_1}{n}$$

in the r-axis and z-axis directions, respectively, where m and n denotes the number of segments, to which the region is divided in the r-axis and z-axis directions, respectively. Let us denote $r_k=R_1+kh_r$, $z_s=Z_1+sh_z$, and $\vartheta_{k,s}=\vartheta(r_k,z_s)$ the warming at the node $[r_k,z_s]$, where $k\in\{0,1,...,m\}$, $s\in\{0,1,...,n\}$.

Let the point $[r_k,z_s]$ be an internal node in the domain Ω (see Fig. 2). Then equation (5) can be approximated at this node using the following balance of heat:

$$\lambda_r\left(r_k+\frac{h_r}{2}\right)\frac{\vartheta_{k+1,s}-\vartheta_{k,s}}{h_r}h_z+\lambda_r\left(r_k-\frac{h_r}{2}\right)\frac{\vartheta_{k-1,s}-\vartheta_{k,s}}{h_r}h_z+$$
$$+\lambda_z r_k\frac{\vartheta_{k,s+1}-\vartheta_{k,s}}{h_z}h_r+\lambda_z r_k\frac{\vartheta_{k,s-1}-\vartheta_{k,s}}{h_z}h_r=-r_k\delta^2(z_s)\rho(1+\alpha_T\vartheta_{k,s})h_rh_z. \qquad (6)$$

The left-hand side of equation (6) describes the approximate quantity of heat supplied from or delivered to surrounding mesh nodes, the right-hand side expresses approximate waste heat arising in the element that pertains to the node $[r_k,z_s]$.

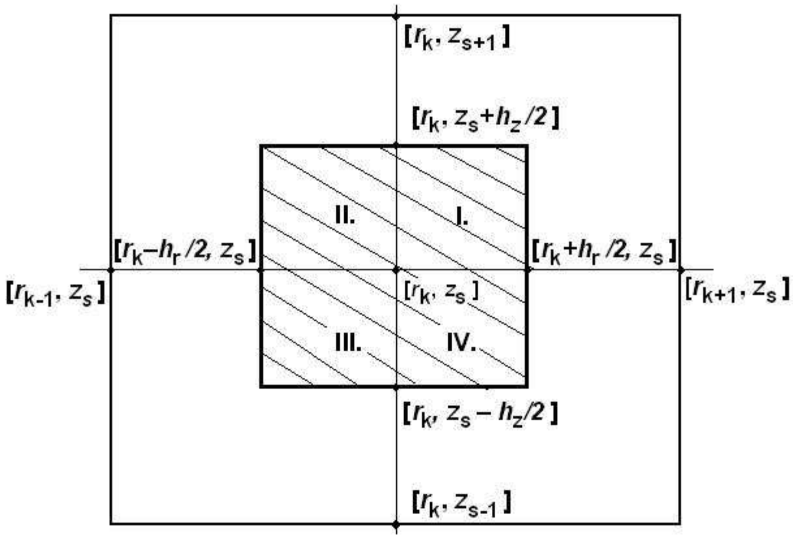


Fig. 2: *The neighborhood of the point* $[r_k,z_s]$.

Fig. 2 shows four parts (I, II, III, and IV) of a square neighborhood of a node $[r_k,z_s]$. Clearly, if the node lies on the boundary of Ω or at the corner then the neighborhood consists of two or one part, only. For the boundary nodes, boundary conditions (2) or (3) will be used to determine the thermal balances. For instance, as long as the neighborhood of the boundary point $[r_k, z_s]$ consists of parts III and IV only, we obtain by means of thermal balances the following equation:

$$\lambda_r \left(r_k + \frac{h_r}{2}\right) \frac{\vartheta_{k+1,s} - \vartheta_{k,s}}{h_r} \frac{h_z}{2} + \lambda_r \left(r_k - \frac{h_r}{2}\right) \frac{\vartheta_{k-1,s} - \vartheta_{k,s}}{h_r} \frac{h_z}{2} + \\ + \lambda_z r_k \frac{\vartheta_{k,s-1} - \vartheta_{k,s}}{h_z} h_r - \alpha_H r_k \left(\vartheta_{k,s} - u(z_s)\right) h_r = -r_k \delta^2(z_s) \rho (1 + \alpha_T \vartheta_{k,s}) h_r \frac{h_z}{2}. \quad (7)$$

Let us set $h = \max(h_r, h_z)$. Then we make the $O(h^2)$-order error by approximating equation (5) in the internal node $[r_k, z_s]$, since central differences are used. In boundary nodes we make the $O(h)$-order error in the approximation (see [2, p. 277]), because the difference

$$\lambda_z r_k \frac{\vartheta_{k,s+1} - \vartheta_{k,s}}{h_z},$$

for example, is substituted in equation (7) by the expression

$$-\alpha_H r_k (\vartheta_{k,s} - u(z_s))$$

from relation (3). This low accuracy is quite sufficient in our case, since all physical constants suffer from large uncertainties. By equations of type (6) and (7) in all mesh points, we obtain a system of linear algebraic equations with a band symmetric and positive definite matrix (for practically used values of physical quantities from equations (1), (2), and (3)). The Choleski decomposition algorithm (see [1]) was used to solve the associated system.

4. One-dimensional problem of heat conduction

For a one-dimensional heat conduction, an analytical solution can be determined and compared with an approximate solution obtained by means of the thermal balance method. Let us examine the case, when a one-dimensional heat conduction is considered in the r-axis direction. The heat transfer coefficient is nonzero only on the vertical part of the boundary of Ω (i.e. $\alpha_L = 0$, $\alpha_P \neq 0$), the current density δ is constant, and $\alpha_T = 0$. Then, equation (1) attains a simple form:

$$\lambda_r \left(\frac{\partial^2 \vartheta(r)}{\partial r^2} + \frac{1}{r} \frac{\partial \vartheta(r)}{\partial r} \right) = -q, \quad (8)$$

where $q = \delta^2 \rho$.

For the solution ϑ of problem (8) in the interior point r we have:

$$\vartheta_r = X \left\{ \frac{2\lambda_r}{\alpha_P R_2} Y + 1 - \left(\frac{r}{R_2}\right)^2 - 2 \left(\frac{R_1}{R_2}\right)^2 \ln \frac{R_2}{r} \right\}. \quad (9)$$

The temperature at boundary nodes is given by:

$$\vartheta_{R_1} = X\left\{\left(\frac{2\lambda_r}{\alpha_P R_2}+1\right)Y - 2\left(\frac{R_1}{R_2}\right)^2 \ln\frac{R_2}{R_1}\right\}, \tag{10}$$

$$\vartheta_{R_2} = \vartheta_{R_1} + q\frac{R_1^2}{4\lambda_r}\left[2\ln\frac{R_2}{R_1}+1-\left(\frac{R_2}{R_1}\right)^2\right], \tag{11}$$

where

$$X = q\frac{R_2^2}{4\lambda_r}, \quad Y = 1-\left(\frac{R_1}{R_2}\right)^2.$$

The proof of relations (9)–(11) is based on the transformation of equation (8) to the form

$$\frac{\partial}{\partial r}\left(r\frac{\partial\vartheta}{\partial r}\right) = -\frac{qr}{\lambda_r},$$

repeatedly using the integration with respect to r and applying the conditions $\alpha_P \neq 0$ and $\alpha_L = 0$.

Table 1 lists approximate values of temperature rise computed numerically by means of the thermal balance method and the values obtained through analytical formulae (9)–(11) for the following input values: $q = 10^5\,\mathrm{W/m^3}$, $R_1 = 1\,\mathrm{m}$, $R_2 = 1.1\,\mathrm{m}$, $\alpha_P = 50\,\mathrm{W/m^2K}$, $\alpha_L = \alpha_H = \alpha_D = 0$, and $\lambda_r = 1\,\mathrm{W/mK}$.

$h_r[m]$	ϑ_{R_1}[K] $R_1 = 1$[m]		ϑ_r[K] $r = 1.05$[m]		ϑ_R[K] $R_2 = 1.1$[m]	
	approx.	exact	approx.	exact	approx.	exact
0.05	673.33	675.37	551.37	552.41	190.91	190.88
0.025	674.88		552.15		190.91	
0.0167	675.17		552.29		190.91	

Tab. 1: *One-dimensional heat transfer, the comparison of the exact and approximate values of warming.*

5. Numerical example

By means of the above mentioned thermal balance method, the real-live problem was solved that involved finding the warming in aluminium transformer screening with the following input parameters: $R_1 = 0.86\,\mathrm{m}$, $R_2 = 0.868\,\mathrm{m}$, $Z_1 = 0.8864\,\mathrm{m}$, $Z_2 = 2.51\,\mathrm{m}$, $\lambda_r = \lambda_z = 220\,\mathrm{W/mK}$, $\rho = 0.3 \times 10^{-7}\Omega\,\mathrm{m}$, $\alpha_L = \alpha_P = \alpha_H = \alpha_D = 50\,\mathrm{W/m^2K}$, $\alpha_T = 0.00409\,\mathrm{K^{-1}}$, $C = 10\,\mathrm{K/m}$ (C is the costant appearing in the definition of the function $u(z)$ in expressions (2) and (3) in Section 2). The domain Ω is divided into 2 segments ($h_r = 0.4 \times 10^{-2}\,\mathrm{m}$) in the r-axis direction and subsequently to 16, 32 and 64 segments ($h_z = 0.10148\,\mathrm{m}$, $h_z = 0.050738\,\mathrm{m}$, $h_z = 0.025369\,\mathrm{m}$) in the z-axis direction. The current density $\delta(z)$ is given by means of 19 values between $0.2498 \times 10^5\,\mathrm{Am^{-2}}$ and $0.3508 \times 10^7\,\mathrm{Am^{-2}}$, the current density

		$R_1 = 0.86$ [m]	$r = 0.864$ [m]	$R_2 = 0.868$ [m]
$Z_2 = 2.51$ [m]	$h_z = 0.101480$ [m]	35.790	35.795	35.790
	$h_z = 0.050738$ [m]	30.908	30.911	30.908
	$h_z = 0.025369$ [m]	29.444	29.446	29.444
$z = 1.6982$ [m]	$h_z = 0.101480$ [m]	19.481	19.482	19.481
	$h_z = 0.050738$ [m]	19.467	19.468	19.467
	$h_z = 0.025369$ [m]	19.466	19.467	19.466
$Z_1 = 0.8864$ [m]	$h_z = 0.101480$ [m]	12.607	12.609	12.607
	$h_z = 0.050738$ [m]	12.764	12.765	12.764
	$h_z = 0.025369$ [m]	12.809	12.811	12.809

Tab. 2: *The screening temperature rise (in K) for selected nodes at $h_r = 0.004$ [m].*

at the other node points is computed by means of linear interpolation. Table 2 lists approximate values of temperature rise $\vartheta_{k,s}$ (at chosen nodes) computed numerically using the thermal balance method.

6. Conclusion

The problem (1)–(3) for specific values of transformer screening was solved by means of the above mentioned thermal balance method. The described method of solving is relatively simple, but still allows to obtain an approximate solution, which is sufficiently exact in technical practice. In numerical calculations of warming in transformer screening, the domain Ω was divided only to 2 segments in the r-axis direction (in view of the thin-wall cylindrical area of screening). The value of the increment $h_z = 0.05\,\mathrm{m}$ in the z-axis direction was sufficient. The described procedure can be used for the examination of transformer parts at various load levels during the development of transformer designs.

References

[1] H.M. Antia: *Numerical methods for scientists and engineers.* Birkhäuser Verlag, Berlin, 2000.

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

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

[4] M. Křížek, K. Segeth: *Numerical modelling of electrical engineering problems.* Karolinum, Praha, 2001, (in Czech).

[5] S.S. Kutateladze: *Foundations of heat interchange theory.* Nauka, Moscow, 1979, (in Russian).

WHY ARE THE MESHLESS METHODS USED?

Vratislava Mošová

1. A bit about meshless methods

Meshless (or meshfree) methods are a useful tool for solving partial differential equations. These methods are often compared with the Finite Element Methods. The FEM are essentially applications of the Galerkin method to the weak formulation of a given problem and use spline spaces as approximating subspaces. The basic difference between the FEM and the meshless methods consists in the construction of the approximating space. In the meshless methods, this space is formed by shape functions. The following property plays a fundamental role in construction of these functions.

Definition 1 Let $x_1, x_2, \ldots, x_N$ be arbitrarily spaced points (called particles) in the domain $\Omega \subset R^n$. The functions $\{\Psi_I\}_{i=1}^N$ that are defined on Ω form the partition of unity of s consistency if for every monomial $p(x) \in \mathcal{P}_s$

$$\sum_{I=1}^{N} \Psi_I(x) p(x_I) = p(x) \quad \forall x \in \Omega. \tag{1}$$

Different meshless methods construct the partition of unity in different ways. Shape functions in Smooth Particle Hydrodynamic Method (SPHM, see [10]), Reproducing Kernel Particle Method (RKPM, see [4], [5]), and Reproducing Kernel Hierarchical Partition of Unity Method (RKHPUM, see [9]) are derived to reproduce the kernel (in the integral form) of the approximated functions. Diffuse Element Method (DEM, see [11]) and Element-Free Galerkin Method (EFGM, see [3]) are based on a moving least squares procedure. Partition of Unity (PU, see [1]), *hp*-clouds (see [7]) and Generalized Finite Element Method (GFEM, see [1], [2]) are methods where functions from the approximating space are products of functions from an extrinsic basis (its components form a partitition of unity) and from an intrinsic basis (its components include important features of the solution in the approximation space).

In this contribution, we give some illustrative examples and we study the behaviour of their solutions obtained by means of the FEM, the RKPM and the RKHPUM, and then we show some problems that can be successfully solved by means of meshless methods.

2. Meshless methods and solution of Helmholtz equation

Example 1 Consider the following 1D boundary value problem:

$$u''(x) + 16^2 u(x) = x, \quad x \in (0,1), \tag{2}$$

$$u'(0) = u'(1) = 0. \tag{3}$$

We seek a weak solution of the given problem, it means $u \in W^{1,2}(0,1)$ such that

$$-\int_0^1 (u'v')\,\mathrm{d}x + 16^2 \int_0^1 uv\,\mathrm{d}x = \int_0^1 xv\,\mathrm{d}x,\ \forall v \in W^{1,2}(0,1). \tag{4}$$

a) Figure 1 shows the situation when we solve problem (2), (3) by the FEM with N=11 nodes and linear approximation of solution. This solution is very poor. Let u be the exact solution and $\overline{u}$ its FEM approximation. The dependence of the approximation error $\max_x|u - \overline{u}|$ on the number of nodes is given in Figure 2. Note that the accuracy can be significantly improved if we use a cubic spline approximation of the solution.

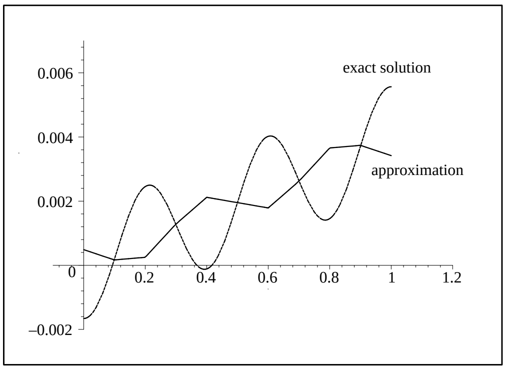


Fig. 1: *FEM – the approximation and the exact solution*

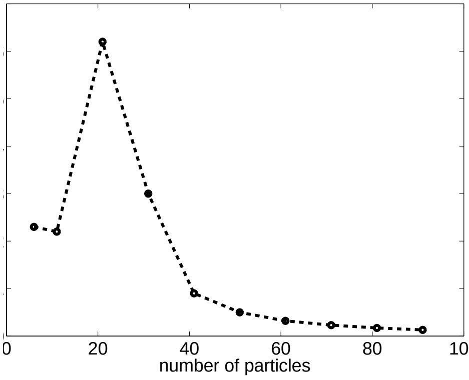


Fig. 2: *FEM – dependence of the approximation error on the number of nodes*

b) We find an approximation of the weak solution of equation (4) by means of the RKPM now. Suppose that uniformly distributed particles $x_1, \dots x_N \in \langle 0,1 \rangle$, the polynomial basis $p(x) = (1, x)$, the weight function

$$w(x) = \begin{cases} (1-x^2)^2 & \text{for } |x| < 1, \\ 0 & \text{otherwise,} \end{cases}$$

and a dilatation parametr R are given. Then the shape functions have the form

$$\Psi_I(x) = p\left(\frac{x - x_I}{R}\right) b(x_I)\, w\left(\frac{x - x_I}{R}\right),\ I = 1, \dots N. \tag{5}$$

Here b is the solution of the system $M(x)b(x) = (1,0)^T$ with the moment matrix

$$M(x) = \begin{pmatrix} m_0(x) & m_1(x) \\ m_1(x) & m_2(x) \end{pmatrix}, \quad m_i(x) = \int_0^1 (y-x)^i w\left(\frac{y-x}{R}\right) \mathrm{d}y, \quad i = 0,1,2. \quad (6)$$

If we replace u in the weak formulation (4) by its approximation $\overline{u}(x) = \sum_{I=1}^N \Psi_I(x)U_I$ and v by $\Psi_J(x)$ for $J = 1, \ldots, N$, we receive the system of linear equations

$$AU = f,$$

where $U = (U_1, ..., U_N)^T$, $f = (f_1, \ldots, f_N)^T$, $f_I = \int_0^1 x\Psi_I(x)\,\mathrm{d}x$,

$$A = (a_{I,J})_{I,J=1}^N, \quad a_{I,J} = \int_0^1 \left(16^2\Psi_I(x)\Psi_J(x) - \Psi_I'(x)\Psi_J'(x)\right) \mathrm{d}x.$$

The approximation $\overline{u}$ for $N = 11$, $R = 0.3$, and the exact analytical solution u are drawn in Figure 3. The behaviour of the error $|u - \overline{u}|$ is ilustrated in Figure 4.

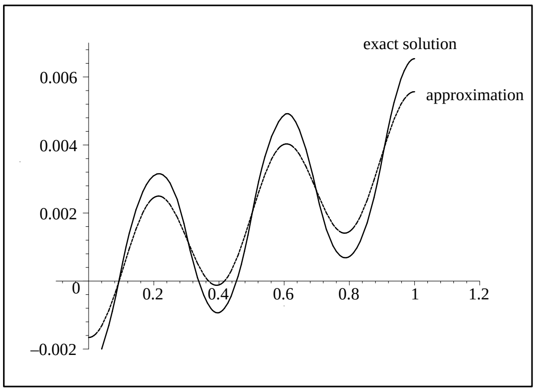


Fig. 3: *RKPM – the approximation and the exact solution*

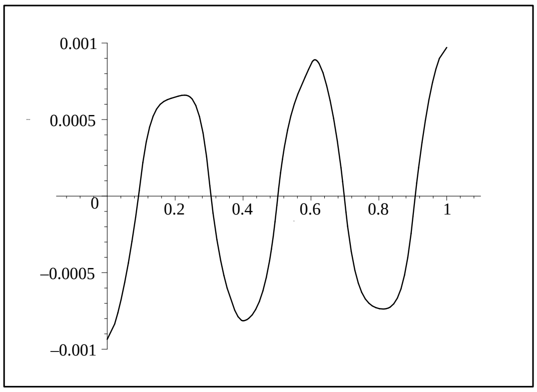


Fig. 4: *RKPM – the error* $\overline{u} - u$

c) We receive a better approximation if we solve the given problem by means of the RKHPUM. In this case, the following shape functions are constructed:

$$\Psi_I^0(x) = p\left(\frac{x-x_I}{R}\right) b^0(x_I)\, w\left(\frac{x-x_I}{R}\right), \ \Psi_I^1(x) = p\left(\frac{x-x_I}{R}\right) b^1(x_I)\, w\left(\frac{x-x_I}{R}\right),$$

where b^0, b^1 are solutions of the systems $M(x)b^0(x) = (1,0)^T$, $M(x)b^1(x) = (0,1)^T$ and the moment matrix M has the form (6). We insert the approximation

$$\overline{u}(x) = \sum_{I=1}^{11} \Psi_I^0(x)U_I^0 + \sum_{I=2}^{10} \Psi_I^1(x)U_I^1$$

into the weak formulation (4) and solve the resulting linear system. We can see in Figure 5 and Figure 6 that the accuracy of the solution has improved considerably.

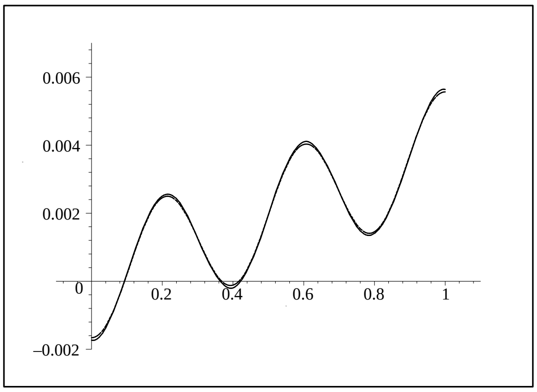


Fig. 5: *RKHPUM – the approximation and the exact solution*

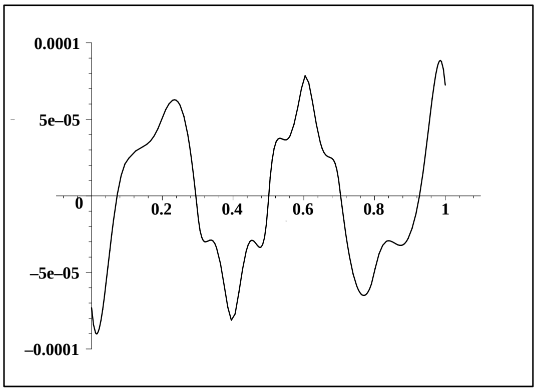


Fig. 6: *RKHPUM – the error* $\overline{u} - u$

Example 2 Let $\Omega = \langle 0,1\rangle \times \langle 0,1\rangle$,

$$\Delta u(x,y) + 16^2 u(x,y) = 1 \quad \text{in } \Omega, \tag{7}$$

$$\frac{\partial u(x,y)}{\partial n} = 2 \quad \text{on } \partial\Omega. \tag{8}$$

We seek a weak solution of this problem, it means $u \in W^{1,2}(\Omega)$ such that

$$-\iint_\Omega \nabla\, u \nabla\, v \,\mathrm{d}x\,\mathrm{d}y + \int_{\partial\Omega} 2v\,\mathrm{d}s + 16^2 \iint_\Omega uv\,\mathrm{d}x\,\mathrm{d}y = \iint_\Omega 1v\,\mathrm{d}x\,\mathrm{d}y,\ \forall v \in W^{1,2}(\Omega). \tag{9}$$

We develop the approximation of this solution by means of the RKHPUM for particles (x_I, y_I), $I = 1,\dots,N$, uniformly distributed inside Ω, the polynomial basis $p(x,y) = (1,x,y)$, a dilatation parametr R, and the weight function

$$w(x,y) = \begin{cases} ((1-x^2)(1-y^2))^2 & \text{for } |x| \le 1,\ |y| \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

In this case, the shape functions are of the form

$$\Psi_I^\alpha(x,y) = p\Big(\frac{x-x_I}{R}, \frac{y-y_I}{R}\Big) b^\alpha(x_I,y_I)\, w\Big(\frac{x-x_I}{R}, \frac{x-y_I}{R}\Big),$$

for $\alpha = (0,0),(1,0),(0,1)$, where vectors b^α satisfy $M(x,y)b^{(0,0)}(x,y) = (1,0,0)^T$, $M(x,y)b^{(1,0)}(x,y) = (0,1,0)^T$, $M(x,y)b^{(0,1)}(x,y) = (0,0,1)^T$, with

$$M(x,y) = \begin{pmatrix} m_{00}(x,y) & m_{10}(x,y) & m_{01}(x,y) \\ m_{10}(x,y) & m_{20}(x,y) & m_{11}(x,y) \\ m_{01}(x,y) & m_{11}(x,y) & m_{02}(x,y) \end{pmatrix}, \tag{10}$$

$$m_{ij}(x,y) = \iint_\Omega \left(\frac{\tilde{x}-x}{R}\right)^i \left(\frac{\tilde{y}-y}{R}\right)^j w\left(\frac{\tilde{x}-x}{R}, \frac{\tilde{y}-y}{R}\right) \mathrm{d}\tilde{x}\,\mathrm{d}\tilde{y},\ i,j = 0,1,2. \tag{11}$$

Here we put

$$\overline{u}(x,y)=\sum_{I=1}^{N}\Psi_I^{(0,0)}(x,y)U_I^{(0,0)}+\sum_{I=1}^{N}(\Psi_I^{(1,0)}(x,y)U_I^{(1,0)}+\Psi_I^{(0,1)}(x,y)U_I^{(0,1)}) \qquad (12)$$

into the weak formulation (9) and solve the resulting system of linear equations. The graph of the approximation for $N = 100$ is plotted in Figure 7. This graph is very close to the graph of the analytical solution. The approximation errors $\sqrt{\sum_{i=0}^{20}\sum_{j=0}^{20}(u(\frac{i}{20},\frac{j}{20})-\overline{u}(\frac{i}{20},\frac{j}{20}))^2}/\sqrt{\sum_{i=0}^{20}\sum_{j=0}^{20}(u(\frac{i}{20},\frac{j}{20}))^2}$ are depicted in Figure 8. The RKHPUM approximation $\overline{u}$ is computed for 25,36,49,64,100 particles.

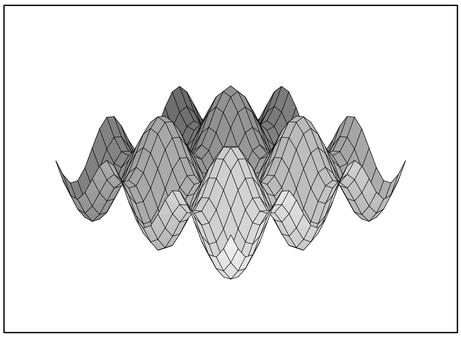


Fig. 7: *RKHPUM approximation ($N = 100$)*

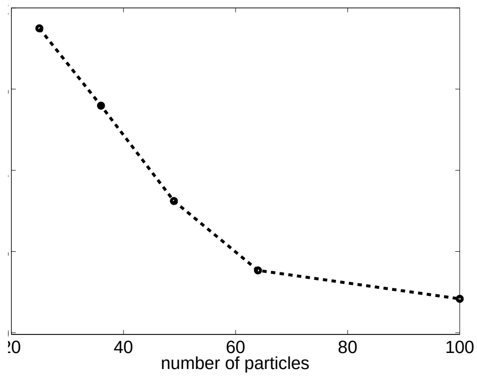


Fig. 8: *RKHPUM – dependence of the approximation error on the number of nodes*

3. Properties and advantages of the meshless methods

We demonstrated the construction of shape functions by means of the RKPM and the RKHPUM in the examples above. We saw that the considered meshless methods produce quite accurate results for $h = 0.1$. To achieve similar or even higher accuracy, we needed a number of particles that was significantly lower than the number of FEM nodes.

To realize meshless methods, no explicitly given mesh is required. The construction of shape functions needs no connectivity information. The size of support and smoothness of shape functions depend on the given dilatation parameter and on the chosen weight function only. The fact that no mesh has to be generated is appreciated in solving 3D structural mechanics problems (see [7]), in dealing with large deformations (see [4], [5]), or when we work with data received from computer tomography (CT) or magnetic resonance imaging (MRI) (see [6]).

The second advantage of meshless methods consists in the range in which shape functions can be constructed. It is possible to build shape functions with high regularity and to successfully solve higher order differential equations (see [8]) or to define shape functions that respect the local behaviour of the solution (see [12], [2]).

The meshless methods can be understood as an alternative to the FEM. For "simple" problems it is better to use the FEM, but in the specific problems mentioned above we prefer the meshless methods.

References

[1] I. Babuška, U. Banerjee, J.E. Osborn: *Survey of meshless and generalized finite element methods: An unified approach.* Acta Numer., 2003, 1–125.

[2] I. Babuška, J.M. Melenk: *The partition of unity finite element method: Basic theory and application.* Comput. Methods Appl. Mech. Engrg. **139**, 1996, 289–314.

[3] T. Belytschko, Y. Lu, I. Gu: *Element-free Galerkin methods.* Internat. J. Numer. Methods Engrg. **37**, 1994, 229–256.

[4] J.S. Chen, C. Pan, C.T. Wu, W.K. Liu: *Reproducing kernel particle methods for large deformation analysis of non-linear structures.* Comput. Methods Appl. Mech. Engrg. **139**, 1996, 195–227.

[5] J.S. Chen, C. Pan, C.T. Wu: *Large deformation analysis of rubber based on a reproducing kernel particle methods.* Comput. Mech. **19**, 1997, 211–227.

[6] E. Cueto, M. Doblaré, L. Gracia: *Imposing essential boundary conditions in the natural element method by means of density-scaled α-shapes.* Internat. J. Numer. Methods Engrg. **49**, 2000, 519–546.

[7] C.A. Duarte, I. Babuška, J.T. Oden: *Generalized finite element methods for three-dimensional structural mechanics problems*, 1999, `http://www.comco.com`.

[8] P. Joyot, J. Trunzier, F. Chinesta: *Enriched reproducing kernel approximation: Reproducing functions with discontinuous derivatives.* Meshfree methods for partial differential equations II, Springer, Berlin, 2004, 93–107.

[9] S. Li, W.K. Liu: *Reproducing kernel hierarchical partition of unity.* Internat. J. Numer. Methods Engrg. **45**, 1999, 251–317.

[10] J.J. Monaghan: *Why particle methods work.* Sci. Stat. Comput. **3**, 1982, 422–433.

[11] B. Nayroles, G. Touzot, P. Vilon: *Generalizing the finite element method: Diffuse approximation and diffuse elements.* Comput. Mech. **10**, 1992, 307–318.

[12] T. Strouboulis, I. Babuška, K. Copps: *The design and analysis of the generalized finite element method.* Comput. Methods Appl. Mech. Engrg. **181**, 2000, 43–69.

NUMERICAL APPROACHES TO PARAMETER ESTIMATES IN STOCHASTIC DIFFERENTIAL EQUATIONS DRIVEN BY FRACTIONAL BROWNIAN MOTION*

Jan Pospíšil

Abstract

We solve the one-dimensional stochastic heat equation driven by fractional Brownian motion using the modified Euler-Maruyama finite differences method. We use the numerical solution as our observation and we show how to estimate the drift parameter from a one path only.

1. Introduction

In this paper we follow [5] where parameter estimates in stochastic evolution equations driven by fractional Brownian motion were studied. The existence and ergodicity of the strictly stationary solution, which is proved there, is crucial for the parameter (especially the drift) estimates. From an observation of the solution on some time interval $[0, T]$, consistent drift estimates are given for $T \to \infty$. Such a constraint is not necessary for the diffusion estimates that can be calculated for $T < \infty$ using the variation of the solution. A presentation of the diffusion estimates is beyond the scope of this paper and only the drift estimates will be considered. In [5], the results are presented in infinite dimension, however, they apply to finite dimensional case as well.

In this paper we give a brief summary of numerical experiments done to support the obtained results in parameter estimates. To simulate the one-dimensional fractional Brownian motion we use the spectral method proposed by Z. Yin in [6]. The problem of numerical simulations of solutions to SDEs and SPDEs has only recently been addressed. Kloeden and Platen wrote a comprehensive book [2] dedicated to numerical solutions to SDEs. Some of the methods were compared by Higham in [1] and by the author in [4]. We solve the one-dimensional SPDE using the Euler-Maruyama finite differences method that has been modified for the purposes of this paper so that the driving process is considered to be a fractional Brownian motion. We will use the numerical solution as our observation and we will show how to estimate the parameters either from a one path or many paths observation.

*This work was partially supported by the GACR Grant 201/04/0750 and by the MSMT Research Plan MSM 4977751301.

2. Parameter estimates in linear SPDEs

In this section we consider the following initial boundary value problem for linear stochastic heat equation

$$\begin{aligned} dX(t,x) &= \alpha\Delta X(t,x)\,dt + \sigma\,dB^H(t), \quad t \geq 0,\ x \in [0,L],\ L > 0\,, \\ X(0,x) &= x_0(x), \quad x \in [0,L], \\ X(t,0) &= X(t,L) = 0, \quad t \geq 0, \end{aligned} \tag{1}$$

where $\alpha > 0$ and $\sigma > 0$ are real constant parameters, $\Delta = \partial^2/\partial x^2$ is the Laplace operator, $x_0 \in L^2([0,L])$ and $B^H(t), t \geq 0$, is a standard cylindrical fractional Brownian motion with Hurst parameter $H \in (1/2, 1)$.

Denote by $e_k(x) = \sqrt{2/L}\sin(k\pi x/L)$ the orthonormal[1] basis for the Laplacian on $[0,L]$ and by $\lambda_k = \alpha k^2\pi^2/L^2$ for $k = 1, 2, \ldots$. Let $S(t)$ be a strongly continuous semigroup generated by the Laplacian. Using this notation, it can be shown that $[S(t)e_k](x) = e_k(x)e^{-\lambda_k t}$. In our estimates below, we can use one function from the basis as our test function $z(x)$. Obviously

$$\begin{aligned} \langle X(t,x), z(x)\rangle_V &= \int_0^L X(t,x)z(x)\,dx \\ |X(t,x)|_V^2 = \langle X(t,x), X(t,x)\rangle_V &= \int_0^L X^2(t,x)\,dx. \end{aligned}$$

We will also need to calculate the covariance operator

$$\begin{aligned} &\langle Q_T e_k(x), e_k(x)\rangle_V \\ &\quad = \sigma^2 \int_0^T\int_0^T \left(\int_0^L \left([S(u)e_k](x)\right)\left([S(v)e_k](x)\right)\,dx\right)\phi(u-v)\,du\,dv \\ &\quad = \sigma^2 \int_0^T\int_0^T e^{-\lambda_k(u+v)}\phi(u-v)\,du\,dv, \end{aligned}$$

where $\phi(u) = H(2H-1)|u|^{2H-2}$ is again the kernel.

Let us now introduce an approach to numerically solve (1). We have to point out that the rest of this section is for illustration purposes only, because there is no result in numerical solution to SPDEs driven by fractional Brownian motion so far. The proposed method below is only a natural modification of a similar method for solving SPDEs driven by the Wiener process, but the method is presented without the knowledge of its convergence.

Define, for $i = 0, 1, \ldots, M$, a space grid by $x_i = ik$, where $k = L/M$. Using the finite difference for Laplacian we obtain the following system of SDEs

$$dX(t,x_i) = \frac{\alpha}{k^2}\left(X(t,x_{i+1}) - 2X(t,x_i) + X(t,x_{i-1})\right)\,dt + \sigma\,d\beta_i^H(t),$$

[1] it means that $\int_0^L e_k^2(x)\,dx = 1$

where $\beta_i^H(t)$ are stochastically independent standard fractional Brownian motions and $i = 1, \dots, M$. We rewrite the system into the matrix form

$$dX(t) = AX(t)\,dt + \sigma\,dB^H(t),$$

where $X(t)$ is now an $M \times 1$ matrix (vector) with elements $X(t, x_i)$, A is an $M \times M$ matrix and $B^H(t)$ an $M \times 1$ vector of the form

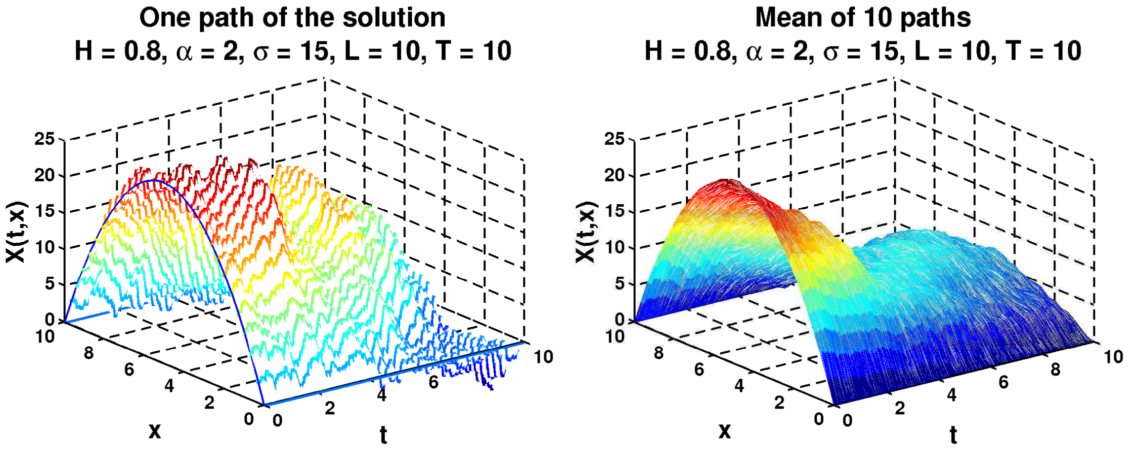
$$A = \frac{\alpha}{k^2} \begin{bmatrix} -2 & 1 & 0 & \cdots & 0 \\ 1 & -2 & 1 & & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & & 1 & -2 & 1 \\ 0 & \cdots & 0 & 1 & -2 \end{bmatrix}, \quad B^H(t) = \begin{bmatrix} \beta_1^H(t) \\ \beta_2^H(t) \\ \vdots \\ \beta_M^H(t) \end{bmatrix}.$$

Now we can use again the Euler-Maruyama method to generate a sequence (of vectors) (Y_j) approximating the solution $X(t_j)$ by the following *explicit* scheme:

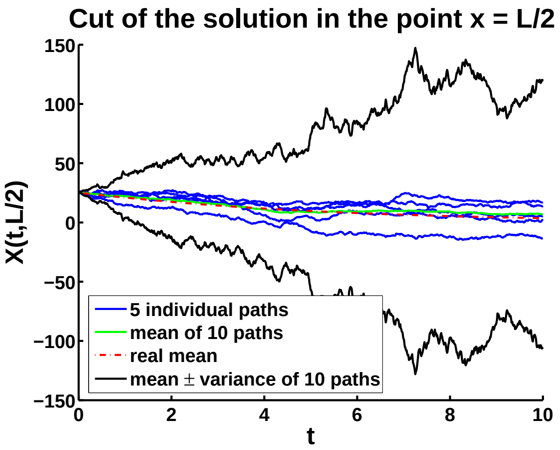
$$\begin{aligned} Y_0 &= x_0 \\ Y_{j+1} &= Y_j + AY_jh + \sigma W_j^H, \quad j = 1, \dots, N, \end{aligned} \tag{2}$$

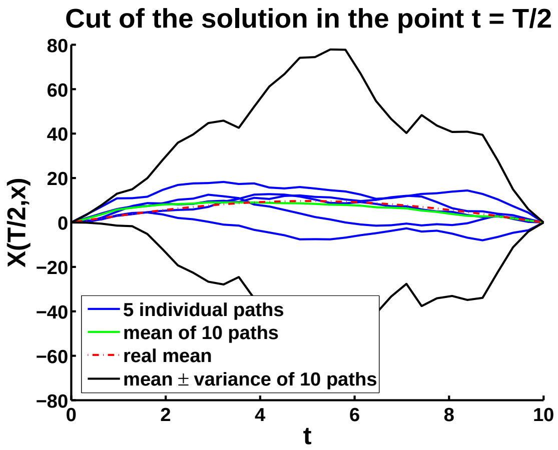
where $W_j^H = B^H(t_{j+1}) - B^H(t_j)$ are the increments of fractional Brownian motion. Like in the previous section, it must be pointed out that it was not the purpose of this paper to study convergence of this numerical scheme.

In the following figure on the left we can see one sample path of the solution $X(t, x)$ to (1) with initial condition $x_0(x) = x(L - x)$, $x \in [0, L]$ for particular values of H, α, σ, L and T. Picture on the right shows the mean of $P = 10$ paths of the solution.

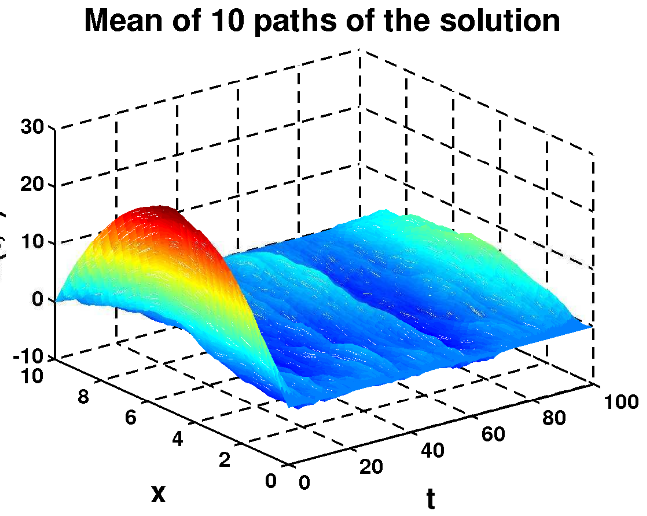


In the next figure we can see the cuts of the solution in the points $x = L/2$ and $t = T/2$ respectively. Several individual paths are drawn together with their mean and variance. Note that some of the path and even the mean could be also negative.





In figure on the right we can see the mean of $P = 10$ paths of the solution to (1) over a larger time interval ($T = 100$). We can see that the influence of the initial condition vanishes rather quickly and the solution converges to the strictly stationary solution.



Remark 2.1. To ensure the convergence in the explicit scheme, it is necessary to control some relation between the time and space steps. For a deterministic PDE, i.e. when $\sigma = 0$, it is known [3] that the relation is the following

$$\alpha \frac{h}{k^2} \leq 1/2. \tag{3}$$

To overcome this difficulty, we can modify (2) to get the *implicit* scheme:

$$\begin{aligned} Y_0 &= x_0 \\ Y_{j+1} &= Y_j + AY_{j+1}h + \sigma W_j^H, \quad j = 1, \ldots, N \end{aligned} \tag{4}$$

and calculate Y_{j+1} by solving the following systems of equations

$$(I - Ah)Y_{j+1} = Y_j + \sigma W_j^H, \quad j = 1, \ldots, N,$$

where I denotes the identity matrix. Instead of calculating each of the unknown vector Y_j by a separate trivial formula, we must now solve this system of equations to give the values simultaneously. This task is however not very difficult, because the matrix $(I - Ah)$ has a special form, it is a three-diagonal symmetric positive definite matrix. From the theory of PDEs, it is known that the implicit scheme has one big advantage, namely there is no such constraint as (3). In [5] it was believed that something similar holds also for this implicit scheme for SPDEs with additive noise. However, additional numerical experiments showed rather unstable behaviour also for the implicit scheme. Therefore, a relation similar to (3) (depending probably also on H) will have to be taken into account.

In both schemes (2) and (4) there has been a slight modification to take into account the boundary conditions.

Let us now suppose that we have one path observation $X^{x_0}(t,x), t \in [0,T], T \gg 1$, of the solution to (1). For the test purposes we use again the already calculated numerical solution as our observation. From this path we want to estimate the value of the parameter α. We may either consider that we know the parameter σ or we can use its estimate from the previous section. To estimate the parameter α we will again use [5], Theorem 3.2.1.

Let $z(x) = e_1(x) = \sqrt{2/L}\sin(\pi x/L)$, $x \in [0,L]$.

First of all we consider a *reference* equation (1) with the parameter $\alpha = 1$. For this equation we calculate numerically

$$\langle Q_T z, z\rangle = \langle Q_T e_1, e_1\rangle_V = \sigma^2 \int_0^T \int_0^T e^{-\lambda_1(u+v)} \phi(u-v)\, du\, dv.$$

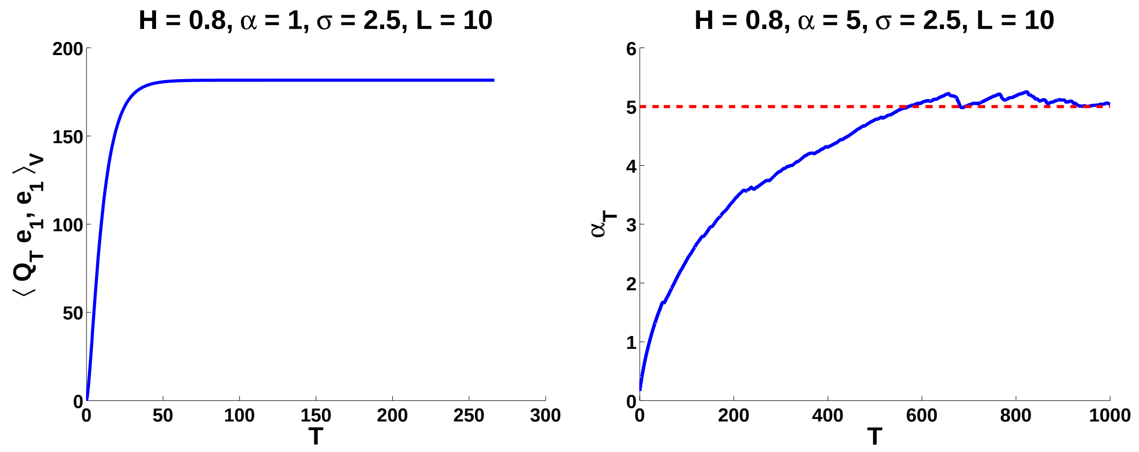
We now turn back to the equation (1) with unknown parameter α (i.e. not necessarily equal to one). From observed path of the solution we have to calculate the average

$$\frac{1}{T}\int_0^T \left|\langle X^{x_0}(t,x), z(x)\rangle_V\right|^2 dt = \frac{1}{T}\int_0^T \left|\int_0^L X^{x_0}(t,x) z(x)\, dx\right|^2 dt$$

for sufficiently large T. Using Theorem 3.2.1 from [5], we are now able to calculate the estimate

$$\hat{\alpha}_T := \left(\frac{\langle Q_\infty z, z\rangle_V}{\frac{1}{T}\int_0^T \left|\langle X^{x_0}(t,x), z\rangle_V\right|^2 dt}\right)^{\frac{1}{2H}} = \left(\frac{\langle Q_\infty e_1, e_1\rangle_V}{\frac{1}{T}\int_0^T \left|\int_0^L X^{x_0}(t,x) e_1(x)\, dx\right|^2 dt}\right)^{\frac{1}{2H}}.$$

In the following figure on the left, we can see how $\langle Q_T e_1, e_1\rangle_V$ converges to $\langle Q_\infty e_1, e_1\rangle_V$ for particular values of α, σ, H and L. In picture on the right, we can see how $\hat{\alpha}_T$ converges to the true value of parameter α for large values of T and for particular value of H (σ and L appears in the solution). We can see that a large time has to be considered to obtain a reasonable estimate and due to the fluctuations also some average.



3. Concluding remarks

It has to be pointed out that there are still no appropriate convergence results for the numerical methods used, therefore we used the modified Euler-Maruyama method only for demonstration purposes. Moreover, for some combinations of parameter constants, especially α and σ, the results of these numerical experiments are not so convincing. Hence, further research in the area of numerical solution to stochastic evolution equations driven by fractional Brownian motion is needed.

References

[1] D.J. Higham: *An algorithmic introduction to numerical simulation of stochastic differential equations.* SIAM Rev. **43**, 3, 2001, 525–546 (electronic).

[2] P.E. Kloeden and E. Platen: *Numerical solution of stochastic differential equations.* Applications of Mathematics (New York), Springer-Verlag, Berlin, **23**, 1992.

[3] K.W. Morton and D.F. Mayers: *Numerical solution of partial differential equations. An introduction.* Cambridge University Press, Cambridge, 1994.

[4] J. Pospíšil: *An introduction to numerical methods for solving stochastic differential equations.* In: J. Chleboun, P. Přiklryl, K. Segeth (eds), *Proceedings to Programms and Algorithms of Numerical Mathematics* **11**, Mathematical Institute AS CR, Prague, 2002, pp. 190–201.

[5] J. Pospíšil: *On parameter estimates in stochastic evolution equations driven by fractional Brownian motion.* Ph.D. Thesis, University of West Bohemia in Plzeň, 2005, iv+88.

[6] Z.-M. Yin: *New methods for simulation of fractional Brownian motion.* J. Comput. Phys. **127**, 1, 1996, 66–72.

NUMERICAL INTEGRATION IN THE DISCONTINUOUS GALERKIN METHOD FOR ELLIPTIC PROBLEMS*

Aleš Prachař, Karel Najzar

1. Introduction

The use of numerical integration is considered as one of *variational crimes* often committed in practical applications of the finite element method. In the theoretical study of the Discontinuous Galerkin method exact integration is almost exclusively considered. We refer to one of exceptions, [5], where the effect of numerical integration applied to the evaluation of nonlinear convective terms is studied while the diffusion term is set in such a way that application of appropriate quadrature formulae yields exact integration.

The aim of this paper is to study various aspects of the use of numerical integration for the evaluation of integrals appearing in Discontinuous Galerkin formulations of a linear elliptic (diffusion) problem. Our aim is to obtain sufficient conditions on quadrature formulae which ensure that there exists a unique solution of the corresponding discrete problem. Moreover, we shall study how the use of numerical integration impacts error estimates.

Let us consider simple model problem

$$\begin{aligned} -\nabla \cdot (A(x)\nabla u) &= f \quad \text{in } \Omega, & (1)\\ u &= g_D \quad \text{on } \Gamma_D, & (2)\\ (A(x)\nabla u)\cdot \boldsymbol{n} &= g_N \quad \text{on } \Gamma_N. & (3) \end{aligned}$$

We assume that $\Omega \subset \mathbb{R}^2$ is a bounded polygonal domain with a Lipschitz-continuous boundary $\partial\Omega$ divided into two disjoint parts Γ_D and Γ_N such that $\partial\Omega = \overline{\Gamma}_D \cup \overline{\Gamma}_N$, where $\mathrm{meas}_1(\Gamma_D) \neq 0$.

We assume that functions f, g_D and g_N are sufficiently regular. Further, let there exists a constant $K > 0$ such that the matrix $A \in [W^{1,\infty}(\Omega)]^{2\times 2}$ satisfies

$$\boldsymbol{\xi}^T A(x)\boldsymbol{\xi} \geq K\ \boldsymbol{\xi}^T \cdot \boldsymbol{\xi} \quad \forall\ \boldsymbol{\xi} \in \mathbb{R}^2,\ \text{a. e. on } \Omega. \tag{4}$$

[*] The research of A. Prachař was a part of the research project MSM 0001066902 financed by the Ministry of Education of the Czech Republic. The research of K. Najzar was a part of the research project MSM 0021620839 financed by the Ministry of Education of the Czech Republic and was partly supported by the project No. 201/04/1503 of the Grant Agency of the Czech Republic.

2. Discontinuous Galerkin formulation

Let $\mathcal{T}_h$ be a conforming triangulation of $\overline{\Omega}$. We shall denote individual triangles of $\mathcal{T}_h$ by T and put $h_T = \operatorname{diam}(T)$. For the theoretical study it is convenient to consider that a family of triangulations $\{\mathcal{T}_h\}_{h>0}$ of a domain Ω is *regular*, see [4].

Let $\mathcal{E}_h$ stand for the set of all *edges* of $\mathcal{T}_h$. These edges represent the interfaces between pairs of adjacent elements, or sides of triangles lying on the boundary of the domain Ω. Let us distinguish sets of *internal edges* ($\mathcal{E}_h^I$), *Dirichlet edges* ($\mathcal{E}_h^D$) and *Neumann edges* ($\mathcal{E}_h^N$). The length of the edge $S \in \mathcal{E}_h$ will be denoted by $|S|$.

Let us define the space $V_h = \{v \in L^2(\Omega)\ ;\ v|_T \in P_p(T)\ \forall T \in \mathcal{T}_h\}$, where $P_p(T)$ is the space of polynomials of degree at most $p \geq 1$ on T.

For $S \in \mathcal{E}_h^I$ let us denote by T_1 and T_2 the two triangles sharing the edge S. Then we define the *average* on the side S by $\{u\} = \frac{1}{2}((u|_{T_1})|_S + (u|_{T_2})|_S)$ and $\{u\} = u|_S$ for $S \in \mathcal{E}_h^D$. The *jump* on $S \in \mathcal{E}_h^I$ is defined by $[\![u]\!] = (u|_{T_1})|_S - (u|_{T_2})|_S$ and again $[\![u]\!] = u|_S$ for $S \in \mathcal{E}_h^D$. Orientation of the vector $\boldsymbol{n}$ is in accord with the orientation of the *jump*.

For the Discontinuous Galerkin formulation let us introduce bilinear forms $a^+, a^- : V_h \times V_h \to \mathbb{R}$,

$$\begin{aligned} a^{\pm}(u,v) = & \sum_{T\in\mathcal{T}_h} \int_T (A\nabla u)\cdot\nabla v\,dx - \sum_{S\in\mathcal{E}_h^I\cup\mathcal{E}_h^D} \int_S \{A\nabla u\}\cdot\boldsymbol{n}[\![v]\!]\,ds \\ & \pm \sum_{S\in\mathcal{E}_h^I\cup\mathcal{E}_h^D} \int_S \{A\nabla v\}\cdot\boldsymbol{n}[\![u]\!]\,ds + \sum_{S\in\mathcal{E}_h^I\cup\mathcal{E}_h^D} \frac{\sigma_S}{|S|}\int_S [\![u]\!][\![v]\!]\,ds \end{aligned} \tag{5}$$

and linear functionals $L^+, L^- : V_h \to \mathbb{R}$ by

$$L^{\pm}(v) = \int_\Omega fv\,dx + \sum_{S\in\mathcal{E}_h^N}\int_S g_N v\,ds + \sum_{S\in\mathcal{E}_h^D}\int_S g_D\left[\frac{\sigma_S}{|S|}v \pm (A\nabla v)\cdot\boldsymbol{n}\right]ds, \tag{6}$$

where $\sigma_S \in \mathbb{R}$, $S \in \mathcal{E}_h^I \cup \mathcal{E}_h^D$, is a chosen penalty parameter. The bilinear form $a^+(\cdot,\cdot)$ introduces the *Nonsymmetric Interior Penalty Galerkin (NIPG)* variant (cf. [8]) while the bilinear form $a^-(\cdot,\cdot)$ is *symmetric* for symmetric matrix A. Therefore, we shall speak of the *Symmetric Interior Penalty Galerkin (SIPG)* variant (cf. [1]). Our discrete Discontinuous Galerkin formulation then becomes:

$$\text{find} \quad u_h \in V_h \quad \text{such that} \qquad a^{\pm}(u_h, v) = L^{\pm}(v) \qquad \forall v \in V_h. \tag{7}$$

It is well-known that there exists a unique solution of (7) if certain properties of penalty parameters are satisfied, see, e. g., [2]. Moreover, if the weak solution u of (1)–(3) satisfies $u \in H^{p+1}(\Omega)$, we are able to show that

$$|\!|\!| u-u_h |\!|\!|^2 := \sum_{T\in\mathcal{T}_h} |u-u_h|^2_{1,2,T} + \sum_{S\in\mathcal{E}_h^I\cup\mathcal{E}_h^D} \frac{1}{|S|}\|[\![u-u_h]\!]\|^2_{0,2,T} \leq C\sum_{T\in\mathcal{T}_h} h_T^{2p}|u|^2_{p+1,2,T} \tag{8}$$

with the constant $C > 0$ independent of u and h.

3. Problem with numerical integration

The core of this paper is to explain what happens if all the terms in (5) and (6) are evaluated with the aid of appropriately chosen quadrature formulae. For $\varphi \in C^0(T)$, $T \in \mathcal{T}_h$ and $\psi \in C^0(S)$, $S \in \mathcal{E}_h$, we use approximations

$$\int_T \varphi(x)\,dx \approx \sum_{\alpha=1}^{n_T} \omega_\alpha^T \varphi(x_\alpha^T), \qquad \int_S \psi\,ds \approx \sum_{\alpha=1}^{n_S} \nu_\alpha^S \psi(x_\alpha^S), \tag{9}$$

where $\omega_\alpha^T, \nu_\alpha^S > 0$ are integration weights and $x_\alpha^T \in T, x_\alpha^S \in S$ are integration points. Let us denote by $a_h^\pm(\cdot,\cdot)$ the result of application of numerical integration to the bilinear form $a^\pm(\cdot,\cdot)$ and similarly for the right-hand side. Related problem

$$\text{find} \quad \tilde{u}_h \in V_h \quad \text{such that} \qquad a_h^\pm(\tilde{u}_h, v) = L_h^\pm(v) \qquad \forall v \in V_h \tag{10}$$

makes sense assuming that all the integrands have their point values well-defined which requires higher regularity of data. The most important step in the verification of assumptions of the Lax–Milgram lemma is the proof of uniform V_h-ellipticity.

Lemma 1 *Let the quadrature formula for the integration of the first term of (5) be exact for polynomials from $P_{2p-2}(T)$ and/or let the set of quadrature points $\{x_\alpha^T\}_{\alpha=1}^{n_T}$ contain a $P_{p-1}(T)$-unisolvent subset. Let us assume that the quadrature formula for the penalty term is exact for polynomials of degree $\le 2p$ and/or let the set of quadrature points $\{x_\alpha^S\}_{\alpha=1}^{n_S}$ contain a $P_p(S)$-unisolvent subset. If penalty parameters σ_S are sufficiently large, there exists a constant $\hat{c} > 0$ independent of h such that*

$$\hat{c}|\!|\!|v|\!|\!|^2 \le a_h^\pm(v,v) \qquad \textit{for all } v \in V_h.$$

Proof: According to Theorem 4.1.2 in [4] there exists a constant $c_1 > 0$ such that

$$K|v|_{1,2,T}^2 \le Kc_1 \sum\nolimits_{\alpha=1}^{n_T} \omega_\alpha^T \sum\nolimits_{i=1}^{2} |\partial_i v(x_\alpha^T)|^2 \le c_1 \sum\nolimits_{\alpha=1}^{n_T} \omega_\alpha^T \sum\nolimits_{i,j=1}^{2} (a_{ij}\partial_j v \partial_i v)(x_\alpha^T).$$

Similar technique is used to show that if the set of quadrature points $\{x_\alpha^S\}_{\alpha=1}^{n_S}$ contains a $P_p(\hat{S})$-unisolvent subset then $\|[\![v]\!]\|_{0,2,S}^2 \le c_2 \sum_{\alpha=1}^{n_S} \nu_\alpha^S [\![v(x_\alpha^S)]\!]^2$ with some $c_2 > 0$. For the *NIPG* variant the proof is finished, because other terms disappear if the same quadrature formula is used for their evaluation. The requirement $\sigma_S > 0$ is necessary. In the case of the *SIPG* formulation we take into account the inequality

$$\sum\nolimits_{\alpha=1}^{n_S} \nu_\alpha^S [\{A\nabla v\} \cdot \boldsymbol{n}[\![v]\!]](x_\alpha^S) \;\; \le \;\; c_3|S|^{-1/2}\|[\![v]\!]\|_{0,2,S} \sum\nolimits_{T:S\subset\partial T} |v|_{1,2,T},$$

where $c_3 > 0$ depends on p, shape regularity, properties of weights of quadrature formulae and properties of the matrix A. By the Young's inequality we find that

$$a_h^-(v,v) \;\; \ge \;\; \sum_{T\in\mathcal{T}_h} K\left(\frac{1}{c_1} - \frac{1}{\delta}\right)|v|_{1,2,T}^2 + \sum_{S\in\mathcal{E}_h^I\cup\mathcal{E}_h^D} \frac{1}{|S|}\|[\![v]\!]\|_{0,2,S}^2 \left(\frac{\sigma_S}{c_2} - \frac{6\delta c_3^2}{K}\right).$$

For $\delta > c_1$ and $\sigma_S > 6\delta c_2 c_3^2/K$ round brackets are positive. □

Since $a_h^{\pm}(\cdot,\cdot)$ is a continuous bilinear form and $L_h^{\pm}(\cdot)$ is a continuous linear functional on the space V_h we find by the Lax–Milgram lemma that:

Theorem 2 *There exists a unique solution of discrete problem (10).*

4. Errors of quadrature formulae

The next step is to express the error induced by the use of numerical integration. We shall denote E_T and E_S error functionals of numerical integration in a similar way as in [4, 5], i. e., $E_T(\varphi) = \int_T \varphi\,dx - \sum_{\alpha=1}^{n_T} \omega_\alpha^T \varphi(x_\alpha^T)$, etc.

Lemma 3 *Let $u, v \in P_p(T)$, $a \in W^{l+1,\infty}(T)$ and $S \subset \partial T$. Let the quadrature formula on the triangle be exact for polynomials of degree $\leq p+l-2$ and let the (edge) quadrature formula be exact for polynomials of degree $\leq p+l-1$. Then there exists a constant $C > 0$ independent of h such that*

$$|E_T(a\partial_j u \partial_i v)| \leq C h_T^l \|a\|_{l,\infty,T} \|\partial_j u\|_{p-1,2,T} \|\partial_i v\|_{0,2,T}, \quad 1 \leq i,j \leq 2, \tag{11}$$

$$\Big| E_S(a\partial_j uv) \Big| \leq C h_T^l \|a\|_{l,\infty,S} \|\partial_j u\|_{p-1,2,T} |S|^{-1/2} \|v\|_{0,2,S}, \quad 1 \leq i,j \leq 2, \tag{12}$$

$$|E_S(a\partial_j vu)| \leq C h_T^l \|a\|_{l+1,\infty,T} \|\partial_j v\|_{0,2,T} \|u\|_{p,2,T}, \quad 1 \leq i,j \leq 2. \tag{13}$$

Proof: Estimate (11) follows as in [4]. Other two terms are also estimated with the aid of suitable transformation to the reference edge, the Bramble-Hilbert lemma (cf. [4]) and also the estimate

$$|v|_{j,r,S} \leq c|S|^{1/r} |T|^{-1/s} |v|_{j,s,T}, \qquad 1 \leq r,s \leq +\infty \tag{14}$$

for all $v \in P_p(T)$, $S \subset \partial T$ and $j \leq p$, see proof of Lemma 1 in [7]. □

Let us now move our attention to the error arising from the integration of terms on the right-hand side. Let us focus on boundary conditions.

Lemma 4 *Let $g_N \in H^{p+1}(\Gamma_N)$ and $g_D \in H^{p+1}(\Gamma_D)$. Let the (edge) quadrature formula be exact for polynomials of degree $\leq 2p$. Then there exists a constant $C > 0$ such that*

$$\begin{aligned} |E_S(g_N v)| &\leq C|S|^{p+1/2} |g_N|_{p+1,2,S} \|v\|_{0,2,T}, \\ \frac{\sigma_S}{|S|} |E_S(g_D v)| &\leq C\sigma_S |S|^{p+1/2} |g_D|_{p+1,2,S} |S|^{-1/2} \|[\![v]\!]\|_{0,2,S}. \end{aligned}$$

If the (edge) quadrature formula is exact for polynomials of degree $\leq 2p-1$ and $A \in [W^{p+1,\infty}(S)]^{2\times 2}$ then

$$|E_S((A\nabla v)\cdot \boldsymbol{n} g_D)| \leq C\|A\|_{p+1,\infty,S} |S|^{p+1/2} \|g_D\|_{p+1,2,S} |v|_{1,2,T}.$$

Proof: It is based on results from [5]. □

5. Error estimate for the problem with numerical integration

In order to estimate the impact of the use of numerical integration, let us state the main idea of the *first Strang lemma* ([4], Theorem 4.1.1) which says that

$$\begin{aligned}\hat{c}|\!|\!|\tilde{u}_h - v_h|\!|\!|^2 \quad \leq \quad & a^{\pm}(u - v_h, \tilde{u}_h - v_h) + \{a^{\pm}(v_h, \tilde{u}_h - v_h) - a_h^{\pm}(v_h, \tilde{u}_h - v_h)\} \\ & + \{L_h^{\pm}(\tilde{u}_h - v_h) - L^{\pm}(\tilde{u}_h - v_h)\}, \qquad (15)\end{aligned}$$

where $\tilde{u}_h$ is defined by (10), v_h is arbitrary element of the space V_h and u is the weak solution of (1)–(3) and $\hat{c}$ comes from Lemma 1. Our aim is to estimate two *consistency errors* arising as the result of the numerical integration.

Theorem 5 *If the quadrature formula on triangles is exact for polynomials of degree $\leq 2p-2$, the (edge) integration formula for the second and third term in (5) is exact for polynomials of degree $\leq 2p-1$ and if the penalty term is integrated exactly, there exists a constant $C > 0$ independent of h such that*

$$|a^{\pm}(v_h, \tilde{u}_h - v_h) - a_h^{\pm}(v_h, \tilde{u}_h - v_h)| \leq C\|A\|_{p+1,\infty,\Omega} \left(\sum_{T \in \mathcal{T}_h} h_T^{2p} \|v_h\|_{p,2,T}^2 \right)^{1/2} |\!|\!|\tilde{u}_h - v_h|\!|\!|,$$

where $\tilde{u}_h, v_h \in V_h$.

Proof: Follows from estimates presented in Lemma 3. □

Theorem 6 *Let the quadrature formula on triangles be exact for polynomials of degree $\leq 2p-2$ and let $f \in W^{p,r}(\Omega)$ with $r \geq 2$. Let assumptions of Lemma 4 be satisfied. There exists a constant $C > 0$ independent of h such that*

$$\begin{aligned}|L_h(\tilde{u}_h - v_h) - L(\tilde{u}_h - v_h)| \quad \leq \quad & Ch^p \|f\|_{p,r,\Omega} |\!|\!|\tilde{u}_h - v_h|\!|\!| + Ch^{p+1/2}\Big(|g_N|_{p+1,2,\Gamma_N} \\ & + |g_D|_{p+1,2,\Gamma_D} + \|A\|_{p+1,\infty,\Omega} \|g_D\|_{p+1,2,\Gamma_D} \Big) |\!|\!|\tilde{u}_h - v_h|\!|\!|,\end{aligned}$$

for $\tilde{u}_h, v_h \in V_h$.

Proof: Is a consequence of Lemma 4, Theorem 4.1.5 in [4] and the Broken Poincaré inequality, see [3]. We also use $|S| \leq h = \max_{T \in \mathcal{T}_h} h_T$. □

Since other terms can be estimated with the aid of the interpolation theory we are ready to write the main theorem.

Theorem 7 *Let all the assumptions of Theorem 5 and Theorem 6 be satisfied and let the approximate bilinear form $a^{\pm}(\cdot,\cdot)$ be uniformly V_h-elliptic. Then there exists a constant $C > 0$ independent of h such that*

$$\begin{aligned}|\!|\!|u - \tilde{u}_h|\!|\!| \quad \leq \quad & Ch^p(|u|_{p+1,2,\Omega} + \|A\|_{p+1,\infty,\Omega}\|u\|_{p+1,2,\Omega} + \|f\|_{p,r,\Omega}) \\ & + Ch^{p+1/2}\Big(|g_N|_{p+1,2,\Gamma_N} + |g_D|_{p+1,2,\Gamma_D} + \|A\|_{p+1,\infty,\Omega}\|g_D\|_{p+1,2,\Gamma_D} \Big),\end{aligned}$$

where $\tilde{u}_h$ is defined in (10) and $u \in H^{p+1}(\Omega)$ is the weak solution of (1)–(3).

6. Conclusion

In this paper the effect of numerical integration in the Discontinuous Galerkin formulations for linear elliptic problem was studied. Sufficient conditions which ensure that the discrete problem is uniquely solvable were found. Moreover, if quadrature formulae of a certain precision are used then the order of accuracy (compared with the case without numerical integration) is not decreased.

If we compare these results with the conforming finite element method (see, e. g., [4]), we find that higher regularity of the matrix A is needed for the proof of error estimate. Theorem 5 has again a simple interpretation: The order of convergence is not decreased if the integration formulae yield exact integration of the bilinear form in the case that A is a constant matrix (cf. Remark 4.1.8 in [4]).

Let us also note that numerical results (not reported here) illustrate reasonable degree of agreement with presented theoretical results.

References

[1] D.N. Arnold: *An interior penalty finite element method with discontinuous element.* SIAM J. Numer. Anal. **19**, 1982, 742–760.

[2] D. Arnold, F. Brezzi, B. Cockburn, D. Marini: *Unified analysis of discontinuous Galerkin methods for elliptic problems.* SIAM J. Numer. Anal. **39**, 2001, 1749–1779.

[3] S.C. Brenner: *Poincaré–Friedrichs inequalities for piecewise H^1 functions.* SIAM J. Numer. Anal. **41**, 2003, 306–324.

[4] P. Ciarlet: *The finite element method for elliptic problems.* North Holland, Amsterdam, 1978.

[5] M. Feistauer, V. Sobotíková: *On the effect of numerical integration in the DGFEM for nonlinear convection-diffusion problems.* Submitted to *Numer. Methods Partial Differential Equations.*

[6] A. Prachař: *On discontinuous Galerkin method and semiregular family of triangulations.* Appl. Math. **51**, 2006, 605–618.

[7] P. Sváček, K. Najzar: *Numerical solution of problems with non-linear boundary conditions.* Math. Comput. Simulation **61**, 2003, 219–228.

[8] B. Rivière, M.F. Wheeler, V. Girault: *Improved energy estimates for interior penalty, constrained and discontinuous Galerkin methods for elliptic problems.* Part I. Computational Geosciences **3**, 1999, 337–360.

AN UNSTEADY NUMERICAL SOLUTION OF VISCOUS COMPRESSIBLE FLOWS IN A CHANNEL*

Petra Punčochářová, Karel Kozel, Jiří Fürst, Jaromír Horáček

Abstract

The work deals with numerical solution of unsteady flows in a 2D channel where one part of the channel wall is changing as a given function of time. The flow is described by the system of Navier-Stokes equations for compressible (laminar) flows. The flow has low velocities (low Mach numbers) and is numerically solved by the finite volume method. Moving grid of quadrilateral cells is considered in the form of conservation laws using ALE (Arbitrary Lagrangian-Eulerian) method.

1. Introduction

This work presents an unsteady numerical solution of the system of Navier-Stokes equations for compressible laminar flow. An unsteady flow is caused by the moving part of the channel wall. The authors investigated flows in two types of channels, in an nonsymmetric channel and in a symmetric channel. The flow in a symmetric channel can represent a very simple model of airflow in a human vocal tract.

The numerical solution was obtained by the explicit central finite volume version of MacCormack scheme on a grid of quadrilateral cells.

2. Mathematical model

The 2D system of Navier-Stokes equations (1) was used as mathematical model to describe an unsteady viscous compressible laminar flow in a channel. The system is expressed in non-dimensional form:

$$W_t + F_x + G_y = \frac{1}{Re}(R_x + S_y), \tag{1}$$

$W = [\rho, \rho u, \rho v, e]^T$ is the vector of conservative variables, F and G are the vectors of inviscid fluxes, R and S are the vectors of viscous fluxes. Variable ρ denotes the density, u and v are the components of the velocity vector, and e is the total energy per unit volume. Static pressure p in the inviscid fluxes is expressed by the equation of state. Reynolds number $Re = \rho_\infty u_\infty H/\eta_\infty$ is computed from the inflow variables: $\rho_\infty = \text{const}$, $u_\infty = \text{const}$, $\eta_\infty = \text{const}$, and H is the inflow width of the channel. Non-dimensional dynamic viscosity $\eta = 1/Re$ is constant in our cases.

[*]This work was supported by grant GA AV ČR No. IAA 2007 60613 and by Research Plan MSM No. 6840770010.

2.1. Mathematical formulation

For the numerical solution, the domain of solution D and the boundary conditions have to be defined. Two channels were tested. The first is an nonsymmetric channel and the second is a symmetric channel. Boundary conditions were considered in the following form:

a) Upstream conditions: three components of W are given, the pressure is extrapolated.

b) Downstream conditions: the pressure is given, the other values are extrapolated or $\partial W/\partial\vec{n} = 0$ where $\vec{n}$ is an outlet normal vector.

c) On the solid wall, the velocity vector and the normal derivative of temperature vanish that is $(u, v)_{\text{wall}} = \vec{0}$ and $\partial T/\partial\vec{n} = 0$.

d) At the axis of symmetry, $(u, v) \cdot \vec{n} = 0$ is considered.

Figure 1 shows D_1, the domain of solution, which is called the nonsymmetric channel. The upper and lower boundary represent solid walls. The lower solid wall of the channel has a time changing part between points A and B that is a given function of time $g_1(t)$.

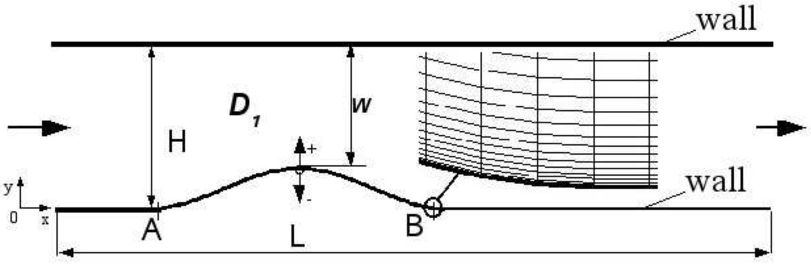


Fig. 1: *Domain of solution D_1 (the nonsymmetric channel).*

Figure 2 shows D_2, the domain of solution in the symmetric channel. The computational domain is only the lower half of the channel. Its upper boundary coincides with the axis of symmetry. The lower boundary represents a solid wall. The part of the wall between points A and B is changing and determined by $g_2(t)$, a given function of time.

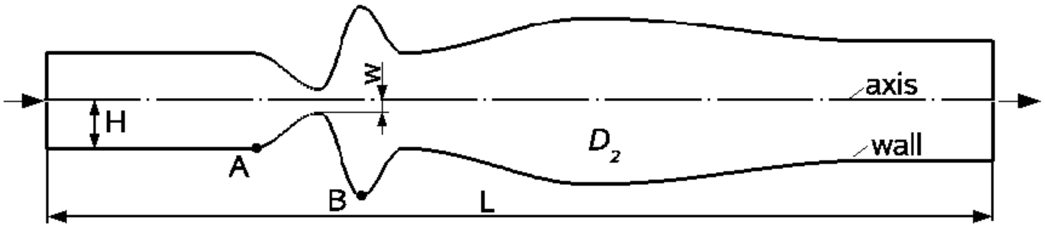


Fig. 2: *Domain of solution D_2 (the symmetric channel).*

3. Numerical solution

The numerical solution of the above two-dimensional problems is obtained by the finite volume method in the cell centered form (FVM) on a grid of quadrilateral cells.

The bounded domain D is divided into mutually disjoint sub-domains $D_{i,j}$ (e.g. quadrilateral cells). Equations (1) are integrated over subdomain $D_{i,j}$. By using the Green formula and the Mean Value Theorem, we can write the integral form of FVM:

$$W_t|_{i,j} = \frac{-1}{\mu_{i,j}} \left[\oint_{\partial D_{i,j}} (Fdy - Gdx) - \oint_{\partial D_{i,j}} (Rdy - Sdx) \right], \tag{2}$$

where $\mu_{i,j} = \int\int_{D_{i,j}} dxdy$ stand for the volumes of the cells. We get FVM in the differential form:

$$\frac{W_{i,j}^{n+1} - W_{i,j}^{n}}{\Delta t} = \frac{-1}{\mu_{i,j}} \sum_k [(\tilde{F}_k - \tilde{R}_k)\Delta y_k - (\tilde{G}_k - \tilde{S}_k)\Delta x_k], \tag{3}$$

where $\Delta t = t^{n+1} - t^n$ is the time step. Physical fluxes F, G, R, S on edge k of cell $D_{i,j}$ are replaced by numerical fluxes $\tilde{F}, \tilde{G}, \tilde{R}, \tilde{S}$. The particular choice of numerical fluxes and of the time derivative approximation depend on a chosen numerical scheme.

3.1. Numerical scheme

The explicit MacCormack (MC) scheme in the predictor-corrector form is used to approximate system (1). This scheme is 2nd order accurate in time and space.

$$\begin{aligned} W_{i,j}^{n+1/2} &= W_{i,j}^{n} - \frac{\Delta t}{\mu_{i,j}} \sum_{k=1}^{4} [(\tilde{F}_k^n - s_{1k}W_k^n - \tilde{R}_k^n)\Delta y_k - (\tilde{G}_k^n - s_{2k}W_k^n - \tilde{S}_k^n)\Delta x_k], \\ \bar{W}_{i,j}^{n+1} &= \frac{1}{2}(W_{i,j}^{n} + W_{i,j}^{n+1/2}) - \frac{\Delta t}{2\mu_{i,j}} \sum_{k=1}^{4} [(\tilde{F}_k^{n+1/2} - s_{1k}W_k^{n+1/2} - \tilde{R}_k^{n+1/2})\Delta y_k \\ &\quad -(\tilde{G}_k^{n+1/2} - s_{2k}W_k^{n+1/2} - \tilde{S}_k^{n+1/2})\Delta x_k]. \end{aligned} \tag{4}$$

Equation (4) represents the MC scheme for a viscous flow in a domain with a moving grid of quadrilateral cells. The moving grid in an unsteady domain is described by using the Arbitrary Lagrangian-Eulerian (ALE) method which defines the projection of reference domain D_0 to a time-dependent domain D_t [1]. Consequently, additional fluxes $\vec{s}_k W_k$ appear in the MC scheme, where vector $\vec{s}_k$ represents the speed of edge k. The approximations of conservative variable W_k and diffusive components R_k, S_k on edge k are central. The second derivatives (dissipative terms) on an edge are approximated using dual volumes [2] as is shown in Figure 3.

The inviscid numerical fluxes are approximated as follows:

$$\begin{aligned} &\tilde{F}_1^n = F_{i,j}^n, \quad \tilde{F}_1^{n+1/2} = F_{i+1,j}^{n+1/2}, \quad \tilde{F}_3^n = F_{i-1,j}^n, \quad \tilde{F}_3^{n+1/2} = F_{i,j}^{n+1/2}, \\ &\tilde{G}_2^n = G_{i,j}^n, \quad \tilde{G}_2^{n+1/2} = G_{i,j+1}^{n+1/2}, \quad \tilde{G}_4^n = G_{i,j-1}^n, \quad \tilde{G}_4^{n+1/2} = G_{i,j}^{n+1/2}, \quad \text{etc.} \end{aligned} \tag{5}$$

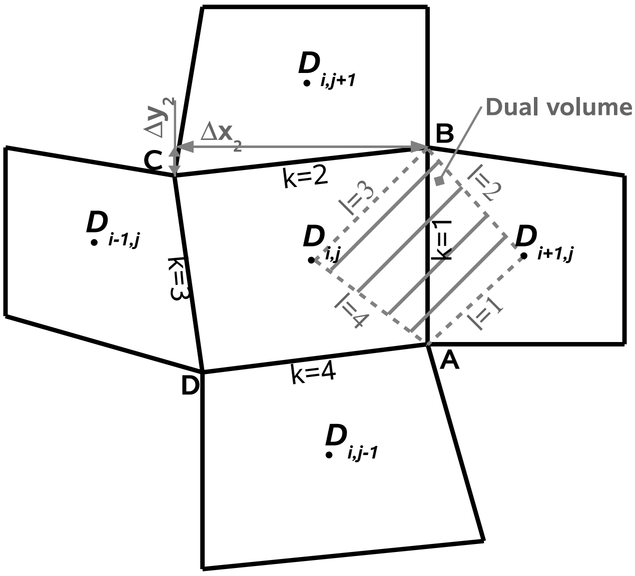


Fig. 3: *Finite volume $D_{i,j}$, dual volume V_k'.*

The last term of the MC scheme is the Jameson artificial dissipation $AD(W_{i,j})^n$, which is added to schemes with higher order of accuracy to stabilize the numerical solution:

$$AD(W_{i,j})^n = C_1\gamma_1(W^n_{i+1,j} - 2W^n_{i,j} + W^n_{i-1,j}) + C_2\gamma_2(W^n_{i,j+1} - 2W^n_{i,j} + W^n_{i,j-1}), \quad (6)$$

where $C_1, C_2 \in R$ are constants and the normed pressure gradients have the form:

$$\gamma_1 = \frac{|p^n_{i+1,j} - 2p^n_{i,j} + p^n_{i-1,j}|}{|p^n_{i+1,j}| + 2|p^n_{i,j}| + |p^n_{i-1,j}|}, \quad \gamma_2 = \frac{|p^n_{i,j+1} - 2p^n_{i,j} + p^n_{i,j-1}|}{|p^n_{i,j+1}| + 2|p^n_{i,j}| + |p^n_{i,j-1}|}. \quad (7)$$

Then we can compute a vector of conservative variables W at a new time level t^{n+1}:

$$W^{n+1}_{i,j} = \bar{W}^{n+1}_{i,j} + AD(W_{i,j})^n. \quad (8)$$

Stability condition of the scheme (on a regular orthogonal grid) limits the time step

$$\Delta t \leq CFL \left(\frac{|u_{\max}| + c}{\Delta x_{\min}} + \frac{|v_{\max}| + c}{\Delta y_{\min}} + \frac{2}{Re}\Big(\frac{1}{\Delta x^2_{\min}} + \frac{1}{\Delta y^2_{\min}}\Big)\right)^{-1}, \quad (9)$$

where c denotes the local speed of sound, $CFL < 1$, and the minimal step of the grid in the y-direction is $\Delta y_{\min} \approx 1/\sqrt{Re}$ due to boundary layer.

4. Numerical results

For numerical computation, domains D_1 and D_2 (see Figures 1, 2) are covered with a grid of quadrilateral cells. The cells near the wall boundary have successive refinement in the y-direction (due to the existing boundary layer) as shown in detail in Figure 1. The results are depicted as Mach number isolines and as the velocity vectors.

4.1. Numerical results in domain D_1

The length and width of domain D_1 are $L = 12$ and $H = 0.5$, and D_1 contains 600 × 50 cells. Parametres considered for computation: the outflow pressure is $p_2 = 0.9p_\infty$ and it corresponds to the inflow Mach number $M_\infty = 0.120$ and $Re = 5 \cdot 10^5$. Figure 4 shows the steady solution of viscous laminar flow in the nonsymmetric channel where the moving part of the solid wall (see Figure 1) is fixed. The maximum Mach number in the domain was computed to be $M_{\max} = 0.345$. Figure 5 (a, b, c, d, e) shows the development of unsteady viscous compressible laminar flows in domain D_1 at several time layers starting by the second period. For the computation of the unsteady solution, the steady solution was used as the initial state.

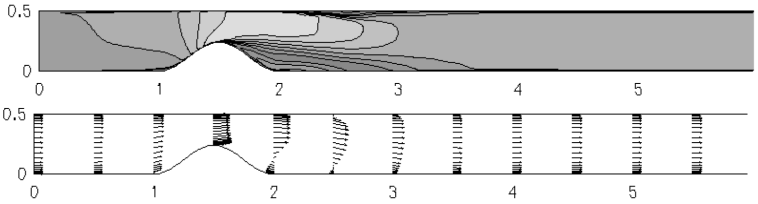
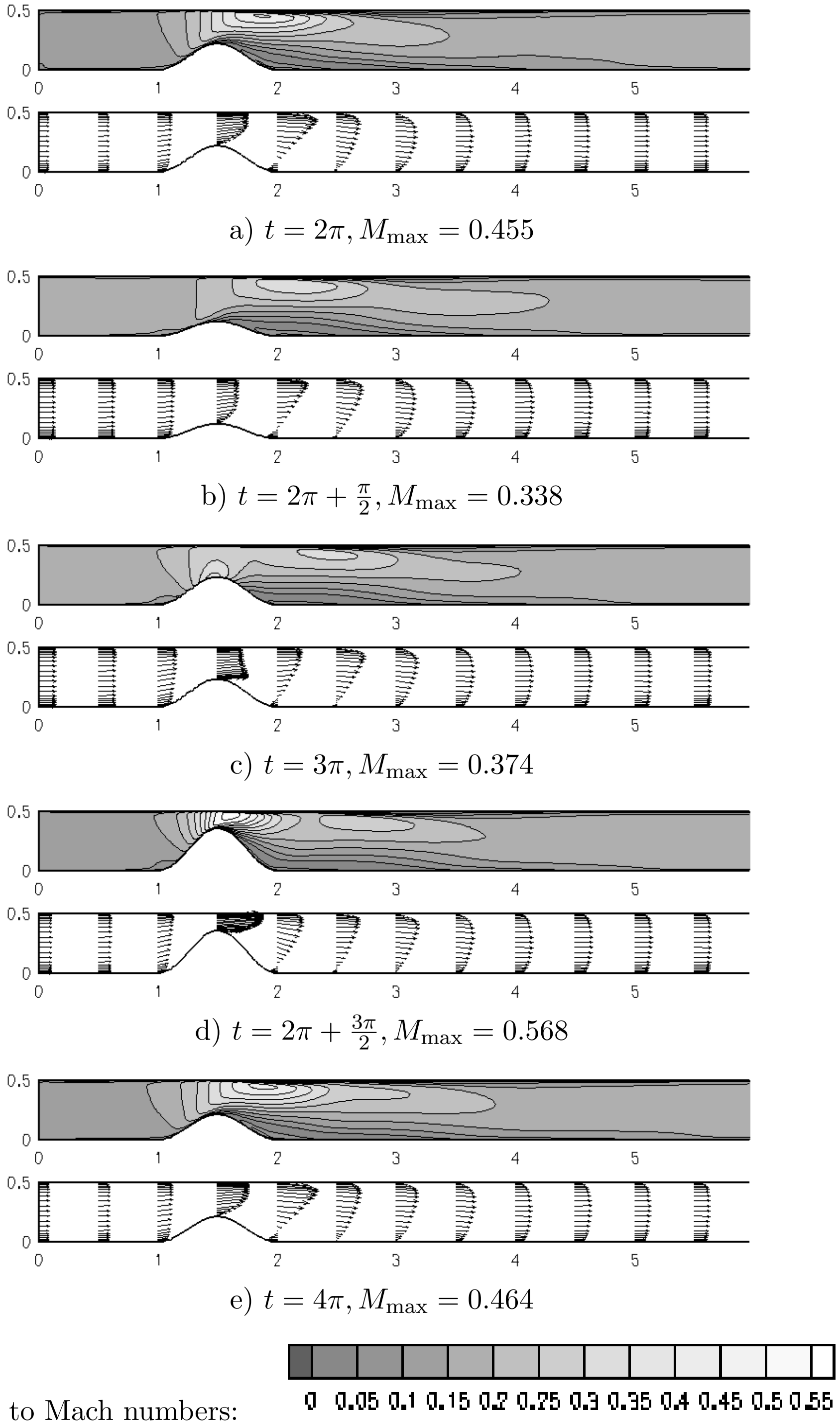


Fig. 4: *The steady solution of a viscous laminar flow in the nonsymmetric channel,* $p_2 = 0.9p_\infty$, $Re = 5 \cdot 10^5$, $M_{\max} = 0.345$, 600 × 50 *cells.*

4.2. Numerical results in domain D_2

The length and width of domain D_2 are $L = 8$ and $H = 0.4$, and D_2 contains 400 × 50 cells. Parametres considered for computation: the inlet Mach number $M_\infty = 0.02$, the dimension frequency of the solid wall between points A, B (see Figure 2) is $f_{\text{dim}} = 20\,\text{Hz}$ and $Re = 9 \cdot 10^3$. These values approximately correspond to the real flow in the human vocal tract. Figure 6a) shows the steady solution of viscous laminar flow in the symmetric channel where the moving part of the solid wall is fixed. The maximum Mach number in the domain was computed, $M_{\max} = 0.096$. Figure 6b) shows convergence to a steady solution that is observed using L_2 norm of momentum residuals (ρu). It seems to be relatively good for this case with a very low Mach number. Figure 7 (a, b, c, d, e) shows development of unsteady viscous compressible laminar flows in domain D_2 at several time layers starting by the third period. For computation of the unsteady solution, the steady solution was used as the initial state. In Figure 7b), one can see typical behaviour with choking flows in a very narrow part of the channel and with the time development of flow including separation domains. The geometry of domain D_2 and the boundary conditions represent a simple model of flow in the human vocal tract [3, 4].

We also tried to compute both cases without the artificial dissipation $AD(W_{i,j})^n$. In this case, however, the convergence to the steady state was not satisfactory.



a) $t = 2\pi, M_{\max} = 0.455$

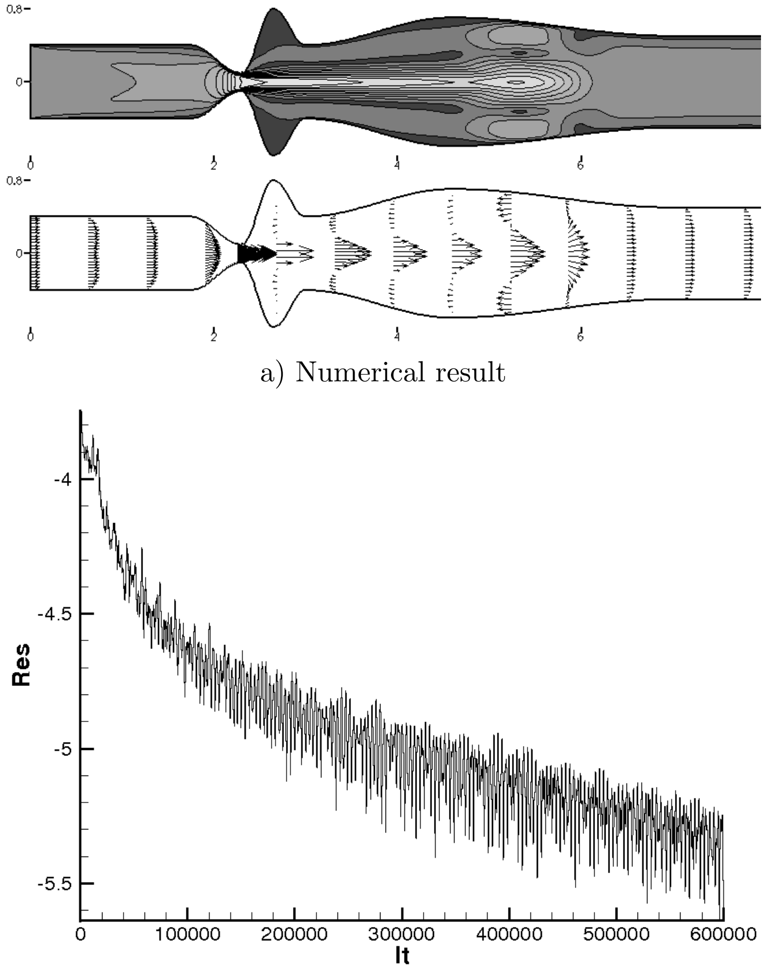
b) $t = 2\pi + \frac{\pi}{2}, M_{\max} = 0.338$

c) $t = 3\pi, M_{\max} = 0.374$

d) $t = 2\pi + \frac{3\pi}{2}, M_{\max} = 0.568$

e) $t = 4\pi, M_{\max} = 0.464$

Fig. 5: *The unsteady solution of a viscous laminar flow in the nonsymmetric channel,* $p_2 = 0.9p_\infty$, $Re = 5 \cdot 10^5$, 600×50 *cells.*



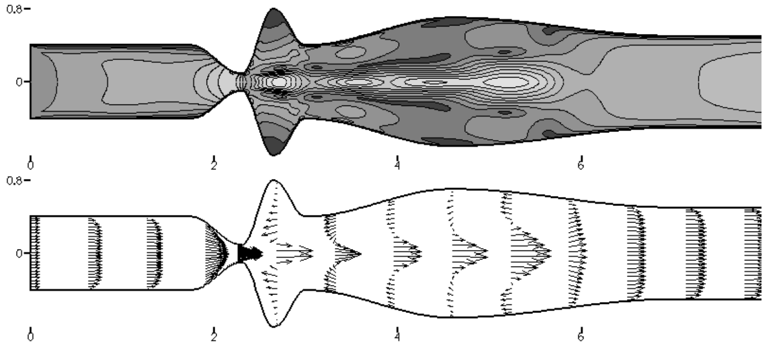
a) Numerical result

b) Convergence to a steady solution – residual vs. number of iterations

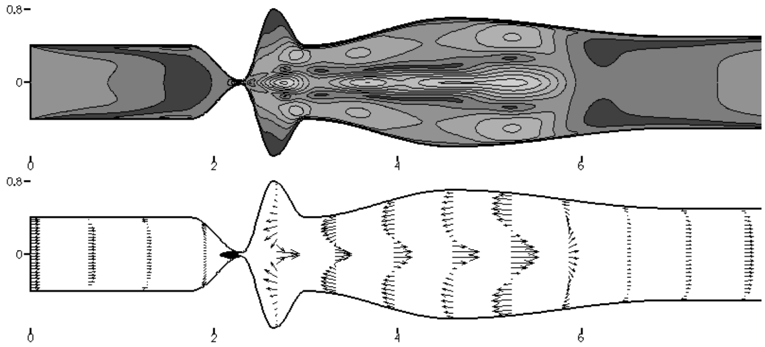
Fig. 6: *The steady solution of a viscous laminar flow in the symmetric channel,* $M_\infty = 0.02$*,* $Re = 9 \cdot 10^3$*,* $M_{\max} = 0.096$*,* 400×50 *cells.*

5. Summary

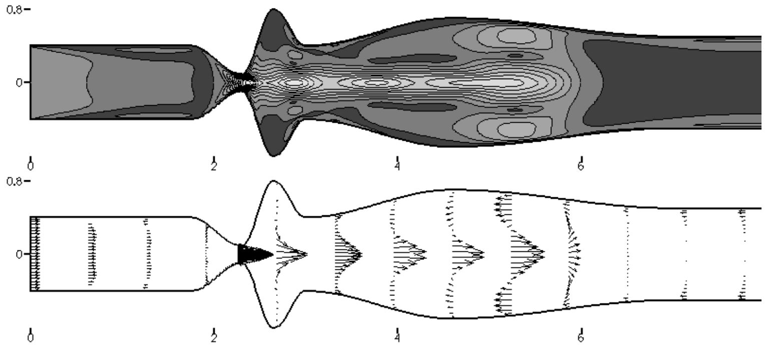
The calculation numerical approximations of steady state solutions for inviscid compressible flows with very low Mach numbers is a very difficult task and special methods have to be used. For viscous compressible problems, the method described above can be successfully used for the steady as well as unsteady numerical solutions of flows with low Mach numbers.



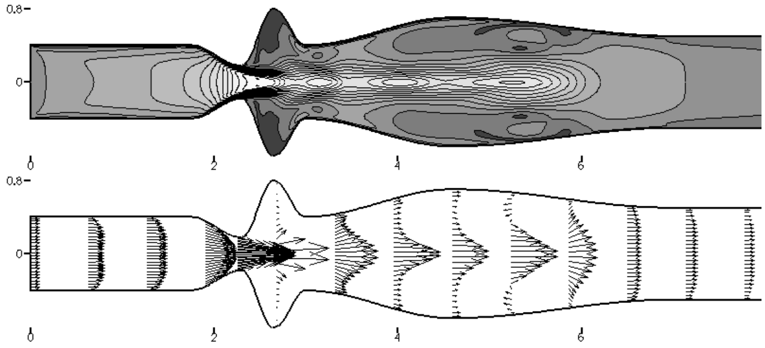
a) $t = 4\pi, M_{\max} = 0.094$



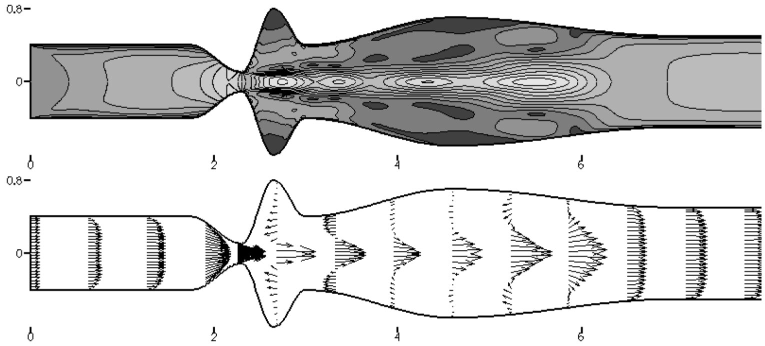
b) $t = 4\pi + \frac{\pi}{2}, M_{\max} = 0.077$



c) $t = 5\pi, M_{\max} = 0.129$



d) $t = 4\pi + \frac{3\pi}{2}, M_{\max} = 0.145$



e) $t = 6\pi, M_{\max} = 0.102$

Legend to Mach numbers:

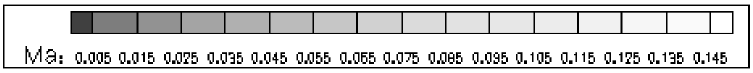


Fig. 7: *The unsteady solution of a viscous laminar flow in the symmetric channel,* $M_\infty = 0.02$*,* $Re = 9 \cdot 10^3$*,* 400×50 *cells.*

References

[1] R. Honzátko, K. Kozel, J. Horáček: *Flow over a profile in a channel with dynamical effects.* In: Proceedings in Applied Mathematics **4**, 1, 2004, 322–323.

[2] J. Fürst, M. Janda, K. Kozel: *Finite volume solution of 2D and 3D Euler and Navier-Stokes equations.* In: J. Neustupa, P. Penel (eds), Mathematical fluid mechanics, Berlin, 2001.

[3] P. Punčochářová, K. Kozel, J. Fürst: *Unsteady, subsonic inviscid and viscous flows in a channel.* In: Fluid Dynamics 2005, IT CAS CZ, 2005, 125–128.

[4] J. Horáček, P. Šidlof, G. Švec: *Numerical simulation of self-oscillations of human vocal folds with Hertz model of impact forces.* Journal of Fluid and Structures **20**, 2005, 853–869.

ON SOME A POSTERIORI ERROR ESTIMATION RESULTS FOR THE METHOD OF LINES*

Karel Segeth, Pavel Šolín

Abstract

The paper is an attempt to present an (incomplete) historical survey of some basic results of residual type estimation procedures from the beginning of their development through contemporary results to future prospects. Recently we witness a rapidly increasing use of the *hp*-FEM which is due to the well-established theory. However, the conventional a posteriori error estimates (in the form of a single number per element) are not enough here, more complex estimates are needed, and this can be the way to obtain them.

1. Introduction

In the 1990's, the subject of a posteriori error estimation with the finite element method and adaptive solution procedures started its very rapid development. Many results for the solution of linear and nonlinear elliptic partial differential equations were reached and first results for the solution of nonlinear parabolic partial differential equations were published. A pioneering paper in this field was [2].

We present some basic results from this time period and continue to contemporary results and future prospects of this approach. A rich contemporary source of knowledge is the book [3].

Recently, we witness a rapidly increasing use of the *hp*-FEM. We are concerned with this subject in the conclusion of this paper. We also refer to some published numerical results and their accuracy.

We introduce a nonlinear parabolic model problem and its finite element solution in Sections 2 and 3 while in Section 4 we are concerned with a posteriori error estimation. We quote some adaptive grid refinement procedures and speak about further prospects in Section 5.

We apologize to all colleagues whose names and contributions to the subject were not, for the lack of space, mentioned in this paper.

2. Model problem

We introduce a nonlinear parabolic model problem. For the sake of brevity, we consider only one equation with a scalar solution u and a single 1D space variable x. All the results can be generalized to a system of parabolic equations and a d-dimensional space variable.

*This work was supported by Grant 201/04/1503 of the Czech Science Foundation and by Institutional Research Plan AV0Z10190503 of the Academy of Sciences of the Czech Republic.

Let us consider the problem

$$\frac{\partial u}{\partial t}(x,t) - \frac{\partial}{\partial x}\Big(a(u)\frac{\partial u}{\partial x}(x,t)\Big) + f(u) = 0 \quad \text{for} \quad 0 < x < 1, \quad 0 < t \le T \tag{1}$$

with the boundary conditions

$$u(0,t) = u(1,t) = 0, \quad 0 \le t \le T, \tag{2}$$

and the initial condition

$$u(x,0) = u_0(x), \quad 0 < x < 1, \tag{3}$$

where u_0 is a given function.

Let us assume

$$\begin{aligned} &0 < \mu \le a(s) \le M, \quad s \in R, \\ &|a(r) - a(s)| \le L|r - s|, \\ &|f(r) - f(s)| \le L|r - s|, \quad r, s \in R, \end{aligned}$$

where μ, M, and L are positive constants. We need some more assumptions for some of the proofs, see [9].

In the standard way we introduce the *weak solution* $u(x,t) \in H^1([0,T], H_0^1(0,1))$ of the model problem by the identity

$$\Big(\frac{\partial u}{\partial t}, v\Big) + \Big(a(u)\frac{\partial u}{\partial x}, \frac{\partial v}{\partial x}\Big) + (f(u), v) = 0 \tag{4}$$

to be satisfied for $t \in (0,T]$ by all test functions $v \in H_0^1(0,1)$ and the identity

$$\Big(a(u_0)\frac{\partial u}{\partial x}, \frac{\partial v}{\partial x}\Big) = \Big(a(u_0)\frac{\partial u_0}{\partial x}, \frac{\partial v}{\partial x}\Big) \tag{5}$$

to be satisfied for $t = 0$ also by all test functions $v \in H_0^1(0,1)$. This latter identity corresponds to the initial condition. Some other weak formulations of the initial condition are also possible. We use the symbol $(\cdot,\cdot)$ for the usual $L_2(0,1)$ inner product and $\|\cdot\|_1$ for the $H^1(0,1)$ norm.

3. Semidiscrete approximate solution

To define the finite element solution of the problem (1) to (3), we start with the space discretization (*semidiscretization*). We choose a partition

$$0 = x_0 < x_1 < \cdots < x_{N-1} < x_N = 1 \tag{6}$$

of the space interval $[0,1]$ and further put

$$h_j = x_j - x_{j-1}, \quad j = 1, \ldots, N, \quad \text{and} \quad h = \max_{j=1,\ldots,N} h_j.$$

We use the notation

$$(v,w)_j = \int_{x_{j-1}}^{x_j} v(x)w(x)\,\mathrm{d}x$$

for the inner product restricted to the interval $[x_{j-1},x_j]$, and similarly $\|v\|_j$ and $\|v\|_{1,j}$ for the restricted $L_2(0,1)$ and $H^1(0,1)$ norms.

On the partition (6), we construct a finite dimensional subspace

$$S_0^{N,p} = \Big\{ V \mid V \in H_0^1(0,1),\ V(x) = \sum_{j=1}^{N-1} V_{j1}\varphi_{j1}(x) + \sum_{j=1}^{N}\sum_{k=2}^{p} V_{jk}\varphi_{jk}(x)\Big\}$$

of the space $H_0^1(0,1)$.

The functions φ_{jk} are chosen to form a *hierarchic basis*. For $k=1$, we put

$$\begin{aligned}\varphi_{j1}(x) &= (x-x_{j-1})/h_j, \quad x_{j-1} \le x < x_j,\\ &= (x_{j+1}-x)/h_{j+1}, \quad x_j \le x \le x_{j+1},\\ &= 0 \quad \text{otherwise.}\end{aligned}$$

These functions are the well known *hat* or *chapeau functions*. For $k>1$, we further put

$$\begin{aligned}\varphi_{jk}(x) &= \frac{\sqrt{2(2k-1)}}{h_j}\int_{x_{j-1}}^{x} \mathrm{P}_{k-1}(y)\,\mathrm{d}y, \quad x_{j-1} \le x \le x_j,\\ &= 0 \quad \text{otherwise,}\end{aligned}$$

where P_k is a *Legendre polynomial* transformed from $[-1,1]$ to $[x_{j-1},x_j]$. These functions (primitive functions to Legendre polynomials) are called the *Lobatto polynomials* or *bubble functions*. The idea of hierarchic basis functions was first introduced in the book [11].

The principal idea of the method of lines is the space semidiscretization while the time variable remains continuous. We look for the *semidiscrete approximate solution* $\bar{U}(x,t) \in H^1([0,T],S_0^{N,p})$ in the form

$$\bar{U}(x,t) = \sum_{j=1}^{N-1} \bar{U}_{j1}(t)\varphi_{j1}(x) + \sum_{j=1}^{N}\sum_{k=2}^{p} \bar{U}_{jk}(t)\varphi_{jk}(x).$$

We require that the identities

$$\Big(\frac{\partial \bar{U}}{\partial t}, V\Big) + \Big(a(\bar{U})\frac{\partial \bar{U}}{\partial x}, \frac{\partial V}{\partial x}\Big) + (f(\bar{U}),V) = 0, \quad t \in (0,T], \quad V \in S_0^{N,p}, \tag{7}$$

$$\Big(a(u_0)\frac{\partial \bar{U}}{\partial x}, \frac{\partial V}{\partial x}\Big) = \Big(a(u_0)\frac{\partial u_0}{\partial x}, \frac{\partial V}{\partial x}\Big), \quad t = 0, \quad V \in S_0^{N,p}, \tag{8}$$

that correspond to the identities (4), (5), be satisfied. The basis functions as well as test functions are thus chosen from the same space $S_0^{N,p}$. Note that after substituting φ_{il} for the test functions $V(x)$ in (7), we obtain an initial value problem for a system of ordinary differential equations with the initial condition (8). Other initial conditions can be employed, too.

The ordinary differential system (7) with the initial condition (8) for the unknown coefficients $\bar{U}_{jk}(t)$ is then solved by standard numerical software.

4. Analysis of residual a posteriori semidiscrete error indicators

Let us denote the error of the semidiscrete solution $\bar{U}(x,t)$ by

$$e(x,t) = u(x,t) - \bar{U}(x,t).$$

We introduce the finite dimensional space

$$\hat{S}_0^{N,p+1} = \Big\{\hat{V} \mid \hat{V} \in H_0^1(0,1), \hat{V}(x) = \sum_{j=1}^{N} \hat{V}_j \varphi_{j,p+1}(x)\Big\}$$

and approximation of the error

$$\bar{E}(x,t) = \sum_{j=1}^{N} \bar{E}_j(t)\varphi_{j,p+1}(x).$$

Note that we look for approximation of the error in the finite element space of piecewise polynomials of the degree $p+1$.

Some results on the semidiscrete error for the case of linear parabolic equations and systems were given in [1], [7].

Some time later, they were generalized to the nonlinear case. If we subtract the identities (7), (8) that define the semidiscrete solution $\bar{U}$ from the identities (4), (5) that define the weak solution u we obtain for $\bar{E}(x,t) \in H^1([0,T], \hat{S}_0^{N,p+1})$ the initial value problem for the system of ordinary differential equations

$$\begin{aligned} &\Big(\frac{\partial \bar{E}}{\partial t}, \hat{V}\Big)_j + \Big(a(\bar{U}+\bar{E})\frac{\partial \bar{E}}{\partial x}, \frac{\partial \hat{V}}{\partial x}\Big)_j \qquad (9)\\ &\quad = -(f(\bar{U}+\bar{E}),\hat{V})_j - \Big(\frac{\partial \bar{U}}{\partial t}, \hat{V}\Big)_j - \Big(a(\bar{U}+\bar{E})\frac{\partial \bar{U}}{\partial x}, \frac{\partial \hat{V}}{\partial x}\Big)_j, \ t \in (0,T],\ \hat{V} \in \hat{S}_0^{N,p+1}, \end{aligned}$$

with the initial condition

$$\Big(a(u_0)\frac{\partial \bar{E}}{\partial x}, \frac{\partial \hat{V}}{\partial x}\Big)_j = \Big(a(u_0)\frac{\partial (u_0 - \bar{U})}{\partial x}, \frac{\partial \hat{V}}{\partial x}\Big)_j, \quad t = 0, \quad \hat{V} \in \hat{S}_0^{N,p+1}. \qquad (10)$$

The quantity $\bar{E}$ defined by (9), (10) is called the *nonlinear parabolic error indicator*. Note that (9), (10) is a nonlinear problem for the unknowns $\bar{E}_j(t)$. For the practical computation, these equations can be added to the system (7), (8) for finding the semidiscrete solution $\bar{U}_{jk}(t)$. Further, note that the equations of the system (9) are uncoupled.

There are some simplifications that allow for more efficient computation while, asymptotically, the error indicator is of the same quality. The *linear parabolic error indicator*, and *nonlinear* and *linear elliptic error indicator* are defined in an analogous way. The detailed description can be found in, e.g., [5] or [9]. The following theorem is proven in [9] for the nonlinear parabolic error indicator.

Theorem. *Let the weak solution $u(x,t)$ given by* (4), (5) *be smooth, let $\bar{U}(x,t)$ and $\bar{E}$ be given by* (7), (8) *and* (9), (10), *respectively. Let $\|e\|_1 \geq Ch^p$. Then*

$$\lim_{h\to 0} \frac{\|\bar{E}\|_1}{\|e\|_1} = 1.$$

The quantity $\|\bar{E}\|_1/\|e\|_1$ is called the *effectivity index.* For the linear parabolic as well as linear elliptic error indicator (but not for the nonlinear elliptic one), this theorem is proven in [9], too.

Analysis of the semidiscrete error does not include analysis of the error of solution of the corresponding system of ordinary differential equations in time. In practice, this system is solved by standard software that admits the required accuracy to be given by the user. This required accuracy is then prescribed several orders less than the total prescribed accuracy of the fully discrete solution. There are several papers concerned with the analysis of fully discrete error, see, e.g., [5], [12], [13].

5. Space h- and hp-adaptive procedures

Procedures that can adapt the space grid are very often used. They are usually based on the *principle of the equidistribution of error* that requires

$$\|e\|_{1,i} = \|e\|_{1,j}, \quad i,j = 1,\dots,N.$$

This requirement is applied to the error indicator $\bar{E}$,

$$\|\bar{E}\|_{1,i} = \|\bar{E}\|_{1,j}, \quad i,j = 1,\dots,N.$$

Several such procedures have been published, e.g. the *dynamic grid adaptation* in [1], *grading function grid adaptation* in [8], etc. We successfully tested the above introduced error indicators on these procedures.

We witness a rapidly increasing use of the *hp*-FEM for solving elliptic as well as parabolic problems. For this adaptive finite element method, however, the conventional error estimates (in the form of a single number per element) are not enough. There are numerous options how a higher-order element can be refined because of the interplay between h and p. Thus the estimates of higher-order derivatives of the error are required. Moreover, these *hp*-procedures are particularly important if the space variable is a vector. In these problems, the *reference solution* usually serves as the source of the a posteriori error estimation. Both the ideas and computational procedures of the *hp*-FEM are presented in, e.g., [4], [6], [10].

References

[1] S. Adjerid, J.E. Flaherty, Y.J. Wang: *A posteriori error estimation with finite element method of lines for one-dimensional parabolic system.* Numer. Math. **65**, 1993, 1–21.

[2] I. Babuška, W.C. Rheinboldt: *A-posteriori error estimates for the finite element method.* Internat. J. Numer. Methods Engrg. **12**, 1978, 1597–1615.

[3] I. Babuška, T. Strouboulis: *The finite element method and its reliability.* Oxford, Clarendon Press 2001.

[4] L. Demkowicz, W. Rachowitz, P. Devloo: *A fully automatic hp-adaptivity.* J. Sci. Comput. **17**, 2002, 117–142.

[5] P.K. Moore: *A posteriori error estimation with finite element semi- and fully discrete methods for nonlinear parabolic equations in one space dimension.* SIAM J. Numer. Anal. **31**, 1994, 149–169.

[6] P.K. Moore: *Applications of Lobatto polynomials to an adaptive finite element method: A posteriori error estimates for hp-adaptivity and grid-to-grid interpolation.* Numer. Math. **94**, 2003, 367–401.

[7] K. Segeth: *A posteriori error estimates for parabolic differential systems solved by the finite element method of lines.* Appl. Math. **39**, 1994, 415–444.

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

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

[10] P. Šolín, K. Segeth, I. Doležel: *Higher-order finite element methods.* Boca Raton, FL, Chapman & Hall/CRC 2004.

[11] B. Szabó, I. Babuška: *Finite element analysis.* New York, John Wiley & Sons 1991.

[12] T. Vejchodský: *Fully discrete error estimation by the method of lines for a nonlinear parabolic problem.* Appl. Math. **48**, 2003, 129–151.

[13] R. Verfürth: *A posteriori error estimates for nonlinear problems: $L^r(0,T;W^{1,\rho}(\Omega))$-error estimates for finite element discretizations of parabolic equations.* Numer. Methods Partial Differential Equations **14**, 1998, 487–518.

ON A FINITE ELEMENT METHOD APPLICATION IN AEROELASTICITY[*]

Petr Sváček

Abstract

The subject of this paper is the numerical simulation of aeroelastic problems. The interaction of two-dimensional incompressible viscous flow and a vibrating airfoil is modelled. The solid airfoil, which can rotate around the elastic axis and oscillate in the vertical direction, is considered. The numerical simulation consists of the finite element solution of the Navier-Stokes equations coupled with the system of ordinary differential equations describing the airfoil motion. The stabilization procedure is of GLS type. The developed numerical approximation is applied on an aeroelastic problem.

1. Introduction

The mathematical model of relevant technical cases consists of (incompressible) fluid model and (elastic) structure model. In this paper mainly the numerical approximation of fluid motion is addressed. In order to approximate the Navier-Stokes equations several methods can be used. Besides finite differences, the finite volume method can be used for the approximation (for application of finite volume method to solution of incompressible flow cf. [5]). In the present paper the finite element method is used for approximation of the fluid motion. In this case one needs to treat several sources of instability: one caused by the fact that Babuška-Brezzi condition needs to be satisfied in order to guarantee the stability of the scheme, the other source of instability related to the fact that extremely large Reynolds numbers are involved in the problem (Re $\approx$ 10^5–10^6).

2. Mathematical model

The incompressible viscous air flow is described with the aid of Navier-Stokes system of equations written in so-called Arbitrary Lagrangian-Eulerian (ALE) form, cf. [6], [2]. In order to clarify the method, we start with the definition of an ALE mapping $\mathcal{A}_t$: We assume that the mapping $\mathcal{A}_t$ is a given C^1 continuous bijective mapping from the reference (original) configuration Ω_0 onto the computational domain at a time t, i.e. the current configuration Ω_t.

$$\mathcal{A}_t : \Omega_0 \mapsto \Omega_t, \qquad Y \mapsto y(t,Y) = \mathcal{A}_t(Y).$$

[*] The author acknowledges the financial support of the Ministry of Education of the Czech Republic by Research Plan MSM 6840770003 and also the financial support of the Grant Agency of the Czech Republic by the project No. 201/05/P142.

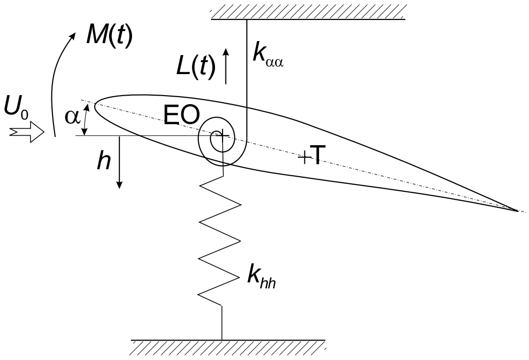


Fig. 1: *The elastic support of the airfoil hanging on translational and rotational springs.*

The time derivative with respect to the reference frame Ω_0 is called the *ALE derivative*, i.e.

$$\frac{D^{\mathcal{A}_t} f}{Dt} = \frac{\partial f}{\partial t} + (\mathbf{w}_g \cdot \nabla) f. \tag{1}$$

With the aid of the ALE derivative $D^{\mathcal{A}_t}\mathbf{u}/Dt$, the Navier-Stokes system of equations is rewritten as follows

$$\frac{D^{\mathcal{A}_t}\mathbf{u}}{Dt} - \nu \triangle \mathbf{u} + \Big((\mathbf{u} - \mathbf{w}_g) \cdot \nabla \Big) \mathbf{u} + \nabla p = 0, \qquad \nabla \cdot \mathbf{u} = 0, \qquad \text{in } \Omega_t\,, \tag{2}$$

where by Ω_t we denote the computational domain occupied by fluid at time $t \in (0, T)$, $\mathbf{u}$ denotes the velocity vector, p denotes the kinematic pressure (i.e. the dynamic pressure divided by the air density), and the domain velocity vector is denoted by $\mathbf{w}_g$. On the boundary $\partial\Omega$ we prescribe suitable boundary conditions. First, the boundary $\partial\Omega$ is decomposed into three distinct parts, i.e. $\partial\Omega = \Gamma_{W_t} \cup \Gamma_D \cup \Gamma_O$. On Γ_D and Γ_{W_t} a Dirichlet boundary conditions are prescribed, i.e.

$$\text{a)} \qquad \mathbf{u} = \mathbf{u}_D \text{ on } \Gamma_D, \qquad\qquad \text{b)} \qquad \mathbf{u} = \mathbf{w}_g \text{ on } \Gamma_{W_t}. \tag{3}$$

The latter part of the boundary is the only moving part of the boundary. The boundary Γ_O represents the outlet, where the following boundary condition is prescribed

$$\left[-(p - p_{ref})\mathbf{n} - \frac{1}{2}(\mathbf{u} \cdot \mathbf{n})^- \mathbf{u} + \nu \frac{\partial \mathbf{u}}{\partial \mathbf{n}} \right] \Bigg|_{\Gamma_O} = 0, \tag{4}$$

where p_{ref} is a reference pressure value (e.g. zero).

If Γ_O is the outflowing part of the boundary, i.e. $(\mathbf{u} \cdot \mathbf{n})^- = 0$, the condition (4) is equivalent to the well known *do-nothing* boundary condition. We consider the weak formulation (2–4) in the Sobolev spaces $(H^1(\Omega))^2$ and $L^2(\Omega)$ for the velocities and pressures, respectively.

The fluid model is coupled with the nonlinear equations of motion for a flexibly supported airfoil, see [7]

$$\begin{aligned} m\,\ddot{h} + S_\alpha\,\ddot{\alpha}\,\cos\alpha - S_\alpha\,\dot{\alpha}^2\,\sin\alpha + k_{hh}\,h &= -L(t), \\ S_\alpha\,\ddot{h}\,\cos\alpha + I_\alpha \ddot{\alpha} + k_{\alpha\alpha}\,\alpha &= M(t). \end{aligned} \tag{5}$$

where h and α denotes the vertical (downwards oriented) and the rotational (clockwise oriented) displacements, respectively, whereas L and M denote the aerodynamical lift force and torsional moment. The mathematical models (5) and (2) are coupled with the evaluation of aerodynamical forces defined by

$$L = -\int_{\Gamma_{W_t}} \sum_{j=1}^{2} \sigma_{2j} n_j dS, \qquad M = -\int_{\Gamma_{W_t}} \sum_{i,j=1}^{2} \sigma_{ij} n_j r_i^{\text{ort}} dS, \tag{6}$$

where $r_1^{\text{ort}} = -(x_{EO2} - x_2)$, $r_2^{\text{ort}} = x_{EO1} - x_1$ and σ_{ij} is the stress tensor, cf. [3].

3. Numerical approximation

First, let us start with an equidistant discretization of the time interval $[0, T]$ with the time step Δt, i.e. $t_k = k \cdot \Delta t$ for $k = 0, 1, 2, \ldots$. Let $\mathbf{u}^n, p^n$ denote approximations of the velocity vector $\mathbf{u}$ and the pressure p evaluated at the time t_n, i.e. $\mathbf{u}^n \approx \mathbf{u}(t_n)$ and $p^n \approx p(t_n)$. The ALE derivative of the velocity vector $\mathbf{u}$ is approximated by

$$\frac{D^{\mathcal{A}_t} f}{Dt} \approx \frac{3\mathbf{u}^{n+1} - 4\hat{\mathbf{u}}^n + \hat{\mathbf{u}}^{n-1}}{2\Delta t}, \tag{7}$$

where the velocity $\mathbf{u}^{n+1}$ denotes the approximate velocity at time t_{n+1} and the velocities $\hat{\mathbf{u}}^n, \hat{\mathbf{u}}^{n-1}$ are the velocities at previous time steps t_n and t_{n-1} transformed from domains $\Omega_{t_n}, \Omega_{t_{n-1}}$ onto the current computational domain $\Omega_{t_{n+1}}$, i.e., $\hat{\mathbf{u}}^n \equiv \mathbf{u}^n \left(\mathcal{A}_{t_n}\left(\mathcal{A}_{t_{n+1}}^{-1}(y)\right)\right)$,$\hat{\mathbf{u}}^{n-1} \equiv \mathbf{u}^{n-1}\left(\mathcal{A}_{t_{n-1}}\left(\mathcal{A}_{t_{n+1}}^{-1}(y)\right)\right)$. The time difference formula is then involved in the problem (2), i.e.

$$\begin{aligned} \frac{3\mathbf{u}^{n+1} - 4\hat{\mathbf{u}}^n + \hat{\mathbf{u}}^{n-1}}{2\Delta t} - \nu \triangle \mathbf{u} + \Big((\mathbf{u} - \mathbf{w}_g) \cdot \nabla\Big)\mathbf{u} + \nabla p &= 0, \\ \nabla \cdot \mathbf{u} &= 0, \qquad \text{in } \Omega_t \end{aligned} \tag{8}$$

and the system of equations (8) is formulated weakly. The components of the approximate solution are sought in the space $X_{\mathbf{\Delta}}$. $X_{\mathbf{\Delta}}$ denotes the finite element space of Taylor-Hood elements, i.e. piecewise quadratic velocity components and linear pressures.

The *stabilized discrete problem* reads: Find $U = (\mathbf{u}, p) \in X_{\mathbf{\Delta}}$ such that

$$\mathbf{a}(U, U, V) + L_{\mathbf{\Delta}}(U, U, V) + P_{\mathbf{\Delta}}(U, V) = f(V) + F_{\mathbf{\Delta}}(V)$$

for all $V = (\mathbf{v}, q) \in X_{\mathbf{\Delta}}^0$ ($X_{\mathbf{\Delta}}^0$ denotes the space of functions from $X_{\mathbf{\Delta}}$ being zero on the Dirichlet part of boundary). The terms $\mathbf{a}(\cdot, \cdot, \cdot)$ and $f(\cdot)$ are the standard Galerkin terms defined as

$$\begin{aligned} \mathbf{a}(U^*, U, V) &= \frac{3}{2\Delta t}(\mathbf{u}, \mathbf{v})_\Omega + \nu (\nabla \mathbf{u}, \nabla \mathbf{v})_\Omega + \Big(\left((\mathbf{u} - \mathbf{w}_g^{n+1}) \cdot \nabla\right)\mathbf{u}, \mathbf{v}\Big)_\Omega \\ &\quad - (p, \nabla \cdot \mathbf{v})_\Omega + (\nabla \cdot \mathbf{u}, q)_\Omega, \\ f(V) &= \frac{1}{2\Delta t}\left(4\hat{\mathbf{u}}^n - \hat{\mathbf{u}}^{n-1}, \mathbf{v}\right)_\Omega - \int_{\Gamma_O} p_{\text{ref}} \mathbf{v} \cdot \mathbf{n}\, dS, \end{aligned} \tag{9}$$

the terms $L_{\boldsymbol{\Delta}}(\cdot,\cdot)$ and $F_{\boldsymbol{\Delta}}(\cdot)$ are GLS (Galerkin Least Squares) additional stabilization terms defined as

$$\begin{aligned}
L_{\boldsymbol{\Delta}}(U^*,U,V) &= \sum_{K\in\tau_{\boldsymbol{\Delta}}} \delta_K\Big(\frac{3}{2\Delta t}\mathbf{u} - \nu\triangle\mathbf{u} + ((\mathbf{u}^* - \mathbf{w}_g)\cdot\nabla)\,\mathbf{u} + \nabla p, \psi(\mathbf{u},q)\Big)_K,\\
F_{\boldsymbol{\Delta}}(V) &= \sum_{K\in\tau_{\boldsymbol{\Delta}}} \delta_K\Big(\frac{1}{2\Delta t}(4\hat{\mathbf{u}}^n - \hat{\mathbf{u}}^{n-1}), \psi(\mathbf{u},q)\Big)_K,
\end{aligned} \tag{10}$$

where $\psi(\mathbf{u},q) \equiv ((\mathbf{u}^* - \mathbf{w}_g)\cdot\nabla)\,\mathbf{v} + \nabla q$, and the term $P_{\boldsymbol{\Delta}}(U,V)$ is the grad-div stabilization term defined as

$$P_{\boldsymbol{\Delta}}(U,V) = \sum_{K\in\tau_{\boldsymbol{\Delta}}} \tau_K(\nabla\cdot\mathbf{u}, \nabla\cdot\mathbf{v})_K, \tag{11}$$

where $U = (\mathbf{u},p)$, $V = (\mathbf{v},q)$, $U^* = (\mathbf{u}^*,p)$ and δ_K and τ_K are suitably chosen parameters, cf. [4].

4. Numerical results

The presented method was applied to several practical problems and the numerical results were validated. Here, the numerical results for the coupled system (2) and (5) is presented for the case of flexibly supported airfoil NACA 0012. The solution was performed for far field velocity $U_\infty = 5\,\mathrm{m\,s^{-1}}$ and modified parameter values were taken from [1]. The critical velocity determined by NASTRAN computations was 30.4 m/s, which corresponds to the results computed by the presented method. The airfoil response can be seen in Figure 2 and 3.

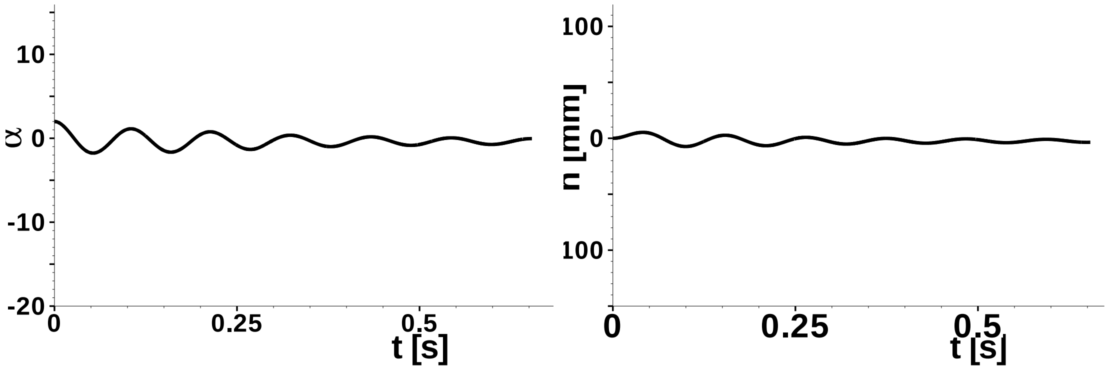


Fig. 2: *Aerodynamical forces acting on airfoil NACA 0012 for far field velocity* $U_\infty = 29\,m\,s^{-1}$ *causes damped vibrations.*

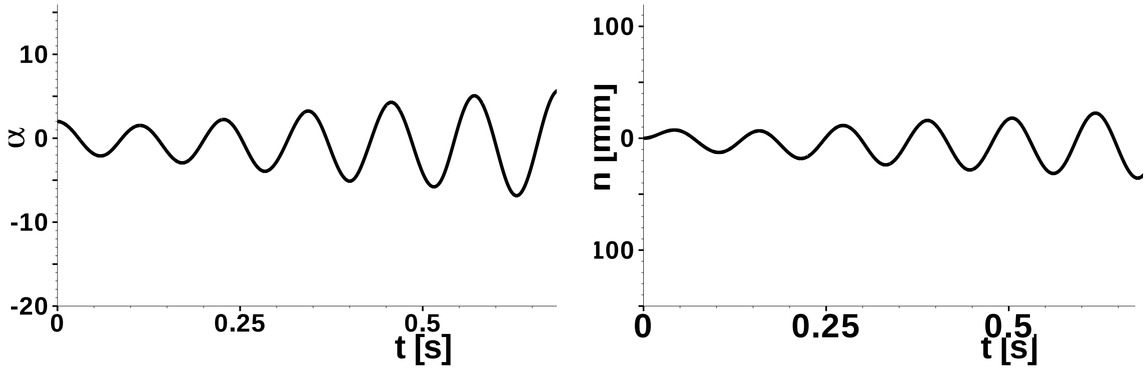


Fig. 3: *The airfoil response of the aerodynamical forces applied on the airfoil NACA 0012 for far field velocity* $U_\infty = 32\,m\,s^{-1}$.

References

[1] J. Čečrdle and J. Maleček: *Verification FEM model of an aircraft construction with two and three degrees of freedom.* Technical Report R-3418/02, Aeronautical Research and Test Institute, Prague, Letňany (in Czech), 2002.

[2] C. Farhat, M. Lesoinne, and N. Maman: *Mixed explicit/implicit time integration of coupled aeroelastic problems: three field formulation, geometric conservation and distributed solution.* International Journal for Numerical Methods in Fluids **21**, 1995, 807–835.

[3] M. Feistauer: *Mathematical methods in fluid dynamics.* Longman Scientific & Technical, Harlow, 1993.

[4] T. Gelhard, G. Lube, M. A. Olshanskii, and J.-H. Starcke: *Stabilized finite element schemes with LBB-stable elements for incompressible flows.* Journal of Computational and Applied Mathematics **177**, 2005, 243–267.

[5] R. J. LeVeque: *Numerical methods for conservation laws: lectures in mathematics.* Birkauser Verlag, 1990.

[6] T. Nomura and T. J. R. Hughes: *An arbitrary Lagrangian-Eulerian finite element method for interaction of fluid and a rigid body.* Computer Methods in Applied Mechanics and Engineering **95**, 1992, 115–138.

[7] P. Sváček, M. Feistauer, and J. Horáček: *Numerical simulation of flow induced airfoil vibrations with large amplitudes.* Journal of Fluids and Structures, 2004, (accepted).

UNCERTAINTIES IN MEASUREMENT OF THERMAL TECHNICAL CHARACTERISTICS OF BUILDING INSULATIONS*

J. Vala, S. Šťastník, H. Kmínová

1. Thermal technical characteristics of building insulations

Most thermal insulation materials used in civil engineering, namely those applied as parts of layered constructions in buildings, have a complicated porous micro- and macrostructure; thus the reliable prediction of their thermal insulation and accumulation properties is rather difficult. Technical standards require the so-called thermal stability, which means in practice i) the preservation of nearly constant temperature $T(t)$ in time $t \geq 0$ in the interior of the building, independent of quasi-periodic (day and year) climatic cycles, and ii) very slow changes of time derivatives of $T(t)$; moreover iii) the minimization of energy consumption by heating (in winter) and air conditioning (in summer) should be guaranteed. The general description of physical processes in porous materials comes from the classical conservation laws for mass, momentum, and energy and contains: i) the heat conduction, convection, and radiation (Fourier equation); ii) the partially irreversible propagation of moisture in various phases (as air humidity, liquid water and ice) and, possibly, of other contaminants; and iii) the compressible viscous air flow in rooms and through walls, roofs, etc. (Navier-Stokes equations). In the corresponding initial and boundary problems for systems of partial differential equations of evolution we need to know a lot of thermal technical characteristics, especially of a) the heat conduction; b) the heat convection; c) the heat radiation; d) the pore space and its availability for air, moisture, and contaminants; and e) the air flow in rooms, walls, roofs, etc. Typically such characteristics depend on T and other quantities, e. g. on the moisture content (and its phase) in applied materials.

One of the research directions at the Faculty of Civil Engineering of the Brno University of Technology is the development of ecological insulation materials, based on the wood waste. The crucial step of such experimental research is some reliable estimate of the basic thermal technical characteristics, at least those denoted as a). Under the assumption that the material is homogeneous and isotropic (due to the technology of its composition) the heat conduction can be described using three constants only: i) the material density ρ, ii) the heat conduction factor λ, iii) the thermal capacity (specific heat) c. Consequently the thermal insulation ability is determined by λ, and the thermal accumulation ability is determined by c, see [6, pp. 52, 57]. Frequently we shall apply the notation $\zeta = c\rho$.

*This work was partially supported by the Ministry of Education, Youth and Sports of the Czech Republic, research & development project No. 0021630511.

2. Laboratory measurements

The classical approach to the simulation of heat transfer is to solve the classical differential equation (see [2, p. 263])

$$\zeta\,\partial T/\partial t - \lambda\,\partial^2 T/\partial x^2 = r \tag{1}$$

(with some "effective characteristics" λ, ζ) in two variables: in t and in one space coordinate x; r represents (in general, as a function of t) the generated heat (per length). The setting of ρ is easy: it can be obtained as the ratio mass / volume. The identification of λ comes usually from standard experiments with the stationary heat transfer; in this case the first additive (time-dependent) term in (1) is missing. However, c must be obtained in another way, using various calorimeters where the contact with water is needed; this brings a danger of mismatched results caused by humidity in pores.

In [5], a new approach to measurements has been suggested: both λ and c (or ζ because ρ in known) can be obtained from an experiment that, unlike the standard experiment for determining λ, takes a non-stationary heat transfer into consideration. A more advanced numerical analysis is then needed for the simultaneous identification of both characteristics.

Fig. 1 presents the scheme of the original measurement device; we use the following notation: 1 – the thermal insulation (foam polystyrene blocks), 2 – two aluminium plates, just the lower one is heated, 3 – the tested sample (with unknown λ and ζ), 4 – the highlighting of the direction of thermal flow, 5 – two temperature sensors, 6 – the data recorder. Fig. 2 shows such a device in practice.

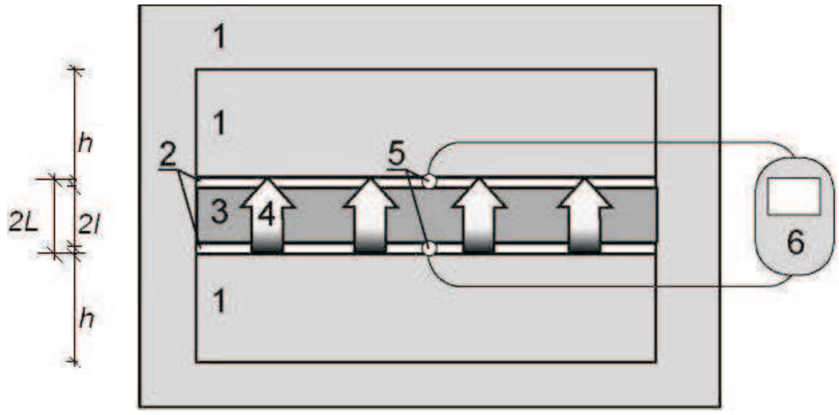


Fig. 1: *Principle of simultaneous measurement of λ and c.*

For the reliable identification of λ and c the technical standards require the analysis of uncertainties of measurements, see [3]. We shall see that in our approach such analysis will be available and using the same numerical technique as is used for the identification of the deterministic values of λ and c.

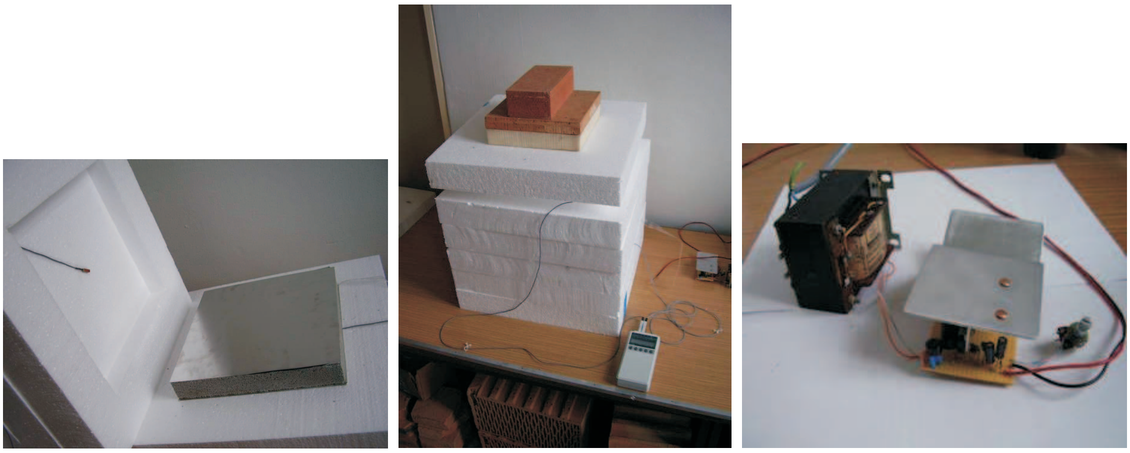


Fig. 2: *Measurement of basic thermal technical characteristics in the laboratory.*

3. An inverse problem in heat conduction

Let l, L and H be three positive numbers, $0 < l < L \ll H$, $L - l \ll l$. The technical interpretation of l, L and H is evident from the simplified scheme at Fig. 1: $2l$ is the thickness of the measured material sample, $L - l$ is the thickness of each of two aluminium plates, $2H = 2(L + h)$ denotes the total size of the whole insulated system where h is the thickness of each of two massive insulation blocks. Let us introduce the set $A = \{(-H, -L), (-L, -l), (-l, l), (l, L), (L, H)\}$ of couples of x-coordinates, too.

The above sketched classical formulation of the heat equation (1) can be (especially for layered structures) converted into the weak formulation

$$\zeta(\varphi, \dot{T}) + \lambda(\varphi', T') = r(\varphi, 1) + \varphi(a_+)q_+ - \varphi(a_-)q_- \,, \qquad (2)$$

satisfied for every test function φ from the Sobolev space $W^{1,2}(a_-, a_+)$ where (a_-, a_+) are elements of A, with (a priori unknown) interface heat fluxes $q_-(t)$ and $q_+(t)$; q_- and q_+ here is the brief notation for heat fluxes at the interfaces $x = a_-$ and $x = a_+$ The dot symbol in (2) is reserved for partial derivatives by t, the prime symbol for partial derivatives by x, $(\psi, \widetilde{\psi})$ are scalar products in the Lebesgue space $L^2(a_-, a_+)$ for any ψ and $\widetilde{\psi}$ from this space, i. e.

$$(\psi, \widetilde{\psi}) = \int_{a_-}^{a_+} \psi(x)\, \widetilde{\psi}(x)\, \mathrm{d}x \,.$$

We suppose that for $t = 0$ the temperature $T_0(x)$ is known everywhere (for $-H \leq x \leq H$), thus the initial condition $T(x, 0) = T_0(x)$ can be prescribed. We also assume that the whole system is perfectly insulated from the external environment (the experiment cannot be too long in practice), thus for $x = -H$ and $x = H$ no heat fluxes q_- or q_+ are considered. The unknowns are $T(x, t)$ everywhere for any positive time t, and (time-independent) λ and ζ only for $-l \leq x \leq l$ (inside the tested sample).

Alternatively the weak formulation (2) could be rewritten for $(a_-, a_+) = (-H, H)$ with test functions $\varphi \in W^{1,2}(-H, H)$; then all q_- and q_+ seemingly vanish. However, in such notation all scalar products $\lambda(.\,,.)$ and $\zeta(.\,,.)$ would obtain more complicated forms $(.\,,\lambda\,.)$ and $(.\,,\zeta\,.)$ with piecewise continuous functions λ and ζ, whose values are not known a priori everywhere. Consequently it is not possible to avoid all explicit calculations of q_- and q_+, at least those corresponding to $x = -l$ and $x = l$.

The identification of λ and ζ is based on the comparison of the temperatures $T(-l, t_s)$ and $T(l, t_s)$ from numerical simulation with the temperatures

$$T_{s-} \approx T(-l, t_s)\,, \qquad T_{s+} \approx T(l, t_s)\,, \tag{3}$$

obtained from sensors in a finite integer number S of times t_s, $s \in \{1, \dots, S\}$; in such discrete time steps the heat generator is able to guarantee the constant values $r_s = r(t_s)$ of r from the right-hand side of (2) inside the whole heated plate (where $l < x < L$). Unlike the much more general approach of [4], thanks to the very simple arrangement of the experiment, we are able to apply the semi-analytical Fourier method here. Following [1, pp. 229, 256], instead of $T(x, t)$ in (2) we can consider $T_N(x, t)$ with a large integer N (theoretically $N \to \infty$) in the form

$$T_N(x, t) = T_N(x, t_*) + \sum_{n=0}^{N} \varphi_n(\tilde{x})\, \alpha_n(t - t_*) \tag{4}$$

with $\tilde{x} = (x - a_-)/(a_+ - a_-)$ and $0 < t_* < t$, and attempt (once the system $\varphi_n(\tilde{x})$, $0 \le \tilde{x} \le 1$, $n \in \{1, \dots, N\}$, is available) to find the approximate solution $T_N(x, t)$ of (2); in practice we are allowed to set (step by step) $t_* = t_{s-1}$ and $t = t_s$.

The one-dimensional discretization in the variable x enables us to apply the method of lines: inserting $T_N(x, t)$ from (4) into (2), we obtain the system of $N + 1$ ordinary differential equations, whose general form is

$$a\zeta M\dot{\alpha} + a^{-1}\lambda K\alpha = \beta_+ q_+ - \beta_- q_- + arg\,, \tag{5}$$

where (for simplicity) $a = a_+ - a_-$ and

$$\alpha(\tau) = [\alpha_0(\tau), \dots, \alpha_N(\tau)]^{\mathrm{T}}, \quad \beta_- = [\varphi_0(0), \dots, \varphi_n(0)]^{\mathrm{T}}, \quad \beta_+ = [\varphi_0(1), \dots, \varphi_n(1)]^{\mathrm{T}}.$$

Let us remind that (5) must be formulated for each couple $(a_-, a_+) \in A$ separately and that $r \neq 0$ for $(a_-, a_+) = (l, L)$ only. The concrete form of the square “mass” and “stiffness” matrices M, K, generated by (φ_m, φ_n) and (φ'_m, φ'_n) with $m, n \in \{0, \dots, N\}$, and of the “load” vector g, generated by $(\varphi_m, 1)$ with $m \in \{0, \dots, N\}$, depends on the practical choice of $\varphi_1, \dots, \varphi_N$. The application of the classical Fourier basis (like [1, p. 139]) brings complications with averaged boundary values; thus the standard finite element technique, the wavelet analysis or other meshless approaches seem to be more efficient.

The solution $\alpha(\tau)$ of the system (5) can be analyzed with the help of real eigenvalues ω_n, $n \in \{0, \dots N\}$, obtained from the characteristic equation

$$\det(\lambda K - a^2\omega_n\zeta M) = 0\,,$$

and of the corresponding real eigenvectors; alternatively this can be rewritten as

$$\lambda K V = a^2\zeta M V \Omega\,,$$

where Ω is a diagonal square matrix of eigenvalues and V is a square matrix compound from column eigenvectors. All particular steps of this calculation can be found in [7]; the final result is

$$\alpha(\tau) = V \begin{bmatrix} \tau/(a\zeta) & & & \\ & \lambda/(a^2\zeta^2\omega_1)(1-\exp\left(-a^2\zeta\omega_1\tau/\lambda\right)) & & \\ & & \cdots & \\ & & & \lambda/(a^2\zeta^2\omega_N)(1-\exp\left(-a^2\zeta\omega_N\tau/\lambda\right)) \end{bmatrix} \qquad (6)$$

$$\times V^{\mathrm{T}} M^{-1}(\beta_+ q_+ - \beta_- q_- + arg)\,.$$

Nevertheless, q_- and q_+ at all material interfaces are still undetermined. No external fluxes are allowed, thus only four unknown values q_-, q_+ at such interfaces occur. At the same interfaces four continuity conditions for T are available, consequently all needed q_-, q_+ can be evaluated formally from the corresponding regular system of 4 linear algebraic equations with 4 variables. Then we have

$$T_N(-l, t_s) = T_N(-l, t_{s-1}) + G_{s-}(\lambda, \zeta)\,, \qquad T_N(l, t_s) = T_N(l, t_{s-1}) + G_{s+}(\lambda, \zeta)$$

with two complicated functions G_{s-}, G_{s+} of two variables λ, ζ, coming from the insertion of (6) into (4) for $x = -l$ and $x = l$; the software code for the evaluation of G_{s-}, G_{s+} makes use of MAPLE. Now we are ready to specify the vague relations (3): the minimum of a function

$$\Phi(\lambda, \zeta) = \frac{1}{2} \sum_{\sigma \in \{-,+\}} \sum_{s=1}^{S} \left(T_N(\sigma l, t_s) - T_{s\sigma}\right)^2$$

can be found with the help of the least squares method and (for a sufficiently good estimate of λ, ζ) of the Newton iterations, completed by an effective algorithm for the evaluation of the first and second partial derivatives of Φ needed in such iterations.

Material engineers in similar situations commonly use an “ad hoc” algorithm: i) set some rough estimate of λ and ζ; ii) by using some “black box” software like ANSYS, calculate the distribution of T in time, including that at the measured points; iii) if the differences between the measured and calculated values of T are large (which is decided from experience), choose another couple (λ, ζ) by using some heuristic technique (bi-sectioning, for example), and return to step ii), otherwise finish. The convergence of such approach is slow and doubtful; our semi-analytic

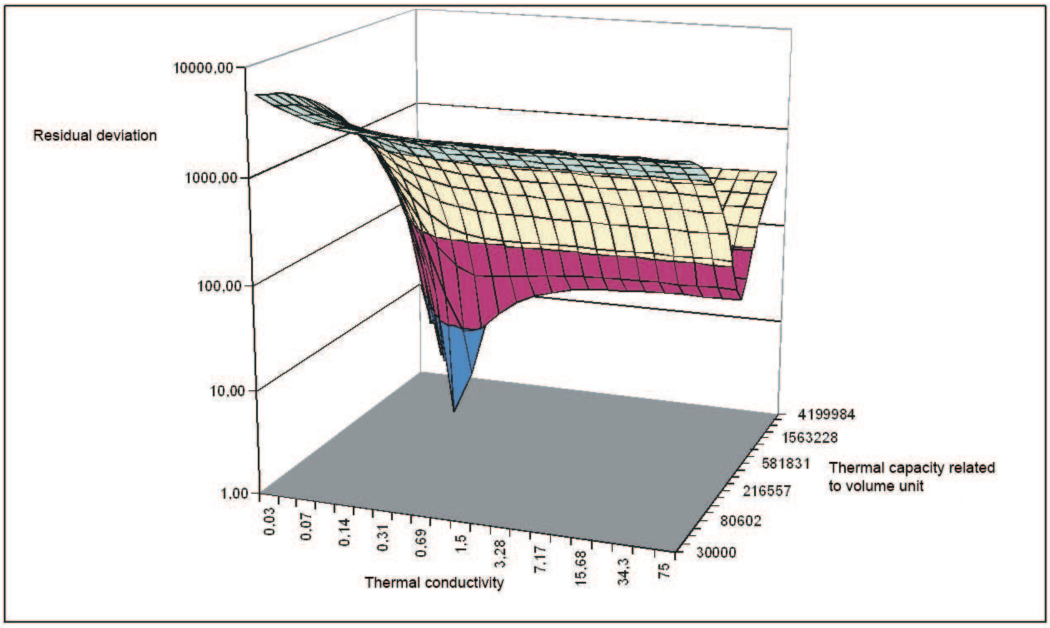


Fig. 3: *Characteristics λ, ζ, obtained by the minimization of $\Phi(\lambda, \zeta)$.*

method seems to be much more efficient. Nevertheless, the new revisions of commercial software packages such as ANSYS involve also certain support of the analysis of inverse problems, more advanced than the above criticized one. A typical distribution of $\Phi(\lambda, \zeta)$ for the (nearly) homogenized insulation layer, making use of the wood waste, is shown in Fig. 3.

4. Stochastic analysis

The analysis of uncertainties in measurements is an important part of the accreditation process of each technical laboratory. Most definitions of uncertainty in technical standards are rather vague, as "uncertainty is a parameter associated with the result of a measurement, that characterises the dispersion of the values that could reasonably be attributed to the measurand" in [3, p. 9]. However, to respect such requirements, the identification of thermal technical characteristics λ and ζ for new insulation materials should contain a deep analysis of uncertainty sources and components and their relation to random and systematic errors in measurements, concerning: i) the size of the material sample and the smoothness of its surface, ii) the correctness of setting of generated heat, iii) the correctness of temperature measurements from both sensors, iii) the preservation of assumed zero boundary fluxes, iv) the validity of homogenized isotropic constant values of the chararacteristics, v) the acceptability of physical, mathematical, and computational simplifications, vi) the numerical error analysis, etc.

For illustration, following [3, pp. 11, 24], we can reduce the analysis of uncertainty components to the analysis of standard deviations, assuming: i) the uncorrelated quantities r_s (adjusted values) and $T_{\sigma s}$ (measured values) with $s \in \{1, \dots, S\}$,

$\sigma \in \{-,+\}$, ii) the normal (Gaussian) probability distribution (justified by the central limit theorem), and iii) the uncertainty w_r of all variables r_s and the uncertainty w_T of all variables T_s. Then the uncertainties w_λ and w_ζ of both material characteristics λ, ζ can be calculated as

$$w_\lambda = \sqrt{w_r^2 \sum_{s=1}^{S} (\partial\lambda/\partial r_s)^2 + w_T^2 \sum_{\sigma\in\{-,+\}} \sum_{s=1}^{S} (\partial\lambda/\partial T_{s\sigma})^2}\,,$$

$$w_\zeta = \sqrt{w_r^2 \sum_{s=1}^{S} (\partial\zeta/\partial r_s)^2 + w_T^2 \sum_{\sigma\in\{-,+\}} \sum_{s=1}^{S} (\partial\zeta/\partial T_{s\sigma})^2}\,.$$

A detailed study shows that evaluation of the above presented uncertainties can use the same algorithms as those in the Newton iteration process; this makes all computations relatively simple and inexpensive. More complicated formulae are needed in some other cases, e. g. in case of the uncertain thickness l.

Recently in [5] the approach presented in this paper has been applied to a room microclimate oriented study of the thermal behaviour of many new experimental materials for insulation layers in buildings. Unfortunately, especially the values of ζ (much more than those of λ) obtained both from the literature and from other experiments under similar conditions have a very large dispersion; thus (although the existence of solutions can be verified formally and the implemented software returns rather low values of $\Phi(\lambda,\zeta)$ – for illustration see Fig. 3 again) the validity of results, taking into account all potential sources of errors and inaccuracies, should be examined properly in the near future.

References

[1] J. Barták, L. Herrmann, V. Lovicar, O. Vejvoda: *Partial differential equations of evolution.* Ellis Horwood, 1991.

[2] M.G. Davies: *Building heat transfer.* John Wiley & Sons, 2004.

[3] S.R.L. Ellison, M. Rosslein, A. Williams (eds.) and al.: *Quantifying uncertainty in analytical measurements.* EURACHEM/CITAC Guide CG4, 2000, available at `http://www.measurementuncertainty.org/mu/guide`.

[4] I. Hlaváček, J. Chleboun, I. Babuška: *Uncertain input data problems and the worst scenario method.* Elsevier Science & Technology, 2004.

[5] H. Kmínová: *Analysis of thermal technical properties of building materials with respect to their influence to the internal microclimate in buildings.* Ph. D. thesis, Faculty of Civil Engineering, Brno University of Technology, 2006, (in Czech).

[6] J. Kuneš: *Modelling of thermal processes.* SNTL Prague, 1989, (in Czech).

[7] S. Šťastník, J. Vala, H. Kmínová: *Identification of basic thermal technical characteristics of building materials.* Kybernetika (Acad. Sci. Czech Rep.), to appear.

DISCRETE GREEN'S FUNCTION AND MAXIMUM PRINCIPLES*

Tomáš Vejchodský, Pavel Šolín

Abstract

In this paper the discrete Green's function (DGF) is introduced and its fundamental properties are proven. Further it is indicated how to use these results to prove the discrete maximum principle for 1D Poisson equation discretized by the *hp*-FEM with pure Dirichlet or with mixed Dirichlet-Neumann boundary conditions and with piecewise constant coefficient.

1. Introduction

The topic of discrete maximum principles (DMP) is already studied for several decades [1]. The problematics of DMP can be simplified to the question under what conditions a numerical method produces nonnegative solution in situations when the exact solution is known to be nonnegative. Numerical methods that satisfies DMP are useful and desirable for problems where naturally nonnegative quantities like temperature, concentration, or density are computed.

Results for the finite element methods (FEM) and for various problems are well known, see e.g. [2, 4, 5, 6] and references therein. These works, however, deal with piecewise linear approximations only. The results about higher order approximations are much scarce, see [3, 11] and recent works of the authors [10, 7, 8, 9]. The reason is that the condition for a piecewise linear function to be nonnegative is trivial but suitable condition for piecewise polynomial function is very difficult to obtain.

In this point of view the discrete Green's function turned out to be a very useful tool for investigation of DMP for higher order finite element methods.

2. Model problem

Although the theory is applicable for very general class of problems, we restrict ourselves for the clarity of explanation to relatively simple linear elliptic problem. The model problem is formulated in the classical way as follows

$$\begin{aligned} -\operatorname{div}(\mathcal{A}\nabla u) + cu &= f \quad \text{in } \Omega \\ u &= 0 \quad \text{on } \Gamma_{\mathrm{D}} \\ \alpha u + (\mathcal{A}\nabla u)\cdot\nu &= g \quad \text{on } \Gamma_{\mathrm{N}}. \end{aligned} \tag{1}$$

*The first author has been supported by grant No. 201/04/P021 of the Grant Agency of the Czech Republic and by the Institutional Research Plan No. AV0Z10190503 of the Academy of Sciences of the Czech Republic. The second author has been supported in part by the U.S. Department of Defense under Grant No. 05PR07548-00, by the NSF Grant No. DMS-0532645, and by Grant Agency of the Czech Republic project No. 102-05-0629. This support is gratefully acknowledged.

Here Ω is a domain with Lipschitz continues boundary in $\mathbb{R}^d$. The boundary $\partial\Omega$ is split into two disjoint parts Γ_{D} and Γ_{N}. The matrix $\mathcal{A} = \mathcal{A}(x) \in \mathbb{R}^{d\times d}$ is uniformly positive definite and the coefficients $c = c(x)$ and $\alpha = \alpha(x)$ are nonnegative. The unit outward normal to $\partial\Omega$ is denoted by ν.

To give rigorous meaning to the model problem, we introduce the concept of weak solution. For that reason we define the space

$$V = \{u \in H^1(\Omega) : u = 0 \text{ on } \Gamma_{\mathrm{D}}\},$$

where the values on $\partial\Omega$ are understood in the sense of traces. The weak solution $u \in V$ of (1) is defined by identity

$$a(u, v) = F(v) \quad \forall v \in V. \tag{2}$$

The bilinear form $a : V \times V \mapsto \mathbb{R}$ and the linear functional $F : V \mapsto \mathbb{R}$ are given by

$$\begin{aligned} a(u,v) &= \int_\Omega (\mathcal{A}\nabla u)\cdot\nabla v \,\mathrm{d}x + \int_\Omega cuv \,\mathrm{d}x + \int_{\Gamma_{\mathrm{N}}} \alpha uv \,\mathrm{d}s, \\ F(v) &= \int_\Omega fv \,\mathrm{d}x + \int_{\Gamma_{\mathrm{N}}} gv \,\mathrm{d}s. \end{aligned}$$

These integrals are well defined if $\mathcal{A} \in \left[L^\infty(\Omega)\right]^{d\times d}$, $c \in L^\infty(\Omega)$, $\alpha \in L^\infty(\Gamma_{\mathrm{N}})$, $f \in L^2(\Omega)$, and $g \in L^2(\Gamma_{\mathrm{N}})$. If $\operatorname{meas}\Gamma_{\mathrm{D}} \neq 0$ or $c \not\equiv 0$ or $\alpha \not\equiv 0$ then by Lax-Milgram lemma the weak solution exists and is unique.

Let us recall the standard definition of Green's function for problem (2). For almost every $y \in \overline{\Omega}$, the Green's function $G_y \in V$ is given as a unique solution to

$$a(w, G_y) = \delta_y(w) \quad \forall w \in V. \tag{3}$$

The symbol δ_y stands for the Dirac functional. This δ_y is well defined for all continous function w by $\delta_y(w) = w(y)$. This definition can be augmented for w from V by the Hahn-Banach theorem.

By (2) and (3) we infer the fundamental Kirchhoff-Helmholtz representation formula

$$u(y) = \delta_y(u) = a(u, G_y) = F(G_y).$$

Hence for our model problem

$$u(y) = \int_\Omega f(x)G_y(x)\,\mathrm{d}x + \int_{\Gamma_{\mathrm{N}}} g(s)G_y(s)\,\mathrm{d}s.$$

3. Discretization by *hp*-FEM

In the *hp* version of the finite element method (*hp*-FEM) we vary both the sizes h and polynomial degrees p of elements. To discretize our model problem (2) by the

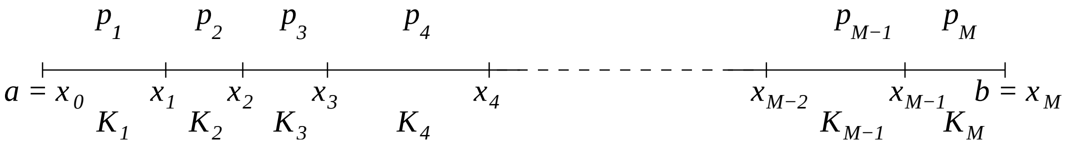


Fig. 1: *A 1D mesh $\mathcal{T}_{hp}$ with elements K_i of polynomial degrees p_i, $i = 1, 2, \dots, M$.*

hp-FEM we assume the domain Ω to be polytopic. We introduce simplicial partition $\mathcal{T}_{hp}$ of Ω into M elements and we endow each element $K_i \in \mathcal{T}_{hp}$, $i = 1, 2, \dots, M$, with an arbitrary polynomial degree $p_i \geq 1$. See Figure 1 for a 1D illustration.

The hp-FEM mesh $\mathcal{T}_{hp}$ defines the finite element space

$$V_{hp} = \{v_{hp} \in V : v_{hp}|_{K_i} \in P^{p_i}(K_i) \text{ for all } K_i \in \mathcal{T}_{hp}\},$$

where $P^{p_i}(K_i)$ stands for the space of polynomials on K_i of degree at most p_i. The hp-FEM solution $u_{hp} \in V_{hp}$ is then defined by identity

$$a(u_{hp}, v_{hp}) = F(v_{hp}) \quad \forall v_{hp} \in V_{hp}. \tag{4}$$

4. Discrete Green's function and its properties

The discrete Green's function (DGF) is defined in analogy with the continuous case, cf. (3). For all $y \in \overline{\Omega}$, define the discrete Green's function $G_{hp,y} \in V_{hp}$ by

$$a(w_{hp}, G_{hp,y}) = \delta_y(w_{hp}) \quad \forall w_{hp} \in V_{hp}. \tag{5}$$

It is convenient to put $G_{hp}(x, y) = G_{hp,y}(x)$. The combination of (4) and (5) gives again the representation formula

$$u_{hp}(y) = \delta_y(u_{hp}) = a(u_{hp}, G_{hp,y}) = F(G_{hp,y}).$$

For our model problem this becomes

$$u_{hp}(y) = \int_\Omega f(x) G_{hp}(x, y)\,\mathrm{d}x + \int_{\Gamma_\mathrm{N}} g(s) G_{hp}(s, y)\,\mathrm{d}s. \tag{6}$$

In contrast to the continuous case the DGF can be easily expressed through the inverse stiffness matrix, cf. [2].

Lemma 4.1. *Let $\{\varphi_1, \varphi_2, \dots, \varphi_N\}$ be a basis in V_{hp}. If $A \in \mathbb{R}^{N\times N}$ be a matrix with entries $A_{ij} = a(\varphi_j, \varphi_i)$, $1 \leq i, j \leq N$, then*

$$G_{hp}(x, y) = \sum_{j=1}^{N} \sum_{k=1}^{N} A^{-1}_{jk} \varphi_k(x) \varphi_j(y), \tag{7}$$

where A^{-1}_{jk} are entries of A^{-1}, i.e., $\sum_{j=1}^{N} A_{ij} A^{-1}_{jk} = \delta_{ik}$ (Kronecker symbol).

Proof. The proof follows from (5) and can be found in [8]. □

The following two corollaries follow directly from Lemma 4.1.

Corollary 4.1. *If $a(\cdot,\cdot)$ is symmetric then $G_{hp}(x,y) = G_{hp}(y,x)$.*

Corollary 4.2. *Let $\{l_1, l_2, \ldots, l_N\}$ be a basis of V_{hp} such that $a(l_i, l_j) = \delta_{ij}$. Then*

$$G_{hp}(x,y) = \sum_{i=1}^{N} l_i(x) l_i(y).$$

Since the nonnegativity of DGF is fundamental for discrete maximum principles, see Theorem 5.1 below, the following lemma is of particular interest.

Lemma 4.2. *If the bilinear form $a(\cdot,\cdot)$ is symmetric and if $a(v_{hp}, v_{hp}) > 0$ for all $0 \neq v_{hp} \in V_{hp}$ then $G_{hp}(x,x) > 0$ for all $x \in \Omega$.*

Proof. Let $\{\varphi_1, \varphi_2, \ldots, \varphi_N\}$ be a basis in V_{hp}. By the assumptions the stiffness matrix $A_{ij} = a(\varphi_j, \varphi_i)$, $1 \leq i,j \leq N$, is symmetric and positive definite as well as its invers matrix. Thus, by Lemma 4.1, $G_{hp}(x,x) = \boldsymbol{\varphi}(x)^T A^{-1} \boldsymbol{\varphi}(x) > 0$, where $\boldsymbol{\varphi}(x) = (\varphi_1(x), \varphi_2(x), \ldots, \varphi_N(x))^T$. Notice that $\boldsymbol{\varphi}(x) \neq \mathbf{0}$ for all $x \in \Omega$ since $\{\varphi_i(x)\}$ is a basis in V_{hp}. □

5. Application to the discrete maximum principles

These results about DGF can be used to proof certain qualitative properties of the discrete solution. Let us start with the comparison principle for our model problem.

Definition 5.1. *The problem* (4) *satisfies the discrete comparison principle if*

$$f \geq 0 \text{ and } g \geq 0 \quad \Rightarrow \quad u_{hp} \geq 0.$$

The following theorem is crucial for the analysis of discrete comparison principle via DGF.

Theorem 5.1. *Problem* (4) *satisfies the discrete comparison principle if and only if the corresponding discrete Green's function $G_{hp}(x,y)$ defined by* (5) *is nonnegative in Ω^2.*

Proof. By (7), the discrete Green's function $G_{hp}(x,z)$ is continuous up to the boundary of Ω^2. The rest follows immediately from representation formula (6). □

For certain problems the DGF can be explicitly expressed and its nonnegativity can be analyzed. We mention two of our results about discrete maximum principle. Both are based on Theorem 5.1. A crucial role in these results plays quantity

$$H^*_{\mathrm{rel}}(p) = 1, \quad \text{for } p = 1,$$

$$H^*_{\mathrm{rel}}(p) = 1 + \frac{1}{2} \min_{(\xi,\eta)\in[-1,1]^2} l_0(\xi) l_0(\eta) \sum_{k=2}^{p} \kappa_k(\xi)\kappa_k(\eta), \quad \text{for } p \geq 2.$$

Here, $l_0(\xi) = (1-\xi)/2$ and $\kappa_k(\xi) = \sqrt{\frac{2k-1}{2}}\frac{4}{k(1-k)}P'_{k-1}(\xi)$, where $P_k(\xi)$ stand for the Legendre polynomials of degree k and prime denotes the derivative.

Theorem 5.2. *Let us consider simplified problem* (4) *in 1D setting with homogeneous Dirichlet boundary conditions, i.e.,* $\Omega = (\bar{a}, \bar{b})$, $\mathcal{A} = 1$, $c = 0$, $\alpha = 0$, $\Gamma_{\mathrm{D}} = \{\bar{a}, \bar{b}\}$, *and* $\Gamma_{\mathrm{N}} = \emptyset$*. Let* $\bar{a} = x_0 < x_1 < \ldots < x_M = \bar{b}$ *be a partition of the domain and let* $p_i \geq 1$ *be polynomial degrees assigned to elements* $K_i = [x_{i-1}, x_i]$, $i = 1, 2, \ldots, M$*. If*

$$\frac{x_i - x_{i-1}}{\bar{b} - \bar{a}} \leq H^*_{\mathrm{rel}}(p_i) \quad \textit{for all } i = 1, 2, \ldots, M, \tag{8}$$

then this problem satisfies the discrete comparison principle.

Theorem 5.3. *Let us consider simplified problem* (4) *in 1D setting with mixed boundary conditions, i.e.,* $\Omega = (\bar{a}, \bar{b})$, $\mathcal{A} = 1$, $c = 0$, $\alpha = 0$, $\Gamma_{\mathrm{D}} = \{\bar{a}\}$ *and* $\Gamma_{\mathrm{N}} = \{\bar{b}\}$*. Let* $\bar{a} = x_0 < x_1 < \ldots < x_M = \bar{b}$ *be a partition of the domain and let* $p_i \geq 1$ *be polynomial degrees assigned to elements* $K_i = [x_{i-1}, x_i]$, $i = 1, 2, \ldots, M$*. If*

$$H^*_{\mathrm{rel}}(p_i) \geq 0 \quad \textit{for all } i = 1, 2, \ldots, M, \tag{9}$$

then this problem satisfies the discrete comparison principle.

Proofs of Theorems 5.2 and 5.3 are given in [8] and [9], respectively. In the same papers we verified that $H_{\mathrm{rel}}(p) \geq 9/10$ for $1 \leq p \leq 100$. Thus, condition (9) is satisfied for these values of p and the condition (8) can be strengthened to $(x_i - x_{i-1})/(\bar{b} - \bar{a}) \leq 9/10$ which means that the discrete comparison principle is valid if all elements are shorter then 90 % of the length of the domain Ω.

Both the results from Theorems 5.2 and 5.3 can be generalized to the case of piecewise constant coefficient $\mathcal{A}$. The case of mixed boundary conditions (Theorem 5.3) remains valid even for piecewise constant $\mathcal{A}$, i.e., the comparison principle is guaranteed for all meshes with polynomial degrees not exceeding 100. The case of pure Dirichlet boundary conditions (Theorem 5.2) needs reformulation of condition (8) in the following way

$$\frac{\tilde{h}_i}{\sum\limits_{k=1}^{M} \tilde{h}_k} \leq H^*_{\mathrm{rel}}(p_i) \quad \text{for all } i = 1, 2, \ldots, M. \tag{10}$$

Here $\tilde{h}_i = (x_i - x_{i-1})/\mathcal{A}_i$, $i = 1, 2, \ldots, M$, mean modified element lengths and $\mathcal{A}_i$ is the constant value of $\mathcal{A}(x)$ on the element K_i. Notice that the sum in the denominator in (10) can be interpretted as a length of a modified domain. More details about the case with piecewise constant coefficient can be found in [10].

Finally, let us recall that for problems treated in Theorems 5.2 and 5.3 the discrete comparison principle implies the discrete maximum principle. The discrete maximum principle states that in the case of nonpositive f and nonpositive g the maximum of u_{hp} is attained in the interior of Ω.

References

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

[2] A. Drăgănescu, T.F. Dupont, L.R. Scott: *Failure of the discrete maximum principle for an elliptic finite element problem.* Math. Comp. **74**, 2005, 1–23 (electronic).

[3] W. Höhn, H.D. Mittelmann: *Some remarks on the discrete maximum principle for finite elements of higher-order.* Computing **27**, 1981, 145–154.

[4] A. Jüngel, A. Unterreiter: *Discrete minimum and maximum principles for finite element approximations of non-monotone elliptic equations.* Numer. Math. **99**, 2005, 485–508.

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

[6] S. Korotov, M. Křížek, P. Neittaanmäki: *Weakened acute type condition for tetrahedral triangulations and the discrete maximum principle.* Math. Comp. **70**, 2000, 107–119.

[7] P. Šolín, T. Vejchodský: *On a weak discrete maximum principle for hp-FEM.* JCAM, 2006 (accepted).

[8] T. Vejchodský, P. Šolín: *Discrete maximum principle for higher-order finite elements in 1D.* Math. Comp., 2006 (accepted).

[9] T. Vejchodský, P. Šolín: *Discrete maximum principle for mixed boundary conditions in 1D.* Research Report No. 2006-09, Department of Math. Sciences, University of Texas at El Paso, July 2006.

[10] T. Vejchodský, P. Šolín: *Discrete maximum principle for 1D problems with piecewise constant coefficient.* Research Report No. 2006-10, Department of Math. Sciences, University of Texas at El Paso, July 2006.

[11] E.G. Yanik: *Sufficient conditions for a discrete maximum principle for high-order collocation methods.* Comput. Math. Appl. **17**, 1989, 1431–1434.

Programs and Algorithms of Numerical Mathematics 13
in honor of Ivo Babuška's 80th birthday

List of Participants

Mark Ainsworth, Professor
Department of Mathematics, University of Strathclyde, Livingstone Tower,
26 Richmond St., Glasgow G1 1XH,UK
e-mail: M.Ainsworth@strath.ac.uk

Ivo Babuška, Professor
The University of Texas at Austin, Department of Aerospace Engineering and Engineering Mechanics,
1 University Station, 201 E. 21st Street, TX 78712 UT Austin, USA
e-mail: babuska@ices.utexas.edu

Luděk Beneš, Ing., Ph.D.
Ústav technické mat., Fak. strojní ČVUT
Karlovo nám. 13, 121 35 Praha 2
e-mail: benes@marian.fsik.cvut.cz

Michal Beneš, Ing., Dr.
Katedra matem., Fak. stavební ČVUT
Thákurova 7, 166 29 Praha 6
e-mail: benes@mat.fsv.cvut.cz

Radim Blaheta, prof. RNDr., CSc.
Ústav geoniky AV ČR
Studentská 1768, 708 00 Ostrava-Poruba
e-mail: blaheta@ugn.cas.cz

Jan Brandts, Dr.
Korteweg-De Vries Institute for Mathematics, Plantage Muidergracht 24,
1018 TV Amsterdam, Netherlands
e-mail: brandts@science.uva.nl

Pavel Burda, prof. RNDr., CSc.
Ústav technické mat., Fak. strojní ČVUT
Karlovo nám. 13, 121 35 Praha 2
e-mail: burda@fsik.cvut.cz

Wei Chen, Dr.
School of Economics, Shandong Univ.
27 Shanda Nanlu, 250 100 Jinan, China
e-mail: weichen@sdu.edu.cn

Jan Chleboun, RNDr., CSc.
Matematický ústav AV ČR
Žitná 26, 115 67 Praha 1; and
Ústav technické mat., Fak. strojní ČVUT
Karlovo nám. 13, 121 35 Praha 2
e-mail: chleb@math.cas.cz

Robert Cimrman, Ing., Ph.D.
Západočeská univerzita,
Nové technologie – výzkumné centrum
Univerzitní 8, 306 14 Plzeň
e-mail: cimrman3@ntc.zcu.cz

Jakub Červený,
Katedra matematiky, ZČU
Univerzitní 20, 306 14 Plzeň
e-mail: ccerv@seznam.cz

Josef Dalík, doc. RNDr., CSc.
Ústav matematiky a deskriptivní geom.,
Fakulta stavební VUT
Žižkova 17, 602 00 Brno
e-mail: dalik.j@fce.vutbr.cz

Jiří Dobeš, Ing.
Ústav technické mat., Fak. strojní ČVUT
Karlovo nám. 13, 121 35 Praha 2
e-mail: Jiri.Dobes@fs.cvut.cz

Jiří Dobiáš, Ing., Csc.
Ústav termomechaniky AV ČR
Dolejškova 5, 182 00 Praha 8
e-mail: jdobias@it.cas.cz

Vít Dolejší, doc. RNDr., Ph.D.
Katedra numerické mat., MFF UK
Sokolovská 83, 186 75 Praha 8
e-mail: dolejsi@karlin.mff.cuni.cz

Zdeněk Dostál, prof. RNDr., CSc.
Katedra aplikované mat., FEI VŠB-TU
tř. 17. listopadu 15, 708 33 Ostrava
e-mail: zdenek.dostal@vsb.cz

Lenka Dubcová, Mgr.
Katedra numerické mat., MFF UK
Sokolovská 83, 186 75 Praha 8
e-mail: dubcova@centrum.cz

Frank Duderstadt, Dr.
WIAS Berlin,
Mohrenstr. 39, 10117 Berlin, Germany
e-mail: dudersta@wias-berlin.de

Miloslav Feistauer,
prof. RNDr., DrSc., dr.h.c.
Katedra numerické mat., MFF UK
Sokolovská 83, 186 75 Praha 8
e-mail: feist@karlin.mff.cuni.cz

Jaroslav Fořt, prof. Ing., CSc.
Ústav technické mat., Fak. strojní ČVUT
Karlovo nám. 13, 121 35 Praha 2
e-mail: fort@marian.fsik.cvut.cz

Jiří Fürst, doc. Ing., Ph.D.
Ústav technické mat., Fak. strojní ČVUT
Karlovo nám. 13, 121 35 Praha 2
e-mail: Jiri.Furst@fs.cvut.cz

Benqi Guo, Professor
Department of Mathematics,
University of Manitoba
Winnipeg, MB R3T 2N2, Canada
e-mail: guo@cc.umanitoba.ca

Antti Hannukainen,
Institute of Mathematics,
Helsinki University of Technology
P.O. Box 1100, Helsinki
FIN-02015 TKK Finland
e-mail: antti.hannukainen@hut.fi

Ivan Hlaváček, Ing., DrSc.
Matematický ústav AV ČR
Žitná 25, 115 67 Praha 1
e-mail: hlavacek@math.cas.cz

Milan Hokr, Ing., Ph.D.
Katedra modelování procesů,
Fakulta mechatroniky TU
Hálkova 6, 461 17 Liberec
e-mail: milan.hokr@tul.cz

Drahoslava Janovská,
doc. RNDr., CSc.
Ústav matematiky VŠCHT,
Technická 5, 166 28 Praha 6
e-mail: Drahoslava.Janovska@vscht.cz

Vladimír Janovský,
doc. RNDr., DrSc.
Katedra numerické mat., MFF UK
Sokolovská 83, 186 75 Praha 8
e-mail: janovsky@karlin.mff.cuni.cz

Petr Knobloch, doc. Mgr., Dr.
Katedra numerické mat., MFF UK
Sokolovská 83, 186 75 Praha 8
e-mail: knobloch@karlin.mff.cuni.cz

Martin Kocurek, Mgr.
Katedra matem., Fak. stavební ČVUT
Thákurova 7, 166 00 Praha 6
e-mail: kocurek@mat.fsv.cvut.cz

Michal Kočvara, RNDr., DrSc.
Ústav teorie informace a automatizace AV ČR
Pod Vodárenskou věží 4, 182 08 Praha 8
e-mail: kocvara@utia.cas.cz

Roman Kohut, RNDr., CSc.
Ústav geoniky AV ČR
Studentská 1768, 708 00 Ostrava-Poruba
e-mail: kohut@ugn.cas.cz

Tomáš Kojecký, RNDr., CSc.
Katedra matem., Fak. stavební ČVUT
Thákurova 7, 166 29 Praha 6
e-mail: kojecky@mat.fsv.cvut.cz

Sergey Korotov, Dr.
Institute of Mathematics,
Helsinki University of Technology
P.O. Box 1100, Helsinki
FIN-02015 TKK Finland
e-mail: skorotov@cc.hut.fi

Karel Kozel, prof. RNDr., DrSc.
Ústav technické mat., Fak. strojní ČVUT
Karlovo nám. 13, 121 35 Praha 2
e-mail: kozelk@fsik.cvut.cz

Tomáš Kozubek, Ing., Ph.D.
Katedra aplikované matem., VŠB-TU
17. listopadu 15, 708 33 Ostrava-Poruba
e-mail: tomas.kozubek@vsb.cz

Michal Křížek, prof. RNDr., DrSc.
Matematický ústav AV ČR
Žitná 25, 115 67 Praha 1
e-mail: krizek@math.cas.cz

Radek Kučera, doc. RNDr., Ph.D.
Katedra matematiky a DG, VŠB-TU
17. listopadu 15, 708 33 Ostrava-Poruba
e-mail: radek.kucera@vsb.cz

Václav Kučera, Mgr.
Katedra numerické mat., MFF UK
Sokolovská 83, 186 75 Praha 8
e-mail: vaclav.kucera@email.cz

Pavel Kůs, Mgr.
Katedra numerické mat., MFF UK
Sokolovská 83, 186 75 Praha 8
e-mail: pavel.kus@gmail.com

Martin Kyncl, RNDr., CSc.
PSP Engineering a.s.,
divize kusových strojů, odd. výpočtů
Kojetínská 71, 750 53 Přerov
e-mail: kyncl@pspeng.cz

Dalibor Lukáš, Ing., Ph.D.
Katedra aplikované matem., VŠB-TU
17. listopadu 15, 708 33 Ostrava-Poruba
e-mail: dalibor.lukas@vsb.cz

Ladislav Lukšan, prof. Ing., DrSc.
Ústav informatiky AV ČR
Pod Vodárenskou věží 2, 182 07 Praha 8
e-mail: luksan@cs.cas.cz

Jiří Maděra, Ing., Ph.D.
Fakulta stavební, ČVUT
Thákurova 6, 166 29 Praha 6
e-mail: madera@fsv.cvut.cz

Ivo Marek, prof. RNDr., DrSc.
Katedra numerické mat., MFF UK
Sokolovská 83, 186 75 Praha 8
e-mail: marek@ms.mff.cuni.cz

Petr Mayer, RNDr., Dr.
Katedra numerické mat., MFF UK
Sokolovská 83, 186 75 Praha 8
e-mail: pmayer@ms.mff.cuni.cz

Jaroslav Mlýnek, RNDr., CSc.
Katedra matematiky a didaktiky mat.,
Fakulta pedagogická TUL
Hálkova 6, 461 17 Liberec
e-mail: jaroslav.mlynek@vslib.cz

Zuzana Morávková, Mgr., Ph.D.
Katedra matematiky a DG, VŠB-TU
17. listopadu 15 , 708 00 Ostrava-Poruba
e-mail: zuzana.moravkova@vsb.cz

Vratislava Mošová, RNDr., CSc.
Ústav exaktních věd,
Moravská vysoká škola Olomouc, o.p.s.
Jeremenkova 1142/42, 772 00 Olomouc
e-mail: vratislava.mosova@mvso.cz

Karel Najzar, doc. RNDr., CSc.
Katedra numerické mat., MFF UK
Sokolovská 83, 186 75 Praha 8
e-mail: knaj@karlin.mff.cuni.cz

Tomáš Neustupa, RNDr.
Ústav technické mat., Fak. strojní ČVUT
Karlovo nám. 13, 121 35 Praha 2
e-mail: tneu@centrum.cz

Miroslav Pospíšek, Ing., CSc.
ANECT, a.s.,
Vinohradská 112, 130 00 Praha 3
e-mail: mpospisek@anect.com

Jan Pospíšil, Ing.
Katedra matematiky, ZČU
Univerzitní 20, 306 14 Plzeň
e-mail: honik@kma.zcu.cz

Aleš Prachař, RNDr.
VZLÚ a.s.,
V Mezihoří 2a, 180 00 Praha 8
e-mail: prachar@vzlu.cz

Milan Práger, RNDr., CSc.
Matematický ústav AV ČR
Žitná 25, 115 67 Praha 1
e-mail: prager@math.cas.cz

Vladimír Prokop, Ing.
Ústav technické mat., Fak. strojní ČVUT
Karlovo nám. 13, 121 35 Praha 2
e-mail: prokop@marian.fsik.cvut.cz

Petr Přikryl, prof. RNDr., CSc.
Matematický ústav AV ČR
Žitná 25, 115 67 Praha 1
e-mail: prikryl@math.cas.cz

Petra Punčochářová, Ing.
Ústav technické mat., Fak. strojní ČVUT
Karlovo nám. 13, 121 35 Praha 2
e-mail: puncocha@marian.fsik.cvut.cz

Eduard Rohan, doc. Dr. Ing.
Katedra matematiky, ZČU
Univerzitní 22, 306 14 Plzeň
e-mail: rohan@kme.zcu.cz

Hans-Görg Roos, Professor
Institute of Numerical Mathematics,
Department of Mathematics,
Technical University of Dresden,
D - 01062 Dresden
e-mail: roos@math.tu-dresden.de

Tomáš Roubíček, doc. Ing., DrSc.
Matematický ústav, MFF UK
Sokolovská 83, 186 75 Praha 8
e-mail: roubicek@karlin.mff.cuni.cz

Miroslav Rozložník, Ing., Dr.
Ústav informatiky AV ČR
Pod Vodárenskou věží 2, 182 07 Praha 8
e-mail: miro@cs.cas.cz

Karel Segeth, prof. RNDr., CSc.
Matematický ústav AV ČR
Žitná 25, 115 67 Praha 8
e-mail: segeth@math.cas.cz

Jitka Segethová, RNDr., CSc.
Katedra numerické mat., MFF UK
Sokolovská 83, 186 75 Praha 8
e-mail: jseg@karlin.mff.cuni.cz

Veronika Sobotíková, RNDr., CSc.
Katedra matematiky, FEL ČVUT
Technická 2, 166 27 Praha 6
e-mail: veronika@math.feld.cvut.cz

Lawrence Somer, Professor
Matematický ústav AV ČR
Žitná 25, 115 67 Praha 1
e-mail: somer@cua.edu

Rolf Stenberg, prof.
Institute of Mathematics,
Helsinki University of Technology
P.O. Box 1100, Helsinki
FIN-02015 HUT Finland
e-mail: Rolf.Stenberg@hut.fi

Zdeněk Strakoš, prof. Ing., DrSc.
Ústav informatiky AV ČR
Pod Vodárenskou věží 2, 182 07 Praha 8
e-mail: strakos@cs.cas.cz

Petr Sváček, RNDr., Ph.D.
Ústav technické mat., Fak. strojní ČVUT
Karlovo nám. 13, 121 35 Praha 2
e-mail: svacek@marian.fsik.cvut.cz

Jakub Šístek, Ing.
Ústav technické mat., Fak. strojní ČVUT
Karlovo nám. 13, 121 35 Praha 2
e-mail: sistek@seznam.cz

Jakub Šolc, Mgr.
Katedra matem., Fak. stavební ČVUT
Thákurova 7, 166 29 Praha 6
e-mail: Jakub.Solc@fsv.cvut.cz

Alena Šolcová, RNDr., Ph.D.
Katedra matem., Fak. stavební ČVUT
Thákurova 7, 166 29 Praha 6
e-mail: solcova@cesnet.cz

Pavel Šolín, RNDr., Ph.D.
Ústav termomechaniky AV ČR
Dolejškova 5, 182 00 Praha 8
e-mail: solin@utep.edu

Miroslav Tůma, prof. Ing., CSc.
Ústav informatiky AV ČR
Pod Vodárenskou věží 2, 182 07 Praha 8
e-mail: tuma@cs.cas.cz

Pavel Váchal, Ing.
Kat. fyzikální elektroniky, FJFI ČVUT
Trojanova 13, 120 00 Praha 2
e-mail: vachal@galileo.fjfi.cvut.cz

Jiří Vala, doc. Ing., CSc.
Ústav matematiky a deskriptivní geom.,
Fakulta stavební VUT
Žižkova 17, 602 00 Brno
e-mail: vala.j@fce.vutbr.cz

Tomáš Vejchodský, RNDr., Ph.D.
Matematický ústav AV ČR
Žitná 25, 115 67 Praha 1
e-mail: vejchod@math.cas.cz

Emil Vitásek, RNDr., CSc.
Matematický ústav AV ČR
Žitná 25, 115 67 Praha 1
e-mail: vitas@math.cas.cz

Jan Vlček, p. m., CSc.
Ústav informatiky AV ČR
Pod Vodárenskou věží 2, 182 07 Praha 8
e-mail: vlcek@cs.cas.cz

Heinrich Voss, Professor
Institute of Numerical Simulation Hamburg, Hamburg University of Technology
D-21071 Hamburg, Germany
e-mail: voss@tu-hamburg.de