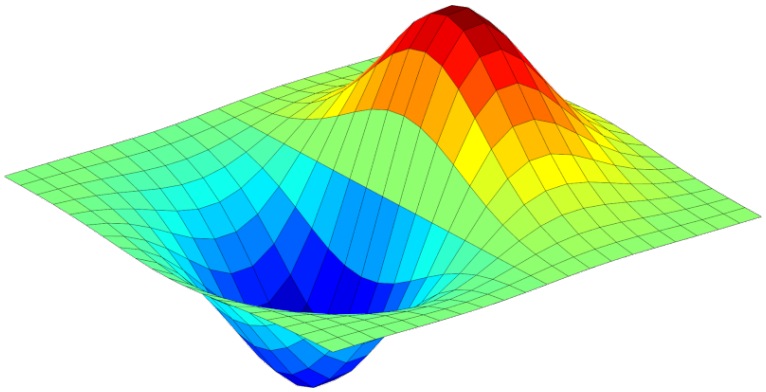
Institute of Geonics of the CAS, Ostrava

SNA'21

Seminar on Numerical Analysis

Modelling and Simulation
of Challenging Engineering Problems



Winter School

Methods of Numerical Mathematics and Modelling,
High-Performance Computing, Numerical Linear Algebra

Ostrava, January 25 – 29, 2021

Programme committee:

Radim Blaheta	Institute of Geonics of the CAS, Ostrava
Stanislav Sysala	Institute of Geonics of the CAS, Ostrava
Dalibor Lukáš	VŠB - Technical University of Ostrava
Jaroslav Kruis	Czech Technical University in Prague
Miroslav Rozložník	Institute of Mathematics of the CAS, Prague
Petr Tichý	Charles University, Prague

Organising committee:

Radim Blaheta	Institute of Geonics of the CAS, Ostrava
Jiří Starý	Institute of Geonics of the CAS, Ostrava
Stanislav Sysala	Institute of Geonics of the CAS, Ostrava
Dagmar Sysalová	Institute of Geonics of the CAS, Ostrava
Marek Pecha	Institute of Geonics of the CAS, Ostrava
Hana Bílková	Institute of Mathematics of the CAS, Prague

Conference secretary:

Dagmar Sysalová	Institute of Geonics of the CAS, Ostrava

Institute of Geonics of the Czech Academy of Sciences
ISBN 978-80-86407-82-1

Preface

Seminar on Numerical Analysis 2021 (SNA'21) is the 15^{th} meeting in a series of SNA events. The previous meetings were held in Ostrava 2003, 2005, Monínec 2006, Ostrava 2007, Liberec 2008, Ostrava 2009, Nové Hrady 2010, Rožnov 2011, Liberec 2012, Rožnov 2013, Nymburk 2014, Ostrava 2015, 2017, and 2019.

SNA events help to bring together the Czech research community working in the field of numerical mathematics and computer simulations. The scope of the seminar ranges from mathematical modelling and simulation of challenging engineering problems, to methods of numerical mathematics, numerical linear algebra, and high performance computing.

SNA'21 has about 80 participants. Its programme includes the traditional Winter School with tutorial lectures focused on selected important topics within the scope of the seminar. This year, the Winter School includes the following lectures:

- *I. Hnětynková* (Charles University, Prague):
  Regularization of large discrete inverse problems by iterative projection methods
- *F. Magoulès* (Université Paris-Saclay):
  Asynchronous iterative methods:
  I – Theory and algorithms
  II – Parallel implementation and applications
- *J. Papež* (Institute of Mathematics of the Czech Academy of Sciences, Prague):
  On the algebraic error in numerical solution of partial differential equations I and II
- *I. Pultarová* (Czech Technical University in Prague):
  Iterative solvers for the stochastic Galerkin method

Beside the Winter School, SNA'21 includes about 25 contributions in the form of short oral presentations.

Due to the current epidemiological circumstances, SNA'21 will differ from the previous meetings. It will take place only remotely in an on-line form. We are aware that such form of the conference cannot fully compensate the classical face-to-face forum including personal discussions or friendly atmosphere with a rich social programme. Despite of this restriction, we believe that SNA'21 will follow the previous events at least on the scientific level. We wish all participants to find lectures interesting and inspiring.

On behalf of the Programme and Organising Committee of SNA'21,

Jiří Starý and Stanislav Sysala

Contents

Winter school lectures

I. Hnětynková:
Regularization of large discrete inverse problems by iterative projection methods

F. Magoulès:
Asynchronous iterative methods:
I – Theory and algorithms
II – Parallel implementation and applications

J. Papež:
On the algebraic error in numerical solution of partial differential equations I and II

I. Pultarová:
Iterative solvers for the stochastic Galerkin method

Stochastic modeling of EGS using continuum-fracture approach

J. Březina, P. Exner, J. Stebel, M. Špetlík

Technical University of Liberec, Faculty of Mechatronics, Informatics and Interdisciplinary Studies

1 Introduction

Although the natural hydrothermal resources are sparse in the Czech Republic, the enhanced geothermal system (EGS) concept is subject of recent research focused on the usage of the geothermal energy. According to the available geological data ([4]), the Litoměřice site is the most attractive due to the presence of a hot crystalline rock in relatively shallow depth. This motivates us to perform a stochastic simulation of a hydraulic stimulation and a long-term operation of an EGS. The main aim is to investigate uncertainty in the operational properties of such EGS caused by the stochastic nature of the fracture system in the rock. The presented EGS model situation is purely artificial, although the concept model and the physical parameters are loosely based on the geologic properties at Litoměřice and other similar EGS sites.

The EGS approach depends on opening and reconnecting the preexisting fracture network in the low permeable rock induced by the hydraulic stimulation and the temperature changes. Since the fracture network is a priori unknown as well as the spatial distribution of the essential parameters, we consider operational properties of the EGS as random variables. In order to predict their distribution, the Monte Carlo method can be applied, providing the forward model and the distribution of input parameters. Considering the forward model, three approaches to the fractured porous media exist: *single continuum* (e.g. in [11] they study uncertainty of the output temperature caused by spatially random input fields in fully coupled THM model), *discrete fracture networks (DFN)*, which essentially use random fractures and thus is often combined with the Monte Carlo method (e.g. [6]) and *continuum-fracture approach* (e.g. in [5] they perform sensitivity analysis of the EGS operation within a fixed fracture network). We adopt the continuum-fracture approach that is much less common especially for significantly more demanding implementation.

2 Forward Model

According to the available geological data for the Litoměřice site, see [4], the crystalline bedrock possibly consisting of granite, gneiss or mica schist is located in the depth below 3 km. Two geological scenarios lead to predicted temperatures in the depth of 5 km: 140 °C and 146 °C. In this depth, we consider two vertical wells in the distance 200 m with opened section 100 m long. The simulation domain is a box with dimensions $600 \times 600 \times 600$ m covering only the close neighborhood of the fractured volume with the two wells placed in its center.

We consider a simple two stage forward model. At first the hydraulic stimulation is described by the hydro-mechanical (HM) poroelastic model combining the discrete fractures and the continuum. Linear elasticity is used, contact conditions and thermal effects are not considered, although these may have significant impact. Based on the stimulation model, the changed aperture and conductivity fields on the fractures are determined. Then the evolution of the temperature and the raw power output is simulated by the thermo-hydraulic (TH) model.

The simulations are computed by our software Flow123d ([3]) which provides both HM and TH coupled models using the reduced dimensional concept [2]. The detailed model formulation can be found in the documentation of the software. Both models use the same computational mesh containing both fractures and the surrounding rock. The meshing itself is not trivial since the fractures can intersect arbitrarily, resulting in a very complex geometry. We use our own Python code together with GMSH meshing tool [8] to create a mesh of sufficient quality. With the used fracture distribution we are able to work with hundreds of fractures.

The fracture stochastic model consists of the Fisher's distribution for the fracture orientation, the power law for the fracture size, Poisson process for the number of fractures and uniform distribution for the fracture centers. Square fractures are considered.

In order to obtain fracture sets of the size tractable by our simulation tools, we represent explicitly only the fractures in a particular scale range, while the smaller ones are treated as equivalent porous media. The upper bound is limited by the diameter of the simulation domain. Since no fracture statistics data for the target locality are available, we use data for the Forsmark site in Sweden ([7]) with a crystalline bedrock.

The HM model is governed by Biot's poroelasticity equations. During the hydraulic stimulation the fluid is injected into both wells in order to open the preexisting fractures. This leads to increase of aperture of the fractures, which is then used in the production phase model. The liquid is injected for 1 day under the stimulation pressure of 50 MPa.

The mechanical and hydromechanical properties for granitic rocks are gathered from different sources. The Biot's coefficient for different rocks is hard to obtain, for our purpose we used estimates for Grimsel granite from [9]. The values of the drained compressibility are based on [12]. The storativity and its relation to porosity and water compressibility follows [1].The values of physical parameters in the fractured zone are mostly unknown. The aperture of the fractures is considered constant in the input of the HM model. The elastic modulus of fracture is assumed much smaller than in the bulk rock.

The TH model describes the heat transport in the EGS during the operation phase within 30 years after the hydraulic stimulation. In this scenario, we consider a steady Darcy flow. The heat transport model is sequentially coupled with the hydraulic model, it assumes thermodynamic equilibrium between the liquid and the rock. The liquid is continuously injected into the left well and pumped out of the production well causing pressure difference of approx. 5 MPa. The outer boundary is considered impermeable. The injecting temperature is set to 15 °C, the natural convection heat flux is prescribed on the production well. The initial temperature field is considered on the outer boundary, which is given by 10 °C at the Earth surface and the geothermal gradient 30 °C·km^{-1}. The hydraulic conductivity in the bulk rock is set similarly to [10] and [5], where they also consider DFN and surrounding rock.

The fracture aperture is updated according to the HM model. The original conductivity of the fractures is adjusted according to the cubic law. The porosity in fractures is considered much higher than in the rock due to the increased conductivity. Additionally we assume that the hydraulic fracturing impacts also the smaller fractures not included in our fracture network, so we increase the conductivity of the bulk rock in the close neighborhood of the fractures.

3 Numerical results

The forward model is used to sample the production temperature and the heat power in the period of 30 years with 1 year time step. Total number of 1000 executions of the complete forward model was performed using a parallel cluster. The samples were used to estimate the mean and the standard deviance for the output temperature and the output power. Evolution of the mean and standard deviance is plotted in Figure 1.

The histograms for the selected times 1, 15 and 30 years are in Figure 2. Absolute values especially for the power may not be realistic, but the important observation is the relation between the standard deviation and the gap between the stimulated and the reference case. As can be seen from the histograms, the distribution for power is close to normal just slightly skewed for the initial times. Thus the standard variance corresponds to about 70 % of samples, say otherwise: doing the stimulation there is a 15 % chance you get the power below the plotted standard deviation band, which is not very small probability and the value is already pretty close to the reference non-stimulated case.

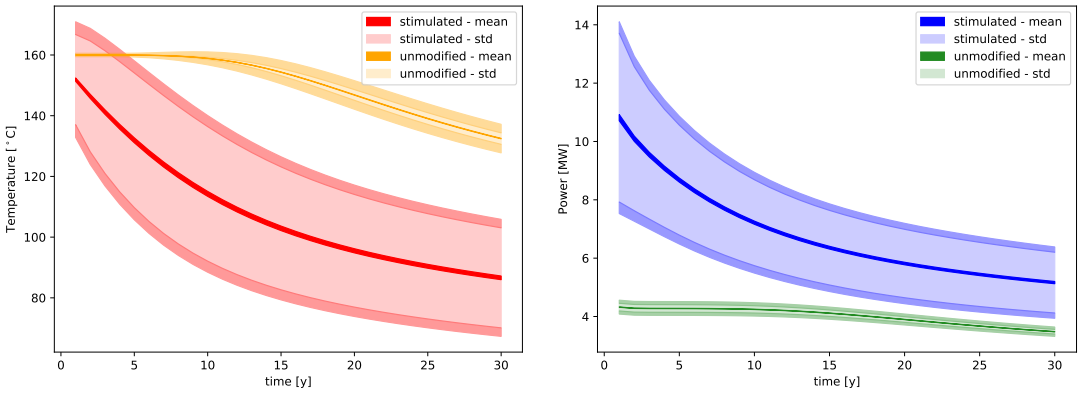


Figure 1: Evolution of the estimated mean output temperature and the output power for the stimulated and unmodified fracture network. The band width is the error of estimation the wide transparent band is the estimate of the standard variance due to the fracture dispersion again with outer band marking the error of the estimation.

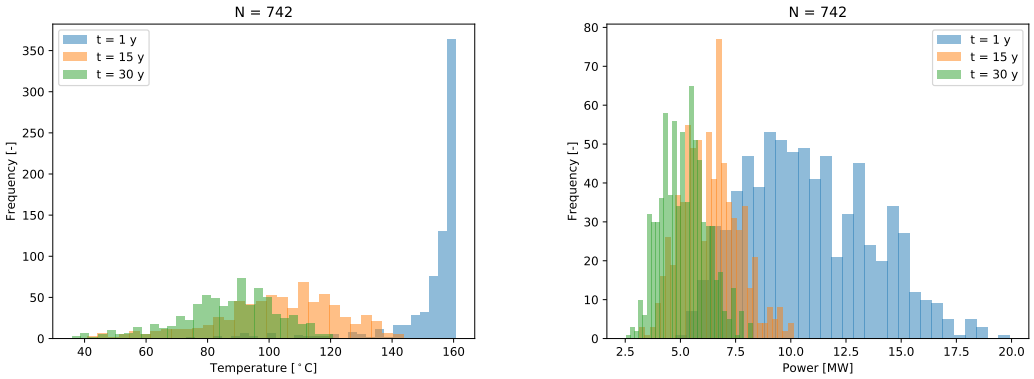


Figure 2: Histogram of the output temperature and power in 1, 15 and 30 years of operation with the stimulated fractures.

4 Conclusions

We developed the HM model for the hydraulic stimulation and the TH model for the heat transfer using the continuum-fracture approach. Although the model is conceptually simple, it neglects important thermal effects and contact conditions during the stimulation, it is yet capable to reproduce a significant effect of the fracture stimulation on the output power and temperature compared to the reference case without stimulation.

We showed that the uncertainty of power due to fracture dispersion is relatively high in comparison with the power increase due to the stimulation. It may indicate relatively high probability that the stimulation may not result in sufficiently high power for the long term production. Further research may investigate if the multiple stimulation can deal with such cases.

Acknowledgments:

The research was supported by the project RINGEN+ No. CZ.02.1.01/0.0/0.0/16_013/0001792, co-funded by the EU Operational Program "Research, Development and Education".

Computational resources were supplied by the project "e-Infrastruktura CZ" (e-INFRA LM2018140) provided within the program Projects of Large Research, Development and Innovations Infrastructures.

References

[1] W.F. Brace, J.B. Walsh, W.T. Frangos: *Permeability of granite under high pressure.* Journal of Geophysical Research (1896-1977), 73(6), 1968, pp. 2225–2236.

[2] J. Březina, J. Stebel: *Analysis of model error for a continuum-fracture model of porous media flow.* In High Performance Computing in Science and Engineering, number 9611 in Lecture Notes in Computer Science, Springer IP, 2015, pp. 152–160.

[3] J. Březina, J. Stebel, P. Exner, J. Hybš: *Flow123d*, 2011–2016, `http://flow123d.github.com`

[4] L. Čápová: *Specification of the Geothermic Model in the Environs of Several Selected Boreholes.* PhD thesis, Charles University in Prague, 2013.

[5] N.G. Doonechaly, R.A. Azim, S.S. Rahman: *Evaluation of recoverable energy potential from enhanced geothermal systems: A sensitivity analysis in a poro-thermo-elastic framework.* Geofluids, 16(3), 2016, pp. 384–395.

[6] S. Ezzedine: *Coupled THMC processes in Geological Media using Stochastic Discrete Fractured Network. Application to HDR Geothermal Reservoirs.* AGU Fall Meeting Abstracts, 41:H41A–0845, Dec. 2008.

[7] S. Follin: *Bedrock hydrogeology Forsmark – Site descriptive modelling, SDM-Site Forsmark.* Technical Report SKB R-08-95, Svensk Kärnbränslehantering AB, 2008.

[8] C. Geuzaine, J.-F. Remacle: *Gmsh: A 3-D finite element mesh generator with built-in pre- and post-processing facilities.* International Journal for Numerical Methods in Engineering, 79(11), 2009, pp. 1309–1331.

[9] P. Selvadurai, P. Selvadurai, M. Nejati: *A Multi-phasic Approach for Estimating the Biot Coefficient for Grimsel Granite.* Solid Earth Discussions, May 2019, pp. 1–17.

[10] J. Šperl, J. Trčková: *Permeability and Porosity of Rocks and Their Relationship Based on Laboratory Testing.* 5(149), 2008, pp. 41–47.

[11] N. Watanabe, W. Wang, C.I. McDermott, T. Taniguchi, O. Kolditz: *Uncertainty analysis of thermo-hydro-mechanical coupled processes in heterogeneous porous media.* Computational Mechanics, 45(4):263, Nov. 2009.

[12] W.A. Zisman: *Compressibility and Anisotropy of Rocks at and near the Earth's Surface.* Proceedings of the National Academy of Sciences, 19(7), July 1933, pp. 666–679â™Ą.

Galerkin method using spline wavelets for second–order integro–differential equations

D. Černá

Technical University of Liberec

1 Introduction

A drawback of most classical numerical methods for integro-differential equations is the full structure of discretization matrices. In comparison, the Galerkin method that uses wavelets as basis functions, often referred to as the wavelet-Galerkin method, leads to sparse matrices; see [2, 3, 4, 5]. The following contribution examines the wavelet-Galerkin method for an equation:

$$-\varepsilon\Delta u + p\,(t)\cdot\nabla u + q\,(t)\,u + \mathcal{K}u = f\,(t) \quad \text{on } \Omega, \tag{1}$$

where $\Omega = (a_1, b_1)\times(a_2, b_2)\times\ldots\times(a_d, b_d)$, $\varepsilon > 0$ is a constant, $p = (p_1, \ldots, p_d)$, ∇u denotes the gradient of u, Δ is the Laplace operator, and $\mathcal{K} : L^2(\Omega) \to L^2(\Omega)$ is given by

$$(\mathcal{K}u)\,(t) = \int_\Omega K\,(t, x)\,u\,(x)\,dx, \tag{2}$$

with the kernel $K \in L^2(\Omega\times\Omega)$. The boundary of Ω is divided into two disjoint pieces $\Gamma^D \neq \emptyset$ and Γ^N such that each facet of Ω belongs either to Γ^D or Γ^N. Equation (1) is equipped with the following boundary conditions

$$u = 0 \text{ on } \Gamma^D \quad \text{and} \quad \frac{\partial u}{\partial \mathbf{n}} = 0 \text{ on } \Gamma^N, \tag{3}$$

where $\mathbf{n}$ stands for the outward pointing unit normal vector.

We use the standard notations for Lebesgue and Sobolev spaces, norms and seminorms, and symbol $\langle\cdot,\cdot\rangle$ for the $L^2(\Omega)$ inner product. Let H be the space of the functions adapted to boundary conditions (3), defined as

$$H = \left\{ f \in H^1(\Omega)\,,\ f = 0 \text{ on } \Gamma^D \right\}, \tag{4}$$

and $a : H\times H \to \mathbb{R}$ be a bilinear form given by

$$a\,(u, v) = \varepsilon\sum_{i=1}^{d}\left\langle\frac{\partial u}{\partial t_i}, \frac{\partial v}{\partial t_i}\right\rangle + \sum_{i=1}^{d}\left\langle p_i\frac{\partial u}{\partial t_i}, v\right\rangle + \langle qu, v\rangle + \langle\mathcal{K}u, v\rangle\,, \quad u, v \in H. \tag{5}$$

The variational formulation of Equation (1) reads as: Find $u \in H$ such that

$$a\,(u, v) = \langle f, v\rangle \quad \forall v \in H. \tag{6}$$

We make the following assumptions:

(A1) The functions p_i, $i = 1, \ldots, d$, and q satisfy $p_i, q \in C\left(\overline{\Omega}\right)$.

(A2) The kernel K is smooth enough, i.e., $K \in C^m\left(\overline{\Omega}\times\overline{\Omega}\right)$ for some $m \in \mathbb{N}$.

(A3) The function f belongs to the space $L^2(\Omega)$.

(A4) The bilinear form a is coercive, which means that there exists a constant $\alpha > 0$ such that

$$a(u,u) \geq \alpha \|u\|_{H^1}^2 \quad \text{for all} \quad u \in H. \tag{7}$$

The following theorem is a consequence of the Lax–Milgram Lemma; for proof, see [3, 4].

Theorem 1. *If the assumptions (A1)–(A4) are satisfied, then there exists a unique solution $u \in H$ of Equation (6).*

2 Wavelet-Galerkin method

Let $\mathcal{J}$ be an index set such that $\lambda \in \mathcal{J}$ takes the form $\lambda = (j,k)$ and let $|\lambda| = j$ denote the level.

Definition 2. *A wavelet basis of a Hilbert space H is defined as the family $\Psi = \{\psi_\lambda, \lambda \in \mathcal{J}\}$ satisfying the following conditions:*

i) *The family Ψ is a Riesz basis for H.*

ii) *The functions are local, i.e., diam supp $\psi_\lambda \leq C2^{-|\lambda|}$ for all $\lambda \in \mathcal{J}$.*

iii) *A wavelet basis has the hierarchical structure*

$$\Psi = \Phi_{j_0} \cup \bigcup_{j=j_0}^{\infty} \Psi_j, \quad \Phi_{j_0} = \{\phi_{j_0,k}, k \in \mathcal{I}_{j_0}\}, \quad \Psi_j = \{\psi_{j,k}, k \in \mathcal{J}_j\}. \tag{8}$$

The functions $\phi_{j_0,k}$ are called scaling functions, and the functions $\psi_{j,k}$ are called wavelets.

iv) *Wavelets have vanishing moments, i.e., $\langle p, \psi_{j,k} \rangle = 0$, $k \in \mathcal{J}_j$, $j \geq j_0$, for any polynomial p of degree less than L, where $L \geq 1$ is dependent on the type of wavelet.*

Hereafter, we assume that Ψ is a wavelet basis of the space $L^2(\Omega)$ such that Ψ when normalized with respect to the H^1-norm is also a wavelet basis in the space H. Furthermore, we assume that this basis on the product domain Ω is constructed via the tensor product of spline-wavelet bases on the interval using the so-called isotropic approach; for details on the construction and examples of wavelet bases, refer to [2, 3, 4].

Let $\Psi^k \subset \Psi$ be a multiscale basis that contains the scaling functions on the coarsest level and the wavelets up to the level k. Then, $X_k = \operatorname{span} \Psi^k$ are the finite-dimensional subspaces of H that are nested and the closure of their union is H.

The Galerkin formulation of (6) reads: Find $u_k \in X_k$ such that

$$a(u_k, v) = \langle f, v \rangle \quad \forall v \in X_k. \tag{9}$$

Theorem 3. *If the assumptions (A1)–(A4) are satisfied, then there exists a unique solution u_k of Equation (9).*

For proof of this theorem, see [4]. Due to well-known Céa's Lemma, the convergence rate depends on spaces X_k and not directly on the chosen bases of these spaces. Since spline wavelet bases typically generate the same spaces as the corresponding splines, the error is similar for the

wavelet-Galerkin method and for the standard Galerkin method with splines. Therefore, for spline-wavelets of order r, we have the error estimate

$$\|u - u_k\|_{H^1} \leq C 2^{-(r-1)k} |u|_{H^r},$$

for any $u \in H \cap H^r(\Omega)$, provided that $r > 1$. For more precise formulation of this statement and its proof, see [3, 4]. Thus, the convergence rate for the wavelet-Galerkin method is $r = 3$ if quadratic spline wavelets are used, and $r = 4$ if cubic spline wavelets are used.

If the integral term is nonzero, then the significant advantage of the wavelet-Galerkin method over classical methods, such as the finite element method, finite difference method, and the Galerkin method with splines, is that discretization matrices can be efficiently approximated by sparse matrices. Another advantage is that a simple diagonal preconditioner is optimal in the sense that diagonally normalized discretization matrices have uniformly bounded condition numbers.

We write the function u_k as

$$u_k = \sum_{\psi_\lambda \in \Psi^k} c_\lambda^k \psi_\lambda. \tag{10}$$

Let matrices $\mathbf{G}^k$ and $\mathbf{K}^k$ and vector $\mathbf{f}^k$ have entries

$$\mathbf{G}_{\mu,\lambda}^k = \varepsilon \langle \nabla\psi_\lambda, \nabla\psi_\mu \rangle + \langle p \cdot \nabla\psi_\lambda, \psi_\mu \rangle + \langle q\psi_\lambda, \psi_\mu \rangle, \quad \mathbf{K}_{\mu,\lambda}^k = \langle \mathcal{K}\psi_\lambda, \psi_\mu \rangle, \quad f_\mu^k = \langle f, \psi_\mu \rangle, \tag{11}$$

for $\psi_\lambda, \psi_\mu \in \Psi^k$. Then, the column vector $\mathbf{c}^k$ of coefficients c_λ^k is the solution of the system

$$\left(\mathbf{G}^k + \mathbf{K}^k\right) \mathbf{c}^k = \mathbf{f}^k. \tag{12}$$

After applying the standard Jacobi diagonal preconditioning on System (12), the condition numbers of the resulting matrices are bounded with a bound independent of the level k, see [3, 4]. The resulting system is typically solved by some iterative method, e.g., by the conjugate gradient method if the system matrix is symmetric and positive definite, or by the GMRES method.

Due to the locality of basis functions, the matrix $\mathbf{G}^k$ corresponding to the differential operator has $\mathcal{O}(N_k \ln N_k)$ nonzero entries; $N_k \times N_k$ being the size of $\mathbf{G}^k$.

As already mentioned, the advantage of wavelet methods is that for some classes of operators and some types of wavelet bases, many entries of the matrix corresponding to the integral operator are small and can be thresholded. Then, the matrix can be approximated with a sparse or quasi-sparse matrix. It was first proven for an integral operator with a singular kernel and orthogonal wavelt bases in [1] and then for other types of wavelet bases and integral operators in [2, 3, 4, 5]. The decay estimate of the entries of the matrix $\mathbf{K}^k$ are for the isotropic wavelet systems and for integral operator with a smooth kernel presented in the following theorem; refer to [4] for proof.

Theorem 4. *Let Ψ be an isotropic wavelet basis of the space $L^2(\Omega)$, $\psi_\lambda, \psi_\mu \in \Psi$ be wavelets with L vanishing moments, and let $K \in C^{2L}\left(\overline{\Omega} \times \overline{\Omega}\right)$. Then there exists a constant C is independent of λ and μ such that*

$$\left|\mathbf{K}_{\mu,\lambda}^k\right| = \left| \int_\Omega \int_\Omega K(x,t)\, \psi_\lambda(x)\, \psi_\mu(t)\, dx\, dt \right| \leq C 2^{-(|\lambda|+|\mu|)(L+d/2)}. \tag{13}$$

3 Application in jump-diffusion option pricing models

Equation (1) represents a wide range of phenomena; e.g., it appears in jump-diffusion models for option pricing. More precisely, let $U(S,t)$ be a value of an option for the price of the underlying asset S and time to maturity T. Then, U can be computed as the solution of an equation

$$\frac{\partial U}{\partial t} - \frac{\sigma^2 S^2}{2}\frac{\partial^2 U}{\partial S^2} - (r - \lambda\kappa) S\frac{\partial U}{\partial S} + (r+\lambda) U - \lambda \int_{-\infty}^{\infty} U(Se^x, t)\, g(x)\, dx = 0 \tag{14}$$

for $S > 0$ and $t \in (0,T)$. The parameters r, σ, λ, and κ are appropriate constants, and g is a probability density function, which depends on the concrete model. By transforming to logarithmic prices and using the Crank-Nicholson scheme for time discretization, the equation of the form (1) is obtained, and the wavelet-Galerkin method can be used for its numerical solution. For more details, numerical experiments, and comparison with other methods, see [2, 4]. Other numerical experiments concerning the wavelet-Galerkin method being used for the solution of integro-differential equations are presented in [3].

4 Conclusions

The great advantage of the Galerkin method with spline wavelets used for the numerical solution of integro-differential equations is the sparse structure of the discretization matrices. This sparsity is a consequence of Theorem 4 about decay estimates. Due to this theorem, many entries of discretization matrices are small and can be thresholded. This process is also called a compression of the matrix. The compressed matrix is then sparse or quasi-sparse. For details about concrete compression strategies, we refer to [3, 4]. The other significant advantage is the uniform boundedness of the condition numbers of diagonally preconditioned matrices. This holds for compressed as well as uncompressed matrices; see [4]. Moreover, in [4], the error and the rate of convergence of the solution was also studied for the compressed system.

Acknowledgement: This work was supported by grant No. PURE-2020-4003 from the Technical University of Liberec.

References

[1] G. Beylkin, R. Coifman, V. Rokhlin: *Fast wavelet transforms and numerical algorithms I.* Comm. Pure Appl. Math. 44, 1991, pp. 141–183.

[2] D. Černá: *Quadratic spline wavelets for sparse discretization of jump-diffusion models.* Symmetry 11, 2019, article no. 999.

[3] D. Černá, V. Finěk: *Galerkin method with new quadratic spline wavelets for integral and integro-differential equations.* J. Comput. Appl. Math. 363, 2020, pp. 426–443.

[4] D. Černá, V. Finěk: *Wavelet-Galerkin method for second-order integro-differential equations on product domains.* In: H. Singh, H. Dutta, M.M. Cavalcanti (eds.): Topics in Integral and Integro-Differential Equations. Springer, 2021, in press.

[5] N. Hilber, O. Reichmann, C. Schwab, C. Winter: *Computational methods for quantitative finance.* Springer, Berlin, 2013.

On the parameter in augmented Lagrangian preconditioning for isogeometric discretizations of the NSE

J. Egermaier, H. Horníková

University of West Bohemia, Faculty of Applied Sciences, Pilsen

1 Introduction

We deal with efficient numerical solution of the incompressible Navier–Stokes equations (NSE) discretized using isogeometric analysis (IgA) approach [1]. IgA exploits the isoparametric approach, i.e., the same basis functions are used for description of the computational domain geometry and also for representation of the solution. One of the main goals of the isogeometric analysis is to keep the exact representation of the geometry independently of the discretization. In practice, the geometry is often described using B-spline/NURBS objects, therefore the B-spline/NURBS basis is also used to represent the solution. The IgA discretization basis has several specific properties different from standard finite element basis, most importantly a higher interelement continuity leading to denser matrices of the resulting linear systems. Our aim is to develop efficient solvers for these systems based on preconditioned Krylov subspace methods. Based on our comparison of several state-of-the-art block preconditioners for linear systems arising from the IgA discretization of the incompressible NSE, the augmented Lagrangian (AL) preconditioner and its modified version (MAL) seems to be very promising. However, their effectiveness is strongly parameter dependent. In this contribution, we focus on the optimal setting of these preconditioners for different IgA discretizations.

2 Problem formulation

Let $\Omega \subset \mathbb{R}^d$ be a bounded domain, d being the number of spatial dimensions, with the boundary $\partial\Omega$ consisting of two complementary parts, Dirichlet $\partial\Omega_D$ and Neumann $\partial\Omega_N$. The steady-state incompressible Navier–Stokes problem is given as a system of $d+1$ differential equations together with boundary conditions

$$\begin{aligned} -\nu\Delta\mathbf{u} + (\mathbf{u}\cdot\nabla)\mathbf{u} + \nabla p &= \mathbf{0} && \text{in } \Omega, \\ \nabla\cdot\mathbf{u} &= 0 && \text{in } \Omega, \\ \mathbf{u} &= \mathbf{g}_D && \text{on } \partial\Omega_D, \\ \nu\tfrac{\partial\mathbf{u}}{\partial\mathbf{n}} - \mathbf{n}p &= \mathbf{0} && \text{on } \partial\Omega_N, \end{aligned} \tag{1}$$

where $\mathbf{u}$ is the flow velocity, p is the kinematic pressure, ν is the kinematic viscosity and $\mathbf{g}_D$ is a given function.

We consider linearization of the nonlinear problem (1) using the Picard's iteration method and discretize it using isogeometric analysis approach, see [2] for details. We limit ourselves to the B-spline discretization basis in this work. The discretization, similarly to the finite element method, leads to a sparse non-symmetric linear system of saddle-point type

$$\begin{bmatrix} \mathbf{F} & \mathbf{B}^T \\ \mathbf{B} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ \mathbf{p} \end{bmatrix} = \begin{bmatrix} \mathbf{f} \\ \mathbf{g} \end{bmatrix}, \tag{2}$$

where $\mathbf{F}$ is block diagonal with the diagonal blocks consisting of the discretization of the viscous term and the linearized convective term, $\mathbf{B}^T$ and $\mathbf{B}$ are discrete gradient and negative divergence operators, respectively. The linear systems are usually very large and need to be solved at every Picard iteration. An efficient iterative solver is necessary because direct solution is unfeasible for real-world problems. Krylov subspace methods are the most commonly used in similar applications and can be very efficient if combined with a good preconditioning technique. Since our matrices are non-symmetric, we choose a Krylov subspace method GMRES.

3 Augmented Lagrangian Preconditioners (AL, MAL)

The augmented Lagrangian approach, proposed by Benzi and Olshanskii [3], belongs to a family of block triangular preconditioners. Block triangular preconditioners are based on the following block decomposition of the system matrix

$$\begin{bmatrix} \mathbf{F} & \mathbf{B}^T \\ \mathbf{B} & \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{B}\mathbf{F}^{-1} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{F} & \mathbf{B}^T \\ \mathbf{0} & \mathbf{S} \end{bmatrix}, \tag{3}$$

where $\mathbf{S} = -\mathbf{B}\mathbf{F}^{-1}\mathbf{B}^T$ is the Schur complement. The preconditioner is in the form

$$\mathcal{M} = \begin{bmatrix} \mathbf{F} & \mathbf{B}^T \\ \mathbf{0} & \hat{\mathbf{S}} \end{bmatrix}, \tag{4}$$

where $\hat{\mathbf{S}}$ is some approximation of $\mathbf{S}$. An overview of these preconditioners can be found, e.g., in [4].

The augmented Lagrangian approach is based on replacing the original system (2) with an equivalent system

$$\begin{bmatrix} \mathbf{F}_\gamma & \mathbf{B}^T \\ \mathbf{B} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ \mathbf{p} \end{bmatrix} = \begin{bmatrix} \mathbf{f}_\gamma \\ \mathbf{g} \end{bmatrix}, \tag{5}$$

where $\mathbf{F}_\gamma = \mathbf{F} + \gamma\mathbf{B}^T\mathbf{W}^{-1}\mathbf{B}, \mathbf{f}_\gamma = \mathbf{f} + \gamma\mathbf{B}^T\mathbf{W}^{-1}\mathbf{g}, \gamma > 0$ is a parameter and $\mathbf{W}$ is a positive definite matrix. The system (5) is then preconditioned with the block triangular preconditioner

$$\mathcal{M}_{AL} = \begin{bmatrix} \mathbf{F}_\gamma & \mathbf{B}^T \\ \mathbf{0} & \hat{\mathbf{S}}_{AL} \end{bmatrix}, \tag{6}$$

where the inverse of the Schur complement approximation is given by

$$\hat{\mathbf{S}}_{AL}^{-1} := -\nu\tilde{\mathbf{M}}_p^{-1} - \gamma\mathbf{W}^{-1} \tag{7}$$

and $\tilde{\mathbf{M}}_p$ is a pressure mass matrix approximation, usually a diagonal matrix. The matrix $\mathbf{W}$ is often chosen to be equal to $\tilde{\mathbf{M}}_p$.

Of course, the choice of the parameter γ is important. A large value would lead to small number of iterations of the preconditioned Krylov method, but for large γ, the block $\mathbf{F}_\gamma$ becomes increasingly ill-conditioned [5]. Hence, it is often set $\gamma \approx 1$. The influence of this parameter on convergence of GMRES method and properties of the solution is tested in our numerical experiments.

We use a direct solver for solving all subsystems in this work. However, the additional term $\gamma\mathbf{B}^T\mathbf{W}^{-1}\mathbf{B}$ makes the matrix $\mathbf{F}_\gamma$ less sparse compared to $\mathbf{F}$ and introduces a coupling between the velocity components which is not present in the discretization of the Picard linearization of

the Navier–Stokes equations. Direct solution of these subsystems becomes very expensive and finding an effective approximate solver can be difficult. A special multigrid method for this block was developed in [3] considering some particular FEM discretizations. Although we have some good results with "standard" geometric multigrid methods, at the same time it turns out that using specialized multigrid methods for IgA will be necessary in the case of higher degrees of B-splines.

One way to simplify the solution of the systems with $\mathbf{F}_\gamma$ is the modified version of AL preconditioner (MAL) [6]. Let us denote the particular blocks of $\mathbf{F}_\gamma$ in two dimensions as follows

$$\mathbf{F}_\gamma = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}. \tag{8}$$

The modified approach suggests to replace this block by its upper block triangle

$$\tilde{\mathbf{F}}_\gamma = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{0} & \mathbf{A}_{22} \end{bmatrix}, \tag{9}$$

such that instead of solving the whole system at once, we solve two smaller systems with the blocks A_{11} and A_{22}. These blocks can be interpreted as discrete anisotropic convection-diffusion operators, thus, applying $\tilde{F}_\gamma^{-1}$ requires solving two anisotropic convection-diffusion problems. The situation is similar in three dimensions, where we have to solve three subsystems.

4 Numerical experiments

We chose several test problems to show and compare convergence properties of GMRES method with the AL preconditioners with respect to the parameter γ. One of the test examples is laminar flow in a 2D blade row obtained by unfolding a cylindrical cross-section of a water turbine runner wheel. The second test example is the standard benchmark problem of 2D flow over a backward facing step domain. Both steady and time-dependent problems have been considered.

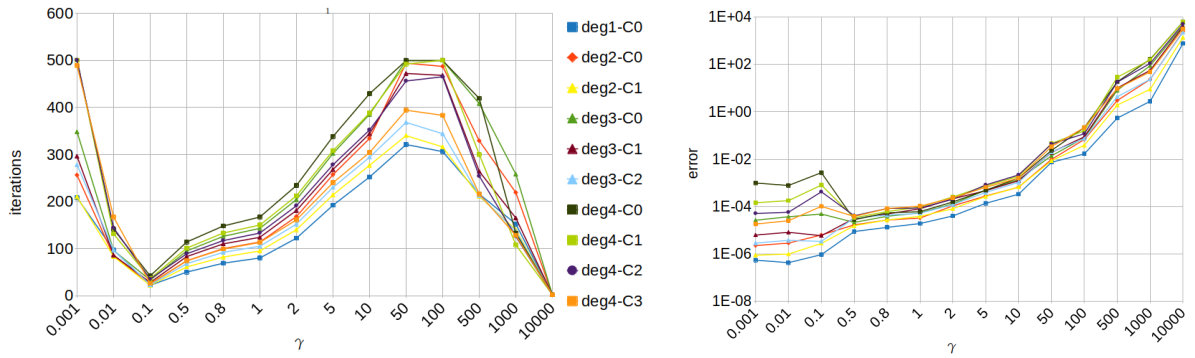


Figure 1: Number of iterations (left) and error (right) of GMRES method dependent on values of parameter γ.

For each test problem, we are interested in iterative solution of one linear system obtained after performing several Picard iterations (steady problem) or time steps (time-dependent problem) iterations. We solve the system using GMRES with no restarts with AL preconditioners with different values of γ. We start from a zero initial solution and stop when the relative residual norm smaller than 10^{-6} is reached. Fig. 1 shows an example of such experiment. We display

the iteration count (left) and the solution "error" depending on the values of the parameter γ. The solution error is defined as

$$\text{error} = \frac{||\mathbf{u}_{\text{GMRES}} - \mathbf{u}_{\text{direct}}||}{||\mathbf{u}_{\text{direct}}||},$$

where $\mathbf{u}_{\text{direct}}$ is a solution of the system (2) obtained with a direct method. This example shows the dependence of GMRES solution with MAL preconditioner for the steady 2D backward facing step problem (Reynolds number Re = 100) discretizied by B-splines of different degrees and continuities.

5 Conclusion

We present a study of the influence of the parameter γ on the convegence of the augmented Lagrangian-based preconditioners for several test problems discretized using B-splines of various degree and interelement continuity. We observe that the optimal value of parameter γ is not dependent on the degree nor continuity of the B-spline discretization, there also seems to be no dependence on the Reynolds number. In the case of AL preconditioner, the number of iterations of GMRES decreases with the increasing value of γ, but the error of the solution increase. In the case of MAL preconditioner, the value of γ which is optimal from the iteration count point of view is sufficiently small to have the error of the solution small too. The errors are larger in time-dependent cases in general. From these experiments, we expect that an optimal value of γ chosen for a particular problem will be suitable also for discretizations with different degree and continuity of the B-spline discretization basis and for different values of Reynolds number.

Acknowledgement: This work was supported by the Czech Science Foundation (GA ČR) grant No. 19-04006S.

References

[1] T. Hughes, J. Cottrel, Y. Bazilevs: *Isogeometric analysis: CAD, finite elements, NURBS, exact geometry and mesh refinement.* Computer Methods in Applied Mechanics and Engineering 194, 2005, pp. 4135–4195.

[2] B. Bastl, M. Brandner, J. Egermaier, K. Michálková, E. Turnerová: *Isogeometric analysis for turbulent flow.* Mathematics and Computers in Simulation 145, 2018, pp. 3–17.

[3] M. Benzi, M.A. Olshanskii: *An augmented Lagrangian-based approach to the Oseen problem.* SIAM J. Sci. Comput. 28(6), 2006, pp. 2095–2113.

[4] A. Segal, M. ur Rehman, C. Vuik: *Preconditioners for incompressible Navier–Stokes solvers.* Numerical Mathematics: Theory, Methods and Applications 3, 2010, pp. 245–275.

[5] X. He, M. Neytcheva, S.S. Capizzano: *On an augmented Lagrangian-based preconditioning of Oseen type problems.* BIT Numer Math 51, 2011, pp. 865–888.

[6] M. Benzi, M.A. Olshanskii, Z. Wang: *Modified augmented Lagrangian preconditioners for the incompressible Navier–Stokes equations.* Int. J. Numer. Meth. Fluids 66(4), 2011, pp. 486–508.

Generalized spectrum of second order differential operators

T. Gergelits[1], *K.-A. Mardal*[2], *B.F. Nielsen*[3], *Z. Strakoš*[1]

[1] Charles University, Prague
[2] University of Oslo
[3] Norwegian University of Life Sciences, Ås

1 Introduction

Standard results in the field of operator preconditioning bound the extreme eigenvalues of the preconditioned matrix independently of the mesh size of the discretization (and eventually also of some problem parameters). However, as Krylov subspace methods are strongly nonlinear in the input data (matrix and the initial residual), convergence bounds based on single number characteristics, such as the condition number, are typically incapable to capture the actual convergence behavior.

This contribution summarizes results from the PhD thesis [4] and the included papers [2] and [3]. Motivated by the results in [1], we investigate the spectra of infinite dimensional operators $-\nabla \cdot (k(x)\nabla)$ and $-\nabla \cdot (K(x)\nabla)$, where $k(x)$ is a scalar coefficient function and $K(x)$ is a symmetric tensor function, preconditioned by the Laplace operator. Subsequently, the focus is on the eigenvalues of the matrices that arise from the discretization using conforming finite elements. Assuming continuity of $K(x)$, it is proved that the spectrum of the preconditioned infinite dimensional operator is equal to the convex hull of the ranges of the diagonal function entries of $\Lambda(x)$ from the spectral decomposition $K(x) = Q(x)\Lambda(x)Q^T(x)$. The other main contribution states that in the discrete case the values of $k(x)$ give close approximations to all individual eigenvalues of the associated preconditioned matrix.

2 Main results

2.1 Scalar coefficient function $k(x)$

Here we consider a second order elliptic PDE

$$\begin{aligned} -\nabla \cdot (k(x)\nabla u) &= f \quad \text{for } x \in \Omega, \\ u &= 0 \quad \text{for } x \in \partial\Omega, \end{aligned} \tag{1}$$

with the uniformly positive and bounded scalar function $k(x)$ on an open and bounded domain $\Omega \subset \mathbb{R}^n$. We consider the standard FEM discretization with nodal polynomial basis functions ϕ_j of the weak formulation of (1) preconditioned by the Laplace operator,

$$[\mathbf{A}]_{ij} = \int_\Omega \nabla\phi_i \cdot k\nabla\phi_j, \tag{2}$$

$$[\mathbf{L}]_{ij} = \int_\Omega \nabla\phi_i \cdot \nabla\phi_j, \quad i,j = 1,\dots,N. \tag{3}$$

Theorem 1. *Let $k(x)$ from (1) be a bounded and piecewise continuous function. Then there exists a (possibly non-unique) permutation π such that the eigenvalues of the matrix $\mathbf{L}^{-1}\mathbf{A}$ given by (2) and (3) satisfy*

$$\lambda_{\pi(j)} \in k(\mathcal{T}_j), \quad j = 1, \dots, N, \tag{4}$$

where $\mathcal{T}_j = supp(\phi_j)$ and where the intervals $k(\mathcal{T}_j)$ are given as

$$k(\mathcal{T}_j) = \left[\inf_{x \in \mathcal{T}_j} k(x), \ \sup_{x \in \mathcal{T}_j} k(x) \right], \quad j = 1, \dots, N. \tag{5}$$

Corollary 2. *Using the notation and assumption of Theorem 1, consider any point $\hat{x}_j$ such that $\hat{x}_j \in \mathcal{T}_j$. Then the associated eigenvalue $\lambda_{\pi(j)}$ of the matrix $\mathbf{L}^{-1}\mathbf{A}$ satisfies*

$$|\lambda_{\pi(j)} - k(\hat{x}_j)| \leq \sup_{x \in \mathcal{T}_j} |k(x) - k(\hat{x}_j)|, \quad j = 1, \dots, N. \tag{6}$$

If, in addition, $k(x) \in \mathcal{C}^2(\mathcal{T}_j)$, then

$$\begin{aligned} |\lambda_{\pi(j)} - k(\hat{x}_j)| &\leq \sup_{x \in \mathcal{T}_j} |k(x) - k(\hat{x}_j)| \\ &\leq \hat{h} \|\nabla k(\hat{x}_j)\| + \tfrac{1}{2} \hat{h}^2 C_j, \quad j = 1, \dots, N, \end{aligned} \tag{7}$$

where $\hat{h} = \operatorname{diam}(\mathcal{T}_j)$ and C_j is the supremum of second derivatives of the function $k(x)$ over the support $\mathcal{T}_j$. In particular, (6) and (7) hold for any discretization mesh node $\hat{x}_j$ such that $\hat{x}_j \in \mathcal{T}_j$.

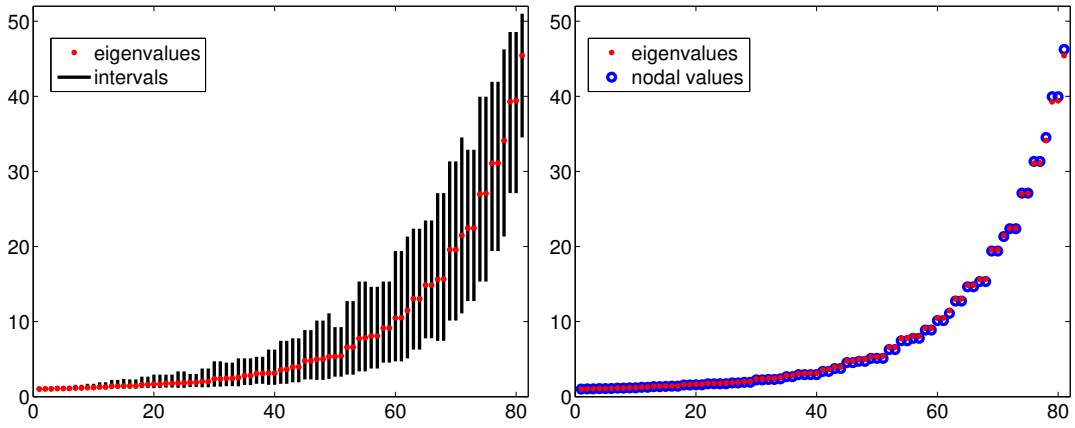


Figure 1: Comparison of the eigenvalues and the increasingly sorted nodal values of $k(x)$ (left) and the associated intervals (right) for test problem (P2) described in [2, Section 4]. The pairing from Theorem 1 can be here defined by sorting of the nodal values of $k(x)$ increasingly.

2.2 Symmetric tensor coefficient function $K(x)$

Here we restrict to two dimensional problems with bounded domain $\Omega \subset \mathbb{R}^2$ and consider more general class of two-dimensional problems with the self-adjoint operator

$$-\nabla \cdot (K(x)\nabla),$$

where the bounded symmetric tensor $K(x)$ has the spectral decomposition

$$K(x) = Q(x) \begin{pmatrix} \kappa_1(x) & 0 \\ 0 & \kappa_2(x) \end{pmatrix} Q^T(x), \tag{8}$$

where $Q(x)$ is an orthogonal matrix function. We fully describe the spectrum

$$\mathrm{sp}(\mathcal{L}^{-1}\mathcal{A}) \equiv \left\{\lambda \in \mathbb{C} : \lambda\mathcal{I} - \mathcal{L}^{-1}\mathcal{A} \quad \text{does not have a bounded inverse}\right\}, \tag{9}$$

of the infinite dimensional operator $\mathcal{L}^{-1}\mathcal{A}$ where

$$\langle \mathcal{A}v, u\rangle = \int_\Omega K\nabla u \cdot \nabla v, \quad u, v \in H_0^1(\Omega), \tag{10}$$

$$\langle \mathcal{L}v, u\rangle = \int_\Omega \nabla u \cdot \nabla v, \quad u, v \in H_0^1(\Omega). \tag{11}$$

Moreover, we present an analogy of Theorem 1 for the spectrum of corresponding discretized operator $\mathbf{L}^{-1}\mathbf{A}$ where

$$[\mathbf{A}]_{ij} = \langle \mathcal{A}\phi_j, \phi_i\rangle = \int_\Omega K\nabla\phi_i \cdot \nabla\phi_j, \tag{12}$$

$$[\mathbf{L}]_{ij} = \langle \mathcal{L}\phi_j, \phi_i\rangle = \int_\Omega \nabla\phi_i \cdot \nabla\phi_j, \quad i, j = 1, \ldots, N. \tag{13}$$

and where ϕ_j are the standard nodal polynomial basis functions.

Theorem 3. *Consider an open and bounded Lipschitz domain $\Omega \subset \mathbb{R}^2$. Assume that the entries of the symmetric tensor $K(x)$ with the spectral decomposition* (8) *are continuous throughout the closure $\overline{\Omega}$. Then the spectrum of the operator $\mathcal{L}^{-1}\mathcal{A}$ given by* (10) *and* (11) *equals*

$$sp(\mathcal{L}^{-1}\mathcal{A}) = \mathrm{Conv}(\kappa_1(\overline{\Omega}) \cup \kappa_2(\overline{\Omega})), \tag{14}$$

where

$$\mathrm{Conv}(\kappa_1(\overline{\Omega}) \cup \kappa_2(\overline{\Omega})) = [\inf_{x\in\overline{\Omega}} \min_{i=1,2} \kappa_i(x), \ \sup_{x\in\overline{\Omega}} \max_{i=1,2} \kappa_i(x)] \tag{15}$$

and where $\kappa_i(x)$, $i = 1, 2$, are the functions from the spectral decomposition (8).

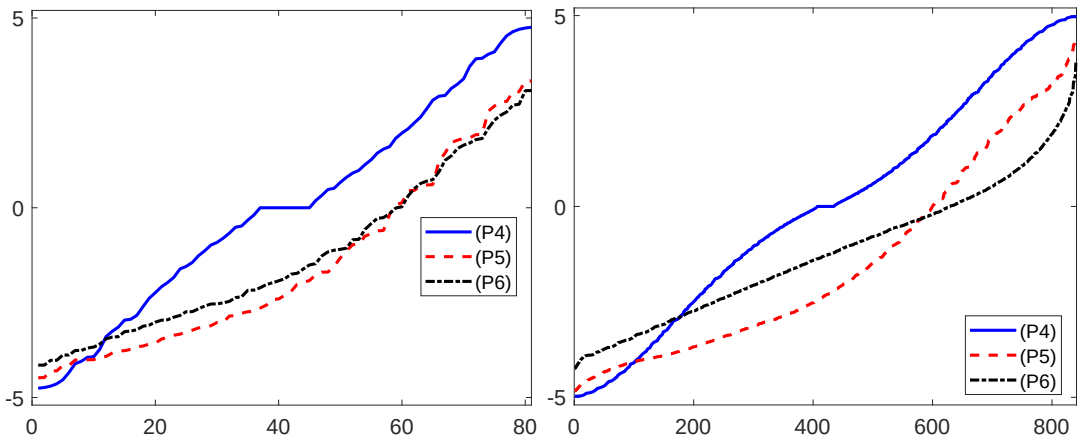


Figure 2: Illustration of Theorem 3 for test problems described in [3, Section 6], for which the interval given in (15) equals to $[-5, 5]$. The eigenvalues tend to fill this whole interval as the resolution of the mesh increases.

Theorem 4. *Let the entries of the symmetric tensor function $K(x)$ with the spectral decomposition* (8) *be bounded and piecewise continuous functions. Then there exists a (possibly non-unique) permutation π such that the eigenvalues of the matrix $\mathbf{L}^{-1}\mathbf{A}$ given by* (12) *and* (13) *satisfy*

$$\lambda_{\pi(j)} \in \kappa(\mathcal{T}_j), \quad j = 1, \ldots, N, \tag{16}$$

where $\mathcal{T}_j = supp(\phi_j)$ *and where the intervals* $\kappa(\mathcal{T}_j)$ *are given as*

$$\kappa(\mathcal{T}_j) \equiv \left[\inf_{x\in\mathcal{T}_j} \min_{i=1,2} \kappa_i(x),\ \sup_{x\in\mathcal{T}_j} \max_{i=1,2} \kappa_i(x) \right], \quad j = 1, \ldots, N. \tag{17}$$

We note that the interval $\kappa(\mathcal{T}_j)$ given by (17) contains the potentially large gap between intervals

$$\kappa_i(\mathcal{T}_j) \equiv \left[\inf_{x\in\mathcal{T}_j} \kappa_i(x), \sup_{x\in\mathcal{T}_j} \kappa_i(x) \right], \quad i = 1, 2, \tag{18}$$

whenever $\kappa_1(\mathcal{T}_j) \cap \kappa_2(\mathcal{T}_j) = \emptyset$.

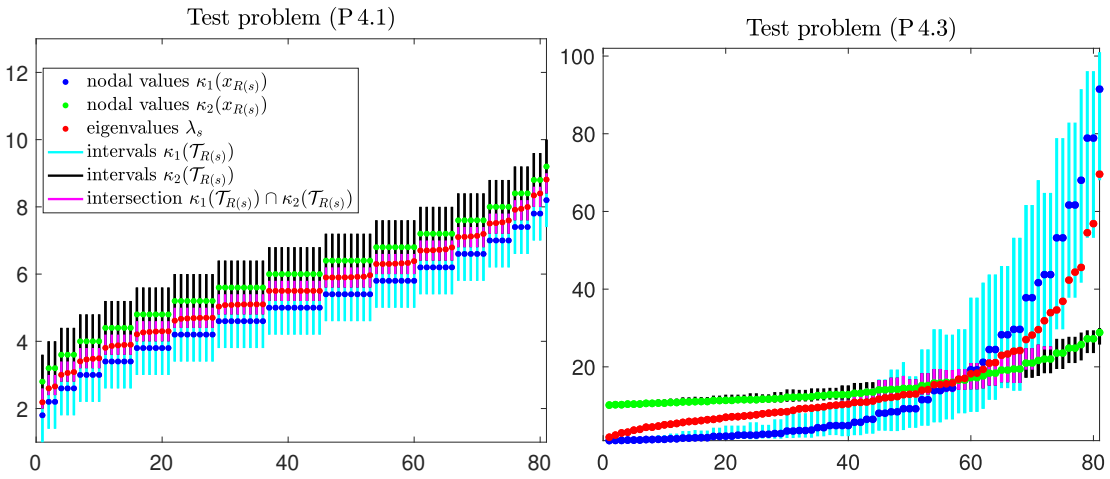


Figure 3: Illustration of Theorem 4 for test problems described in [4, Section 4.5]. The reordering R given by increasingly sorted nodal values of $\kappa_1(x)$ can play the role of a pairing from Theorem 4.

3 Conclusion

Summarizing, the presented results show that detailed information about the spectrum of the preconditioned infinite dimensional operator and the eigenvalues of the preconditioned matrix resulting from its discretization is readily available from the data of the problem. These results seem to be of principal theoretical and practical importance.

References

[1] B.F. Nielsen, A. Tveito, W. Hackbusch: *Preconditioning by inverting the Laplacian: an analysis of the eigenvalues*. IMA Journal of Numerical Analysis 29(1), 2009, pp. 24–42.

[2] T. Gergelits, K.-A. Mardal, B.F. Nielsen, Z. Strakoš: *Laplacian preconditioning of elliptic PDEs: Localization of the eigenvalues of the discretized operator*. SIAM Journal on Numerical Analysis 57(3), 2019, pp. 1369–1394.

[3] T. Gergelits, B.F. Nielsen, Z. Strakoš: *Generalized spectrum of second order differential operators*. SIAM Journal on Numerical Analysis 58(4), 2020, pp. 2193–2211.

[4] T. Gergelits: *Krylov Subspace Methods: Analysis and Application*. PhD thesis, Charles University, Prague, Czech Republic, 2020.

Curve integral of Filippov vector field

T. Hanus, D. Janovská

University of Chemistry and Technology, Prague

1 Simple planar Filippov system

We consider the Filippov system [1, 2] and accept the following simplifications. Let the state space S be the plane, $S = \mathbb{R}^2$. Let it be divided into two unions S_1 and S_2. These unions are open sets in $\mathbb{R}^2$. They are disjoint, $S_1 \cap S_2 = \emptyset$, and when closed they cover the whole state space, $\overline{S_1 \cup S_2} \supset S$. Let Σ_{12} be the regular boundary, which separates S_1 and S_2.

In the simple planar Filippov system, there are given two continuous vector fields $\mathbf{F}_1$ on S_1 and $\mathbf{F}_2$ on S_2. They can be continuously extended to $\overline{S_1}$ and $\overline{S_2}$.

The vector field $\mathbf{F} : S \to \mathbb{R}^2$ defined by parts,

$$\mathbf{F}(\mathbf{x}) = \begin{cases} \mathbf{F}_1(\mathbf{x}), & \mathbf{x} \in S_1, \\ \mathbf{F}_2(\mathbf{x}), & \mathbf{x} \in S_2, \\ \mathbf{G}_{12}(\mathbf{x}), & \mathbf{x} \in \Sigma_{12}, \end{cases} \qquad \mathbf{x} = [x, y] \in S\,, \tag{1}$$

is called the simple planar Filippov vector field.

In the simple planar Filippov system, there is given a scalar function $h : S \to \mathbb{R}$. It decides whether a particular point $\mathbf{x} \in S$ is an element of S_1 or S_2 or Σ_{12}:

$$\begin{array}{lll} \mathbf{x} \in S_1 & \Leftrightarrow & h(\mathbf{x}) > 0\,, \\ \mathbf{x} \in \Sigma_{12} & \Leftrightarrow & h(\mathbf{x}) = 0\,, \\ \mathbf{x} \in S_2 & \Leftrightarrow & h(\mathbf{x}) < 0\,, \end{array} \qquad \mathbf{x} = [x, y] \in S\,.$$

The boundary Σ_{12} is the zero level set of the function h, $\Sigma_{12} = h^{-1}(0)$.

The vector fields $\mathbf{F}_1$ and $\mathbf{F}_2$ imply the scalar function $\sigma : \Sigma_{12} \to \mathbb{R}$,

$$\sigma(\mathbf{x}) = \left(\mathbf{\Sigma}^N(\mathbf{x}) \cdot \mathbf{F}_1(\mathbf{x})\right)\left(\mathbf{\Sigma}^N(\mathbf{x}) \cdot \mathbf{F}_2(\mathbf{x})\right)\,, \quad \mathbf{x} = [x, y] \in \Sigma_{12}\,,$$

where $\mathbf{\Sigma}^N(\mathbf{x})$ is a non-zero normal vector to Σ_{12} at the point $\mathbf{x}$. The function σ decides whether a particular point $\mathbf{x} \in \Sigma_{12}$ is an element of Σ_{12}^C or Σ_{12}^S:

$$\begin{array}{lll} \mathbf{x} \in \Sigma_{12}^C & \Leftrightarrow & \sigma(\mathbf{x}) > 0, \\ \mathbf{x} \in \Sigma_{12}^S & \Leftrightarrow & \sigma(\mathbf{x}) \le 0, \end{array} \qquad \mathbf{x} = [x, y] \in \Sigma_{12}.$$

The function σ divides the boundary Σ_{12} into two disjoint subsets Σ_{12}^C and Σ_{12}^S.

At the points of the crossing set Σ_{12}^C, both vectors $\mathbf{F}_1, \mathbf{F}_2$ point to the same side of Σ_{12} and we define the vector field $\mathbf{G}_{12}$ as their arithmetic mean,

$$\mathbf{G}_{12}(\mathbf{x}) = \frac{1}{2}(\mathbf{F}_1(\mathbf{x}) + \mathbf{F}_2(\mathbf{x})).$$

At the points of the sliding set Σ_{12}^S, the vectors $\mathbf{F}_1, \mathbf{F}_2$ are in other configurations and we apply Filippov method to define the vector field $\mathbf{G}_{12}$,

$$\mathbf{G}_{12}(\mathbf{x}) = \lambda(\mathbf{x})\mathbf{F}_1(\mathbf{x}) + (1 - \lambda(\mathbf{x}))\mathbf{F}_2(\mathbf{x}), \quad \text{where}$$

$$\lambda(\mathbf{x}) = \frac{\boldsymbol{\Sigma}^N(\mathbf{x}) \cdot \mathbf{F}_2(\mathbf{x})}{\boldsymbol{\Sigma}^N(\mathbf{x}) \cdot (\mathbf{F}_2(\mathbf{x}) - \mathbf{F}_1(\mathbf{x}))},$$

and $\boldsymbol{\Sigma}^N(\mathbf{x})$ is a non-zero normal vector to Σ_{12} at the point $\mathbf{x}$.

At the point of double tangency, both vectors $\mathbf{F}_1$ and $\mathbf{F}_2$ are tangent to Σ_{12}. In this case, λ is not defined, "$\lambda = \frac{0}{0}$" , and we define $\mathbf{G}_{12}$ as the arithmetic mean of $\mathbf{F}_1$ and $\mathbf{F}_2$.

The simple planar Filippov vector field $\mathbf{F}$ in (1) is defined by two vector fields $\mathbf{F}_1, \mathbf{F}_2$ and by one scalar function h.

The simple planar Filippov system is the system of differential equations

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = \mathbf{F}(\mathbf{x}), \quad \mathbf{x} = [x, y] \in S\,, \tag{2}$$

where $\mathbf{F}$ is the Filippov vector field (1).

2 Curve integral of a vector field

In the following subsections, we present mathematical statements that apply to continuous vector fields, and generalizations of these statements to Filippov vector fields. The statements are restricted to a region in a plane.

2.1 Continuous vector field

Let S be a region in $\mathbb{R}^2$, on which a continuous vector field $\mathbf{F}$ is defined. The curve integral of the vector field $\mathbf{F}$ along the smooth curve $\mathcal{I} \subset S$ from point $A \in \mathcal{I}$ to point $B \in \mathcal{I}$ is defined as,

$$\int_{\mathcal{I}} \mathbf{F} \cdot \mathrm{d}\Phi\,,$$

where Φ is a smooth parameterization of the curve $\mathcal{I}$, the curve $\mathcal{I}$ is oriented in accordance with the parameterization Φ,

$$\mathcal{I} = \Phi(I)\,,\ \Phi : I \to \mathbb{R}^2\,,\ I = \langle a, b\rangle\,,\ \Phi(a) = A\,,\ \Phi(b) = B\,.$$

Let the integration path from point A to point B be a piecewise smooth curve $\mathcal{K}$, which is an oriented sum of curves $\mathcal{K}_1, \ldots, \mathcal{K}_r$,

$$\mathcal{K} = \mathcal{K}_1 \dot{+} \cdots \dot{+} \mathcal{K}_r\,,$$

where each part $\mathcal{K}_i$ of the curve $\mathcal{K}$ has a smooth parameterization Φ_i,

$$\mathcal{K}_i = \Phi_i(I_i)\,,\ \Phi_i : I_i \to \mathbb{R}^2\,,\ I_i = \langle a_i, b_i\rangle\,,$$

$$\begin{array}{lcccl} \Phi_1(a_1) & = & A_1 & = & A\,, \\ \Phi_1(b_1) & = & B_1 & = & A_2\,, \\ \Phi_i(b_i) & = & B_i & = & A_{i+1}\,, \quad i = 1, \ldots, r-1\,, \\ \Phi_r(b_r) & = & B_r & = & B\,. \end{array}$$

The curve integral of the vector field $\mathbf{F}$ along a piecewise smooth curve $\mathcal{K}$ from point A to point B is the sum

$$\int_{\mathcal{K}} \mathbf{F} \cdot \mathrm{d}\Phi = \sum_{i=1}^{r} \int_{\mathcal{K}_i} \mathbf{F} \cdot \mathrm{d}\Phi_i\,.$$

2.2 Filippov vector field

Let S be a region in $\mathbb{R}^2$, where a Filippov vector field $\mathbf{F}$ is defined. We insert a piecewise smooth curve $\mathcal{K} = \mathcal{K}_1 \dot{+} \cdots \dot{+} \mathcal{K}_r$ into the region S. If the curve $\mathcal{K}$ crosses the boundary Σ_{12}, then the intersections cut the curve $\mathcal{K}$ into a few new parts $\widetilde{\mathcal{K}}_i$. If a part of the curve $\mathcal{K}$ lies completely on the boundary Σ_{12}, then it becomes a newly formed part $\widetilde{\mathcal{K}}_i$. Thus a new subdivision of the curve $\mathcal{K}$ will be created,

$$\mathcal{K} = \widetilde{\mathcal{K}}_1 \dot{+} \cdots \dot{+} \widetilde{\mathcal{K}}_s, \quad r \leq s\,.$$

Each new part $\widetilde{\mathcal{K}}_i\,,\ i = 1, \ldots, s$, is a smooth curve and the Filippov vector field $\mathbf{F}$ is continuous on it up to some endpoints $A_i, B_i\,,$ [4]\,.

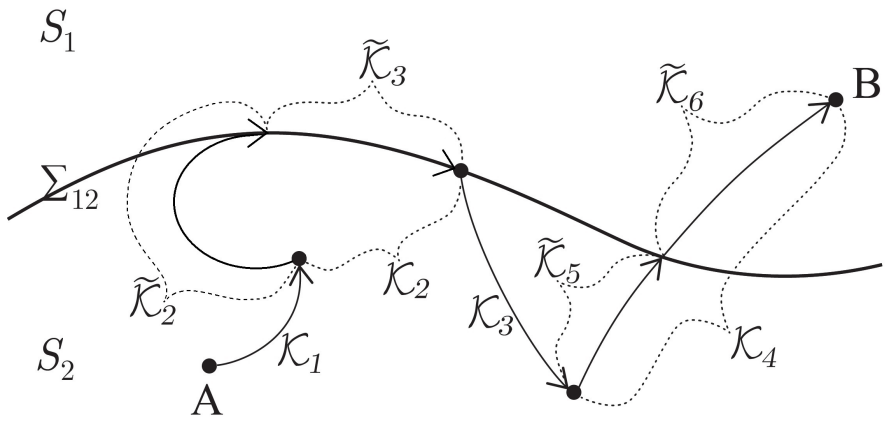


Figure 1: The piecewise smooth curve $\mathcal{K}$ inserted into the state space S.

The curve integral of the Filippov vector field $\mathbf{F}$ along a piecewise smooth curve $\mathcal{K}$ from point A to point B is the sum

$$\int_{\mathcal{K}} \mathbf{F} \cdot \mathrm{d}\Phi = \sum_{i=1}^{s} \int_{\widetilde{\mathcal{K}}_i} \mathbf{F} \cdot \mathrm{d}\Phi_i\,, \tag{3}$$

where Φ_i is a smooth parameterization of the part $\widetilde{\mathcal{K}}_i$ of the curve $\mathcal{K}$.

3 Conclusions

We have shown that the curve integral can be defined and calculated, even if the vector field is only piecewise continuous. Also the scalar potential of the piecewise continuous vector field can be defined and calculated. All detailed formulations, definitions, lemmas and theorems on the existence and properties of the scalar potential of the simple planar Filippov vector field can be found in [4]. This book is already in print. Some applications of Filippov systems in chemistry and biology can be found in [3, 5, 6, 7].

References

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

[2] A.F. Filippov: *Differential equations with discontinuous righthand sides.* Kluwer academic publishers, Dordrecht, 1988.

[3] T. Hanus, D. Janovská: *Discontinuous, piecewise–smooth dynamical systems.* CD-ROM of full texts CHISA 2006, Prague, 2006.

[4] T. Hanus, D. Janovská: *Scalar Potential in Planar Filippov Systems*, to appear in Advances & applications in computational mathematics, River Publishers, 2020.

[5] D. Janovská, M. Biák, T. Hanus: *Some applications of Filippov's dynamical systems.* Journal of Computational and Applied Mathematics 254, 2013, pp. 132–143.

[6] D. Janovská, T. Hanus, M. Biák: *Some applications of piece-wise smooth dynamical systems.* ICNAAM 2010, AIP conference proceedings, Vol. 1281, 2010, pp. 728–731.

[7] D. Janovská, T. Hanus: *Qualitative methods in discontinuous dynamical systems.* ICNAAM 2011, AIP conference proceedings, Vol. 1389, 2011, pp. 1252–1255.

Parameter choice methods for inner-outer regularization in Single Particle Analysis

E. Havelková [1], *I. Hnětynková* [2]

[1] Charles University, Faculty of Mathematics and Physics, Prague and Eyen SE
[2] Charles University, Faculty of Mathematics and Physics, Prague

In this contribution, we concentrate on approaches for the choice of regularization parameters in inner-outer regularization methods. We specifically focus on discrete inverse problems of the form $Ax \approx b$ arising in cryo-electron microscopy single particle analysis. These problems have very specific properties such as an extremely large level of noise in the input data or an atypical form of the point spread function that represents effects of the optics of an electron microscope. We describe a variant of an inner-outer regularization method combining Golub-Kahan iterative bidiagonalization with inner Tikhonov regularization and discuss how the non-standard properties of the studied problem affect its behavior. Namely, we analyze two approaches for the choice of regularization parameter for the inner Tikhonov regularization and study its influence on stopping criteria and the quality of the obtained solution.

Acknowledgement: The authors would like to thank the Eyen SE company for providing the testing data and both hardware and software for GPU computations.

References

[1] E. Havelková: *Regularization methods for discrete inverse problems in single particle analysis.* Masters Thesis. Charles University, Prague, 2019.

[2] C.C. Paige, M.A. Saunders: *LSQR: An algorithm for sparse linear equations and sparse least squares.* ACM Trans. Math. Softw., Vol. 8, No. 1, 1982, pp. 43–71.

[3] P.C. Hansen, D.P. O'Leary: *The Use of the L-Curve in the Regularization of Discrete Ill-Posed Problems.* SIAM Journal on Scientific Computing, Vol. 14, No. 6, 1993, pp. 1487–1503.

Flow and transport in a single fracture calculated from laserscan or tomography data

M. Hokr, A. Balvín, P. Rálek

Technical University of Liberec

1 Introduction

One of the challenges of modelling flow and transport phenomena in porous or fractured rock is capturing its heterogeneity, e.g. the microstructure geometry of pores and fractures, and obtaining the related parameters in the process equations. Standard simulations consider homogenization, i.e. setting scalar or tensor permeability value homogeneous over some volume, which can be obtained by measurement or by an inverse model. Modern measurement methods allow capturing detailed geometry, in particular the laser scanning for surfaces and the computed tomography (CT) for volumes. Use of the microstructure measurement data for studying (not only) rock materials fits to current trends of research, with the aim to predict the macroscopic phenomena and quantities (upscaling). The methods are more developed for quasi-periodic porous media structures, where a small representative pattern can be used for numerical simulation. Solution becomes more difficult for a fracture, when the irregular surface and its opening needs to be represented over whole rock sample volume, rather than for defining homogenized properties. Examples of work are [2] for flow and [3] for rock mechanics using CT data and [1] using laser scan data.

Resolution for laser scan and CT data is often finer than what is possible to use within numerical calculation for practical problem scale under reasonable computing/memory cost. On the other hand, the fine scale is necessary to capture the details of structure controlling the phenomena of interest. Processing of raw CT data into numerical model is also not straightforward and includes several ad-hoc settings and other difficulties.

The presented work is composed of two case studies, based on experiments on splitted granite block with artificial fracture and on granite drill core with natural fracture (kept in a single piece) [4], each processed in a different way into a numerical model.

2 Large-scale artificial fracture problem

The block 80 × 50 × 40 cm is split by an artificial fracture along the largest side and equiped with 8 inlet/outlet holes and a grid of boreholes from one side for measusing sensors. The laser scanning provided the geometry of each fracture side, in a form of x, y, z point cloud in 0.1 mm resolution. The separate coordinate systems were connected afterwards from the scan of the assembled block. The fracture aperture, as a distance of the two surfaces, is therefore determined with uncertainty in the mutual movement of the surfaces, so various parameterization of its correction is considered as a task for inverse modelling.

The single fracture domain is defined as a plane with variable aperture $b(x, y)$ (the spatial variation of position is neglected in this case). The flow in such 2D domain in transversally integrated form is governed by $\vec{q} = -T\nabla(\frac{p}{\varrho g})$ and $\nabla \cdot \vec{q} = 0$ where $\vec{q}$ is the flux density per unit

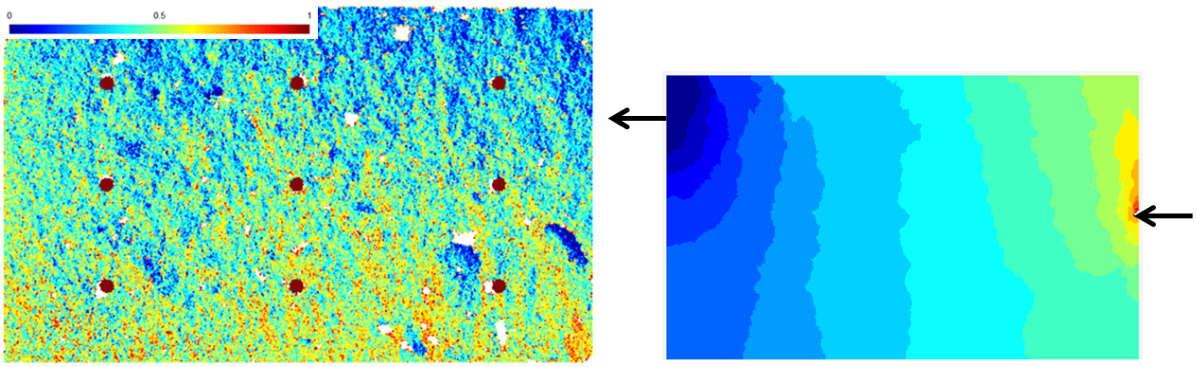


Figure 1: Left: Artificial fracture aperture field (mm) obtained from laser scanning, projected to the numerical mesh. Right: Calculated pressure field for particular inlet and outlet position (arrows).

length in the plane domain [m^2/s], p is the pressure, ϱ is the density, g the gravity accelleration, and $T = \frac{\varrho g}{12\mu} b^3$ is the transmissivity derived by the Hagen-Poiseulle (cubic) law. The distrubution $T(x, y)$ is input element-wise (Fig.1 left). To avoid singularity in T (or complicated domain boundaries), certain minimal aperture $b = 1\mu$m is defined to fully cover the rectangular domain by a permeable fracture. It still generates the variability of three orders of magnitude for b and nine orders of magnitude for T.

The transport model is a standard case of advection and hydrodynamic dispersion. The dispersivity parameters are meant as representing the dispersion in a smaller scale then the aperture field variations captured by the laser scanning data.

The flow and transport simulations by the mixed-hybrid finite elements and discontinuous Galerkin method, respectively, are made by Flow123d [6], the in-house open-source code of the Technical University of Liberec. The inverse solver UCODE (freeware of the US Geological Survey) uses a gradient based method with parameter perturbation sensitivity evaluation.

The calculated pressure field is shown in Fig.1 (right). In the inverse model, aperture field is varied by adding a constant or linear function to get the effective hydraulic resistance corresponding to the measured condition (relation between the pressure difference and the total flow rate). The median of raw measured aperture 0.44 mm is corrected to 0.22 mm with some variation given by individual experiments. Hypothetical uniform fracture of the same resistance is 0.18 mm. A different value of the "transport aperture" can be independently determined from the tracer breakthrough (travel time and effective mobile water volume).

3 Small-scale natural fracture problem

The experimental sample is a cylinder of 71.5 mm diameter and 89.5 mm length, with natural fracture approximately along the cylinder axis. The model domain is a block fitting inside the cylinder and covering the most part of the fracture (88 × 70 × 26 mm). The fracture volume close to the cylinder surface was filled with an isolation material, so not important for the model.

The model problem is constituted on CT data, i.e. a 3D grid of voxels with assigned quantity representing a measure of beam attenuation ("density"). The void space of pores and fractures corresponds to lower value while the solid mineral grains to higher value (for this case, a separation of signal for different minerals is not of interest). Scanning of the sample was made with voxel size of 45 μm and approximately 2000 voxels in one direction.

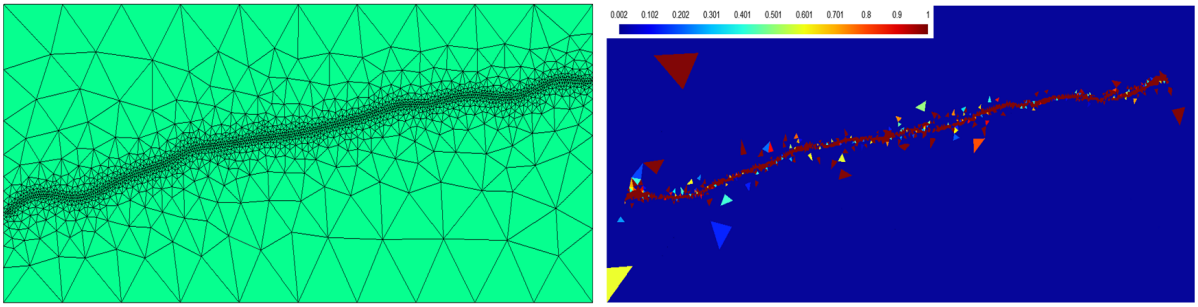


Figure 2: Projecting the CT data into unstructured 3D numerical mesh – porosity field section.

The presented approach is based on transforming and projecting the source CT values into a numerical 3D mesh of heterogeneous permeability for Darcy flow problem, so with use of the same equations and numerical methods as for other groundwater flow problems. This is in contrast with approach when the void space is defined as a complex geometry by the voxel grid assignment to the individual materials (segmentation) and the Stokes equation is solved in this domain, which requires different software tools and numerical methods [5].

We define two thresholds, one representing the fully open space (porosity 1) and one the full solid volume (porosity 0.001 for regularity purposes), the transition in between is a linear range of porosity ($0.001 < \varepsilon < 1$). The porosity is then converted to the local permeability value $k(x, y, z)$, considering each voxel as a single channel (Hagen-Poiseulle law) with its aperture defined as relative part of the voxel size a, i.e. $k = b^2/12$, $b = \varepsilon a$. These values are projected to the numerical discretisation: the element barycentre is assinged with the appropriate voxel value (Fig. 2). To save computing, the numerical mesh is unstructured and refined along the estimated position of the fracture. For the mixed-hybrid FEM in Flow123d, the element number was limited to about 4 millions with 90GB memory computer (cluster Charon at TUL, part of Metacentrum) and the finest size along the fracture could be reached to 3-5 voxels (150-250 μm).

Examples of results are shown in Fig. 3: The velocity field is very nonuniform as expected and confirms the known phenomenon of "channeling" in the fracture plane. The dependence of transmissivity (in fact calculated flow rate) on the chosen CT value threshold shows the uncertainty in the CT data: the voxels of matrix and the fracture do not have well separated peaks on the histogram and therefore other supporting information is necessary to define the actual fracture volume (here the overall hydraulic property, i.e. the sample transmissivity). The results also depend on the discretisation, which could be still too coarse.

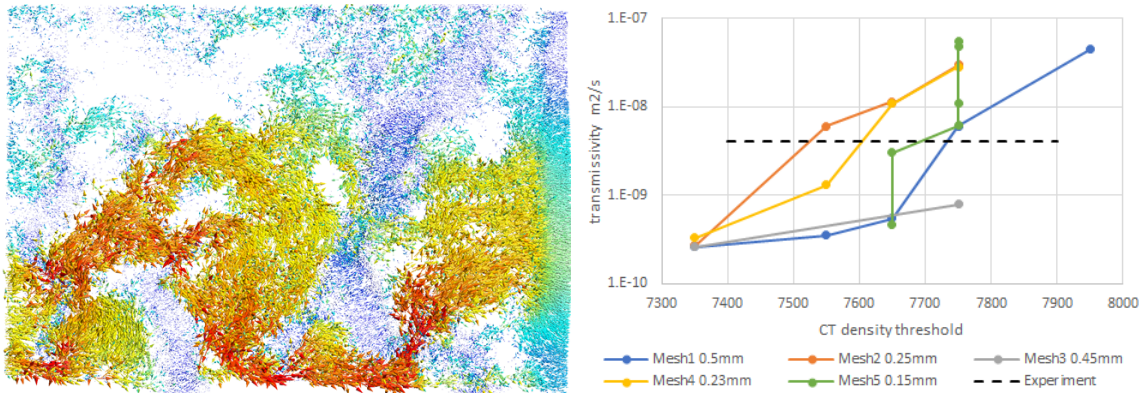


Figure 3: Velocity field calculated from volumetric CT data (a view perpendicular to the fracture, only velocities in the fracture are enough large to be visible, logarithmic range 10^{-9}–10^{-4} m/s). Comparison of experiment and model flow depending on choice of CT threshold.

Acknowledgement: The work was funded by Czech Technological Agency under Project No. TH02030543. The experimental data used for computation are a joint work of the team from ÚJV Řež, CV Řež, TU Liberec and PROGEO lead by F. Jankovský. The CT scanning was made at HZDR Leipzig under leadership of J. Kulenkampff.

Computational resources were supplied by the project "e-Infrastruktura CZ" (e-INFRA LM2018140) provided within the program Projects of Large Research, Development and Innovations Infrastructures.

References

[1] A.S. Aljehani, Y. Da Wang, S.S. Rahman: *An innovative approach to relative permeability estimation of naturally fractured carbonate rocks.* Journal of Petroleum Science and Engineering 162, 2018, pp. 309–324.

[2] M.B. Bird, S.L. Butler, C.D. Hawkes, T. Kotzer: *Numerical modeling of fluid and electrical currents through geometries based on synchrotron X-ray tomographic images of reservoir rocks using Avizo and COMSOL.* Computers & Geosciences 73, 2014, pp. 6–16.

[3] R. Blaheta, R. Kohut, A. Kolcun, K. Souček, L. Staš, L. Vavro: *Digital image based numerical micromechanics of geocomposites with application to chemical grouting.* Int. J. Rock Mech. Min. Sci. 77, 2015, pp. 77–88.

[4] F. Jankovský, M. Zuna, V. Havlová, J. Kotowski, J. Jankovec, M. Hokr, J. Kulenkampff: *Contaminant migration in fractured rocks: insight from tracer tests in core scale.* EGU General Assembly 2019, poster, Vienna.

[5] Math2Market: *GeoDict - the digital material laboratory.* Online: `http://www.geodict.com`

[6] TUL: *FLOW123D version 2.2.1, Documentation of file formats and brief user manual.* NTI TUL, 2019, online: `http://flow123d.github.io`

Analysis of pattern formation using numerical continuation

V. Janovský

Charles University, Faculty of Mathematics and Physics, Prague

1 Solution manifolds and numerical continuation

We consider the Turing instability in the context of reaction-diffusion system for two species $\mathbf{u}$ and $\mathbf{v}$ in the 1D domain, $x \in [0, l]$. The objective of the analysis is *domain size driven instability*, see [4]. The domain can be scaled to the unit interval $0 \le x \le 1$ introducing parameter L. Hence, we consider

$$\mathbf{u}_t = \frac{d_1}{L^2}\,\mathbf{u}_{xx} + f(\mathbf{u},\mathbf{v}) \tag{1}$$

$$\mathbf{v}_t = \frac{d_2}{L^2}\,\mathbf{v}_{xx} + g(\mathbf{u},\mathbf{v}) \tag{2}$$

in the domain $0 \le x \le 1$. Here L is the length of the interval. We consider Neumann boundary conditions (zero flux)

$$\mathbf{u}_x(0,t) = \mathbf{u}_x(1,t) = 0\,, \quad \mathbf{v}_x(0,t) = \mathbf{v}_x(1,t) = 0\,. \tag{3}$$

We seek for *steady states* of the system (1)&(2) which satisfy (3). We assume the existence of *homogeneous steady state*: There exists $u^* \in \mathbb{R}^1$, $v^* \in \mathbb{R}^1$, such that

$$f(\mathbf{u}(x,0),\mathbf{v}(x,0)) = f(u^*,v^*) = g(\mathbf{u}(x,0),\mathbf{v}(x,0)) = g(u^*,v^*) = 0\,,\ 0 \le x \le 1\,.$$

Note that, in this case, $\mathbf{u}_{xx} = 0$ and $\mathbf{v}_{xx} = 0$ in the domain $0 \le x \le 1$.

In order to discretize the above problem we use *method of lines*, i.e. semi-discretization in the spatial variable x. We define the equidistant mesh on the interval $0 \le x \le 1$

$$x_j = jh\,, \quad h = \frac{1}{N+1}\,, \quad j = 1,\dots,N\,, \tag{4}$$

N is the number of meshpoints. The state variables $\mathbf{u}$, $\mathbf{v}$ are approximated by discrete state variables

$$\mathbf{u} \approx [u_1,\dots,u_i,\dots,u_N]^T \in \mathbb{R}^N\,, \quad \mathbf{v} \approx [v_1,\dots,v_i,\dots,v_N]^T \in \mathbb{R}^N, \tag{5}$$

We seek for discrete steady states. They depend on the parameter L^2. The problem can be formulated as a set of $2N$ nonlinear algebraic equations depending on the parameter L^2. We define

$$F : \mathbb{R}^{2N} \times \mathbb{R}^1 \longrightarrow \mathbb{R}^{2N} \tag{6}$$

and seek for the roots

$$F(w,L^2) = 0\,, \quad w \in \mathbb{R}^{2N}\,, \quad w_i = u_i\,, \quad w_{N+i} = v_i\,, \quad i = 1,\dots,N\,. \tag{7}$$

The set (7) is called *solution manifold*. We assume the existence of *homogeneous steady state*

$$F(w^*,L^2) = 0\,, \quad w^* \in \mathbb{R}^{2N}\,, \quad w^*_i = u^*\,, \quad w^*_{N+i} = v^*\,, \quad i = 1,\dots,N\,. \tag{8}$$

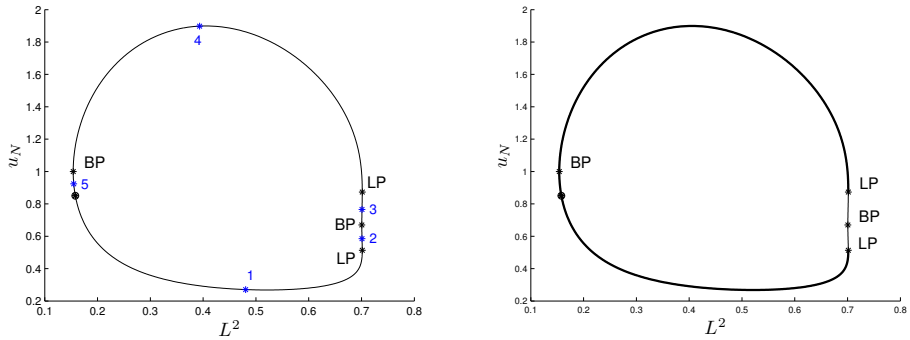


Figure 1: On the left: Branch of inhomogeneous steady states projected onto pairs $\big(u_N, L^2\big)$. The circle on the branch marks the starting point of the continuation procedure. In points $1, \ldots, 5$ we test stability. On the right: thick and thin segments of curves indicate stable and unstable steady states.

Definition 1. *Consider the particular homogeneous steady state* $w^* \in \mathbb{R}^{2N}$, $(L^*)^2 \in \mathbb{R}^1$,

$$F(w^*, (L^*)^2) = 0 \in \mathbb{R}^{2N}\,, \quad A \equiv F_w(w^*, (L^*)^2)\,, \quad \dim \mathrm{Ker}\, A = 1\,. \tag{9}$$

Let ξ *and* η *be right and left eigenvectors corresponding to the zero eigenvalue*

$$A\xi = 0 \in \mathbb{R}^{2N}\,, \quad \|\xi\| = 1\,, \quad A^T\eta = 0 \in \mathbb{R}^{2N}\,, \quad \|\eta\| = 1\,, \quad \eta^T\xi \neq 0,$$

with an algebraic multiplicity equal to one. Then the point $(w^*, (L^*)^2) \in \mathbb{R}^{2N+1}$ *is called* primary bifurcation point *of the system (6).*

Example 1. *Consider the Schnackenberg model [6] for the parameter setting* $a = 0.1$, $b = 0.9$, $\gamma = 10$, $d_1 = 0.1$, $d_2 = 1.6$ *and* $N = 20$. *The coordinate* w^*, *see (8), can be computed as* $u^* = a + b = 1$ *and* $v^* = b/(a+b)^2 = 0.9$. *The aim is to compute the branch of inhomogeneous steady states (7) emanating from a particular primary bifurcation point.*

The primary bifurcation point $w^*, (L^*)^2$ with least $(L^*)^2 > 0$ is $(L^*)^2 = 0.153969537228066$. In Figure 2 this point is labeled as *L2_1_up*. Figure 1 shows the branch (7) projected on pairs $\big(u_N, L^2\big)$. The closed curve on the left is computed by means of continuation software, see [1], [2]. The branch is oriented anticlockwise. The continuation software computes bifurcation points which are (in the anticlockwise direction) LP, BP, LP and BP. It is understood that the primary bifurcation point BP is caused by the crossing of homogeneous steady states (8) with inhomogeneous steady states (7), while the *secondary bifurcation point* BP is not related to such crossing. Both primary and secondary are called *branching points*. The continuation software identifies branch points BP. Distinction (primary/secondary) depends on the context. Branching points BP will be further classified in the Section 3 as symmetry-breaking bifurcation points. Note that the abbreviation LP means *limit point*.

2 Critical wavelengths: primary bifurcation points

We consider the system (1)&(2). Let $u^* \in \mathbb{R}^1$, $v^* \in \mathbb{R}^1$ be a homogeneous steady state. In particular,

$$u^* = a + b\,, \quad v^* = b/(a+b)^2\,, \quad a > 0, b > 0, \tag{10}$$

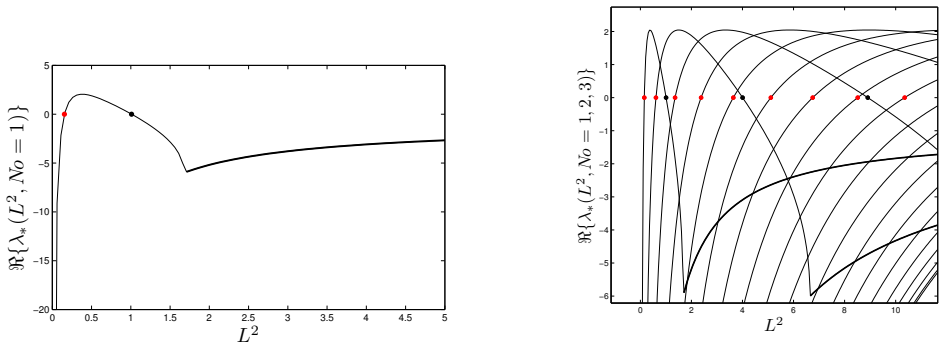


Figure 2: Critical wavelengths, Example 1: On the left: $L2\_1\_up \approx 0.1542$, $L2\_1\_down \approx 1.0098$. On the right: sorted in ascending order: $L2\_1\_up$, $L2\_2\_up$, $L2\_1\_down$, $L2\_3\_up$, $L2\_4\_up$, $L2\_5\_up$, $L2\_2\_down$, $L2\_6\_up$, etc.

is a steady state due to the Schnakenberg model.

The equation

$$\det\left(\mathbf{J} - k^2 \begin{bmatrix} d_1 & 0 \\ 0 & d_1 \end{bmatrix} - \lambda \mathbf{I}\right) = 0\,, \quad \mathbf{I} = \mathbf{I}_{2\times 2} \in \mathbb{R}^{2\times 2} \tag{11}$$

is called *dispersion relation*, see e.g. [5] p. 382., where $\mathbf{J}$ is Jacobian at the steady state. The equation depends on *wavenumber* k^2 and *frequency* λ. The dispersion relation (11) implicitly defines the relationship $\lambda = \lambda(k^2)$. The aim is to analyze the spatial pattern formation via the linear stability analysis. We follow [5], 14.3. General conditions for diffusion-driven instability are well known, see formula (14.29), p. 384. To estimate the stable range of wavelengths in the PDE formulation, we use Fourier analysis ($1{-}D$ Laplacian with Neumann boundary conditions), [5], 14.4. We arrive at the notion of *critical wavelengths*. They appear in pairs. However, when it comes to the discrete model (6), we must consistently consider Laplacian on an equidistant grid (4). We provide closed-form expressions for the calculation of all critical wavelengths. The critical wavelengths are labeled as $L2\_No\_up$ and $L2\_No\_down$ related to the mode number $No = j \in \{1, \ldots, N-1\}$. Examples of critical wavelengths are shown in Figure 2. They are related to Example 1.

We can link the critical wavelength with the primary bifurcation point. Primary bifurcation points appear (generically) in pairs. Thus, we can calculate and sort all $2(N-1)$ bifurcation points.

3 Symmetries of steady states

Consider abstract group $\Gamma = \mathbb{Z}_2 \oplus \mathbb{Z}_2 = \{\iota, \kappa_1, \kappa_2, \kappa_1\kappa_2\}$, $\kappa_1\kappa_2 = \kappa_2\kappa_1$. Here $\mathbb{Z}_2$ is a cyclic group of order 2 and $\mathbb{Z}_2 \oplus \mathbb{Z}_2$ is a direct sum of groups. Γ is therefore an Abelian group. In the state space $\mathbb{R}^{2N}$ there exists a faithful matrix representation of the group Γ. The relevant matrices are not presented in this extended abstract. A key feature of Schnackenberg's model is its Γ-*equivariance*:

$$F(\gamma w, L^2) = \gamma F(w, L^2) \tag{12}$$

for $(w, L^2) \in \mathbb{R}^{2N} \times \mathbb{R}^1$, for all $\gamma \in \{\iota, \kappa_1, \kappa_2, \kappa_1\kappa_2\}$.

The group Γ has proper subgroups $\Sigma_{\kappa_1} = \{\iota, \kappa_1\}$, $\Sigma_{\kappa_2} = \{\iota, \kappa_2\}$, $\Sigma_{\kappa_1\kappa_2} = \{\iota, \kappa_1\kappa_2\}$ a $\Sigma_0 = \{\iota\}$.

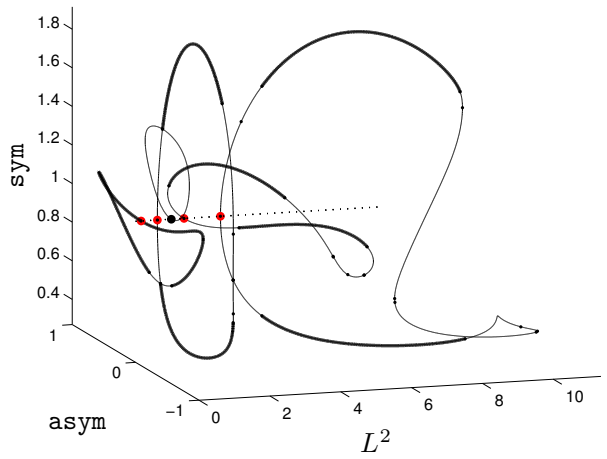


Figure 3: Schnakenberg model, Example 1. Homogeneous steady states (dashed). Branches of inhomogeneous steady states emanating from *L2_1_up*, *L2_2_up*, *L2_1_down*, *L2_3_up*, *L2_4_up*.

Moreover, Σ_{κ_2} and $\Sigma_{\kappa_1\kappa_2}$ are *maximal isotropy subgroups*, see [3], Definition 2.6., p. 78. On that account the steady states $[u_1, \ldots, u_N] \in \mathbb{R}^N$ and $[v_1, \ldots, v_N] \in \mathbb{R}^N$ are

- discrete *odd functions* = *antisymmetric* with a proper shift, in the former case
- discrete *even functions* = *symmetric* in the latter case.

In order to present bifurcation diagram in Figure 3, we apply a symmetry-adapted filter

$$L^2\,, \quad \texttt{asym} \equiv \frac{u_1 - u_N}{2}\,, \quad \texttt{sym} \equiv \frac{u_1 + u_N}{2}\,.$$

The filter reflects properties of the groups Σ_{κ_2} and $\Sigma_{\kappa_1\kappa_2}$.

References

[1] E. Allgower, K. Georg: *Introduction to numerical continuation methods*. SIAM, Philadelphia, 2003.

[2] A. Dhooge, W. Govaerts, Yu.A. Kuznetsov: *MATCONT: a MATLAB package for numerical bifurcation analysis of ODEs*. ACM Transactions on Mathematical Software, Vol. 29, No. 2, 2003, pp 141–164.

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

[4] V. Klika, M. Kozák, E.A. Gaffney: *Domain Size Driven Instability: Self-Organization in Systems with Advection*. SIAM J. Appl. Math., 2018, pp. 2298–2322.

[5] J.D. Murray: *Mathematical biology. II*. Springer-Verlag, New York, 2003.

[6] J. Schnakenberg: *Simple chemical reaction systems with limit cycle behaviour*. J. Theoret. Biol., Vol. 81, 1979, pp. 389–400.

Guaranteed two-sided bounds on all eigenvalues of preconditioned elliptic problems

M. Ladecký, I. Pultarová, J. Zeman

Czech Technical University in Prague

1 Introduction

In 2009, Nielsen, Tveito, and Hackbusch studied in [1] spectra of elliptic differential operators of the type $\nabla \cdot k\nabla$ that are preconditioned using the inverse of the Laplacian. They proved that the range of the scalar coefficient k is contained in the spectrum of the preconditioned operator, provided that k is continuous. Ten years later, Gergelits, Mardal, Nielsen, and Strakoš showed in [2] without any assumption about the continuity of the scalar function k that there exists a one-to-one pairing between the eigenvalues of the preconditioned discretized operator of the type $\nabla \cdot k\nabla$ preconditioned by the inverse of the discretized Laplacian and the intervals determined by the images under k of the supports of the conforming finite element (FE) nodal basis functions used for the discretization.

We introduce guaranteed two-sided bounds on all individual eigenvalues. Similarly as in [2], the bounds can be obtained solely from the data of the original and preconditioning problems defined on supports of the FE basis functions.

2 Diffusion problem

Let $\Omega \subset \mathbb{R}^d$ be a polygonal bounded domain, where $d = 2$ or 3. We consider the diffusion equation with homogeneous Dirichlet boundary conditions

$$\nabla \cdot \boldsymbol{A}\nabla u = f \text{ in } \Omega, \quad u = 0 \text{ on } \partial\Omega.$$

The weak form of the problem reads: find $u \in V = \{v \in H^1(\Omega);\ v = 0 \text{ on } \partial\Omega\}$ such that

$$\int_\Omega \nabla v \cdot \boldsymbol{A}\nabla u \,\mathrm{d}\boldsymbol{x} = \int_\Omega f v \,\mathrm{d}\boldsymbol{x}, \quad v \in V, \tag{1}$$

for $u, v \in V$ and $f \in L^2(\Omega)$; see, e.g., [4] for details. The coefficients $\boldsymbol{A} : \Omega \to \mathbb{R}^{d\times d}$ are assumed to be essentially bounded, i.e. $\boldsymbol{A} \in L^\infty(\Omega; \mathbb{R}^{d\times d})$ is symmetric, and uniformly elliptic (positive definite) in Ω.

3 Discretization and preconditioning

We assume that a conforming FE method is employed to discretize the diffusion (1) problem. The domain Ω is thus decomposed into a finite number of elements $\mathcal{E}_j$, $j = 1, \ldots, N_e$, and continuous FE basis functions (with compact supports) denoted by φ_k, $k = 1, \ldots, N$, are used as approximation and test functions. By $\mathcal{P}_k$ we denote the smallest patch of elements covering the

support of φ_k, $k = 1, \ldots, N$. The stiffness matrix $\mathsf{K} \in \mathbb{R}^{N \times N}$ of the system of linear equations of the discretized problem (1) has elements

$$\mathsf{K}_{kl} = \int_\Omega \nabla\varphi_l(\boldsymbol{x}) \cdot \boldsymbol{A}(\boldsymbol{x}) \nabla\varphi_k(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x}, \quad k, l = 1, \ldots, N.$$

The idea of preconditioning of a system

$$\mathsf{M}\mathsf{u} = \mathsf{B}$$

with a matrix $\widetilde{\mathsf{M}}$ is based on assumptions that a system of linear equations with a matrix $\widetilde{\mathsf{M}}$ is relatively easily solvable and that the spectrum of $\widetilde{\mathsf{M}}^{-1}\mathsf{M}$ is more favorable than that of M regarding some iterative solution method. Considering

$$\widetilde{\mathsf{M}}^{-1}\mathsf{M}\mathsf{u} = \widetilde{\mathsf{M}}^{-1}\mathsf{B} \quad \text{or} \quad \widetilde{\mathsf{M}}^{-1/2}\mathsf{M}\widetilde{\mathsf{M}}^{-1/2}\mathsf{v} = \widetilde{\mathsf{M}}^{-1/2}\mathsf{B},\ \mathsf{u} = \widetilde{\mathsf{M}}^{-1/2}\mathsf{v},$$

can thus lead to equivalent problems that can be solved more efficiently than the original one. Our approach is built on the preconditioning matrix $\widetilde{\mathsf{K}} \in \mathbb{R}^{N \times N}$ obtained for the material data $\widetilde{\boldsymbol{A}}$, such that

$$\widetilde{\mathsf{K}}_{kl} = \int_\Omega \nabla\varphi_l(\boldsymbol{x}) \cdot \widetilde{\boldsymbol{A}}(\boldsymbol{x}) \nabla\varphi_k(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x}.$$

4 Bounds on eigenvalues of preconditioned problems

The lower and upper bounds on the eigenvalues of $\widetilde{\mathsf{K}}^{-1}\mathsf{K}$ for any uniformly positive definite measurable data $\boldsymbol{A}, \widetilde{\boldsymbol{A}} : \Omega \to \mathbb{R}^{d \times d}$ are introduced in this part. Let us build two sequences of positive real numbers

$$\begin{aligned} \lambda_k^{\mathrm{L}} &= \operatorname{ess\,inf}_{\boldsymbol{x} \in \mathcal{P}_k} \lambda_{\min}\left(\widetilde{\boldsymbol{A}}^{-1}(\boldsymbol{x})\boldsymbol{A}(\boldsymbol{x})\right), \\ \lambda_k^{\mathrm{U}} &= \operatorname{ess\,sup}_{\boldsymbol{x} \in \mathcal{P}_k} \lambda_{\max}\left(\widetilde{\boldsymbol{A}}^{-1}(\boldsymbol{x})\boldsymbol{A}(\boldsymbol{x})\right), \end{aligned}$$

every function φ_k having its support inside the patch $\mathcal{P}_k$, $k = 1, \ldots, N$. Thus λ_k^{L} and λ_k^{U} are in the above sense the smallest and largest, respectively, eigenvalues of $\widetilde{\boldsymbol{A}}^{-1}(\boldsymbol{x})\boldsymbol{A}(\boldsymbol{x})$ on the patch $\mathcal{P}_k$.

After inspecting all patches, we sort the two series in non-decreasingly. Thus we obtain two bijections $r, s : \{1, \ldots, N\} \to \{1, \ldots, N\}$ such that

$$\lambda_{r(1)}^L \le \lambda_{r(2)}^L \le \cdots \le \lambda_{r(N)}^L, \qquad \lambda_{s(1)}^U \le \lambda_{s(2)}^U \le \cdots \le \lambda_{s(N)}^U. \tag{2}$$

The lower and upper bounds on the eigenvalues $0 < \lambda_1 \le \lambda_2 \le \cdots \le \lambda_N$ of $\widetilde{\mathsf{K}}^{-1}\mathsf{K}$ are given by (2), i.e.,

$$\lambda_{r(k)}^{\mathrm{L}} \le \lambda_k \le \lambda_{s(k)}^{\mathrm{U}}, \qquad k = 1, \ldots, N. \tag{3}$$

The proof for this claim can be found in our recent publication [5].

5 Numerical experiments

Example 1. *Assume* $d = 2$, $\Omega = (-\pi, \pi)^2$,

$$\boldsymbol{A}(\boldsymbol{x}) = \begin{pmatrix} 1 + 0.3\,\mathrm{sign}(\sin(x_2)) & 0.3 + 0.1\cos(x_1) \\ 0.3 + 0.1\cos(x_1) & 1 + 0.3\,\mathrm{sign}(\sin(x_2)) \end{pmatrix},$$

and a simple and a more sophisticated preconditioning operators with

$$\widetilde{\boldsymbol{A}}_1(\boldsymbol{x}) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \textit{and} \quad \widetilde{\boldsymbol{A}}_2(\boldsymbol{x}) = \begin{pmatrix} 1 & 0.3 \\ 0.3 & 1 \end{pmatrix},$$

respectively. Let us consider uniform grid with piece-wise bilinear FE functions, $N = 10^2$ *or* 30^2*, (Figure 1).*
The numerical experiments illustrate that the bounds on the eigenvalues are guaranteed. We also notice that because $\widetilde{\boldsymbol{A}}_2$ *is point-wise closer to* $\boldsymbol{A}$*, than* $\widetilde{\boldsymbol{A}}_1$*, the spectrum of the second preconditioned problem (together with its bounds) is closer to unity than the spectrum of the problem preconditioned with* $\widetilde{\boldsymbol{A}}_1$*.*

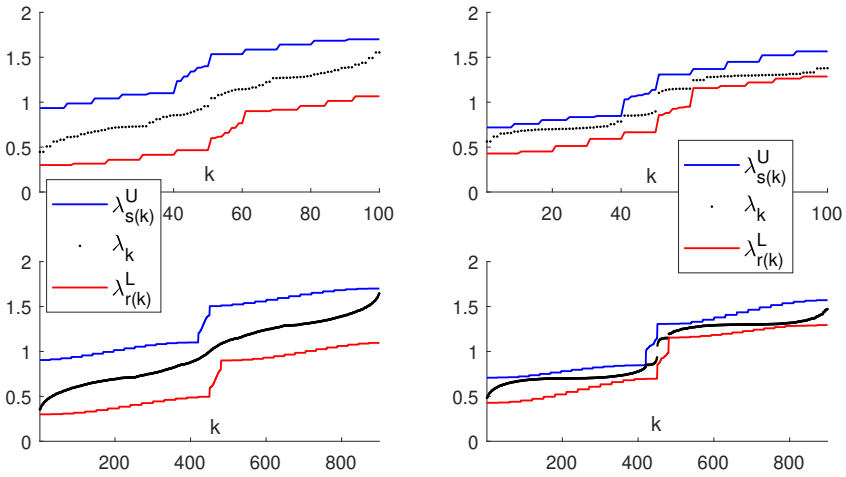


Figure 1: Lower ($\lambda^{\mathrm{L}}_{r(k)}$) and upper ($\lambda^{\mathrm{U}}_{s(k)}$) bounds on eigenvalues λ_k with $N = 10^2$ (top graphs) and $N = 30^2$ (bottom graphs) preconditioned by operators with $\widetilde{\boldsymbol{A}}_1$ (left) and $\widetilde{\boldsymbol{A}}_2$ (right).

Example 2. *Let us consider the test problem of [2, Section 4]: the diffusion equation,* $\Omega = (0,1) \times (0,1)$*,* $\boldsymbol{A}(x_1, x_2) = \sin(x + y)\boldsymbol{I}$*, and homogeneous Dirichlet boundary conditions on* $\partial\Omega$*. Let us use the uniform grid with piece-wise bilinear FE functions,* $N = 9^2$ *or* $N = 19^2$*. For preconditioning we use* $\widetilde{\boldsymbol{A}} = \boldsymbol{I}$*. The appropriatelly ordered bounds provided by [2] and the bounds obtained by our method coincide. They are presented in Figure 2.*

6 Conclusion

Up to our knowledge, [2] is the first paper on characterising all eigenvalues of a preconditioned discretized diffusion operator. Motivated by [2, 3], we further contributed in [5] to this theory by generalising some of these results to vector valued equations with tensor data and with more general boundary conditions preconditioned by arbitrary operators of the same type. Moreover, we provide bounds to every particular eigenvalue. Analogously to [2], the bounds are easily

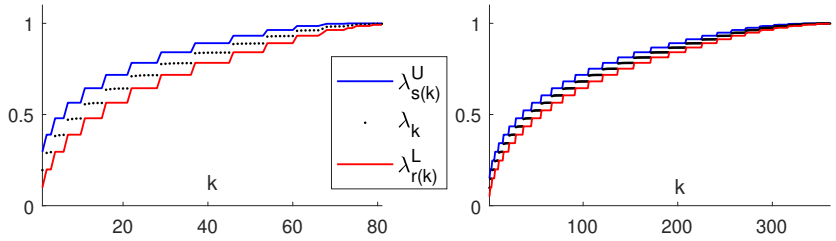


Figure 2: Lower ($\lambda^{\mathrm{L}}_{r(k)}$) and upper ($\lambda^{\mathrm{U}}_{s(k)}$) bounds on eigenvalues λ_k of Example 2 with $N = 9^2$ (left) and $N = 19^2$ (right).

accessible and obtained solely from the data defined on supports of the FE basis functions. If the data are element-wise constant, only $O(N)$ arithmetic operations and sorting of two series of N numbers must be performed.

Acknowledgement: The authors acknowledge the financial support received from the Center of Advanced Applied Sciences, the European Regional Development Fund (project No.CZ.02.1.01/0.0/0.0/16 019/0000778). Martin Ladecký was supported also by the Grant Agency of the Czech Technical University in Prague (project No.SGS20/002/OHK1/1T/11) and by the Czech Science Foundation (project No. 20-14736S).

References

[1] B.F. Nielsen, A. Tveito, W. Hackbusch: *Preconditioning by inverting the Laplacian: an analysis of the eigenvalues*. IMA Journal of Numerical Analysis **29**, 2009, pp. 24–42.

[2] T. Gergelits, K.-A. Mardal, B.F. Nielsen, Z. Strakoš: *Laplacian preconditioning of elliptic PDEs: Localization of the eigenvalues of the discretized operator*. SIAM Journal on Numerical Analysis **57**, 2019, pp. 1369–1394.

[3] T. Gergelits, B.F. Nielsen, Z. Strakoš: *Generalized spectrum of second order differential operators*. SIAM J. Numer. Anal. **58**, 2020, pp. 2193–2211.

[4] A. Ern, J.-L. Guermond: *Theory and Practice of Finite Elements*. Springer, New York, 2004.

[5] M. Ladecký, I. Pultarová, J. Zeman: *Guaranteed Two-Sided Bounds on All Eigenvalues of Preconditioned Diffusion and Elasticity Problems Solved By the Finite Element Method*. Accepted to Applications of Mathematics in October 2020.

[6] I. Pultarová, M. Ladecký: *Two-sided guaranteed bounds to individual eigenvalues of preconditioned finite element and finite difference problems*. Numerical Linear Algebra with Applications, submitted in October 2020.

Results solving the ill-conditioned Hilbert equation systems of rank 36

E.J. Kansa

Convergent Solutions, Livermore, CA

The Hilbert matrix, H, is notoriously ill-conditioned on present computers even though the matrix has an analytic inverses valid to all orders. Consider an arbitrary problem of rank N,

$$Hx = b. \tag{1}$$

If x is a column vector,

$$x = [1, 1, \ldots, 1]^T , \tag{2}$$

then the right hand side b can be found.

In practice, computers have finite word sizes, finite memory, and finite execution speed. However, there are strategies that permits the calculation of ill-conditioned linear systems arising from full matrices, such as the Hilbert matrix and C^∞ radial basis functions. A combination of strategies are employed in this example: use of extended arithmetic precision software found either in Advanpix, or MATHEMATICA. Splitting a large system into many smaller sized systems controls the accumulation of rounding errors.

The rank 36 Hilbert system has a condition number of `6.01e+54` given in (1) is solved with the Advanpix package. The results will be presented.

References

[1] D. Hilbert: *Ein Beitrag zur Theorie des Legendre'schen Polynoms.* Acta Mathematica, Vol. 18, 1893, pp. 155–159.

[2] `http://www.advanpix.com`

[3] E.J. Kansa, P. Holoborodko: *On the ill-conditioned nature of C^∞RBF strong collocation.* Engineering Analysis with Boundary Elements, Vol. 78, 2017, pp. 26–30.

[4] E.J. Kansa, P. Holoborodko: *Fully and sparsely supported radial basis functions.* International Journal of Computational Methods and Experimental Measurements 8(3), 2020, pp. 208–219.

Uncertainties in geotechnical problems described by fuzzy sets

J. Kruis, T. Koudelka, T. Krejčí

Czech Technical University in Prague, Faculty of Civil Engineering, Department of Mechanics

1 Introduction

Soils and rocks are important materials studied in civil engineering. Every structure is founded on soil or rock. Soils are used in embankments of roads and railways, tunnels are located in rocks, etc.

Description of soil and rock materials is usually complicated by very significant uncertainty. The uncertainty can be divided into aleatoric uncertainty and epistemic one. The aleatoric uncertainty stems from the fact that several repetitions of an experiment lead to different results although the conditions of the experiment are tried to be the same. This type of uncertainty could be reduced by higher number of experiments or measurements. Unfortunately, only very limited number of measurements, especially in situ, are usually performed in connection with soils and rocks. The reason is caused by limited budget.

The epistemic uncertainty stems from very complicated properties of any material and any process in nature. Further improvement of measurements encounters new challenges and it reveals new uncertainties. Therefore, the epistemic uncertainty cannot be removed. In connection of soils, the structure of soil is very complicated because there are solid grains and pore space is filled with liquid and gas. The shape of grains as well as shape of the pore space are very complicated which results in obstacles in modelling of liquid and gas transport through the pore space.

Application of methods of probability and mathematical statistics is possible in cases, where reasonable number of measurements and experiments of soils or rocks are available. Because it is not a common case, description of uncertainty based on the fuzzy sets can be used. More precisely, the uncertainty are described by fuzzy numbers. The fuzzy numbers can be constructed from very limited number of measurements which are accompanied by an expert opinion. It has to be emphasized that the opinion of geotechnical engineers are commonly used in civil engineering due to lack of other sources of information. In the classical civil engineering approach, deterministic models and simulations prevail. In these simulations, very conservative values of loads (large quantiles, e.g. 98%) and material strength (small quantiles, e.g. 1%) are used.

2 Fuzzy Numbers

Theory of fuzzy sets was introduced by L.A. Zadeh in 1965 in [1]. The classical set theory is based on the possibility to uniquely decide whether an object is or is not a member of a set. It means, either an element x belongs to a set A which is written in the form $x \in A$, or an element x does not belong to a set A which is written in the form $x \notin A$. A set is an ensemble of elements with property $V(x)$, the set is denoted by $\{x; V(x)\}$. The classical sets are called crisp sets. In fuzzy approach, a membership function, $\mu(x)$, which expresses the degree of truth of the statement $x \in A$, is introduced. The membership function has to be non-negative ($\mu(x) \geq 0$) but usually

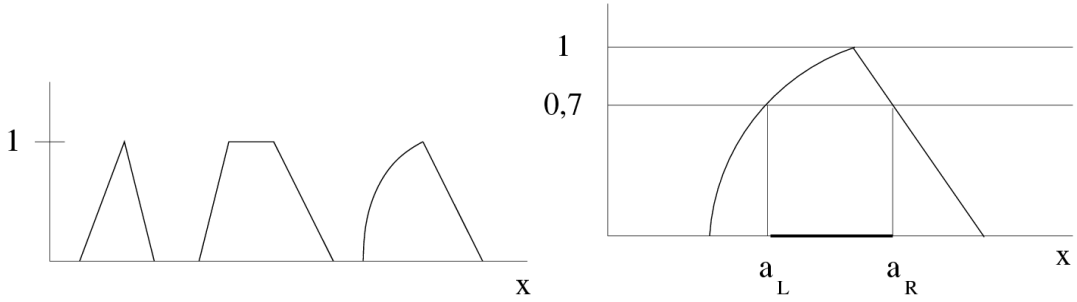


Figure 1: Fuzzy sets, fuzzy numbers (on the left) and α-cuts (on the right).

the so called normalized membership functions are used, where $\mu(x) \in \langle 0; 1\rangle$. If U represents a fundamental set and x are the elements of this fundamental set, to be assessed according to an uncertain proposition and assigned to a subset A of U, the set $A = \{(x, \mu(x)) : x \in U\}$ is referred to as the uncertain set or fuzzy set on U. More details can be found in book [2]. For people who command of Czech language, book [3] is recommended.

There are two limit cases $\mu_A(x) = 1 \Leftrightarrow x \in A$ and $\mu_A(x) = 0 \Leftrightarrow x \notin A$. Interpretation of the membership function can be shown by the following three cases:

$\mu_A(x) = 0$	element x does not belong to the set A
$\mu_A(x) = 1$	element x belongs to the set A
$\mu_A(x) = 0,3$	element x belongs partially to the set A

The support of a fuzzy set A is a crisp set supp $A = \{x; \mu_A(x) > 0\}$. The kernel of a fuzzy set A is a crisp set ker $A = \{x; \mu_A(x) = 1\}$. A fuzzy set A is called normal if ker $A \neq \emptyset$. A fuzzy set A is convex if its membership function monotonically decreases on each side of the maximum value

$$\mu_A(x_2) \geq \min\{\mu_A(x_1), \mu_A(x_3)\} \quad \forall x_1, x_2, x_3 \in X,\ x_1 \leq x_2 \leq x_3 \tag{1}$$

A fuzzy number $\tilde{a}$ is a convex, normalized fuzzy set $A \subseteq R$ whose membership function is at least segmentally continuous and has the functional value $\mu_A(x) = 1$ at precisely one of the x values. On the left in figure 1, there is a fuzzy number, a fuzzy set (because $\mu(x) = 1$ for more than one x) and a fuzzy number.

Manipulation with the fuzzy numbers can be efficiently done with the help of α-cuts. α-cut of fuzzy set A, where $\alpha \in \langle 0; 1\rangle$, is a crisp set $A_\alpha = \{x; \mu_A(x) \geq \alpha\}$. α-level of fuzzy set A, where $\alpha \in \langle 0; 1\rangle$, is a crisp set $A^\alpha = \{x; \mu_A(x) = \alpha\}$. On the right of figure 1, there is an $\alpha = 0.7$-cut of a fuzzy number.

Arithmetic operations with fuzzy numbers can be performed with the help of α-cuts. Several values of α are selected, therefore several α-cuts are constructed (these are crisp intervals) and interval arithmetic can be used on them. Sum of two fuzzy numbers is depicted on the left of figure 2. Product of two fuzzy numbers is depicted on the right of figure 2.

3 Material Models for Soils

Many material models are used for description of soils. The most frequent material models are the Mohr-Coulomb model and the Drucker-Prager model. Both of them are models based on theory of plasticity. The Mohr-Coulomb model assumes plastic yielding caused by frictional

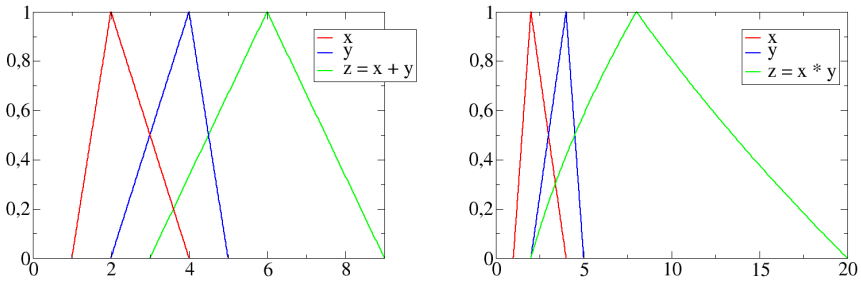


Figure 2: Sum and product of two fuzzy numbers.

sliding between material particles [4]. So called critical line is defined in the form

$$\tau = c - \sigma_N \tan\phi \tag{2}$$

where τ is the shear stress, c is the cohesion, σ_N is the normal stress and ϕ is the angle of internal friction. Yield condition can be written with the help of the principal stresses in the form

$$\sigma_{max} - \sigma_{min} = 2c\cos\phi - (\sigma_{min} + \sigma_{max})\sin\phi \tag{3}$$

and therefore the yield function has the form

$$f(\boldsymbol{\sigma}, c) = (\sigma_{max} - \sigma_{min}) + (\sigma_{min} + \sigma_{max})\sin\phi - 2c\cos\phi = 0 \tag{4}$$

where σ_{min} and σ_{max} are the minimum and maximum principal stresses, respectively.

The Drucker-Prager model is based on the yield function in the form

$$f(\boldsymbol{\sigma}, c) = \sqrt{J_2(\boldsymbol{\sigma})} + \eta p - \xi c \tag{5}$$

where J_2 is the invariant of the deviatoric stress, p is the hydrostatic stress and η and ξ are material parameters and c is the cohesion. If the Drucker-Prager model has to approximate the Mohr-Coulomb model, the following relationships have to be satisfied

$$\eta = \frac{6\sin\phi}{\sqrt{3}(3 \pm \sin\phi)} \qquad \xi = \frac{6\sin\psi}{\sqrt{3}(3 \pm \sin\psi)} \tag{6}$$

4 Material Parameters

Angle of internal friction, ϕ, and cohesion, c, are needed for the Mohr-Coulomb model. In various geotechnical literature, e.g. [5], the angle of internal friction and cohesion are summarized. Table 1 contains values of cohesive soils with water saturation less than 80%. Figure 3 shows possible shapes of fuzzy numbers describing the angle of internal friction of the cohesive soil D20. The limit values are given in handbook [5] and the form of the membership function can be obtained from an expert.

Table 1: Angle of internal friction and cohesion of cohesive soils with saturation less than 80%.

soil	ϕ	c (kPa)
D19	28°-33°	0-20
D20	22°-28°	10-40
D21	11°-17°	50-80

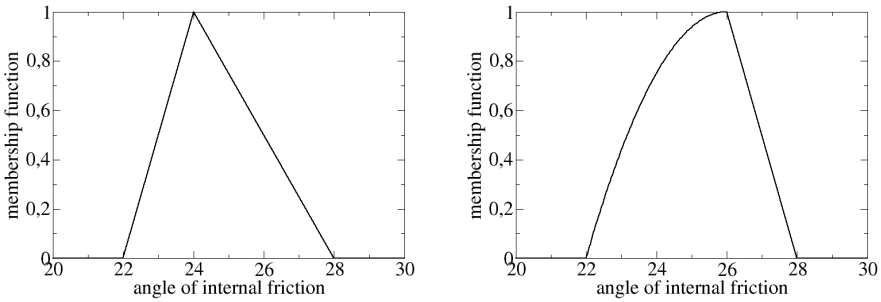


Figure 3: Possible fuzzy numbers describing the angle of internal friction.

5 Conclusion

The fuzzy set theory can be used for description of uncertainties if other theories are not applicable especially because of a lack of measurements or tests. Numerical simulations with fuzzy numbers are usually based on the α-cut method which will be described in our next contributions.

Acknowledgement: This outcome has been achieved with the financial support of the Grant Agency of the Czech Republic, project No. 19-11441S. The financial support is gratefully acknowledged.

References

[1] L.A. Zadeh: *Fuzzy Sets*. Information and Control 8, 1965, pp. 338–353.

[2] B. Möller, M. Beer: *Fuzzy Randomness. Uncertainty in Civil Engineering and Computational Mechanics*. Springer-Verlag, Berlin, 2004.

[3] V. Novák: *Fuzzy sets and their applications* (in Czech). Seminar of Mathematics SNTL 23, SNTL, Prague, 1990.

[4] E.A. de Souza Neto, D. Perić, D.R.J. Owen: *Computational Methods for Plasticity. Theory and Applications*. John Wiley and Sons Ltd., 2008.

[5] J. Hořejší, J. Šafka: *Civil Engineering Handbook* (in Czech). SNTL-Nakladatelství technické literatury, Prague, 1987.

3-dimensional wire-basket domain decomposition combined with multigrid

D. Lukáš

VŠB-Technical University of Ostrava, Faculty of Electrical Engineering and Computer Science

Nonoverlapping Schur complement domain decomposition methods for the elliptic PDEs can be viewed as a block Gaussian elimination of local Dirichlet problems. The arising Schur complement system on the skeleton is approximated by solution to local Dirichlet problems over pairs of neighbouring subdomains, the so-called edge or face problems in 2 and 3 dimensions, respectively, and by a global system on the wire-basket. In 2 dimensions the wire-basket problem, after a local-to-global transformation of the vertex basis functions, is nothing but the coarse discretization of the original PDE. The method is referred to as the vertex preconditioner. The method dates back to the pioneering work of Bramble, Pasciak, and Schatz in 1986 [1]. The number of iterations grows only poly-logarithmically with respect to H/h, where H and h are the coarse and fine finite element discretization steps, respectively. The method is also robust with respect to coefficient jumps that are aligned with subdomains.

In 3 dimensions the situation changes as the number of iterations of the vertex method grows super-linearly with H/h. An efficient remedy was proposed by B. Smith in 1991 [2]. The wire-basket is approximated by an auxiliary solver resulting in a coarse problem for the subdomain average values. The method again enjoys the poly-logarithmic complexity independently of coefficient jumps aligned with subdomains.

In this talk I describe a novel parallel implementation of the wire-basket method of Smith, which, to my best knowledge, has not been reported in literature yet. In the implementation the data as well as the computation is distributed over pairs of the neighbouring subdomains, rather than the subdomains themselves. The reason is that the local face problem, which is nowadays referred to as the deluxe scaling, is more expensive than the local subdomain problem. At the end I will discuss replacing local direct solvers with geometric multigrid.

This work continues in the direction of our previous work [3, 4].

References

[1] J.H. Bramble, J.E. Pasciak, A.H. Schatz: *The construction of preconditioners for elliptic problems by substructuring, I.* Mathematics of Computation **47**(175), 1986, pp. 103–134.

[2] B. Smith: *A domain decomposition algorithm for elliptic problems in three dimensions.* Numerische Mathematik **60**, 1991, pp. 219–234.

[3] D. Lukáš, J. Bouchala, P. Vodstrčil, L. Malý: *2-dimensional primal domain decomposition theory in detail.* Applications of Mathematics **60**(3), 2015, pp. 265–283.

[4] L. Foltyn, D. Lukáš, I. Peterek: *Domain decomposition methods coupled with parareal for the transient heat equation in 1 and 2 spatial dimensions.* Applications of Mathematics **65**(2), 2020, pp. 173–190.

Determination of initial stress tensor

J. Malík, A. Kolcun

Institute of Geonics of the Czech Academy of Sciences, Ostrava

The knowledge of initial stress tensor is very important when one evaluates the stability of underground openings like tunnels, compressed gas tanks or radioactive waste deposits. The knowledge of initial stress tensor enables to optimize the reinforcement of tunnels, choose the suitable shape of underground openings and their orientation in the rock environment. The mathematical modeling of stress fields in the vicinity of underground openings requires precise boundary conditions, which can be derived from initial stress tensor. Extensive literature is devoted to the determination of initial stress tensor. An overview of these methods can be found in the papers [1]-[3] that describe the development of these methods to the present. These methods are based on the installation of probes equipped with sensors that measure deformations occurring after removal rock, overcoring, in their vicinity. Due to the stress in the rock, the removal of a part of the rock causes deformation of the remaining rock, which is transfered to the sensors. The probes are relatively small, a few centimeters, and the accuracy of such measurements is not high. In this paper we present a new method, which is based on measuring the distances between pairs of selected points on the walls of the underground opening. When a part of the rock is excavated, the distance between these points changes and the magnitude of these changes depends on the initial stress tensor. A procedure which allows to determine the initial stress tensor from the measured distances is developed. A criterion showing how to select measuring points so that errors of measurement do not affect the results very much is presented. In this section we will describe the method of obtaining the initial stress tensor from measuring distances between suitably selected pairs of points. The solution to our problem will be based on the first boundary problem of the theory of elasticity and the approach used is shown in Fig. 1a-c.

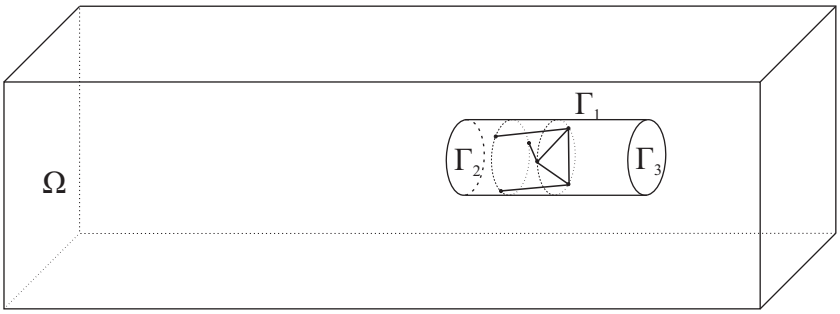


Figure 1a: First step of measuring process.

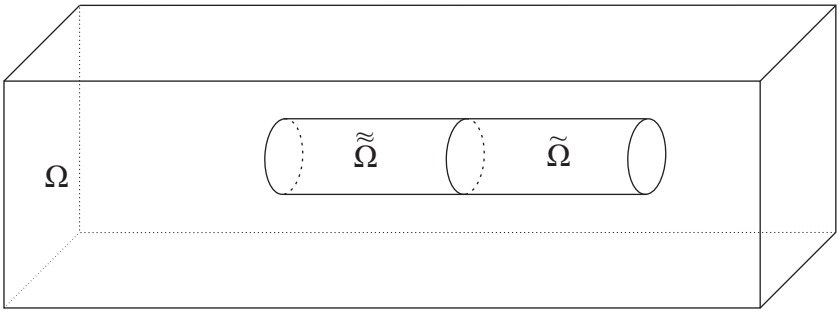


Figure 1b: Second step of measuring process.

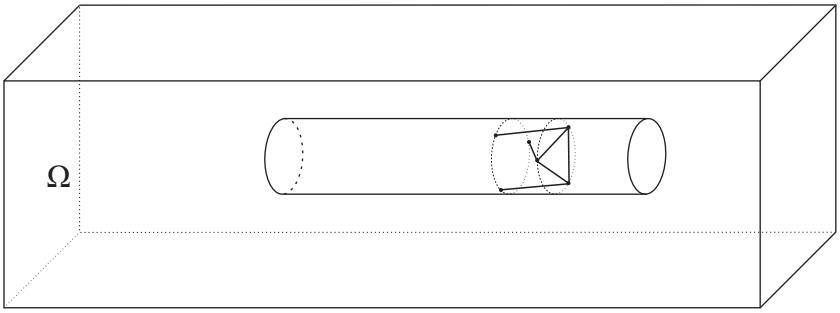


Figure 1c: Third step of measuring process.

Let us formulate our task as follows:

- Suppose that short steel bars are installed on the boundary of the tunnel.
- We measure the distances between the points in the situation shown in Fig. 2a.
- After removing part of the rock, (Fig. 2b) we re-measure these distances (Fig. 2c) and through mathematical modeling determine the initial stress tensor.

Our problem is the proper assembly of the matrix Z, which is the least square matrix connected to the selection of measuring points and their pairs so that the conditional number $\kappa(Z)$ is small as possible.

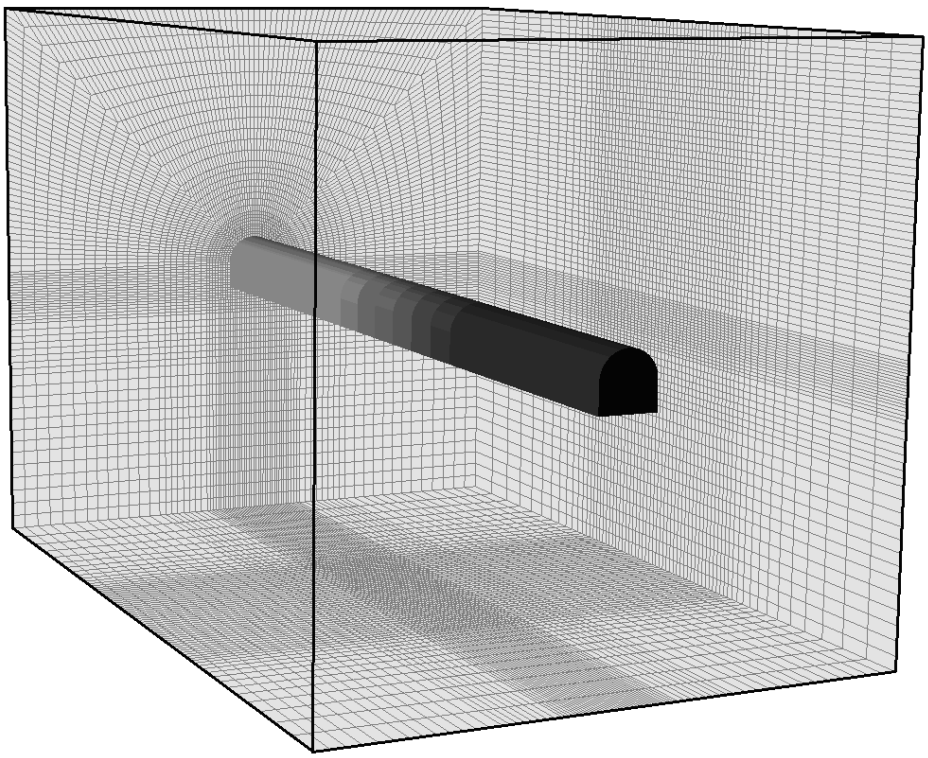
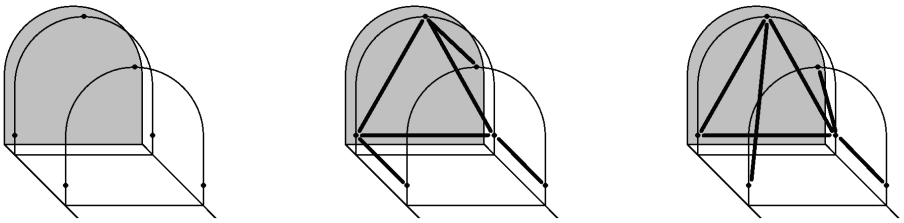


Figure 2: Finite element net of the numerical example.

The choice of measuring points and pairs of these points plays an important role. The correct choice can significantly reduce the condition number $\kappa(Z)$ and thus affect the reliability of the determination of the initial stress tensor. This fact is demonstrated by a simple numerical

example shown in Fig. 2. This figure shows a domain measuring $40 \times 40 \times 90\,m$ with a tunnel measuring $4 \times 4 \times 50\,m$. The darker part of the tunnel corresponds to the first phase and shows the part of the tunnel in which the measuring points were located. The lighter part of the tunnel corresponds to the second phase when the rock is extracted and the distances between the selected pairs of measuring points are re-measured. The tunnel is excavated in a homogeneous isotropic rock with Young's modulus $E = 65GPa$ and Poisson's ratio $\nu = 0.25$.



a) b) c)

Figure 3: Example of selection of measuring points.

The selection of measuring points and pairs of points at which distance measurements were made is shown in Fig. 3. The set X of six measuring points is shown in Fig. 3a). The first three measuring points are located $40cm$ behind the tunnel face and the second three points at a distance of $3m$ from the first three points. Two different sets of measuring points and their pairs are shown in Fig. 3b) and Fig. 3c). The performed numerical calculations lead to the following values of the conditional number are 58 and 5747. These values show that the choice of measuring points plays an essential role in the evaluation of measurements. Using mathematical modeling allows you to select suitable sets of measuring points and their pairs before installing the measuring points and measuring the distances between the selected pairs of these points.

Acknowledgement: This work was supported by Czech Science Foundation (GAČR) through project No. 19-11441S.

References

[1] J.A. Hudson, F.H. Cornet, R. Christiansson: *ISRM suggested methods for rock stress estimation - Part 1: strategy for rock stress estimation.* International Journal of Rock Mechanics and Mining Sciences 40, 2003, pp. 991–998.

[2] J. Sjőberg, R. Christiansen, J.A. Hudson: *ISRM suggested methods for rock stress estimation - Part 2: overcoring methods.* International Journal of Rock Mechanics and Mining Sciences 40, 2003, pp. 999–1010.

[3] B.C. Haimson, F.H. Cornet: *ISRM suggested methods for rock stress estimation - Part 3: hydraulic fracturing (HF) and/or hydraulic testing of pre-existing fractures (HTPF).* International Journal of Rock Mechanics and Mining Sciences 40, 2003, pp. 1011–1020.

Simple finite elements and multigrid for efficient mass-consistent wind downscaling in a coupled fire-atmosphere model

J. Mandel[1], *A. Farguell*[2], *A.K. Kochanski*[2], *D.V. Mallia*[3], *K. Hilburn*[4]

[1] University of Colorado Denver, Denver, CO
[2] San José State University, San José, CA
[3] University of Utah, Salt Lake City, UT
[4] Colorado State University, Fort Collins, CO

1 Introduction

In the coupled atmosphere-fire model WRF-SFIRE [6, 5], the weather model runs at 300–1km horizontal resolution, while the fire model runs at the resolution of 30m or finer. The wind has a fundamental effect on fire behavior and the topography details have a strong effect on the wind, but WRF does not see the topography on the fire grid scale. We want to downscale the wind from WRF to account for the fine-scale terrain. For this purpose, we fit the wind from WRF with a divergence-free flow over the detailed terrain. Such methods, called mass-consistent approximations, were originally proposed on regular grids [9, 10] for urban and complex terrain modeling, with terrain and surface features modeled by excluding entire grid cells from the domain. For fire applications, WindNinja [11] uses finite elements on a terrain-following grid. The resulting equations are generally solved by iterative methods such as SOR, which converge slowly, so use of GPUs is of interest [2]. A multigrid method with a terrain-following grid by a change of coordinates was proposed in [13].

The method proposed here is to be used in every time step of WRF-SFIRE in the place of interpolation to the fire model grid. Therefore, it needs to have the potential to (1) scale to hundreds or thousands of processors using WRF parallel infrastructure [12]; (2) scale to domains size at least 100km by 100km horizontally, with $3000 \times 3000 \times 15$ grid cells and more; (3) have reasonable memory requirements per grid point; (4) not add to the cost of the time step significantly when started from the solution in the previous time step; and, (5) adapt to the problem automatically, with minimum or no parameters to be set by the user.

2 Finite element formulation

Given vector field $\boldsymbol{u}_0$ on domain $\Omega \subset \mathbb{R}^d$, subset $\Gamma \subset \partial\Omega$, and $d \times d$ symmetric positive definite coefficient matrix $\boldsymbol{A} = \boldsymbol{A}(\boldsymbol{x})$, we want to find the closest divergence-free vector field $\boldsymbol{u}$ by solving the problem

$$\min_{\boldsymbol{u}} \frac{1}{2} \int_\Omega (\boldsymbol{u} - \boldsymbol{u}_0) \cdot \boldsymbol{A}(\boldsymbol{u} - \boldsymbol{u}_0)\, d\boldsymbol{x} \quad \text{subject to } \operatorname{div} \boldsymbol{u} = 0 \text{ in } \Omega \text{ and } \boldsymbol{u} \cdot \boldsymbol{n} = 0 \text{ on } \Gamma, \tag{1}$$

where Γ is the bottom of the domain (the surface), and $\boldsymbol{A}(\boldsymbol{x})$ is a 3×3 diagonal matrix with penalty constants a_1^2, a_2^2, a_3^2 on the diagonal. Enforcing the constraints in (1) by a Lagrange

multiplier λ, we obtain the solution $(\boldsymbol{u}, \lambda)$ as a stationary point of the Lagrangean

$$\mathcal{L}(\boldsymbol{u}, \lambda) = \frac{1}{2} \int_\Omega \boldsymbol{A}(\boldsymbol{u} - \boldsymbol{u}_0) \cdot (\boldsymbol{u} - \boldsymbol{u}_0)\, d\boldsymbol{x} + \int_\Omega \lambda \operatorname{div} \boldsymbol{u} d\boldsymbol{x} - \int_\Gamma \lambda \boldsymbol{n} \cdot \boldsymbol{u} d\boldsymbol{s}. \tag{2}$$

Eliminating $\boldsymbol{u}$ from the stationarity conditions $\partial\mathcal{L}(\boldsymbol{u}, \lambda)/\partial\lambda = 0$ and $\partial\mathcal{L}(\boldsymbol{u}, \lambda)/\partial\boldsymbol{u} = 0$ by

$$\boldsymbol{u} = \boldsymbol{u}_0 + \boldsymbol{A}^{-1} \operatorname{grad} \lambda \tag{3}$$

leads to the generalized Poisson equation for Lagrange multiplier λ,

$$-\operatorname{div} \boldsymbol{A}^{-1} \operatorname{grad} \lambda = \operatorname{div} \boldsymbol{u}_0 \text{ on } \Omega, \quad \lambda = 0 \text{ on } \partial\Omega \setminus \Gamma, \quad \boldsymbol{n} \cdot \boldsymbol{A}^{-1} \operatorname{grad} \lambda = -\boldsymbol{n} \cdot \boldsymbol{u}_0 \text{ on } \Gamma. \tag{4}$$

Multiplication of (4) by a test function μ, $\mu = 0$ on $\partial\Omega \setminus \Gamma$, and integration by parts yields the variational form to find λ such that $\lambda = 0$ on $\partial\Omega \setminus \Gamma$ and

$$\int_\Omega \boldsymbol{A}^{-1} \operatorname{grad} \lambda \cdot \operatorname{grad} \mu\, d\boldsymbol{x} = -\int_\Omega \operatorname{grad} \mu \cdot \boldsymbol{u}_0 d\boldsymbol{x} \tag{5}$$

for all μ such that $\mu = 0$ on $\partial\Omega \setminus \Gamma$. The solution is then recovered from (3). We proceed formally and avoid the language of functional spaces; the theory will be presented elsewhere.

The variational problem (5) is discretized by standard isoparametric 8-node hexahedral finite elements, e.g., [4]. The integral on the left-hand side of (5) is evaluated by tensor-product Gauss quadrature with two nodes in each dimension, while for the right-hand side, one-node quadrature at the center of the element is sufficient. The same code for the derivatives of a finite element function is used to evaluate grad λ in (3) at the center of each element.

The unknown λ is represented by its values at element vertices, and the wind vector is represented naturally by its values at element centers. No numerical differentiation of λ from its nodal values, computation of the divergence of the initial wind field $\boldsymbol{u}_0$, or explicit implementation of the boundary condition on grad λ in (4) is needed. These are all taken care of by the finite elements naturally.

3 Multigrid iterations

The finite element method for (5) results in a system of linear equations $Ku = f$. The values of the solution are defined on a grid, which we will call a *fine grid*. One cycle of the multigrid method consists of several iterations of a basic iterative method, such as Gauss-Seidel, called a *smoother*, followed by a *coarse-grid correction*. A prolongation matrix P is constructed to interpolate values from a coarse grid, in the simplest case consisting of every other node, to the fine grid. For a given approximate solution u after the smoothing, we seek an improved solution in the form $u + Pu_c$ variationally, by solving

$$P^\top K (u + Pu_c) = P^\top f \tag{6}$$

for u_c, and obtain the coarse-grid correction procedure as

$$\begin{aligned} f_c &= P^\top (f - Ku) && \text{form the coarse right-hand side} \\ K_c &= P^\top K P && \text{form the coarse stiffness matrix} \\ K_c u_c &= f_c && \text{solve the coarse-grid problem} \\ u &\leftarrow u + Pu_c && \text{insert the coarse-grid correction} \end{aligned} \tag{7}$$

The coarse grid correction is followed by several more smoothing steps, which completes the multigrid cycle.

In the simplest case, P is a linear interpolation and the coarse stiffness matrix K_c is the stiffness matrix for a coarse finite element discretization on a grid with each coarse-grid element taking the place of a $2 \times 2 \times 2$ agglomeration of fine-grid elements. That makes it possible to apply the same method to the coarse-grid problem (7) recursively. This process creates a hierarchy of coarser grids. Eventually, the coarsest grid problem is solved by a direct method, or one can just do some more iterations on it.

Multigrid methods gain their efficiency from the fact that simple iterative methods like Gauss-Seidel change the values of the solution at a node from differences of the values between this and neighboring nodes. When the error values at neighboring nodes become close, the error can be well approximated in the range of the prolongation P and the coarse-grid correction can find u_c such that $u + Pu_c$ is a much better approximation of the solution. For analysis of variational multigrid methods and further references, see [1, 7].

Multigrid methods are very efficient. For simple elliptic problems, such as the Poisson equation on a regular grid, convergence rates of about 0.1 (reduction of the error by a factor of 10) at the cost of 4 to 5 Gauss-Seidel sweeps on the finest grid are expected [3]. However, the convergence rates get worse on more realistic grids, and adaptations are needed. We choose as the smoother vertical sweeps of Gauss-Seidel from the bottom up to the top, with red-black ordering horizontally into 4 groups. For the base method, we use $2\times2\times2$ coarsening and construct P so that the vertices of every $2\times2\times2$ agglomeration of elements interpolate to the fine-grid nodes in the agglomeration, with the same weights as the trilinear interpolation on a regular grid. The interpolation is still trilinear on a stretched grid, but only approximately trilinear on a deformed terrain-following grid.

The base method works as expected as long as some grid layers are not tightly coupled. If they are, we mitigate the slower convergence by semicoarsening [8]: After smoothing, the error is smoother in the tightly coupled direction(s), which indicates that we should not coarsen the other direction(s). When the grid is stretched vertically away from the ground, the nodes are relatively closer and thus tightly coupled in the horizontal direction. Similarly, when the penalty coefficient a_3 in the vertical direction is larger than a_1 and a_2 in the horizontal directions, the neighboring nodes in the vertical direction are tightly coupled numerically. The algorithm to decide on coarsening we use is: Suppose that the penalty coefficients are $a_1 = a_2 = 1$ and $a_3 \geq 1$, and at the bottom of the grid, the grid spacing is $h_1 = h_2$ (horizontal) and h_3 (vertical). If $h_3/(h_1 a_3) > 1/3$, coarsen in the horizontal directions by 2, otherwise do not coarsen. Then, replace h_1 and h_2 by their new values, corsened (multiplied by 2) or not, and for every horizontal layer from the ground up, if $h_3/(h_1 a_3)] < 3$, coarsen about that layer vertically, otherwise do not coarsen. This algorithm retains the coarse grids as logically cartesian, which is important for computational efficiency and keeping the code simple, and it controls the convergence rate to remain up to about 0.28 with four smoothing steps per cycle.

4 Conclusion

We have presented a simple and efficient finite element formulation of mass-consistent approximation, and a multigrid iterative method with adaptive semicoarsening, which maintains the convergence of iteration over a range of grids and penalty coefficients. A prototype code is available at: `https://github.com/openwfm/wrf-fire-matlab/tree/femwind/femwind`

Acknowledgement: This work has been supported by NSF grant ICER-1664175 and NASA grant 80NSSC19K1091.

References

[1] R.E. Bank, T. Dupont: *An optimal order process for solving finite element equations.* Math. Comp. 36, 1981, pp. 35–51.

[2] B. Bozorgmehr, Z. Patterson, P. Willemsen, J.A. Gibbs, R. Stoll, J.J. Kim, E.R. Pardyjak: *A CUDA-based implementation of a fast response urban wind model.* 100th American Meteorological Society Annual Meeting, 2020. `https://ams.confex.com/ams/2020Annual/meetingapp.cgi/Paper/366583` accessed December 28, 2020.

[3] A. Brandt: *Multi-level adaptive solutions to boundary-value problems.* Math. Comp. 31, 1977, pp. 333–390.

[4] T.J.R. Hughes: *The finite element method.* Prentice Hall, Inc., Englewood Cliffs, NJ, 1987.

[5] J. Mandel, S. Amram, J.D. Beezley, G. Kelman, A.K. Kochanski, V.Y. Kondratenko, B.H. Lynn, B. Regev, M. Vejmelka: *Recent advances and applications of WRF-SFIRE.* Natural Hazards and Earth System Sciences 14, 2014, pp. 2829–2845.

[6] J. Mandel, J.D. Beezley, A.K. Kochanski: *Coupled atmosphere-wildland fire modeling with WRF 3.3 and SFIRE 2011.* Geoscientific Model Development 4, 2011, pp. 591–610.

[7] J. Mandel, S. McCormick, R. Bank: *Variational multigrid theory*, in Multigrid methods, Vol. 3 of Frontiers Appl. Math., SIAM, Philadelphia, PA, 1987, pp. 131–177.

[8] E. Morano, D.J. Mavriplis, V. Venkatakrishnan: *Coarsening strategies for unstructured multigrid techniques with application to anisotropic problems.* SIAM J. Sci. Comput. 20, 1998, pp. 393–415.

[9] C.A. Sherman: *A mass-consistent model for wind fields over complex terrain.* Journal of Applied Meteorology 17, 1978, pp. 312–319.

[10] B. Singh, B.S. Hansen, M.J. Brown, E.R. Pardyjak: *Evaluation of the QUIC-URB fast response urban wind model for a cubical building array and wide building street canyon.* Environmental Fluid Mechanics 8, 2008, pp. 281–312.

[11] N.S. Wagenbrenner, J.M. Forthofer, B.K. Lamb, K.S. Shannon, B.W. Butler: *Downscaling surface wind predictions from numerical weather prediction models in complex terrain with WindNinja.* Atmospheric Chemistry and Physics 16, 2016, pp. 5229–5241.

[12] W. Wang, C. Bruyère, M. Duda, J. Dudhia, D. Gill, M. Kavulich, K. Werner, M. Chen, H.C. Lin, J. Michalakes, S. Rizvi, X. Zhang, J. Berner, D. Munoz-Esparza, B. Reen, S. Ha, K. Fossell, J.D. Beezley, J.L. Coen, J. Mandel: *ARW version 4 modeling system user's guide.* Mesoscale & Miscroscale Meteorology Division, National Center for Atmospheric Research, January 2019.

[13] Y. Wang, C. Williamson, D. Garvey, S. Chang, J. Cogan: *Application of a multigrid method to a mass-consistent diagnostic wind model.* Journal of Applied Meteorology 44, 2005, pp. 1078–1089.

A correct and efficient algorithm for impacts of bodies

I. Němec[1], *H. Štekbauer*[1], *R. Lang*[1], *M. Zeiner*[2], *D. Burkart*[2]

[1] Brno University of Technology, Faculty of Civil Engineering, Brno
[2] FEM consulting, s.r.o., Brno

1 Introduction

Modelling of contact is still one of the most difficult aspects of nonlinear analysis. From the mechanical point of view, contact is the interaction between bodies that touch and exchange loads and energy. In the finite element method the number of different approaches were developed. The node-to-segment (NTS) algorithm is a multipurpose discretization technique [1]. The algorithm is widely used due to its simplicity, clear physical meaning and flexibility [3]. The contact is defined between a master segment and a slave node (Figure 1). The slave node C lies on the unit normal $\mathbf{n}$ of the master segment AB in the distance g_N, the normal lies in the distance ξ from the node 1, ξ can take values in the interval $\langle 0; 1\rangle$ [2].

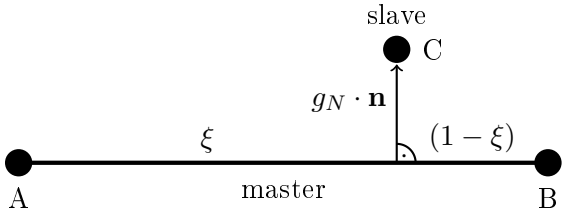


Figure 1: Geometry of the node-to-segment contact. [2]

For two bodies in the contact a contact element is created between the slave node and the master segment, with the weak formulation

$$\int_{\Omega_{1,2}} \sigma \colon \delta\varepsilon d\Omega - \int_{\Omega_{1,2}} \mathbf{f}_v \cdot \delta u d\Omega - \int_{\Gamma^N_{1,2}} f_0 \cdot \delta u d\Gamma + \Pi_c = 0 \tag{1}$$

where is σ - stress, ε - strain, f_v - volume forces, f_0 - surface forces, u - deformations, Ω_1 and Ω_2 - body regions, $\Gamma_{1,2}$ - body boundaries, extended by a contact contribution Π_c which differs based on the used method [2].

There are basically two methods that are used for Π_c calculation, penalty method and Lagrange multipliers. The main drawback of the penalty method is choosing the penalty weight. The penalty method is not exact, because constraint violation is dependent on chosen penalty weight. It can be shown that constraint violation is proportional to 1/P for P large enough, but with larger P solution becomes more unstable and forces calculated in contact element may be much higher than the correct ones. The method of Lagrange multipliers is exact, but has a disadvantage in need to expand the original system of equations that is not positive definite. [4]

The paper introduces a correct and efficient algorithm for calculation of impacts of bodies by explicit method, fulfilling the correct conservation of both, momentum and energy. The algorithm is described in 2D formulation in this paper, but the principles introduced there holds also for the formulation in 3D space. The algorithm has been implemented in a computer program and its correctness and efficiency has been fully proved.

2 Calculation of impacts of bodies

Calculation of impacts of bodies consists of solution of impacts of particular nodes of one body with a surface of elements of other bodies. The principle of the analysis of impacts of bodies introduced in this paper assumes that the surfaces of elements of the bodies are plane. The mathematical description described in this paper is formulated for 2D problem, which implies that the surfaces are represented by line segments. Particular bodies are arbitrarily moving (and rotating) in the space. For each node of a body it is needed to investigate if it had impacted on a surface of an element of another body in the current time step. An impact which occurs earlier is analyzed first.

2.1 Calculation of magnitude of the contact force

The change of the total potential energy Π is the sum of all its three components.

$$\Pi = \Pi_k + \Pi_\sigma + \Pi_p \tag{2}$$

All of them are functions of one unknown variable which is the magnitude of the contact force $\mathbf{f}_C$. Determination of its direction has been described above. The direction of the force $\mathbf{f}_C$ can be defined as the base of the vector $\mathbf{f}_C$, $\mathbf{e}_C = \frac{\mathbf{f}_C}{\|\mathbf{f}_C\|}$. Thus for $\mathbf{f}_C$ we can write

$$\mathbf{f}_C = \mathbf{e}_C \|\mathbf{f}_C\| \tag{3}$$

The magnitude of this force $\|\mathbf{f}_C\|$ can be determined from the law of conservation of energy, so we can write

$$\Pi'(\|\mathbf{f}_C\|) - \Pi(\|\mathbf{f}_C\|) = \Delta\Pi(\|\mathbf{f}_C\|) = \Delta\Pi_k(\|\mathbf{f}_C\|) + \Delta\Pi_\sigma(\|\mathbf{f}_C\|) + \Delta\Pi_p(\|\mathbf{f}_C\|) = 0 \tag{4}$$

where $\Pi(\|\mathbf{f}_C\|)$ and $\Pi'(\|\mathbf{f}_C\|)$ is the total potential energy of the structure at the beginning and at the end of the fictive time step. The formulas for calculation of the kinetic energy and the potential energy of position contains only linear and quadratic members with unknown magnitude of the impact force $\|\mathbf{f}_C\|$. Only the calculation of elastic potential energy needs calculation of the square root of the member with the unknown variable. The issue is the calculation of the member length at the end of the fictive time step. If we want to obtain the quadratic equation for the unknown variable and thus to obtain the magnitude of the impact force $\|\mathbf{f}_C\|$in the closed form, we need to calculate the change of the lenght of the member in linearized form based on adding linear projections of the nodal displacement components to the exact calculation of the end of the member in the beginning of the fictive time step. Then the calculation of $\|\mathbf{f}_C\|$ would lead to solution of a quadratic equation without an absolute member which would have been vanished by detraction of the energy at the beginning of the time step. Physical sense has only the positive root of the equation. It should be substitute as a force at the point C and then determine with it also forces of the points A and B. Regarding that the following relations holds

$$\begin{aligned} \Delta\mathbf{v}_A &= \frac{\mathbf{f}_A}{m_A} dt \\ \Delta\mathbf{v}_B &= \frac{\mathbf{f}_B}{m_B} dt \end{aligned} \tag{5}$$

by substitution for $\Delta\mathbf{v}_A$ and $\Delta\mathbf{v}_B$ from the equations

$$\begin{aligned}\Delta\mathbf{v}_A =&\Delta\mathbf{v}_Q\frac{m_A+m_B}{m_A}\left(1-\xi\right)=-\frac{\mathbf{f}_C}{m_A}\left(1-\xi\right)dt\\ \Delta\mathbf{v}_B =&\Delta\mathbf{v}_Q\frac{m_A+m_B}{m_B}\xi=-\frac{\mathbf{f}_C}{m_B}\xi dt\end{aligned}\tag{6}$$

we obtain relations for the forces $\mathbf{f}_A$ and $\mathbf{f}_B$

$$\begin{aligned}\mathbf{f}_A =&\Delta\mathbf{v}_A\frac{m_A}{dt}=-\mathbf{f}_C\left(1-\xi\right)\\ \mathbf{f}_B =&\Delta\mathbf{v}_B\frac{m_B}{dt}=-\mathbf{f}_C\xi\end{aligned}\tag{7}$$

In case that the linearization of the equation 4 was not performed, then it is possible to calculate the unknown variable $\|\mathbf{f}_C\|$ by an iterative process using variation of the Newton's method. The Newton's method of solution of a nonlinear algebraic equation $f\left(x\right)=0$ supposes knowledge of the fist derivative of the function $f'\left(x\right)$.

The unknown variable x can then be calculated by the iterative formula

$$x_{i+1}=x_i-\frac{f_i}{f'_i}\tag{8}$$

Because we are not able to calculate the first derivative of the function$\Delta\Pi$ analytically, it is then determined only numerically from the last two iterations as follows. The iterative algorithm then reads:

$$\|\mathbf{f}_C\|_{i+1}=\|\mathbf{f}_C\|_i-\frac{\Delta\Pi_i\left(\|\mathbf{f}_C\|_i-\|\mathbf{f}_C\|_{i-1}\right)}{\Delta\Pi_i-\Delta\Pi_{i-1}}\tag{9}$$

2.2 Satisfying the conservation laws

The issue deals with the laws of conservation of mass, energy and momentum.

a) Conservation of mass: There must be satisfied these following simple equations:

$$\begin{aligned}dm =&dm_0\\ \rho dV =&\rho_0 dV_0\Rightarrow\rho=\rho_0\frac{dV_0}{dV}\end{aligned}\tag{10}$$

dm_0, $d\rho_0$ a dV_0 is mass, density and volume of a mass element on the initial configuration and dm, $d\rho$ a dV is mass, density and volume of the same mass element on the current configuration (deformed) configuration. Thus, when e.g. decreasing the element volume, the density of the element must be proportionally increased to satisfy the same element mass. This law is automatically fulfilled regardless the presented algorithm and is not influenced by this algorithm.

b) Conservation of energy: The law of conservation of energy is explicitly satisfied in calculation of the impact of bodies using the presented algorithm, because the magnitude of the opposite impact forces $\|\mathbf{f}_C\|$ and $\|\mathbf{f}_Q\|$ acting on the points C and Q of the impacted bodies are calculated on the base of just this law (equation 4).

c) Conservation of momentum: This law is satisfied implicitly, because the unknown contact force is introduced in the algorithm by the opposite forces acting on the impacted points in the sense of the 3rd Newton's law, so the global momentum will not be influenced.

3 Conclusion

The paper has introduced a theoretical correct and efficient algorithm for impact of bodies. The presented algorithm was implemented in a computer program and correctness of the presented algorithm was proved by number of numerical examples where all the energy components were monitored. It was shown that during each impact the law of conservation of energy is perfectly satisfied. The conservation of momentum is fulfilled implicitly. The numerical tests also have shown the stability of the presented algorithm. Although that the algorithm was presented and implemented in 2D, the principles of the algorithm hold also for 3D. This will be developed and presented in the next phase of the research. In the computer tests the presented algorithm was compared with the penalty method which is widely used in another computer programs and the superiority of the algorithm presented in this paper over the penalty method was shown.

References

[1] T.R.J. Hughes, R.L. Taylor, W. Kanoknukulchai: *A finite element method for large displacement contact and impact problems.* In: K. Bathe, J. Oden, W. Wunderlich, E. Wilson (eds.): Formulations and Computational Algorithms in FE Analysis, MIT Press, 1977, pp. 468–495.

[2] H. Štekbauer, I. Němec: *Modeling of Welded Connections Using Lagrange Multipliers.* In: AIP Conference Proceedings: Vol. 2293, No. 1, 2020, pp. 340013, DOI: 10.1063/5.0031396.

[3] G. Zavarise, L. De Lorenzis: *A modified node-to-segment algorithm passing the contact patch test.* In: International Journal for Numerical Methods in Engineering: Vol. 79, 2009, pp. 379–416.

[4] H.Štekbauer: *The pulley element.* In: Transactions of the VSB - Technical University of Ostrava: Civil Engineering Series, Vol. 16, Issue 2, 2016, pp. 161–164.

Preconditioning the stage equations of implicit Runge Kutta methods

M. Outrata, M.J. Gander

University of Geneva

When using implicit Runge-Kutta methods for solving parabolic PDEs, solving the stage equations is often the computational bottleneck, because the dimension of the stage equations is related to the spatial discretization and can thus become very large. The solution of the stage equations hence often requires the use of iterative solvers, whose convergence can be less than satisfactory. Using spectral analysis, we study the properties of two recently introduced preconditioners for the stage equations, and their dependence on the associated Butcher tableau of the Runge-Kutta method. We then try to optimize the Butcher tableau for the performance of the entire solution process, rather than only the order of convergence of the Runge-Kutta method. To do so requires to carefully balance the numerical stability of the Runge-Kutta method, its order of convergence, and also the convergence of the iterative solver for the stage equations. We illustrate our result on a simple test problem and then outline possible generalizations.

Mathematics and Optimal control theory meet Pharmacy: Towards application of special techniques in modeling, control and optimization of biochemical networks

Š. Papáček[1], *C. Matonoha*[2], *J. Duintjer Tebbens*[2,3]

[1] Institute of Information Theory and Automation of the Czech Academy of Sciences, Prague
[2] Institute of Computer Science of the Czech Academy of Sciences, Prague
[3] Charles University, Faculty of Pharmacy in Hradec Králové

1 Introduction

Similarly to other scientific domains, the expenses related to *in silico* modeling in pharmacology need not be extensively apologized. *Vis à vis* both *in vitro* and *in vivo* experiments, physiologically-based pharmacokinetic (PBPK) and pharmacodynamic models represent an important tool for the assessment of drug safety before its approval, as well as a viable option in designing dosing regimens.

In this contribution, we present some special techniques related to the mathematical modeling, control and optimization of biochemical networks. We continue in direction of papers devoted to mathematical models describing the drug-induced enzyme production networks, see [3] and references within there. Again, we deal with the inverse problem of model parameter estimation. Furthermore, being aware of the expected growing complexity of PBPK models, here, we mainly focus on the problem of model reduction [6] and related techniques, e.g. Quasi-Steady-State Assumption (QSSA) [5], delayed QSSA (D-QSSA) [7], Singular Perturbation Method [2]. Newly, we consider optimal control problems, e.g. output regulation *via* a periodic drug intake.

2 Case study: Enzymatic process with external dosing (leading to Michaelis-Menten kinetics)

Let us consider perhaps the most studied example of the enzyme-substrate transport-reaction (T-R) network, see e.g. [5]. The model for the action of an enzyme E on another chemical (so-called substrate S transported from a cell exterior to interior) is described in Tab. 1.

Table 1: Description of transport and reaction systems (enzyme-substrate reaction network) leading to Michaelis-Menten kinetics when QSSA is employed.

Description of related T-R process	**Chem. notation**	**Model parameters**
R_1: Substrate (drug) dosing (model input)	$\emptyset \to S_{ext}$	$a_{dose}(t)$
R_2: Drug enters the cell e.g. by permeation	$S_{ext} \to S_{int}$	$k_{up} \equiv k_3$
R_3: Enzyme binds to drug, formation of C	$S_{int} + E \rightleftharpoons C$ (reaction R_3 is reversible)	$k_{assoc} \equiv k_1,\ k_{dis} \equiv k_2$
R_4: Complex breaks down (P formation)	$C \to P + E$	$k_{cat} \equiv k_4$
R_5: Product P removal (or degradation)	$P \to \emptyset$	$a_{deg}(t)$

The state variables of substance concentrations are collected in a 5-dimensional vector x as follows

$$x(t) = \begin{pmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \\ x_4(t) \\ x_5(t) \end{pmatrix} \equiv \begin{pmatrix} \mathsf{E}(t) \\ \mathsf{C}(t) \\ \mathsf{S_{ext}}(t) \\ \mathsf{S_{int}}(t) \\ \mathsf{P(t)} \end{pmatrix}.$$

Then the system of ODEs describing T-R process can be written as

$$\frac{\mathrm{d}x(t)}{\mathrm{d}t} = \begin{pmatrix} x_1'(t) \\ x_2'(t) \\ x_3'(t) \\ x_4'(t) \\ x_5'(t) \end{pmatrix} = Dx(t) + z(t), \tag{1}$$

with the constant matrix

$$D = \begin{pmatrix} 0 & k_2 + k_4 & 0 & 0 & 0 \\ 0 & -(k_2 + k_4) & 0 & 0 & 0 \\ 0 & 0 & -k_3 & 0 & 0 \\ 0 & k_2 & k_3 & 0 & 0 \\ 0 & k_4 & 0 & 0 & 0 \end{pmatrix} \tag{2}$$

representing the linear part of the system, and the vector

$$z(t) = \begin{pmatrix} -k_1 \cdot x_1(t) \cdot x_4(t) \\ k_1 \cdot x_1(t) \cdot x_4(t) \\ a_{dose} \\ -k_1 \cdot x_1(t) \cdot x_4(t) \\ -a_{deg} \end{pmatrix}$$

representing the nonlinear (quadratic) and constant parts. The initial conditions (given the normalization with the initial enzyme concentration E_0) are

$$x(0) = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \end{pmatrix}^T. \tag{3}$$

An appealing modification of the above relations using two sets of state variables of substance concentrations was introduced in [1] and further developed by prof. Ivo Marek (R.I.P.) in [4].

$$x^1(t) = \begin{pmatrix} x_1(t) \\ x_2(t) \end{pmatrix}, \quad x^2(t) = \begin{pmatrix} x_2(t) \\ x_3(t) \\ x_4(t) \\ x_5(t) \end{pmatrix}.$$

Then the linear system of differential equations for modified state variable vector $\tilde{x}(t)$ is

$$\frac{\mathrm{d}\tilde{x}(t)}{\mathrm{d}t} = \begin{pmatrix} x_1'(t) \\ x_2'(t) \\ x_2'(t) \\ x_3'(t) \\ x_4'(t) \\ x_5'(t) \end{pmatrix} = A\tilde{x}(t) + \tilde{z}(t), \tag{4}$$

with the block diagonal system matrix of special form (M-matrix)

$$A = \begin{pmatrix} -k_1 \cdot x_4 & k_2 + k_4 & 0 & 0 & 0 & 0 \\ k_1 \cdot x_4 & -(k_2 + k_4) & 0 & 0 & 0 & 0 \\ 0 & 0 & -(k_2 + k_4) & 0 & k_1 \cdot x_1 & 0 \\ 0 & 0 & 0 & -k_3 & 0 & 0 \\ 0 & 0 & k_2 & k_3 & -k_1 \cdot x_1 & 0 \\ 0 & 0 & k_4 & 0 & 0 & 0 \end{pmatrix} \tag{5}$$

representing (in some sense) the linear part of the system, and the vector

$$\tilde{z}(t) = \begin{pmatrix} 0 & 0 & 0 & a_{dose} & 0 & -a_{deg} \end{pmatrix}^T$$

representing the nonlinear (constant) parts (including the initial/boundary conditions for state variable x_4 in a_{dose} term).

3 Expected results – Future prospects

Taking our benchmark system, either in form (1) or (4), we plan to perform numerical simulations for two types of problems: (i) IVP (initial value problem) when the initial dose a_{dose} should determine the speed of product formation (*Michaelis-Menten* like kinetics is being expected); (ii) BVP (boundary value problem) when the initial dose a_{dose} is a periodic function. For the latter case, we aim to formulate and solve an optimization problem.

Moreover, we aim to study some numerical issues related to model reduction techniques. Namely, we shall compare the numerical results obtained from the full (non-reduced) problem with the results obtained using two different model reduction methods. Both IVP and BVP shall be performed on the following models:

1. Non-reduced model, i.e. (1) or (4),
2. Reduced model A (old QSSA/SPM method),
3. Reduced model B (new delayed QSSA method).

Actually, we are looking for a similar result as in [7], where the delayed QSSA (D-QSSA) method was firstly presented, justified and employed on more complex systems.

As it was stated in [7]: *Reduction methods can produce a significant simplification of complex Systems Biology models whilst retaining a high degree of predictive accuracy. The essential first step for both the standard QSSA and the D-QSSA is the identification of the fast variables. However, in some systems none of the variables can be considered as fast, while a suitable combination can.*

Here, for the enzyme-substrate reaction network, the slow-fast variables separation is well known and the technique of the D-QSSA can be applied successfully.

Once having determined the numerical issues concerning the appropriate method for numerical integration, the efficient and reliable simulation of state variables would be possible and eventually the formulation and solving of an **optimization** problem should crown our efforts.

Last comment on D-QSSA method: The authors of [7] point out that in most biochemical systems the delay in the D-QSSA depends on the rate constants of the chemical reactions involved. Thus,

the technique of the D-QSSA, applied for complex biochemical processes, should be suitably 'tuned'.[1]

Thus, we plan, in the future work, to apply the delayed QSSA method to the model describing the action of pregnane X receptor (PXR) causing the xenobiotic (drug) metabolizing enzyme induction, see [3] and references within there. The reason is that the transcriptional delay on the rates of the elementary chemical reactions is inherently present there, and moreover, some related work has been already done.

Acknowledgement: This work was supported by the grant No. GA19-05872S of the Czech Science Foundation and by the long-term strategic development financing of the Institute of Computer Science (RVO:67985807).

Moreover, we thank Prof. Ivo Marek, R.I.P. for his inspiring and original work in the field of Cell Biology. He put not only the basis of a novel approach (making nonlinear ODEs from linear ones) but he also traced new directions in the spirit of Control Engineering.

References

[1] E. Bohl, I. Marek: *Existence and Uniqueness Results for Nonlinear Cooperative Systems.* In: I. Gohberg, H. Langer (eds.): Linear Operators and Matrices. Operator Theory: Advances and Applications, Vol. 130, Birkhäuser, Basel, 2002.

[2] E. Bohl, I. Marek: *Input-output systems in biology and chemistry and a class of mathematical models describing them.* Appl. Math. 50, 2005, pp. 219–245.

[3] J. Duintjer Tebbens, C. Matonoha, A. Matthios, Š. Papáček: *On parameter estimation in an in vitro compartmental model for drug-induced enzyme production in pharmacotherapy.* Applications of Mathematics 64, 2019, pp. 253–277.

[4] I. Marek: *On a Class of Stochastic Models of Cell Biology: Periodicity and Controllability.* In: R. Bru, S. Romero-Vivó (eds.): Positive Systems. Lecture Notes in Control and Information Sciences, Vol. 389, Springer, Berlin, Heidelberg, 2009.

[5] L.A. Segel, M. Slemrod: *The Quasi-Steady-State Assumption: A Case Study in Perturbation.* SIAM Review, Vol. 31, No. 3, 1989, pp. 446–477.

[6] T.J. Snowden, P.H. van der Graaf, M.J. Tindall: *Methods of Model Reduction for Large-Scale Biological Systems: A Survey of Current Methods and Trends.* Bull Math Biol 79, 2017, pp. 1449–1486.

[7] T. Vejchodský, R. Erban, P.K. Maini: *Reduction of chemical systems by delayed quasi-steady state assumptions.* 2014. arXiv preprint arXiv:1406.4424.

[1] What is presenting another *sui generis* optimization problem.

Inverse problem for nonlinear Gao beam and elastic foundation

J. Radová, J. Machalová

Palacký University Olomouc, Faculty of Science

1 Introduction of identification problem

Identification problem is a framework of mathematical problems dealing with the identification of unknown coefficients of a given differential equation. In general, the identification of parameters consists of using experimentally measured data such that the differential equation's coefficients can be determined. In practice, there exist many interesting problems in which the equation's coefficients are not exactly known. The aim of this contribution is to find unknown coefficients of the nonlinear Gao beam model and an obstacle situated in some distance under the beam. The obstacle is considered as an elastic foundation governed by the Winkler one-parametric model.

2 Gao beam equation

The Gao beam model, which was firstly introduced in [1], is given by the fourth-order differential equation

$$E\,I\,w^{IV} - E\,\alpha\,(w')^2\,w'' + P\,\mu\,w'' = f, \qquad \text{in } (0,L), \tag{1}$$

where

$$\alpha = 3\,t\,b\,(1-\nu^2), \qquad \mu = (1-\nu^2)(1+\nu), \qquad f = (1-\nu^2)\,q,$$

w is an unknown deflection, E is Young's modulus, ν is the Poisson's ratio. The area moment of inertia $I = \frac{2}{3}\,t^3\,b$ is constant with $2t$ as a thickness and b as a width of the beam. The symbol L stands for the length of the beam. The distributed transverse load is denoted by q and P represents the constant axial force acting at the end point $x = L$. We distinguish two types of axial force cases, the case with an axial force causing compression $P > 0$ and an axial force causing tension $P < 0$.

In the recent paper [6] small correction of the Gao beam model was presented. The correction modifies the constant μ. Instead of the constant $\mu = (1-\nu^2)(1+\nu)$ is proposed the coefficient $\overline{\mu} = (1-\nu^2)$. With respect to this fact we will consider the modified Gao beam equation, i.e.

$$E\,I\,w^{IV} - E\,\alpha\,(w')^2\,w'' + P\,\overline{\mu}\,w'' = f, \qquad \text{in } (0,L). \tag{2}$$

In this contribution an identification problem for a contact problem for the Gao beam and an elastic deformable foundation governed by the Winkler one-parametric model with a foundation modulus $k_F > 0$ is studied. See in [2], [4] and [5] for detailed information about solution of the contact problem for the Gao beam model. It is assumed that the beam is situated above the foundation, i.e. there is a gap g between the foundation and the beam, see Fig. 1. Thus, the equation (2) is modified in the following way

$$E\,I\,w^{IV} - E\,\alpha\,(w')^2\,w'' + P\,\overline{\mu}\,w'' = f + T(w), \qquad \text{in } (0,L), \tag{3}$$

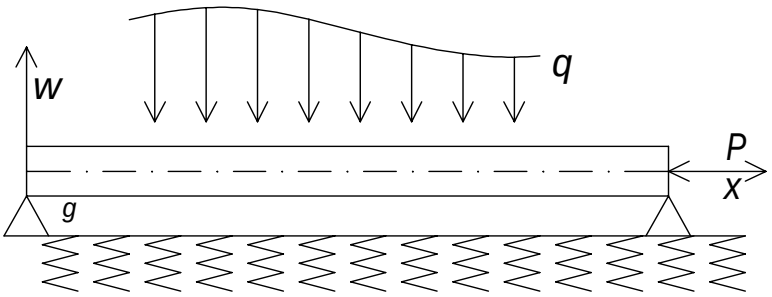


Figure 1: Beam situated above elastic foundation

where

$$T(w) = c_F(g - w)^+$$

with $c_F = (1-\nu^2)k_F$ represents contact forces between the beam and the foundation. The term $(g-w)^+$ is defined as follows

$$(g-w)^+ = \max\{0, g-w\}.$$

The variational formulation of the contact problem (3) reads as

$$\begin{cases} \text{Find } w \in V \text{ such that} \\ a(w,v) + \pi(w,v) - \kappa(w,v) = \mathcal{L}(v) \qquad \forall v \in V, \end{cases} \tag{4}$$

where V is the space of displacements with respect to the boundary conditions and where

$$a(w,v) = \int_0^L EIw''v''\mathrm{d}x - P\,\overline{\mu}\int_0^L w'v'\mathrm{d}x, \qquad \pi(w,v) = \frac{1}{3}\int_0^L E\,\alpha\,(w')^3\,v'\mathrm{d}x, \tag{5}$$

$$\kappa(w,v) = \int_0^L c_F(g-w)^+ v\,\mathrm{d}x, \qquad \mathcal{L}(v) = \int_0^L fv\,\mathrm{d}x, \tag{6}$$

for more information see e.g. [5].

3 Identification of parameter

The main idea of the parameter identification problem is to determine coefficients of the static Gao beam equation by using an *optimal control approach*, see e.g. [8]. The similar problem including the identification problem for the Gao beam model was studied in [7], where the identification problem for deflection of Gao beam model was analyzed. The identification problem is formulated as the minimization of a cost functional which depends on a solution of the state problem. In this case the state problem is represented by the Gao beam equation and the elastic deformable foundation. The cost functional $\mathcal{J}$ is defined as follows

$$\mathcal{J} : H_0^2((0,L)) \longrightarrow \mathbb{R}, \qquad \mathcal{J}(w(c)) = \frac{1}{2}\|w(c) - z\|^2, \tag{7}$$

where $\|\cdot\|$ is L^2-norm, $z \in L^2((0,L))$ is given function, $H_0^2((0,L))$ is Sobolev space and U_{ad} is so called admissible set defined as

$$\begin{aligned} U_{ad} := \{c \in L^\infty((0,L)) \times L^\infty((0,L)) \times L^\infty((0,L)) :\ & 0 < c_{\min} \le c \le c_{\max} < \infty \text{ in } (0,L), \\ & c|_{K_i} \in P_0(K_i) \times P_0(K_i) \times P_0(K_i)\, i = 1, \dots, r\}, \end{aligned}$$

where $c = (E, \nu, k_F)$ and $c_{\min}, c_{\max}$ are given vectors. We suppose that the interval $(0, L)$ is decomposed into mutually disjoint open intervals K_i, $i = 1, \ldots, r$, i.e. $K_i \cap K_j = \emptyset, \forall i \neq j$, and $\langle 0, L\rangle = \bigcup_{i=1}^{r} \overline{K}_i$. Further $P_0(K_i)$ is the set of constant functions on the subintervals K_i. It is easy to see that U_{ad} is the closed, convex subset of triplets of piecewise constant functions on the partition of $(0, L)$. The final identification problem is defined as follows

$$\begin{cases} \text{Find unknown parameter } c^* \in U_{ad} \text{ such that} \\ J(w(c^*)) \;=\; \min\limits_{c \in U_{ad}} J(w(c)), \\ \text{where } w(c) \text{ solves the state problem } (4). \end{cases} \tag{$\mathbb{P}$}$$

If $c^* = (E^*, \nu^*, k_F^*)$ is the solution to ($\mathbb{P}$) and $w^* := w(E^*, \nu^*, k_F^*)$ solves the state problem (4) then the pair $((E^*, \nu^*, k_F^*), w^*)$ is called *an optimal pair* of the problem ($\mathbb{P}$).

Numerical realization of the identification problem ($\mathbb{P}$) is based on using finite element method. Discretization is composed of two parts. The first part is the discretization of the state problem. The second part concerns the discretization of the cost functional (7) that it is described in [7]. Finally, the discretization of the identification problem leads to a nonlinear programming problem as follows

$$\begin{cases} \text{Find vector } \; \boldsymbol{c}^* \in U_{ad} \; \text{ such that} \\ \boldsymbol{J}(\boldsymbol{c}^*) = \min\limits_{\boldsymbol{c} \in U_{ad}} \boldsymbol{J}(\boldsymbol{c}), \\ \text{where } \; \boldsymbol{w}(\boldsymbol{c}) \; \text{ solves the discrete state problem} \end{cases}$$

and where the discrete cost functional is given by

$$\boldsymbol{J}(\boldsymbol{c}) = \frac{1}{2} \|\boldsymbol{S}\boldsymbol{w}(\boldsymbol{c}) - \boldsymbol{z}\|^2,$$

where $\boldsymbol{S}$ is a matrix representing the restriction mapping, $\boldsymbol{w}(\boldsymbol{c})$ is a solution of the discrete state problem, $\boldsymbol{z}$ denotes the vector of given measured values. For more information we refer to [7].

The minimization process is performed by a gradient method and it is based on generating a sequence $\{\boldsymbol{c}^k\}$. The new iteration $\boldsymbol{c}^{k+1}$ is found in the form $\boldsymbol{c}^{k+1} = \boldsymbol{c}^k + \alpha \boldsymbol{d}^k$, where $\boldsymbol{d}^k$ is a descent direction. This direction is chosen in such a way that $\boldsymbol{J}(\boldsymbol{c}^k + \alpha \boldsymbol{d}^k) < \boldsymbol{J}(\boldsymbol{c}^k)$ for $\alpha \in (0, \overline{\alpha})$, where $\overline{\alpha} > 0$ and α is a suitable step which is obtained by using line search techniques. In gradient type methods, the descent direction is computed by means of the first order derivatives of the minimized functional. It is obvious that

$$\boldsymbol{J}'(\boldsymbol{c}) = \left(\boldsymbol{S}\boldsymbol{w}(\boldsymbol{c}) - \boldsymbol{z}, \boldsymbol{S}\boldsymbol{w}'(\boldsymbol{c})\right) = \left(\boldsymbol{S}^\top \left(\boldsymbol{S}\boldsymbol{w}(\boldsymbol{c}) - \boldsymbol{z}\right), \boldsymbol{w}'(\boldsymbol{c})\right).$$

The problematic part $\boldsymbol{w}'(\boldsymbol{c})$ can be eliminated by using adjoint state problem by using properties of an implicit function. For more details see e.g. [3], [7] or [8].

Acknowledgement: The authors gratefully acknowledge the support by the IGA UPOL grant IGA_Prf_2020_015.

References

[1] D.Y. Gao: *Nonlinear elastic beam theory with application in contact problems and variational approaches*. Mechanics Research Communications 23 (1), 1996, pp. 11–17.

[2] D.Y. Gao, J. Machalová, H. Netuka: *Mixed finite element solutions to contact problems of nonlinear Gao beam on elastic foundation.* Nonlinear Analysis: Real World Applications 22, 2015, pp. 537–550.

[3] J. Haslinger, R. Blaheta, R. Hrtus: *Identification problems with given material interfaces.* Journal of Computational and Applied Mathematics 310, 2017, pp. 129–142.

[4] J. Machalová, H. Netuka: *Control variational method approach to bending and contact problems for Gao beam.* Applications of Mathematics 62 (6), 2017, pp. 661–677.

[5] J. Machalová, H. Netuka: *Solution of contact problems for Gao beam and elastic foundation.* Mathematics and Mechanics of Solids 23 (3), 2018, pp. 473–488.

[6] J. Machalová, H. Netuka: *Comments on the large deformation elastic beam model developed by D.Y. Gao.* Mechanics Research Communications 110, 2020.

[7] J. Radová, J. Machalová, J. Burkotová: *Identification Problem for Nonlinear Gao Beam.* Mathematics 8 (11), 2020.

[8] F. Tröltzsch: *Optimal control of partial differential equations: theory, methods, and applications*, Vol. 112, American Mathematical Soc., 2010.

On a distributed computing platform for a class of contact - impact problems

V. Rek[1], *J. Vala*[2]

[1] Mews Systems Ltd., Prague
[2] Brno University of Technology, Faculty of Civil Engineering

1 Introduction

Dynamic contact problems of several deformable bodies belongs to the set of engineering issues with potential to be solved using the finite element technique and the explicit integration in time in a distributed manner effectively. Their analysis offers a capability to work in dynamically changing environment, provided by some general computer network.

The open-source platform, introduced in this paper, is based on the domain-level decomposition method, with complete dynamic behaviour both in scope of current accessible hardware for computation and of data content. Such platform follows the recent trends of cloud technologies, built on the computer network stack as a core for inter-process communication.

The illustrative example presents an impact of structures assembled from a finite number of shells. It should be understood as an indicator of current capabilities and performance of the suggested strategy for distributed computing.

2 Physical and mathematical background

Following [4], p. 77, the principle of balance of linear momentum, applied to a deformable body $\mathcal{B}$ in the 3-dimensional Euclidean space $\mathcal{R}^3$, reads

$$\begin{aligned} \nabla \cdot \sigma + \varrho f = \varrho a \quad & \text{in} \quad \mathcal{B} \times [0,\tau]\,, \\ u = u_D \quad & \text{on} \quad \partial\mathcal{B}_D \times [0,\tau]\,, \\ b = \sigma \cdot \nu \quad & \text{on} \quad \partial\mathcal{B}_N \times [0,\tau]\,, \\ u(.,0) = \overline{u}\,, \quad v(.,0) = \overline{v} \quad & \text{in} \quad \mathcal{B}\,. \end{aligned} \tag{1}$$

The Hamilton operator $\nabla = (\partial/\partial x_1, \partial/\partial x_2, \partial/\partial x_3)$ in (1) is related to certain fixed Cartesian coordinate system $x = (x_1, x_2, x_3)$, as well as the components of all quantities; moreover $a = \dot{v} = \ddot{u}$ where the upper dot symbol replaces $\partial/\partial t$ for a time t from a finite time interval $[0, \tau]$ and ν denotes the outward unit normal vector to certain part $\partial\mathcal{B}_N$ of a boundary of $\partial\mathcal{B}$ in $\mathcal{R}^3$, whereas $\partial\mathcal{B}_D$ means the rest of such boundary. The following quantities are prescribed: the material density $\rho(x)$, the body forces $b(x,t)$, the surface forces $g(x,t)$, due to the Neumann boundary conditions (3rd equation), the boundary displacements $u_D(x,t)$, due to the Dirichlet boundary conditions (2nd equation), the initial displacements $\overline{u}(x)$ and velocities, due to the couple of Cauchy initial conditions (last equations); $u(x,t)$ are unknown displacements and $\sigma(x,t)$ unknown stresses. The principle of balance of angular momentum by [4], p. 78, forces the symmetry of stress tensors σ.

To enable the evaluation of u from (1), the strain - stress relation can be considered in the sense of Cauchy elasticity by [4], p. 178, using the strain energy function $\Psi(\varepsilon)$, as $\sigma = \partial\Phi(\varepsilon)/\partial\varepsilon$,

taking the usual strain components $\varepsilon_{ij} = u_{i,j} + u_{j,i} + u_{k,i}u_{k,j}$ for $i, j \in \{1, 2, 3\}$ where $u_{i,j}$ means $\partial u_i / \partial x_j$, etc., and k represents the Einstein summation index from $\{1, 2, 3\}$ again. In particular, $\Psi(\varepsilon) = \lambda\,(\mathrm{tr}\ \varepsilon)^2 + 2\mu\varepsilon : \varepsilon$ is well-known as the Hooke law for isotropic materials with 2 positive Lamé constants λ and μ.

The contact constraints between particular bodies, namely the master $\mathcal{B}$ and the slave $\tilde{\mathcal{B}}$ ones, can be quantified with help of a normal gap function

$$g(x,t) = (x - \tilde{x}) \cdot \nu \geq 0\,, \qquad g(x,t)\,p(x,t) = 0\,, \qquad p(x,t) \leq 0\,, \tag{2}$$

which is known as the Hertz - Signorini - Moreau conditions. The total stress traction vector $\sigma \cdot \nu$ on $\partial\mathcal{B}$ can be decomposed into its contact pressure component p, occurring in (2), and the tangential one; $p(g)$ should be evaluated from experiments, at least in the simple form $p = -\varsigma g$, used in the illustrative example here, with just 1 positive material parameter ς. Clearly the effective contact detection is crucial for the design and implementation of any relevant numerical approach working with (1) and (2).

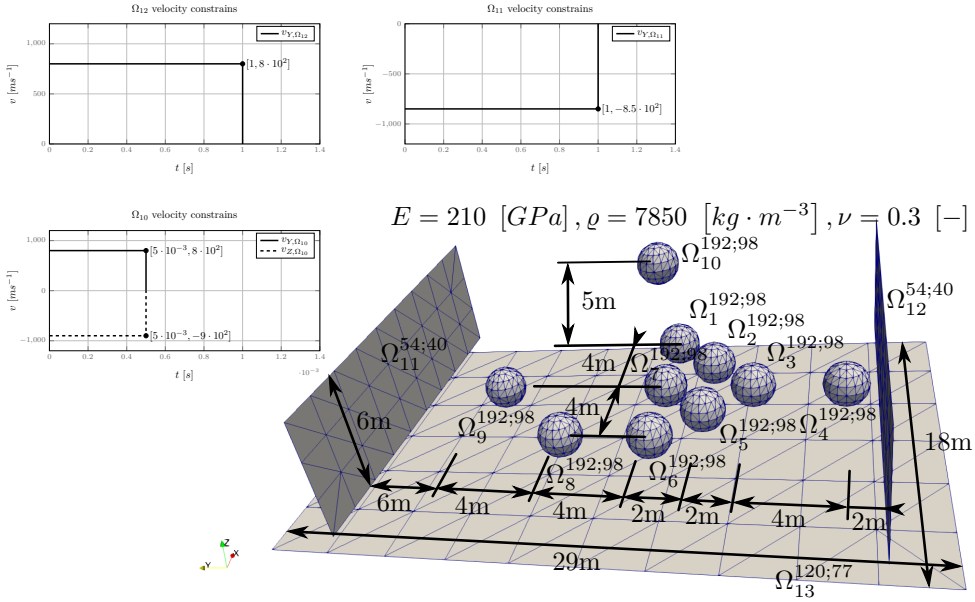


Figure 1: Velocity constraints and model geometry.

3 Numerical approach and computational algorithms

The application of finite element techniques to the weak formulation of (1) and (2) leads to a system of nonlinear ordinary differential equations, whose general form is

$$M\ddot{\mathcal{U}} + F_I(\mathcal{U}) = F_E(\mathcal{U}) + F_C(\mathcal{U}) \quad \text{on } [0, \tau]\,. \tag{3}$$

The symbols M, F_I, F_E and F_C represent the lumped mass matrix and all internal, external and contact forces, respectively, whereas $\mathcal{U}$ is certain vector of a priori unknown real parameters, time-dependent only. Because of the nonlinearity of all additive terms in (3) except the 1st one, some stable explicit time discretization scheme is required, like [8].

Such numerical approach is suitable for application in ah platform that ensure capability to do computations regardless of their environment, i. e. whether they are run in sequential, parallel

or in a hybrid manner on a computer network, as discussed by [5]. Each cluster node considered in the hybrid form of computation is represented by some single workstation, which processes computation of a set of associated macro-entities. It can be assumed that each cluster node comprises a multi-core CPU, capable of executing computational instructions in a fully parallel form.

The platform can be characterized (without technical details) by the following computational algorithms:

1) the main time loop: hybrid parallel explicit finite element analysis process running in the scope of one computer cluster node,

2) explicit integration of contact forces, involving both the detection of pairs of finite element nodes causing the contact force formation and the determination of the magnitude of contact forces: building the kd-tree map T, implementing the nearest neighbour search, inspired by [6], Chap. 9, [2] and [7], and inserting the d-dimensional node into T,

3) explicit integration of the internal and external forces for each particular finite element with the determination of the resultant acting force: application of a range searching query within T computations in co-rotated coordinates regarding a particular finite element from specified domains,

4) final explicit integration of equations of motion (3).

The parallel solution is divided into 2 following levels: i) integration of internal, external and contact forces, respectively, by mapping finite element ranges on the individual cores of a multi-core CPU, ii) processing of the Macro Entity Interaction Multi-graph (MEIM), here the entire domain under the solution, on computer cluster. In the data transfer over such cluster 2 stages must be handled carefully: a) input data migration: both retrieving model data (finite element mesh, material, etc.) and MEIM-based taking care of migrating computational data (serialization and de-serialization), and b) merging of results.

The CAP theorem (i. e. the Brewer's theorem, stemming from [1]) states that a distributed database system can only guarantee two out of the following three characteristics: Consistency, Availability, and Partition tolerance. The CAP characteristics for the above sketched finite element explicit solver have been analyzed properly and documented on benchmark solutions. Only 1 illustrative example follows, due to the limited extent of this presentation.

4 Illustrative example

For the purpose of testing the behaviour of a model during the numerical computations, a dynamic simulation of structures was performed, as evident from Fig. 1, where impacts of 10 spheres and 3 surfaces were observed. The corresponding numerical simulation includes 13 separated discretized objects, composed from C^0-triangular flat shell finite elements that interact with each other by contact forces during the simulation. Superscripts of such objects Ω refer to "number of finite elements; number of finite element nodes" Interactions are initiated by the movement of side-walls. against each other at a constant velocity.

Fig. 2 shows that in the approximately 1st half of solution time 8 successively following data transfers were initiated by penetration of bounding box volumes belonging to the respective macro objects. Such penetration indicates possible subsequent invocation of detailed contact

events between the macro objects. This contact interaction cannot be currently solved in the distributed manner, thus only the parallel solution of this task on one workstation is possible.

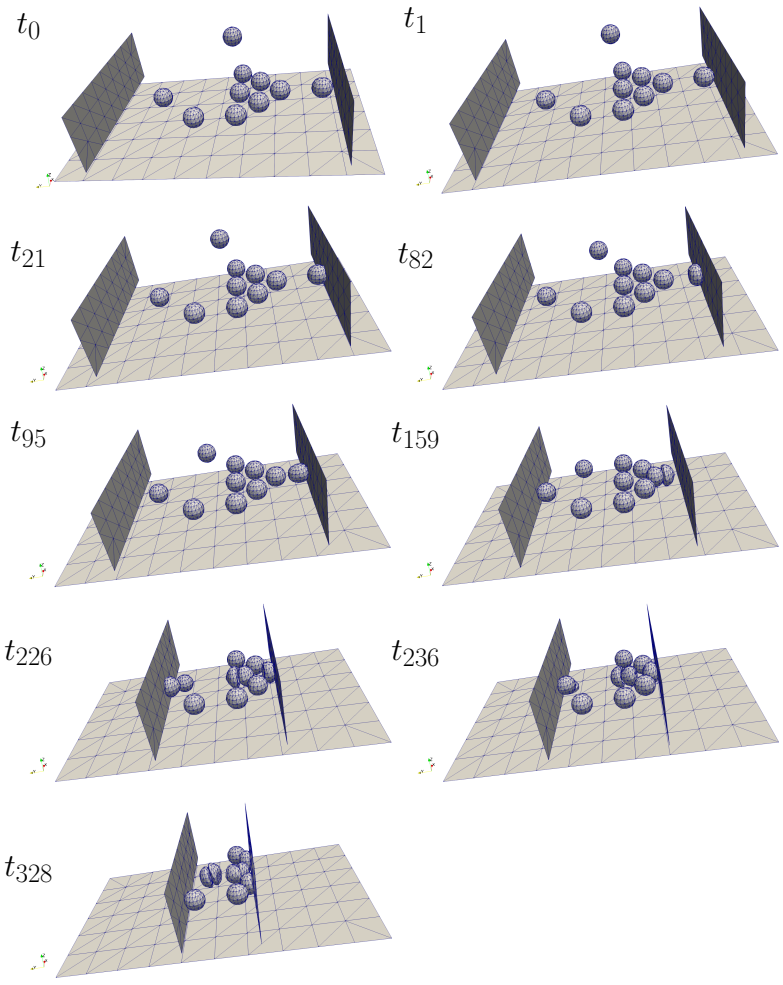


Figure 2: Eight time steps including initial state of the dynamic simulation.

5 Conclusions

The testing structure was computed not in a real distributed environment, but on a single workstation, thus further progress is needed, involving optimization of communication protocol, as well as an improved algorithms for setting the size and shape of bounding boxes encapsulating macro entities (spheres), etc. The expected applications are addressed namely to massive car crashes and large deformation problems with contact of 3-dimensional printed structures.

Acknowledgement: This work was supported from the project of specific university research FAST-S-20-6294 at Brno University of Technology.

References

[1] E.A. Brewer: *Towards robust distributed systems.* 19th PODC (Symposium on Principles of Distributed Computing) – Abstracts in Portland (Oregon, USA), 2000, Association for Computing Machinery, 2000, p. 7.

[2] Y. Chen, L. Zhou, Y. Tang, J.P. Singh, N. Bouguila, C. Wang, H. Wang, J. Du: *Fast neighbor search by using revised k-d tree.* Information Sciences 472, 2019, pp. 145–162.

[3] X. Dong, X. Yin, Q. Deng, B. Yu, H. Wang, P. Weng, C. Chen, H. Yuan: *Local contact behavior between elastic and elasticâ€“plastic bodies.* International Journal of Solids and Structures 150, 2018, pp. 22–39.

[4] K. Hashiguchi, *Elastoplasticity Theory.* Springer, 2014.

[5] V. Rek, *Explicit integration of equations of motion solved on computer cluster.* 19th PANM (Programs and Algorithms of Numerical Mathematics) – Proceedings of Seminar in Hejnice (Czech Republic), 2018, Institute of Mathematics CAS, 2019, pp. 119–131.

[6] W.J. Stronge, *Impact Mechanics.* Cambridge University Press, 2000.

[7] D. Wehr, R. Radkowski: *Parallel kd-tree construction on the GPU with an adaptive split and sort strategy.* International Journal of Parallel Programming 46, 2018, pp. 1139–1156.

[8] M. Zheng, Z. Yuan, Q. Tong, G. Zhang, W. Zhu: *A novel unconditionally stable explicit integration method for finite element method.* The Visual Computer 34, 2018, pp. 721–733.

Geometry algebra and conditionality of linear system of equations

V. Skala

University of West Bohemia, Pilsen

1 Introduction

A new approach to the matrix conditionality and the solvability of the linear systems of equations is presented. It is based on the application of the *geometric algebra* with the projective space representation using homogeneous coordinates representation. There are two main groups:

- non-homogeneous systems of linear equations, i.e. $\mathbf{Ax} = \mathbf{b}$
- homogeneous system of equations, i.e. $\mathbf{Ax} = \mathbf{0}$

Using the *principle of duality* and *projective extension of the Euclidean space* the first type of the linear system, i.e. $\mathbf{Ax} = \mathbf{b}$, can be easily transformed to the second type, i.e. $\mathbf{Ax} = \mathbf{0}$. The *geometric algebra* offers more general formalism, which can be used for a better understanding of the linear system of equations properties and behavior.

1.1 Geometric algebra

The *Geometric Algebra* (GA) uses a "new" product called *geometric product* defined as:

$$\mathbf{ab} = \mathbf{a} \cdot \mathbf{b} + \mathbf{a} \wedge \mathbf{b} \tag{1}$$

where $\mathbf{ab}$ is the new entity. It should be noted, that it is a "bundle" of objects with different dimensionalities and properties, in general. In the case of the n-dimensional space, the vectors are defined as $\mathbf{a} = (a_1\mathbf{e}_1 + ... + a_n\mathbf{e}_n)$, $\mathbf{b} = (b_1\mathbf{e}_1 + ... + b_n\mathbf{e}_n)$ and the $\mathbf{e}_i$ vectors form orthonormal vector basis in E^n. In the E^3 case, the following objects can be used in geometric algebra: [5]:

1	0-vector (scalar)	$\mathbf{e}_{12}, \mathbf{e}_{23}, \mathbf{e}_{31}$	2-vectors (bivectors)
$\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3,$	1-vector (vectors)	$\mathbf{e}_{123}$	3-vector (pseudoscalar)

The significant advantage of the geometric algebra is, that it is more general that than the Gibbs algebra and can handle all objects with dimensionality up to n. The geometry algebra uses the following operations, including the inverse of a vector.

$$\mathbf{a} \cdot \mathbf{b} = \frac{1}{2}(\mathbf{ab} + \mathbf{ba}) \qquad \mathbf{a} \wedge \mathbf{b} = -\mathbf{b} \wedge \mathbf{a} \qquad \mathbf{a}^{-1} = \mathbf{a}/||\mathbf{a}||^2 \tag{2}$$

It should be noted, that geometric algebra is *anti-commutative* and the "pseudoscalar" I in the E^3case has the basis $\mathbf{e}_1\mathbf{e}_2\mathbf{e}_3$ (briefly as $\mathbf{e}_{123}$), i.e.

$$\mathbf{e}_i\mathbf{e}_j = -\mathbf{e}_j\mathbf{e}_i \qquad \mathbf{e}_i\mathbf{e}_i = 1 \qquad \mathbf{e}_1\mathbf{e}_2\mathbf{e}_3 = I \qquad \mathbf{a} \wedge \mathbf{b} \wedge \mathbf{c} = q \tag{3}$$

where q is a scalar value (actually a pseudoscalar).

2 Solution of linear systems of equations

The linear system of equations $\mathbf{Ax} = \mathbf{b}$ can be transformed to the homogeneous system of linear equations, i.e. to the form $\mathbf{D}\xi = \mathbf{0}$, where $\mathbf{D} = [\mathbf{A}|-\mathbf{b}]$, $\xi = [\xi_1, ..., \xi_n : \xi_w]^T$, $x_i = \xi_i \ / \ \xi_w$, $i = 1, ..., n$. If $\xi_w \mapsto 0$ then the solution is in infinity and the vector $(\xi_1, ..., \xi_n)$ gives the "direction", only.

As the solution of a linear system of equations is equivalent to the outer product (generalized cross-vector) of vectors formed by rows of the matrix $\mathbf{D}$, the solution of the system $\mathbf{D}\xi = \mathbf{0}$ is defined as:

$$\xi = \mathbf{d}_1 \wedge \mathbf{d}_2 \wedge ... \wedge \mathbf{d}_n \qquad \mathbf{D}\xi = 0 \quad , \ i.e. \quad [\mathbf{A}| - \mathbf{b}]\xi = 0 \tag{4}$$

where: $\mathbf{d}_i$ is the i-th row of the matrix $\mathbf{D}$, i.e. $\mathbf{d}_i = (a_{i1}, ..., a_{in}, -b_i)$, $i = 1, ..., n$. The application of the projective extension of the Euclidean space enables us to transform the non-homogeneous system of linear equations $\mathbf{Ax} = \mathbf{b}$ to the homogeneous linear system $\mathbf{D}\xi = \mathbf{0}$, i.e.:

$$\begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix} \quad \overset{\text{conversion}}{\Longleftrightarrow} \quad \begin{bmatrix} a_{11} & \cdots & a_{1n} & -b_1 \\ \vdots & \ddots & \vdots & \vdots \\ a_{n1} & \cdots & a_{nn} & -b_n \end{bmatrix} \begin{bmatrix} \xi_1 \\ \vdots \\ \xi_n \\ \xi_w \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix} \tag{5}$$

It is an important result as a solution of a linear system of equations is formally the same for both types, i.e. homogeneous linear systems $\mathbf{Ax} = \mathbf{0}$ and non-homogeneous systems $\mathbf{Ax} = \mathbf{b}$.

2.1 Angular criterion

Both types of the linear systems of equations, i.e. $\mathbf{Ax} = \mathbf{b}$ ($\mathbf{A}$ is $n \times n$) and $\mathbf{Ax} = \mathbf{0}$ ($\mathbf{A}$ is $(n+1) \times n$), actually have the same form $\mathbf{Ax} = \mathbf{0}$ ($\mathbf{A}$ is $(n+1) \times n$), now, if the projective representation is used. Therefore, it is possible to show the differences between the matrix conditionality and conditionality (solvability) of a linear system of equations, see Fig.1.

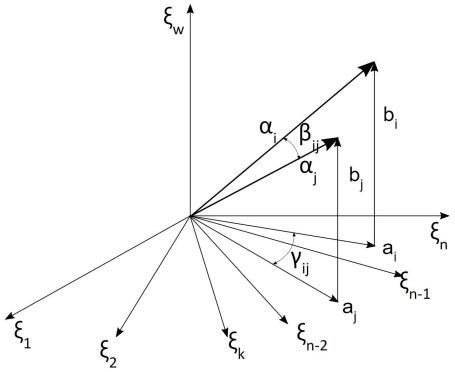


Figure 1: Difference between matrix and linear system conditionality

The eigenvalues are usually used and the ratio $rat_\lambda = |\lambda_{max}|/|\lambda_{min}|$ & $\lambda \in C$ is mostly used as a criterion. If the ration rat_λ is high, the matrix is said to be ill-conditioned, especially in the case of large data with a large span of data. There are two cases, which are needed to be taken into consideration:

- non-homogeneous systems of linear equations, i.e. $\mathbf{Ax} = \mathbf{b}$. In this case, the matrix conditionality is considered as a criterion for the solvability of the linear system of equations. It depends on the matrix $\mathbf{A}$ properties, i.e. on eigenvalues.

$$\begin{bmatrix} 10^2 & 0 & 0 \\ 0 & 10^0 & 0 \\ 0 & 0 & 10^{-2} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_3 \end{bmatrix} = \begin{bmatrix} b_1 \\ \vdots \\ b_3 \end{bmatrix} \tag{6}$$

  A conditionality number $\kappa(\mathbf{A}) = |\lambda_{max}|/|\lambda_{min}|$ is usually used as the solvability criterion. In the case of the Eq.6, the matrix conditionality is $\kappa(\mathbf{A}) = 10^2/10^{-2} = 10^4$. However, if the 1^{st} row is multiplied by 10^{-2} and the 3^{rd} row is multiplied by 10^2, then the conditionality is $\kappa(\mathbf{A}) = 1$.

- a homogeneous system of equations $\mathbf{Ax} = \mathbf{0}$, when the system of linear equations $\mathbf{Ax} = \mathbf{b}$ is expressed in the projective space. In this case, the vector $\mathbf{b}$ is taken into account and bivector area and bivector angles properties can be used for solvability evaluation.

The only *angular* criterion is invariant to the row multiplications, while only the column multiplication changes angles of the bivectors. There are several significant consequences:

- the solvability of a linear system of equations can be improved by the column multiplications, only, if unlimited precision is considered. Therefore, the matrix-based preconditioners might not solve the solvability problems and might introduce additional numerical problems.

- the precision of computation is significantly influenced by addition and subtraction operations, as the exponents must be the same for those operations with mantissa. Also, the multiplication and division operations using exponent change by $2^{\pm k}$ should be preferred.

2.2 Preconditioning simplified

There are several methods used to improve the ratio $\kappa(\mathbf{A}) = |\lambda_{max}|/|\lambda_{min}|$ of the matrix $\mathbf{A}$ of the linear system, e.g. matrix eigenvalues shifting or preconditioning [1] [2]. The preconditioning is usually based on solving a linear system $\mathbf{Ax} = \mathbf{0}$:

$$\mathbf{PAS}\ \ \mathbf{S}^{-1}\mathbf{x} = \mathbf{Pb} \tag{7}$$

where $\mathbf{P}$ is a matrix, which can cover complicated computation, including Fourier transform. The inverse operation, i.e. $\mathbf{P}$, is computationally very expensive as it is of $O(n^3)$ complexity. Therefore, they are not easily applicable for large systems of linear equations used nowadays. There are methods based on incomplete factorization, etc., which might be used [3]. The proposed matrix conditionality improvement method requires only the diagonal matrices values $\mathbf{P}$ and $\mathbf{S}$, i.e. multiplicative coefficients $p_i \neq 0$, $s_j \neq 0$, which have to be optimized. This is a significant reduction of computational complexity, as it decreases the cost of finding sub-optimal p_i, s_j values. The proposed approach was tested on the Hilbert's matrix as conditionality can be estimated as $\kappa(\mathbf{H_n}) \simeq e^{3.5n}$. The experimental results of the original conditionality $\kappa(\mathbf{H}_{orig})$ and conditionality using the proposed method $\kappa(\mathbf{H}_{new})$ are presented in Tab.1.

Table 1: Conditionality of modified the Hilbert matrix: Experimental results (*with Octave warnings).

N	cond($\mathbf{H}_{orig}$)	cond($\mathbf{H}_{new}$)		N	cond($\mathbf{H}_{orig}$)	cond($\mathbf{H}_{new}$)
3	5.2406e+02	2.5523e+02		7	4.7537e+08	1.4341e+08
4	1.5514e+04	6.0076e+03		8	1.5258e+10	6.0076e+03
5	4.7661e+05	1.6099e+05		9	4.9315e+11	1.3736e+11
6	1.4951e+07	5.0947e+06		10	1.6024e+13	4.1485e+12
				20	1.6024e+13*	4.1485e+12

The experiments proved, that the conditionality cond($\mathbf{H}_{new}$) of the modified matrix using the proposed approach was decreased by more than half of the magnitude for higher values of n, see Tab.1. This is consistent with the recently obtained results [4], where the inverse Hilbert matrix computation using the modified Gauss elimination without division operation was analyzed.

The Hilbert matrix conditionality improvement also improved the angular criterion based on maximizing the ratio $\kappa_{rat}(\mathbf{H})$ defined as:

$$\kappa_{rat}(\mathbf{H}) = \frac{\cos\gamma_{min}}{cos\gamma_{max}} \qquad \kappa_{rat}(\mathbf{H}) = \frac{\cos\beta_{min}}{cos\beta_{max}} \tag{8}$$

It says, how the angles $\cos\gamma_{ij}$, formed by the vectors $\mathbf{a}_{ij}$ of the bivectors are similar, see Fig.1. It means, that if the ratio $\kappa_{rat}(\mathbf{A}) \simeq 1$ the angles of all bivectors are nearly equal. In the case of conditionality assessment of the linear system of equations $\mathbf{Ax} = \mathbf{0}$, the angles β_{ij}, formed by the angels α_{ij} have to be taken into account, see Fig.1. The results presented in Tab.2 reflects the improvement of the Hilbert matrix by proposed approach using the diagonal matrices $\mathbf{P}$ and $\mathbf{S}$ used as the multipliers.

Table 2: Conditionality of modified the Hilbert matrix: Experimental results (*with Octave warnings).

N	3	4	5	6	7
$\kappa_{rat}(\mathbf{H}_{orig})$	0.54464	0.39282	0.31451	0.26573	0.23195
$\kappa_{rat}(\mathbf{H}_{new})$	0.98348	0.97740	0.98173	0.96283	0.87961
N	8	9	10		20
$\kappa_{rat}(\mathbf{H}_{orig})$	0.20694	0.18755	0.17199	$\cdots$	0.09917*
$\kappa_{rat}(\mathbf{H}_{new})$	0.92500	0.96435	0.96322	$\cdots$	0.74701*

3 Conclusion

The advantage of the angular criterion is that it is common for the conditionality evaluation of the matrix and of the linear system of equations. It should be noted, that this conditionality assessment method gives different values of conditionality of those two different cases, as in the first case only the matrix is evaluated, while in the second one the value of the $\mathbf{b}$ in the $\mathbf{Ax} = \mathbf{b}$ is taken into account.

Acknowledgement: The author would like to thank their colleagues and students at the University of West Bohemia, Plzen, for their discussions and suggestions and implementations.

References

[1] M. Benzi: *Preconditioning techniques for large linear systems: A survey.* Journal of Computational Physics 182(2), 2002, pp. 418–477.

[2] K. Chen: *Matrix Preconditioning Techniques and Applications.* Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, 2005.

[3] Y. Saad: *Numerical problems for Large Eigenvalue Problems.* SIAM, 2 edition, 2011.

[4] V. Skala: *Modified Gaussian Elimination without Division Operations.* In ICNAAM 2013 Proceedings, Vol. 1558 of AIP Conf.Proceedings, 2013, pp. 1936–1939.

[5] J. Vince: *Geometric Algebra for Computer Graphics.* Springer, 2008.

An abstract inf-sup problem with bilinear Lagrangian and convex constraints and its applications

S. Sysala

Institute of Geonics of the Czech Academy of Sciences, Ostrava

This contribution is concerned with analysis of the abstract duality problem

$$\lambda^* := \sup_{x\in P} \inf_{\substack{y\in Y \\ L(y)=1}} a(x,y) \stackrel{?}{=} \inf_{\substack{y\in Y \\ L(y)=1}} \sup_{x\in P} a(x,y) =: \zeta^*, \tag{1}$$

where $P \subset X$ is a closed, convex set with $0 \in P$, X, Y are Banach spaces, L is a non-trivial continuous linear functional in Y, and $a\colon X \times Y \to \mathbb{R}$ is a bilinear form continuous with respect to both arguments. Henceforth the problem in the right hand side of (1) is called primal, while the one in the left hand side is called dual. It is easy to check that $0 \le \lambda^* \le \zeta^* \le +\infty$. In general, necessary and sufficient conditions for $\lambda^* = \zeta^*$ are unknown.

In classical perfect plasticity, (1) represents the limit analysis problem enabling to determine the critical (limit) load and related failure zones. It is used in geotechnical stability assessment and in other applications. In [1], the limit analysis problem has been analyzed by using the so-called *inf-sup condition on convex cones* which generalizes the well-known Babuška-Brezzi condition. If such a condition holds then one can, for example, prove the equality $\lambda^* = \zeta^*$ or find a computable majorant of ζ^*.

Recently in [2], it was shown that the problem (1) can also be used in strain-gradient plasticity for finding the elastic threshold or plastically admissible stresses. Limit analysis for the strain-gradient plasticity can be defined by (1), too, see [3].

The fact that (1) has more applications motivates us to study this abstract problem in more details and extend our results from [1, 2] to the abstract setting for other eventual applications. The abstract form is also more convenient for readers which are not too familiar with the theory of plasticity. Beside the mentioned inf-sup condition on convex cones and the computable majorant, a regularization method convenient for numerical solution of (1) is also presented and analyzed. The analysis of (1) can be found in [3], in more details.

References

[1] J. Haslinger, S. Repin, S. Sysala: *Inf-sup conditions on convex cones and applications to limit load analysis.* Mathematics and Mechanics of Solids 24, 2019, pp. 3331–3353.

[2] B.D. Reddy, S. Sysala: *Bounds on the elastic threshold for problems of dissipative strain-gradient plasticity.* Journal of the Mechanics and Physics of Solids 143, 2020, 104089.

[3] S. Sysala, J. Haslinger, B.D. Reddy, S. Repin: *An abstract inf-sup problem inspired by limit analysis in perfect plasticity and related applications.* Submitted, `https://arxiv.org/pdf/2009.03535.pdf`

Finite element phase-field model for multivariant martensitic transformation at finite-strain

K. Tůma[1], *M. Rezaee-Hajidehi*[2], *J. Hron*[1], *P.E. Farrell*[3], *S. Stupkiewicz*[2]

[1] Charles University, Faculty of Mathematics and Physics, Prague
[2] Institute of Fundamental Technological Research of the Polish Academy of Sciences, Warsaw
[3] University of Oxford, Mathematical Institute

Reversible martensitic phase transformation is the mechanism for two important properties in shape memory alloys, namely shape memory effect and pseudoelasticity. We study CuAlNi in which austenite transforms to six variants of martensite within a cubic to orthorhombic transformation and creates an interesting fine microstructure.

Our previously developed finite strain phase-field model for martensitic transformation [1, 2] was enhanced such that is capable of capturing all six variants of martensite in CuAlNi described above. A 2D version of the model has been used to study the martensitic transformation in 2D nano-indentation problems [3]. The model admits an arbitrary crystallography of transformation and arbitrary elastic anisotropy of all phases. It incorporates Hencky-type elasticity that is implemented using the Padé approximants [5], a penalty-regularized double-obstacle potential [4], and viscous dissipation.

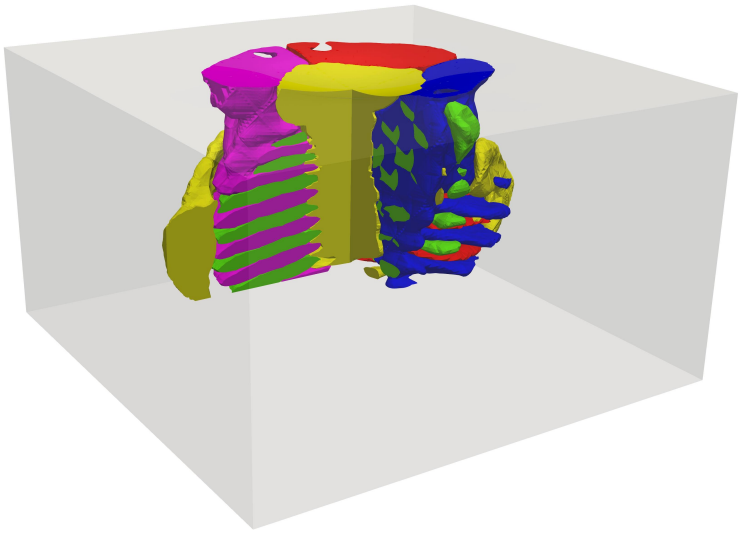


Figure 1: Fine microstructure of CuAlNi shape memory alloy, that appears during the nano-indentation, illustrates the saw-tooth and twinning morphologies.

The finite-element discretization of the model is performed in Firedrake and relies on the PETSc solver library. The large systems of linear equations arising are efficiently solved using GMRES and a geometric multigrid preconditioner with a carefully chosen relaxation. The modeling capabilities are illustrated through a 3D simulation of the microstructure evolution during nano-indentation, with all six orthorhombic martensite variants taken into account. To capture the fine miscrostructure it is needed to solve a very large problem, see Figure 1.

Robustness and a good parallel scaling performance have been demonstrated, with the problem size reaching 150 million degrees of freedom. All finite-element simulations were carried out on the high-performance clusters operated by the IT4Innovations National Supercomputing Center in Ostrava, Czech Republic, namely, the Barbora cluster (BullSequana XH2000) consisting of 200 computing nodes, where each node possesses two 18-core Intel Xeon Gold 6240 processors (2.60 GHz, 192 GB RAM) with InfiniBand HDR, connected in a fat tree topology, running Red Hat Enterprise Linux Server release 7.

Acknowledgement: J.H. and K.T. have been supported by the Charles University Research program No. UNCE/SCI/023. K.T. has been supported by the Czech Science project 18-12719S. M.R.H. and S.S. have been supported by the National Science Center (NCN) in Poland through Grant No. 2018/29/B/ST8/00729. P.E.F. has been supported by EPSRC grants EP/R029423/1 and EP/V001493/1. This work was supported by the Ministry of Education, Youth and Sports of the Czech Republic from the Large Infrastructures for Research, Experimental Development and Innovations project "IT4Innovations National Supercomputing Center (LM2015070)".

References

[1] K. Tůma, S. Stupkiewicz, H. Petryk: *Size effects in martensitic microstructures: Finite-strain phase field model versus sharp-interface approach.* J. Mech. Phys. Solids 95, 2016, pp. 284–307.

[2] K. Tůma, S. Stupkiewicz: *Phase-field study of size-dependent morphology of austenite-twinned martensite interface in CuAlNi.* Int. J. Solids Struct. 97, 2016, pp. 89–100.

[3] M. Rezaee-Hajidehi, S. Stupkiewicz: *Phase-field modeling of multivariant martensitic microstructures and size effects in nano-indentation.* Mech. Mat. 141, 2020, 103267.

[4] I. Steinbach: *Phase-field models in materials science.* Modelling Simul. Mat. Sci. Engng. 17, 2009, 073001.

[5] M. Rezaee-Hajidehi, K. Tůma, S. Stupkiewicz: *A note on Padé approximants of tensor logarithm with application to Hencky-type hyperelasticity.* Comp. Mech., 2020.

Stabilized IgA-based method for RANS equations and $k-\omega$ turbulence model

E. Turnerová, K. Slabá, M. Brandner

University of West Bohemia, Pilsen

1 Introduction

The contribution is focused on application of isogeometric analysis (IgA) to incompressible turbulent flow problems solving RANS (Reynolds-Averaged Navier-Stokes) equations closed with $k-\omega$ turbulence model. Since IgA is continuous Galerkin-based method, the numerical solution of convection dominated problems (containing sharp layers where the solution gradients are very large) is usually polluted by spurious (unphysical) oscillations, which cause loss of accuracy and stability. The stabilization techniques, which improve the stability, however, without degrading accuracy are investigated.

2 Reynolds–Averaged Navier–Stokes equations

The initial boundary value Navier–Stokes problem is given as a system of $d+1$ equations together with initial and mixed boundary conditions as follows

$$\begin{aligned} \frac{\partial \mathbf{u}}{\partial t} + \mathbf{u}\cdot\nabla\mathbf{u} &= \mathbf{f} - \nabla p + \nabla\cdot\left[\nu(\nabla\mathbf{u} + \nabla\mathbf{u}^T)\right] \quad \text{in } \Omega\times(0,\bar{t}),\\ \nabla\cdot\mathbf{u} &= 0 \quad \text{in } \Omega\times(0,\bar{t}),\\ \mathbf{u}(\mathbf{x},0) &= \mathbf{u}_0(\mathbf{x}) \quad \text{in } \Omega,\\ \mathbf{u} &= \mathbf{g} \quad \text{in } \partial\Omega_D\times[0,\bar{t}],\\ \nu(\nabla\mathbf{u} + \nabla\mathbf{u}^T)\cdot\mathbf{n} - \mathbf{n}p &= \mathbf{0} \quad \text{in } \partial\Omega_N\times[0,\bar{t}], \end{aligned} \tag{1}$$

where $\mathbf{u}$ is the velocity, p is the pressure, ν is the given kinematic viscosity of the fluid and $\mathbf{g}$ and $\mathbf{u}_0(\mathbf{x})$ are given functions. In the case of flow with dominant convection, we use the Reynolds decomposition of velocity and pressure and substitute this decomposition into the Navier-Stokes equations (1). The turbulent flow is considered to be the sum of the mean $\bar{\mathbf{U}} = [\bar{\mathbf{u}}, \bar{p}]^T$ and fluctuating $\mathbf{U}' = [\mathbf{u}', p']^T$ components. It yields

$$\begin{aligned} \frac{\partial \bar{\mathbf{u}}}{\partial t} + \bar{\mathbf{u}}\cdot\nabla\bar{\mathbf{u}} &= -\nabla\bar{p} + \nabla\cdot\left[\nu(\nabla\bar{\mathbf{u}} + \nabla\bar{\mathbf{u}}^T)\right] - \overline{\mathbf{u}'\cdot\nabla\mathbf{u}'},\\ \nabla\cdot\bar{\mathbf{u}} &= 0. \end{aligned} \tag{2}$$

Applying the Boussinesq hypothesis to the equations (2), the Reynolds–Averaged Navier–Stokes problem is written

$$\begin{aligned} \frac{\partial \bar{\mathbf{u}}}{\partial t} + \bar{\mathbf{u}}\cdot\nabla\bar{\mathbf{u}} &= -\nabla\bar{p} + \nabla\cdot\left[(\nu+\nu_T)(\nabla\bar{\mathbf{u}} + \nabla\bar{\mathbf{u}}^T)\right] \quad \text{in } \Omega\times(0,\bar{t}),\\ \nabla\cdot\bar{\mathbf{u}} &= 0 \quad \text{in } \Omega\times(0,\bar{t}),\\ \bar{\mathbf{u}}(\mathbf{x},0) &= \bar{\mathbf{u}}_0(\mathbf{x}) \quad \text{in } \Omega,\\ \bar{\mathbf{u}} &= \mathbf{g} \quad \text{in } \partial\Omega_D\times[0,\bar{t}],\\ (\nu+\nu_T)(\nabla\bar{\mathbf{u}} + \nabla\bar{\mathbf{u}}^T)\cdot\mathbf{n} - \mathbf{n}\bar{p} &= \mathbf{0} \quad \text{in } \partial\Omega_N\times[0,\bar{t}], \end{aligned} \tag{3}$$

where ν_T is eddy (turbulent) viscosity introduced below. The Boussinesq approach provides simplification which allows the simulation of the effects of the turbulence on the mean flow with relatively low memory requirements.

In essentially all practical formulations of the RANS equations, the time derivative term is included, despite the fact that $\bar{\mathbf{u}}$ is independent of time. This formulation is only auxiliary. The solution of the RANS equations is understood as stationary in our case.

3 SST $k-\omega$ turbulence model

The SST (shear stress transport) $k-\omega$ turbulence model is used for the numerical computations in this work and is written as (cf. e.g. [3])

$$\begin{aligned}
\frac{\partial k}{\partial t} + \bar{\mathbf{u}}\cdot\nabla k &= P_k + \nabla\cdot\left[\left(\sigma_k\nu_T+\nu\right)\nabla k\right] - \beta^* k\omega \quad \text{in } \Omega\times(0,\bar{t}),\\
\frac{\partial \omega}{\partial t} + \bar{\mathbf{u}}\cdot\nabla \omega &= \alpha S^2 + \nabla\cdot\left[\left(\sigma_\omega\nu_T+\nu\right)\nabla\omega\right] - \beta\omega^2 + 2\left(1-F_1\right)\sigma_{\omega 2}\frac{1}{\omega}\nabla k\cdot\nabla\omega \quad \text{in } \Omega\times(0,\bar{t}),\\
k(\mathbf{x},0) &= k_0(\mathbf{x}), \quad \omega(\mathbf{x},0) = \omega_0(\mathbf{x}) \quad \text{in } \Omega,\\
k &= g_k, \qquad \omega = g_\omega \quad \text{in } \partial\Omega_D\times[0,\bar{t}],\\
\nabla k\cdot\mathbf{n} &= 0, \quad \nabla\omega\cdot\mathbf{n} = 0 \quad \text{in } \partial\Omega_N\times[0,\bar{t}],
\end{aligned} \tag{4}$$

where k is the turbulence kinetic energy, ω is the specific dissipation rate, S is the strain rate tensor,

$$F_1 = \tanh\left(\left[\min\left[\max\left(\frac{\sqrt{k}}{\beta^*\omega y}, \frac{500\nu}{y^2\omega}\right), \frac{4\sigma_{\omega 2}k}{CD_{k\omega}y^2}\right]\right]^4\right),$$

$$CD_{k\omega} = \max\left(2\rho\sigma_{\omega 2}\frac{1}{\omega}\nabla k\cdot\nabla\omega, 10^{-10}\right), \quad P_k = \min\left(\nu_T f, 10\beta^* k\omega\right),$$

$\beta^* = \frac{9}{100}$, $\sigma_{\omega 2} = 0.856$. The values of the remaining parameters σ_k, σ_ω, α and β are dependent on the wall distance y. Let ϕ_1 and ϕ_2 be two given parameters. Then define a parameter ϕ, whose value depends on the wall distance y, such that it varies between the given parameters ϕ_1, ϕ_2 as

$$\phi = \phi_1 F_1 + \phi_2\left(1-F_1\right). \tag{5}$$

This relation is applied to calculate appropriate values of the parameters σ_k, σ_ω, α and β using

$$\begin{aligned}
\sigma_{k1} &= 0.85, & \sigma_{k2} &= 1, & \sigma_{\omega 1} &= 0.5, & \sigma_{\omega 2} &= 0.856,\\
\alpha_1 &= \tfrac{5}{9}, & \alpha_2 &= 0.44, & \beta_1 &= \tfrac{3}{40}, & \beta_2 &= 0.0828,
\end{aligned} \tag{6}$$

where the parameter $\sigma_{\omega 2}$ is already given above. Since the two-equation model switches according to the wall distance y, the eddy viscosity also has to be dependent on the wall distance. The form of the eddy viscosity is given by

$$\nu_T = \frac{k}{\max\left(\omega, SF_2\right)}, \tag{7}$$

where

$$F_2 = \tanh\left(\left[\max\left(\frac{2\sqrt{k}}{\beta^*\omega y}, \frac{500\nu}{y^2\omega}\right)\right]^2\right). \tag{8}$$

The SST turbulence model belongs to the group of LRN models. Thus, it is not necessary to use wall functions. This means that we apply damping functions for certain terms in the equations of the turbulence model.

4 Isogeometric analysis

Our numerical approach is based on the Galerkin method. We therefore start from a weak formulation of the system of partial differential equations (3) and (4). The function spaces taken in this formulation are approximated by finite-dimensional subspaces, which will be used to approximate the solution of the problem. Isogeometric analysis (IgA) shares a lot of features with the finite element method (FEM) and it is even usually understood as a modification of FEM such that other basis functions are chosen in the Galerkin approximation. In contrast to FEM, IgA is closely related to the description of geometry and takes inspiration from Computer Aided Design (CAD), which allows exact geometry representation. Indeed, the computational domain with a boundary represented as B-spline/NURBS objects can be exactly discretized and then the isoparametric approach is applied, meaning the solution spaces of the velocity and pressure approximation are generated by the same basis functions which represent the geometry. This is the main advantage of IgA, which cannot be achieved by a FEM polynomial description of the boundaries. More details about the B-spline basis can be found in [1].

5 Stabilization techniques for IgA

The turbulence model consists of a system of two convection-diffusion-reaction equations. The Navier-Stokes equations contain a convection term, which plays a crucial role in the case of large Reynolds numbers. It follows that in many cases it is necessary to stabilize the numerical scheme so that it is not polluted by artificial oscillations. We focus on CSD (Classical Streamline Diffusion), SUPG, isotropic artificial diffusion methods and crosswind methods. We are also proposing a new consistent stabilization technique called tanh-CSD. A description of a number of stabilization techniques can be found in [2].

6 Numerical experiment

As part of the contribution, we will present the simulation of the fluid flow in the blade cascade. The geometry in Figure 1 (left) is the B-spline representation of the computational domain used in the experiments. It consists of three conforming patches with B-spline basis of the degree $q = 3$. The middle (red) patch displayed in Figure 1 (left) represents the part of the domain between two neighboring blade profiles. The blade profile in Figure 1 is a chosen unfolded cylindrical slice of the runner blade of the Kaplan turbine.

The discrete RANS problem is solved decoupled from the discrete turbulence model. In the first time step, the linearized RANS problem is solved until the iterations of the numerical solution of $\bar{\mathbf{u}}$ and $\bar{p}$ converge (or until the maximum number of the iterative process is achieved). Then, we continue solving the discrete turbulence model such that the computed velocity and pressure solutions are used for the evaluation of the turbulence model terms. This sequence is repeated for each time step. To achieve a stable iteration process, we first use a suitable SOLD technique for stabilization of the RANS equations and/or SOLD method for the turbulence model. The numerical simulation is stopped at a suitable time and the obtained result is used as the initial condition for the subsequent computation of the RANS problem such that any stabilization is used neither for RANS equations nor for turbulence model. We indicate this approach by 'init.stab.'. Another stable simulation is achieved such that any stabilization technique is applied for the RANS equations from the beginning of the computation, only turbulence model is

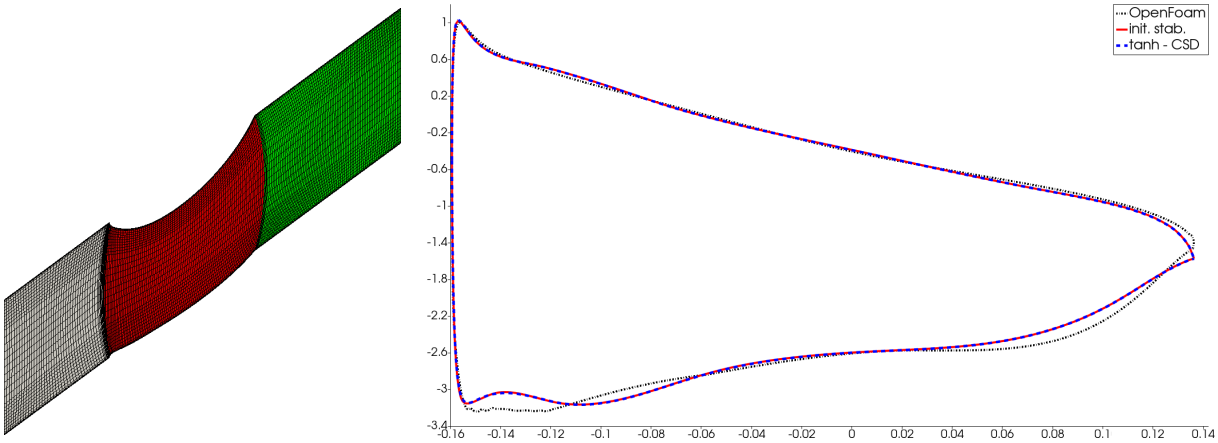


Figure 1: B-spline geometry representation of the blade cascade and the computational mesh with 74879 DOFs (left). Pressure coefficient in t = 2s (right).

stabilized using tanh-CSD method. This approach is indicated by 'tanh-CSD'. We compare our solutions with OpenFOAM result (free, open source CFD software) in Figure 1 (right), where the pressure coefficient is shown in t = 2s. Evidently, our both approaches give almost identical numerical solution.

Acknowledgement: This work was supported by the Czech Science Foundation [grant number 19-04006S].

References

[1] T.J. Hughes, J.A. Cottrell, Y. Bazilevs: *Isogeometric analysis: CAD, finite elements, nurbs, exact geometry and mesh refinement.* Computer methods in applied mechanics and engineering 194, 2005, pp. 4135–4195.

[2] V. John, P. Knobloch: *On spurious oscillations at layers diminishing (sold) methods for convection–diffusion equations: Part I–A review.* Computer methods in applied mechanics and engineering 196, 2007, pp. 2197–2215.

[3] R. Nichols: *Turbulence models and their application to complex flows.* Revision 4.01, University of Alabama at Birmingham, 2014.

Numerical solution of macroscopic traffic flow models on networks using numerical fluxes at junctions

L. Vacek, V. Kučera

Charles University, Faculty of Mathematics and Physics, Prague

1 Introduction

Modelling of traffic flows will have an important role in the future. With a rising number of cars on the roads, we must optimize the traffic situation. That is the reason we started to study traffic flows. It is important to have working models which can help us to improve traffic flow. We can model real traffic situations and optimize e.g. the timing of traffic lights or local changes in the speed limit. The benefits of modelling and optimization of traffic flows are both ecological and economical.

Let us have a road and an arbitrary number of cars. We would like to model the movement of cars on our road. There are two main ways how to describe traffic flow. The first way is the *microscopic model*. Microscopic models describe every car and we can specify the behaviour of every driver and type of car. The basic microscopic models are described by ordinary differential equations. The second approach is the *macroscopic model*. In that case, we view our traffic situation as a continuum and study the density of cars in every point of the road. This model is described by partial differential equations.

2 Macroscopic traffic flow models

Our work [1] describes the numerical solution of traffic flows on networks. We solve especially the macroscopic models. Using these models, it is possible to make simulations on big networks with a large number of cars. These models are described by partial differential equations in the form of conservation laws:

$$\frac{\partial}{\partial t}\rho\left(x,t\right)+\frac{\partial}{\partial x}Q\left(x,t\right)=0, \tag{1}$$

where $\rho\left(x,t\right)$ and $Q\left(x,t\right)$ are the unknown traffic density and traffic flow at position x and time t, respectively. Equation (1) must be supplemented by the initial condition $\rho(x,0)=\rho_0(x)$ and $Q(x,0)=Q_0(x)$ and an inflow boundary condition. We have only one equation (1) for two unknowns. Thus, we use the *Lighthill–Whitham–Richards model* (abbreviated LWR) where $Q\left(x,t\right)$ is taken as the equilibrium flow $Q_e(\rho(x,t))$, cf. [1].

Following [2], we consider a complex *network* represented by a directed graph. The graph is a finite collection of directed edges, connected together at vertices. Each vertex has a finite set of incoming and outgoing edges. On each road (edge) we consider the LWR model, while at junctions (vertices) we consider a *Riemann solver*.

3 Discontinuous Galerkin method

Due to the character of equation (1), we can expect discontinuity of the traffic density $\rho\left(x,t\right)$. Therefore, for the numerical solution of our models, we choose the *discontinuous Galerkin* (ab-

breviated DG) method, which is essentially a combination of finite volume and finite element techniques, cf. [3].

Consider an interval $\Omega = (a, b)$. Let $\mathcal{T}_h$ be a partition of $\overline{\Omega}$ into a finite number of intervals (elements). We denote the set of all boundary points of all elements by $\mathcal{F}_h$. We seek the numerical solution in the space of discontinuous piecewise polynomial functions $S_h = \{v;\ v|_K \in P^p(K),\ \forall K \in \mathcal{T}_h\}$, where $P^p(K)$ denotes the space of all polynomials on K of degree at most $p \in \mathbb{N}$. For a function $v \in S_h$ we denote the *jump* in the point s as $[v]_s = v^{(L)}(s) - v^{(R)}(s)$, where we use the notation of spatial limits $v^{(L)}(s) := \lim_{x \to s_-} v(x)$ and $v^{(R)}(s) := \lim_{x \to s_+} v(x)$.

The DG formulation of equation (1) then reads: Find $\rho_h : [0, T] \to S_h$ such that

$$\int_\Omega (\rho_h)_t \varphi \,\mathrm{d}x - \sum_{K \in \mathcal{T}_h} \int_K Q_e(\rho_h) \varphi_x \,\mathrm{d}x + \sum_{s \in \mathcal{F}_h} H(\rho_h^{(L)}, \rho_h^{(R)})\, [\varphi]_s = \int_\Omega g\varphi \,\mathrm{d}x,$$

for all $\varphi \in S_h$. In the boundary terms on $\mathcal{F}_h$ we use the approximation $Q_e(\rho_h) \approx H(\rho_h^{(L)}, \rho_h^{(R)})$, where H is a *numerical flux*. We use the *Godunov flux*, cf. [4]:

$$H(u_h^{(L)}, u_h^{(R)}) = \begin{cases} \min_{u_h^{(L)} \le u \le u_h^{(R)}} f(u), & \text{if } u_h^{(L)} < u_h^{(R)}, \\ \max_{u_h^{(R)} \le u \le u_h^{(L)}} f(u), & \text{if } u_h^{(L)} \ge u_h^{(R)}. \end{cases} \tag{2}$$

4 Implementation

For time discretization of the DG method we use the *forward Euler method*. As a basis for S_h, we use *Legendre polynomials*. We use *Gauss–Legendre quadrature* to evaluate integrals over elements. The implementation is in the C++ language.

Because we calculate physical quantities (density and velocity), the result must be in some interval, e.g. $\rho \in [0, \rho_{\max}]$. Thus, we use *limiters* in each time step to obtain the solution in the admissible interval. Here it is important not to change the total number of cars. Following [4], we also apply limiting to treat spurious oscillations near discontinuities and sharp gradients in the numerical solution.

All the above was performed on networks. Thus, we had to deal with the problem of boundary conditions at the junctions. In [1] we introduce our own approach to boundary conditions at junctions, which uses special numerical flux choices. This approach is new and the behavior of the resulting model can be interpreted as the introduction of turning lanes in front of the junction. This is a different approach to the models in [2] and [5], which correspond to single-lane roads where overtaking is prohibited. Moreover, the presented construction of the traffic flux at junctions allows the simulation of arbitrary traffic light combinations instead of only full red/green lights as in [2] and [5].

We prove several important properties of our proposed numerical scheme, such as a discrete analogue to the Rankine–Hugoniot conditions for the numerical fluxes at the junction, conservation property of the DG scheme and traffic distribution error, cf. [1, Lemma 2, Theorems 1 and 2].

5 Numerical results

In this section we demonstrate how our program computes traffic on networks. We define the simple network from Fig. 1. This network is closed. We have three roads and two junctions. The

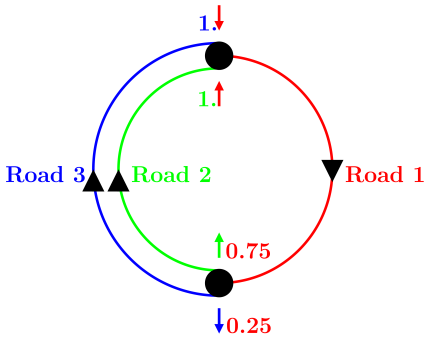
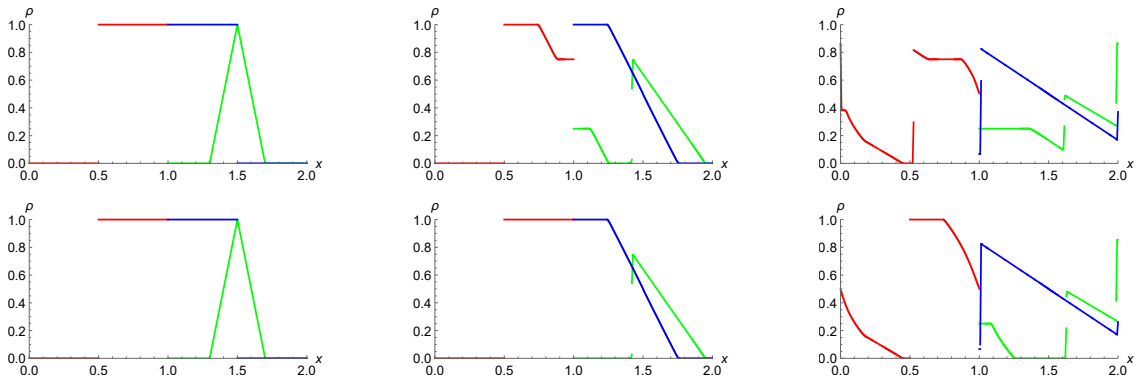


Figure 1: Test network with Road 1, Road 2 and Road 3.

length of all roads is 1. At the first junction we have one incoming and two outgoing roads. At the second junction we have the opposite situation. We use a different distribution of cars at the first junction: $\frac{3}{4}$ go from the first road to the second and $\frac{1}{4}$ from the first road to the third. The initial condition is shown in Figure 2a. We use LWR (specifically the *Greenshields model*, see [1]) on all roads.

We compare our approach with that of [5] which uses the maximum possible flux. In both approaches we use the Godunov flux (2), the forward Euler method with the step size $\tau = 10^{-4}$ and the number of elements is $N = 150$ on each road. A right of way parameter q must be prescribed for the junction with two incoming roads in the case of the maximum possible flux, cf. [2, Section 5.2.2]. We use $q = 0.5$, so the roads are equal. In our approach, we do not have a defined right of way (in the sense of yielding rules at main or side roads), so the roads are equal as well. We can see the comparison in Figure 2. Our approach is in the top row while the approach using the maximum possible flux is in the bottom row. We point out the different behavior in both junctions.

First, we notice the first junction with one incoming and two outgoing roads, i.e. $x = 1$ in the figures. The approach using the maximum possible flux through the junction is zero up to time $t = 0.5$ because one of the outgoing roads (Road 3) reaches the maximal traffic density, hence the flux is zero (traffic jam) cf. Figure 2b. Our approach has nonzero traffic flow through this junction for $t \in [0, 0.5]$ because the numerical flux is nonzero between Road 1 and Road 2 allowing flows between these two roads. For times $t > 0.5$, the maximal traffic density is not attained on Road 3 and the traffic flow is nonzero through the junction in both cases, cf. Figure 2c. If we



(a) $t = 0$. (b) $t = 0.25$. (c) $t = 0.75$.

Figure 2: Comparison of network with Road 1, Road 2 and Road 3. Top column – Our approach using numerical flux. Bottom column – Approach using maximum possible flux.

compare both approaches, we see completely different results on Roads 1 and 2 while the results on Road 3 are almost identical.

Now we focus on the second junction with two incoming and one outgoing road, i.e. $x = 0$ which is identified with $x = 2$ in the figures, due to periodicity. At first glance, there is no difference between the two approaches. Let's compare $\rho_1^{(R)}(0,1)$, i.e. the limit from the right of traffic density on the outgoing Road 1 at $x = 0$ and $t = 1$. Our approach gives us $\rho_1^{(R)}(0,1) \approx 0.4$ while the approach using the maximum possible flux gives us $\rho_1^{(R)}(0,1) \approx 0.5$, which is the maximal traffic flow. The reason for this difference is that we do not have a defined right of way in our approach. Road 2 and Road 3 push too many cars into the junction congesting it slightly. The approach using the maximum possible flux takes into account the whole situation and selects the best solution for both roads. From a real point of view, this approach could be viewed as simulating the behavior of communicating autonomous vehicles which optimize the traffic situation globally, while our approach could be interpreted as simulating the behavior of human drivers without the right of way.

6 Conclusion

We have presented an overview of our paper [1] and demonstrated the numerical solution of macroscopic traffic flow models on network using the discontinuous Galerkin method. On individual roads, we use the Godunov numerical flux, while on junctions, we construct a new numerical flux based on the preferences of drivers. We compare our approach with the paper [5] by Čanić, Piccoli, Qiu and Ren, where Runge-Kutta methods are used along with a different choice of numerical fluxes at junctions. We discuss the differences between the two approaches, where that of [5] corresponds to single-lane roads with a strict enforcement of a priori traffic distribution, while the presented approach corresponds to having dedicated turning-lanes and/or flexibility of the drivers' preferences in extreme situations such as congestions. In future work, we would like to implement right of way rules (with regard to main and side roads) into the numerical flux.

Acknowledgement: The work of L. Vacek is supported by the Charles University, project GA UK No. 1114119. The work of V. Kučera is supported by the Czech Science Foundation, project No. 20–01074S.

References

[1] L. Vacek, V. Kučera: *Discontinuous Galerkin method for macroscopic traffic flow models on networks*. Comm. Appl. Math. Comput. (submitted), arXiv: 2011.10862.

[2] M. Garavello, B. Piccoli: *Traffic flow on networks*, AIMS Series on Applied Mathematics, 2006.

[3] V. Dolejší, M. Feistauer: *Discontinuous Galerkin Method â€“ Analysis and Applications to Compressible Flow*. Springer, Heidelberg, 2015.

[4] C.-W. Shu: *Discontinuous Galerkin methods: general approach and stability*. Numerical solutions of partial differential equations, Birkhäuser Basel, 201, 2009.

[5] S. Čanić, B. Piccoli, J. Qiu, T. Ren: *Runge–Kutta Discontinuous Galerkin Method for Traffic Flow Model on Networks*. Journal of Scientific Computing 63, 2014.

Multilevel methods with inexact solver on the coarsest level

P. Vacek, Z. Strakoš

Charles University, Faculty of Mathematics and Physics, Prague

The analysis of the convergence behavior of the multilevel methods is in the literature typically carried out under the assumption that the problem on the coarsest level is solved exactly. This assumption is, however, not satisfied in practical computation either due to the use of an iterative solver on the coarsest level, or due to the finite precision arithmetic, or both. In this talk we present an abstract description of the multilevel methods which allows inexact solve on the coarsest level and discuss its convergence behavior. In particular, we show that even under these weaker assumptions it is still possible to derive a bound on the rate of convergence, which is independent of the number of levels. Further, we consider application of the multilevel methods to the elliptic partial differential equations and their finite element discretization. We discuss both exact and inexact solvers on the coarsest level. We show that the convergence behavior of the multilevel method with inexact solver on the coarsest level may depend on the mesh size of the initial triangulation.

Numerical approximation between fluid flow and a vibrating airfoil

O. Winter, P. Sváček

Czech Technical University, Faculty of Mechanical Engineering, Prague

1 Introduction

The usage of the finite volume (FV) and the finite element (FE) methods is common approach in the technical practice for the numerical simulations of the fluid flow problems. Both FV and FE methods were used to solve multi physical problems such as fluid-structure interaction, see e.g. [4, 6]. Another option being the discontinuous Galerkin (DG) method which uses ideas of both the FV and the FE methods. The DG method is based on piecewise polynomials but discontinuous approximations, which provides robust numerical processes and high-order accurate solutions. For an overview of DGM see e.g. [1] and references inside. This work presents the numerical solutions of a selected cases of the inviscid gas dynamics approximated with aid of high order discontinuous Galerkin method.

2 Mathematical Model

Mathematical model consists of formulation of the initial-boundary value problem describing the interaction of the fluid flow with an oscillating airfoil. The formulation consists of the Eulerâ€™s equations and formulae prescribing a motion of an airfoil. In order to enable the computations on the moving domain, the arbitrary Lagrangian-Eulerian (ALE) formulation of the governing equations is given.

Let us denote the computational domain $\Omega_t \subset \mathbb{R}^2$ occupied by the fluid at time instant t, see Figure 1. The boundary of Ω_t is decomposed into distinct parts $\partial\Omega_t = \Gamma_{\mathrm{I}} \cup \Gamma_{\mathrm{O}} \cup \Gamma_{\mathrm{W}} \cup \Gamma_{\mathrm{W}t}$, where Γ_{I} is the inlet part, Γ_{O} is the outlet part, Γ_{W} is the static wall, and $\Gamma_{\mathrm{W}t}$ is the moving wall, of the boundary $\partial\Omega_t$. The inviscid gas dynamics in computational domain Ω_t, described

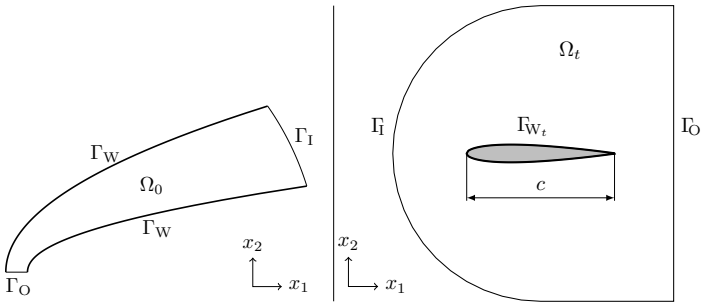


Figure 1: Sketch of the computational domains for the two considered cases, i.e., Ringleb case and fluid flow around the NACA 0012 profile.

by the Euler's equations, i.e., a set of three coupled nonlinear conservation laws, see [1], in the

ALE conservative form written in vector form reads

$$\frac{1}{J}\frac{D^{\mathcal{A}}(\boldsymbol{q}J)}{Dt} + \frac{\partial \boldsymbol{F}^{\mathcal{A}}}{\partial x_1} + \frac{\partial \boldsymbol{G}^{\mathcal{A}}}{\partial x_2} = \boldsymbol{0}, \tag{1}$$

where $\boldsymbol{q} = [\varrho, \varrho v_1, \varrho v_2, \varrho e]^T$ is the state vector composed of the conserved variables, i.e., the density ϱ, the momentum $\varrho \boldsymbol{v}$, and the total energy ϱe and

$$\boldsymbol{F}^{\mathcal{A}}(\boldsymbol{q}, \boldsymbol{w}) = \begin{bmatrix} \varrho(v_1 - w_1) \\ \varrho v_1(v_1 - w_1) + p \\ \varrho v_2(v_1 - w_1) \\ \varrho e(v_1 - w_1) + p v_1 \end{bmatrix}, \quad \boldsymbol{G}^{\mathcal{A}}(\boldsymbol{q}, \boldsymbol{w}) = \begin{bmatrix} \varrho(v_2 - w_2) \\ \varrho v_1(v_2 - w_2) \\ \varrho v_2(v_2 - w_2) + p \\ \varrho e(v_2 - w_2) + p v_2 \end{bmatrix}. \tag{2}$$

are two nonlinear fluxes. The symbols $\boldsymbol{v}$, and p denote the velocity vector with components v_1, v_2, and the pressure. The symbol $D^{\mathcal{A}}/Dt$ in equation (1) denotes so-called ALE derivative defined as

$$\frac{D^{\mathcal{A}} f}{Dt} = \frac{\partial f}{\partial t} + \mathrm{grad}(f) \cdot \boldsymbol{w}. \tag{3}$$

and $\boldsymbol{w} = (w_1, w_2)$ is the domain velocity.

The system needs to be closed by a constitutive equation. The fluid is assumed to be ideal gas, i.e, the internal energy ε and the pressure are related through the equation of the state for the ideal gas. The total energy of the gas is the sum of the internal energy and kinetic energy, i.e., $e = \varepsilon + \frac{v_k v_k}{2}$ where for the ideal gas the internal energy is $\varepsilon = c_v \vartheta$. Here c_v is the specific heat at constant volume and ϑ is the thermodynamic temperature. The equation of the state is $p = (\gamma - 1)\varrho\varepsilon$, where γ is the adiabatic index. The local speed of sound is then defined as $c = \sqrt{\gamma p/\varrho}$.

The system (1) is hyperbolic, it is equipped with the initial condition $\boldsymbol{q}(\boldsymbol{x}, 0) = \boldsymbol{q}_0(\boldsymbol{x}), \quad \boldsymbol{x} \in \Omega_0$, and boundary conditions chosen in such a way that problem (1) is well-posed (see, e.g. [2]). Inlet Γ_{I} and outlet Γ_{O} boundary conditions are determined according to a regime of flow, i.e., subsonic/supersonic (see, e.g. [2]). We assume three types of boundary conditions:

1. Subsonic inlet: the direction of the velocity (given by the inlet angle), the value of the stagnation density ϱ_0 and the stagnation pressure p_0 are prescribed.

   Necessary quantities for evaluation of the speed of sound c and velocity magnitude $|\boldsymbol{v}|$ are extrapolated. Using $|\boldsymbol{v}|$ and c the static pressure p_I and static density ϱ_I and other variables are computed using the relation between the stagnation and the static quantities, i.e.

   $$p_0 = p_I \left(1 + \frac{\gamma - 1}{2}\frac{|\boldsymbol{v}|^2}{c^2}\right)^{\frac{\gamma}{\gamma-1}} \qquad \text{and} \qquad \varrho_0 = \varrho_I \left(1 + \frac{\gamma - 1}{2}\frac{|\boldsymbol{v}|^2}{c^2}\right)^{\frac{1}{\gamma-1}}. \tag{4}$$

2. Subsonic outlet: the conservative variable ϱe is prescribed using constitutive equation through given pressure p_O. Other quantities are extrapolated.

3. Solid wall: reflective boundary condition for all quantities, see [3].

3 Numerical Method

We define a partition $\mathcal{T}_h$ (triangulation) of the closure of the computational domain Ω_t into a finite number of closed simplexes (cells) D_k, $k \in \mathcal{M}_h$, with mutually disjoint interiors, where $\mathcal{M}_h$

is an index set, such that $\bar{\Omega} = \bigcup_{k \in \mathcal{M}_h} D_k$. The solution $\boldsymbol{q}(\boldsymbol{x}, t)$ is approximated by $\boldsymbol{q}_h(\boldsymbol{x}, t) \in \boldsymbol{V}_h(\Omega, \mathcal{T}_h)$, $\boldsymbol{V}_h(\Omega, \mathcal{T}_h) = \{\boldsymbol{q} \in L^2(\Omega); \boldsymbol{q}|_{D_k} \in \boldsymbol{P}_N(D_k), \forall D_k \in \mathcal{T}_h\}$, where $\boldsymbol{P}_N(D_k)$ is the space of all polynomials of degree $\leq N$ on D_k. The volume and surface integrals are realized via Gauss quadrature formulas with special treatment of the curved elements forming the boundary, see e.g. [3]. The time integration is done with aid of the second order strong stability preserving Runge-Kutta scheme.

4 Numerical Results

4.1 Ringleb's Flow

This case considers transonic Ringleb's flow. Ringleb's flow is an exact solution to the Euler's equations for $\gamma = 1.4$ obtained by Ringleb in 1940. The subsonic inlet is assumed: the stagnation pressure $p_0 = 100000$, the stagnation density $\varrho_0 = 1.1684$, and the inlet angle $\alpha_I = \alpha_I(x), x \in \Gamma_\mathrm{I}$, obtained from exact solution and outlet is assumed to be subsonic: the pressure $p_O = p(x), x \in \Gamma_0$, obtained from the exact solution. Figure 2 shows the contours of density for different order of polynomials $N = 1, 2, 10$. One can see tremendous improvement between the numerical solution for $N = 1$ and the numerical solution for $N = 2$ and for $N = 10$ the numerical solution is very close to the exact solution.

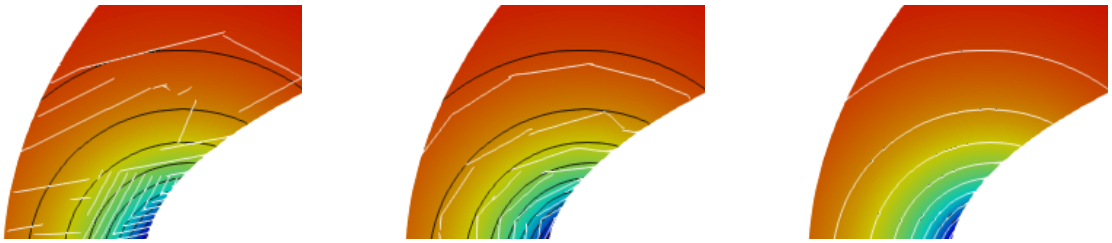


Figure 2: Details of contours of density for different order of polynomials. Black lines indicate exact solution and the white lines indicate the numerical solution. (Left) $N = 1$, (Middle) $N = 2$, (Right) $N = 10$.

4.2 Flow around Vibrating NACA 0012 Profile

This subsection presents the results of the computation of the flow past the oscillating NACA 0012 profile. The following notations is used. The length of the cord is denoted by c and the frequency of the oscillation by f. The motion of the airfoil is prescribed by the harmonic motion, i.e., $\alpha(t) = \alpha_0 + \Delta\alpha \sin(2\pi f t)$. The pressure coefficient $c_p = \frac{p - p_\infty}{0.5 U_\infty^2}$, index ∞ denotes the inlet average, is decomposed according to $c_p = c_{p\mathrm{mean}} + c_p' \sin(2\pi f t) + c_p'' \cos(2\pi f t)$, where $c_{p,\mathrm{mean}}$ is the time-mean values of the c_p, c_p' is the real part of the c_p and c_p'' is the imaginary part of the c_p. The series of calculations was setup as follows. The initial angle of attack $\alpha_0 = 0\,\mathrm{deg}$, the oscillation amplitude $\Delta\alpha = 1\,\mathrm{deg}$ and the frequency of the oscillation $f = 30\mathrm{Hz}$, position of the elastic axis $x_{\mathrm{EA}} = 0.25c$ measured from the leading edge. The free-stream March number $\mathrm{Ma} = 0.4$, and $c = 0.3\mathrm{m}$. Figure 3 shows mean value and real part of the pressure coefficient for different orders of polynomials N, respectively. One can see that lower orders of polynomials do not capture the flow sufficiently for given grid resolution and in case of $N = 5$ the numerical solution is very close to the experimental data presented in [5].

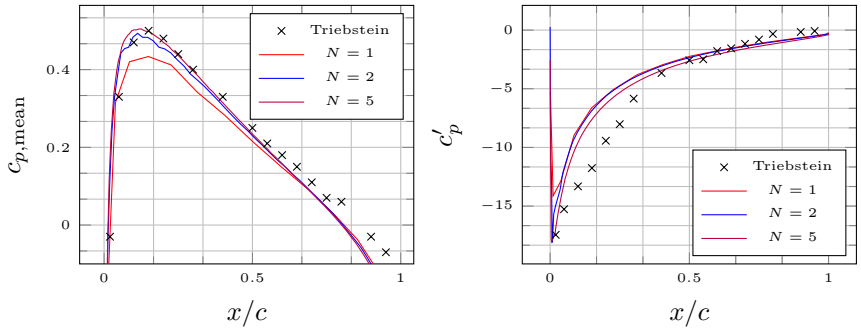


Figure 3: Mean value of the pressure coefficient $c_{p,\mathrm{mean}}$ (Left) and real part of the pressure coefficient c'_p (Right).

5 Conclusion

In this contribution the in-house implementation of the high order discontinuous Galerkin method is used to compute fluid flow problems. Obtained results correspond very well both with theoretical and experimental data.

Acknowledgement: This work was supported by the Grant Agency of the Czech Technical University in Prague, grant SGS19/154/OHK2/3T/12. Authors acknowledge support from the EU Operational Programme Research, Development and Education, and from the Center of Advanced Aerospace Technology (CZ.02.1.01/0.0/0.0/16019/0000826), Faculty of Mechanical Engineering, Czech Technical University in Prague.

References

[1] V. Dolejší, M. Feistauer: *Discontinuous Galerkin Method: Analysis and Applications to Compressible Flow.* Springer Series in Computational Mathematics, Vol. 48, Springer, 2015.

[2] M. Feistauer, J. Felcman, I. Straškraba: *Mathematical and Computational Methods for Compressible Flow.* Clarendon Press, 2003.

[3] J.S. Hesthaven, T. Warburton: *Nodal Discontinuous Galerkin Methods: Algorithms, Analysis, and Applications.* Springer, 2008.

[4] P. Sváček: *Finite Element Approximation of Flow Induced Vibrations of Human Vocal Folds Model: Effects of Inflow Boundary Conditions and the Length of Subglottal and Supraglottal Channel on Phonation Onset.* Applied Mathematics and Computation 319, 2018, pp. 178–194.

[5] H. Triebstein: *Steady and Unsteady Transonic Pressure Distributions on NACA 0012.* Journal of Aircraft 23(3), 1986, pp. 213–219.

[6] O. Winter, P. Sváček: *On Numerical Simulation of Flexibly Supported Airfoil in Interaction with Incompressible Fluid Flow using Laminar-Turbulence Transition Model.* Computers & Mathematics with Applications, 2020.

Winter school lectures

I. Hnětynková:
Regularization of large discrete inverse problems by iterative projection methods

F. Magoulès:
Asynchronous iterative methods:
I – Theory and algorithms
II – Parallel implementation and applications

J. Papež:
On the algebraic error in numerical solution of partial differential equations I and II

I. Pultarová:
Iterative solvers for the stochastic Galerkin method

Regularization of large discrete inverse problems by iterative projection methods

I. Hnětynková

Faculty of Mathematics and Physics, Charles University in Prague

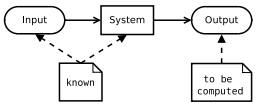
SNA 21 - January 2021

Outline

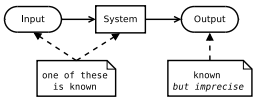
Outline

Basic illustration

Forward Problem



Inverse Problem



Fredholm integral equation

Given the continuous smooth kernel $K(s,t)$ and the (measured) data $g(s)$, the aim is to find the (source) function $f(t)$ such that

$$g(s) = \int_I K(s,t) f(t) \mathrm{d}t + e(s).$$

Fredholm integral has smoothing property, i.e. high frequency components in g are dampened compared to f.

1D example: Barcode reading

Example: Fredholm integral equation - discretization

1D example: Barcode reading

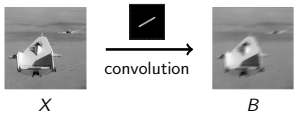
$$g(s) = \int_I K(s,t) f(t) \mathrm{d}t + e(s).$$

Using numerical quadrature formulas, we get a linearized model

$$b = Ax + e, \quad \text{with} \quad A \in \mathbb{R}^{N \times M}, \quad b, e \in \mathbb{R}^N, \quad x \in \mathbb{R}^M,$$

where A has the smoothing property.

2D Example: image deblurring problem



The data $B \in \mathbb{R}^{m\times n}$ are naturally discrete. Using the vectorization $x = vec(X), b = vec(B)$, we obtain a deconvolution problem

$$b = Ax + e, \quad \text{with} \quad A \in \mathbb{R}^{N\times N}, \quad N = mn.$$

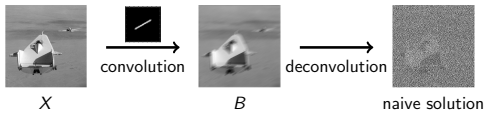
The model matrix is typically large, sparse and structured.

Naive solution

If A is square nonsingular, a naive approach is to solve directly

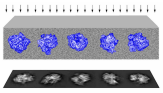
$$Ax^{\text{naive}} = b.$$

2D Example: image deblurring



3D Example: Electron microscopy

$$PSF_\omega * (P_\omega f + e^s_\omega) + e^b_\omega = g_\omega$$



- f: Unknown function representing the particle
- ω: Projection angle.
- PSF_ω: Point Spread Function.
- P_ω: X-Ray transform: $P_\omega f(s) := \int_{-\infty}^{\infty} f(t\cdot\omega + s)\mathrm{d}t, \quad s \in \omega^\perp$.
- e^s_ω, e^b_ω: Structure and background noise functions.
- g_ω: Measured data.
- $*$: Convolution operator.

Discrete model (one projection)

$$PSF_\omega * (P_\omega f + e^s_\omega) + e^b_\omega = g_\omega \quad \text{Continuous model}$$
$$C_\omega\,(\bar{P}_\omega \bar{f} + \bar{e}^s_\omega) + \bar{e}^b_\omega = \bar{g}_\omega \quad \text{Discrete model}$$

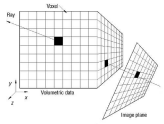


Figure: 3D grid discretization with unknown voxel values.

Linear model

Consider a linear ill-posed problem

$$b = Ax + e,$$

where the noise vector e

- is an unknown perturbation (rounding errors, errors of measurement, noise with physical sources, etc.),
- with the unknown noise level

$$\delta^{\text{noise}} \equiv \|e\|/\|b\| << 1$$

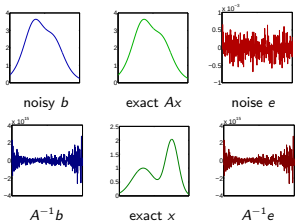
Properties of the problem:

- A dampens high frequencies (smoothing property),
- exact right-hand side is smooth, but noise is not,
- small changes in b cause large changes in the solution.

Naive solution - noise amplification

$$b = Ax + e, \quad \text{where} \quad \|Ax\| \gg \|e\| \quad \textbf{BUT}$$
$$A^{-1}b = x + A^{-1}e, \quad \text{where} \quad \|x\| \ll \|A^{-1}e\|$$

1D Example: `shaw(400)`, $\delta^{\text{noise}} \approx 1e-4$, white noise



Naive solution - noise amplification

Singular value decomposition (SVD): $R = rank(A)$

$$A = U\Sigma V^T = \sum_{j=1}^{R} u_j^T \sigma_j v_j,$$

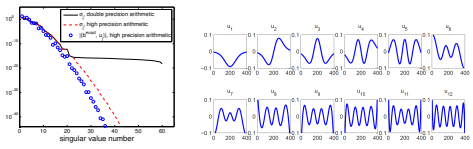
$$\Sigma = \mathrm{diag}\{\sigma_1, \ldots, \sigma_R, 0, \ldots, 0\},$$

where $U = [u_1, \ldots, u_N]$ and $V = [v_1, \ldots, v_M]$ are unitary matrices. Then

$$x^{\text{naive}} \equiv A^\dagger b = \underbrace{\sum_{j=1}^{R} \frac{u_j^T b^{\text{exact}}}{\sigma_j} v_j}_{x^{\text{exact}}} + \underbrace{\sum_{j=1}^{R} \frac{u_j^T e}{\sigma_j} v_j}_{\text{noise component}}.$$

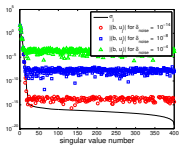
Discrete Picard condition (DPC)

- singular values of A decay quickly without a noticeable gap;
- singular vectors u_i, v_j of A represent increasing frequencies;
- for the exact right-hand side, $|(b^{\text{exact}}, u_j)|$ decay faster than the singular values σ_j of A (**DPC**)



Noise amplification

White noise: $|(e, u_j)|$, $j = 1, \ldots, N$ do not exhibit any trend



$$x^{\text{naive}} \equiv A^\dagger b = \underbrace{\sum_{j=1}^{R} \frac{u_j^T b^{\text{exact}}}{\sigma_j} v_j}_{x^{\text{exact}}} + \underbrace{\sum_{j=1}^{R} \frac{u_j^T e}{\sigma_j} v_j}_{\text{amplified noise}}$$

Components corresponding to small σ_j's are dominated by e^{HF}.

2D imaging problem

For a blurred image B

$$x^{\text{naive}} = \sum_{j=1}^{R} \underbrace{\frac{u_j^T \mathrm{vec}(B)}{\sigma_j}}_{\text{scalar}} v_j, \qquad X = \mathrm{mtx}(x),$$

is a linear combination of right singular vectors v_j.

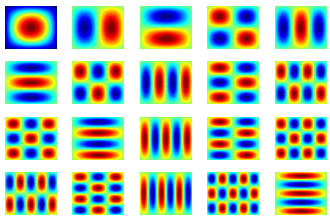
It can be further rewritten as

$$X^{\text{naive}} = \sum_{j=1}^{R} \frac{u_j^T \mathrm{vec}(B)}{\sigma_j} V_j, \qquad V_j = \mathrm{mtx}(v_j) \in \mathbb{R}^{m \times n}$$

using singular images V_j (the reshaped right singular vectors).

2D imaging problem: Singular images

Singular images $V_j \in \mathbb{R}^{m \times n}$ for 2D image deblurring model (Gaußian blur, zero BC, artificial colors).



Filtered solution

Unwanted components can be suppressed by

$$x^{\text{filtered}} = \sum_{j=1}^{R} \phi_j \frac{u_j^T b}{\sigma_j} v_j, \qquad x^{\text{filtered}} = V \Phi \Sigma^{-1} U^T b,$$

where $\Phi = \mathrm{diag}(\phi_1, \ldots, \phi_N)$. In image deblurring problem

$$X^{\text{filtered}} = \sum_{j=1}^{R} \phi_j \frac{u_j^T \mathrm{vec}(B)}{\sigma_j} V_j.$$

The filter factors are given by some filter function

$$\phi_j = \phi(j, A, b, \ldots).$$

Classical regularization approaches

Spectral filtering (e.g., truncated SVD, Tikhonov): suitable for solving small ill-posed problems.

Projection on smooth subspaces: suitable for solving large ill-posed problems. The dimension of projection space represents a regularization parameter.

Hybrid techniques: combination of outer iterative regularization with a spectral filtering of the projected small problem.

... etc.

Large scale problems

- Direct filtering of SVD is too costly.
- The method should avoid work with full A.
- The method should take advantage of data properties (sparsity, structure, ...).
- The approximation must be dominated by low frequencies, high frequencies must be dumped.

We try to look for an approximation in some low dimensional subspace $\mathcal{W}_k$ dominated by low frequencies.

Outline

Projection methods

Consider a subspace

$$\mathcal{W}_k = \mathrm{span}(w_1, \ldots, w_k) \subset \mathbb{R}^N, \qquad W_k = [w_1, \ldots, w_k] \in \mathbb{R}^{N \times k},$$

such that $W_k^T W_k = I_k$ and w_j are dominated by low frequencies.

Then we solve the projected problem

$$\min_{x \in \mathcal{W}_k} \|b - Ax\| \quad \Leftrightarrow \quad \min_{y \in \mathbb{R}^k} \|b - (AW_k)y\|$$

$$\Leftrightarrow \quad W_k^T (A^T A) W_k y = W_k^T A^T b.$$

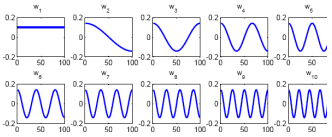
The question is, how to choose the basis w_j?

Projection using DCT basis

An example of a suitable basis is the DCT basis

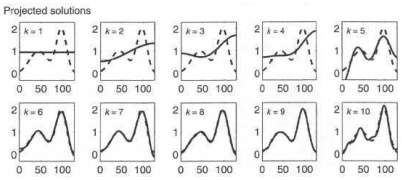
$$w_1 = \sqrt{\frac{1}{N}}\,(1, 1, \ldots, 1)^T,$$
$$w_j = \sqrt{\frac{2}{N}}\left(\cos\left(\frac{(j-1)\pi}{2N}\right), \cos\left(\frac{3(j-1)\pi}{2N}\right), \ldots \cos\left(\frac{(2N-1)(j-1)\pi}{2N}\right)\right)^T,$$

for $j > 1$.



Projection using DCT basis

Example: Solutions computed using the DCT basis $w_1, \ldots, w_k$, $k = 1, \ldots, 10$



A-priori known properties of the true solution (symmetry, periodicity, etc.) can be imposed by well-chosen basis.

Projection using DCT basis

Advantage:

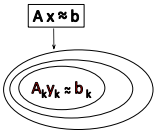
With a fixed set of basis Fourier-type vectors, computations can be performed efficiently, the basis is not stored.

Disadvantage:

The basis vectors are not always adapted to the particular problem.

Krylov subspace methods

When A is large/sparse/not given explicitly, approximation by projection onto a low dimensional Krylov subspace is advantageous.



$$\mathcal{K}_k(C,d) \equiv Span\{d, Cd, \ldots, C^{k-1}d\}$$

$$\mathcal{K}_1(C,d) \subseteq \mathcal{K}_2(C,d) \subseteq \ldots$$

For A square: $\mathcal{K}_k(A,b)$ ... GMRES, CG, MINRES
$\vec{\mathcal{K}}_k(A,b)$ ... RRGMRES, MINRES-II

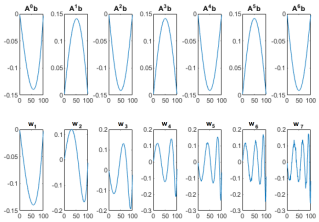
For A general: $\mathcal{K}_k(A^TA, A^Tb)$ ... LSQR, LSMR, CGLS

$$x_k \longrightarrow x^{\text{naive}}$$

Key role of orthonormal basis

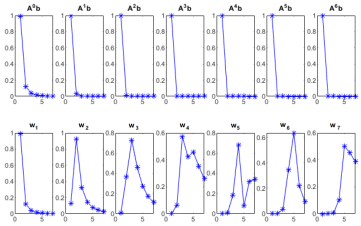
Generating Krylov vectors are smooth. In order to approximate less smooth features, it is necessary to use orthonormal basis.

Example: Generating vectors and orthonormal basis vectors w_i (computed by Arnoldi process) for $\mathcal{K}_5(A,b)$



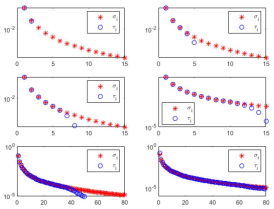
Key role of orthonormal basis

Example: Generating vectors and orthonormal basis vectors w_i in frequency basis U (left singular vectors of A)



Inheritance of DPC

Example: Singular values σ_i of A and singular values τ_i of H_k from the Arnoldi process for $k = 2, 5, 8, 5, 50, 80$



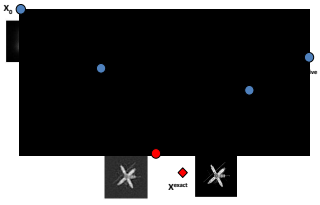
The projected problem $A_k y_k \approx b_k$ then subsequently inherits DPC properties of the original problem.

Semiconvergence of Krylov subspace methods

With growing k:

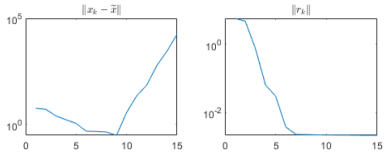
- we include HF features to the solution,
- noise e propagates to the projection.

small k = over-smoothed solution **large** k = noisy solution



Semiconvergence of Krylov subspace methods

Example: True errors and residual norms of LSQR approximations x_k for the problem `shaw(400)` contaminated by white noise e



Number of iterations = regularization parameter

Stopping criteria

Since $b - Ax^{exact} = e$, a reasonable requirement could be

$$r_k \equiv b - Ax_k \quad \approx \quad e.$$

Stopping criteria: this idea can be used if a priori information is available, e.g., $\|e\|$ in DP, spectral properties of e (white) in NCP.

However, e is often not known.

Understanding noise propagation:

- consider $\mathcal{K}_k(A^TA, A^Tb)$ for a general A,
- study how e propagates to the projections,
- study the relation between e and $r_1, r_2, \ldots$.

Outline

Golub-Kahan iterative bidiagonalization (GK)

Given $w_0 = 0$, $s_1 = b/\beta_1$, $\beta_1 = \|b\|$, for $j = 1, 2, \ldots$

$$\begin{aligned} \alpha_j w_j &= A^T s_j - \beta_j w_{j-1}, \quad \|w_j\| = 1, \\ \beta_{j+1} s_{j+1} &= A w_j - \alpha_j s_j, \quad \|s_{j+1}\| = 1. \end{aligned}$$

Output:

- $S_k = [s_1, \ldots, s_k]$ - orthonormal bases of $\mathcal{K}(AA^T, b)$,
- $W_k = [w_1, \ldots, w_k]$ - orthonormal bases of $\mathcal{K}(A^TA, A^Tb)$,
- bidiagonal matrices of the normalization coefficients

$$L_k = \begin{bmatrix} \alpha_1 & & & \\ \beta_2 & \alpha_2 & & \\ & \ddots & \ddots & \\ & & \beta_k & \alpha_k \end{bmatrix}, \quad L_{k+} = \begin{bmatrix} L_k \\ e_k^T \beta_{k+1} \end{bmatrix}.$$

Regularization based on GK

$x_k = W_k y_k$, where the columns of W_k span $\mathcal{K}_k(A^TA, A^Tb)$

LSQR method: minimize the residual

$$\min_{x \in \mathcal{K}_k(A^TA, A^Tb)} \|Ax - b\| = \min_{y \in \mathbb{R}^k} \|L_{k+}y - \beta_1 e_1\|$$

CRAIG method: minimize the error

$$\min_{x \in \mathcal{K}_k(A^TA, A^Tb)} \|x^* - x\| = \min_{y \in \mathbb{R}^k} \|L_k y - \beta_1 e_1\|$$

LSMR method: minimize $A^T r_k$

$$\min_{x \in \mathcal{K}_k(A^TA, A^Tb)} \|A^T(Ax - b)\| = \min_{y \in \mathbb{R}^k} \|L_{k+1}^T L_{k+} y - \beta_1 \alpha_1 e_1\|$$

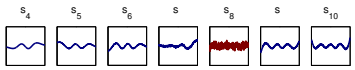
Noise propagation in GK

Recall that we are interested in the relation between

$$\tilde{r} \equiv b - A\tilde{x} \quad \longleftrightarrow \quad e.$$

Since $x_k = W_k y_k \in \mathcal{K}_k(A^TA, A^Tb)$, then

$$r_k \equiv b - AW_k y_k = \beta_1 s_1 - S_{k+1} L_{k+} y_k = S_{k+1} p_k \in \mathcal{K}_k(AA^T, b).$$



Analyzed in [H., Plešinger, Strakoš - 09], [H., Plešinger, Kubínová - 17].

Exact and noise component in s_k

- $s_1 = b/||b|| = Ax/||b|| + e/||b||$
- for $k = 2, 3, \ldots$

$$\beta_{k+1} s_{k+1} = A w_k - \alpha_k s_k$$

Thus

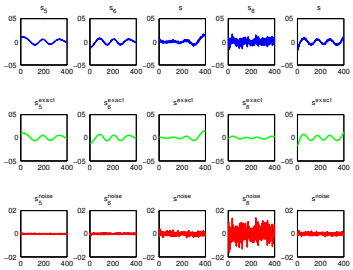
$$s_k = (\cdot) + \gamma_k e^{HF}, \quad \text{where} \quad \gamma_k \equiv \varphi_{k-1}(0) = (-1)^{k-1} \frac{1}{\beta_k} \prod_{j=1}^{k-1} \frac{\alpha_j}{\beta_j}.$$

Here $(\cdot)$ is smooth and the amplification factor γ_k of e^{HF} is the absolute term of the Lanczos polynomial,

$$s_{k+1} = \varphi_k(AA^T)b, \qquad \varphi_k \in \mathcal{P}_k.$$

Exact and noise component in s_k

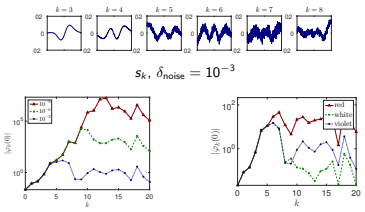
$$s_k = s_k^{exact} + s_k^{noise}$$



Noise propagation in GK - behavior

The size of γ_k (on average) rapidly grows until it reaches the noise revealing iteration k_{rev}. Then it decreases.

Example: `shaw(400)`, reortogonalization in GK

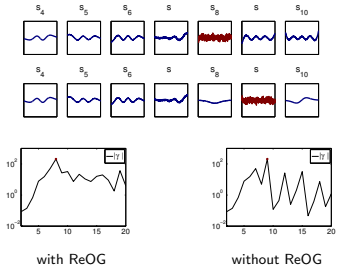


s_k, $\delta_{\text{noise}} = 10^{-3}$

white noise colored noise

Influence of the loss of orthogonality

Comparison GK with and without reorthogonalization:

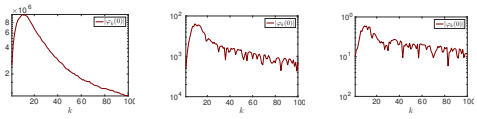


with ReOG without ReOG

Aggregation may be necessary [Gergelits, H., Kubínová - 18].

Noise propagation in GK - large 2D problems

Example: $\delta_{\text{noise}} \approx 10^{-2}$, various physical noise, without ReOG

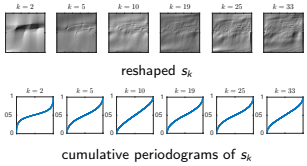


vargaussianblur, $N = 99856$ seismictomo, $N = 20000$ paralleltomo, $N = 65160$

There is no particular noise revealing iteration k, but rather a noise revealing phase represented by a group of subsequent iterations k, see [H., Plešinger, Kubínová - 17].

Noise propagation in GK - large 2D problems

Example: `seismictomo`, $\delta_{\text{noise}} \approx 10^{-2}$, without ReOG



reshaped s_k

cumulative periodograms of s_k

Cumulative periodogram (examining distribution of frequencies) of s_{10} is flatter, thus s_{10} belong to the noise revealing phase.

Application in regularization process

- Stopping criterion - before noise propagates seriously to s_k.
- If k_{rev} can be identified, we can estimate the high frequency part of e:

$$s_{k_{\text{rev}}} \equiv (\cdot) + \gamma_{k_{\text{rev}}} e^{HF} \approx \gamma_{k_{\text{rev}}} e^{HF}$$

gives the estimate by scaled left bidiagonalization vector

$$\tilde{e} \equiv \gamma_{k_{\text{rev}}}^{-1} s_{k_{\text{rev}}}.$$

- We can obtain a cheap estimate of the unknown noise level $\| e \| / \| b \|$, see [H., Kubínová, Plešinger - 16] for application in image deblurring.

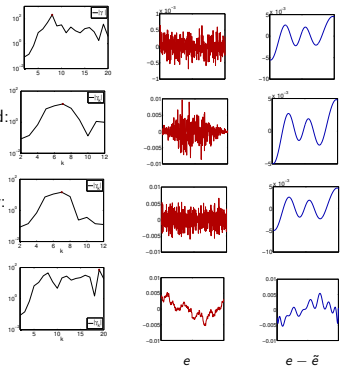
Noise estimate for `shaw(400)`

White:

Data correlated:

High-frequency:

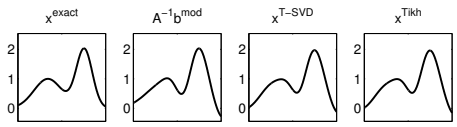
Low-frequency:



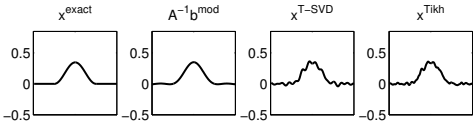
e $\quad$ $e - \tilde{e}$

Comparison of noise reduction to spectral filtering

`shaw(400)`, white noise



`phillips(400)`, white noise



Outline

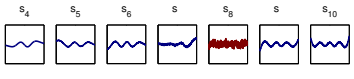
Regularization based on GK

Recall that we are interested in the relation between

$$\tilde{r} \equiv b - A\tilde{x} \qquad \longleftrightarrow \qquad e.$$

For GK based methods with $x_k = W_k y_k \in \mathcal{K}_k(A^T A, A^T b)$, we have

$$r_k = S_{k+1} p_k.$$



Based on noise propagation in S_k, we can analyze CRAIG, LSQR, LSMR by studing p_k, see [H., Kubínová, Plešinger - 17].

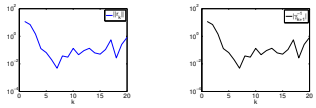
Residual of CRAIG method

$$\min_{x \in \mathcal{K}_k(A^T A, A^T b)} \| x^* - x \| = \min_{y \in \mathbb{R}^k} \| L_k y - \beta_1 e_1 \|, \quad x_k = W_k y_k$$

Theorem: x_k^{CRAIG} is the exact solution to the consistent system

$$A x_k^{\text{CRAIG}} = b - \varphi_k(0)^{-1} s_{k+1}.$$

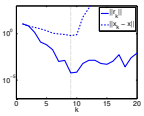
Consequently, $\| r_k^{\text{CRAIG}} \| = |\varphi_k(0)^{-1}| \equiv |\gamma_{k+1}|^{-1}$ reaches its minimum in the noise revealing iteration.



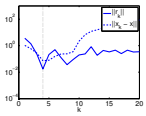
CRAIG residual $\quad$ inverted ampl. factor

Comparison of the error and the residual

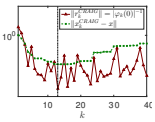
Measuring the size of the residual seems to be a valid stopping criterion for CRAIG. The minimal error is reached approximately at the iteration with the minimal residual.



`shaw(400)`



`phillips(1000)`



`phillips`, no ReOG

Residual of LSQR method

$$\min_{x\in\mathcal{K}_k(A^TA,A^Tb)} \|Ax-b\| = \min_{y\in\mathbb{R}^k} \|L_{k+}y-\beta_1 e_1\|\,, \quad x_k = W_k y_k$$

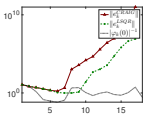
Theorem: The residual corresponding to x_k^{LSQR} satisfies

$$r_k^{\mathrm{LSQR}} = \frac{1}{\sum_{l=0}^{k}\varphi_l(0)^2}\sum_{l=0}^{k}\varphi_l(0)s_{l+1}.$$

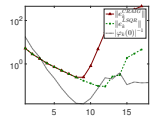
Consequently, the size of the component of r_k in the direction of s_j is proportional to the amount of propagated noise e^{HF} in s_j.

Comparison of CRAIG and LSQR

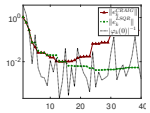
Typically, LSQR can reach better approximation than CRAIG.



`shaw(400)`, white



`gravity(400)`, Poisson

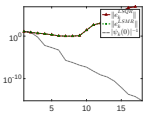


`phillips(400)`, white, no ReOG

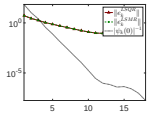
Residual of LSMR method

$$\min_{x\in\mathcal{K}_k(A^TA,A^Tb)} \|A^T(Ax-b)\| = \min_{y\in\mathbb{R}^k} \|L_{k+1}^T L_{k+}y-\beta_1\alpha_1 e_1\|$$

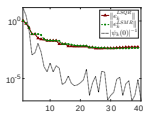
Components of r_k in LSMR behave similarly as in LSQR. The errors resemble as long as $|\psi_k(0)|$ (the absolute term of the Lanczos polynomial for GK vectors w_k) grows rapidly.



`shaw(400)`, white

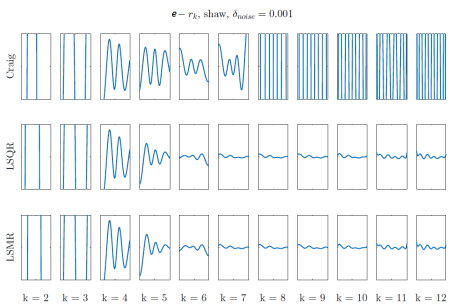


`gravity(400)`, Poisson



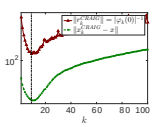
`phillips(400)`, white, no ReOG

Comparison of noise and residuals

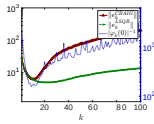


Comparison of the methods - large 2D problems

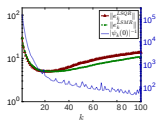
Example: `seismictomo(100,100,200)`, white noise, $\delta_{\mathrm{noise}} = 0.01$, $A \in \mathbb{R}^{20000\times 10000}$, no ReOG



CRAIG error and residual



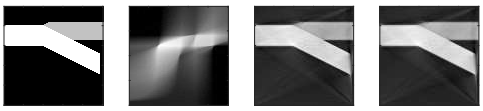
CRAIG vs LSQR



LSQR vs LSMR

Comparison of reconstructions

Reconstructions for `seismictomo(100,100,200)`. Iteration is selected as $k = \text{argmax}_{k=1,2,\ldots} |\varphi_k(0)|$.



exact solution | noisy projections | LSQR | LSMR

Outline

Basic idea

Two stage inner (Krylov projection) - outer (direct) regularization.

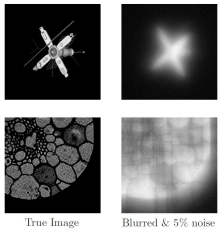
Algorithm: Hybrid LSQR

- Golub-Kahan iterative bidiagonalization
  - $L_{k+}y_k \approx \beta_1 e_1$
- Tikhonov regularization of the projected problem
  - $y_k^\lambda = \arg\min_y \{\|L_{k+}y - \beta_1 e_1\|_2^2 + \lambda^2 \|y\|_2^2\}$
  - Parameter selection approach.
- Back projection $x_k^\lambda = W_k y_k^\lambda$
- Stopping criterion.

See [Calvetti, Reichel - 03], [Chung, Nagy, O'Leary - 08], [Kilmer, Hansen, Español - 07], [Renaut, H., Mead - 10], [Chung, Palmer - 15], ...

2D image deblurring

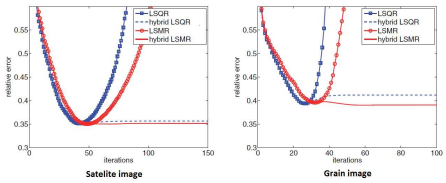
Examples: Satelite and grain test image, Gaussian blur, white noise with $\delta_{\text{noise}} = 0.05$.



True Image | Blurred & 5% noise

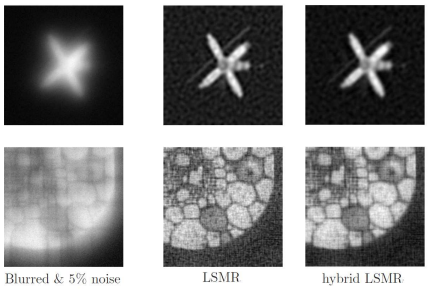
2D image deblurring

Example: LSQR and LSMR with inner Tikhonov regularization



- overcomes the semiconvergence phenomenon,
- two regularization parameters (outer - number of iterations, inner - direct regularizer) must be tuned.

2D image deblurring - reconstructions



Blurred & 5% noise | LSMR | hybrid LSMR

Outline

Conclusion

- Iterative projective regularization is a powerfull tool to solve large problems.
- Noise propagates subsequentially to the projections, early stopping is necessary.
- Combinations of projection and direct regularization is advantageous.
- Constraints (e.g. nonnegativity or sparsity of the solution) can be incorporated.

Selected references

Software:

- S. Gazzola, P. C. Hansen, and J. G. Nagy: IR Tools Version 1.0, 2019.
- P. C. Hansen, and J. S. Jrgensen: AIR Tools II Version 1.0, 2018.
- P. C. Hansen: Regularization Tools Version 4.0, 2007.

Overview books:

- R C. Gonzalez, R. E. Woods: Digital Image Processing, Pearson, 4th Edition 2018.
- M. Hanke: A Taste of Inverse Problems: Basic Theory and Examples, SIAM, 2017.
- P. C. Hansen: Discrete Inverse Problems: Insight and Algorithms, SIAM, 2010.
- P. C. Hansen, J. G. Nagy, and D. P. O'Leary: Deblurring Images: Matrices, Spectra, and Filtering, SIAM, 2006.

Thank you for your attention!

Asynchronous iterative methods:

I – Theory and algorithms

II – Parallel implementation and applications

F. Magoulès

Université Paris-Saclay

The traditional scheme for parallel iterative algorithms is a synchronous method where a new iteration is only started when all the data from the previous one has been received. Such algorithms meet serious scalability limitation due to the synchronization procedure occurring between the processors at the end of each iteration. Another kind of iterative scheme, called asynchronous iterations, can help solve these scalability problems, but lead to several convergence issues, as presented in the first talk.

Modifying an iterative scheme to make asynchronous iterations leads to several convergence difficulties, but the implementation of such methods is also a challenge. Indeed, two different parallel executions will lead to different numbers of iterations, and the asynchronous behavior will make difficult the computation of any stopping criteria. These aspects are discussed in the second talk and illustrated on numerical experiments performed in parallel on large scale engineering problems. Besides, the programming library developed is proved stable and powerful for the implementation of any asynchronous iterative methods.

On the algebraic error in numerical solution of partial differential equations

J. Papež

Institute of Mathematics of the Czech Academy of Sciences, Prague

In the lectures we will present some difficulties and results in the numerical solution of algebraic systems stemming from the discretization of partial differential equations (PDE).

In the first part, we will discuss several topics that are chosen to illustrate the role of algebraic solvers in the overall solution procedure and to emphasize the interconnections with other phases of the solution, such as discretization and preconditioning. We will try to point out that an efficient procedure requires thorough understanding of mutual relationships and close interaction between all the phases.

We will use simple examples to illustrate some risks of using inappropriate error measures or error estimates as well as to demonstrate possible differences between the errors of different origin. We will see that the algebraic and discretization errors can have very different spatial distribution over the computational domain. This provides a challenge for (not only) estimating the errors.

Then we will discuss a widely used residual-based error estimator that has been derived for the Galerkin solution, i.e. asuming the exact solution of the associated algebraic system. We will see that a generalization of the estimator in order to be used for computed, inexact approximation requires a careful analysis and it results in a form with an additional unknown multiplicative factor. We will also illustrate how an adaptive mesh refiment based on this error estimator can be affected when the estimator is evaluated for the Galerkin solution or a computed approximation.

Recalling briefly the result of Málek and Strakoš [1], we will show that any algebraic preconditioning can be interpreted as a transformation of the discretization basis and, at the same time, transformation of the inner product in the given (infinite-dimensional) Hilbert space. This underlines that discretization and preconditioning are tightly coupled.

We will also present an idea of backward interpretation of the algebraic error in the context of numerical solution of PDEs. In numerical linear algebra, the backward error analysis is a widely used concept of representing the inexactness of the computed approximation as a perturbation of the original system. When solving a system arising from a finite element discretization of a PDE, we can ask ourselves if we can represent the (algebraic) inexactness as a transformation of the discretization basis. On a simple example, we will see that such transformed basis can lose its locality, i.e. the supports of transformed basis functions can be over the entire discretization domain.

Finally, to end the first part with some more positive results, we will discuss two ideas on how the information given by error estimates can be used to steer and speed-up the computation of a new approximation. First, we will show that a lower bound for the algebraic error and a simple iterative solver can be derived simultaneously using one lifting of the algebraic residual. We have used this to derive an estimator and a multilevel solver that are robust with respect to the polynomial degree of the finite element discretization. Second, we will present an adaptive preconditioner that aims at efficiently reducing the algebraic error in the regions where it was indicated as large.

In the second lecture, we will present an error estimator that uses a unified framework to bound from above the total and the algebraic errors. It involves no unknown or uncomputable factors and allows to estimate the local distribution of the errors in the solution domain. The estimator is based on so-called quasi-equilibrated flux reconstruction and we will detail its construction. The numerical results will be presented to demonstrate very good behaviour of the estimator. On the other hand, not to claim that a question of error estimation is a fully resolved issue, we must admit that the flux reconstruction is a complex procedure and the evaluation of the error estimator can be constly.

References

[1] J. Málek, Z. Strakoš: *Preconditioning and the Conjugate Gradient Method in the Context of Solving PDEs*. SIAM Spotlights, Chapter 8, 2015, https://doi.org/10.1137/1.9781611973846

Iterative solvers for stochastic Galerkin method

Ivana Pultarová

Faculty of Civil Engineering, CTU in Prague

SNA 2021, January 2021

Outline

Problems with parametric (uncertain) data

As an example,

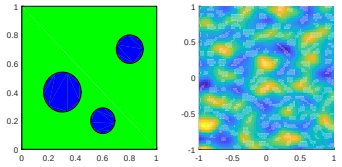
$$-\nabla\cdot a(\boldsymbol{x},\boldsymbol{\xi})\nabla u(\boldsymbol{x},\boldsymbol{\xi}) = f(\boldsymbol{x}), \qquad (\boldsymbol{x},\boldsymbol{\xi})\in D\times\Gamma,$$

with $u(\boldsymbol{x},\boldsymbol{\xi})=0$ on $\partial D\times\Gamma$, data $0<\alpha_1\le a(\boldsymbol{x},\boldsymbol{\xi})\le\alpha_2<\infty$.

Parametric data/random field $a(\boldsymbol{x},\boldsymbol{\xi})$

a) Left:
well separated domains
with different characteristics
$a(\boldsymbol{x},\boldsymbol{\xi})$
$= a_0(\boldsymbol{x})+\chi_1(\boldsymbol{x})\xi_1+\chi_2(\boldsymbol{x})\xi_2$

b) Right:
Karhunen-Loéve expansion,
truncated
(covariance, eigenvectors)



Karhunen-Loéve expansion

Numerical computation needs discrete finite random field $a(\boldsymbol{x},\omega)$.
Covariance operator c,

$$C(g)(\boldsymbol{x})=\int_D c(\boldsymbol{x},\boldsymbol{y})g(\boldsymbol{y})\,\mathrm{d}\boldsymbol{y}, \qquad c(\boldsymbol{x},\boldsymbol{y})=\mathrm{cov}(a(\boldsymbol{x},\omega),a(\boldsymbol{y},\omega))$$

with eigenvalues λ_k and eigenfunctions $a_k(\boldsymbol{x})$, then

$$a(\boldsymbol{x},\omega)=a_0(\boldsymbol{x})+\sum_{k=1}^{\infty}\xi_k(\omega)\sqrt{\lambda_k}a_k(\boldsymbol{x}),$$

where ξ_k are uncorelated random variables with zero mean and unit variance.
Truncation, check $a_{\mathrm{trunc}}(\boldsymbol{x},\omega)>0$.

Measure space $L^2_\rho(\Gamma)$

Doob-Dynkin lemma: measure space $L^2_\rho(\Gamma)$, $\rho(\xi)=\mathrm{d}P/\mathrm{d}\xi$ (instead of (Ω,Σ,P))

$$a(\boldsymbol{x},\omega):=a(\boldsymbol{x},\boldsymbol{\xi}(\omega)), \qquad \boldsymbol{\xi}(\omega)=(\xi_1(\omega),\dots,\xi_{N_\xi}(\omega)),$$

where the random variables $\xi_i(\omega)$ are iid with the joint probability density

$$\rho(\boldsymbol{\xi})=\prod_{i=1}^{N_\xi}\rho_i(\xi_i) \quad\text{and}\quad \Gamma=\prod_{i=1}^{N_\xi}\Gamma_i=\prod_{i=1}^{N_\xi}\mathrm{Im}(\xi_i).$$

Data $a(\boldsymbol{x},\boldsymbol{\xi})$

a) linear w.r.t. stochastic part

$$a(\boldsymbol{x},\boldsymbol{\xi})=a_0(\boldsymbol{x})+\sum_{i=1}^{N_\xi}a_i(\boldsymbol{x})\xi_i, \qquad\text{or}\qquad a(\boldsymbol{x},\boldsymbol{\xi})=a_0(\boldsymbol{x})+\sum_{i=1}^{N_\xi}\chi_i(\boldsymbol{x})\xi_i$$

b) non-linear w.r.t. stochastic part

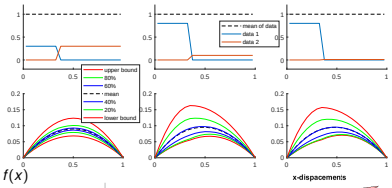
$$a(\boldsymbol{x},\boldsymbol{\xi})=\exp\left(\widetilde{a}_0(\boldsymbol{x})+\sum_{i=1}^{N_\xi}\widetilde{a}_i(\boldsymbol{x})\xi_i\right)=\sum_{j=0}^{N_a}a_j(\boldsymbol{x})p_j(\boldsymbol{\xi}),$$

Stochastic variables / parameters $\boldsymbol{\xi}=(\xi_1,\dots,\xi_{N_\xi})\in\Gamma$

Probability distribution $N(0,1)$, $U(-1,1)$, etc.
Probability density / weight function ρ

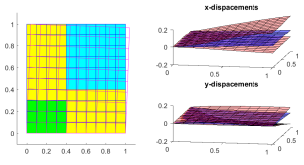
Where we can meet parametric problems

a) Studying
dependency of u
on the data



iso-lines
$-(a(x,\boldsymbol{\xi})u(x,\boldsymbol{\xi})')'=f(x)$
$\boldsymbol{\xi}=(\xi_1,\xi_2)$

iso-surfaces
linear elasticity
stiffness in three domains $2:3:1$
$\boldsymbol{\xi}=(\xi_1,\xi_2,\xi_2)$



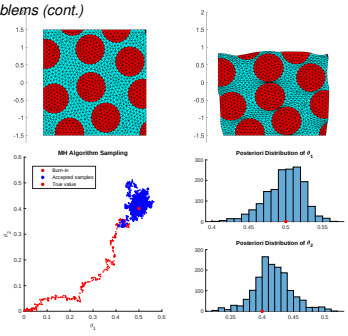
Where we meet parametric problems (cont.)

b) Identifying parameters, inverse problems

linear elasticity
$\boldsymbol{\xi} = (\xi_1, \xi_2)$

Bayes methods
Metropolis-Hastings method
surrogate models

(figures by L.Gaynutdinova)



Blaheta, Béreš, Domesová, Pan, A comparison of deterministic and Bayesian inverse with application in micromechanics, Applications of Mathematics, 2018

Solution methods for parametric problems

a) Monte Carlo methods

+ nonintrusive, multilevel Monte Carlo methods, universal
- time consuming (unless applied in parallel), generating samples, lack of guaranteed error bounds

b) Collocation methods (w.r.t. stochastic variables)

+ nonintrusive, sparse grids, nested grids (Clenshaw-Curtis quadrature)
- curse of dimensionality, discrete approximation measure, lack of guaranteed error bounds

c) Stochastic Galerkin methods / stochastic finite element methods

+ integral approximation measure, various post-processing of results, guaranteed error bounds
- intrusive (unless using double-orthogonal approximation polynomials), large linear systems, coupled problem, curse of dimensionality

Stochastic Galerkin method / stochastic FE method

As example

$$-\nabla \cdot a(\boldsymbol{x},\boldsymbol{\xi}) \nabla u(\boldsymbol{x},\boldsymbol{\xi}) = f(\boldsymbol{x}), \qquad (\boldsymbol{x},\boldsymbol{\xi}) \in D \times \Gamma,$$

with $u(\boldsymbol{x},\boldsymbol{\xi}) = 0$ on $\partial D \times \Gamma$.

Stochastic variational form. Find $u \in V = H_0^1(D) \times L_\rho^2(\Gamma) = L_\rho^2(\Gamma, H_0^1(D))$ (Bochner space) such that

$$\mathscr{A}(u,v) = \mathscr{F}(v), \quad v \in V.$$

Equality of moments.
Energy inner product and linear functional

$$\begin{aligned} \mathscr{A}(u,v) &= \int_\Gamma \int_D \nabla v(\boldsymbol{x},\boldsymbol{\xi}) \cdot a(\boldsymbol{x},\boldsymbol{\xi}) \nabla u(\boldsymbol{x},\boldsymbol{\xi}) \rho(\boldsymbol{\xi}) \mathrm{d}\boldsymbol{x} \mathrm{d}\boldsymbol{\xi} \\ \mathscr{F}(v) &= \int_\Gamma \int_D f(\boldsymbol{x}) v(\boldsymbol{x},\boldsymbol{\xi}) \rho(\boldsymbol{\xi}) \mathrm{d}\boldsymbol{x} \mathrm{d}\boldsymbol{\xi}. \end{aligned}$$

Regularity and a priori convergence estimates
Babuska, Nobile, Tempone, Ghanem, Zouraris; Gittelson, 2009; Bespalov, Powell, Silvester, 2012

Discretization

Solution $u(\boldsymbol{x},\boldsymbol{\xi}) = \sum_{r,j=1}^{N_{\mathrm{FE}},N_{\mathrm{pol}}} u_{(j-1)N_{\mathrm{FE}}+r} \phi_r(\boldsymbol{x}) \Psi_j(\boldsymbol{\xi})$

Approximation basis functions $\phi_r(\boldsymbol{x})\Psi_j(\boldsymbol{\xi}) \in V^{\mathrm{FE}} \times V^{\mathrm{pol}} \subset V$

Finite element basis functions $\phi_r(\boldsymbol{x}) \in V^{\mathrm{FE}} \subset H_0^1(D)$

Orthogonal w.r.t. weight ρ polynomials $\Psi_j(\boldsymbol{\xi}) = \psi_{j1}(\xi_1) \cdots \psi_{jN_\xi}(\xi_{N_\xi}) \in V^{\mathrm{pol}} \subset L_\rho^2(\Gamma)$

Polynomials

Hermite pol. $\rho(\xi) = \frac{1}{\sqrt{\pi}} \mathrm{e}^{-\xi^2/2}$, Legendre pol. $\rho(\xi) = \frac{1}{2}\chi_{\langle -1,1 \rangle}$, etc.

Complete polynomials (total degree $\le d$) $N_{\mathrm{pol}} = \binom{N_\xi + d}{N_\xi}$

Tensor product (degrees $\le d_i$ at ξ_i) $N_{\mathrm{pol}} = \prod_{i=1}^{N_\xi} (d_i + 1)$

Matrices

$$\begin{aligned} \mathsf{A}_{(k-1)N_{\mathrm{FE}}+s,(j-1)N_{\mathrm{FE}}+r} &= \mathscr{A}(\psi_s \Phi_k, \psi_r \Phi_j) \\ &= \int_\Gamma \int_D \nabla \psi_s \Phi_k \cdot a(\boldsymbol{x},\boldsymbol{\xi}) \nabla \psi_r \Phi_j \rho \mathrm{d}\boldsymbol{x} \mathrm{d}\boldsymbol{\xi} \\ &= \int_\Gamma \Phi_k \Phi_j \rho \int_D \nabla \psi_s \cdot a(\boldsymbol{x},\boldsymbol{\xi}) \nabla \psi_r \mathrm{d}\boldsymbol{x} \mathrm{d}\boldsymbol{\xi} \end{aligned}$$

depend on data $a(\boldsymbol{x},\boldsymbol{\xi})$; e.g. (slide 5)

$$\begin{aligned} \mathsf{A}_{(k-1)N_{\mathrm{FE}}+s,(j-1)N_{\mathrm{FE}}+r} &= \int_\Gamma \Phi_k \Phi_j \rho \int_D \sum_{l=0}^{N_a} a_l(\boldsymbol{x}) p_l(\boldsymbol{\xi}) \nabla \psi_s \cdot \nabla \psi_r \mathrm{d}\boldsymbol{x} \mathrm{d}\boldsymbol{\xi} \\ &= \sum_{l=0}^{N_a} \int_\Gamma p_l(\boldsymbol{\xi}) \Phi_k \Phi_j \rho \mathrm{d}\boldsymbol{\xi} \int_D a_l(\boldsymbol{x}) \nabla \psi_s \cdot \nabla \psi_r \mathrm{d}\boldsymbol{x} \\ &= \sum_{l=0}^{N_a} (\mathsf{G}_l)_{kj} \cdot (\mathsf{K}_l)_{sr} \end{aligned}$$

A is s.p.d. (unless, e.g., Gauss distribution and high degree approximation)

$$\mathsf{b}_{(k-1)N_{\mathrm{FE}}+s} = \mathscr{F}(\psi_s \Phi_k) = \int_\Gamma \int_D f \psi_s \Phi_k \rho \mathrm{d}\boldsymbol{x} \mathrm{d}\boldsymbol{\xi}$$

Ernst, Ullmann, 2008

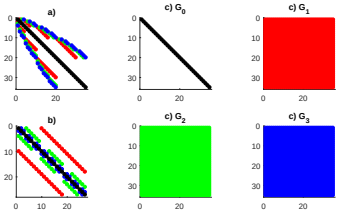
Structures of matrices

Matrix A is not built. Sum of tensor products $\mathsf{A} = \sum_{l=0}^{N_\xi} \mathsf{G}_l \otimes \mathsf{K}_l$

Examples:
a) $a(\boldsymbol{x},\boldsymbol{\xi}) = a_0(\boldsymbol{x}) + \sum_{i=1}^{3} a_i(\boldsymbol{x}) \xi_i$, complete polynomials, $d = 4$
b) $a(\boldsymbol{x},\boldsymbol{\xi}) = a_0(\boldsymbol{x}) + \sum_{i=1}^{3} a_i(\boldsymbol{x}) \xi_i$, tensor product polynomials, $d_1 = d_2 = d_3 = 2$
c) $a(\boldsymbol{x},\boldsymbol{\xi}) = \sum_{i=0}^{3} a_i(\boldsymbol{x}) p_i(\boldsymbol{\xi})$, $N_a = 3$, complete polynomials, $d = 4$

G_0, G_1, G_2, G_3

Every dot is a multiple of an $N_{\mathrm{FE}} \times N_{\mathrm{FE}}$ "stiffness matrix" K_l



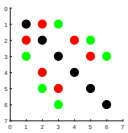
Structure of matrices

Linear $a(\boldsymbol{x},\xi) = a_0(\boldsymbol{x}) + \xi_1 a_1(\boldsymbol{x}) + \xi_2 a_2(\boldsymbol{x})$:

$N_\xi = 2$, uniform distribution of ξ_i, Legendre polynomials, complete pol. $d = 2$

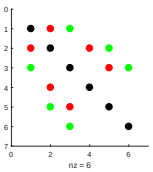
$$\mathsf{A} = \left(\begin{array}{c|cc|ccc} \mathsf{K}_0 & \frac{1}{\sqrt{3}}\mathsf{K}_1 & \frac{1}{\sqrt{3}}\mathsf{K}_2 & 0 & 0 & 0 \\ \hline \frac{1}{\sqrt{3}}\mathsf{K}_1 & \mathsf{K}_0 & 0 & \frac{2}{\sqrt{15}}\mathsf{K}_1 & \frac{1}{\sqrt{3}}\mathsf{K}_2 & 0 \\ \frac{1}{\sqrt{3}}\mathsf{K}_2 & 0 & \mathsf{K}_0 & 0 & \frac{1}{\sqrt{3}}\mathsf{K}_1 & \frac{2}{\sqrt{15}}\mathsf{K}_2 \\ \hline 0 & \frac{2}{\sqrt{15}}\mathsf{K}_1 & 0 & \mathsf{K}_0 & 0 & 0 \\ 0 & \frac{1}{\sqrt{3}}\mathsf{K}_2 & \frac{1}{\sqrt{3}}\mathsf{K}_1 & 0 & \mathsf{K}_0 & 0 \\ 0 & 0 & \frac{2}{\sqrt{15}}\mathsf{K}_2 & 0 & 0 & \mathsf{K}_0 \end{array}\right),$$

K_0, K_1, and K_2 are "stiffness matrices" corresponding to $a_0(\boldsymbol{x})$, $a_1(\boldsymbol{x})$, and $a_2(\boldsymbol{x})$, respectively



stiffness matrices
$(\mathsf{K}_i)_{rs} = \int_D a_i(\boldsymbol{x}) \nabla\psi_r(\boldsymbol{x}) \nabla\psi_s(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}$
Jacobi matrices
$(\mathsf{G}_i)_{jk} = \int_\Gamma \xi_i \Phi_j(\boldsymbol{\xi}) \Phi_k(\boldsymbol{\xi}) \rho(\boldsymbol{\xi}) \mathrm{d}\boldsymbol{\xi}$
$(\mathsf{G}_0)_{jk} = \int_\Gamma \Phi_j(\boldsymbol{\xi}) \Phi_k(\boldsymbol{\xi}) \rho(\boldsymbol{\xi}) \mathrm{d}\boldsymbol{\xi} = \delta_{jk}$

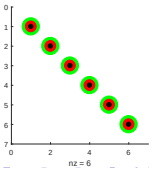
Double orthogonal polynomials



$p_0, p_1, p_2, \ldots$ (infinite set) orthogonal on $\Gamma \subset \mathbb{R}$

$$\int_\Gamma p_j(z) p_k(z) \rho(z) \mathrm{d}z = \delta_{jk}$$

$\widetilde{p}_0, \widetilde{p}_1, \widetilde{p}_2, \ldots \widetilde{p}_m$ (finite set) double-orthogonal on $\Gamma \subset \mathbb{R}$, Lagrange polynomials with the set of nodes - roots of p_m, all of the degree $m-1$



$$\int_\Gamma \widetilde{p}_j(z) \widetilde{p}_k(z) \rho(z) \mathrm{d}z = \delta_{jk}$$
$$\int_\Gamma z\, \widetilde{p}_j(z) \widetilde{p}_k(z) \rho(z) \mathrm{d}z = \delta_{jk}$$

A is block diagonal matrix with different blocks.

Solution methods

Conjugate gradient method with preconditioning

Preconditioning of $\mathsf{Au} = \mathsf{b}$ is getting M such that $\mathsf{M}^{-1}\mathsf{Au} = \mathsf{M}^{-1}\mathsf{b}$ is better solvable than $\mathsf{Au} = \mathsf{b}$, and $\mathsf{Mv} = \mathsf{c}$ is easy to solve. Also in an abstract form.

Preconditioning

1) Multigrid w.r.t. physical variable
Brezina, et al., 2014, Elman, Furnival, 2007

2) Multilevel w.r.t. stochastic variable . . . focused in this talk

3) Reduced basis, low rank approximations, rational Krylov subspace, etc.

vector $u \to$ tensorised matrix U, and

$$\mathsf{b} = \mathsf{Au} = \left(\sum_{i=0}^{N_a} \mathsf{G}_i \otimes \mathsf{K}_i\right) u \qquad \text{is the same as} \qquad \mathsf{B} = \sum_{i=0}^{N_a} \mathsf{K}_i \mathsf{U} \mathsf{G}_i^T$$

Matthies et al., 2014; Powell, Silvester, Simoncini, 2016; Powell, Silvester, Simoncini, 2018; Audouze, Nair, 2019

Multilevel preconditioning with respect to stochastic variable $\left(\mathsf{A} = \sum_{i=0}^{N_a} \mathsf{G}_i \otimes \mathsf{K}_i\right)$

Mean based - diagonal blocks - Powell, Elman, 2009; etc.

$$\mathsf{M}^{\text{mean}} = \mathsf{G}_0 \otimes \mathsf{K}_0$$

Kronecker product preconditioner - Ullmann, 2010

$$\mathsf{M}^{\text{Kron}} = \mathsf{G}_0 \otimes \mathsf{K}_0 + \sum_{i=1}^{N_a} \beta_i \mathsf{G}_i \otimes \mathsf{K}_0, \qquad \beta_i = \frac{\mathrm{tr}(\mathsf{K}_0{}^T \mathsf{K}_i)}{\mathrm{tr}(\mathsf{K}_0{}^T \mathsf{K}_0)}$$

Symmetric block Gauss-Seidel - Bespalov, Loghin, Youngnoi, arXiv 2020

$$\mathsf{M}^{\text{SBGS}} = \left(\mathsf{G}_0 \otimes \mathsf{K}_0 + \sum_{i=1}^{N_a} \mathsf{L}_i \otimes \mathsf{K}_i\right) (\mathsf{G}_0 \otimes \mathsf{K}_0)^{-1} \left(\mathsf{G}_0 \otimes \mathsf{K}_0 + \sum_{i=1}^{N_a} \mathsf{L}_i{}^T \otimes \mathsf{K}_i\right), \quad \mathsf{L}_i + \mathsf{L}_i{}^T = \mathsf{G}_i$$

Two-by-two blocks and Schur complement - Sousedík, Ghanem, Phips, 2013

Block diagonal precoditioner with large blocks - P., Kubínová, 2020

Overview - Crowder, Adaptive and Multilevel Stochastic Galerkin Finite Element Methods, Ph.D. Thesis, 2020

Numerical experiments

1D diffusion equation, $N_{\text{FE}} = 20$, complete polynomials, maximum degree d, number of stoch. variables N_ξ, ξ_i uniformly distributed in $[-1, 1]$

Table: Mean based, Kronecker and SBGS preconditioning.

		$\kappa(\mathsf{M}^{-1}\mathsf{A})$				CG steps			
N_ξ	d	no	mean	Kron	SBGS	no	mean	Kron	SBGS
1	1	252.8	2.1	1.6	1.1	39	10	8	5
1	3	317.5	3.1	2.0	1.2	71	13	9	5
1	7	345.4	3.7	2.3	1.2	103	14	10	5
2	1	256.2	2.1	1.6	1.1	56	10	8	5
2	3	335.5	3.9	2.5	1.3	99	14	11	6
2	7	387.9	6.1	3.6	1.5	125	17	13	6
3	1	262.6	2.1	1.6	1.2	60	10	7	5
3	3	372.6	4.2	2.8	1.4	112	14	11	6
3	7	462.6	7.8	5.0	1.8	142	20	15	7

Our contribution

Improving guaranteed spectral bounds for preconditioned matrix $\mathsf{M}^{-1}\mathsf{A}$ based on

— orthogonal polynomial properties
— data of problems associated to A and M - element-by-element
— for many kinds of distribution of data
P., 2016; Kubínová, P., 2020, Plešinger, P., 2018

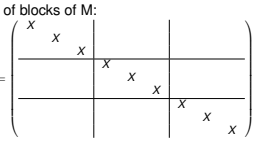
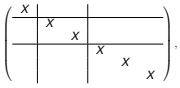
Our approach - connected to and based on classical condition number estimates for algebraic multi-level (AML) preconditioning
Eijkhout, Vassilevski, Axelsson, Neytcheva, Blaheta, Kraus

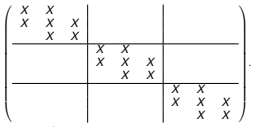
Patterns of blocks of M:

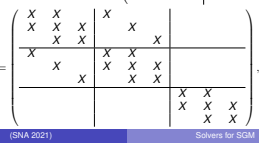
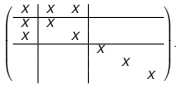
$$\mathsf{M}^{\mathsf{C}} = \left(\begin{array}{cc|cc|cc} X & & & & & \\ & X & & X & & \\ \hline & & X & & & \\ & & & X & X & \\ \hline & & & & & X \end{array}\right), \qquad \mathsf{M}^{\mathsf{C}} = \left(\begin{array}{cc|cc|cc} X & X & X & & & \\ X & X & & & & \\ \hline X & & X & & & \\ & & & X & X & \\ \hline & & & & & X \end{array}\right)$$

Our contribution (cont.)

Patterns of blocks of M:

$\mathrm{M}^{\mathrm{TP}} =$  , $\mathrm{M}^{\mathrm{C}} =$  ,

$\mathrm{M}^{\mathrm{TP}} =$  .

$\mathrm{M}^{\mathrm{TP}} =$  , $\mathrm{M}^{\mathrm{C}} =$  .

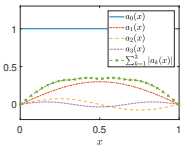
Our contribution (cont.)

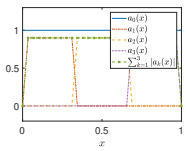
Numerical experiment.

$$-(a(x,\boldsymbol{\xi})u(x,\boldsymbol{\xi})')' = f(x),$$

$D = (0,1)$, $N_\xi = 3$, $a(x,\boldsymbol{\xi}) = a_0(x) + \xi_1 a_1(x) + \xi_2 a_2(x) + \xi_3 a_3(x)$, $\xi_i \in [-1,1]$, $i = 1,2,3$.

Problem 1: Problem 2:





Our contribution (cont.)

Table: Block-diagonal preconditioning of Problem 1 and Problem 2. New and classical bounds.

d	$\kappa(\mathrm{A})$	$\underline{c}_{\mathrm{class}}$	$\underline{c}$	$\lambda_{\min}(\mathrm{M}^{-1}\mathrm{A})$	$\lambda_{\max}(\mathrm{M}^{-1}\mathrm{A})$	$\overline{c}$	$\overline{c}_{\mathrm{class}}$	$\overline{c}/\underline{c}$
Problem 1								
1	458.42	0.76	0.80	0.83	1.17	1.20	1.24	1.51
2	498.47	0.68	0.73	0.76	1.24	1.27	1.32	1.75
...								
6	546.55	0.61	0.67	0.69	1.31	1.33	1.39	2.26
7	550.80	0.61	0.66	0.68	1.32	1.34	1.39	2.29
Problem 2								
1	947.79	-0.65	0.45	0.45	1.56	1.56	2.65	3.43
2	1596.34	-1.21	0.26	0.26	1.74	1.74	3.21	6.57
...								
6	4576.93	-1.71	0.10	0.10	1.90	1.90	3.71	19.34
7	5294.63	-1.74	0.09	0.09	1.91	1.91	3.74	21.80

P., 2015, 2016, 2017; Kubínová, P., 2020

Conclusion

- PDE with (stochastic) parameters - many methods, demanding
- Variational approach - stochastic Galerkin method
  Preconditioning – a posteriori error estimates – adaptivity
  A posteriori error estimates
  Bespalov, Powell, Silvester, 2014; Eigel, Merdon, 2014; Khan, Bespalov, Powell, Silvester, 2020
  Adaptivity
  Eigel, Gittelson, Schwab, Zander, 2014; P., 2015; Bespalov, Praetorius, Ruggeri, 2020
- Relatively short history; many recent results; still developing
  Using as much information about A and M as possible
- Spectral estimates of $\mathrm{M}^{-1}\mathrm{A}$ (matrix pencil) - sharp new bounds based on element-by-element approach
  P., 2016; Kubínová, P., 2017; Plešinger, P., 2018
- Estimates of all eigenvalues of preconditioned matrices
  Gergelits, Mardal, Nielsen and Strakoš, 2019; Gergelits, Nielsen, Strakoš, 2020;
  Ladecký, P., Zeman, 2020; P., Ladecký, submitted
  Our new method described in Martin Ladecký's talk at SNA 2021.

Title: Seminar on Numerical Analysis & Winter School
Proceedings of the conference SNA'21
Ostrava, January 25 – 29, 2021

Editors: Jiří Starý, Stanislav Sysala, Dagmar Sysalová

Published by: Institute of Geonics of the CAS, Ostrava

First edition

Ostrava, 2021

ISBN 978-80-86407-82-1