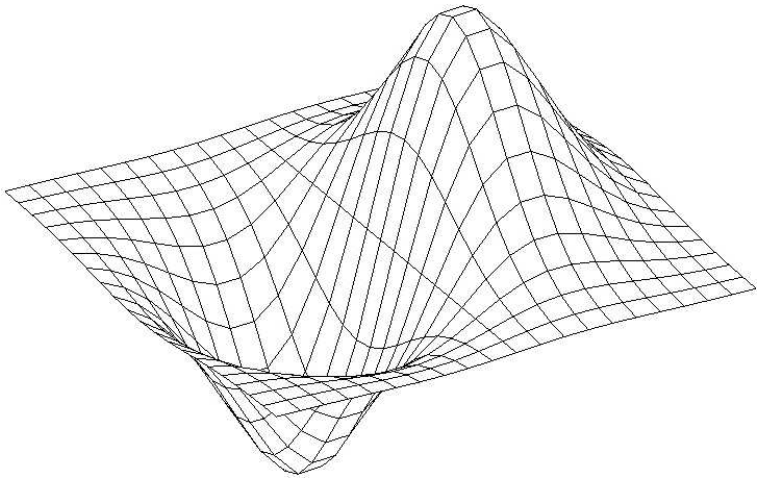
Institute of Geonics AS CR, Ostrava

SNA'09

Seminar on Numerical Analysis

Modelling and Simulation
of Challenging Engineering Problems



Winter School

High-performance and Parallel Computers,
Programming Technologies & Numerical Linear Algebra

Ostrava, February 2 – 6, 2009

Programme committee:

Radim Blaheta	*Institute of Geonics AS CR, Ostrava*
Zdeněk Dostál	*VŠB-Technical University, Ostrava*
Ivo Marek	*Czech Technical University, Prague*
Zdeněk Strakoš	*Institute of Computer Science AS CR, Prague*

Organizing committee:

Radim Blaheta	*Institute of Geonics AS CR, Ostrava*
Petr Harasim	*Institute of Geonics AS CR, Ostrava*
Alexej Kolcun	*Institute of Geonics AS CR, Ostrava*
Jiří Starý	*Institute of Geonics AS CR, Ostrava*
Stanislav Sysala	*Institute of Geonics AS CR, Ostrava*

Conference secretary:

Jaroslava Vávrová	*Institute of Geonics AS CR, Ostrava*

Institute of Geonics AS CR Ostrava
ISBN 978-80-86407-60-9

Preface

Seminar on Numerical Analysis 2009 (SNA'09) is the sixth meeting of a series of events started in Ostrava 2003 and devoted to numerical methods necessary for mathematical modelling of problems in sciences and engineering. In this respect, it was natural that in period 2005 - 2008 the SNA conferences became a part of the project MSTEP (`http://www2.cs.cas.cz/mweb/`) *Modelling and simulation of complex engineering problems: effective numerical algorithms and parallel implementation using new information technologies* within the program Information society administrated by the Academy of Sciences of the Czech Republic. We hope that the tradition of SNA conferences will be preserved even after finishing the MSTEP project in 2008.

Since 2005, a part of SNA has been devoted to the so called Winter school with tutorial lectures devoted to selected topics within the conference scope. In this year, the Winter school includes lectures devoted to the discontinuous Galerkin method and compressible flow (V. Dolejší and M. Feistauer), direct methods for solving indefinite systems (M. Rozložník), duality for variational inequalities (Z. Dostál) and to the shape optimization (J. Haslinger). The Winter school also includes a series of lectures devoted to problems with uncertain input data, namely interval computing (S. Ratschan), fuzzy approach (J. Kruis), worst scenario (J. Chleboun), Monte Carlo approach (D. Novák, M. Vořechovský) and polynomial chaos (T. Kozubek).

The SNA conferences also cover the topics of computer implementation of numerical methods, parallel and high performance computing. Despite some contributions devoted to these topics, this year we would like to inform the participants about the supercomputing project *IT for Innovations*, which is under preparation for the EU funded Operational Programme Research and Development for Innovations by VSB - Technical University of Ostrava, University of Ostrava, Silesian University of Opava and the Institute of Geonics AS CR Ostrava.

On behalf of the Programme and Organizing Committee of SNA'09,

Radim Blaheta and Jiří Starý

Contents

Winter school lectures

V. Dolejší, M. Feistauer:
Discontinuos Galerkin methods and applications to compressible flow

Z. Dostál:
Duality for QP problems with semidefinite Hessian and contact problems

J. Haslinger:
Structural optimization

J. Chleboun:
What is the role of the worst scenario method in solving problems with uncertain input data ?

J. Kruis:
Uncertainty in engineering problems described by fuzzy sets

T. Kozubek:
A numerical solution of elliptic boundary value problems with uncertain data and geometry

D. Novák, M. Vořechovský:
Small-sample simulační metody typu Monte Carlo

M. Rozložník:
Numerical stability of symmetric indefinite solvers: direct methods

S. Ratschan:
Interval computation: Why ? When ? How ?

IT for Innovations

Tvarová optimalizace pro 3D kontaktní problém s Coulombovým třením - o citlivostní analýze

P. Beremlijski, J. Haslinger, M. Kočvara, R. Kučera, J. V. Outrata*

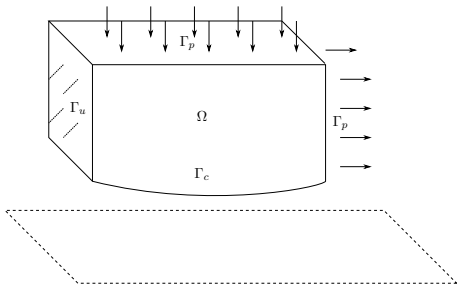
* VŠB - Technická univerzita Ostrava

1 Úvod

V příspěvku se zabýváme diskretizovanou úlohou tvarové optimalizace trojrozměrného pružného tělesa v jednostranném kontaktu s tuhou překážkou. Tření mezi tělesem a překážkou modelujeme Coulombovým zákonem. Matematický model problému s Coulombovým třením vede na řešení kvazivariační nerovnosti. Pro malý koeficient tření má diskrétní kontaktní úloha s Coulombovým třením jediné řešení. Navíc řešení této úlohy je závislé lokálně lipschitovsky na řídící proměnné popisující tvar pružného tělesa. Díky jedinému řešení diskrétní úlohy pro fixovanou řídící proměnnou, můžeme použít tzv. přístup implicitního programování. Ten je založen na minimalizaci nehladké funkce složené z cenové funkce a jednoznačného zobrazení, které řídící proměnné přiřazuje řešení diskrétní úlohy, tzn. stavové proměnné. Pro minimalizaci nehladké funkce lze efektivně použít bundle trust metodu. K výpočtu subgradientní informace, kterou metoda vyžaduje je výhodné použití Clarkeova kalkulu (viz [3]). Implicitní programování spolu s Clarkeovým kalkulem bylo použito pro řešení diskretizované úlohy tvarové optimalizace pro 2D kontaktní problém s Coulombovým třením (viz [1]). Pro 3D úlohu není možné jednoduše modifikovat stejný postup (subdiferenciál eukleidovské normy v $\mathbb{R}^2$ není polyhedrální). Cílem příspěvku je naznačení hledání subgradientu, tj. citlivostní analýza, pro tvarovou optimalizaci 3D kontaktní úlohy s Coulombovým třením (podrobně v [2]).

2 Stavová úloha

Nechť $\Omega \subset \mathbb{R}^3$ je pružné těleso s lipschitzovskou hranicí $\partial\Omega$. Hranice $\partial\Omega$ je složena ze tří nepřekrývajících se částí Γ_u, Γ_p a Γ_c. Viz obrázek 1.



Obrázek 1: 3D pružné těleso.

Γ_u je hranice s Dirichletovskou podmínkou. Povrchové síly $F = (F_1, F_2, F_3)$ působí na hranici Γ_p, $F \in L^2(\Gamma_p)$. Těleso je zdola *podepřeno* podél hranice Γ_c (její tvar je určen řídící proměnnou $\boldsymbol{\alpha} \in \mathbb{R}^d$) tuhou překážkou. Na této hranici je předepsáno Coulombovo tření s koeficientem tření

$\mathcal{F}$. Řešením diskrétního kontaktního problému s Coulombovým třením nazveme uspořádanou dvojici $(\boldsymbol{u}, \boldsymbol{\lambda}) \in \mathbb{R}^n \times \mathbb{R}^p_+$ splňující

$$(\boldsymbol{A}\boldsymbol{u}, \boldsymbol{v} - \boldsymbol{u})_n + \mathcal{F}(\boldsymbol{\lambda}, |T\boldsymbol{v}| - |T\boldsymbol{u}|)_p \geq (\boldsymbol{L}, \boldsymbol{v} - \boldsymbol{u})_n + (\boldsymbol{\lambda}, \boldsymbol{N}\boldsymbol{v} - \boldsymbol{N}\boldsymbol{u})_p \quad \forall \boldsymbol{v} \in \mathbb{R}^n$$
$$(\boldsymbol{\mu} - \boldsymbol{\lambda}, \boldsymbol{N}\boldsymbol{u} + \boldsymbol{\alpha})_p \geq 0 \quad \forall \boldsymbol{\mu} \in \mathbb{R}^p_+,$$

kde $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ a $\boldsymbol{L} \in \mathbb{R}^n$ jsou matice tuhosti a vektor sil závislé na řídící proměnné $\boldsymbol{\alpha}$. Vektor $(\boldsymbol{u}, \boldsymbol{\lambda})$ nazveme stavovou proměnnou. Nyní zavedeme rozdělení vektoru posunutí $\boldsymbol{u}$ na $(\boldsymbol{u}_t, \boldsymbol{u}_\nu)$, kde $\boldsymbol{u}_t$ přísluší tečnému posunutí a $\boldsymbol{u}_\nu$ odpovídá normálovému posunutí. Dále zredukujeme naši úlohu a budeme se zabývat pouze kontaktními uzly (jejich počet je p). Stavová úloha realizuje zobrazení $\mathcal{S} : \boldsymbol{\alpha} \in \mathbb{R}^d \to (\boldsymbol{u}_t, \boldsymbol{u}_\nu, \boldsymbol{\lambda}) \in \mathbb{R}^{4p}$ (řídícímu vektoru $\boldsymbol{\alpha} \in U_{ad}$ je přiřazeno řešení kontaktní úlohy s Coulombovým třením $(\boldsymbol{u}_t, \boldsymbol{u}_\nu, \boldsymbol{\lambda})$). $\mathcal{S}$ je pro malé koeficienty tření lokálně lipschitzovské. Diskretizovanou stavovou úlohu lze ekvivalentně popsat zobecněnou rovností

$$\begin{aligned}
\mathbf{0} &\in \boldsymbol{A}_{tt}(\boldsymbol{\alpha})\boldsymbol{u}_t + \boldsymbol{A}_{t\nu}(\boldsymbol{\alpha})\boldsymbol{u}_\nu - \boldsymbol{L}_t(\boldsymbol{\alpha}) + \tilde{\boldsymbol{Q}}(\boldsymbol{u}_t, \boldsymbol{\lambda}) \\
\mathbf{0} &= \boldsymbol{A}_{\nu t}(\boldsymbol{\alpha})\boldsymbol{u}_t + \boldsymbol{A}_{\nu\nu}(\boldsymbol{\alpha})\boldsymbol{u}_\nu - \boldsymbol{L}_\nu(\boldsymbol{\alpha}) - \boldsymbol{\lambda} \\
\mathbf{0} &\in \boldsymbol{u}_\nu + \boldsymbol{\alpha} + N_{\boldsymbol{R}^p_+}(\boldsymbol{\lambda}),
\end{aligned}$$

kde

$$\tilde{\boldsymbol{Q}}(\boldsymbol{u}_t, \boldsymbol{\lambda}_\nu) = \partial_{u_t} j(\boldsymbol{u}_t, \boldsymbol{\lambda}_\nu), \quad j(\boldsymbol{u}_t, \boldsymbol{\lambda}_\nu) = \mathcal{F} \sum_{i=1}^{p} \lambda^i ||\boldsymbol{u}_t^i||$$

a $N_{\boldsymbol{R}^p_+}$ je standardní normálový kužel. Tuto zobecněnou rovnost můžeme zapsat stručněji takto

$$\mathbf{0} \in F(\boldsymbol{\alpha})\boldsymbol{y} - l(\boldsymbol{\alpha}) + Q(\boldsymbol{y}),$$

kde

$$F(\boldsymbol{\alpha}) = \begin{bmatrix} \boldsymbol{A}_{tt}(\boldsymbol{\alpha}) & \boldsymbol{A}_{t\nu}(\boldsymbol{\alpha}) & \mathbf{0} \\ \boldsymbol{A}_{\nu t}(\boldsymbol{\alpha}) & \boldsymbol{A}_{\nu\nu}(\boldsymbol{\alpha}) & -\mathbf{E} \\ \mathbf{0} & \mathbf{E} & \mathbf{0} \end{bmatrix},$$

$$\boldsymbol{y} = (\boldsymbol{u}_t, \boldsymbol{u}_\nu, \boldsymbol{\lambda})^T, \; l(\boldsymbol{\alpha}) = (\boldsymbol{L}_t(\boldsymbol{\alpha}), \boldsymbol{L}_\nu(\boldsymbol{\alpha}), -\boldsymbol{\alpha})^T, \; Q(\boldsymbol{y}) = \left(Q_t(\boldsymbol{u}, \boldsymbol{\lambda}), \mathbf{0}, N_{\mathbb{R}^p_+}(\boldsymbol{\lambda})\right)^T,$$

$\mathbf{E}$ je jednotková matice.

$F(\boldsymbol{\alpha})\boldsymbol{y} - l(\boldsymbol{\alpha})$ je jednoznačná část zobecněné rovnosti, $Q(\boldsymbol{y})$ je její víceznačná část.

3 Tvarová optimalizace pro kontaktní úlohu s Coulombovým třením

Naším úkolem je nalézt řídící proměnnou $\boldsymbol{\alpha}$ určující tvar Beziérovy plochy, kterou je popsána kontaktní hranice Γ_c, pro kterou nabývá cenový funkcionál $\mathcal{J}(\boldsymbol{\alpha}, \mathcal{S}(\boldsymbol{\alpha}))$ svého minima. Úlohu diskrétní tvarové optimalizace zavedeme jako řešení

$$\min_{\boldsymbol{\alpha} \in U_{ad}} \Theta(\boldsymbol{\alpha}) = \mathcal{J}(\boldsymbol{\alpha}, \mathcal{S}(\boldsymbol{\alpha})).$$

Předpokládejme, že funkcionál $\mathcal{J}$ je spojitě diferencovatelný. K řešení této nehladké úlohy použijeme bundle trust metodu (podrobně viz [5]).

4 Citlivostní analýza pro úlohu tvarové optimalizace

Bundle trust metoda potřebuje rutinu, která v každém kroce vypočte hodnotu cenového funkcionálu (k tomu potřebujeme vyřešit stavovou úlohu) a jeden (libovolný) Clarkeův subgradient z Clarkeova zobecněného gradientu $\partial\Theta(\boldsymbol{\alpha})$. Pro jeho konstrukci použijeme tvrzení

$$\partial\Theta(\boldsymbol{\alpha}) = \nabla_1\mathcal{J}(\boldsymbol{\alpha},\mathcal{S}(\boldsymbol{\alpha})) + \{\boldsymbol{C}^T\nabla_2\mathcal{J}(\boldsymbol{\alpha},\mathcal{S}(\boldsymbol{\alpha}))|\,\boldsymbol{C}\in\partial\mathcal{S}(\boldsymbol{\alpha})\}$$

(viz [3]). Dále využijeme nehladkého kalkulu B. Morduchoviče (viz [4]).

Protože platí $\emptyset \neq D^*\mathcal{S}(\boldsymbol{\alpha})(\boldsymbol{y}^*)$ pro všechna $\boldsymbol{y}^*$ a conv $(D^*\mathcal{S}(\boldsymbol{\alpha}))\,(\boldsymbol{y}^*) = \{\boldsymbol{C}^T\boldsymbol{y}^*|\,\boldsymbol{C}\in\partial\mathcal{S}(\boldsymbol{\alpha})\}$, stačí nalézt jeden prvek z množiny $D^*\mathcal{S}(\boldsymbol{\alpha})(\nabla_2\mathcal{J}(\boldsymbol{\alpha},\mathcal{S}(\boldsymbol{\alpha})))$. Prvky limitní koderivace

$$D^*\mathcal{S}(\boldsymbol{\alpha})(\boldsymbol{y}^*) := \{\boldsymbol{x}^*\in\mathbb{R}^d\,|\,(\boldsymbol{x}^*,-\boldsymbol{y}^*)\in N_{\mathrm{Gr}\,\mathcal{S}}(\boldsymbol{\alpha})\},$$

kde Gr $\mathcal{S}$ je graf $\mathcal{S}$ a $N_{\mathrm{Gr}\,\mathcal{S}}$ je limitní normálový kužel, najdeme použitím následujícího tvrzení.

Teorém 4.1 *Nechť máme $(\boldsymbol{\alpha},\boldsymbol{y})$, kde $\boldsymbol{\alpha}\in U_{ad}, \boldsymbol{y}=\mathcal{S}(\boldsymbol{\alpha})$. Potom pro všechna $\boldsymbol{y}^*\in\mathbb{R}^{4p}$ platí*

$$D^*\mathcal{S}(\boldsymbol{\alpha})(\boldsymbol{y}^*)\subset(\nabla_1(F(\boldsymbol{\alpha})\boldsymbol{y}-l(\boldsymbol{\alpha})))^T\mathcal{V},$$

kde $\mathcal{V}$ je množina řešení $\boldsymbol{v}$ limitní adjungované zobecněné rovnosti

$$\boldsymbol{0}\in\boldsymbol{y}^*+(F(\boldsymbol{\alpha}))^T\boldsymbol{v}+D^*Q(\boldsymbol{y},-F(\boldsymbol{\alpha})\boldsymbol{y}+l(\boldsymbol{\alpha}))(\boldsymbol{v}).$$

Abychom vypočetli koderivaci $D^*Q(\boldsymbol{y},-F(\boldsymbol{\alpha})\boldsymbol{y}+l(\boldsymbol{\alpha}))(\boldsymbol{v})$ přeuspořádáme víceznačnou část zobecněné rovnosti $Q(\boldsymbol{y})$ následujícím způsobem

$$Q(\boldsymbol{y}) = \begin{bmatrix}\Phi(\boldsymbol{y}^1)\\ \Phi(\boldsymbol{y}^2)\\ \vdots\\ \Phi(\boldsymbol{y}^p)\end{bmatrix},$$

kde $\boldsymbol{y}^i = (\boldsymbol{u}^i_\tau, u^i_\nu, \lambda^i)\in\mathbb{R}^2\times\mathbb{R}\times\mathbb{R}_+$ obsahuje hodnoty všech stavových proměnných v i-tém kontaktním uzlu a

$$\Phi(\boldsymbol{y}^i) = \begin{bmatrix}\mathcal{F}\lambda_i\partial\|\boldsymbol{u}^i_\tau\|_2\\ 0\\ N_{\mathbb{R}_+}(\lambda_i)\end{bmatrix},\ i=1,2,\ldots,p.$$

Pro výpočet $D^*Q(\boldsymbol{y},-F(\boldsymbol{\alpha})\boldsymbol{y}+l(\boldsymbol{\alpha}))(\boldsymbol{v})$ je nutné provést diskusi polohy bodu $(\boldsymbol{y},-F(\boldsymbol{\alpha})\boldsymbol{y}+l(\boldsymbol{\alpha}))$ vzhledem ke Gr Q, tj. diskusi poloh bodů $(\boldsymbol{y}^i,-F^i(\boldsymbol{\alpha})\boldsymbol{y}+l^i(\boldsymbol{\alpha}))$ vzhledem ke Gr Φ, $i=1,2,\ldots,p$. Pro zjednodušení zaveďme místo $(\boldsymbol{y}^i,-F^i(\boldsymbol{\alpha})\boldsymbol{y}+l^i(\boldsymbol{\alpha}))$ dvojici vektorů $(\boldsymbol{a},\boldsymbol{b})\in$ Gr Φ (tzn. $b_3=0$) a označme symbolem $\boldsymbol{a}_{12}$ dvojrozměrný vektor $(a_1,a_2)^T$ a symbolem $\boldsymbol{b}_{12}$ vektor $(b_1,b_2)^T$.

Množinu Gr Φ můžeme zapsat

$$\mathrm{Gr}\ \Phi = L\cup M_1\cup M_2\cup M_3^+\cup M_3^-\cup M_4,$$

kde

$$\begin{aligned}
L &= \{(\boldsymbol{a},\boldsymbol{b})\in\mathrm{Gr}\ \Phi\mid b_4<0\},\\
M_1 &= \{(\boldsymbol{a},\boldsymbol{b})\in\mathrm{Gr}\ \Phi\mid \boldsymbol{a}_{12}\neq 0, a_4>0\},\\
M_2 &= \{(\boldsymbol{a},\boldsymbol{b})\in\mathrm{Gr}\ \Phi\mid \boldsymbol{a}_{12}\neq 0, a_4=0, b_4=0\},\\
M_3^+ &= \{(\boldsymbol{a},\boldsymbol{b})\in\mathrm{Gr}\ \Phi\mid \boldsymbol{a}_{12}=0, a_4>0, \|\boldsymbol{b}_{12}\|<\mathcal{F}a_4\},\\
M_3^- &= \{(\boldsymbol{a},\boldsymbol{b})\in\mathrm{Gr}\ \Phi\mid \boldsymbol{a}_{12}=0, a_4>0, \|\boldsymbol{b}_{12}\|=\mathcal{F}a_4\},\\
M_4 &= \{(\boldsymbol{a},\boldsymbol{b})\in\mathrm{Gr}\ \Phi\mid \boldsymbol{a}_{12}=0, a_4=0, \|\boldsymbol{b}_{12}\|=\mathcal{F}a_4, b_4=0\}.
\end{aligned}$$

Všimněme si významu předchozích množin. Pokud $\boldsymbol{a}_{12} \neq 0$ hovoříme o prokluzu, zatímco když $\boldsymbol{a}_{12} = 0$ o přilepení. L znamená stav bez kontaktu a tedy i bez tření. M_1 odpovídá prokluzu s kontaktem, M_2 popisuje prokluz se slabým kontaktem, M_3^+ přilepení s kontaktem, M_3^- slabé přilepení s kontaktem a M_4 slabé přilepení se slabým kontaktem.

Množiny L, M_1 a M_3^+ popisují stabilní chování, tj. platí následující implikace

$$\left.\begin{array}{l}(\bar{\boldsymbol{a}}, \bar{\boldsymbol{b}}) \in L(\text{ or } M_1 \text{ or } M_3^+) \\ (\boldsymbol{a}, \boldsymbol{b}) \in \text{ Gr } \Phi \\ (\boldsymbol{a}, \boldsymbol{b}) \text{ je } \textit{blízko } (\bar{\boldsymbol{a}}, \bar{\boldsymbol{b}})\end{array}\right\} \Rightarrow (\boldsymbol{a}, \boldsymbol{b}) \in L(\text{ or } M_1 \text{ or } M_3^+)$$

Pro jednotlivé množiny $L, M_1, M_2, M_3^+, M_3^-, M_4$ lze pak odvodit vztahy pro výpočet koderivace $D^*\Phi((\boldsymbol{a}, \boldsymbol{b}))(\boldsymbol{v})$ a z nich pak zkonstruovat $D^*Q(\boldsymbol{y}, -F(\boldsymbol{\alpha})\boldsymbol{y} + l(\boldsymbol{\alpha}))(\boldsymbol{v})$.

5 Závěr

Ve 2D verzi výše popsané úlohy tvarové optimalizace bylo využito toho, že stavové zobrazení $\mathcal{S}$ je po částech spojitě diferencovatelné. O stavovém zobrazení ve 3D případě to již není známo. Proto je použití Morduchovičova kalkulu nezbytné pro řešení optimalizační úlohy, kterou se zabýváme v této práci. Při implementaci navrženého postupu je možno udělat určité úpravy, které mohou ještě zefektivnit řešení dané úlohy.

Tato práce byla podpořena GA ČR 201/07/0294, MŠMT MSM6198910027.

Literatura

[1] P. Beremlijski, J. Haslinger, M. Kočvara and J. Outrata: *Shape Optimization in Contact Problems with Coulomb Friction*. In: SIAM Journal on Optimization 12/3, 2002, pp. 561 - 587.

[2] P. Beremlijski, J. Haslinger, M. Kočvara, R. Kučera and J. Outrata: *Shape Optimization in Three-Dimensional Contact Problems with Coulomb Friction*. In: SIAM Journal on Optimization (accepted).

[3] F. H. Clarke: *Optimization and Nonsmooth Analysis*. J. Wiley & Sons, 1983.

[4] B. S. Mordukhovich, *Variational Analysis and Generalized Differentiation, Volumes I and II*. Springer-Verlag, 2006.

[5] J. Outrata, M. Kočvara, and J. Zowe: *Nonsmooth Approach to Optimization Problems with Equilibrium Constraints: Theory, Applications and Numerical Results*. Kluwer Acad. Publ., 1998.

Multiscale modelling of geomaterials and iterative solvers

R. Blaheta, P. Byczanski, P. Harasim

Institute of Geonics AS CR, Ostrava

1 Introduction

Standard geomaterials as well as other construction materials are mostly considered as homogeneous or piecewise homogeneous at the application scale. On the other hand, these materials are heterogeneous when we consider a finer scale and this heterogeneous structure gives insight to many properties of the materials and processes occurring in them. Further, we shall speak about microstructure of materials despite of the size of the objects considered at the finer scale, which can be different constituents, grains but also just homogeneous pieces of a rock mass.

Let us mention specific geotechnical problems. First, the properties of a rock mass can be influenced by grouting the rock matrix with a polyurethane resin. The mechanical properties as well as permeability then depend on degree of filling the fractures and properties of the used resin. The numerical upscaling and evaluation of properties of homogenized material enable to assess the effect of grouting and can be also used for optimization of the grouting process.

Other problems are in assessment of the influence of microstructure to processes at the microlevel. For example, for porous media flow and even more for transport and reaction of chemicals, it may be important to know so called hydrogeological dispersion due to different properties in microstructure. Further, for investigation of the mechanical damage of material, it is again important to investigate initialization of the damage due to heterogeneity in microstructure and subsequently heterogeneity in the stress field.

The knowledge of microstructure, necessary for the modelling, can be deterministic or stochastic. The deterministic knowledge can be derived from microscope observation, using X-ray CT scans, ultrasound tomography etc. The stochastic information can be derived from a partial knowledge of material and can be also readily used for assessment of sensitivity to the microstructure variation.

There are also important computational aspects of solving boundary value problems with microstructure representation:

- the discretization capable to represent the microstructure should normally be very dense and, consequently, requires solving very large problems,
- the oscillation of coefficients causes very ill conditioning of the solved problems.

2 Problems with deterministic microstructure

For optimization of the grouting process, it is possible to use numerical upscaling. It means that a cubic samples with edge 75mm are scanned by X-ray computer tomography and discretized with an uniform voxel grid with $251 \times 251 \times 76$ grid giving 4 788 076 nodes and nearly 15 million DOFs in the case of investigation of elastic properties. The CT scans are used for determining

the properties in different voxels and representing the computational microstructure. An use of finite element analysis then allow to compute homogenized properties.

For the solution of the FE systems, arising from upscaling elastic properties, we used two-level Schwarz method with a coarse space created by the non-smoothed aggregation [1]. In this case, the heterogeneity as well as the size of jumps in coefficients influence very negatively the convergence of the method, see [2].

A remedy can be found in using other coarse spaces, which can be constructed in different ways. First way uses multiscale finite element basis functions which are a-harmonic on the coarse elements, see [4, 3]. Another way, is to use a coarse space defined by basis functions with prescribed supports and energy minimization property, see [3]. These approaches are also close to multilevel methods with elementwise Schur complements, see [5, 6].

3 Problems with stochastic microstructure

We shall consider an academic model problem of saturated Darcy flow through a representative volume $\Omega = \langle 0,\, 1\rangle \times \langle 0,\, 1\rangle$, see [7]. The flow is described by the equations

$$\nabla \cdot u \,=\, 0, \quad u \,=\, -k\nabla(p) \qquad \text{in} \quad \Omega, \tag{1}$$

$$u_n = 0 \quad \text{on} \quad \Gamma_u = \{x :\ x_1 = 0 \text{ and } x_1 = 1\}, \tag{2}$$

$$p = 1,\ p = 0 \quad \text{on} \quad \Gamma_{p1} = \{x :\ x_2 = 0\} \text{ and } \Gamma_{p2} = \{x :\ x_2 = 1\}, \text{respectivelly.} \tag{3}$$

The stochastic character is given by the permeability coefficient k. We shall assume that k is a random field with the following properties:

- for all $x \in \Omega$ the quantity $z(x) = \log k(x)$ has normal distribution with the mean value 0,
- there is a correlation given by the covariance with the parameters σ (the variance) and λ (the length scale),

$$\Sigma_{xy} = \mathrm{cov}(z(x), z(y)) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \xi\eta\, \phi_{\xi,\eta}\, d\xi\, d\eta \,= \sigma^2 \exp(-\mid x - y \mid /\lambda), \tag{4}$$

  where $\phi_{\xi,\eta}$ is the conjugate probability density of $(z(x),\, z(y))$, $x, y \in \Omega$.

The defined problem is discretized by a mixed finite element method on a regular grid Ω_h created by a division of Ω into small congruent squares and subsequent division of the squares into triangles. Then we use the lowest order Thomas-Raviart finite elements for discretization of the problem.

In the stochastic FE approach (e.g. [8]), a specific problem is the generation of the correlated random fields giving values of k (constant) on the square grid elements. Our approach starts with generation of an uncorrelated random field (denoted as $\lambda = 0$) at an extended grid Ω_h^+ and smoothing this field with the aid of a prepared stencil. This approach will be more thoroughly described in a forthcoming paper. For another approach see e.g. [9].

4 Iterative solution

The stochastic microstructure problem enables easily to investigate the robustness of iterative methods with respect to oscillation of the PDE's coefficient. Let us use the mixed formulation of the porous media flow, which means that both pressure p and Darcy velocity u are considered as independent unknowns. This formulation is an origin for the mixed finite element methods with two big advantages over the standard approach: better approximation of fluxes and preservation of the local mass conservation for the approximate solution.

The mixed variational formulation and its simplest discretization with lowest order Raviart-Thomas leads to the system

$$A \begin{bmatrix} u \\ p \end{bmatrix} = \begin{bmatrix} 0 \\ -\varphi \end{bmatrix}, \quad A = \begin{bmatrix} M & B^T \\ B & 0 \end{bmatrix}. \tag{5}$$

which is symmetric, indefinite and regular. We shall solve this system by MINRES method preconditioned first by a block diagonal preconditioner C_η,

$$C_\eta = \begin{bmatrix} M_\eta & 0 \\ 0 & \eta I \end{bmatrix}, \text{ where } M_\eta = M + \eta^{-1} B^T B. \tag{6}$$

This preconditioner is introduced and analysed in [12], where we can found a proof of h-independent spectral equivalence between C_η and A. For solving problems with strong heterogeneity, it is important that the coefficient oscillation does not influence the $B^T B$ term.

The second step consists in use of a Schwarz type preconditioner for M_η, i.e.

$$M_\eta^{-1} \sim G_\eta = \sum_{i=1}^{s} R_i^T M_{\eta,i}^{-1} R_i, \qquad C_\eta^{-1} \sim \begin{bmatrix} G_\eta & 0 \\ 0 & \eta^{-1} I \end{bmatrix}. \tag{7}$$

The construction of the Schwarz preconditioner to M_η is described e.g. in [11, 10]. For testing the robustness of the MINRES with the above block diagonal (BD) and one-level additive Schwarz (AS) preconditioners we solve the model problem of Section 3 with 101×101 grid (h=1/100) and the most oscillatory uncorrelated random field ($\lambda = 0$). The number of subdomains used for construction of the Schwarz prconditioner is $s = 4$ and the subdomains are vertical strips with the overlap 2h. The numbers of iterations can be found in the following Table.

	$\sigma=0$		$\sigma=1$		$\sigma=2$		$\sigma=3$		$\sigma=4$	
η	BD	AS	BD	AS	BD	AS	BD	AS	BD	AS
1e-1	47	222	66	260	82	320	158	631	545	2216
1e-2	17	94	23	124	26	149	53	316	179	1044
1e-3	8	56	9	82	11	96	20	191	77	695
1e-4	5	29	6	66	6	89	8	142	29	393
1e-5	4	19	4	59	4	78	5	134	12	257
1e-6	3	14	3	52	3	69	4	110	6	211
1e-8	4	8	4	47	4	68	4	92	4	170

Note that this Table shows relatively very good robustness and efficiency of the method. The coefficients are oscillatory with jumps $2 \cdot 10^{-2}$ to $6 \cdot 10^{2}$ for $\sigma = 2$, $1 \cdot 10^{-4}$ to $8 \cdot 10^{3}$ for $\sigma = 3$, $1 \cdot 10^{-7}$ to $9 \cdot 10^{6}$ for $\sigma = 4$.

5 Conclusions

The paper described reasons for considering the microstructure of materials and solving boundary value problems with oscillatory coefficients. For solving the mixed FE problems with oscillating coefficients, a preconditioned MINRES method is suggested and efficiency and robustness of this method is shown. In paper also a stochastic description of microstructure is introduced with aims to investigate sensitivity of processes and robustness of iterative solvers.

Acknowledgement: This work is supported by the grants GACR105/09/1830 of the Grant Agency CR and the research plan AV0Z30860518 of the Academy of Sciences of the Czech Republic.

References

[1] R. Blaheta, Algebraic Multilevel Methods with Aggregations: An Overview, In: I. Lirkov, S. Margenov, and J. Wasniewski (Eds.) LSSC 2005, LNCS 3743, Springer, Berlin, Heidelberg 2006, pp. 3-14, 2006.

[2] J. Starý, R. Blaheta, R. Kohut, A. Kolcun, S. Margenov, Micro FEM analysis of geocomposites. Conference Parallel Matrix Algorithms and Applications 2008, section Robust multilevel methods and parallel algorithms, Neuchatel, June 2008

[3] Jan Van lent, Robert Scheichl and Ivan G. Graham, Energy minimizing coarse spaces for two-level Schwarz methods for multiscale PDEs, submitted to Numerical Linear Algebra with Applications

[4] Graham IG, Lechner PO, Scheichl R. Domain decomposition for multiscale PDEs. Numer. Math. 2007; 106(4):589-626

[5] J. Kraus, Algebraic multilevel preconditioning of finite element matrices using local Schur complements, Numerical Linear Algebra with Applications, 13 (2006), pp. 49-70

[6] O. Axelsson, R. Blaheta, M. Neytcheva, Preconditioning of boundary value problems using elementwise Schur complements, submitted to SIAM Matrix Analysis 7/2008

[7] Cliffe KA, Graham IG, Scheichl R, Stals L. Parallel computation of flow in heterogeneous media modelled by mixed finite elements. J. Comput. Phys. 2000; 164(2): 258-282

[8] M.A. Gutiérrez, S. Krenk, Stochastic Finite Element Methods. In: Encyclopedia of Computational Mechanics, E. Stein, R. de Borst, T.J.R. Hughes eds. Volume 2: Solids and Structures. John Wiley, Chichester 2004

[9] B. Kozintsev, Computations with Gaussian Random Fields, Ph.D. dissertation, University of Maryland at College Park, 1999

[10] R. Blaheta, *Schwarz Methods for Simulation of THM Processes*, In: Proceedings of SIMONA: 3rd Int. Workshop on Simulation and Modelling and Numerical Analysis, TU Liberec, September 2006.

[11] D.N. Arnold, R.S. Falk, R. Winther, Preconditioning in H(div) and applications, Mathematics of Computations 66(1997), 957–984

[12] P.S. Vassilevski, R.D. Lazarov, Preconditioning mixed finite element saddle–point elliptic problems, Num. Linear Alg. Appl. 3(1996), 1–20.

Řešení Bernoulliho úlohy s volnou hranicí pomocí BEM

J. Bouchala, T. Kozubek, M. Sadowská

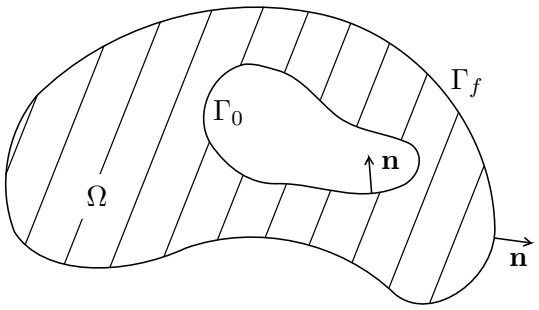
VŠB - Technická univerzita Ostrava

1 Úvod

Budeme se zabývat vnější Bernoulliho úlohou s volnou hranicí, se kterou se setkáme například při řešení problémů mechaniky tekutin, galvanizace kovů, elektrostatiky (viz [1],[4],[6]). Cílem bude ukázat efektivní způsob řešení této úlohy založený na kombinaci technik tvarové optimalizace a metody hraničních prvků (BEM). Tento přístup spočívá v přeformulování Bernoulliho úlohy na úlohu tvarové optimalizace, jejíž stavový problém budeme diskretizovat pomocí BEM. Řešení stavové úlohy bude přímo reprezentovat Neumannova data na volné hranici oblasti.

2 Formulace problému

Buď $\mathcal{O}$ třída omezených dvojnásobně souvislých oblastí $\Omega \subset \mathbb{R}^2$ s lipschitzovskou hranicí $\partial\Omega = \Gamma_0 \cup \Gamma_f$, kde Γ_0 je pevná hranice a $\Gamma_f = \Gamma_f(\Omega)$ je volná hranice (viz obrázek 1). Volná hranice oblasti leží ve vnějšku pevné části hranice. Hledejme oblast $\Omega^* \in \mathcal{O}$ a funkci $u : \Omega^* \mapsto \mathbb{R}$ takové,



Obrázek 1: Geometrie stavové úlohy.

že

$$\triangle u = 0 \ \text{ v } \ \Omega^*, \quad u = g \ \text{ na } \ \partial\Omega^*, \quad \frac{\mathrm{d}u}{\mathrm{d}\mathbf{n}} = Q \ \text{ na } \ \Gamma_f(\Omega^*), \tag{1}$$

kde $g = 1$ na Γ_0, $g = 0$ na $\Gamma_f(\Omega^*)$, $Q = \text{konst.} < 0$ a $\mathbf{n}$ je vnější jednotkový normálový vektor k $\partial\Omega^*$. Lze ukázat [2], že pokud je Γ_0 hranicí C^2 oblasti *hvězdicového typu*, existuje jednoznačné (klasické) řešení úlohy (1).

Je zřejmé, že pro předem danou oblast $\Omega \in \mathcal{O}$ není výše uvedená okrajová úloha korektní, protože předepsané okrajové podmínky na volné hranici tvoří přeurčený systém. Abychom odstranili tuto obtíž, přeformulujeme úlohu (1) pomocí metod optimálního řízení, kdy tvar oblasti Ω bude hrát roli řídící proměnné. Základní myšlenka tohoto přístupu je velmi jednoduchá: přebývající okrajovou podmínku zahrneme do vhodného cenového funkcionálu a zbylou pak budeme splňovat a priori jako součást dané stavové úlohy, která již bude zadaná korektně.

Namísto (1) budeme tedy uvažovat následující optimalizační problém: nalezněme $\Omega^* \in \mathcal{O}$ tak, aby

$$J(\Omega^*, u(\Omega^*)) \leq J(\Omega, u(\Omega)) \quad \forall \Omega \in \mathcal{O}, \tag{2}$$

kde

$$J(\Omega, u(\Omega)) := \frac{1}{2} \left\| \frac{\mathrm{d}u(\Omega)}{\mathrm{d}\mathbf{n}} - Q \right\|^2_{H^{-1/2}(\Gamma_f(\Omega))} \tag{3}$$

a $u(\Omega)$ řeší stavový problém

$$\triangle u = 0 \ \text{ v } \ \Omega, \quad u = g \ \text{ na } \ \partial\Omega. \tag{4}$$

Vztah mezi problémy (1) a (2) je snadno vidět: oblast $\Omega^* \in \mathcal{O}$ je řešením (1) právě tehdy, jestliže Ω^* řeší (2) a současně $J(\Omega^*, u(\Omega^*)) = 0$.

3 Slabá hraniční formulace stavové úlohy

Pro slabé řešení $u \in H^1(\Omega)$ Laplaceovy rovnice v $\Omega \in \mathcal{O}$ platí Gaussův reprezentační vztah, tj.

$$u(x) = \int_{\partial\Omega} \gamma_1 u(y)\, U(x,y)\, \mathrm{d}s_y - \int_{\partial\Omega} \gamma_0 u(y)\, \gamma_{1,y}\, U(x,y)\, \mathrm{d}s_y, \quad x \in \Omega, \tag{5}$$

kde

$$U(x,y) := -\frac{1}{2\pi} \ln \|x - y\|, \quad x,\, y \in \mathbb{R}^2,$$

je fundamentální řešení Laplaceova operátoru v rovině, $\gamma_0 : \ H^1(\Omega) \mapsto H^{1/2}(\partial\Omega)$ je operátor stopy a $\gamma_1 : \ \{v \in H^1(\Omega) : \ \triangle v \in L^2(\Omega)\} \mapsto H^{-1/2}(\partial\Omega)$ je operátor příslušné normálové derivace, který je pro $v \in C^\infty(\overline{\Omega})$ dán vztahem

$$\gamma_1 v = \frac{\mathrm{d}v}{\mathrm{d}\mathbf{n}} \quad \text{na} \quad \partial\Omega.$$

Aplikací operátoru stopy na (5) získáme (viz [11]) vztah

$$\gamma_0 u = (\frac{1}{2} I - K)\gamma_0 u + V\gamma_1 u \quad \text{na } \partial\Omega$$

s dobře známými hraničními integrálními operátory [3, 11]:

$$V : \ H^{-1/2}(\partial\Omega) \mapsto H^{1/2}(\partial\Omega), \ (V\lambda)(x) := \int_{\partial\Omega} \lambda(y)\, U(x,y)\, \mathrm{d}s_y \ \text{(operátor jednoduché vrstvy)},$$

$$K : \ H^{1/2}(\partial\Omega) \mapsto H^{1/2}(\partial\Omega), \ (Kv)(x) := \int_{\partial\Omega} v(y)\, \gamma_{1,y}\, U(x,y)\, \mathrm{d}s_y \ \text{(operátor dvojvrstvy)},$$

$x \in \partial\Omega$.

Slabou hraniční formulací Dirichletova problému (4) rozumíme úlohu: nalezněme $\lambda \in H^{-1/2}(\partial\Omega)$ splňující

$$V\lambda = (\frac{1}{2} I + K)g. \tag{6}$$

Je známo, že pokud $\operatorname{diam}\Omega < 1$, je úloha (6) jednoznačně řešitelná [3].

K vyčíslení cenového funkcionálu (3) tedy použijeme řešení $\lambda = \gamma_1 u(\Omega)$ úlohy (6).

4 Numerické výsledky

V numerických experimentech jsme použili toto nastavení:

$$\Gamma_0 := \{x \in \mathbb{R}^2 : \ \|x\| = 0,1\} \quad \text{a} \quad Q := -\frac{1}{0,3 \ln 3}. \tag{7}$$

Přesné řešení úlohy (1), které dále využijeme k porovnání s jeho vypočtenou aproximací, má potom tvar

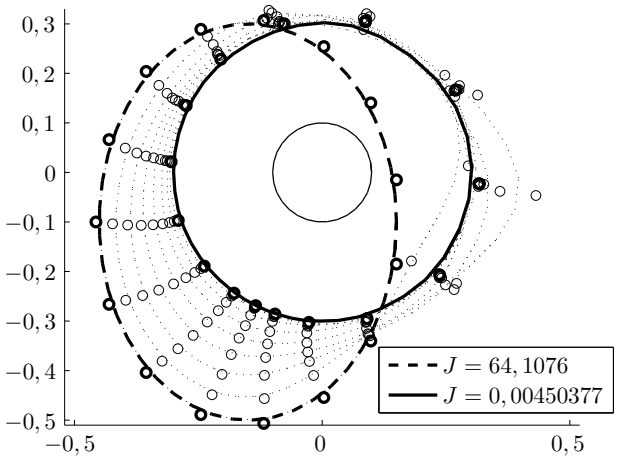
$$\Omega^* = \{x \in \mathbb{R}^2 : \ 0,1 < \|x\| < 0,3\} \quad \text{a} \quad u(x) = -\frac{1}{\ln 3} \cdot \ln \frac{\|x\|}{0,3}. \tag{8}$$

Úlohu (6) jsme diskretizovali pomocí Galerkinovy metody, přičemž jsme použili dělení s 30 uzly na Γ_0 a 45 uzly na $\Gamma_f(\Omega)$. Pro aproximaci normálové derivace na volné hranici byly zvoleny po částech konstantní testovací funkce. BEM je zde vhodnou metodou pro řešení stavového problému, jelikož dává normálovou derivaci na volné hranici s velmi dobrou přesností a navíc není nutné diskretizovat celou oblast, ale pouze její hranici.

Třídu $\mathcal{O}$ jsme nahradili množinou omezených dvojnásobně souvislých oblastí v $\mathbb{R}^2$ s předepsanou pevnou hranicí a s volnou hranicí realizovanou po částech Bézierovou křivkou nejvýše druhého řádu. Při řešení jsme zvolili 15 řídících bodů pro určení tvaru volné hranice.

Pro minimalizaci cenového funkcionálu (3) jsme použili metodu největšího spádu v kombinaci s jednorozměrným vyhledáváním na bázi půlení intervalu [10]. Vztah pro gradient cenového funkcionálu je odvozen v [10].

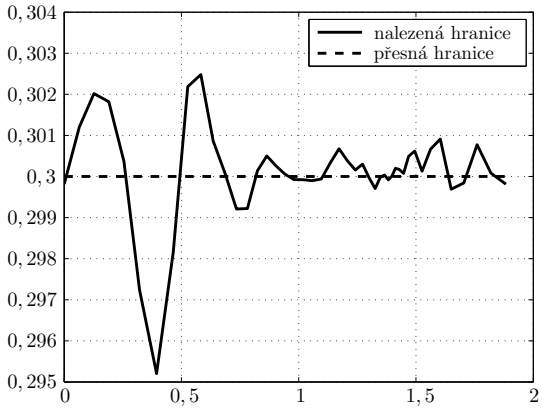
Na obrázku 2 je tlustou plnou čarou vykreslen nalezený tvar volné hranice odpovídající zadaným hodnotám (7). Tlustou přerušovanou čarou je znázorněn počáteční tvar volné hranice. Uvádíme i hodnoty funkcionálu J odpovídající výchozímu a nalezenému tvaru volné hranice.



Obrázek 2: Optimalizovaný tvar volné hranice.

Výsledný tvar volné hranice odpovídá řešení (8), což ukazuje i obrázek 3, kde je znázorněna vzdálenost uzlů na přesné a nalezené volné hranici od středu $(0,0)$.

Poděkování: Tato práce byla podpořena granty GAČR 201/07/0294, AVČR IAA100750802 a výzkumným záměrem MSM 6198910027.



Obrázek 3: Vzdálenost uzlů na přesné a nalezené volné hranici od bodu $(0, 0)$.

Literatura

[1] A. Acker: *An extremal problem involving distributed resistance.* SIAM J Math Anal 12, 1981, 169–172.

[2] A. Acker, R. Meyer: *A Free Boundary Problem for the p-Laplacian: Uniqueness, Convexity, and Successive Approximation of Solutions.* Electronic J Diff Eq 8, 1995, 1-20.

[3] M. Costabel: *Boundary integral operators on Lipschitz domains: Elementary results.* SIAM J Math Anal 19, 1988, 613–626.

[4] A. Fasano: *Some free boundary problems with industrial applications.* Delfour M. C., Sabidussi G. (eds.), Shape optimization and free boundaries, 1992, 113–142.

[5] M. Flucher, M. Rumpf: *Bernoulli's free-boundary problem, qualitative theory and numerical approximation*, J Reine Angew Math 486, 1997, 165–204.

[6] A. Friedman: *Free-boundary problem in fluid dynamics.* Astérisque, Soc Math France 118, 1984, 55–67.

[7] H. Harbrecht: *A Newton method for Bernoulli's free boundary problem in three dimensions.* Computing 82, 2008, 11–30.

[8] J. Haslinger, T. Kozubek, K. Kunisch, G. Peichl: *Shape optimization and fictitious domain approach for solving free boundary value problems of Bernoulli type.* Comput Optim Appl 26, 2003, 231–251.

[9] J. Haslinger, K. H. Hoffmann, R. Mäkinen: *Optimal control/dual approach for the numerical solution of a dam problem.* Advances Math Sci Appl 2, 1993, 189–213.

[10] J. Haslinger, K. Ito, T. Kozubek, K. Kunisch, G. Peichl: *On the shape derivative for problems of Bernoulli type.* Přijato k publikaci v IFB.

[11] O. Steinbach: *Numerical Approximation Methods for Elliptic Boundary Value Problems, Finite and Boundary Elements.* Springer – New York, 2008.

Two Views on Discrete Approximation of Balance Laws

M. Brandner, J. Egermaier, H. Kopincová

University of West Bohemia, Pilsen

1 Introduction

There are many finite volume schemes with different properties (for example central, upwind or central-upwind schemes) for solving conservation laws represented by hyperbolic system of equations

$$\mathbf{q}_t + [\mathbf{f}(\mathbf{q})]_x = \mathbf{0}, \tag{1}$$

where $\mathbf{q}(x,t)$ is the vector of conserved quantities and $\mathbf{f}(\mathbf{q})$ is the flux function. The different approaches based on the simplified wave decompositions are used to construct these schemes.

The main goal of this work is to show two types of constructions same for all these schemes and thus describe the connections between the central and central-upwind schemes and the approximate Riemann solvers. The first way is based on information about the structure of solution of Riemann problem. This information is used in decomposition of flux function. The second way is based on decomposition of space interval to subintervals by the speeds of the waves. The described ideas are also useful for numerical solving of nonhomogeneous systems with spatially varying flux functions.

2 Finite volume methods

The finite volume methods are suitable for solving conservation laws, because the numerical solution is modified only by the intercell fluxes. These methods are based on the integral formulation

$$\int_{x_1}^{x_2} \mathbf{q}(x,t_{n+1})\,dx - \int_{x_1}^{x_2} \mathbf{q}(x,t_n)\,dx + \int_{t_n}^{t_{n+1}} \mathbf{f}(\mathbf{q}(x_2,t_n))\,dt - \int_{t_n}^{t_{n+1}} \mathbf{f}(\mathbf{q}(x_1,t_n))\,dt = \mathbf{0}, \quad \forall (x_1,x_2)\times(t_n,t_{n+1}) \subset \mathbf{R}\times(0,T). \tag{2}$$

They use approximations of the integral averages of the unknown functions instead of the approximations of the unknown functions.

Fully discrete conservative method can be written as relation between approximations of the flux averages and approximations of the integral averages of the conserved quantities

$$\bar{\mathbf{Q}}_j^{n+1} = \bar{\mathbf{Q}}_j^n - \frac{\Delta t}{\Delta x}(\bar{\mathbf{F}}_{j+1/2}^{n+1/2} - \bar{\mathbf{F}}_{j-1/2}^{n+1/2}). \tag{3}$$

We can also derive the semidiscrete form of this method

$$\frac{d}{dt}\bar{\mathbf{Q}}_j = -\frac{1}{\Delta x}[\mathbf{F}_{j+1/2} - \mathbf{F}_{j-1/2}], \tag{4}$$

or semidiscrete method in the fluctuation form

$$\frac{d\bar{\mathbf{Q}}_j}{dt} = -\frac{1}{\Delta x}[\mathbf{A}^-(\Delta\mathbf{Q}_{j+1/2}) + \mathbf{A}(\Delta\mathbf{Q}_j) + \mathbf{A}^+(\Delta\mathbf{Q}_{j-1/2})]. \tag{5}$$

3 Decomposition of the flux function

All standard schemes like central schemes, upwind schemes or central-upwind schemes can be represented and understood by the same way. The amount of information about the structure of the solution of the Riemann problem included into schemes causes the differences between schemes. This information is employed in decomposition of the difference of the flux function.

The semidiscrete central schemes use estimate of upper bound of maximal local speed of the propagating discontinuities. They are based on the following decomposition

$$\mathbf{f}(\mathbf{Q}^+_{j+1/2}) - \mathbf{f}(\mathbf{Q}^-_{j+1/2}) = s_{j+1/2}(\mathbf{Q}^+_{j+1/2} - \mathbf{Q}^*_{j+1/2}) - s_{j+1/2}(\mathbf{Q}^*_{j+1/2} - \mathbf{Q}^-_{j+1/2}) = \sum_{p=1}^{2} \mathbf{Z}^p_{j+1/2}, \quad (6)$$

where

$$s_{j+1/2} = \max_p\{\max\{|\lambda^p(\mathbf{Q}^-_{j+1/2})|, |\lambda^p(\mathbf{Q}^+_{j+1/2})|\}\},$$

and

$$\begin{array}{rcl} \mathbf{Z}^2_{j+1/2} & = & s_{j+1/2}(\mathbf{Q}^+_{j+1/2} - \mathbf{Q}^*_{j+1/2}), \\ \mathbf{Z}^1_{j+1/2} & = & -s_{j+1/2}(\mathbf{Q}^*_{j+1/2} - \mathbf{Q}^-_{j+1/2}). \end{array} \quad (7)$$

We can express

$$\mathbf{Q}^*_{j+1/2} = \frac{1}{2s_{j+1/2}}[\mathbf{f}(\mathbf{Q}^-_{j+1/2}) - \mathbf{f}(\mathbf{Q}^+_{j+1/2})] + \frac{1}{2}(\mathbf{Q}^-_{j+1/2} + \mathbf{Q}^+_{j+1/2}), \quad (8)$$

and we define

$$\mathbf{A}^-(\Delta\mathbf{Q}_{j+1/2}) = \sum_{p=1, s^p_{j+1/2}<0}^{2} \mathbf{Z}^p_{j+1/2}, \qquad \mathbf{A}^+(\Delta\mathbf{Q}_{j+1/2}) = \sum_{p=1, s^p_{j+1/2}>0}^{2} \mathbf{Z}^p_{j+1/2}. \quad (9)$$

For evaluating $\mathbf{F}_{j+1/2} = \mathbf{f}(\mathbf{Q}^*_{j+1/2})$ we use the Rankine–Hugoniot jump condition in the form

$$\mathbf{f}(\mathbf{Q}^+_{j+1/2}) - \mathbf{f}(\mathbf{Q}^*_{j+1/2}) = s_{j+1/2}(\mathbf{Q}^+_{j+1/2} - \mathbf{Q}^*_{j+1/2}), \quad (10)$$

and together with (8) we get

$$\mathbf{F}_{j+1/2} = \mathbf{f}(\mathbf{Q}^*_{j+1/2}) = \frac{1}{2}[\mathbf{f}(\mathbf{Q}^-_{j+1/2} + \mathbf{f}(\mathbf{Q}^+_{j+1/2}] - \frac{1}{2}s_{j+1/2}(\mathbf{Q}^+_{j+1/2} - \mathbf{Q}^-_{j+1/2}). \quad (11)$$

This scheme we can derive from fully discrete form (3) where the x-axis is partitioned to subintervals of the following types

$$\langle x_{j-1/2,R}, x_{j+1/2,L}\rangle \quad \text{and} \quad \langle x_{j+1/2,L}, x_{j+1/2,R}\rangle,$$

where $x_{j+1/2,L} = x_{j+1/2} - s_{j+1/2}\Delta t$, $x_{j+1/2,R} = x_{j+1/2} + s_{j+1/2}\Delta t$. On these intervals we use the integral balance law (2). The points where the solution is discontinuous lie inside these intervals and this scheme is Riemann solver free.

The central-upwind schemes (for example in [1]) can be identified with HLL solver (see [2]). The decomposition has the form

$$\mathbf{f}(\bar{\mathbf{Q}}_{j+1}) - \mathbf{f}(\bar{\mathbf{Q}}_j) = s^2_{j+1/2}(\bar{\mathbf{Q}}_{j+1} - \bar{\mathbf{Q}}_{j+1/2}) + s^1_{j+1/2}(\bar{\mathbf{Q}}_{j+1/2} - \bar{\mathbf{Q}}_j) = \sum_{p=1}^{2} \mathbf{Z}^p_{j+1/2}, \quad (12)$$

where $s^1_{j+1/2} = a^+_{j+1/2}$, $s^1_{j+1/2} = a^-_{j+1/2}$ and

$$\begin{array}{rcl} \mathbf{Z}^2_{j+1/2} & = & s^2_{j+1/2}(\mathbf{Q}^+_{j+1/2} - \mathbf{Q}^*_{j+1/2}), \\ \mathbf{Z}^1_{j+1/2} & = & -s^1_{j+1/2}(\mathbf{Q}^*_{j+1/2} - \mathbf{Q}^-_{j+1/2}). \end{array} \tag{13}$$

As in the previous cases we can express $\mathbf{Q}^*_{j+1/2}$

$$\mathbf{Q}^*_{j+1/2} = \frac{\mathbf{f}(\mathbf{Q}^+_{j+1/2}) - \mathbf{f}(\mathbf{Q}^-_{j+1/2})}{s^1_{j+1/2} - s^2_{j+1/2}} + \frac{s^1_{j+1/2}\mathbf{Q}^-_{j+1/2} - s^2_{j+1/2}\mathbf{Q}^+_{j+1/2}}{s^1_{j+1/2} - s^2_{j+1/2}}. \tag{14}$$

We define

$$\mathbf{A}^-(\Delta\mathbf{Q}_{j+1/2}) = \sum_{p=1, s^p_{j+1/2}<0}^{2} \mathbf{Z}^p_{j+1/2}, \qquad \mathbf{A}^+(\Delta\mathbf{Q}_{j+1/2}) = \sum_{p=1, s^p_{j+1/2}>0}^{2} \mathbf{Z}^p_{j+1/2}. \tag{15}$$

The Relation (14) with the Rankine–Hugoniot jump condition in the form

$$\mathbf{f}(\mathbf{Q}^+_{j+1/2}) - \mathbf{f}(\mathbf{Q}^*_{j+1/2}) = s^2_{j+1/2}(\mathbf{Q}^+_{j+1/2} - \mathbf{Q}^*_{j+1/2}) \tag{16}$$

give us the following

$$\mathbf{F}_{j+1/2} = \mathbf{f}(\mathbf{Q}^*_{j+1/2}) = \frac{s^1_{j+1/2}\mathbf{f}(\mathbf{Q}^+_{j+1/2}) - s^2_{j+1/2}\mathbf{f}(\mathbf{Q}^-_{j+1/2})}{s^1_{j+1/2} - s^2_{j+1/2}} + \frac{s^1_{j+1/2}s^2_{j+1/2}}{s^1_{j+1/2} - s^2_{j+1/2}}(\mathbf{Q}^-_{j+1/2} - \mathbf{Q}^+_{j+1/2}). \tag{17}$$

As in the previous cases we can derive these schemes from fully discrete method (3) by limiting process ($\Delta t \to 0$). The x–axis is partitioned to subintervals of following types

$$\langle x_{j-1/2,R}, x_{j+1/2,L}\rangle \quad \text{and} \quad \langle x_{j+1/2,L}, x_{j+1/2,R}\rangle,$$

where $x_{j+1/2,L} = x_{j+1/2} - s^1_{j+1/2}\Delta t$, $x_{j+1/2,R} = x_{j+1/2} + s^2_{j+1/2}\Delta t$. In analogy with previous cases we formulate the integral balance law (2) on each of defined intervals. The solution is discontinuous in the points lying inside of these intervals and no Riemann problem we need to solve.

The previous schemes contain only one middle state $\mathbf{Q}^*_{j+1/2}$ between states $\mathbf{Q}^-_{j+1/2}$ and $\mathbf{Q}^+_{j+1/2}$. It is possible derive schemes with two or more than two middle states. For example, **the Roe solver** (see [3]) is based on the decomposition with (m-1) middle states

$$\mathbf{f}(\mathbf{Q}^+_{j+1/2}) - \mathbf{f}(\mathbf{Q}^-_{j+1/2}) = \sum_{p=1}^{m} s^p_{j+1/2}\mathbf{W}^p_{j+1/2}, \tag{18}$$

where $s^p_{j+1/2} = \lambda^p_{j+1/2}$ are eigenvalues and $\mathbf{r}^p_{j+1/2}$ are eigenvectors of the approximate Jacobian matrix, $s^1_{j+1/2} < s^2_{j+1/2} < \cdots < s^m_{j+1/2}$, $\mathbf{W}^p_{j+1/2} = \gamma^p_{j+1/2}\mathbf{r}^p_{j+1/2}$, $\gamma^p_{j+1/2} = \mathbf{R}^{-1}_{j+1/2}\Delta\mathbf{Q}_{j+1/2}$.

The middle states can be express in the following form

$$\mathbf{Q}^{p,*}_{j+1/2} = \mathbf{Q}^-_{j+1/2} + \sum_{k=1}^{p} \mathbf{W}^k_{j+1/2}. \tag{19}$$

Next we define

$$\mathbf{Z}^p_{j+1/2} = s^p_{j+1/2}\mathbf{W}^p_{j+1/2}. \tag{20}$$

and than the following holds

$$\mathbf{A}^-(\mathbf{Q}^-_{j+1/2},\mathbf{Q}^+_{j+1/2}) = \sum_{p=1,s^p_{j+1/2}<0}^{m} \mathbf{Z}^p_{j+1/2}, \qquad \mathbf{A}^+(\mathbf{Q}^-_{j+1/2},\mathbf{Q}^+_{j+1/2}) = \sum_{p=1,s^p_{j+1/2}>0}^{m} \mathbf{Z}^p_{j+1/2}. \quad (21)$$

From the conservativity and some relations (see [4]) we get the following results

$$\begin{array}{rcl} \mathbf{F}_{j+1/2} & = & \mathbf{f}(\mathbf{Q}^-_{j+1/2}) + \mathbf{A}^-(\mathbf{Q}^-_{j+1/2},\mathbf{Q}^+_{j+1/2}), \\ \mathbf{F}_{j-1/2} & = & \mathbf{f}(\mathbf{Q}^+_{j-1/2}) - \mathbf{A}^+(\mathbf{Q}^-_{j-1/2},\mathbf{Q}^+_{j-1/2}). \end{array} \quad (22)$$

The numerical flux function can be express in the form

$$\mathbf{F}_{j+1/2} = \mathbf{f}(\mathbf{Q}^*_{j+1/2}) = \frac{1}{2}[f(\mathbf{Q}^-_{j+1/2}) + f(\mathbf{Q}^+_{j+1/2})] - \frac{1}{2}|\mathbf{A}_{j+1/2}|(\mathbf{Q}^+_{j+1/2} - \mathbf{Q}^-_{j+1/2}) \quad (23)$$

This scheme can be derived in the same way as the previous. We define the partition of the x–axis

$$\langle x_{j-1/2,m}, x_{j+1/2,1}\rangle, \langle x_{j-1/2,m}, x_{j+1/2,1}\rangle, \langle x_{j+1/2,1}, x_{j+1/2,2}\rangle, \ldots, \langle x_{j+1/2,m-1}, x_{j+1/2,m}\rangle,$$

where $x_{j+1/2,p} = x_{j+1/2} + s^p_{j+1/2}\Delta t$. The speeds $s^p_{j+1/2}$ was getting from linearized problem and it cannot be said that the discontinuities lie inside of the intervals. It is not possible to interpret this scheme as a scheme without Riemann solver.

4 Conclusion

It was shown that all described schemes can be understood in the same way and it is possible to construct them by two different manners. The first starts from general formulation of semidiscrete method. It is formulated decomposition based on generalized Rankine-Hugoniot condition. Than it is possible to formulate the scheme in fluctuation form. For the scheme in conservation form it is used the classical Rankine-Hugoniot condition and the numerical flux for semidiscrete scheme is constructed from these relations. The second uses adaptive dividing x-axis based on speeds of the waves. The scheme can be interpreted as Riemann free only in the certain cases.

Acknowledgement: This work has been supported by the Research Plan MSM 4977751301 and by Moravian-Silesian region.

References

[1] A. Kurganov, S. Noelle, G.Petrova: *Semidiscrete Central-Upwind Schemes for Hyperbolic Conservation Laws and Hamilton-Jacobi Equations.* Journal of Scientific Computation, 2001, Vol. 23, No. 3, pp. 707–740.

[2] A. Harten, P. D. Lax, B. Van Leer: *On Upstream Differencing and Godunov-Type Schemes for Hyperbolic Conservation Laws.* Society for Industrial and Applied Mathematics, 1983. SIAM Review, Vol. 25, No 1, pp. 35–61

[3] P. L. Roe: *Approximate Riemann Solvers, Parameter Vectors, and Difference Schemes.* Journal of Computational Physics 135, 1997. pp. 250–258

[4] M. Brandner, J. Egermaier, H. Kopincová: *Numerical Modelling of River Flow (Numerical Schemes for One Type of Nonconservative Systems).* Proceedings of Seminars PANM 14, Dolní Maxov, 2008, pp. 23–36, http://www.math.cas.cz/ panm/

Fixing Nodes Method for Stabilization of Generalized Inverse Arising in Total FETI Algorithms

T. Brzobohatý, P. Kabelíková, T. Kozubek, A. Markopoulos

VSB - Technical University of Ostrava

1 Introduction

A typical example where we can exploit generalized inverse is a system of consistent linear equations with symmetric positive semidefinite (SPS) matrix arising in the stress analysis of a "floating" static structure whose essential boundary conditions are not sufficient to prevent its rigid body motions [3, 4, 8]. This system can be solved by standard direct methods for the solution of systems with positive definite matrices, such as the Cholesky decomposition, adapted to the solution of systems with only positive semidefinite matrix. The only modification comprises setting to zero the columns which correspond to zero pivots. However, in agreement with the theoretical results of Pan [7], it turns out that it is very difficult to recognize the positions of such pivots in the presence of rounding errors when the nonsingular part of A is ill-conditioned. Due to the rounding errors, the main difficulty in implementation of the FETI method is effective elimination of the displacements, in particular evaluation of the action of generalized inverse of the SPS stiffness matrices of "floating" subdomains. To alleviate this problem, Farhat and Géradin [3] proposed to combine the Cholesky decomposition with the SVD decomposition of a relatively small matrix. The method was developed further by Papadrakakis and Fragakis [8]. An improved modification of Farhat-Géradin algorithm is proposed by Dostál, Kozubek, Markopoulos, Brzobohatý in [2]. This modification based on the active choice of the SVD part uses fixing nodes strategy to make the system as stiff as possible and has been implemented in our Total FETI solver. This solver uses the Lagrange multipliers not only for gluing of the subdomains along auxiliary interfaces, but also for implementation of the essential boundary conditions; first considered by Felipa, Park, Justino, and Gumaste [4]; then by Dostál, Horák, and Kučera [1]. The main advantage of this approach is that it makes all the subdomains floating, so that the null spaces of the stiffness matrices are a priori known.

2 Stable Computation of the Generalized Inverse Matrix

Let us consider the problem $Ax = b$, with symmetric positive semidefinite matrix (SPS) of the order n and with $b \in ImA$. Thus a solution $x = A^+b$ exists, where A^+ denotes a generalized inverse matrix. We shall assume that the sparsity pattern of A enables its effective triangular decomposition $A = LL^T$. The method of evaluation of the factor L is known as the *Cholesky factorization.*

In the following, we assume that A is an SPS stiffness matrix of a floating 2D or 3D elastic body. If we choose M mesh nodes that are neither near each other nor placed near any line, $M < N$, $M \geq 2$ in 2D, and $M \geq 3$ in 3D, then the submatrix A_{JJ} of the stiffness matrix A defined by the set J with the indices of the displacements of the other nodes is "reasonably" nonsingular. Of course, this is not surprising, as A_{JJ} can be considered as the stiffness matrix of the body that is fixed in the chosen nodes. Using the arguments of mechanics, it is natural to assume that if fixing of the chosen nodes makes the body uniformly stiff, then A_{JJ} is well-conditioned.

Our starting point is the following decomposition of the matrix $A \in R^{n\times n}$:

$$PAP^T = \begin{bmatrix} \widetilde{A}_{JJ} & \widetilde{A}_{JI} \\ \widetilde{A}_{IJ} & \widetilde{A}_{II} \end{bmatrix} = \begin{bmatrix} L_{JJ} & O \\ L_{IJ} & I \end{bmatrix} \begin{bmatrix} L_{JJ}^T & L_{IJ}^T \\ O & S \end{bmatrix}, \tag{1}$$

where $L_{JJ} \in R^{r\times r}$ is a lower factor of the Cholesky factorization of $\widetilde{A}_{JJ}$, $L_{IJ} \in R^{s\times r}$, $L_{IJ} = \widetilde{A}_{IJ}L_{JJ}^{-T}$, $S \in R^{s\times s}$ is a singular matrix, $r = n - s$ and s is the number of displacements corresponding to the fixing nodes. Finally, P is a permutation matrix which corresponds to both preserving sparsity and fixing nodes reordering.

Then

$$A^+ = P^T \begin{bmatrix} L_{JJ}^{-T} & -L_{JJ}^{-T}L_{IJ}^T S^\dagger \\ O & S^\dagger \end{bmatrix} \begin{bmatrix} L_{JJ}^{-1} & O \\ -L_{IJ}L_{JJ}^{-1} & I \end{bmatrix} P, \tag{2}$$

where $S^\dagger$ denotes the Moore–Penrose generalized inverse, $S^\dagger = V\Sigma^\dagger U^T$, computed by the SVD of matrix S of the defect d, where $U, V \in R^{s\times s}$ are orthogonal matrices, $UU^T = I$, $VV^T = I$, $\Sigma^\dagger = \mathrm{diag}\{\sigma_1^{-1}, \ldots, \sigma_{s-d}^{-1}, 0, \ldots, 0\} \in R^{s\times s}$ and $\sigma_1 \geq \cdots \geq \sigma_{s-d} > \sigma_{s-d+1} = \cdots = \sigma_s = 0$ are singular values of S.

To find P, we shall proceed in two steps. We first form a permutation matrix P_1 to decompose A into blocks

$$P_1^T AP_1 = \begin{bmatrix} A_{JJ} & A_{JI} \\ A_{IJ} & A_{II} \end{bmatrix}, \tag{3}$$

where the submatrix A_{JJ} is nonsingular and A_{II} corresponds to the degrees of freedom of the M fixing nodes. Then we apply a suitable reordering algorithm on $P_1^T AP_1$ to get a permutation matrix P_2 which leaves the part A_{II} without changes and enables the sparse Cholesky factorization of A_{JJ}. Further, we decompose PAP^T with $P = P_2P_1$ as in (1). To preserve sparsity we may use well-known sparse reordering algorithms such as SYMAMD, SYMRCM, SLOAN etc.

3 Detection of Fixing Nodes for Generalized Inverse

Next we show how to find the mesh fixing nodes to make the system as stiff as possible, i.e., to minimize condition number of the regular part A_{JJ}.

3.1 Fixing Nodes as Graph Centers

In our programs we work with the adjacency matrix D_A of the original mesh corresponding to the matrix A. Conditioning of the regular part A_{JJ} and A^+ seems to be related to positioning of fixing nodes in the original mesh such that the Dirichlet conditions imposed in the fixing nodes make the structure as stiff as possible.

We have tested different positions (variants (a)-(d)) of the fixing nodes in the mesh of the 3D elastic body depicted in Figure 1. The testing criterion was the regular condition number denoted as κ. In our case, $\kappa = \overline{\mathrm{cond}}(A^+) = \lambda_{max}/\lambda_{min}$, where λ_{max} and λ_{min} correspond to the largest and the nonzero smallest eigenvalues, respectively. As we have three-dimensional problem, the minimum number of fixing nodes are three to prevent rigid body motions in all three directions. As we can see in Figure 1, the best result is obtained when we place the fixing nodes inside the object as uniformly as possible (see the variant (d)). This result leads to idea to consider the problem of finding the fixing nodes as the problem of *finding graph centers*.

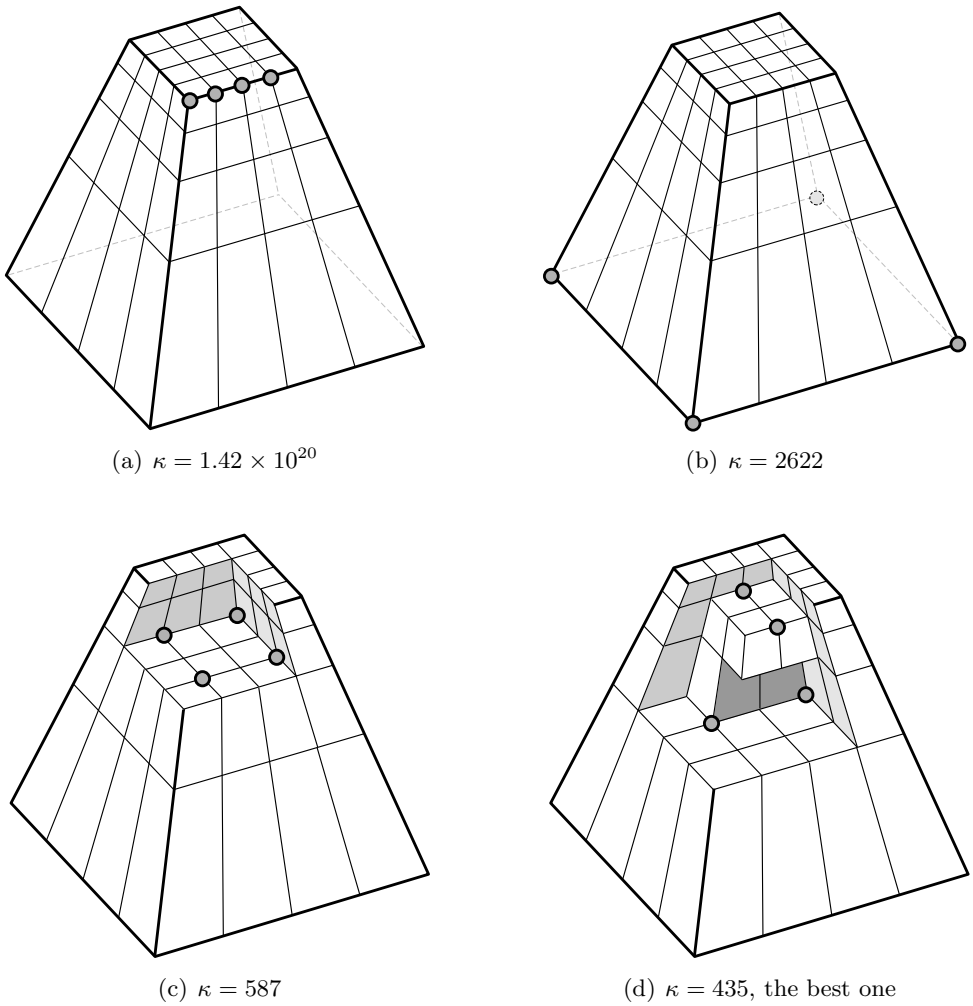
By extension of the common definition of graph center to a set of vertices we get the following definition.

Definition 1 (A "graph center" as a set of k vertices)

$$\min_{\substack{C\subset V(G)\\|C|=k}} \max_{v\in V(G)} dist(C,v) = \min_{\substack{C\subset V(G)\\|C|=k}} \max_{v\in V(G)} \left(\min_{x\in C} dist(x,v) \right), \tag{4}$$

where k is the number of graph centers, $V(G)$ is a vertex set of a graph G, $dist(x,v)$ is a distance between vertices x and v (length of the shortest path between those vertices).

In general, there could be more k-sets of vertices that fit Definition 1 but not all of them fit the requirement on minimum condition number. Thus, we have to remark that the regular condition number of A^+ depends mainly on the graph topology and only slightly on the geometry of the mesh. A method of finding fixing nodes as graph centers is described in [5].



(a) $\kappa = 1.42 \times 10^{20}$ (b) $\kappa = 2622$

(c) $\kappa = 587$ (d) $\kappa = 435$, the best one

Figure 1: Pyramid: dependance of $\kappa = \overline{\mathrm{cond}}(A^+)$ on positioning of fixing nodes.

Natural requirement to these nodes is that they are not near any straight line and not close to each other. The results of experiments also agree with the intuitive rule that placing the fixing nodes inside the body can result in more stable generalized inverse than placing them at the corners as in the FETI-DP methods.

3.2 Fast Algorithm for Fixing Nodes Finding

We do not strictly require the optimal solution. A sub-optimal solution obtained in a short time suffices for purposes of fast computation of generalized inverse.

In our software, we use the following algorithm consisting of two steps:

1. Dividing the graph into k parts using some suitable graph/mesh partitioning software (for example METIS, see [6]).
2. Finding one graph center in each part using the results of spectral theory. From the vertices that fit the basic definition we choose the nearest vertex to the geometrical center.

Our experiments show that the spectral theory is very powerful case for finding graph center. Especially, we use the (Perron) eigenvector corresponding to the largest eigenvalue of the adjacency matrix D_A. The maximum entry of this eigenvector (in absolute value) corresponds to the graph center. Finding the eigenvector of a sparse symmetric adjacency matrix D_A using some iterative method such as power method or Lanczos method is very fast comparing to the standard graphs methods.

Acknowledgement: This work has been supported by the grant GAČR 101/08/0574 and by the project of Ministry of Education of the Czech Republic MSM6198910027.

References

[1] Z. Dostál, D. Horák, R. Kučera. *Total FETI – an easier implementable variant of the FETI method for numerical solution of elliptic PDE.* Communications in Numerical Methods in Engineering, 2006, 22: 1155–1162.

[2] T.Brzobohatý, Z. Dostál, T. Kozubek. A. Markopoulos, *Combining Cholesky decomposition with SVD to stable evaluation of a generalised inverse of the stiffness matrix of a floating structure*, 2009, submited.

[3] C. Farhat, M. Géradin. *On the general solution by a direct method of a large scale singular system of linear equations: application to the analysis of floating structures.* International Journal for Numerical Methods in Engineering, 1998, 41:675–696.

[4] C.A. Felipa, K.C. Park. *The construction of free–free flexibility matrices for multilevel structural analysis.* Computer Methods in Applied Mechanics and Engineering, 2002, 191:2111-2140.

[5] P. Kabelíková. *Graph centers used for stabilization of matrix factorizations*, 2009, submited.

[6] G. Karypis, V. Kumar. *METIS manual Version 4.0*, University of Minnesota, 1998. http://glaros.dtc.umn.edu/gkhome/views/metis (online)

[7] C.T. Pan. *On the existence and factorization of rank–revealing LU factorizations.* Linear Algebra and Its Applications, 2000, 316:199-222.

[8] M. Papadrakakis, Y. Fragakis. *An integrated geometric–algebraic method for solving semi-definite problems in structural mechanics.* Computer Methods in Applied Mechanics and Engineering, 2001, 190:6513–6532.

Adaptive wavelet methods for two-dimensional elliptic operator equations

D. Černá, V. Finěk

Technical University in Liberec

1 Introduction

In recent years adaptive wavelet methods have been successfully used for solving partial differential as well as integral equations, both linear and nonlinear. It has been shown that these methods converge and that they are asymptotically optimal in the sense that storage and number of floating point operations, needed to resolve the problem with desired accuracy, remain proportional to the problem size when the resolution of the discretization is refined. Thus, the computational complexity for all steps of the algorithm is controlled.

The effectiveness of adaptive wavelet methods is strongly influenced by the choice of a wavelet basis, in particular by the condition of the basis. In our contribution, we compare the number of iterations needed to resolve the problem with desired accuracy for wavelet bases adapted to homogeneous Dirichlet boundary conditions of the first order from [1, 4]. Numerical examples are presented for two-dimensional elliptic problems with a singular right-hand side.

2 Adaptive wavelet scheme

In this section, we briefly review adaptive wavelet methods for the elliptic operator equations similar to the method proposed by Cohen, Dahmen and DeVore in [2, 3].

Let H be a real Hilbert space with the inner product $\langle\cdot,\cdot\rangle_H$ and the induced norm $\|\cdot\|_H$. Let $A : H \to H'$ be the selfadjoint and H-elliptic operator, i.e.

$$a(v,w) := \langle Av, w\rangle \lesssim \|v\|_H \|w\|_H \quad \text{and} \quad a(v,v) \sim \|v\|_H^2 . \tag{1}$$

By the Lax-Milgram theorem, A is an isomorphism from H to H', i.e. there exist positive constants c_A and C_A such that

$$c_A \|v\|_H \le \|Av\|_{H'} \le C_A \|v\|_H , \quad v \in H. \tag{2}$$

Therefore, the equation

$$Au = f \tag{3}$$

has for any $f \in H'$ a unique solution. If (2) holds, then (3) is called *well-posed* (on H). Typical examples are second order elliptic boundary value problems with homogeneous Dirichlet boundary conditions on some open domain $\Omega \subset \mathbb{R}^d$. In this case $H = H_0^1(\Omega)$ and $H' = H^{-1}(\Omega)$. Other examples are for instance singular integral equations on the boundary $\partial\Omega$ with $H = H^{-1/2}(\partial\Omega)$, $H' = H^{1/2}(\partial\Omega)$.

Thus H is typically a Sobolev space. In the following, we assume that

$$H \subset L^2 \subset H' \quad \text{or} \quad H' \subset L^2 \subset H. \tag{4}$$

We assume that $\mathbf{D}^{-t}\Psi$, $\Psi = \{\psi_\lambda, \lambda \in \mathcal{J}\}$, is a wavelet basis in the energy space H. Thus, we have

$$c_\psi \|\mathbf{v}\|_{l^2} \leq \left\|\mathbf{v}^T \mathbf{D}^{-t}\Psi\right\|_H \leq C_\psi \|\mathbf{v}\|_{l^2}, \quad \mathbf{v} \in l^2(\mathcal{J}), \tag{5}$$

where $c_\psi > 0$. Then the original equation (3) can be reformulated as an equivalent biinfinite matrix equation

$$\mathbf{A}\mathbf{u} = \mathbf{f}, \tag{6}$$

where $\mathbf{A} = \mathbf{D}^{-t} \langle A\Psi, \Psi \rangle \mathbf{D}^{-t}$ is a diagonally preconditioned stiffness matrix, $u = \mathbf{u}^T D^{-t}\Psi$ and $\mathbf{f} = \mathbf{D}^{-t} \langle f, \Psi \rangle$.

Under the above assumptions, u solves (3) if and only if $\mathbf{u}$ solves the matrix equation (6). Moreover, the matrix $\mathbf{A}$ satisfies

$$\|\mathbf{A}\| \leq \frac{C_\psi^2 C_A}{c_\psi^2 c_A} < +\infty. \tag{7}$$

As an immediate consequence all finite sections

$$\mathbf{A}_\Lambda := \mathbf{D}^{-t} \langle A\Psi_\Lambda, \Psi_\Lambda \rangle \mathbf{D}^{-t}, \quad \Psi_\Lambda := \{\psi_\lambda, \lambda \in \Lambda\}, \quad \Lambda \subset \mathcal{J}, \tag{8}$$

have uniformly bounded condition numbers

$$\kappa(\mathbf{A}_\Lambda) \leq \frac{C_\psi^2 C_A}{c_\psi^2 c_A}, \quad \Lambda \subset \mathcal{J}. \tag{9}$$

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

1. One starts with a variational formulation but instead of turning to a finite dimensional approximation, using the suitable wavelet basis the continuous problem is transformed into an infinite-dimensional l^2-problem (6), which is well-conditioned.
2. One then tries to devise a *convergent iteration* for the l^2-problem.
3. Finally, one derives a practicle version of this idealized iteration. All infinite-dimensional quantities have to be replaced by finitely supported ones and the routine for the application of the biinfinite-dimensional matrix $\mathbf{A}$ approximately have to be designed.

The simplest convergent iteration for the l^2-problem is a *Richardson iteration* which has the following form:

$$\mathbf{u}_0 := 0, \quad \mathbf{u}_{n+1} := \mathbf{u}_n + \omega(\mathbf{f} - \mathbf{A}\mathbf{u}_n), \quad n = 0, 1, \ldots. \tag{10}$$

For the convergence, the relaxation parameter ω has to satisfy

$$\rho := \|\mathbf{I} - \omega\mathbf{A}\|_{\mathcal{L}(l^2)} < 1. \tag{11}$$

Then the iteration (10) convergence with an error reduction per step

$$\|\mathbf{u}_{n+1} - \mathbf{u}\|_{l^2} \leq \rho \|\mathbf{u}_n - \mathbf{u}\|_{l^2}. \tag{12}$$

In the case that $\mathbf{A}$ is symmetric and positive definite, then (11) is satisfied if

$$0 < \omega < \frac{2}{\lambda_{max}}, \tag{13}$$

where λ_{max} is the largest eigenvalue of $\mathbf{A}$. It is known that the optimal relaxation parameter is given by

$$\hat{\omega} = \frac{2}{\lambda_{min} + \lambda_{max}}, \tag{14}$$

where λ_{min} is the smallest eigenvalue of $\mathbf{A}$. For $\hat{\omega}$ the estimate of the error reduction can be computed as

$$\rho(\hat{\omega}) = \frac{\lambda_{max} - \lambda_{min}}{\lambda_{max} + \lambda_{min}} = \frac{\kappa(\mathbf{A}) - 1}{\kappa(\mathbf{A}) + 1} = 1 - \frac{1}{\kappa(\mathbf{A}) + 1} \leq 1 - \frac{1}{\frac{C_\psi^2 C_A}{c_\psi^2 c_A} + 1}. \tag{15}$$

We use the following implementable version of the ideal iteration (10). It was proved that such an algorithm converge and is asymptotically optimal.

SOLVE [A, f, ϵ] $\rightarrow \mathbf{u}_\epsilon$

Let $\theta < 1/3$ and $K \in \mathbb{N}$ be fixed such that $3\rho^K < \theta$.

1. Set $j := 0$, $\mathbf{u}_0 := 0$, $\epsilon_0 := \left\|\mathbf{A}^{-1}\right\|_{\mathcal{L}(l^2)} \|\mathbf{f}\|_{l^2}$.

2. While $\epsilon_j > \epsilon$ do

   $j := j + 1$,

   $\epsilon_j := 3\rho^K \epsilon_{j-1}/\theta$,

   $\mathbf{f}_j := \mathbf{RHS}[\mathbf{f}, \frac{\theta\epsilon_j}{6\omega K}]$,

   $\mathbf{z}_0 := \mathbf{u}_{i-1}$,

   For $l = 1, \ldots, K$ do

   $$\mathbf{z}_l := \mathbf{z}_{l-1} + \omega\left(\mathbf{f}_j - \mathbf{APPLY}[\mathbf{A}, \mathbf{z}_{l-1}, \tfrac{\theta\epsilon_j}{6\omega K}]\right),$$

   end for,

   $\mathbf{u}_j := \mathbf{COARSE}[\mathbf{z}_K, (1-\theta)\epsilon_j]$,

end while,

$\mathbf{u}_\epsilon := \mathbf{u}_j$.

For the subroutines **RHS, APPLY**, and **COARSE** we refer to [2].

3 Numerical examples

Quantitative behaviour of the above algorithm depends on the used wavelet basis, namely on its condition. The optimally conditioned linear and quadratic wavelet bases were constructed in [4]. In [1] we propose a construction which leads to optimally conditioned wavelet bases also in the cubic case. In this section, our intention is to compare the quantitative behaviour of the adaptive wavelet method for cubic wavelet bases from [1] and [4].

Example 1 *We consider two-dimensional Poisson equation*

$$-\Delta u = f, \quad in \quad \Omega = (0,1)^2, \quad \partial\Omega = 0, \tag{16}$$

with the solution u given by

$$u\left(x,y\right)=16\frac{\left(e^{40x}-1\right)\left(e^{40y}-1\right)}{e^{40}-1}\left(1-\frac{e^{40x}-1}{e^{40}-1}\right)\left(1-\frac{e^{40y}-1}{e^{40}-1}\right),\quad (x,y)\in\Omega. \tag{17}$$

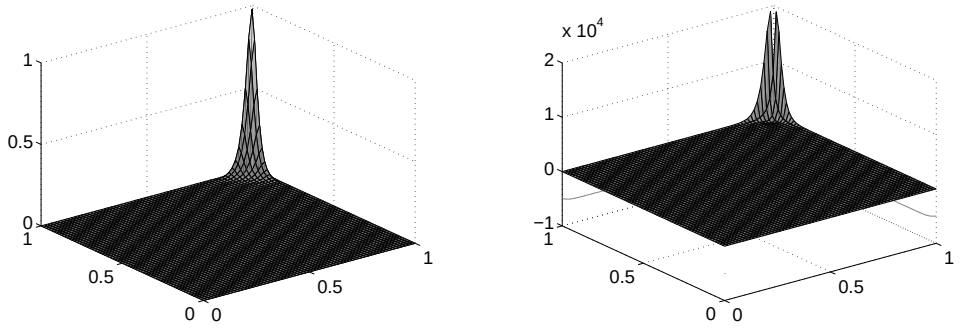


Figure 1: The solution and the right-hand side of the equation (16)

We use the above adaptive wavelet scheme with the cubic wavelet basis adapted to homogeneous Dirichlet boundary conditions of the first order from [1, 4].

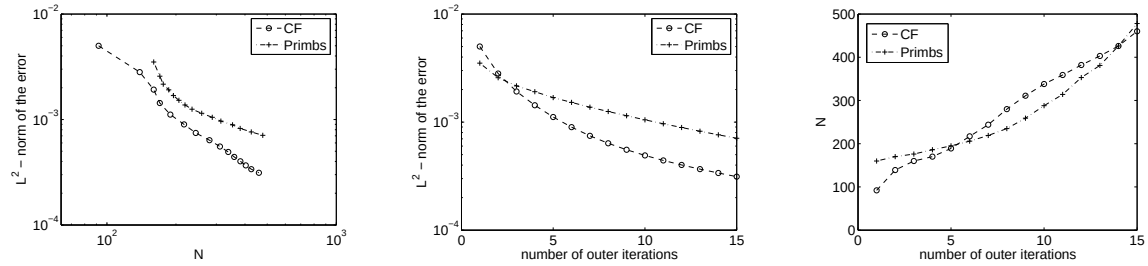


Figure 2: The convergence history for wavelet bases from [1, 4], the number of basis functions is denoted by N.

Acknowledgement: This work has been supported by the Ministry of Education, Youth and Sports of the Czech Republic through the Research Centers LC06024 and 1M06047.

References

[1] D. Černá; V. Finěk: *Optimized Construction of Biorthogonal Spline-Wavelets,* in: ICNAAM 2008 (Simos T.E. et al., eds.), AIP Conference Proceedings **1048**, American Institute of Physics, New York, 2008, pp. 134-137.

[2] A. Cohen, W. Dahmen; R. DeVore: *Adaptive Wavelet Schemes for Elliptic Operator Equations - Convergence Rates.* Math. Comput. **70** (2001), 27-75.

[3] A. Cohen, W. Dahmen; R. DeVore: *Adaptive Wavelet Methods II - Beyond the elliptic case.* Found. Math. **2** (2002), 203-245.

[4] M. Primbs: *New Stable Biorthogonal Spline-Wavelets on the Interval.* Preprint, Universität Duisburg-Essen, 2007.

Matematické modelování kompozitních materiálů s nedokonalým rozhraním složek

P. Gruber, J. Zeman

České vysoké učení technické v Praze

1 Úvod

Moderní inženýrské konstrukce si žádají moderní materiály a ty zase vývoj v jejich matematickém modelování. Dnes je hlavní pozornost zaměřena na materiály kompozitní, které jsou schopny využít vynikajících mechanických vlastností v nich použitých složek a naopak jejich negativní mechanické vlastnosti potlačit. Tato práce ukazuje jednu z možných cest v matematickém modelování kompozitních materiálů – víceúrovňový matematický model založený na homogenizační teorii, jež na mikroúrovni využívá numerickou metodu doménové dekompozice Finite Element Tearing and Interconnecting (FETI) method.

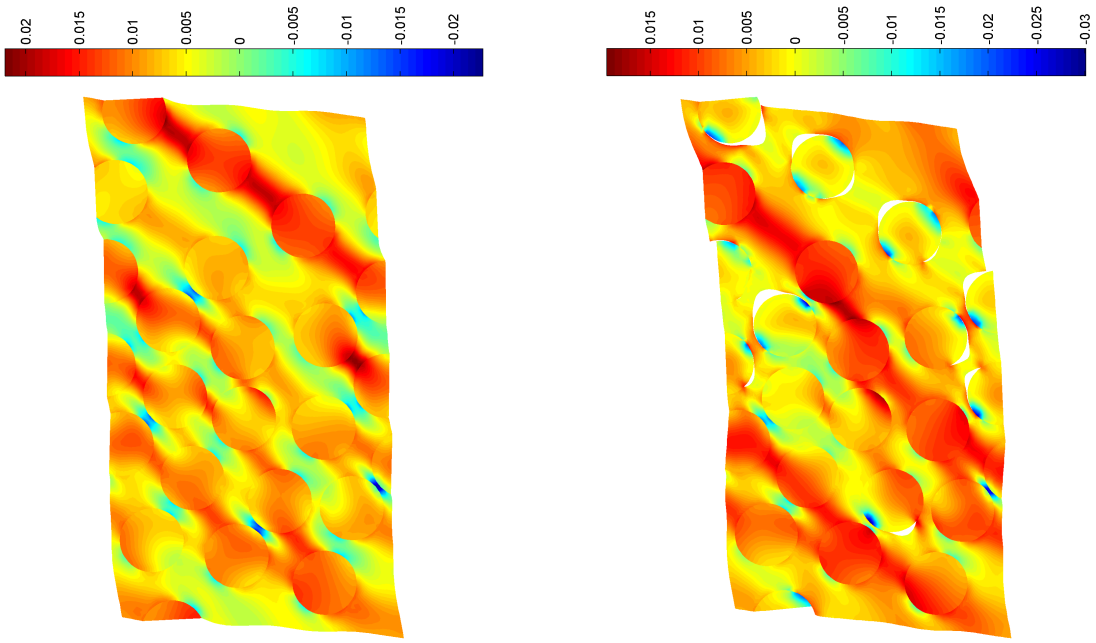
2 Mikroskopický průměr vláken – homogenizace

Vlákna moderních kompozitních materiálů (např. uhlíková) mají příčný rozměr v řádech mikrometrů, tedy zanedbatelný v porovnání s rozměry konstrukce z tohoto materiálu vytvořené. Při výběru matematického modelu výše uvedené konstrukce je v dnešní době přirozená volba modelu numerického, konkrétně modelu založeném na metodě konečných prvků (MKP). Z výpočetních důvodů je zřejmě nemožné diskretizovat model konstrukce natolik jemnou sítí, která by nám umožnila přímo modelovat heterogenní mikrostrukturu materiálu. Tento problém je řešen homogenizační teorií periodických mikrostruktur [5], která rozdělí pohled na konstrukci na mikroměřítko (v němž popisujeme mikrostrukturu materiálu pomocí periodické jednotkové buňky) a makroměřítko (na němž je popsána geometrie celé konstrukce). Výsledkem aplikace homogenizační teorie jsou efektivní materiálové charakteristiky homogenního materiálu (nejen lineárně pružného), kterým lze nahradit daný materiál heterogenní a získat velmi přesnou makroskopickou odezvu konstrukce na předepsané zatížení.

3 Nedokonalé spojení složek – metoda FETI

Spojení vlákna s matricí na jejich rozhraní však není dokonalé, má zřejmě omezenu pevnost a jsou zde přítomny také počáteční imperfekce. Nedokonalé spojení složek na rozhraní způsobí nelineární odezvu materiálu a není tak možné odhadnout efektivní vlastnosti celé konstrukce v jednom kroku. Homogenizační teorie však nabízí možnost paralelního výpočtu. Pokud chceme jev rozpojování složek do výpočtu zahrnout, je nutné ho uvažovat již na mikroúrovni. K modelování rozhraní je vhodné použít metodu FETI [7], pomocí níž lze modelovat dokonale spojené či rozpojené složky, předepsat konstitutivní zákon na rozhraní a to včetně aplikace kontaktní úlohy. Z inženýrského pohledu lze říci, že se jedná o metodu, která algoritmizuje přechod z deformační metody, ve které jsou v tomto případě neznámé posuny uzlů v diskretizované mikroúrovni – celé jednotkové buňce, k silové metodě s neznámými silami pouze v uzlech na rozhraní složek. Více

informací o modelování rozhraní složek pomocí metody FETI z několika *nezávislých* pohledů a od různých autorů lze najít například v [1, 2, 3, 4].



Obrázek 1: Smykové porušení mikrostruktury (jednotkové buňky) reálného kompozitu s křehkým rozhraním před a po překonání počáteční pevnosti. Izolinie znázorňují hlavní tahová napětí. Deformace 20x zvětšena.

4 Závěr

Homogenizační teorie je vhodnou cestou k matematickému modelování reálných konstrukcí z kompozitních materiálů a ve spojení s metodou FETI, použitou k numerickému modelování na mikroúrovni – jednotkové buňky, se jeví jako velice efektivní. Doposud je náš výzkum soustředěn především na možnosti, které nabízí metoda FETI v oblasti modelování konstitutivního vztahu na rozhraní složek. Z tohoto hlediska se jako vhodná strategie jeví kombinace metody FETI buďto s teorií izotropního porušování, nebo sekvenční lineární analýzou (SLA) [6]. Obrázek 1 znázorňuje porušení mikrostruktury (jednotkové buňky) reálného kompozitu s křehkým rozhraním, které bylo modelováno pomocí zde naznačených principů.

Poděkování: Práce vznikla za podpory projektů GAČR 106/08/1379 a GAČR 106/07/1244.

Literatura

[1] Z. Dostál, David Horák, Oldřich Vlach: *FETI-based algorithms for modelling of fibrous composite materials with debonding*. In: Mathematics and Computers in Simulation, 76, 1–3, pp. 57–64, 2007.

[2] J. Kruis, Z. Bittnar: *Reinforcement-matrix Interaction Modelled by FETI Method*. In: 17th International Conference on Domain Decomposition Methods, Linz, 2006.

[3] O. Vlach: *Modelování kompozitů pomocí řešičů založených na dualitě*. In: Diplomová práce, VŠB, Ostrava, 2001.

[4] P. Gruber: *Homogenizace kompozitů s možností nedokonalého spojení složek*. In: Diplomová práce, ČVUT, Praha, 2007.

[5] J. C. Michel, H. Moulinec, P. Suquet: *Effective properties of composite materials with periodic microstructure: a computational approach*. In: Computer Methods in Applied Mechanics and Engineering, 172, 1–4, 1999, pp. 109–143.

[6] Matthew J. DeJong, Max A.N. Hendriks, Jan G. Rots: *Sequentially linear analysis of fracture under non-proportional loading*. In: Engineering Fracture Mechanics, 75, 18, 2008, pp. 5042–5056.

[7] J. Kruis: *Domain Decomposition Methods for Distributed Computing*. Saxe-Coburg Publications, 2006.

On the Worst Scenario Method: A Modified Convergence Theorem and Its Application to an Uncertain Differential Equation

P. Harasim

Institute of Geonics AS CR, Ostrava

1 Introduction

We propose a theoretical framework for solving a class of worst scenario problems. The existence of the worst scenario is proved through the convergence of a sequence of approximate worst scenarios. The main convergence theorem modifies and corrects the relevant results already published in the literature. The theoretical framework is applied to a particular problem with an uncertain boundary value problem for a nonlinear ordinary differential equation with an uncertain coefficient.

Quasilinear elliptic boundary value problems with uncertain coefficients were studied in [3, 4, 7, 8], see also [5, Chapter III]. In these works the coefficient of the state equation is a u-dependent function. The state problem that has motivated this work is different: the coefficient is a function of the squared derivative of the state solution u. Equations of this kind describe some electromagnetic phenomena, fluid flow phenomena, and the elastoplastic deformation of a body, see [9, page 212].

2 Worst scenario problem

Let V be a real, separable, and reflexive Banach space and let V^* denote its dual space. We deal with the following nonlinear operator state equation

$$A(a)u = b, \qquad u \in V, \tag{1}$$

where $A(a) : V \to V^*$, $b \in V^*$. We assume that the operator $A(a)$ depends on a parameter a that belongs to a set of admissible input parameters $\mathcal{U}_{\rm ad} \subset U$, where U is a Banach space. We assume that

(i) the set $\mathcal{U}_{\rm ad}$ is compact in U;

(ii) a unique state solution $u(a)$ of equation (1) exists for any parameter $a \in \mathcal{U}_{\rm ad}$;

(iii) a criterion-functional $\Phi : \mathcal{U}_{\rm ad} \times V \to \mathbb{R}$ is given such that :
if $a_n \in \mathcal{U}_{\rm ad}$, $a_n \to a$ in U and $v_n \to v$ in V as $n \to \infty$, then

$$\Phi(a_n, v_n) \to \Phi(a, v).$$

The goal is to solve the following worst scenario maximization problem: Find $a^0 \in \mathcal{U}_{\rm ad}$ such that

$$a^0 = \arg \max_{a \in \mathcal{U}_{\rm ad}} \Phi(a, u(a)). \tag{2}$$

We will prove the existence of a solution to problem (2) by means of a sequence of solutions to approximate worst scenario problems.

We resort to a discretization of both the set $\mathcal{U}_{\rm ad}$ and the space V. Let $\mathcal{U}_{\rm ad}^M \subset \mathcal{U}_{\rm ad} \subset U$ be a finite-dimensional approximation of the set $\mathcal{U}_{\rm ad}$ and let V_h be a finite-dimensional subspace of V. Let us consider the Galerkin approximation $u_h(a) \in V_h$ of the state solution $u(a)$. We set the following approximate worst scenario problem: Find $a_h^{M0} \in \mathcal{U}_{\rm ad}^M$ such that

$$a_h^{M0} = \arg \max_{a^M \in \mathcal{U}_{\rm ad}^M} \Phi(a^M, u_h(a^M)). \tag{3}$$

Next, we assume that

(iv) the set $\mathcal{U}_{\rm ad}^M$ is compact in U;

(v) for any $a \in \mathcal{U}_{\rm ad}$, there exists a unique Galerkin approximation $u_h(a)$ of the state solution $u(a)$;

(vi) if $a_n \in \mathcal{U}_{\rm ad}$ and $a_n \to a$ in U as $n \to \infty$, then $u_h(a_n) \to u_h(a)$ in V_h;

(vii) if $a_n \in \mathcal{U}_{\rm ad}$, $a_n \to a$ in U as $n \to \infty$, and if $h_n \to 0$ as $n \to \infty$, then $u_{h_n}(a_n) \to u(a)$ in V, where $\{u_{h_n}(a_n)\}$ is an n-controlled sequence of the Galerkin approximations;

(viii) for any $a \in \mathcal{U}_{\rm ad}$, there exists a sequence $\{a^M\}$, $a^M \in \mathcal{U}_{\rm ad}^M$, $M \to \infty$, such that $a^M \to a$ in U as $M \to \infty$.

To show that the approximate worst scenario problem (3) has at least one solution, we can proceed analogously to the proof of [5, Theorem 3.3].

3 Main Result

Theorem 1 *Let $\{V_h\}$, $h \to 0$, be a sequence of finite-dimensional subspaces of the space V. For any fixed $h > 0$, let $\{a_h^{M0}\}$, where $a_h^{M0} \in \mathcal{U}_{\rm ad}^M$ and $M \to \infty$, be a sequence of solutions to the approximate worst scenario problem (3). Let the assumptions (i)-(viii) be fulfilled. Then a sequence $\{a_{h_n}^{M_n 0}\}$, $a_{h_n}^{M_n 0} \in \mathcal{U}_{\rm ad}^{M_n}$, exists such that $h_n \to 0$ and $M_n \to \infty$ as $n \to \infty$, and*

$$a_{h_n}^{M_n 0} \to a^0 \qquad \text{in } U, \tag{4}$$

$$u_{h_n}(a_{h_n}^{M_n 0}) \to u(a^0) \qquad \text{in } V, \tag{5}$$

$$\Phi\big(a_{h_n}^{M_n 0}, u_{h_n}(a_{h_n}^{M_n 0})\big) \to \Phi\big(a^0, u(a^0)\big) \tag{6}$$

as $n \to \infty$, where $a^0 \in \mathcal{U}_{\rm ad}$ solves problem (2) and $u(a^0)$ is the corresponding state solution mentioned in (ii).

Proof: We fix a subspace V_h for a while and consider a sequence $\{a_h^{M0}\}$, $a_h^{M0} \in \mathcal{U}_{\rm ad}^M$, $M \to \infty$, i.e., a sequence of the solutions of the approximate worst scenario problem (3). Since $\{a_h^{M0}\} \subset \mathcal{U}_{\rm ad}$ and $\mathcal{U}_{\rm ad} \subset U$ is compact, a convergent subsequence $\{a_h^{M_k 0}\} \subset \{a_h^{M0}\}$ exists such that

$$a_h^{M_k 0} \to a_h^0 \qquad \text{in } U \quad \text{as} \quad k \to \infty, \tag{7}$$

where $a_h^0 \in \mathcal{U}_{\rm ad}$. By virtue of assumption (vi) of the previous section, we obtain

$$u_h(a_h^{M_k 0}) \to u_h(a_h^0) \qquad \text{in } V_h \quad \text{as} \quad k \to \infty. \tag{8}$$

Let $a \in \mathcal{U}_{\rm ad}$ be arbitrary and chosen independently of h. It follows from assumption (viii) that there exists a sequence $\{a^M\}$, $a^M \in \mathcal{U}_{\rm ad}^M$, such that

$$a^M \to a \qquad \text{in } U \quad \text{as} \quad M \to \infty. \tag{9}$$

By virtue of assumption (vi), we infer

$$u_h(a^M) \to u_h(a) \qquad \text{in } V_h \quad \text{as} \quad M \to \infty. \tag{10}$$

For any k, it holds

$$\Phi\big(a_h^{M_k 0}, u_h\big(a_h^{M_k 0}\big)\big) \geq \Phi\big(a^{M_k}, u_h\big(a^{M_k}\big)\big). \tag{11}$$

By virtue of (7)-(10), and assumption (iii), we obtain

$$\Phi(a_h^0, u_h(a_h^0)) \geq \Phi(a, u_h(a)). \tag{12}$$

Inequality (12) is valid for any $h > 0$.

Let us release h and consider the sequences $\{a_h^0\}$, $\{u_h(a_h^0)\}$, and $\{u_h(a)\}$, where $h \to 0$. Since $\{a_h^0\} \subset \mathcal{U}_{\rm ad}$ and $\mathcal{U}_{\rm ad} \subset U$ is compact, there exists a convergent subsequence $\{a_{h_l}^0\} \subset \{a_h^0\}$, $h_l \to 0$ as $l \to \infty$, such that

$$a_{h_l}^0 \to a^0 \qquad \text{in } U \quad \text{as} \quad l \to \infty, \tag{13}$$

where $a^0 \in \mathcal{U}_{\rm ad}$. By virtue of assumption (vii), we get for the corresponding sequence of the Galerkin approximations

$$u_{h_l}(a_{h_l}^0) \to u(a^0) \qquad \text{in } V \quad \text{as} \quad l \to \infty. \tag{14}$$

If we set $a_n := a \in \mathcal{U}_{\rm ad}$ for $n = 1, 2, \ldots$, then it follows from assumption (vii) that

$$u_{h_l}(a) \to u(a) \qquad \text{in } V \quad \text{as} \quad l \to \infty. \tag{15}$$

By virtue of (12)–(15), and assumption (iii), we obtain

$$\Phi(a^0, u(a^0)) \geq \Phi(a, u(a)). \tag{16}$$

The inequalities (11), (12), and (16), hold for any $a \in \mathcal{U}_{\rm ad}$, so that a^0 is a solution of problem (2).

The existence of the sequence $\{a_{h_n}^{M_n 0}\}$ appearing in (4) is a direct consequence of the existence of the solution a^0. By virtue of assumption (vii) we infer (5), and by assumption (iii), we obtain (6). □

4 Application

In this section, we show an application of the proposed theoretical framework to the following boundary value problem: Find a function $u \in C^1(\overline{\Omega}) \bigcap C^2(\Omega)$ such that

$$-(a(u'^2)u')' = f \quad \text{in} \quad \Omega, \tag{17}$$

$$u = 0 \quad \text{on} \quad \Gamma, \tag{18}$$

where $\Omega = (0,1)$, $\Gamma = \{0,1\}$, a is a Lipschitz continuous function on $\mathbb{R}_0^+$ (nonnegative real numbers), and $f \in C(\Omega)$. The prime stands for $\mathrm{d}u/\mathrm{d}x$.

For more detailed treatment, see [2].

Instead of (17)–(18), we will deal with the following weakly formulated problem: Find $u \in H_0^1(\Omega)$ such that

$$\int_0^1 a(u'^2)u'v'\mathrm{d}x = \int_0^1 fv\,\mathrm{d}x \qquad \forall v \in H_0^1(\Omega), \tag{19}$$

where $H_0^1(\Omega)$ is usual Sobolev space, $f \in L^2(\Omega)$. We assume that the function a belongs to the admissible set

$$\mathcal{U}_{\mathrm{ad}} := \big\{a \in \mathcal{U}_{\mathrm{ad}}^0 : 0 < a_{\min} \le a(x) \le a_{\max} \quad \forall x \in \mathbb{R}_0^+\big\},$$

which models the uncertainty in a and where

$$\mathcal{U}_{\mathrm{ad}}^0 := \Big\{a \in C^{(0),1}(\mathbb{R}_0^+) : 0 \le \frac{\mathrm{d}a}{\mathrm{d}x} \le C_{\mathrm{L}} \quad \text{a.e.,} \quad a(x) = a(x_{\mathrm{C}}) \quad \text{for} \quad x \ge x_{\mathrm{C}}\Big\},$$

C_{L}, $a_{\min}$, $a_{\max}$, x_{C} are positive constants, and $C^{(0),1}(\mathbb{R}_0^+)$ stands for the Lipschitz continuous functions defined on $\mathbb{R}_0^+$.

We observe that $\mathcal{U}_{\mathrm{ad}} \subset U$, where U is the Banach space of functions continuous on $\mathbb{R}_0^+$ and constant for $x \ge x_{\mathrm{C}}$, with the norm $\|w\|_U := \max\limits_{x\in[0,x_{\mathrm{C}}]} |w(x)|$ for $w \in U$.

The operator equation (1) stems from (19) if we set $V := H_0^1(\Omega)$ and define $A(a) : V \to V^*$ and $b \in V^*$ by

$$\langle A(a)u, v\rangle := \int_0^1 a(u'^2)u'v'\,\mathrm{d}x,$$

$$\langle b, v\rangle := \int_0^1 fv\,\mathrm{d}x,$$

where $u, v \in V$.

Let us define the set $\mathcal{U}_{\mathrm{ad}}^M \subset \mathcal{U}_{\mathrm{ad}}$ and a finite- dimensional space V_h. Let T_i, $i = 1,\dots,M$, are equally spaced points in $[0, x_C]$, $T_1 = 0$ and $T_M = x_C$.

$$\mathcal{U}_{\mathrm{ad}}^M := \big\{a \in \mathcal{U}_{\mathrm{ad}} : a|_{[T_i,T_{i+1}]} \in P_1([T_i, T_{i+1}]),\, i = 1,\dots,M-1\big\},$$

where $P_1([T_i, T_{i+1}])$ denotes the linear polynomials on the interval $[T_i, T_{i+1}]$.

To approximate the space V, we introduce points $x_0, x_1, \dots, x_{N+1}$ into the interval $[0,1]$, $x_0 = 0$, $x_{N+1} = 1$. We define the discretization parameter h as

$$h := \max_{i=1,\dots,N+1}(x_i - x_{i-1}).$$

The space V_h is defined as

$$V_h := \{v_h \in V : v_h|_{[x_i,x_{i+1}]} \in P_1([x_i, x_{i+1}]),\, i = 0,\dots,N\}.$$

To be able to apply the Theorem 1, we have to verify its assumptions. By the Arzelà–Ascoli theorem [10, page 35] the assumptions (i) and (iv) of Section 2 are fulfilled.

The operator A is continuous [2, Lemma 4.1], strongly monotonie [2, Lemma 4.2] and coercive [2, the proof of Theorem 4.1] on V. It follows from [11, Theorem 2.K] that the problem (19) has

a solution, the uniqueness of the state solution follows from [11, p. 93, Corollary 1]; see also [2, Theorem 4.1]. Thus, the assumption (ii) is fulfilled.

The assumptions (v), (vi), (vii) and (viii) are also fulfilled, see [2, Theorem 4.2, Theorem 4.3, Theorem 4.4, Lemma 4.5].

Acknowledgments. The author would like to thank Dr. J. Chleboun for his great help during the work. This work has been supported by the project MSM4781305904 from the Ministry of Education, Youth and Sports of the Czech Republic and by the research plan AV0Z30860518 of Institute of Geonics AS CR.

References

[1] J. Franců: *Monotone operators (A survey directed to applications to differential equations).* Appl. Math., 35, 1990, 257-301.

[2] P. Harasim: *On the worst scenario method: a modified convergence theorem and its application to an uncertain differential equation.* Appl. Math., 53, 2008, 583- 598.

[3] I. Hlaváček: *Reliable solution of a quasilinear nonpotential elliptic problem of a nonmonotone type with respect to uncertainty in coefficients.* J. Math. Anal. Appl., 212, 1997, 452-466.

[4] I. Hlaváček: *Reliable solution of elliptic boundary value problems with respect to uncertain data.* Nonlinear Anal., 30, 1997, 3879-3890.

[5] I. Hlaváček, J. Chleboun, and I. Babuška: *Uncertain Input Data Problems and the Worst Scenario Metod.* Elsevier, Amsterdam, 2004.

[6] I. Hlaváček, M. Křížek, and J. Malý: *On Galerkin approximations of a quasilinear nonpotential elliptic problem of a nonmonotone type.* J. Math. Anal. Appl. 184, 1994, 168-189.

[7] J. Chleboun: *Reliable solution for a 1D quasilinear elliptic equation with uncertain coefficients.* J. Math. Anal. Appl., 234, 1999, 514-528.

[8] J. Chleboun: *On a reliable solution of a quasilinear elliptic equation with uncertain coefficients: Sensitivity analysis and numerical examples.* Nonlinear Anal. 44, 2001, 375-388.

[9] M. Křížek, P. Neittaanmäki: *Finite Element Approximation of Variational Problems and Applications.* Longman Scientific & Technical, New York, 1990

[10] E. Zeidler: *Applied Functional analysis (Applications to Mathematical Physics).* Springer-Verlag, New York, 1995

[11] E. Zeidler: *Applied Functional analysis (Main Principles and Their Applications).* Springer-Verlag, New York, 1995

Fictitious domain method for linear elasticity

J. Haslinger, T. Kozubek, R. Kučera

Charles University in Prague
VSB-Technical University of Ostrava
VSB-Technical University of Ostrava

1 Introduction

The contribution deals with numerical realization of elliptic boundary value problems arising in linear elasticity by a fictitious domain method. Any fictitious domain formulation [2] extends the original problem defined in a domain ω to a new (fictitious) domain Ω with a simple geometry (e.g. a box) which contains $\overline{\omega}$. The main advantage consists in the fact that an uniform mesh can be constructed on $\overline{\Omega}$. Consequently, the stiffness matrix has a structure that enables us to use highly efficient multiplying procedures. We will apply multiplying procedures based on a correspondence between circulant matrices and the discrete Fourier transform (DFT).

The original fictitious domain method based on Lagrange multipliers [1] enforces boundary conditions by Lagrange multipliers defined on the boundary of the original domain γ. Therefore the fictitious domain solution has a singularity on γ that can result in an intrinsic error of the computed solution. Our modified version [3] uses an auxiliary curve Γ located outside of $\overline{\omega}$, on which we introduce a new control variable in order to satisfy the boundary conditions on γ. In this case the singularity is moved away from $\overline{\omega}$ so that the computed solution is smoother in ω. We have illustrated experimentally in [3] that the discretization error is significantly smaller in the second case and corresponding rate of convergence is higher.

2 Formulation of the problem

Let us consider an elastic body represented by a bounded domain $\omega \subset \mathbb{R}^2$ with the sufficiently smooth boundary γ consisting of two disjoint parts γ_u and γ_p, $\gamma = \overline{\gamma}_u \cup \overline{\gamma}_p$ (see Figure 4.1). The zero displacements are prescribed on γ_u while surface tractions of density $\boldsymbol{p} \in (L^2(\gamma_p))^2$ act on γ_p. Finally we suppose that the body ω is subject to volume forces of density $\boldsymbol{f}_{|\omega}$, $\boldsymbol{f} \in (L^2_{loc}(\mathbb{R}^2))^2$. We seek a displacement field $\boldsymbol{u}$ in ω satisfying the *equilibrium equation* and the *Dirichlet and Neumann boundary conditions*:

$$\left.\begin{aligned} -\mathrm{div}\,\boldsymbol{\sigma}(\boldsymbol{u}) &= \boldsymbol{f} \quad \text{in} \quad \omega, \\ \boldsymbol{u} &= \boldsymbol{0} \quad \text{on} \quad \gamma_u, \\ \boldsymbol{\sigma}(\boldsymbol{u})\boldsymbol{\nu} &= \boldsymbol{p} \quad \text{on} \quad \gamma_p, \end{aligned}\right\} \tag{1}$$

where $\boldsymbol{\sigma}(\boldsymbol{u})$ is the stress tensor in ω and $\boldsymbol{\nu}$ stands for the unit outward normal vector to γ. The stress tensor is related to the linearized strain tensor $\boldsymbol{\varepsilon}(\boldsymbol{u}) := 1/2(\nabla \boldsymbol{u} + \nabla^\top \boldsymbol{u})$ by the Hooke law for linear isotropic materials:

$$\boldsymbol{\sigma}(\boldsymbol{u}) := c_1 \mathrm{tr}(\boldsymbol{\varepsilon}(\boldsymbol{u}))\boldsymbol{I} + 2c_2\boldsymbol{\varepsilon}(\boldsymbol{u}) \quad \text{in} \quad \omega,$$

where "tr" denotes the trace of matrices, $\boldsymbol{I} \in \mathbb{R}^{2\times 2}$ is the identity matrix and $c_1, c_2 > 0$ are the Lamè constants.

Denote

$$\mathbb{V}(\omega) = \{\boldsymbol{v} \in (H^1(\omega))^2 |\ \boldsymbol{v} = \boldsymbol{0} \text{ on } \gamma_u\}.$$

The *weak formulation* of (1) reads as follows:

$$\textit{Find } \boldsymbol{u} \in \mathbb{V}(\omega) \textit{ such that } \ a_\omega(\boldsymbol{u}, \boldsymbol{v}) = f_\omega(\boldsymbol{v}) + (\boldsymbol{p}, \boldsymbol{v})_{\gamma_p} \quad \forall \boldsymbol{v} \in \mathbb{V}(\omega), \tag{2}$$

where

$$a_\omega(\boldsymbol{u}, \boldsymbol{v}) = \int_\omega \boldsymbol{\sigma}(\boldsymbol{u}) : \boldsymbol{\varepsilon}(\boldsymbol{v})\, \mathrm{d}\boldsymbol{x}, \quad f_\omega(\boldsymbol{v}) = \int_\omega \boldsymbol{f} \cdot \boldsymbol{v}\, \mathrm{d}\boldsymbol{x}$$

and $(\cdot,\cdot)_{\gamma_p}$ is the scalar product in $(L^2(\gamma_p))^2$.

Let us consider a box Ω such that $\overline{\omega} \subset \Omega$ and construct a closed curve Γ surrounding ω (see Figure 4.1). Instead of (2), we propose to solve the following *fictitious domain formulation* of (1) in Ω:

$$\left.\begin{array}{l} \textit{Find } (\hat{\boldsymbol{u}}, \boldsymbol{\lambda}) \in (H^1_{per}(\Omega))^2 \times \boldsymbol{\Lambda}(\Gamma) \textit{ such that} \\ a_\Omega(\hat{\boldsymbol{u}}, \boldsymbol{v}) + b_\Gamma(\boldsymbol{\lambda}, \boldsymbol{v}) = f_\Omega(\boldsymbol{v}) \quad \forall \boldsymbol{v} \in (H^1_{per}(\Omega))^2, \\ b_{\gamma_u}(\boldsymbol{\mu}_u, \hat{\boldsymbol{u}}) = 0 \quad \forall \boldsymbol{\mu}_u \in \boldsymbol{\Lambda}(\gamma_u), \\ b_{\gamma_p}(\boldsymbol{\mu}_p, \boldsymbol{\sigma}(\hat{\boldsymbol{u}})\boldsymbol{\nu}) = b_{\gamma_p}(\boldsymbol{\mu}_p, \boldsymbol{p}) \quad \forall \boldsymbol{\mu}_p \in \boldsymbol{\Lambda}(\gamma_p), \end{array}\right\} \tag{3}$$

where $H^1_{per}(\Omega)$ is the space of periodic functions from $H^1(\Omega)$; $\boldsymbol{\Lambda}(\Gamma) := (H^{-1/2}(\Gamma))^2$, $\boldsymbol{\Lambda}(\gamma_u) := (H^{-1/2}(\gamma_u))^2$, $\boldsymbol{\Lambda}(\gamma_p) := (H^{1/2}(\gamma_p))^2$ and b_Γ, b_{γ_u}, b_{γ_p} are the respective duality pairings between these spaces and their duals. It is readily seen that $\hat{\boldsymbol{u}}_{|\omega}$ solves (2).

3 Algebraic solvers

A discretization of (3) based on a mixed finite element method leads typically to the following algebraic *saddle-point* problem: find a pair $(u, \lambda) \in \mathbb{R}^{2n} \times \mathbb{R}^{2m}$ such that

$$\left(\begin{array}{c|c} A & B_\Gamma^\top \\ \hline B_{\gamma_u} & 0 \\ C_{\gamma_p} & 0 \end{array}\right) \left(\begin{array}{c} u \\ \lambda \end{array}\right) = \left(\begin{array}{c} f \\ \hline 0 \\ p \end{array}\right), \tag{4}$$

where $A \in \mathbb{R}^{2n\times 2n}$ is the stiffness matrix, $B_\Gamma \in \mathbb{R}^{2m\times 2n}$ and $B_{\gamma_u} \in \mathbb{R}^{2m_u\times 2n}$ are the Dirichlet trace matrices on Γ and γ_u, respectively, $C_{\gamma_p} \in \mathbb{R}^{2m_p\times 2n}$ is the Neumann trace matrix (representing the trace of $\boldsymbol{\sigma}(\boldsymbol{u})\boldsymbol{\nu}$) on γ_p, $f \in \mathbb{R}^{2n}$, $p \in \mathbb{R}^{2m_p}$ and $m = m_u + m_p$.

The system (4) can be solved by the algorithm presented in [3] that combines the Schur complement reduction with the null-space method. It requires a multiplying procedure to perform the matrix-vector products $A^\dagger y$, where $A^\dagger$ is a generalized inverse to A and $y \in \mathbb{R}^{2n}$. Let us note that A is singular due to the presence of $H^1_{per}(\Omega)$ in (3). On the other hand, the periodic boundary condition on $\partial\Omega$ leads to a block circulant structure of A that enables us to handle the spectral decomposition of A by the DFT. Therefore one can evaluate $A^\dagger y$ by the FFT-algorithm without necessity to assemble and store A.

We introduce the main ideas of our multiplying procedure. First note that the differential operator in (1) reads as follows:

$$\operatorname{div} \boldsymbol{\sigma}(\boldsymbol{u}) = \left(\begin{array}{c|c} (c_1+2c_2)\dfrac{\partial^2 u_1}{\partial x_1^2} + c_2\dfrac{\partial^2 u_1}{\partial x_2^2} & (c_1+c_2)\dfrac{\partial^2 u_2}{\partial x_1 \partial x_2} \\ \hline (c_1+c_2)\dfrac{\partial^2 u_1}{\partial x_1 \partial x_2} & c_2\dfrac{\partial^2 u_2}{\partial x_1^2} + (c_1+2c_2)\dfrac{\partial^2 u_2}{\partial x_2^2} \end{array}\right),$$

where $\boldsymbol{u} = (u_1, u_2)$. Let us consider equidistant partitions of the sides of $\Omega := (0, l_1) \times (0, l_2)$ into n_1, n_2 segments with stepsizes $h_1 = l_1/n_1$, $h_2 = l_2/n_2$, respectively. Thus, Ω is partitioned into $n := n_1 n_2$ rectangles. On such a partition we define the finite element subspace of $H^1_{per}(\Omega)$ formed by piecewise bilinear functions. Then the stiffness matrix A takes the form:

$$A = \left(\begin{array}{c|c} (c_1 + 2c_2)A_1 \otimes M_2 + c_2 M_1 \otimes A_2 & (c_1 + c_2)B_1 \otimes B_2 \\ \hline (c_1 + c_2)B_1 \otimes B_2 & c_2 A_1 \otimes M_2 + (c_1 + 2c_2)M_1 \otimes A_2 \end{array} \right), \tag{5}$$

where $A_k, M_k, B_k \in \mathbb{R}^{n_k \times n_k}$ are the circulants with the first columns $a_k, m_k, b_k \in \mathbb{R}^{n_k}$, $a_k = \frac{1}{h_k}(2, -1, 0, \ldots, 0, -1)^\top$, $m_k = \frac{h_k}{6}(4, 1, 0, \ldots, 0, 1)^\top$, $b_k = \frac{1}{2}(0, -1, 0, \ldots, 0, 1)^\top$, $k = 1, 2$, respectively, and $\otimes$ stands for the Kronecker product. It is well-known that the eigenvalues of any circulant can be obtained by the DFT of its first column while the eigenvectors are the columns of the inverse to the DFT matrix [2]. Introducing notation X_k for the DFT matrix of order n_k, we can write $A_k = X_k^{-1} D_{A_k} X_k$, $M_k = X_k^{-1} D_{M_k} X_k$, $B_k = X_k^{-1} D_{B_k} X_k$, where D_{A_k}, D_{M_k}, D_{B_k}, $k = 1, 2$, are the respective diagonal matrices of eigenvalues. Substituting into (5), we obtain:

$$A = \left(\begin{array}{c|c} X_1^{-1} \otimes X_2^{-1} & 0 \\ \hline 0 & X_1^{-1} \otimes X_2^{-1} \end{array} \right) \left(\begin{array}{c|c} D_{11} & D_{12} \\ \hline D_{21} & D_{22} \end{array} \right) \left(\begin{array}{c|c} X_1 \otimes X_2 & 0 \\ \hline 0 & X_1 \otimes X_2 \end{array} \right), \tag{6}$$

where $D_{11} = (c_1 + 2c_2)D_{A_1} \otimes D_{M_2} + c_2 D_{M_1} \otimes D_{A_2}$, $D_{22} = c_2 D_{A_1} \otimes D_{M_2} + (c_1 + 2c_2)D_{M_1} \otimes D_{A_2}$, $D_{12} = (c_1 + c_2)D_{B_1} \otimes D_{B_2}$, $D_{21} = D_{12}$. Denote D the second matrix on the right hand-side of (6). The generalized inverse $A^\dagger$ may be obtained replacing D by $D^\dagger$ in (6). Let us note that the actions of $D^\dagger$ can be easily performed using the following factorization of D:

$$D = \left(\begin{array}{c|c} I & 0 \\ \hline D_{21} D_{11}^\dagger & I \end{array} \right) \left(\begin{array}{c|c} D_{11} & 0 \\ \hline 0 & D_{22} - D_{21} D_{11}^\dagger D_{12} \end{array} \right) \left(\begin{array}{c|c} I & D_{11}^\dagger D_{12} \\ \hline 0 & I \end{array} \right), \tag{7}$$

where $D_{11}^\dagger = \mathrm{diag}(\tilde{d}_1, \ldots, \tilde{d}_n)$ with $\tilde{d}_i = 1/d_i$, if $d_i \neq 0$, and $\tilde{d}_i = 0$, if $d_i = 0$. Taking into account the fact that all blocks in (7) are diagonal, we obtain the following result.

Lemma 3.1 *Let n_1 and n_2 be powers of two. Then the matrix-vector product $A^\dagger v$, $v \in \mathbb{R}^{2n}$, can be evaluated by the total complexity $\mathcal{O}(4n \log_2 n + 4n)$.*

4 Numerical experiments

Let ω be given by the interior of the circle (see Figure 4.1):

$$\omega = \{(x, y) \in \mathbb{R}^2 |\ (x - 0.5)^2 + (y - 0.5)^2 < 0.3^2\}$$

and $\Omega = (0, 1) \times (0, 1)$. The right hand-side in (1) are chosen as $\boldsymbol{f} = -\mathrm{div}\, \boldsymbol{\sigma}(\hat{\boldsymbol{u}})$, $\boldsymbol{p} = \boldsymbol{\sigma}(\hat{\boldsymbol{u}})\boldsymbol{\nu}$, where $\hat{\boldsymbol{u}}(x, y) = (0.1 \ln(x + y + 1), 0.1xy)$, $(x, y) \in \mathbb{R}^2$. The approximation of $H^1_{per}(\Omega)$ in (3) has been described in the previous section while $\boldsymbol{\Lambda}(\gamma_u)$, $\boldsymbol{\Lambda}(\gamma_p)$ and $\boldsymbol{\Lambda}(\Gamma)$ are replaced by their subspaces of piecewise constant functions on partitions of polygonal approximations of γ_u, γ_p and Γ, respectively. The stepsizes H on γ_u, γ_p and Γ are chosen to guarantee the requirement $\dim \boldsymbol{\Lambda}(\gamma_u) + \dim \boldsymbol{\Lambda}(\gamma_p) = \dim \boldsymbol{\Lambda}(\Gamma)$. The auxiliary boundary Γ is constructed by shifting γ four h units in the direction of the outward normal vector with $h := h_1 = h_2$. The original and deformed geometries are depicted in Figure 4.2 and the difference between the exact and computed displacements is shown in Figure 4.3 for $h = 1/256$.

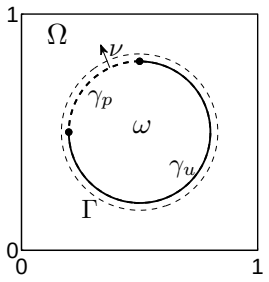


Figure 4.1: Geometry of ω.

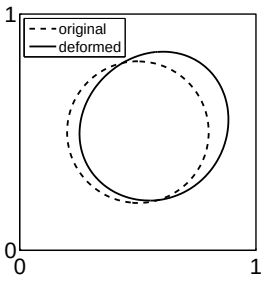


Figure 4.2: Original and deformed geometry.

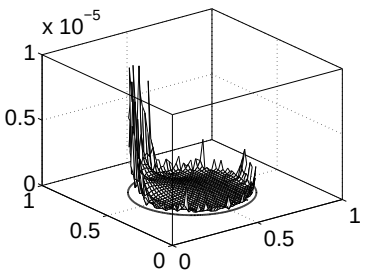


Figure 4.3: Differ. $|\hat{\boldsymbol{u}}_h - \hat{\boldsymbol{u}}|$ in ω.

In Table 4.1, we report the number of primal ($2n$) and control ($2m$) variables, the number of BiCGSTAB iterations, the computational time and the relative errors in the following norms:

$$\mathrm{Err}_{(L_2(\omega))^2} = \frac{\|\hat{\boldsymbol{u}}_h - \hat{\boldsymbol{u}}\|_{(L_2(\omega))^2}}{\|\hat{\boldsymbol{u}}\|_{(L_2(\omega))^2}},\ \mathrm{Err}_{(H^1(\omega))^2} = \frac{\|\hat{\boldsymbol{u}}_h - \hat{\boldsymbol{u}}\|_{(H^1(\omega))^2}}{\|\hat{\boldsymbol{u}}\|_{(H^1(\omega))^2}},\ \mathrm{Err}_{(L_2(\gamma))^2} = \frac{\|\hat{\boldsymbol{u}}_h - \hat{\boldsymbol{u}}\|_{(L_2(\gamma))^2}}{\|\hat{\boldsymbol{u}}\|_{(L_2(\gamma))^2}}.$$

From the computed errors, we determine the convergence rates of the fictitious domain solution in the $(L_2(\omega))^2$, $(H^1(\omega))^2$ and $(L_2(\gamma))^2$-norm, respectively. We consider partitions with the non-constant ratio of stepsizes $H/h = |\log_2(h)|$ found experimentally which leads to a smooth behavior of the approximations of control variables as $H \to 0+$.

Table 4.1: Results of the FD approach (3).

Step h	$2n/2m$	Iters.	C.time[s]	$\mathrm{Err}_{(L^2(\omega))^2}$	$\mathrm{Err}_{(H^1(\omega))^2}$	$\mathrm{Err}_{(L^2(\gamma))^2}$
1/64	8450/44	20	0.2808	4.2348e-004	5.2662e-001	9.7813e-004
1/128	33282/68	19	0.39	1.7261e-004	3.3539e-001	3.4267e-004
1/256	132098/124	34	2.371	3.8171e-005	1.5851e-001	1.4673e-004
1/512	526338/212	46	16.26	1.0374e-005	8.2440e-002	2.9814e-005
1/1024	2101250/384	77	109	4.7117e-006	5.5679e-002	1.1683e-005
Convergence rates:				1.7036	0.8508	1.6298

Acknowledgement: This work has been supported by the grant 101/08/0574 of the Grant Agency of the Czech Republic, by the grant IAA100750802 of the Grant Agency of the Czech Academy of Sciences and by the Research Project MSM6198910027 of the Czech Ministry of Education.

References

[1] Girault, V.; Glowinski, R.: Error analysis of a fictitious domain method applied to a Dirichlet problem. Japan J. Indust. Appl. Math. 12(1995), 487-514.

[2] Golub, G. H.; Van Loan, C. F.: Matrix computation, 3rd ed. The Johns Hopkins University Press, Baltimore 1996.

[3] Haslinger, J.; Kozubek, T.; Kučera; Peichl, G.: Projected Schur complement method for solving non-symmetric systems arising from a smooth fictitious domain approach. Lin. Algebra Appl., 14(2007), 713-739.

Použití T-FETI pro řešení 3D kvazistatických kontaktních úloh s Coulombovým třením

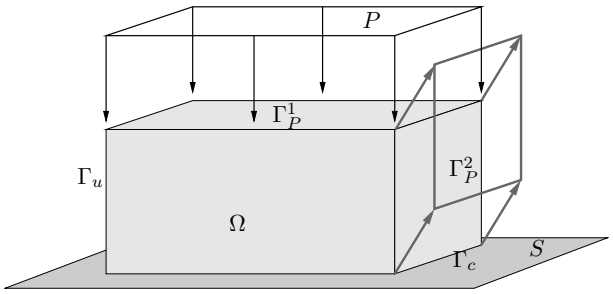
J. Haslinger, O. Vlach, R. Kučera

Univerzita Karlova, Praha
VŠB-Technická univerzita Ostrava

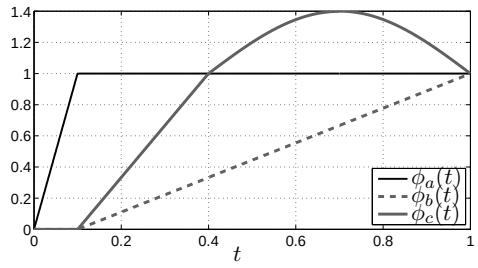
Řešíme kontaktní úlohu pružného tělesa ležícím na tuhém podloží. Klasické okrajové podmínky jsou z tohoto důvodu rozšířeny o podmínky nepronikání a podmínky Coulombova tření. Statický případ je poměrně dobře prozkoumán. Je však známo, že výsledná deformace po ustálení působících sil nezávisí pouze na jejich konečném stavu, ale také na historii jejich průběhu, což popisuje mimo jiné kvazistatický model. Ten předpokládá, že deformace při změně působících sil je okamžitá, a také zanedbává setrvačné síly.

Vhodná časová diskretizace kvazistatického modelu vede na posloupnost statických úloh, u kterých je zatížení modifikováno o člen závislý na řešení z předchozí časové úrovně. Tento přístup teoreticky analyzuje Rocca a Coccu v [4]. Stejně jako v [2] nahrazujeme jednotlivé statické úlohy s Coulombovým třením posloupností úloh s daným třením, ve kterých je postupně iterována mez skluzu. Užitím Lagrangeových multiplikátorů které odstraňují jednostranné okrajové podmínky a regularizují třecí člen, přeformulujeme úlohy s daným třením do duální podoby. Diskretizace vede na minimizaci kvadratické funkce s jednoduchými a se separovatelnými kvadratickými omezeními. Užitím vhodné metody rozložení oblastí pro 3D úlohy (zde T-FETI [1]), navíc zefektivníme násobení duální maticí, které zahrnuje řešení soustavy s maticí tuhosti. Za cenu rozšíření neznámých o další Lagrangeovy multiplikátory (na které není kladeno omezení) stačí řešit pouze soustavy s maticemi tuhostí jednotlivých podoblastí. Pro minimizaci používáme algoritmus [3]. Výsledky navíc porovnáme s řešením pomocí nehladké Newtonovy metody.

Numerický příklad



Obrázek 1: Geometrie modelové úlohy

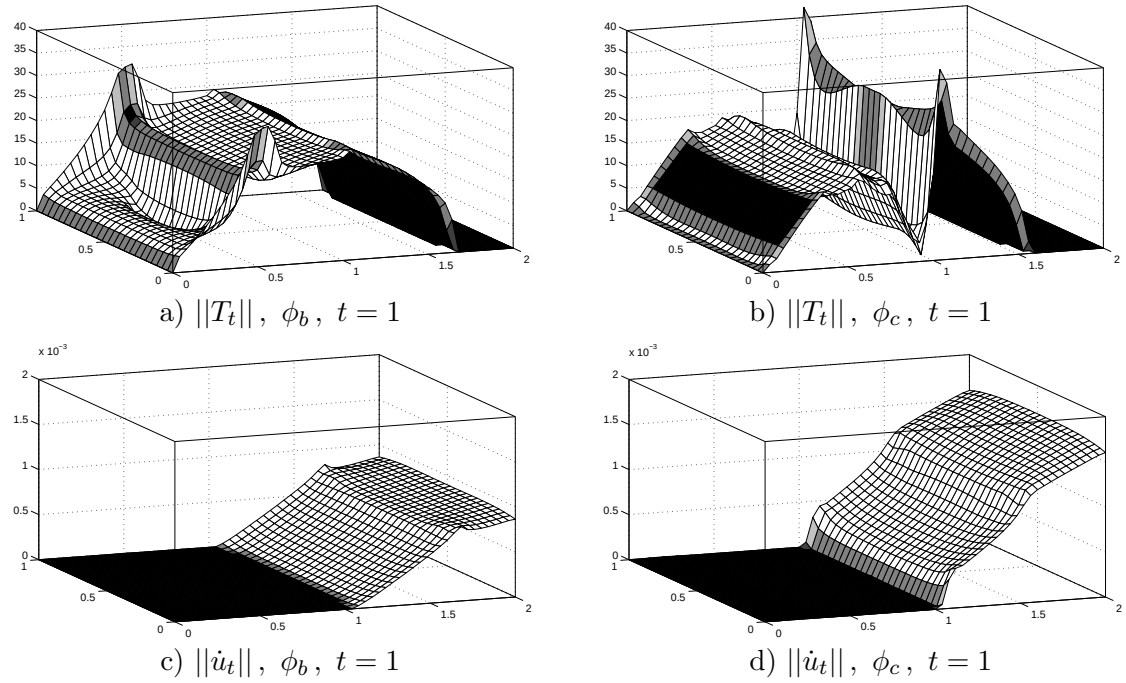


Obrázek 2: Grafy funkcí ϕ_a, ϕ_b a ϕ_c

Pružné těleso je tvořeno kvádrem $\Omega = (0,2) \times (0,1) \times (0,1)[m]$, materiálové vlastnosti jsou charakterizované Youngovým modulem $E = 211.9e9[Pa]$ a Poissonovou konstantou $\sigma = 0.277$. Rozložení hranice $\partial\Omega$ na jednotlivé části je patrno z obr. 1. Na části Γ_u jsou předepsána nulová posunutí, na Γ_c podmínky nepronikání a tření. Část hranice $\Gamma_P = \Gamma_P^1 \cup \Gamma_P^2$ je zatížena silami o hustotách:

$$\left.\begin{array}{ll} P(t) = \phi_a(t)(\ \ 0\,,\ 0\,,\ 10)1e7 & \text{na } \Gamma_P^1 \\ P(t) = \phi_x(t)(\ \ 3\,,\ 0\,,\ 5)1e7 & \text{na } \Gamma_P^2, x \in \{b,c\}. \end{array}\right\} \quad (1)$$

Grafy funkcí ϕ_a, ϕ_b a ϕ_c jsou na obrázku 2. Sledujeme dvě historie zatěžování, kde namísto ϕ_x použijeme v prvním případě ϕ_b a ve druhém ϕ_c. V obou případech je zatížení v koncovém čase stejné. Výsledné rozložení velikosti tečného napětí (tření) v čase $t = 1$ pro obě historie zatěžování je zobrazeno na obr. 3.a,b). Velikost rychlosti tečného posunutí na kontaktu je zobrazena na obr. 3.c,d).



a) $||T_t||\,,\ \phi_b\,,\ t = 1$ b) $||T_t||\,,\ \phi_c\,,\ t = 1$

c) $||\dot{u}_t||\,,\ \phi_b\,,\ t = 1$ d) $||\dot{u}_t||\,,\ \phi_c\,,\ t = 1$

Obrázek 3: Průběhy $||T_t||$ a $||\dot{u}_t||$ na Γ_c

Poděkování: Autoři byli podporováni grantem č. 201/07/0294 GAČR a MSM 6198910027.

Literatura

[1] Z. Dostál, D. Horák, R. Kučera: *Total FETI - an easier implementable variant of the FETI method for numerical solution of elliptic PDE.* In: Comm. in Num. Meth in Eng., 22(2006), 12, pp. 1155-1162.

[2] J. Haslinger: *Approximation of the Signorini problem with friction obeying Coulomb's law.* In: Math. Meth. in Appl. Sci., 5(1983), pp. 422-437.

[3] R. Kučera: *Convergence rate of an optimization algorithm for minimizing quadratic functions with separable convex constraints.* In: SIAM J. Optim. 19(2008), 2, pp. 846-862.

[4] R. Rocca, M. Coccu: *Numerical analysis of quasistatic unilateral contact problems with local friction.* In: SIAM J. Numer. Anal. 39(2001), pp. 1324-1342.

On the Golub-Kahan Iterative Bidiagonalization and Revealing the Size of the Noise in a Data

I. Hnětynková, M. Plešinger, Z. Strakoš

Institute of Computer Science AS CR, Prague

Regularization techniques based on the Golub-Kahan iterative bidiagonalization belong among popular approaches for solving large ill-posed problems. First, the original problem is projected onto a lower dimensional subspace using the bidiagonalization algorithm, which by itself represents a form of regularization by projection. The projected problem however inherits a part of the ill-posedness of the original problem, and therefore some form of inner regularization must be applied. Stopping criteria for the whole process are then based on the regularization of the projected (small) problem.

In this lecture we consider an ill-posed problem with a noisy right-hand side (observation vector), where the size of the noise is unknown. We show how the information from the Golub-Kahan iterative bidiagonalization can be used for revealing the unknown level of the noise. Such information can be useful in construction of an efficient stopping criteria in solving large ill-posed problems.

Acknowledgement: This work was supported by the GAAS grant IAA100300802, and by the Institutional Research Plan AV0Z10300504.

References

[1] G. H. Golub, W. Kahan, *Calculating the singular values and pseudo-inverse of a matrix*, SIAM J. Numer. Anal., Ser. B 2 (1965), pp. 205–224.

[2] P. C. Hansen, *Regularization Tools – version 3.2 for MALTAB 6.0, a package for analysis and solution of discrete ill-posed problems*, (`http://www2.imm.dtu.dk/~pch/Regutools/index.html`).

[3] P. C. Hansen, *Rank-Deficient and Discrete Ill-Posed Problems, Numerical Aspects of Linear Inversion*. Philadelphia, SIAM Publications, 1998.

[4] I. Hnětynková, M. Plešinger, Z. Strakoš, *Golub-Kahan Iterative Bidiagonalization and Revealing the Size of the Noise in a Data*, To appear to BIT.

[5] G. Meurant, Z. Strakoš, *The Lanczos and conjugate gradient algorithms in finite precision arithmetic*, Acta Numerica, 15 (2006), pp. 471–542.

[6] D. P. O'Leary, J. A. Simmons, *A bidiagonalization–regularization procedure for large scale discretizations of ill-posed problems*, SIAM J. Sci. Stat. Comp., 2 (1981), pp. 474–489.

[7] M. Plešinger, *The total least squares problem and reduction of data in* $AX \approx B$. Ph.D. thesis, Faculty of Mechatronics, Technical University of Liberec, 2008.

Řešení úlohy proudění v rozsáhlé diskrétní síti puklin v kontextu sdružených úloh proudění-mechanika

M. Hokr, J. Kopal, J. Havlíček

Technická univerzita v Liberci

1 Úvod

Numerické výpočty mají nezastupitelné místo v geovědních aplikacích jako je modelování proudění vody, vedení tepla a napjatosti v horninovém prostředí. Jedním z významných problémů je analýza funkčnosti a bezpečnosti konceptu hlubinného ukládání vyhořelého jaderného paliva, kde typicky jde o řešení sdružených úloh z výše jmenovaných fyzikálních procesů (T-H-M). Specifika, která vedou na složité úlohy matematické formulace a numerického řešení, jsou například

- velký geometrický rozsah úloh zároveň s požadavky na přesnost v lokálním měřítku,
- dlouhý časový interval pro studované procesy,
- složité chování geomateriálů při větším zatížení (nelinearity), vliv nehomogenity a mikrostruktury

Kompaktní horninový masiv, který je uvažován jako možné prostředí pro hlubinné úložiště, je ve skutečnosti geometricky velmi složité prostředí, kde dominantní význam na mechanické a hydraulické vlastnosti mají existující pukliny - v modelech je prostředí reprezentováno jako ekvivalentní kontinuum nebo diskrétní síť puklin. V článku se zabýváme otázkou stanovení makroskopických hydro-mechanických vlastností na základě řešení modelu v detailní škále – je analyzován vliv struktury a materiálových parametrů jednotlivých puklin jak na globální hydraulické vlastnosti, tak na kvantitativní vyjádření nehomogenity toku a vlivu mechanického zatížení na tyto vlastnosti.

2 Definice úlohy

Geometrie úlohy je zadána jako diskrétní puklinová síť ve 2D, tj. systém vzájemně se protínajících úseček v rovině xy, vyplňující čtverec s rozměry 20x20m se středem v počátku souřadné soustavy [1, 3]. Uvažujeme ustálené proudění, které se na jednotlivých puklinách řídí rovnicemi

$$\begin{aligned} \mathbf{u} &= -K\nabla p \\ \nabla \cdot \mathbf{u} &= q, \end{aligned} \tag{1}$$

kde $\mathbf{u}(x,t)$ je neznámá rychlost, $p(x,t)$ neznámá tlaková výška a K je hydraulická vodivost a q zdroje/propady. V místě průsečíků puklin je předpokládána spojitost tlaků a bilance toku.

Pro úlohu proudění je předepsána Dirichletova okrajová podmínka na celé hranici, odpovídající konstantnímu gradientu ve směru x resp. y (Obr. 1), rozdíl tlakové výšky mezi protilehlými hranami je $p_2 - p_1 = 20\text{m}$. Pro úlohy mechaniky je předepsán normálový tlak na celé hranici, konstantní podél každé hrany, rozdílný ve směrech x a y.

Puklinová síť byla vygenerována podle stochastického modelu [1], který proti běžným modelům reflektuje korelaci mezi délkou pukliny a jejím rozevřením (šířkou), která koresponduje s reálným pozorováním hornin. Celkový počet puklin je 7786. Každé puklině přísluší jiná šířka (rozevření) b, která určuje podle standardního vztahu $K = \frac{\varrho g}{12\mu} b^2$ její hydraulickou vodivost. Mechanické konstitutivní vztahy jsou schematizací reálného pozorování, ve formě nelineárního vztahu mezi rozevřením a normálovým a smykovým napětím jsou uvedeny v [3] a hodnota rozevření tak zároveň udává vazbu mezi úlohou napjatosti a úlohou proudění.

V tomto článku se zabýváme v první fázi řešením úlohy proudění a stručně je naznačeno zahrnutí vazby na úlohu mechaniky. Protože mezi jednotlivými průsečíky puklin nedochází k dalším jevům ovlivňujícím proudění, jde v jednorozměrném případě úseku pukliny o jednoduchý lineární vztah, který nevyžaduje další diskretizaci – pro referenční případ je numerická diskretizace ve smyslu metody konečných prvků tedy dána pouze vzájemnou polohou puklin a jednotlivé liniové elementy jsou úsečky mezi průsečíky.

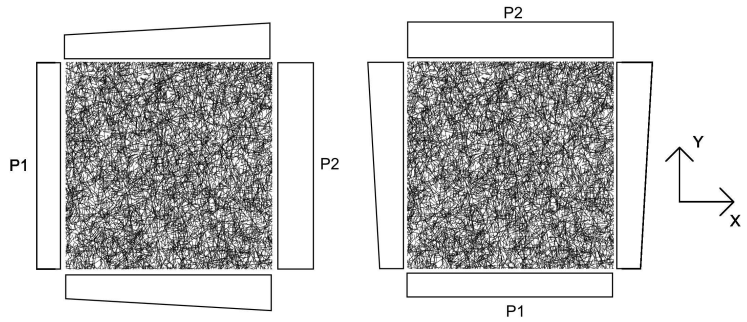
3 Numerické řešení a výsledky

Úloha je numericky řešena smíšenou hybridní metodou konečných prvků s lineárními vektorovými bázovými funkcemi pro rychlost a po částech konstantními funkcemi pro tlaky. Pro úlohu s kombinací 1D, 2D a 3D elementů je metoda formulována v [4] a implementována v kódu Flow123D vyvinutém na Technické univerzitě v Liberci [5]. Kód používá externí řešič soustavy lineárních rovnic, která pro uvedenou formulaci má indefinitní symetrickou matici. Pro prezentované úlohy je použit řešič ISOL vyvinutý P. Jiránkem (TU Liberec), založený na metodě GMRES. Pro postprocesing je použit program GMSH.

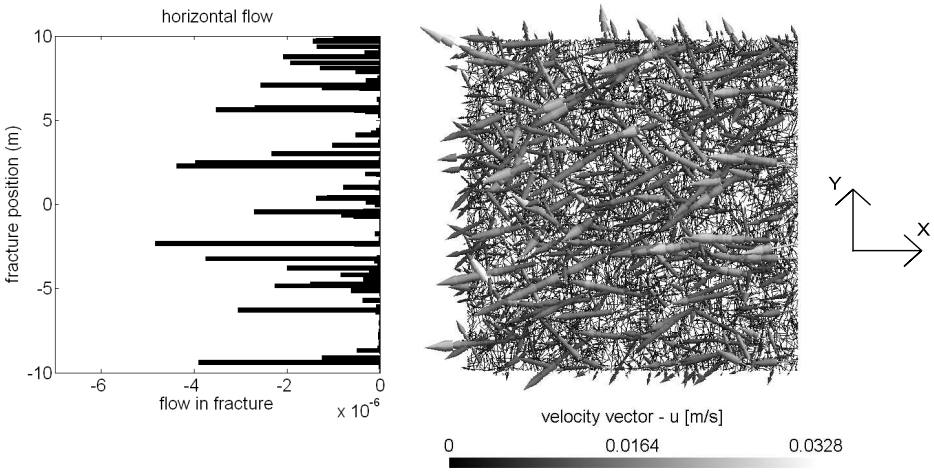
Vlivem složité struktury protínaní puklin jde o rozsáhlou úlohu s 74826 elementy s velkým poměrem délky nejdelšího a nejkratšího (Tab.1). Vodivost elementu korespondující s hodnotami prvků ve výsledné soustavě lineárních rovnic je dána $\frac{Kb}{\Delta x} \sim \frac{b^3}{\Delta x}$ (kde Δx je délka elementu), podmíněnost úlohy je pak ovlivněna kombinací geometrických a materiálových koeficientů a s přihlédnutím k pozitivní korelaci mezi délkou pukliny (jako celku) a rozevřením b jsou tedy nejméně příznivé případy krátkých elementů na velkých (dlouhých i širokých) puklinách a dlouhých elementů na malých (krátkých a tenkých) puklinách. Hodnoty pro konkrétní případ puklinové sítě ilustruje tabulka 1. Velikost soustavy rovnic ze smíšené hybridní metody je 273570, se 773994 nenulovými prvky.

Pro řešení úlohy mechaniky je uvažována standardní úloha pružnosti na spojité oblasti s puklinami vyjádřenými jako pásy konečné šířky a příslušnými nelineárními konstitutivními vztahy. Z výsledného pole posunutí lze určit změnu rozevření pukliny (která vstupuje jako materiálový parametr do úlohy proudění) jako průmět rozdílu posunutí protilehlých bodů na puklině do normály, $\Delta b = (\vec{u_1} - \vec{u_2}) \cdot \frac{\vec{b}}{||\vec{b}||}$, kde $\vec{b}$ je vektor spojující protilehlé body pukliny.

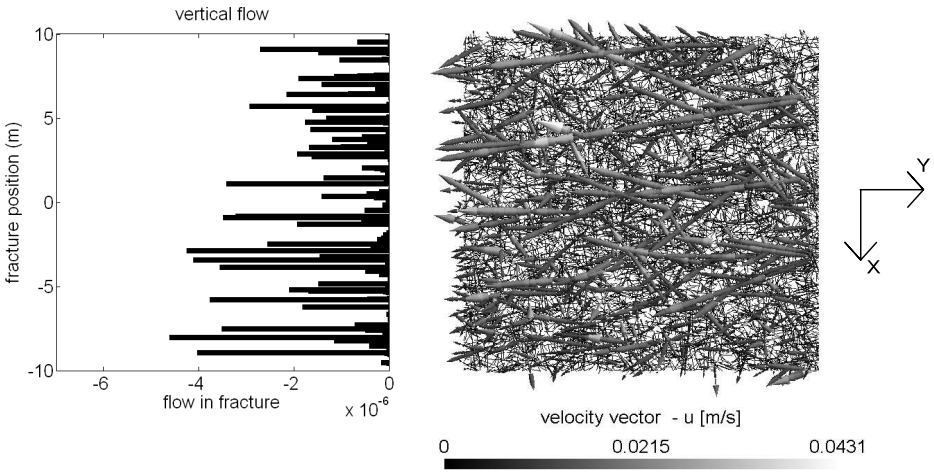
Výsledky potvrzují předpokládanou výraznou nehomogenitu toku oblastí (Obr. 2 a 3). Dlouhé dobře vodivé pukliny vytvářejí tzv. preferenční cesty a ve velké části drobných puklin jsou rychlosti o několik řádů nižší (nejsou vidět ve vizualizaci). Z důvodu výše zmíněné špatné podmíněnosti byl výpočet testován i s rovnoměrně přeškálovanými parametry hydraulické vodivosti K a rozevření b a byla potvrzena úměrnost hodnot rychlosti. Další testy byly zaměřeny na vliv úprav puklinové sítě zmírňující nepříznivé geometrické vlastnosti a velikost úlohy – sloučení velmi blízkých bodů a vynechání slepých úseků puklin.



Obrázek 1: Typy zadání Dirichletovy okrajové podmínky (předepsaná tlaková výška).



Obrázek 2: Výsledky úlohy s horizontálním tokem - hodnoty rychlosti v puklinách a celkový tok koncovými body na odtokové hranici [m^3/s].



Obrázek 3: Výsledky úlohy s vertikálním tokem - hodnoty rychlosti v puklinách a celkový tok koncovými body na odtokové hranici [m^3/s]

Table 1: Geometrické a materiálové parametry elementů s vlivem na podmíněnost úlohy.

	K	b	Δx	$\frac{Kb}{\Delta x}$ skutečné	$\frac{Kb}{\Delta x}$ nejnepříznivější
Maximum	2.4e-2	2e-4	8.4e-1	2.6e-1	$\frac{\max Kd}{\min \Delta x}$=2.4
Minimum	1.1e-5	4.1e-6	2e-6	1.1e-10	$\frac{\min Kd}{\max \Delta x}$=5.4e-11
Max./Min.	2.2e+3	4.7e+1	4.1e+5	2.4e+9	4.5e+10

4 Závěr

Prezentované výpočty jsou úvodním krokem řešení zadání rozsáhlejšího projektu sdružených úloh proudění-mechanika na puklinových sítích a potvrdily schopnost navržených numerických metody a vytvořeného kódu řešit úlohu takového rozsahu, s výraznými nehomogenitami materiálových koeficientů způsobujících špatnou podmíněnost.

V další fázi řešení bude provedeno spojení výpočtu napjatosti a proudění: Předpokládaný efekt mechanického zatížení v rámci sdružené úlohy je, že v závislosti na poměru tlaků v různých směrech, dojde buď k rovnoměrnému snížení toku nebo zvýšení v puklinách určité (nepříznivé) orientace a tedy k dalšímu zesílení nehomogenity. Použitý koncept řešení úlohy napjatosti z důvodu plné 2D diskretizace neumožní řešit úlohu s tak velkým počtem puklin a bude využit koncept částečné homogenizace (náhrady ekvivalentním kontinuem), tj. řešení úlohy na kombinaci 2D kontinua a 1D diskrétní sítě. Tento koncept který se osvědčil u úloh proudění jako kompromis mezi přesností a výpočetní a datovou náročností, bude zobecněn pro úlohu napjatosti, kde v daném kontextu není běžně používán.

Další navazující oblastí je vývoj pokročilých metod řešení výsledné soustavy lineárních rovnic. V práci [2] byla odvozena konstrukce paralelizovatelného předpodmínění technikou rozšířených Lagrangiánů a Schwarzovou metodou, pro indefinitní matici jako celek. Druhým možným postupem je využití blokové struktury matice a pomocí Schurova doplňku převést soustavu na pozitivně definitní, řešenou sdruženými gradienty, rovněž s možností paralelizace výpočtu.

Poděkování: Tento článek vznikl za podpory MŠMT, projekt č.1M0554, a Správy úložišť radioaktivních odpadů, projekt č. 2008/031/Slo, v rámci účasti v mezinárodním projektu Decovalex.

Literatura

[1] A. Baghbanan, L. Jing: *Hydraulic properties of fractured rock masses with correlated fracture length and aperture*. Int. J. Rock Mech. Min. Sci. 44 (2007), pp. 704–719.

[2] R. Blaheta, P. Byczanski, R. Kohut, J. Starý: Modelling THM Processes in Rocks with the Aid of Parallel Computing, In: Thermo-Hydromechanical and Chemical Coupling in Geomaterials and Applications (Proceedings of the 3rd Int. Symp. GeoProc 2008), Jian-Fu Shao, Nicolas Burlion eds., ISTE and J. Wiley, London 2008, pp.373-380.

[3] Hudson, Jing, Neretnieks: Technical Definition of the 2-D BMT Problem for Task C, DECOVALEX-2011 project, 5 May 2008

[4] J. Šembera, J. Maryška, J. Královcová, O. Severýn: *A novel approach to modelling of flow in fractured porous medium*, Kybernetika (2007), Vol. 43/4, str. 577-588. ISSN 0023-5954

[5] O. Severýn, M. Hokr, J. Královcová, J. Kopal, M. Tauchman: Flow123D: Numerical simulation software for flow and solute transport problems in combination of fracture network and continuum, Technical report, TU Liberec, 2008.

Our Blue Gene Experience

O. Jakl, J. Starý

Institute of Geonics AS CR, Ostrava

1 The Blue Gene project

The original *Blue Gene* project was a computer architecture project aiming at moving the frontiers in supercomputing, to produce supercomputers with operating speeds in the petaFLOPS range, see e.g. [3], [4]. It was announced for the years 1999 – 2004, operating with a budget of $ 100 million. The main participators in the project were IBM, the Lawrence Livermore National Laboratory (LLNL), the United States Department of Energy and academia. Recall that 2002 – 2004 were the years of the Japanese Earth Simulator domination in the TOP500 supercomputer list.

The project originally focused to advance the understanding of important biological processes such as protein folding, later the design became more general purpose. Nevertheless it retains the goal of an extreme scalability appropriate for molecular modelling, with the first system in the series called *Blue Gene/L* (BG/L) targeting at least 65 536 nodes at full scale. Another unique aspects include:

- trading the speed of the (PowerPC) processors for lower power consumption;
- system-on-a-chip design;
- 3D torus interconnect with auxiliary networks for global communications, I/O, and management;
- lightweight operating system per node for minimum system overhead.

A BG/L prototype reached a speed of 70.72 TFLOPS by November 2004, taking first place in the TOP500 list. Since then, thanks to continuous development, a machine of this type, installed at LLNL, has been occupying the premier positions, nowadays being No. 4 with 478 TFLOPS Rmax Linpack performance, delivered by its 212 992 processing elements.

IBM now offers the *Blue Gene/P* (BG/P) solution as a more commercial follow-on product to the successful BG/L generation. It provides ultrascale performance within a standard programming environment and high efficiency in power, cooling and floor-space consumption. BG/P extends the performance through a density and frequency jump, doubling the performance of the processors and interconnects.

2 Blue Gene installation in Bulgaria

It was exactly the Blue Gene/P system deployed in Sofia by mid 2008 through which Bulgaria became one of the first countries from the former socialist block (after Russia and Poland) that made an investment into supercomputing facilities of the highest ranking, see [5]. The following goals are given as motivation: (1) to solve high-end computing-intensive projects in life sciences,

new drugs discovery, financial modelling and education, (2) to allow Bulgarian businesses and research institutes to join European partners in research and other projects and (3) to become the regional supercomputing centre for South East Europe. The project is run by a consortium of Bulgarian state and academic institutions (including Bulgarian Academy of Sciences) and IBM.

The machine consists of two BG/P racks with 2048 four-processor chips, in total 8192 PowerPC 450 processing elements running on 850 MHz and accelerated by a double-precision, dual pipe floating-point unit. The chips reside on compute cards with 2 GB of shared memory, which are connected through several types of interconnects: 3D torus for two-point communications, collective network for collective communications and global interrupt network for fast barriers. With Linpack performance Rmax = 23.42 TFLOPS the machine would have shared the 74th place in the TOP500 list at the installation time (July 2008); in the latest 32nd list it is ranked the 126th.

3 Initial experiments

Thanks to our colleagues in BAS we have got a precious short-time opportunity to make some hands-on experience with the Bulgarian Blue Gene/P machine (let us call it BG) at the end of 2008. In this period, BG was entering the production mode, but rather frequent drop-outs of the backend, some missing utilities, lack of tailored documentation, etc. revealed that the new system (and its administration) has not been tuned yet.[1]

As far as the working environment on BG is concerned, the user has to make himself familiar with several points. In general, although he will find common tools for code development and execution on BG, most of them have been adapted and those specifics have to be learned and considered to get good performance. The user will access BG through a front-end, which is an IBM System p 64-bit server, but not binary compatible with BG. Without having direct access to the nodes of BG, one has to create jobs for batch processing, which is controlled by the IBM LoadLeveler job scheduler. When preparing the (parallel) application codes on the front-end, cross-compilers have to be used.

Compilers and sequential performance. BG supports two alternatives for compilations, GNU Compilers and IBM XL Compilers. In the benchmarks based on our (sequential) solvers, the latter showed slightly (25 %) better performance on the front-end, so we made use of them, namely of the XLF 11.1 Fortran compiler, in the following tests. In this phase, we also studied the influence of the various optimization options offered by the compiler on the performance of our codes. To our surprise, the best timings of our sequential solver on a (single) processor of BG were by more than 40 % longer than on our seven years old Thea cluster with AMD Athlon 1400 MHz processors and Fast Ethernet interconnect, and almost five times longer than on the front-end. This observation confirms that the Blue Gene architecture can be advantageous mainly for highly scalable parallel applications capable to utilize a great number of processors.

Processors versus cores. BG provided us with an interesting opportunity to carry out the 4×1 vs. 1×4 processor $\times$ core performance comparison. Recall that the compute nodes on BG/P have four processing elements (cores) with 2 GB of local RAM on a single chip. The user can specify the run mode, e.g. the *virtual node* (VN) mode, when the compute node can

[1] But the conditions were evidently improving.

host four parallel processes, each assigned to one core and 1/4 of the RAM (0.5 GB), or the *SMP* mode, when not more than one process is assigned to the compute node, using the entire RAM available.[2] We chose a benchmark problem sufficiently small to match the restricted memory and a (displacement decomposition) solver giving rise to four parallel processes, and observed that there is almost no difference between the VN and SMP modes, i.e. that at least in our application, the interchip communication is comparable in performance with the intrachip communication. A good message since in the VN mode the amount of RAM per process is too restrictive.

Parallel scalability of the DD solver. Most interestingly, BG allowed us also to test the parallel scalability of our domain decomposition (DD) finite element solver on a greater number of processors. We prepared a benchmark problem (called FOOT240z) with a large number of nodes especially in the Z direction ($61 \times 61 \times 241$ nodes, i.e. 2 690 283 DOF), along which the problem is decomposed in the solver, that allowed us to employ up to 64 processors before the domains became too "thin". The selected results of a fairly large number of runs are summarized in Fig. 1.

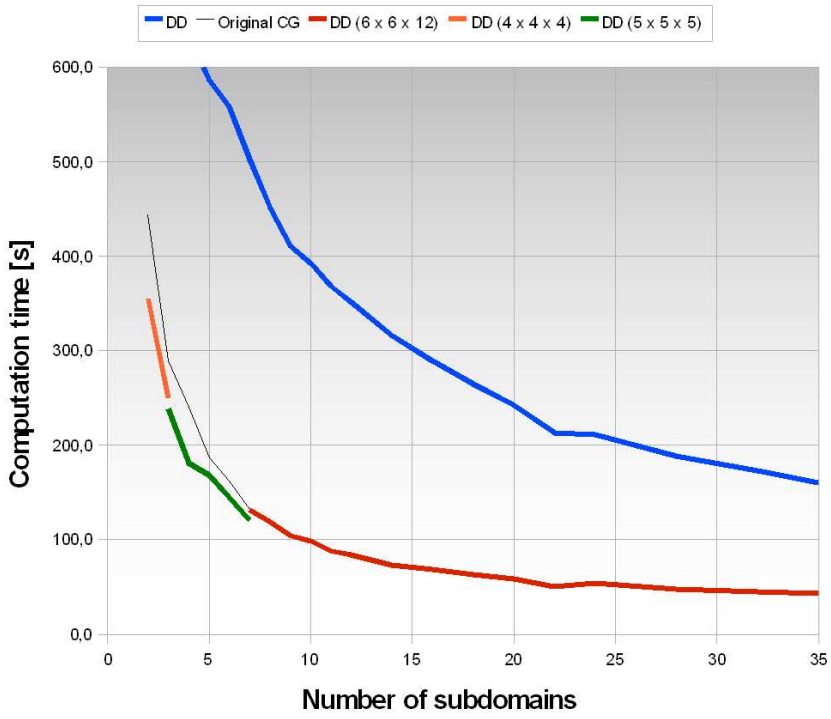


Figure 1: DD solver on BG: Computing times with increasing number of processors.

When no coarse grid in used (the upper curve), we can observe almost uninterrupted decrease of the computing time, up to 112 s, which is more than 11 times shorter than the 1252 s of the sequential solver (the relative efficiency is about 0.17). With a coarse grid, the absolute times are much favourable, however the curves are "dentate". The reason can be found in the fact that each coarse grid is appropriate only for a limited range of decompositions, when its process matches in computing time the processes of the domains. Outside this range, another coarse grid should be applied. In the lower curve in Fig. 1, we tried to compose a "ideal" scalable

[2]The process can start up to four parallel threads.

coarse grid computation making use of three different coarse grids. With such a construction, the computing time on 64 processors is 30 s and the relative efficiency does not drop below 0.48.

4 Microstructure computations

We completed our BG experience with a number of tests related to the microstructure 3D finite element modelling, the current topic of our research, see e.g. [2]. The problem under consideration deals with the analysis of mechanical behaviour of geocomposites that arose from the injection of polyurethane resin into the coal environment. Such technology can be used e.g. to reinforce coal pillars during mining. The questions to be answered based on the modelling results can read as follows: What are the upscaled elastic properties of coal geocomposites due to their complicated microstructure given by porous and disturbed coal? How sensitive are these properties on the quality of filling of the coal matrix by the polyurethane?

The homogenized properties are determined by numerical upscaling. The structure of a geocomposite sample is digitalized by X-ray computer tomograph (CT), then the upscaled properties are obtained via numerical implementation of loading tests. We consider strain and stress driven tests implemented numerically by means of the FE analysis of the microstructure, when the standard conforming linear tetrahedral finite elements are applied.

We considered two kinds of parallel iterative solution methods for the FEM system arising from a CT scan of $231 \times 231 \times 37$ voxels: conjugate gradients with parallel displacement decomposition – MIC(0) factorization preconditioning (DiD–MIC(0) – fixed number of subdomains) and conjugate gradients with two-level Schwarz type preconditioning (DD–ACG - varying number of subdomains), where coarse subproblems created by aggregation are used. The timings of the solution on the Thea cluster and on BG are presented in Table 1.

Solver	# Subd.	# Iter.	**Thea** T [s]	**BG** T [s]
seq–MIC(0)	1	75	544	1473
DiD–MIC(0)	3	75	678	387
DD–ACG	2	47	361	425
	4	43	196	209
	8	41	119	123
	16	41		103

Table 1: Solution times of the microstructure problem on BG and Thea.

There are some strange values among in the results, e.g. that of the sequential run on BG, which we did not manage to explain in the short period of BG's accessibility. But in total, as one can see, BG has not contributed much to the speed of solution of this particular problem, because on 16 BG's processors, the computation was only slightly shorter than on Thea with 8 processors and the discretization provided not enough nodes in the Z direction to employ more processors.

5 Conclusions

The Blue Gene/P platform, including its Bulgarian installation, is without question very interesting and powerful — cf. the TOP500 lists. However, to take full advantage of its potential,

it may not be enough to make technical tuning of the existing codes. Remember that this architecture, in accordance with the aims of the original Blue Gene project, is distinguished by a large number of relatively slow processors with fast interconnects, cf. e.g. [1]. As a consequence, it is most appropriate fine-grained parallel decompositions, which may be unsuitable for some problems. Anyway, porting an application to BG (and, in general, to a parallel system with thousands of processors) is usually a great challenge: One has to reconsider the potential of the problem to be parallelized and make it match with the strong sides of the target supercomputer.[3] Of course, this fully holds for our solvers.

Acknowledgement: We acknowledge the support of the Grand Agency of the Czech Reublic (GAČR) under the grant No. 105/09/1830.

References

[1] S. Alam et al.: *Early Evaluation of IBM BlueGene/P.* In: SC08: International Conference High Performance Computing, Networking, Storage, and Analysis. Austin, ACM/IEEE, 2008.

[2] R. Blaheta, P. Byczanski, P. Harasim: *Multiscale modelling of geomaterials and iterative solvers.* In this proceedings.

[3] Blue Gene – from Wikipedia, the free encyclopedia. http://en.wikipedia.org/wiki/Blue_Gene, 8/12/2008.

[4] Blue Gene. http://domino.research.ibm.com/comm/research_projects.nsf/pages/bluegene.index.html, 2/12/2008.

[5] BGSC – Bulgarian Supercomputing Centre. http://www.scc.acad.bg/, 13/12/2008.

[3] BTW: The "ability to demonstrate application scalability at least up to 512 processor cores" is one of the requirements posted by BG administration to allow execution on BG.

On a stable variant of Simpler GMRES and GCR

P. Jiránek, M. Rozložník

CERFACS, Toulouse

Technical University of Liberec
Institute of Computer Science AS CR, Prague

1 Introduction

Systems of linear algebraic equations

$$Ax = b, \qquad A \in \mathbb{R}^{N\times N},\ b \in \mathbb{R}^N, \tag{1}$$

where A is a large and sparse nonsingular matrix, arise in a large variety of scientific problems. Almost all modern iterative solvers for treating large-scale sparse systems belong to the wide class of Krylov subspace methods. Starting from an arbitrary initial guess x_0, such iterative methods seek at the nth iteration the approximate solution x_n in the affine subspace $x_0 + \mathcal{K}_n(A, r_0)$, where $r_0 := b - Ax_0$ is the initial residual vector corresponding to x_0 and $\mathcal{K}_n(A, r_0) := \mathrm{span}\{r_0, Ar_0, \ldots, A^{n-1}r_0\}$ stands for the nth Krylov subspace generated by A from r_0. Minimum residual methods like GMRES [11] or GCR [5], which are usual methods of choice for solving general nonsymmetric problems, construct the approximate solution $x_n \in x_0 + \mathcal{K}_n(A, r_0)$ such that its corresponding residual vector $r_n := b - Ax_n$ has a minimal Euclidean norm:

$$\|r_n\| = \|b - A(x_0 + d_n)\| = \min_{d\in\mathcal{K}_n(A,r_0)} \|b - A(x_0 + d)\|. \tag{2}$$

The minimum norm property (2) is equivalent to the orthogonality of the r_n to the residual subspace $A\mathcal{K}_n(A, r_0)$:

$$\langle r_n, w\rangle = 0 \qquad \forall w \in A\mathcal{K}_n(A, r_0). \tag{3}$$

Here and henceforth, $\langle\cdot,\cdot\rangle$ stands for the standard Euclidean inner product and $\|\cdot\|$ denotes the Euclidean vector norm as well as the induced spectral matrix norm.

The classical implementation of GMRES [11] is based on the Arnoldi process [1] providing an orthonormal basis Q_n of the Krylov subspace $\mathcal{K}_n(A, r_0)$. The norm of the residual r_n is minimized in (2) by solving an $(n + 1) \times n$ upper Hessenberg least squares problem. The implementations based on the modified Gram-Schmidt and the Householder QR were shown to be backward stable [4] and [10]. Another implementation of GMRES called Simpler GMRES and proposed by Walker and Zhou [13] generates an orthonormal basis V_n of $A\mathcal{K}_n(A, r_0)$ and carries out the relation (3) simply by projecting the initial residual r_0 onto the orthogonal complement of the column-span of V_n. In particular, let $Z_n := [z_1, \ldots, z_n]$ be a basis of $\mathcal{K}_n(A, r_0)$ such that $\mathrm{span}\{z_1, \ldots, z_k\} = \mathcal{K}_k(A, r_0)$ for $k = 1, \ldots, n$ and assume that all z_k are normalized. The orthonormal basis of $A\mathcal{K}_n(A, r_0)$ can be formed by computing the QR factorization

$$AZ_n = V_nU_n, \tag{4}$$

where $V_n := [v_1, \ldots, v_n]$ has orthonormal columns and U_n is an $n \times n$ nonsingular and upper triangular matrix. The residual vector is then computed as $r_n = (I - V_nV_n^T)r_0 = r_{n-1} - \alpha_n v_n$,

$\alpha_n = \langle r_{n-1}, v_n \rangle$ and the corresponding minimum residual norm approximation $x_n = x_0 + Z_n t_n$ is found by solving the upper triangular system $U_n t_n = V_n^T r_0 = [\alpha_1, \ldots, \alpha_n]^T$.

In the original Simpler GMRES implementation [13] the basis Z_n of $\mathcal{K}_n(A, r_0)$ consists of the normalized initial residual r_0 and the first $n-1$ columns of V_n. As it was shown in [9] and partially also in [13], the condition number of $Z_n = [r_0/\|r_0\|, V_{n-1}]$ (equal to the ratio of its extremal singular values) can be bounded as

$$\frac{\|r_0\|}{\|r_{n-1}\|} \leq \kappa([r_0/\|r_0\|, V_{n-1}]) \leq 2\frac{\|r_0\|}{\|r_{n-1}\|}.$$

The fast convergence in the residual norm leads hence to a poor conditioning of the Krylov subspace basis possibly resulting to the numerical instability. On the other hand, the basis conditioning is not dramatically deteriorated by poor convergence of the residuals. When the Simpler GMRES basis $[r_0/\|r_0\|, V_{n-1}]$ is replaced by scaled residual vectors $[\frac{r_0}{\|r_0\|}, \ldots, \frac{r_{n-1}}{\|r_{n-1}\|}]$ as in RB-SGMRES proposed in [8] (and similarly to GCR [5]), we can observe essentially the opposite behavior and give the following bounds

$$\max_{k=1,\ldots,n} \left(\frac{\|r_{k-1}\|^2 + \|r_k\|^2}{\|r_{k-1}\|^2 - \|r_k\|^2} \right)^{\frac{1}{2}} \leq \kappa \left(\left[\frac{r_0}{\|r_0\|}, \ldots, \frac{r_{n-1}}{\|r_{n-1}\|} \right] \right) \leq n^{\frac{1}{2}} \left(1 + \sum_{k=1}^{n-1} \frac{\|r_{k-1}\|^2 + \|r_k\|^2}{\|r_{k-1}\|^2 - \|r_k\|^2} \right)^{\frac{1}{2}}.$$

Hence the fast convergence implies well-conditioning of the Krylov subspace basis and vice versa. It is not surprising since the residuals obtained by GMRES in this case are close to the orthogonal residuals computed using FOM [11, 3]. For the similar result see also [12]. In the following section we try to combine the good properties of both bases by proposing an adaptive version of Simpler GMRES.

2 Adaptive Simpler GMRES

In this section we propose an adaptive variant of Simpler GMRES computing the Krylov subspace basis Z_n with condition number kept at a reasonably small level. This is achieved by an adaptive switching between the bases of Simpler GMRES and RB-SGMRES (GCR) using an intermediate residual norm decrease criterion: if the residual norm at given step is sufficiently reduced the Krylov subspace basis is extended with the normalized residual vector as in RB-SGMRES or GCR; otherwise we use the last available vector of the orthonormal basis as in Simpler GMRES. We introduce a threshold parameter $\nu \in [0, 1]$ and at the nth step ($n > 1$) we use either $z_n = r_{n-1}/\|r_{n-1}\|$ provided that $\|r_{n-1}\| \leq \nu \|r_{n-2}\|$ or $z_n = v_{n-1}$ in the latter case. The algorithm can be formulated as follows:

Adaptive Simpler GMRES:

1: *Initialization:* Choose an initial guess x_0 and the threshold parameter $\nu \in [0, 1]$, compute the initial residual $r_0 = b - Ax_0$ and $\rho_0 = \|r_0\|$.

2: *Compute the bases Z_m and V_m:*
for $n = 1, \ldots, m$ (until convergence) do

2.1: Compute the new direction vector z_n:

$$z_n = \begin{cases} r_0/\rho_0 & \text{if } n = 1, \\ r_{n-1}/\rho_{n-1} & \text{if } n > 1 \text{ and } \rho_{n-1} \leq \nu \rho_{n-2}, \\ v_{n-1} & \text{otherwise.} \end{cases}$$

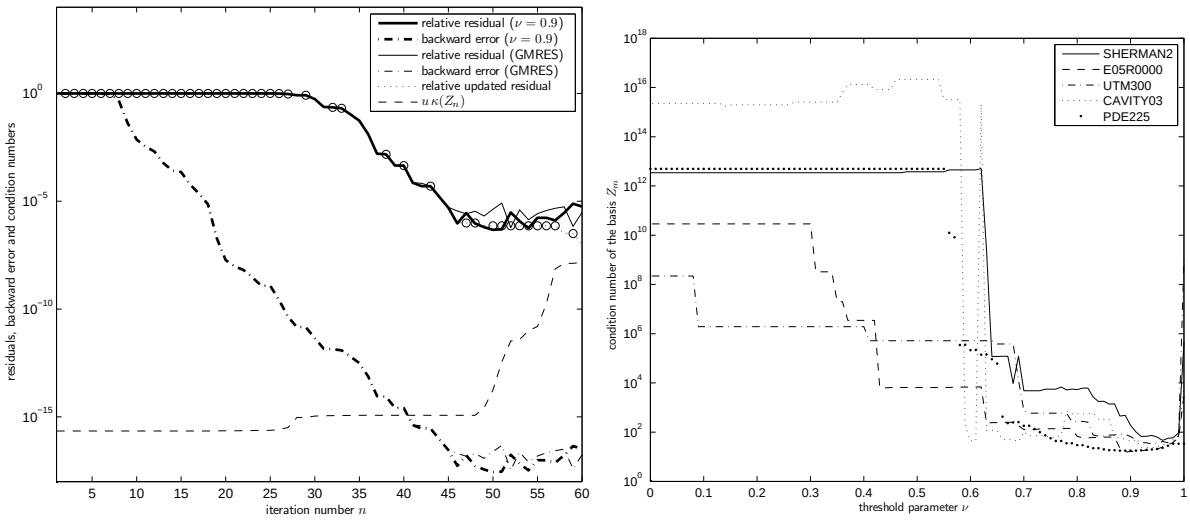


Figure 1: Left plot: Test problem with the matrix `FS1836` and b equal to the left singular vector corresponding to the smallest singular value of A solved by adaptive Simpler GMRES with $\nu = 0.9$. Right plot: The dependence of the condition number of Z_m on the choice of the threshold parameter ν for various problems from Matrix Market.

2.2: Update the QR factorization $AZ_n = V_nU_n$.

2.3: Compute $\alpha_n = \langle r_{n-1}, v_n \rangle$.

2.4: Update $r_n = r_{n-1} - \alpha_n v_n$ and $\rho_n = \|r_n\|$.

3: *Compute the approximate solution:* Solve the upper triangular system $U_m t_m = [\alpha_1, \ldots, \alpha_m]^T$ and compute $x_m = x_0 + Z_m t_m$.

If $\nu = 0$ then $Z_n = [r_0/\rho_0, V_{n-1}]$ we obtain the algorithm identical to Simpler GMRES [13]. The case $\nu = 1$ results in $Z_n = [r_0/\rho_0, \ldots, r_{n-1}/\rho_{n-1}]$ and corresponds to RB-SGMRES, closely related to the GCR method. It is known that in the minimal residual method the residuals can be linearly dependent if the stagnation occurs, in particular when 0 belongs to the field of values of A resulting in the breakdown of RB-SGMRES and GCR. However, setting $\nu < 1$ prevents extending the basis with a linearly dependent residual vector and hence the adaptive Simpler GMRES does not break down until the exact solution has been computed.

3 Conditioning of Z_m and accuracy of the adaptive variant

It was shown in [8] that the condition number of the basis Z_m can significantly affect the accuracy of the computed approximate solution in algorithms based on (4). Such algorithms deliver a backward stable solution provided that Z_m is well-conditioned. In particular, if $cu\kappa(A)\kappa(Z_m) < 1$ the gap between the true residual corresponding to the approximate solution $\hat{x}_m$ and the updated residual vector r_m can be estimated as follows:

$$\frac{\|b - A\hat{x}_m - r_m\|}{\|A\|\|\hat{x}_m\|} \leq cu\kappa(Z_m)\left(1 + \frac{\|x_0\|}{\|\hat{x}_m\|}\right).$$

Here c denotes a moderate generic constant dependent on N and m and u stands for the unit roundoff of the underlying finite precision arithmetic.

For the sake of simplicity, we assume that Z_m computed in adaptive Simpler GMRES is equal to $[r_0/\rho_0, v_1, \ldots, v_{q-1}, r_q/\rho_q, \ldots, r_{m-1}/\rho_{m-1}]$: we have $\|r_n\| > \nu\|r_{n-1}\|$ for $n = 2, \ldots, q-1$ and $\|r_n\| \leq \nu\|r_{n-1}\|$ for $n = q, \ldots, m-1$. Hence in the first stage of the convergence we use the Simpler GMRES basis and the residual basis in the second stage. Such a convergence behavior, i.e., the initial stagnation of the residual norm, appears often in practical computations. Then the conditioning of Z_m can be bounded as

$$\underline{\gamma}_{m,q} \frac{\|r_0\|}{\|r_{q-1}\|} \leq \kappa(Z_m) \leq 2\,\overline{\gamma}_{m,q} \frac{\|r_0\|}{\|r_{q-1}\|}, \tag{5}$$

where

$$\underline{\gamma}_{q,m} := \max_{n=q,\ldots,m-1} \left(\frac{\|r_{n-1}\|^2 + \|r_n\|^2}{\|r_{n-1}\|^2 - \|r_n\|^2} \right)^{\frac{1}{2}}, \quad \overline{\gamma}_{q,m} := (m-q+1)^{\frac{1}{2}} \left(1 + \sum_{n=q}^{m-1} \frac{\|r_{n-1}\|^2 + \|r_n\|^2}{\|r_{n-1}\|^2 - \|r_n\|^2} \right)^{\frac{1}{2}}.$$

In addition we can obtain the stronger bound in terms of the parameter ν

$$1 \leq \kappa(Z_m) \leq \frac{2\sqrt{2}}{\nu^{q-1}} \frac{1+\nu}{1-\nu}. \tag{6}$$

Note that (6) does not (entirely) follow from (5); for more details as well as for the analysis of a more general Z_m, we refer to [7].

The left plot in Figure 1 shows the relative residual norms as well as the normwise backward errors for adaptive Simpler GMRES with the threshold parameter $\nu = 0.9$ and for the modified Gram-Schmidt implementation of the GMRES method for a matrix `FS1836` from the Matrix Market [2] with the right-hand side equal to the left singular vector corresponding to the smallest singular value of A. For this problem, the residual basis is nearly rank deficient in the initial stage of the convergence, which leads to the numerical instability in RB-SGMRES and GCR. On the other hand, an adaptive basis with $\nu = 0.9$ provides a well-conditioned basis. By circles on the residual curve, we denote the iteration steps, where the Simpler GMRES basis is used. In the right plot we show the dependence of $\kappa(Z_m)$ on the value of the parameter $\nu = [0, 1]$ at the iteration step, where the normwise backward error is smaller than 10^{-14} for various problems from the same repository. It is clear, that the values of ν close to 1 should be preferred. This is also apparent from the value ν_{opt} minimizing the right-hand side in (6) for a fixed iteration number m corresponding to the maximum number of iterations or the restart parameter. This leads to

$$\nu_{\text{opt}} = \frac{\sqrt{1+m^2} - 1}{m} \to 1 \quad \text{as} \quad m \to \infty$$

and

$$\kappa(Z_m)|_{\nu=\nu_{\text{opt}}} = O(m).$$

Even though the value ν_{opt} does not necessarily lead to an optimal conditioning of Z_m, it still provides a well-conditioned basis growing at most linearly with m. It was also observed that the poor conditioning of Z_m does not always cause numerical instability. Nevertheless, the adaptive switching providing a well-conditioned Krylov subspace basis should be used in order to develop a robust iterative solver based on Simpler GMRES, which has a guarantee of delivering the accurate approximate solutions to (1) as well as other quantities like the harmonic Ritz values [6], which is however beyond the scope of this contribution.

Acknowledgement: This work has been supported by the grant No. 201/09/P464 of the GACR, by the project IAA100300802 of the GAAS, and by the Institutional Research Plan AV0Z10300504 "Computer Science for the Information Society: Models, Algorithms, Applications".

References

[1] W. E. Arnoldi. The principle of minimized iteration in the solution of the matrix eigenproblem. *Quart. Appl. Math.*, 9:17–29, 1951.

[2] R. F. Boisvert, R. Pozo, K. Remington, R. Barrett, and J. J. Dongarra. The Matrix Market: A web resource for test matrix collections. In R. F. Boisvert, editor, *Quality of Numerical Software, Assessment and Enhancement*, pages 125–137, London, UK, 1997. Chapman & Hall.

[3] J. Cullum and A. Greenbaum. Relations between Galerkin and norm-minimizing iterative methods for solving linear systems. *SIAM J. Matrix Anal. Appl.*, 17(2):223–247, 1996.

[4] J. Drkošová, A. Greenbaum, M. Rozložník, and Z. Strakoš. Numerical stability of GMRES. *BIT*, 35(3):309–330, 1995.

[5] S. C. Eisenstat, H. C. Elman, and M. H. Schultz. Variational iterative methods for nonsymmetric systems of linear equations. *SIAM J. Numer. Anal.*, 20(2):345–357, 1983.

[6] S. Goossens and D. Roose. Ritz and harmonic Ritz values and the convergence of FOM and GMRES. *Numer. Linear Algebra Appl.*, 6:281–293, 1999.

[7] P. Jiránek and M. Rozložník. Adaptive version of Simpler GMRES. Technical Report TP/PA/08/101, CERFACS, 2008. submitted for publication to *Numerical Algorithms*.

[8] P. Jiránek, M. Rozložník, and M. H. Gutknecht. How to make Simpler GMRES and GCR more stable. *SIAM J. Matrix Anal. Appl.*, 30:1483–1499, 2008.

[9] J. Liesen, M. Rozložník, and Z. Strakoš. Least squares residuals and minimal residual methods. *SIAM J. Sci. Comput.*, 23(5):1503–1525, 2002.

[10] C. C. Paige, M. Rozložník, and Z. Strakoš. Modified Gram-Schmidt (MGS), least squares, and backward stability of MGS-GMRES. *SIAM J. Matrix Anal. Appl.*, 28(1):264–284, 2005.

[11] Y. Saad and M. H. Schultz. GMRES: a generalized minimal residual algorithm for nonsymmetric linear systems. *SIAM J. Matrix Anal. Appl.*, 13:856–869, 1986.

[12] J. van den Eshof, G. L. G. Sleijpen, and M. B. van Gijzen. Iterative linear solvers with approximate matrix-vector products. In A. Boriçi, A. Frommer, B. Joó, A. D. Kennedy, and B. Pendleton, editors, *QCD and Numerical Analysis III*, volume 47 of *Lecture Notes in Computational Science and Engineering*, pages 133–141, Heidelberg, 2005. Springer-Verlag.

[13] H. F. Walker and L. Zhou. A simpler GMRES. *Numer. Linear Algebra Appl.*, 1:571–581, 1994.

Graph partitioning

K. Jurková

Technical university of Liberec

1 Introduction

The problem of proper graph partitioning is one of the classical problems of the parallel computing. It is well-known that the process of computing high-quality graph partitionings arising in most practical situations is reasonably understood. This is, e.g., if we consider *standard* criteria for partitionings expressed by balancing sizes of domains and minimizing separator sizes. However, the situation may be different if we need to balance, for example, the time to perform some specific operations, as the time to compute matrix decompositions, incomplete factorizations, or some auxiliary numerical transformations used by linear equations solvers. It can happen, that a partitioning which is well-balanced partitioning with respect to the standard criteria may be completely unbalanced with respect to some time-critical operations on the domains.

The graph partitioning is tightly coupled with the general problems of load balancing. In particular, the partitioning represents a *static* load balancing. In practice, also as mentioned above, work distribution in the computation may be completely different from what was assumed at its beginning. In general context, dynamic load balancing strategies can redistribute the work dynamically. A lot of interest was devoted to analysis and possible cure of such problems [2], [6], [8]. In some situations, in order to allow complicated and unpredictably time-consuming operations on the domains, we can talk about minimizing with respect to the *complex objectives* [7]. A strategy which was proposed in this case is to improve the partitioning iteratively during the course of the computation. Nevertheless, in some cases we know more about these critical operations, and we may be able to include this knowledge into the the graph partitioner, or we may be able to use this information to improve the graph partitioning in one simple step, while having some guarantees on its quality, at the same time. Both these strategies have their own advantages and disadvantages. Probably the most efficient strategy is to integrate additional knowledge on the matrix factorization or other desired operations into the graph partitioner, but this approach may not be very flexible. In addition, its analysis may not be simple. A redistribution in one subsequent step which follows the partitioning provides the useful flexibility, and may not add too much additional computational effort.

The paper introduces a new approach to the graph partitioning problem, where we assume that a full or incomplete matrix factorization will be performed on the domains. Our strategy is based on analyzing the factorizations using graph-theoretic tools. In particular, we will deal with solving the most simple problem of this kind. Namely, we will discuss the complete factorization of a matrix which is symmetric and positive definite. In this case, the underlying model of the factorization is the elimination tree. Based on its properties, we will provide related theoretical and algorithmic results for a post-processing of a given graph partitioning such that the new, redistributed graph is better balanced with respect to the factorization.

Section 2 of the paper summarizes some basic terminology and states the problem we would like to solve. Section 3 is devoted to summarizing our future goals.

2 Basic terminology and our limitations

In order to describe our contribution we are forced to introduce some basic definitions and concepts related to the sparse matrix factorizations. As mentioned above, we will treat the case of complete factorization of symmetric and positive definite matrices.

The decomposition of an SPD matrix A is directed by the *elimination tree*. This tree and its subtrees represent and provide most of the structural information which is relevant to the sparse factorization. For example, based on the elimination tree, we are able quickly determine sizes of matrix factors, their sparsity structure, or other useful counts [1], [3]. In our case a *subtree* of the elimination tree corresponds to a connected subgraph of the original undirected graph. The elimination tree $T(A)$ is the rooted tree with the same node set as $G(A)$ and vertex n as the root. It may be represented by the vector PARENT[.] defined as follows:

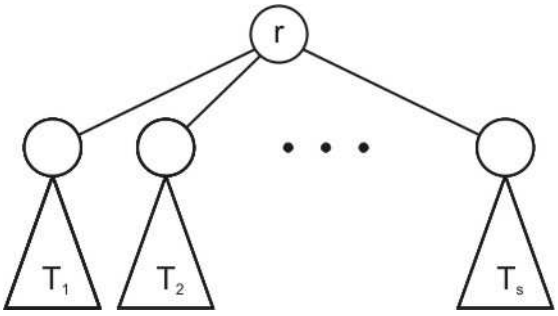
$$PARENT[j] = \begin{cases} \min\{i > j |\ l_{ij} \neq 0\}, & j < n, \\ 0, & j = n, \end{cases}$$

where l_{ij} are entries of L. The n-th column is the only column which does not have any offdiagonal entries.

Many quantities related to the sparse factorization of SPD matrices can be efficiently computed only if the matrix is preordered by some specific reorderings. One of those is a *postordering*. It is induced by a postordering of the elimination tree of the matrix. In particular, a postordering of a tree is its *topological ordering*.

For a given rooted tree, we define a topological ordering of the tree to be an ordering that numbers children nodes before their parent node. A topological ordering of a directed acyclic graph (directed graphs without cycles) is one such that for every directed edge from a node u to v, u is ordered before v. In the case of a rooted tree, if we treat each tree edge as a directed edge that goes from a child to its parent, our definition of a topological ordering of a rooted tree is the same as that used for directed acyclic graphs. Note that the root of a subtree will always be labeled last among nodes in the subtree.

The postorder sequence of a rooted tree T is can be computed recursively as we demonstrate in the following figure, and in the subsequent algorithm.



Algorithm 1 *Postordering(T)*

```
if s=0
   then sequence is r
   else sequence is Postordering(T1), Postordering(T2), .., Postordering(Ts), r
```

Note that any reordering of a sparse matrix that numbers a node ahead of its parent node in its elimination tree is equivalent to the original ordering in terms of fills and computation. In particular, postorderings are equivalent reorderings in this sense.

Let us now summarize additional assumptions which we adopt in this paper. First, we will explicitly assume that the graph was divided just into two domains which are separated by an edge separator. Further, we will deal only with the standard graph model instead of the factorgraph model which would capture matrix blocks. Nevertheless, note that considering matrix blocks seems to be a must in many practical cases, e.g., for a typical matrix arising in finite element computations.

3 Our goals

As mentioned above, we are interested in solving problems of numerical linear algebra. In particular, the operations of our interest are incomplete or full decompositions of large sparse matrices. Such decompositions offer a lot of tools to estimate their sizes even before actual decompositions are performed. Many of them are based on the elimination tree of the related decomposition. We believe that in this case it is not always necessary to use an outer loop to balance the computation as mentioned in [7]. It may be possible to use cheaper tools instead.

We will briefly explain the basic steps of our new approach. The approach is applied as a postprocessing of a given partitioning, that is, in the form of a *repartitioning*. It is considered if we encounter a lack of balance between the sizes of the Cholesky factors. The result of this repartitioning step will be the new distribution of the graph into the domains which implicitly defines the graph separator as well. The repartitioning problem can be split into the two simpler subproblems. First, we need to decide *which vertices should be removed* from one domain and added to the other domain. Second, we need to find *where these removed nodes should be placed* in the reordering sequence of the other domain. This second reveals slight symmetry of these two tasks and shows that here we couple the standard graph partitioning problem with the problem of graph reordering. In the other words, in order to compare the above mentioned theoretical quantities with the help of the elimination tree we need to assume that the separated subgraphs were reordered, and we need to get the new reorderings as well. Note that first steps were considered in [5]. Here we will present some theoretical results along this line.

References

[1] T. A. Davis. *Direct Methods for Sparse Linear Systems*, SIAM, Philadelphia, Sept. 2006.

[2] B. Hendrickson and T. G. Kolda. *Graph partitioning models for parallel computing*, Parallel Computing, volume 26, number 12, pp. 1519–1534, 2000.

[3] J. W. H. Liu. *The role of elimination trees in sparse factorization* SIAM J. Matrix Anal. Appl. 11, pp. 134–172, 1990.

[4] J. W. H. Liu, E. G. Ng and B. W. Peyton. *On Finding Supernodes for Sparse Matrix Computations* SIAM J. Matrix Anal. Appl. 14, pp. 242–252, 1993.

[5] T. Pěnička. *"Dělení rozsáhlých a řídkých grafů v numerické lineární algebře*, draft of PhD. Thesis, 2006.

[6] A. Pinar and B. Hendrickson. *Combinatorial Parallel and Scientific Computing*, chapter in Parallel Processing for Scientific Computing, SIAM, 2006.

[7] A. Pinar and B. Hendrickson. *Partitioning for complex objectives*. Proceedings of Irregular'01, 2001.

[8] K. Schloegel, G. Karypis and V. Kumar. *A Unified Algorithm for Load-balancing Adaptive Scientific Simulations*. Proceedings of the International Conference on Supercomputing, Dallas, TX, November 2000.

The solution of problems with the pure Neumann boundary conditions on the outer boundary

R. Kohut

Institute of Geonics AS CR, Ostrava

Introduction

For some practical problems it is necessary to use the pure Neumann boundary conditions (see [1]). The pure Neumann boundary conditions are more flexible and we can expect higher accuracy of the results if we use this kind of boundary conditions, especially in cases when the outer boundary is not far enough from the considered sites. On the other hand this type of boundary conditions can cause some computational troubles.

The pure Neumann boundary value problem is solvable only if all the applied extenal forces (i.e. surface forces and volume forces given by the weight of rocks) are balanced which means that the resultants of all forces and their moments vanish. If the domain is not homogeneous (material with different weight, holes etc.) it is not simple to determine balanced forces and the condition of balanced forces can be disturbed. This disturbation although not very big, indicates some incorrectness in the model formulation and causes divergence of the used iterative method. For obtaining some numerical results, we must seek generalized solution and modify the numerical techniques.

Projection of rhs

The FE analysis of the boundary value problems of elasticity requires numerical solution of the linear system

$$Au = f, u, f \in R^n \tag{1}$$

where A is a large sparse symmetric matrix which is singular and positive semidefinite for pure Neumann boundary conditions.

Due to symmetry of A, the space R^n can be decomposed as

$$R^n = R(A) \oplus N(A), \tag{2}$$

where $R(A)$ is the range and $N(A)$ is the null space of the matrix A. We shall say that the system (1) is consistent if $f \in R(A)$. The consistent system has a (non unique) solution. Generally, the rhs f can be decomposed into consistent and inconsistent parts,

$$f = f_p + \hat{f}, f_p \in R(A), \hat{f} \in N(A). \tag{3}$$

Any solution of $Au = f_p$ is then called the generalized solution of (1).

Due to a not ideal balance between volume and boundary forces and also due to roundoff errors, we can obtain a slightly inconsistent system $Au = f$. In this case the PCG method converges at an initial phase and then starts to diverge. If the initial phase provides an iteration u_i with a

sufficiently small residual then u_i gives also a suitable approximation for the generalized solution of (1) because (due to orthogonality)

$$\| Au^i - f_p \| \leq \sqrt{\| Au^i - f_p \|^2 + \| \hat{f} \|^2} = \| Au^i - f \| . \tag{4}$$

For the elasticity problems, we know that the nullspace $N(A)$ consists from rigid translations and rotations. Thererefore, we can also project the rhs f into $R(A)$ ($P_R f = f_p$) and/or stabilize the iterative process by projecting the transformed residuals $w_i = B(r^i)$, where B is preconditioner into $R(A)$.

If w_1, w_2, w_3 are three independent rigid body translations,

$$w_1 = (1, 0, 0, 1, 0, 0, ...), \tag{5}$$
$$w_2 = (0, 1, 0, 0, 1, 0, 0, 1, 0, ...), \tag{6}$$
$$w_3 = (0, 0, 1, 0, 0, 1, ...), \tag{7}$$

and w_4, w_5, w_6 are three independent rigit body rotations,

$$w_4 = (0, -z_1, y_1, 0, -z_2, y_2, ...), \tag{8}$$
$$w_5 = (z_1, 0, -x_1, z_2, 0, -x_2, ...), \tag{9}$$
$$w_6 = (-y_1, x_1, 0, -y_2, x_2, 0, ...) \tag{10}$$

(vectors $w_1, .., w_6$ form the base of the nulspace $N(A)$), the projection $f_p = \in R(A)$ can be constructed numerically

$$f_p = f - \hat{f} = b - \sum \alpha_i w_i, for\, i = 1, .., 6 \tag{11}$$
$$f_p \perp \hat{f} \Rightarrow \sum \alpha_i < w_i, w_j > = < f, w_j > . \tag{12}$$

The coeficients $\alpha_i, i = 1, ..., 6$ can be determined solving system in (12). During PCG iterations the roundoff errors may cause instability and/or divergence which leads to finding a more substantial stabilization of the PCG algorithm. This can be done by projecting all the computed residuals back to the theoretical range $R(A)$, i.e.all computations of the residuals $r_i, i = 0, 1, ...$ are folowed by the projection

$$r_i := r_i - P_N(r_i). \tag{13}$$

Becase we'll do the projections after each iterations it's useful to orthonormalize the basis $\{w_1, w_2, w_3, w_4, w_5, w_6\}$. We obtain new orthonormalized basis $\{\tilde{w}_1, \tilde{w}_2, \tilde{w}_3, \tilde{w}_4, \tilde{w}_5, \tilde{w}_6\}$ and for the coeficients α_i it holds $\alpha_i = < f, \tilde{w}_i >, i = 1, ..., 6$.

In some practical problems (e.g. uniaxial pressure tests on specimens) we suppose normal Dirichlet nonhomogeneous boundary conditions on the two oposite faces of the hexahedral domain and homogeneous Neumann conditions on the other faces. In this case the nullspace $N(A)$ consists from two rigid translations and one rotation, both in the direction orthogonal to the direction of the given normal Dirichlet BC. If e.g. the Dirichlet BC are in the direction of z-coordinate, the corresponding nullspace vectors are w_1, w_2, w_6 (see (5),(8)).

The matrix A is singular and positive semidefinite for pure Neumann boundary conditions and for vectors $v \in N(A)$ the relations $Av = 0$ holds. But in the case of Dirichlet BC on two opposite sides we modify the matrix A in our software in such a way that the columns and rows corresponding to components of nodal vectors where the normal Dirichlet BC are given have all members equal zero except of diagonal member which is equal to 1. In the same way the

columns and rows corresponding to nodes in "empty" area (a hole in the domain) are modified. After this modification we receive matrix A_M which has following form:

$$A_M = \begin{pmatrix} A_S & 0 \\ 0 & D \end{pmatrix}, \tag{14}$$

where A_S is symmetric, singular and positive semidefinite, D is the unit diagonal matrix. The vectors v_M from nullspace $N(A_M)$ have form

$$v_M = \begin{pmatrix} v_S \\ 0 \end{pmatrix},$$

where the vectors $v_S \in N(A_S)$.

Numerical tests

The computed residuals are projecting back to the range $R(A)$ (see (13)). Due to the roundoff errors the functions v from the theoretical nullspace $N(A)$ do not fulfill exactly the condition $Av = 0$. If all matrices and vectors are stored in the single precision ($real * 4$), the l_2 norm of Aw_i, $i = 1, 6$ is between 10^{-2} and 10^{0}, in the case of the double precision ($real * 8$) the l_2 norm of Aw_i is between 10^{-10} and 10^{-6}. How these precisions influenced the results we can see on Figures 1 - 2. The test were done on model task with 1842750 unknowns for accuracy $\varepsilon = 1.0 \times 10^{-6}$. In the case of real*4 the "exact" residual vector $r_{ex} = Au_i - f_p$ shows the

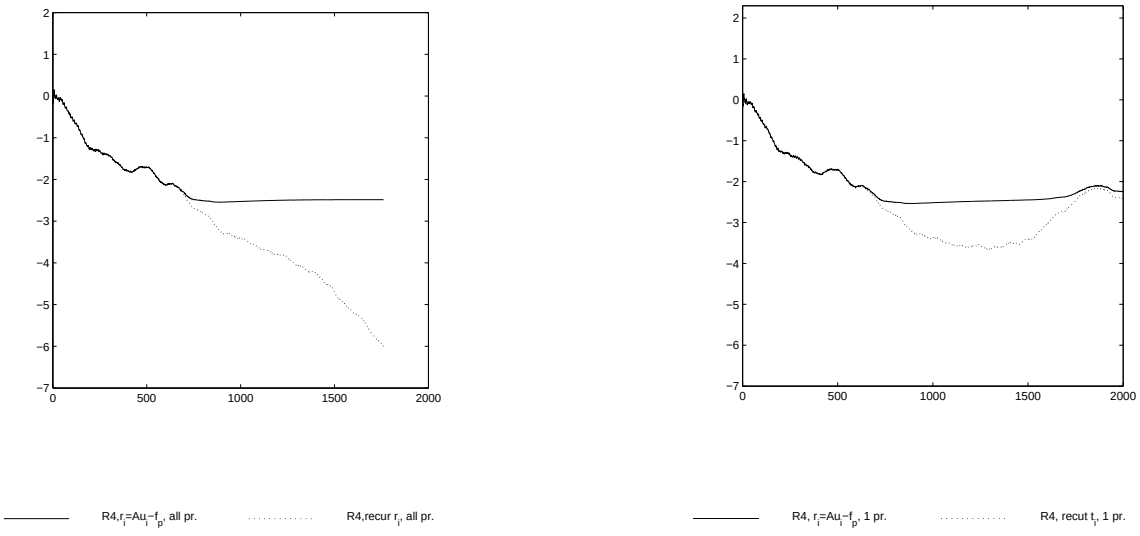


Figure 1: The behaviour of $\|r^k\|_{l_2}$, for "exact" r_k, "recurent" r_k, the precision real*4 : a)(left) the projection of all r_i, b)(right) the projection only for rhs f.

convergence till the value 5.0×10^{-3} both in Figure 1a and 1b, while $\|r^k_{rec}\|$ ($r^k = r^{k-1} + ...$) shows incorrectly the permanent convergence in the case with the projection in each iteration (Figure 1a).

In the case of real*8 the "exact" residual vector r_{ex} behaves in the same way as "recurent" r_k and the figures show that projection only of initial rhs is sufficient. If we compare the computed

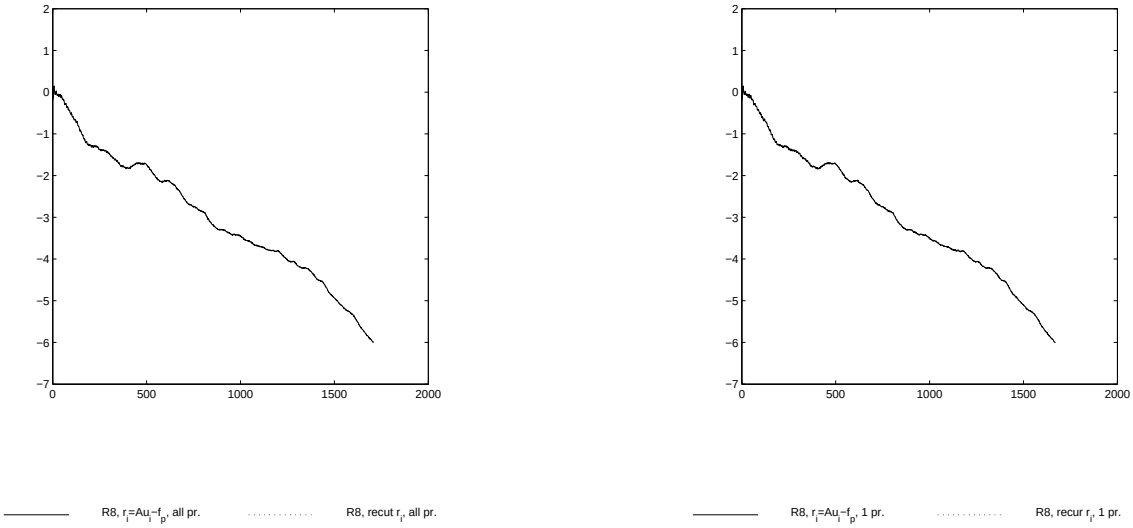


Figure 2: The behaviour of $\|r^k\|_{l_2}$, for "exact" r_k, "recurent" r_k, the precision real*8: a)(left) the projection of all r_k, b)(right) the projection only for rhs f.

stress fields for single and double precision, the difference is about 0.2%. It means that in our model example the accuracy close to 5.0×10^{-3} is sufficient and we can use the code using single precision.

Acknowledgment: The work was supported by the Grant Agency of the Czech Republic through the project 103/08/1700.

References

[1] R. Blaheta, O. Jakl, R. Kohut, J. Starý: Iterative Displacement Decomposition Solvers for HPC in Geomechanics. In: Large-Scale Scientific Computations of Engineering and Environment Problems 2. Proceedings of the Second Workshop on "Large-Scale Scientific Computations" Sozopol, Bulgaria, June 2-6, 1999. - (Ed.Griebel, M.; - Margenov, S.; - Yalamov, P.). - Wiesbaden, Vieweg 1999. - S. 347-356.

Adaptive *hp*-FEM for 3D Problems

P. Kus[1], *D. Andrs*[1,2], *P. Solin*[1,2]

[1]Institute of Thermomechanics AS CR, Prague
[2]University of Texas at El Paso

1 Introduction

We present a new adaptive *hp*-FEM based on arbitrary-level hanging nodes. The goal of this work is to create a general framework for solving partial differential equations corresponding to various physical fields. Achievement of this goal is essential for the next step of our work – solving coupled problems. Each field usually exhibits different behavior, such as singularities or boundary layers. *hp*-FEM method allowes us to use optimal type of elements for each field and each part of the computational domain, such as large higher-order elements for areas, where solution is smooth and small low-order elements close to singularities and boundary layers. This leads to better convergence, compared to standard *h*-adaptivity.

2 Constrained approximation

Constrained *hp*-FEM approximation was first introduced by Demkowicz [1] who uses one-level hanging nodes (both in 2D and 3D). It was demonstrated in [1] that the *hp*-FEM with one-level hanging nodes was more efficient than the approximation on regular meshes, but he still has been reporting problems with *forced refinements*. By forced refinements we mean refinements of elements which are not marked for refinement because of a large approximation error, but which are refined for technical reasons (to preserve mesh regularity). Forced refinements slow down the convergence of the adaptive process since the error is not reduced optimally, and moreover, they induce additional degrees of freedom whose numbers cannot be predicted easily due to their recursive nature.

3 Arbitrary level hanging nodes

In order to eliminate the forced refinements completely, we proposed a new *hp*-FEM with arbitrary-level hanging nodes for two-dimensional elliptic problems in [3] and generalized it later to two-dimensional time-harmonic Maxwell's equations solved by higher-order edge elements in [4]. In both cases, the absence of forced refinements improved the performance of automatic *hp*-adaptivity while simplifying its algorithmic treatment significantly.

4 Extension to 3D

The extension of the technique to 3D was nontrivial due to the more complex structure of higher-order shape functions and also because the structure of direct and indirect constraints in 3D is more complicated. Nevertheless, we can confirm once more that the technique was worth

developing – the algorithmic treatment of automatic *hp*-adaptivity in 3D (referred to as "programmer's nightmare" by Demkowicz) becomes modular and very simple, and the performance is much better compared to algorithms which need to deal with forced refinements. Numerical examples and comparisons are presented.

5 Example Application

Despite its higher programming complexity, adaptive *hp*-FEM is becoming increasingly popular in engineering circles due to its unconditional extremely fast convergence. In this study, we illustrate this fact using the standard benchmark example called Fichera corner. We solve the problem

$$\begin{aligned} -\Delta u &= f \quad \text{in } \Omega, \\ u &= u_D \quad \text{on } \partial\Omega, \end{aligned}$$

where $\Omega = (-1,1)^3 \setminus [0,1]^3$ and f and u_D are chosen to comply with the exact solution

$$u(x_1, x_2, x_3) = (x_1^2 + x_2^2 + x_3^2)^{1/4}.$$

The missing part of the cube represents a metallic object. The solution, representing the electric potential in the surrounding air itself is smooth, but it's gradient exhibits a strong singularity near the re-entrant corner and edges. The convergence curve shown in Fig. 1 was obtained after several iterations of the automatic adaptive algorithm, starting with seven hexahedral elements only. It can be seen that the *hp*-FEM outperforms both piecewise-linear and piecewise-quadratic FEM significantly.

From the graph shown in Fig. 1 it can be seen, that if one does not require very small relative error, quadratic, or even linear elements can be used. On the other hand, if one needs really good approximation with relative error below 0.1 percent, both linear and quadratic approximations become too expensive and *hp*-FEM is the most suitable. We can conclude that the future of the adaptive *hp*-FEM lies in large problems, where high accuracy is requested, such as singular or multiscale problems in 3D.

6 Conclusion

In this work we showed several aspects of *hp*-FEM adaptivity with arbitrary level hanging nodes. Despite its rather difficult algorithmic treatment, numerical results suggest, that this method can be successfully used. We believe, that its advantages will be even more significant for more complicated and coupled problems.

Acknowledgement: This research was supported by the Czech Science Foundation (Projects No. 102/05/0629 and 102/07/0496) and by the Grant Agency of the Academy of Sciences of the Czech Republic (Project No. IAA100760702)

References

[1] L. Demkowicz: Computing with *hp*-Adaptive Finite Elements, Volume 1, Chapman & Hall/CRC, 2007.

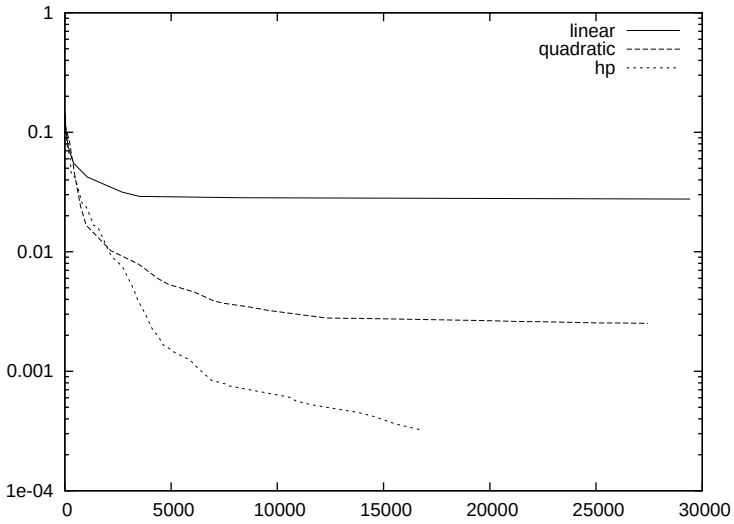


Figure 1: The convergence curves for linear FEM, quadratic FEM, and *hp*-FEM. Relative error in energy norm is shown on the vertical axis as a function of the number of DOF.

[2] P. Solin, K. Segeth, I. Dolezel: *Higher-Order Finite Element Methods*, Chapman & Hall/CRC, Boca Raton, 2003.

[3] P. Solin, J. Cerveny, I. Dolezel: Arbitrary-Level Hanging Nodes and Automatic Adaptivity in the *hp*-FEM, MATCOM, in press, doi:10.1016/j. matcom.2007.02.011.

[4] P. Solin, L. Dubcova, I. Dolezel: Adaptive *hp*-FEM with Arbitrary-Level Hanging Nodes for Time-Harmonic Maxwell's Equations, Research Report No. 2007-09, Department of Mathematical Sciences, University of Texas at El Paso, 2007, http://www.math.utep.edu/preprints/.

On pressure boundary conditions for steady flows of incompressible fluids with pressure and shear rate dependent viscosities

M. Lanzendörfer, J. Stebel

Institute of Computer Science AS CR, Prague
Institute of Mathematics AS CR, Prague

Abstract

We consider a class of incompressible fluids whose viscosities depend on the pressure and the shear rate. Suitable boundary conditions on the surface force at the inflow/outflow part of boundary are given. As an advantage of this, the mean value of the pressure over the domain is no more a free parameter which would have to be prescribed otherwise. We prove the existence and the uniqueness of weak solutions (the later for small data) and discuss particular applications of the results.

1 Introduction

A well-known property of the Navier-Stokes equations describing the motion of an incompressible Newtonian fluid is that the fluid pressure is determined to within a constant. This degree of freedom does not play important role as far as only the pressure gradient is present in the equations of motion. It is however not the case of fluids whose viscosities depend on the pressure and the shear rate. Since the value of the pressure affects the whole solution of the equations, one has to provide an additional parameter in order to fix this value.

In previous theoretical studies, such as [5], the mean value of the pressure either over the whole domain or over its nontrivial subdomain was prescribed as one of the input parameters. A difficulty of this approach lies in the fact that the pressure mean value is not a proper quantity from the practical point of view, i.e. there is no hint on the value which should be prescribed for a particular application. The objective of this paper is to propose an alternative way of fixing the pressure, namely to use a suitable inflow/outflow boundary condition. Proofs of the results can be found in [7].

2 Definition of the problem and the main result

We investigate the following system of PDEs:

$$\left.\begin{aligned} \operatorname{div}(\boldsymbol{v}\otimes\boldsymbol{v}) - \operatorname{div}\mathbf{S} + \nabla p &= \boldsymbol{f} \\ \operatorname{div}\boldsymbol{v} &= 0 \end{aligned}\right\} \quad \text{in } \Omega,$$

where

$$\mathbf{S} \equiv \mathbf{S}(p, \mathbf{D}(\boldsymbol{v})) = \nu(p, |\mathbf{D}(\boldsymbol{v})|^2)\mathbf{D}(\boldsymbol{v})\,. \tag{2.1}$$

Here $\boldsymbol{v}$, p, $\boldsymbol{f}$, $\nu(p, |\mathbf{D}(\boldsymbol{v})|^2)$ is the velocity, the kinematic pressure, the body force and the kinematic viscosity, respectively. The equations describe the motion of an incompressible homogeneous fluid in a bounded domain $\Omega \subset \mathbb{R}^d$, $d = 2$ or 3. The domain boundary consists of three parts: $\partial\Omega := \Gamma_D \cup \Gamma_1 \cup \Gamma_2$, on which we prescribe the following boundary conditions:

$$\boldsymbol{v} = \mathbf{0} \qquad \text{on } \Gamma_D\,, \tag{2.2}$$

$$p\boldsymbol{n} - \mathbf{S}\boldsymbol{n} = \boldsymbol{b}_1(\boldsymbol{v}) \qquad \text{on } \Gamma_1\,, \tag{2.3}$$

$$\left.\begin{aligned} \boldsymbol{v} &= (\boldsymbol{v}\cdot\boldsymbol{n})\,\boldsymbol{n} \\ p - \mathbf{S}\boldsymbol{n}\cdot\boldsymbol{n} &= b_2(\boldsymbol{v}) \end{aligned}\right\} \quad \text{on } \Gamma_2\,. \tag{2.4}$$

Throughout the paper we will assume that Ω has the Lipschitz boundary. Further we will denote $\Gamma := \Gamma_1 \cup \Gamma_2$ and suppose that $|\Gamma_D| > 0$ and $|\Gamma| > 0$, i.e. the Dirichlet condition (2.2) and at least one of the conditions (2.3), (2.4) is present.

2.1 Structural assumptions

The following assumptions on $\mathbf{S}$ are considered.

(A1) For a given $r \in (1,2)$, there are positive constants C_1 and C_2 such that for all symmetric linear transformations $\mathbf{B}$, $\mathbf{D} \in \mathbb{R}^{d\times d}$ and all $p \in \mathbb{R}$:

$$C_1(1+|\mathbf{D}|^2)^{\frac{r-2}{2}}|\mathbf{B}|^2 \le \frac{\partial \mathbf{S}(p,\mathbf{D}(\boldsymbol{v}))}{\partial \mathbf{D}} \cdot (\mathbf{B}\otimes\mathbf{B}) \le C_2(1+|\mathbf{D}|^2)^{\frac{r-2}{2}}|\mathbf{B}|^2\,,$$

where $(\mathbf{B}\otimes\mathbf{B})_{ijkl} = \mathbf{B}_{ij}\mathbf{B}_{kl}$.

(A2) For all symmetric linear transformations $\mathbf{D} \in \mathbb{R}^{d\times d}$ and for all $p \in \mathbb{R}$:

$$\left|\frac{\partial \mathbf{S}(p,\mathbf{D}(\boldsymbol{v}))}{\partial p}\right| \le \gamma_0(1+|\mathbf{D}|^2)^{\frac{r-2}{4}} \le \gamma_0\,,$$

with $\gamma_0 > 0$ specified later.

2.2 Boundary assumptions

Concerning the boundary conditions (2.3)–(2.4), we define

$$\langle \boldsymbol{b}(\boldsymbol{v}), \boldsymbol{\varphi}\rangle := \langle \boldsymbol{b}_1(\boldsymbol{v}), \boldsymbol{\varphi}\rangle_{\Gamma_1} + \langle b_2(\boldsymbol{v}\cdot\boldsymbol{n}), \boldsymbol{\varphi}\cdot\boldsymbol{n}\rangle_{\Gamma_2}$$

and assume the following:

(B1) With some $\gamma_1 \in \langle 3, r^*)$, the mapping

$$\boldsymbol{b}_1(\cdot) : \mathbf{L}^{\gamma_1}(\Gamma_1) \to \mathbf{L}^{\gamma_1}(\Gamma_1)^* \tag{2.5}$$

is continuous and bounded. Here $r^* := \frac{(d-1)r}{d-r}$ denotes the exponent for which $\mathbf{W}^{1,r}(\Omega) \hookrightarrow \mathbf{L}^{r^*}(\partial\Omega)$.

(B2) With some $\beta_1 \ge 0$,

$$\langle \boldsymbol{b}_1(\boldsymbol{u}), \boldsymbol{u}\rangle_{\Gamma_1} \ge -\frac{1}{2}\int_{\Gamma_1} (\boldsymbol{u}\cdot\boldsymbol{n})|\boldsymbol{u}|^2\,\mathrm{d}\boldsymbol{x} - \beta_1\|\boldsymbol{u}\|_{\gamma_1,\Gamma_1} \tag{2.6}$$

for all $\boldsymbol{u} \in \mathbf{L}^{\gamma_1}(\Gamma_1)$.

(B3) With some $\gamma_2 \ge 3$, the mapping

$$b_2(\cdot) : \mathrm{L}^{\gamma_2}(\Gamma_2) \to \mathrm{L}^{\gamma_2}(\Gamma_2)^* \tag{2.7}$$

is continuous and bounded.

(B4) With some $\beta_2 \ge 0$ and $\underline{\beta_2} > 0$,

$$\langle b_2(\boldsymbol{u}\cdot\boldsymbol{n}), \boldsymbol{u}\cdot\boldsymbol{n}\rangle_{\Gamma_2} \ge -\frac{1}{2}\int_{\Gamma_2} (\boldsymbol{u}\cdot\boldsymbol{n})|\boldsymbol{u}|^2\,\mathrm{d}\boldsymbol{x} + \underline{\beta_2}\,\|\boldsymbol{u}\|^{\gamma_2}_{\gamma_2,\Gamma_2} - \beta_2 \tag{2.8}$$

for all $\boldsymbol{u} \in \mathbf{L}^{\gamma_2}(\Gamma_2)$.

(B5) With some continuous function $m : \mathbb{R}^+ \to \mathbb{R}^+$, where $\lim\limits_{x\searrow 0} m(x) = 0$, b_2 is uniformly monotone:

$$\langle b_2(w) - b_2(z), w - z\rangle_{\Gamma_2} \ge m(\|w-z\|_{\gamma_2,\Gamma_2})\,, \tag{2.9}$$

for all $w \ne z \in \mathrm{L}^{\gamma_2}(\Gamma_2)$.

Additionally, in order to prove uniqueness of solutions we will require that the following stronger properties hold:

(B6) With some $\lambda_1 > 0$ and $K_1 > 0$ (specified later),

$$\|\boldsymbol{b}_1(\boldsymbol{u}^1) - \boldsymbol{b}_1(\boldsymbol{u}^2)\|_{\gamma_1',\Gamma_1} \le \lambda_1 \|\boldsymbol{u}^1 - \boldsymbol{u}^2\|_{\gamma_1,\Gamma_1} \tag{2.10}$$

for all $\boldsymbol{u}^1, \boldsymbol{u}^2 \in \mathbf{L}^{\gamma_1}(\Gamma_1)$, $\|\boldsymbol{u}^i\|_{\gamma_1,\Gamma_1} \le K_1$, $i = 1, 2$.

(B7) With some $\lambda_2 > 0$ and $K_2 > 0$ (specified later),

$$\|b_2(\boldsymbol{u}^1 \cdot \boldsymbol{n}) - b_2(\boldsymbol{u}^2 \cdot \boldsymbol{n})\|_{1,\Gamma_2} \le \lambda_2 \|\boldsymbol{u}^1 - \boldsymbol{u}^2\|_{r^*,\Gamma_2} \tag{2.11}$$

for all $\boldsymbol{u}^1, \boldsymbol{u}^2 \in \mathrm{L}^{\gamma_2}(\Gamma_2)$, $\|\boldsymbol{u}^i\|_{\gamma_2,\Gamma_2} \le K_2$, $i = 1, 2$.

2.3 Weak formulation

We define the following function spaces:

$$\begin{aligned}
\mathbf{W}^{1,r}_{\mathrm{b.c.}}(\Omega) &:= \left\{ \boldsymbol{v} \in \mathbf{W}^{1,r}(\Omega)\,;\ \operatorname{tr}\boldsymbol{v}\,\big|_{\Gamma_D} = \mathbf{0}\,,\ \operatorname{tr}\boldsymbol{v}\,\big|_{\Gamma_2} = (\operatorname{tr}\boldsymbol{v}\cdot\boldsymbol{n})\boldsymbol{n} \in \mathrm{L}^{\gamma_2}(\Gamma_2) \right\}\,, \\
\mathbf{W}^{1,r}_{\mathrm{b.c,div}}(\Omega) &:= \left\{ \boldsymbol{v} \in \mathbf{W}^{1,r}_{\mathrm{b.c}}(\Omega)\,;\ \operatorname{div}\boldsymbol{v} = 0 \text{ a.e. in } \Omega \right\}\,.
\end{aligned}$$

Definition 2 (Problem (P)) *A pair $(\boldsymbol{v}, p) \in \mathbf{W}^{1,r}_{\mathrm{b.c.,div}}(\Omega) \times \mathrm{L}^{r'}(\Omega)$ is said to be a weak solution of Problem (P) iff for every $\boldsymbol{\varphi} \in \mathbf{W}^{1,r}_{\mathrm{b.c.}}(\Omega)$*

$$\int_\Omega \operatorname{div}(\boldsymbol{v}\otimes\boldsymbol{v})\cdot\boldsymbol{\varphi}\,\mathrm{d}\boldsymbol{x} + \int_\Omega \mathbf{S}(p, \mathbf{D}(\boldsymbol{v})) : \mathbf{D}(\boldsymbol{\varphi})\,\mathrm{d}\boldsymbol{x} - \int_\Omega p \operatorname{div}\boldsymbol{\varphi}\,\mathrm{d}\boldsymbol{x} + \langle \boldsymbol{b}(\boldsymbol{v}), \boldsymbol{\varphi}\rangle = \langle \boldsymbol{f}, \boldsymbol{\varphi}\rangle. \tag{2.12}$$

2.4 Main result

Theorem 3 (Well-posedness of (P)) *Let $\boldsymbol{f} \in \mathbf{W}^{-1,r'}(\Omega)$ and assume that* **(A1)**–**(A2)** *hold for the viscosity,* **(B1)**–**(B5)** *hold for the boundary data, with*

$$\frac{3d}{d+2} < r < 2 \qquad \text{and} \qquad \gamma_0 < \frac{1}{\tilde{C}_{\mathrm{div}}(\Omega, \Gamma_1, \Gamma_2, 2)} \frac{C_1}{C_1 + C_2}. \tag{2.13}$$

Then

(i) *there exists a weak solution to (P);*

(ii) *for any weak solution $(\boldsymbol{v}, p)$ of (P), the velocity $\boldsymbol{v}$ satisfies the estimate*

$$\|\boldsymbol{v}\|_{1,r} + \|\boldsymbol{v}\|_{\gamma_2,\Gamma_2} \le K\,, \tag{2.14}$$

where $K \searrow 0$ whenever $(\|\boldsymbol{f}\|_{-1,r'}, \beta_1, \beta_2) \searrow \mathbf{0}$, the other problem data being fixed;

(iii) *if additionally* **(B6)**–**(B7)** *are satisfied and if K and λ_1, λ_2 are small enough, then the weak solution to (P) is unique.*

Remark 4 (Pressure is fixed by velocity) *Let $(\boldsymbol{v}, p^1)$ and $(\boldsymbol{v}, p^2)$ be weak solutions to (P). Then, under the assumptions of Theorem 3, $p^1 = p^2$.*

3 Boundary conditions in applications

Although the assumptions **(B1)**–**(B7)** seem to be motivated mainly by PDE analysis, they cover important engineering applications; we mention three types of them in the following.

Artificial boundary.

In numerical simulations, large or even unbounded domains arising from the physical model must be truncated and the boundary condition for artificial boundaries has to be provided. For example in [3], an application to the flow through a cascade of profiles with the outflow condition

$$-\mathbf{T}\boldsymbol{n} = \boldsymbol{h}(\boldsymbol{x}) + \frac{1}{2}(\boldsymbol{v}\cdot\boldsymbol{n})^{-}\boldsymbol{v} \tag{3.1}$$

is considered (see also Section 1). In [1], several b.c. including (3.1) were proposed (for unsteady incompressible Navier-Stokes equations) in order to perform long-time simulations at high Reynolds numbers.

Conditions involving Bernoulli's pressure.

In some applications, the quantity $p+\frac{1}{2}|\boldsymbol{v}|^2$, referred to as *total pressure* or *Bernoulli pressure*, is used for prescribing the inflow/outflow boundary conditions on artificial boundaries (see e.g. [6, 2]). Note that this class of conditions:

$$\left(p + \frac{1}{2}|\boldsymbol{v}|^2\right)\boldsymbol{n} - \mathbf{S}\boldsymbol{n} = \boldsymbol{h}(\boldsymbol{x}) \tag{3.2}$$

is covered by our theory.

Porous wall.

Boundary conditions of the type (2.4) are applicable to the flows, where an inflow/outflow is possible through a porous wall (*filtration* boundary conditions). In most studies, for the flow through an isotropic porous medium the linear law of Darcy is considered. However, Darcy's law is valid only for slow flows. It can be in fact derived from the Stokes equation, i.e. neglecting the inertia of the fluid, see e.g. [8]. For higher Reynolds numbers, the experimental observations "did not allow to find a universally accepted formula" [8]. Nevertheless, the relation

$$-\nabla p = \frac{\mu}{k}\boldsymbol{v} + d_2|\boldsymbol{v}|\boldsymbol{v} + d_3|\boldsymbol{v}|^2\boldsymbol{v}\,, \qquad \text{with } d_2, d_3 > 0, \tag{3.3}$$

was proposed more than a century ago in [4]. Here, the last two terms were added to make the equation fit the experimental results. Formula (3.3) with $d_3 = 0$ is well established as the Forchheimer equation.

As an analogy of (3.3), the boundary condition of the type

$$-\mathbf{T}\boldsymbol{n}\cdot\boldsymbol{n} = p_{\text{out}} + (c_1 + c_2|\boldsymbol{v}\cdot\boldsymbol{n}| + c_3|\boldsymbol{v}\cdot\boldsymbol{n}|^2)\,\boldsymbol{v}\cdot\boldsymbol{n} \qquad \text{with } c_1, c_2, c_3 \geq 0, \tag{3.4}$$

can be prescribed for the normal component of velocity. If $c_3 > 0$ or $c_2 > \frac{1}{2}$ then **(B3)**–**(B5)** and **(B7)** are satisfied (the last property for $K_2 > 0$ small enough).

Acknowledgements

Jan Stebel was supported by the Nečas Center for Mathematical Modelling project LC06052 financed by MŠMT. Martin Lanzendörfer acknowledges the support of Czech Science Foundation project GA201/06/0352.

References

[1] C. Bruneau and P. Fabrie, *Effective downstream boundary-conditions for incompressible Navier-Stokes equations*, International Journal for Numerical Methods in Fluids, 19 (1994), pp. 693–705.

[2] C. Conca, F. Murat, and O. Pironneau, *The Stokes and Navier-Stokes equations with boundary conditions involving the pressure*, Japan J. Math, 20 (1994), pp. 279–318.

[3] M. Feistauer and T. Neustupa, *On non-stationary viscous incompressible flow through a cascade of profiles*, Mathematical Methods in the Applied Sciences, 29 (2006), pp. 1907–1941.

[4] P. Forchheimer, *Wasserbewegung durch boden*, Z. Ver. Deutsch. Ing., 45 (1901), pp. 1781–1788.

[5] M. Franta, J. Málek, and K. R. Rajagopal, *On steady flows of fluids with pressure- and shear-dependent viscosities*, Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci., 461 (2005), pp. 651–670.

[6] J. Heywood, R. Rannacher, and S. Turek, *Artificial boundaries and flux and pressure conditions for the incompressible Navier-Stokes equations*, Internat. J. Numer. Methods Fluids, 22 (1996), pp. 325–352.

[7] M. Lanzendörfer and J. Stebel, *On Pressure Boundary Conditions for Steady Flows of Incompressible Fluids with Pressure and Shear Rate Dependent Viscosities*, Preprint 15, Nečas Center for Mathematical Modeling, 2008.

[8] A. Mikelić, *Homogenization theory and applications to filtration through porous media*, in Filtration in Porous Media and Industrial Application, A. Fasano, ed., vol. 1734 of Lecture notes in mathematics, Springer-Verlag, 2000.

Mathematical modeling of geosynthetic tubes

J. Malík

Institute of Geonics AS CR, Ostrava

1 Introduction

Geosynthetic tubes are comprised of thin sheets and pumped with water or slurry. The tubes are made of synthetic fabrics (geotextile). They have been used as dikes or breakwaters and to prevent beach erosion. They have many other applications in geoengineering (see [2]).

Geosynthetic tubes on rigid foundation are studied, for instance, in [1, 3]. These results are generalized for tubes on elastic foundation [5]. Geosynthetic tubes in mutual contact are studied in [6]. Some problems connected with 3D modeling are solved in [7]. Similar techniques have been applied for solving some quite different problems. Floating liquid filled membranes are studied in [8, 9]. The shape of a towed boom of logs is studied in [4].

The main purpose of this paper is to give the strict mathematical formulation and analysis of some problems connected with the geosynthetic tubes on the rigid foundation. These problems can be of practical use and their solutions can contribute to the optimal design. The basic governing equations to the problems are presented. These problems are studied in [11].

Similar problems are studied in [1]. First of all the authors deal with extensible elastic membranes holding liquid and gases. Inextensible membranes are studied as a limit case of extensible ones. In this paper inextensible membranes are studied, which is quite natural for the problems connected with geosynthetical tubes.

2 Basic hypotheses and setting up problems

Geosynthetic tubes have diameters ranging from one to several meters and have theoretically infinite length. Let us consider that all cross sections are identical, so we can study the geosynthetic tubes as a two-dimensional problem. The modeling is based on the following hypotheses:

1. The geosynthetic is inextensible and flexible and its weight can be neglected.
2. The filling medium (water or slurry) behaves as an ideal liquid which generates hydrostatic pressure in every point and act in the perpendicular direction to the geosynthetic.
3. There is no friction between the foundation and the geosynthetic.

The geosynthetic tube is filled through the inlets on the top of the tube, which results in the process, where the certain part of the geosynthetic rises and the other part of the geosynthetic rests on the rigid foundation (see Fig.1.).

Let us consider the coordinates in Fig.1 with the origin in the point O and with the axes x, y oriented in the way depicted in Fig.1. Let us use the notation

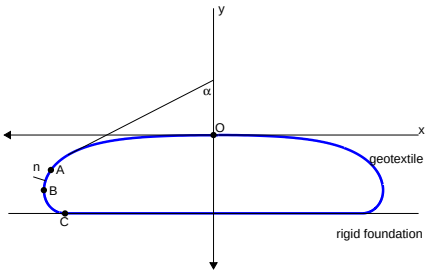


Fig.1 Cross section of a geosynthetic tube

ρ the density of the water or slurry.
g the gravitational acceleration.
p the pressure of the water or slurry at the point O.

The pressure p can be interpreted as the pumping pressure of the water or slurry which is transported into the tube. Let us set up equilibrium conditions on the curve representing the shape of the cross section of the geosynthetic tube.

Let s be the parameter representing the length of the curve. The parameter is equal 0 in the point O and is oriented in the anticlockwise direction.

Let $n = (n_x, n_y)$ be the normal vector to the curve, $H(s)$ be the tension force in the geosynthetic in the point corresponding to the parameter s, and the functions $x(s)$, $y(s)$ describe the shape of the curve between the points O, C.

The basic equlibrium equations read as follows

$$\begin{aligned} \frac{d}{ds}\left(H\frac{dx}{ds}\right) + \frac{dy}{ds}\left(g\rho y + p\right) = 0, \\ \frac{d}{ds}\left(H\frac{dy}{ds}\right) - \frac{dx}{ds}\left(g\rho y + p\right) = 0, \end{aligned} \tag{2.1}$$

which hold on the interval OC (see Fig. 1.).

3 Numerical solutions.

This section contains some examples connected with the mathematical modeling og geosynthetical tubes.

The algorithms were implemented in MATLAB. Now let us apply the MATLAB code for solving some model problems. Let us consider that we have a tube with the perimeter $10m$ filled with water ($\rho = 1000kg/m^3$) and $g = 10m/s^2$. The graphs in Fig. 2 describe the shapes of the tube for some values of the parameters p, H and $L = 10m$. The graphs in Figs. 3-5 describe the functional dependences between h and p, H, V for $L = 10m$.

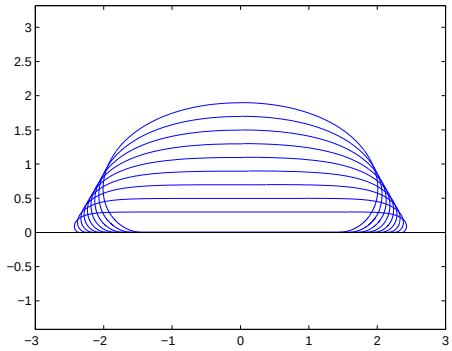


Fig.2 The shapes of the tube with the perimeter $10m$ for some values of height

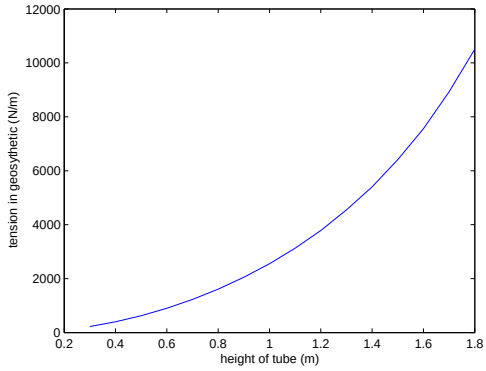


Fig.3 The functional dependence between h and H for the tube with the perimeter $10m$

4 Conclusion

From the graphs above it is clear that the dependence between the parameters p, H, L, h, V is nonlinear. The result show how to choose some parameters of the geotextile so that the tension H does not exceed the limits which can result in a destruction of the tube. Such information can contribute to the optimal design.

Acknowledgment

The research was supported by Grant GACR 103/08/1700.

References

[1] S. S. Antman, M. Schagerl, *Slumping instabilities of elastic membranes holding liquids and gases,* International Journal of Non-linear Mechanics, 40 (2005) pp. 1112-1038.

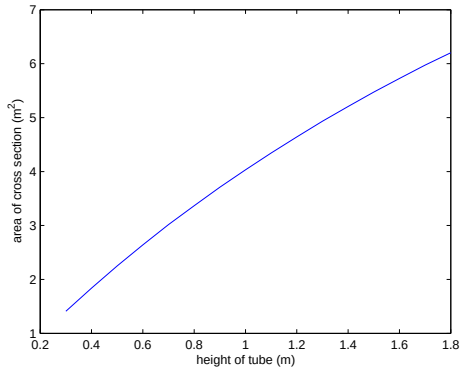


Fig.4 The functional dependence between h and p for the tube with the perimeter $10m$

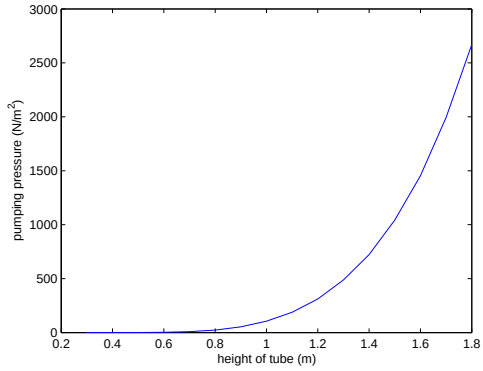


Fig.5 The functional dependence between h and V for the tube with the perimeter $10m$

[2] K. Kazimierowicz, *Simple analysis of deformation of sand-sausages,* Proceedings of 5th Conference on Geotextile, Geomembranes, and Related Products, Singapore, 1994, pp. 775-778.

[3] R. M. Koerner, *Designing with Geosynthetics,* Prentice Hall, New York, 1994.

[4] D. Leshchinsky, O. Leshchinsky, H. I. Ling, P. A. Gilbert, *Geosynthetic tubes for confining pressurized slurry: some design aspects,* Journal of Geotechnical Engineering, 122 (1996), pp. 682-690.

[5] B. G. Newman, *Shape of a towed boom of logs,* Proceedings of the Royal Society of London, A.346 (1975), pp. 329-348.

[6] R. H. Plaut, S. Suherman, *Two-dimensional analysis of geosynthetic tubes,* Acta Mechanica, 129 (1998), pp. 207-218.

[7] R. H. Plaut, C. R. Klusman, *Two-dimensional analysis of stacked geosynthetic tubes on deformable foundations,* Thin-Walled Structures, 34 (1999), pp. 179-194.

[8] P. A. Seay, R. H. Plaut, *Three-dimensional behavior of geosynthetic tubes,* Thin-Walled Structures, 32 (1998, pp. 263-274.

[9] R. A. Zhao, *A complete linear theory for two dimensional floating and liquid-filled membrane structures in waves,* Journal of Fluids and Structures, 9 (1995), pp. 937-956.

[10] R. A. Zhao, J. V. Aarsnes, *Numerical and experimental studies of a floating and liquid-filled membrane structure in waves,* Ocean Engineering, 25 (1998) pp. 753-765.

[11] J. Malík, *Some problems connected with 2D modiling og geosythetical tubes,* Nonlinear Analysis:Real World Applications, 10 (2009) pp. 810-823.

Modifications of IAD methods for large scale computing

I. Pultarová

Czech Technical University in Prague

1 Introduction

The iterative aggregation - disaggregation (IAD) methods attract an attention especially due to their ability to solve large or not well conditioned problems. Nevertheless, the convergence analysis has not brought enough satisfactory results yet. There are available a lot of modifications of the aggregation approach, still the convergence of these algorithms is mostly controlled by checking the error and by changing the number of basic iterations and this in general cannot be estimated in advance. In this short contribution we introduce some new observations and estimates on the spectra of the error matrices which are connected to the IAD methods.

2 Solving Perron eigenvector by the IAD metod

We assume an $N \times N$ column stochastic matrix B. We want to get an eigenvector $\hat{x}$ of B for which $B\hat{x} = \hat{x}$, $e^T\hat{x} = 1$, where e is an all ones vector. We will assume that B is irreducible which implies that $\hat{x}$ is unique.

We may divide the set of indices $\{1, 2, \ldots, N\}$ into $n \leq N$ subgroups $G_1, \ldots, G_n$ and consider them as for a new set of macro-states. Let the ordering fulfils that if $i \in G_k, j \in G_m, k < m$ then $i < j$. We need the following notation. Let R be an $n \times N$ matrix for which $R_{ij} = 1$ if $j \in G_i$ and $R_{ij} = 0$ otherwise. For any positive vector x we define a matrix $S(x)$ with the elements $S(x)_{ij} = x_i / \sum_{k \in G_j} x_k$ if $i \in G_j$ and $S(x)_{ij} = 0$ otherwise. Let $P(x) = S(x)R$.

Let us denote $B(x)_a$ the aggregated matrix $B(x)_a = RBS(x)$. Starting with some positive vector x^0 we solve the equation

$$B(x^0)_a z = z$$

for z. Then z is prolonged to the size N by $y = S(x^0)z$ and several steps of some basic iterative method is performed. We may use e.g. the power method, Jacobi or Gauss-Seidel methods or their block forms. Let us denote $M - W$ some weak nonnegative splitting of $I - B$, where I is an identity matrix. Then the basic iteration matrix will be $T = M^{-1}W$. Then $x^1 = T^m y$ for some chosen integer m. This finishes one loop of the IAD method.

It was derived that for the sequence of the computed approximations it is $x^{k+1} - \hat{x} = J(x^k)(x^k - \hat{x})$, where the error matrix is

$$J(x) = T^m(I - P(x)Z)^{-1}(I - P(x)),$$

where $Z = B - \hat{x}e^T$. In the next section we show some properties of $J(\hat{x})$ for some special structures of data.

3 Convergence and divergence in local sense

We know that cycles in B may cause divergence of IAD methods even in local sense. That is why we study such kind of matrices in this paper. Matrix B is assumed to be cyclic, $B_{1,n} = 1$, $B_{i+1,i} = 1$ for $i = 1, 2, \ldots, N-1$, and $B_{ij} = 0$ otherwise. Now we study the spectral radius of $J(\hat{x})$ in order to determine the asymptotic rate of convergence of the IAD method or to prove divergence. Since we want to distinquish among several types of the IAD methods in which different basic iteration matrices are used, we denote the corresponding error matrix $J(T^m, \hat{x})$. Let us denote B_1 the block diagonal of B where the indices of the particular blocks correspond to the aggregation groups $G_1, \ldots, G_n$.

Lemma 1. Asymptotic spectral radii of the error matrices corresponding to the IAD methods for the basic iteration matrices B, B^N and $(I - B_1)^{-1}B_2$, respectively, are

$$\rho(J(B, \hat{x})) = 1,$$

$$\rho(J(B^N, \hat{x})) = 1,$$

and

$$\rho(J((I - B_1)^{-1}B_2, \hat{x})) = 0,$$

respectively.

It is assumed that the IAD methods converge for great part of the set of irreducible stochastic matrices. But we introduce examples, that for cyclic B the spectral radius of $J(B^{N-1}, \hat{x})$ can be arbitrarily close to 2. In Figure 1 one can see the spectra (dots in bold) of the error matrices $J(B^{N-1}, \hat{x})$ for the partitioning with two blocks each including 100 elements, and for 30 blocks each of 20 elements. There are also displayed two thin circles in each figure, which help to recognize the location of the eigenvalues.

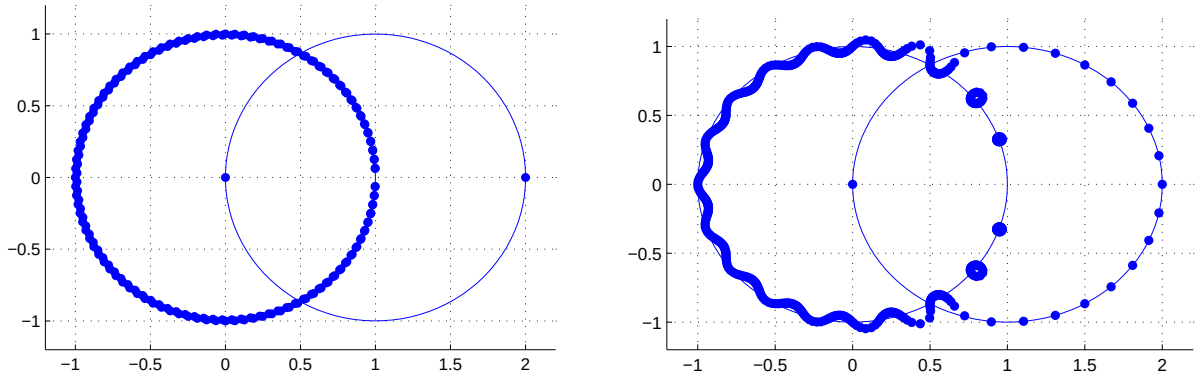


Figure 1: Spectra of $J(B^{N-1}, \hat{x})$ for $N = 200$, $n = 2$ (left) and for $N = 600$, $n = 30$ (right).

From the above considerations we may conclude that two following properties of the IAD methods are important. Firstly, when B is close to a cyclic matrix, the proper ordering is desirable. And secondly, an appropriate basic iteration matrix in this case is $(I-B_1)^{-1}B_2$, where $(I-B_1)^{-1}$ can be substituted by $I + B_1 + B_1^2 + \cdots + B_1^m$.

Acknowledgement: This work has been supported by the project CEZ MSM 6840770001.

Implementation of the BDDC method based on the frontal and multifrontal algorithm

J. Šístek, J. Novotný, P. Burda, M. Čertíková

Institute of Mathematics AS CR, Prague
Institute of Thermomechanics AS CR, Prague
Czech Technical University in Prague
Czech Technical University in Prague

1 Introduction

Numerical solution of linear problems arising from isotropic elasticity discretized by finite elements is important in many areas of engineering. The matrix of the system is typically large, sparse, and ill-conditioned. The classical frontal solver [2] has become a popular direct method for solving problems with such matrices arising from finite element analyses. For large problems, iterative methods such as the preconditioned conjugate gradients (PCG) are usually less expensive in terms of memory and computational time. However, their convergence rate deteriorates with growing condition number of the solved linear system and good preconditioning becomes essential. The need of first-rate preconditioners tailored to the solved problem that can be implemented in parallel gave rise to the field of domain decomposition methods [3].

The Balancing Domain Decomposition based on Constraints (BDDC) [4, 5] is one of the most advanced preconditioners of this class. However, the additional custom coding effort required represents a difficulty in incorporating the method to an existing finite element code. We propose an implementation of BDDC built on top of common components of existing finite element codes – the frontal solver and the element stiffness matrix generation. The implementation requires only a minimal amount of additional code.

2 The BDDC method

After discretization by the finite element method (FEM), the linear system $Ku = f$ is to be solved for a vector u of unknown values of displacements at nodes of a given domain.

The domain is split into nonoverlapping subdomains with the *interface* formed by unknowns common to at least two subdomains. Then the problem is reduced to the *Schur complement* problem with respect to the interface and this reduced problem is solved by PCG method. The BDDC method is used as a preconditioner, that splits the computation of the preconditioned residual needed in every iteration of PCG to solution of independent *subdomain problems* (2.1) and the global *coarse problem* (2.2). The preconditioned residual is obtained as a combination of their solutions (for details see [3, 4, 5]).

The subdomain problems can be expressed as saddle point problems

$$\begin{bmatrix} K^i & C^{iT} \\ C^i & 0 \end{bmatrix} \begin{bmatrix} u^i \\ \mu^i \end{bmatrix} = \begin{bmatrix} r^i \\ 0 \end{bmatrix}, \tag{2.1}$$

where K^i denotes the subdomain local stiffness matrix, matrix C^i enforces zero values of *coarse degrees of freedom* and so ensures continuity constraints at coarse degrees of freedom across the interface, and μ^i is the vector of Lagrange multipliers. Matrix C^i contains both constraints enforcing continuity across corners (point constraints), and constraints enforcing equality of averages over edges and faces of subdomains. The former type corresponds to just one nonzero entry equal to 1 on a row of C^i, while the latter leads to several nonzero entries on a row.

The coarse problem for coarse unknowns u_c is

$$K_c u_c = r_c, \tag{2.2}$$

where the *coarse matrix* K_c can be assembled from local coarse matrices in a similar way as the global stiffnes matrix is assembled from element matrices in standard FEM. Construction of a local coarse matrix also relies on efficient solution of problem (2.1).

3 The implementation

The frontal solver implements the solution of a square linear system $Ax = f$ with some of the variables having prescribed values. Equations that correspond to these fixed variables are omitted and the values of these variables are substituted into the solution vector directly. The output of the solver consists of the solution and the resulting imbalance in the equations, called reaction forces. In matrix notation this can be expressed as

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} f_1 \\ f_2 \end{bmatrix} + \begin{bmatrix} 0 \\ r_2 \end{bmatrix}, \tag{3.1}$$

where fixed variable values x_2 and the load vectors f_1 and f_2 are the inputs, the solution x_1 and the reaction r_2 are the outputs. System matrix A is not assembled or stored as a whole, instead stiffness matrices of elements are subsequently assembled and eliminated as needed during the factorization.

As the frontal solver treats naturally only point constraints, the implementation relies on the separation of point constraints and enforcing the rest by Lagrange multipliers, as suggested already in [4]. An early version of the implementation that used simplified coarse problem based only on point constraints was presented in [6].

The local substructure problems (2.1) can be written in the frontal solver form (3.1) (with index i omitted for simplicity) as

$$\begin{bmatrix} K_{ff} & K_{fc} & C_f^T \\ K_{cf} & K_{cc} & 0 \\ C_f & 0 & 0 \end{bmatrix} \begin{bmatrix} v_f \\ 0 \\ \mu \end{bmatrix} = \begin{bmatrix} r \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ R \\ 0 \end{bmatrix}, \quad \text{where} \quad K^i = \begin{bmatrix} K_{ff} & K_{fc} \\ K_{cf} & K_{cc} \end{bmatrix}, \tag{3.2}$$

subscript $_c$ denotes coarse variables representing point constraints, subscript $_f$ denotes the rest of the variables, R is the residual at point constraints and the block $[C_f \;\; 0]$ involves only the rows of C^i that represent constraints on averages (point constraints are omitted).

From (3.2) the problem for Lagrange multipliers μ can be extracted as

$$C_f K_{ff}^{-1} C_f^T \mu = C_f K_{ff}^{-1} r,$$

the matrix of which is dense but small with the order equal to the number of averages on the subdomain and can be factorized directly. After computing μ and substituting it into (3.2), the

subdomain problem takes form suitable for the frontal solver:

$$\begin{bmatrix} K_{ff} & K_{fc} \\ K_{cf} & K_{cc} \end{bmatrix} \begin{bmatrix} v_f \\ 0 \end{bmatrix} = \begin{bmatrix} r - C_f^T \mu \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ R \end{bmatrix},$$

from which the subdomain correction v_f can be computed.

Coarse problem (2.2) is solved by the multifrontal algorithm using the package MUMPS ([1]) just like an ordinary finite element problem, with subdomains playing the role of elements. Thus the coarse matrix in not assembled as a whole but stored distributed among processors as local coarse matrices.

Detailed description of the implementation can be found in [7].

4 Numerical results

The method was applied to a problem of stress analysis of a mine reel. The computational mesh consists of 140 816 quadratic elements, 579 737 nodes and 1 739 211 degrees of freedom. Its division into 16 subdomains is presented in Figure 1 with a detail of computational mesh of the steel rope. Here, the neighbouring subdomain is hidden to reveal the difficult interface.

An experiment with adding constraints on averages to the optimal set (2 000) of coarse nodes is summarized in Table 1. The iterface is divided into 2 edges and 22 faces. We can see, that the choice of averages is rather delicate task. The effect of edges is negligible in comparison to the effect of faces due to their number. However, although additional averages improve the condition number and reduce the number of PCG iterations, they may not necessarily reduce the computational time, since the time saved on iterations may be spent in factorization of the larger coarse problem.

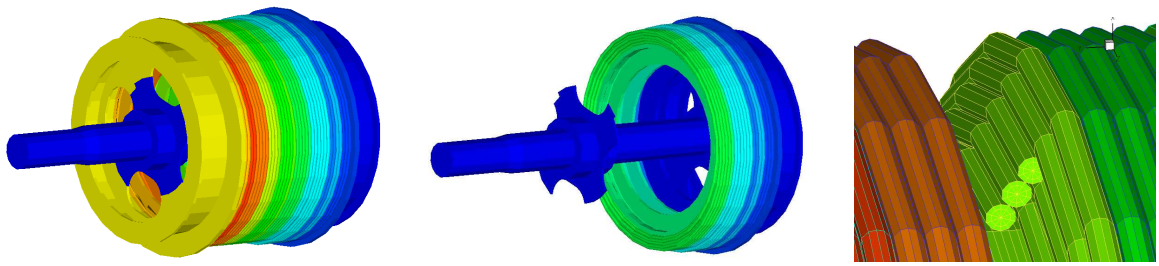


Figure 1: Mine reel problem, division into 16 subdomains (left), central part of 8 subdomains (centre), and detail of the computational mesh of the steel rope (right)

5 Conclusion

We have presented an approach to implementation of the recent BDDC method using common components of finite element codes, such as frontal solver and matrix assembly process. The key idea here is the different treatment of pointwise continuity constraints and equality of averages over edges and faces across subdomains. In this way, we are able to minimize the amount of additional code that is necessary for the BDDC method. The approach was implemented into our previous implementation based only on coarse nodes.

coarse problem	c	c+e	c+f	c+e+f
iterations	142	141	117	112
cond. number est.	13 982	13 982	1 287	1 272
factorization (sec)	12 694	12 956	15 142	15 309
pcg iter (sec)	4 138	4 097	3 124	3 406
total (sec)	17 532	17 753	18 965	19 423

Table 1: Mine reel problem, 16 subdomains, 2 000 coarse nodes, 'c' – continuity in coarse nodes, 'e' – equivalence of averages over edges, 'f' – equivalence of averages over faces

Acknowledgement: This research has been supported by the Czech Science Foundation under grant GA CR 106/08/0403, by the Grant Agency of the Academy of Sciences of the CR under grant IAA200600801, and by projects MSM 6840770001 and MSM 6840770010. It has also been supported by Institutional Research Plans AV0Z 10190503 and AV0Z 20760514. A part of this work was done while Jakub Šístek was visiting professor Jan Mandel at the University of Colorado Denver. The mesh of the problem of mine reel was kindly provided by Jan Leština, Vamet, Ltd.

References

[1] Amestoy, P. R., Duff, I. S., and L'Excellent, J.-Y.: Multifrontal parallel distributed symmetric and unsymmetric solvers, Comput. Methods Appl. Mech. Engrg. 184 (2000), pp. 501–520.

[2] Irons, B. M.: *A frontal solution scheme for finite element analysis* Internat. J. Numer. Methods Engrg. 2 (1970), pp. 5–32.

[3] Toselli, A. and Widlund, O.: *Domain decomposition methods—algorithms and theory* vol. 34 of Springer Series in Computational Mathematics, Springer-Verlag, Berlin (2005).

[4] Dohrmann, C. R.: *A preconditioner for substructuring based on constrained energy minimization.* SIAM J. Sci. Comput. 25 (2003), pp. 246–258.

[5] Mandel, J., Dohrmann, C. R. and Tezaur, R.: *An algebraic theory for primal and dual substructuring methods by constraints* Appl. Numer. Math. 54 (2) (2005) 167–193.

[6] Šístek, J., Čertíková, M., Burda, P., Neumanová, E., Pták, S., Novotný, J. and Damašek, A.: Development of an efficient parallel BDDC solver for linear elasticity problems, in: Blaheta, R. and Starý, J. (ed.), Proceedings of Seminar on Numerical Analysis, SNA'07, Ostrava, Czech Republic, January 22–26, Institute of Geonics AS CR, Ostrava, 2007, pp. 105–108.

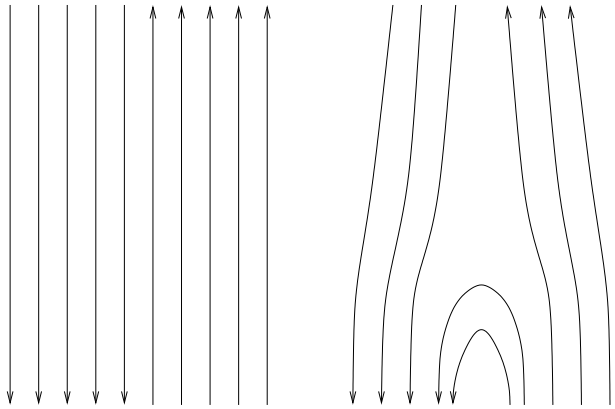
[7] Šístek, J., Novotný, J., Mandel, J., Čertíková, M. and Burda, P.: BDDC by a frontal solver and stress computation in a hip joint replacement, (2008), to appear, preprint: http://arxiv.org/abs/0802.4295.

Modelování rekonexe magnetických polí ve sluneční koróně metodou konečných prvků

J. Skála[1], *M. Bárta*[2], *M. Varady*[1,2]

[1] Univerzita J. E. Purkyně, Ústí nad Labem
[2] Astronomický ústav AV ČR, Ondřejov

Při rekonexi magnetického pole dochází ke změně jeho topologie tak, že pole přejde z konfigurace které odpovídá vyšší energie do konfigurace s energií nižší (viz. obr. 1). Energie uvolněná při rekonexi hraje klíčovou roli v řadě dynamických procesů ve sluneční atmosféře. Nejenergetičtější z těchto procesů jsou sluneční erupce, kdy se během rekonexe uvolní na časových škálách $\sim 10 - 100$ s ohromné množství energie v řádech $10^{22} - 10^{25}$ J. Z pozorování je známo, že energie uvolněná při rekonexi se transformuje do vysoce energetických svazků elektronů a protonů, magneto-hydrodynamických (MHD) vln, do energie plazmoidů vyvržených ze sluneční korony a podobně. Kinetická energie svazků částic směrovaných ke sluneční fotosféře se termalizuje v hustých vrstvách atmosféry a v případě slunečních erupcí zde dochází k prudkému ohřevu plazmatu a mohutnému explozivnímu vypařování hustého plazmatu podél magnetických siločar směrem do korony (předpokládá se plazma s nízkým β parametrem[4]). V důsledku toho se okolí magnetických siločar v koróně naplní plazmatem s teplotou až 30 MK a hustotou řádově 10^{16} m^{-3}. Takto popisuje vznik erupce tzv. standardní model slunečních erupcí.



Obrázek 1: Schématické znázornění rekonexe magnetického pole ve 2D geometrii. Na levé části obrázku je magnetické pole před rekonexí – ve vyšším energetickém stavu, na pravé je magnetické pole po rekonexi – v nižším energetickém stavu. Uzavřené siločáry mg. pole (ve slunečních erupcích představují erupční smyčky) vznikly přepojením, neboli rekonexí původních anti-paralelních otevřených siločar.

Velkoškálová dynamika magnetického pole a plazmatu při rekonexi se standardně modeluje v tzv. magnetohydrodynamickém (MHD) přiblížení. Chování plazmatu popisují zjednodušené Maxwellovy rovnice společně s Ohmovým zákonem a soustavou hydrodynamických zákonů zachování [1]. Základními předpoklady nerelativistické magnetohydrodynamiky jsou dostatečně velké prostorové a časové škály popisovaných procesů a nízké rychlosti plazmatu (v porovnání s rychlostí světla), což umožňuje zanedbání posuvného proudu v Maxwellových rovnicích. Rovnice kontinuity pro elektrický náboj se tak zjednoduší na tvar $\nabla \cdot \mathbf{j} = 0$. Základní rovnice MHD

[4]Parametr β je poměr tlaku plazmatu ku magnetickému tlaku.

jsou

$$\begin{aligned}
\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{v}) &= 0 \,, \\
\frac{\partial \rho \mathbf{v}}{\partial t} + \nabla \cdot (\rho \mathbf{v}\mathbf{v}) &= -\nabla p + \mathbf{j} \times \mathbf{B} \,, \\
\frac{\partial E}{\partial t} + \nabla \cdot \mathbf{S} &= 0 \,, \\
\frac{\partial \mathbf{B}}{\partial t} &= \nabla \times (\mathbf{v} \times \mathbf{B}) + \nabla \times (\eta \mathbf{j}) \,, \\
\mathbf{j} &= \nabla \times \mathbf{B}/\mu_0 \,, \\
p &= \frac{k_\mathrm{B}}{m} \rho T \,,
\end{aligned}$$

kde $\mathbf{B}$ je magnetická indukce, μ_0 je permeabilita vakua, v je makroskopická rychlost plazmatu, ρ je hustota plazmatu, p tlak plazmatu, η je rezistivita plazmatu, m je střední hmotnost částic, k_B je Boltzmanova konstanta a T je termodynamická teplota plazmatu. E je celková energie a S je tok energie [3]

$$\begin{aligned}
E &= \frac{p}{\gamma - 1} + \frac{1}{2}\rho v^2 + \frac{B^2}{2\mu_0} \,, \\
\mathbf{S} &= \left(U + p + \frac{B^2}{2\mu_0} \right) \mathbf{v} + \frac{\mathbf{v} \cdot \mathbf{B}}{\mu_0} \mathbf{B} + \frac{\eta}{\mu_0} \mathbf{j} \times \mathbf{B} \,.
\end{aligned}$$

Pro účely simulací je vhodné přepsat rovnice MHD do konzervativního tvaru

$$\begin{aligned}
\frac{\partial \rho}{\partial t} &= -\frac{\partial}{\partial x_j} \left(\rho v_j \right) \,, \\
\frac{\partial \rho v_i}{\partial t} &= -\frac{\partial}{\partial x_j} \left[\rho v_i v_j - \frac{B_i B_j}{\mu_0} + \delta_{ij} \left(\frac{|\mathbf{B}|^2}{2\mu_0} + p \right) \right] \,, \\
\frac{\partial E}{\partial t} &= -\frac{\partial}{\partial x_j} S_j \,, \\
\frac{\partial B_i}{\partial t} &= \varepsilon_{ijk} \frac{\partial}{\partial x_j} \left(\varepsilon_{klm} v_l B_m - \eta j_k \right) \,.
\end{aligned}$$

Tato soustava rovnic se v současnosti standardně řeší metodou konečných diferencí, přičemž v místech s velkými gradienty, a tedy současně také v oblastech kde lze očekávat zajímavé fyzikální procesy jako například urychlování svazků částic [4], se síť zjemňuje pomocí různých adaptivních metod [2], které v závislosti na vývoji simulace generují strukturované (e.g. PARAMESH [5]) nebo nestrukturované sítě. Tento způsob řešení problému naráží na řadu obtíží například při správném ošetření okrajových podmínek na hranici sítí s různým rozlišením, při volbě časového kroku nebo při paralelní implementaci kódu. Na druhou stranu metoda konečných prvků (FEM) umožňuje bez problému konstrukci nestrukturované sítě a přesnější implementování okrajových podmínek. Přestože jsou metody FEM hojně užívány pro numerické modelování v mnoha fyzikálních i technických oborech, ve fyzice plazmatu a magnetohydrodynamice je jejich využití teprve v počátcích. Protože cílem projektu je porozumění přenosu energie v rekonexi od makroskopických škál (globální škála erupce je zhruba 10000 km) směrem ke škálám na nichž dochází k vlastní disipaci a urychlování částic (řádově 10 m), je potřeba současně studovat procesy jak na velkých tak i malých měřítkách. Metoda konečných prvků s možností měnit hustotu sítě a řád bázových funkcí je proto velmi vhodná pro řešení této úlohy.

Literatura

[1] E. R. Priest: *Solar Magnetohydrodynamics*, D. Reidel Publishing Company, 1982

[2] T. J. Chung: *Computational Fluid Dynamics*, Cambridge University Press, 2002

[3] B. Kliem, M. Karlický, A.O. Benz: *Solar flare radio pulsations as a signature of dynamic magnetic reconnection*, Astronomy and Astrophysics, 360, 715-728

[4] M. Karlický, M. Bárta: *Drifting pulsating structures generated during tearing and coalescence processes in a flare current sheet*, Astronomy and Astrophysics, 464, 2, 735-740

[5] *Parallel Adaptive Mesh Refinement*,
`http://www.physics.drexel.edu/ olson/paramesh-doc/Users_manual/amr.html`
(cit. 2008.1.5)

Using triangular preconditioner updates in matrix-free implementations

J. Duintjer Tebbens, M. Tůma

Institute of Computer Science AS CR, Prague

1 Introduction

We consider sequences of linear systems of the form

$$A^{(i)}x = b^{(i)},\ i = 1, \ldots, \tag{1.1}$$

where $A^{(i)} \in R^{n\times n}$ are general nonsingular sparse matrices and $b^{(i)} \in R^n$ are corresponding right-hand sides. Such sequences arise, for example, when a system of nonlinear equations is solved by a Newton or Broyden-type method [11], [12]. Among the most successful approaches for solving the arising linear systems are Krylov subspace methods. They have the property that the system matrix is needed only in the form of matrix-vector products; in a *matrix-free* implementation of a Krylov subspace method the matrix is not represented explicitly. Krylov subspace methods must be preconditioned in order to be efficient and robust. However most of the strong preconditioners require the system matrix explicitly. To reduce the costs of the computation of preconditioners, we may reuse a preconditioner over several systems of the given sequence of systems of linear equations. In addition, the quality of the reused preconditioner may be enhanced through updates containing information extracted from the sequence of matrices, or from previous application of the Krylov subspace method. In this extended abstract we briefly describe the main idea of two techniques to solve a sequence of general nonsymmetric systems by preconditioned Krylov subspace methods, where the preconditioners are based on incomplete LU decompositions, they use triangular rank-n updates, and all the computations are done in matrix-free environment. The techniques are described in more detail in [7].

Due to the costs that are related to estimate the system matrix, avoiding frequent recomputations of the preconditioner from scratch seems to be even more important in matrix-free environment than if the matrices are given explicitly. Some new approaches to approximate preconditioner updates were introduced recently, see e.g. [13]. The authors in [1] introduced approximate diagonal updates to solve parabolic PDEs, see also [2]. Nonsymmetric updates of general incomplete LU decompositions were proposed in [8, 9], see also some results in solving CFD problems in [3]. So far, neither of these approaches has addressed the challenges related to updating in matrix-free environment.

This extended abstract deals with matrix-free algorithms to solve the sequences of linear systems based on the general triangular preconditioner updates introduced in [8]. The abstract is organized as follows. In Section 2 the general preconditioner updates are briefly recapped. In Section 3 the main idea of the two matrix-free approaches are described.

2 Triangular preconditioner updates

The triangular preconditioner updates for nonsymmetric sequences from [8] can be described as follows. Let A be the system matrix of a *reference* system and let A^+ be the current system

matrix. Assume LDU is an incomplete triangular decomposition of A and let $B = A - A^+$ be the difference matrix. Then a triangularly updated preconditioner for the current system is defined as

$$(LD - tril(B))\, U, \tag{2.1}$$

where *tril* denotes the lower triangular part. A preconditioner which updates the upper triangular part DU can be defined analogously, see [8]; we here concentrate on lower triangular updates. We assume that $(LD - tril(B))$ is nonsingular.

In matrix-free environment the factorization LDU has been obtained through estimating the reference matrix A, and it is stored explicitly. The update needs in addition the difference matrix B which is not given explicitly (only A has been estimated). The two strategies we will describe enable application of (2.1) without or with very cheap partial estimation of B.

3 Matrix-free updates

3.1 Partial matrix estimation

As the straightforward estimation of the difference matrix B may be expensive, one possible strategy which we propose is based on using enhanced *partial* and *approximate* matrix estimation. The classical matrix estimation problem is the problem of estimating a sparse matrix by a small number of well-chosen matrix-vector multiplications (matvecs). In [6] it was shown that all nonzero entries of a sparse matrix can be estimated, given the sparsity pattern, using a number of matvecs which is often much smaller than its dimension. Coleman and Moré [4] demonstrated the relation of estimating a matrix with a minimum number of matvecs to the coloring of a related graph G by a minimum number of colors. So-called direct methods for solving the matrix estimation problem for a matrix B use as G the intersection graph of B, that is the adjacency graph G_{B^TB} of B^TB. For an (undirected) adjacency graph G_C of a square and symmetric matrix C the set of vertices is defined as $V(G_C) = \{1, \ldots, n\}$ and its set of edges as $E(G_C) = \{\{i, j\} \mid c_{ij} \text{ is nonzero}\}$. A *coloring* of the intersection graph labels every vertex with a color such that no two adjacent vertices have the same color. Then the number of groups of vertices of the related graph with the same color corresponds to the number of matvecs needed to estimate the entries of the matrix. If we need to estimate only a part of a given matrix, we speak about a *partial matrix estimation problem* [10], [5].

To use the triangular updates described above we only have to estimate, in addition to A which was estimated earlier, the lower triangular part of A^+. This leads to a particular partial matrix estimation problem. We will formulate this problem as a graph coloring problem for a graph which is different from the intersection graph of A^+. The following theorem describes this graph.

Theorem 3.1 *Consider the graph*

$$G_T(B) = G_U(L_B) \cup G_K,$$

where $G_U(L_B) = (V_U, E_U)$ is the intersection graph of the lower triangular part of the matrix B and G_K is defined as

$$G_K = \cup_{i=1}^{n} G_i, \qquad G_i = (V_i, E_i) = (V_U, \{\{k, j\} |\ b_{ik} \neq 0 \wedge b_{ij} \neq 0 \wedge k \leq i < j\}).$$

If the graph $G_T(B)$ can be colored by p colors, then the entries of the lower triangular part L_B of B can be computed by p matvecs of B with vectors $d_1, \ldots d_p$ such that for each nonzero entry l_{ij} of L_B there is a vector $d_k, 1 \leq k \leq p$, satisfying $(Bd_k)_i = l_{ij}(d_k)_j$.

P r o o f : See [7, Section 3].

Note that the graph $G_T(B)$ contains only a subset of edges of the adjacency graph $G(B^TB)$. Consequently, in order to estimate only a triangular part of A^+ we may need a smaller number of matvecs than in the case of estimation of the whole B. In combination with a sparsification strategy for the nonzero entries of $tril(A^+)$, the estimation of $tril(A^+)$ needed in (2.1) is considerably less expensive than the estimation of A^+. For examples demonstrating the gain in computational costs, see [7, Section 5].

3.2 Mixed implicit/explicit forward solves

Another strategy to use the triangular updates in matrix-free environment is beneficial only when function components are *separable*. Let us explain what we mean here by separability. Consider a matrix-free implementation of a Krylov subspace method where the product of the system matrix A with a vector v is replaced by the value of a function $\mathcal{F}$ evaluated at v. We say that $\mathcal{F}$ is separable if the evaluation of $\mathcal{F}$ can be separated in the evaluation of its function components with low costs. That is, if the components of the function $\mathcal{F} : R^n \to R^n$ can be written as $\mathcal{F}_i : R^n \to R$, where $e_i^T\mathcal{F}(v) = \mathcal{F}_i(v)$, and computing $\mathcal{F}_i(v)$ costs about one n-th of the full function evaluation $\mathcal{F}(v)$.

Every application of (2.1) requires a forward solve with $LD - tril(B)$ and a backward solve with U, which is trivial as U is stored explicitly. With separable function components we propose the following *mixed explicit-implicit* strategy for the forward solve: Split the lower triangular matrix $LD - tril(B)$ as $LD - tril(B) = E + tril(A^+)$, i.e. $E \equiv LD - tril(A)$ is stored explicitly and the implicit part $tril(A^+)$ contains entries of the new system matrix. Let the function $\mathcal{F}^+$ represent A^+ implicitly. We have to solve triangular systems of the form

$$\left(E + tril(A^+)\right) z = y,$$

which yields the forward solve loop

$$z_i = \frac{y_i - \sum_{j<i} e_{ij}z_j - \sum_{j<i} a^+_{ij}z_j}{e_{ii} + a^+_{ii}}, \qquad i = 1, 2, \ldots, n. \tag{3.1}$$

Note that the values e_{ii} are known explicitly. The values a^+_{ii} can be obtained with the n function component evaluations

$$a^+_{ii} = \mathcal{F}^+_i(e_i), \qquad 1 \le i \le n.$$

In the numerator of (3.1), the first sum can be computed explicitly and the second sum can be computed by the function evaluation

$$\sum_{j<i} a^+_{ij}z_j = \mathcal{F}^+_i\left((z_1, \ldots, z_{i-1}, 0, \ldots, 0)^T\right). \tag{3.2}$$

With this technique one avoids estimation and storage of A^+ (except for its main diagonal). The cost of every forward solve is that of a forward solve with LD plus about one full function evaluation.

Acknowledgements: The work of the authors has been supported by the Institutional Research Plan AV0Z10300504 "Computer Science for the Information Society: Models, Algorithms, Applications". The work of the first author was also supported by project number KJB100300703 of the Grant Agency of the Academy of Sciences of the Czech Republic. The work of the second author was also supported by project number IAA100300802 of the Grant Agency of the Academy of Sciences of the Czech Republic.

References

[1] M. Benzi and D. Bertaccini. Approximate inverse preconditioning for shifted linear systems. *BIT*, 43(2):231–244, 2003.

[2] D. Bertaccini. Efficient preconditioning for sequences of parametric complex symmetric linear systems. *Electronic Transactions on Numerical Mathematics*, 18:49–64, 2004.

[3] P. Birken, J. Duintjer Tebbens, A. Meister, and M. Tůma. Preconditioner updates applied to CFD model problems. *Appl. Num. Math.*, 58(11):1628–1641, 2008.

[4] T. F. Coleman and J. J. Moré. Estimation of sparse Jacobian matrices and graph coloring problems. *SIAM J. Numer. Anal.*, 20:187–209, 1983.

[5] J. Cullum and M. Tůma. Matrix-free preconditioning using partial matrix estimation. *BIT Numer. Math.*, 46:711–729, 2006.

[6] A. R. Curtis, M. J. D. Powell, and J. K. Reid. On the estimation of sparse Jacobian matrices. *J. Inst. Maths. Applics.*, 13:117–119, 1974.

[7] J. Duintjer Tebbens and M. Tůma. Preconditioner updates for solving sequences of nonsymmetric linear systems in matrix-free environment. Preprint submitted in 2008.

[8] J. Duintjer Tebbens and M. Tůma. Efficient preconditioning of sequences of nonsymmetric linear systems. *SIAM J. Sci. Comput.*, 29(5):1918–1941, 2007.

[9] J. Duintjer Tebbens and M. Tůma. Improving triangular preconditioner updates for nonsymmetric linear systems. *LNCS*, 4818:737–744, 2008.

[10] A. H. Gebremedhin, F. Manne, and A. Pothen. What color is your Jacobian? Graph coloring for computing derivatives. *SIAM Review*, 47:629–705, 2005.

[11] C. T. Kelley. *Iterative Methods for Linear and Nonlinear Equations*. SIAM, Philadelphia, 1995.

[12] C. T. Kelley. *Solving nonlinear equations with Newton's method*. Fundamentals of Algorithms. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2003.

[13] G. Meurant. On the incomplete Cholesky decomposition of a class of perturbed matrices. *SIAM J. Sci. Comput.*, 23(2):419–429 (electronic), 2001. Copper Mountain Conference (2000).

Survey of Discrete Maximum Principles for Higher-Order Finite Elements

T. Vejchodský

Institute of Mathematics AS CR, Prague

1 Introduction

Many second order (both linear and nonlinear) elliptic and parabolic problems satisfy the maximum principle. Besides the theoretical importance, the maximum principle mirrors the natural property of the modelled physics. For example, in the heat conduction problem, the maximum principle guarantees the nonnegativity of the obtained temperature. Similarly, the maximum principle is important for modeling of the other naturally nonnegative quantities, like concentration, density, etc.

The natural question if the maximum principle is satisfied after the discretization by a suitable method has been studied for several decades. The first result (up to the author's knowledge) about the discrete maximum principle (DMP) for linear elliptic problems was published by Varga [3] in 1966. Since that time many generalizations to different problems and methods appeared. Majority of these results concern with the lowest-order finite difference and finite element methods and the results are based on the special properties of the system matrices (theory of M-matrices, cf. [2]).

Surprisingly, much less attention was paid to the DMPs for higher order approximations [1, 8]. Let us emphasize the negative result of Höhn and Mittelman [1] which shows that a strong version of the DMP is satisfied in 2D for quadratic and cubic finite elements under unrealistic assumptions on the triangulation only. The standard version of the DMP was proved recently [4, 5, 6] for 1D diffusion problems discretized by the *hp*-version of the finite element method (*hp*-FEM). Realistic conditions for the validity of the DMP in higher dimension are still unknown.

In the talk, we present the DMP result for the 1D Poisson equation [4] discretized by higher-order finite elements. We also mention generalizations to the mixed boundary conditions and to the case with piecewise constant coefficients [5, 6]. The main emphasis will be put on the generalization of the DMP result from the Poisson problem to the diffusion-reaction problem.

2 Diffusion-Reaction Problem and its Discretization

In particular, we will consider 1D diffusion-reaction problem

$$-(au')' + \kappa^2 u = f \quad \text{in } \Omega = (a_\Omega, b_\Omega), \qquad u(a_\Omega) = u(b_\Omega) = 0, \tag{2.1}$$

where the coefficients a and κ are assumed to be piecewise constant. This problem is discretized by the higher-order finite element method, where various polynomial degrees on different elements are allowed (*hp*-FEM). Hence, we consider a partition $\mathcal{T}_{hp}$ of the interval Ω into a finite number of elements and a polynomial degree p_K for each element $K \in \mathcal{T}_{hp}$. This defines the finite element space

$$V_{hp} = \{v_h \in H_0^1(\Omega) : v_h|_K \in P^{p_K}(K) \text{ for all } K \in \mathcal{T}_{hp}\},$$

where $H_0^1(\Omega) = \left\{v \in L^2(\Omega) : v' \in L^2(\Omega),\ v = 0 \text{ on } \partial\Omega\right\}$ denotes the Sobolev space and $P^{p_K}(K)$ stands of the space of polynomials of degree at most p_K in the interval K. The hp-FEM solution is then define as $u_{hp} \in V_{hp}$ such that

$$\int_\Omega \left(a u'_{hp} v'_{hp} + \kappa^2 u_{hp} v_{hp}\right) \,\mathrm{d}x = \int_\Omega f v_{hp} \,\mathrm{d}x \quad \forall v_{hp} \in V_{hp}. \tag{2.2}$$

3 Discrete Maximum Principle

The original problem (2.1) satisfies the well-known maximum principle. However, since we consider homogeneous Dirichlet boundary conditions, the standard maximum principle for problem (2.1) is equivalent to the conservation of nonnegativity

$$f \geq 0 \text{ a.e. in } \Omega \quad \Rightarrow \quad u \geq 0 \text{ a.e. in } \Omega.$$

However, if we replace u by u_{hp} then it is not difficult to find counterexamples violating this implication.

On the other hand, it is possible to characterize a suitable class of finite element meshes (and consequently a class of finite element spaces V_{hp}) such that implication

$$f \geq 0 \text{ a.e. in } \Omega \quad \Rightarrow \quad u_{hp} \geq 0 \text{ in } \Omega \tag{3.1}$$

holds true for all $f \in L^2(\Omega)$ and for $u_{hp} \in V_{hp}$ given by (2.2). Thus, if implication (3.1) is satisfied for a fixed mesh $\mathcal{T}_{hp}$ (consequently for a fixed space V_{hp}) then we say that the discretization (2.2) satisfies the discrete maximum principle (DMP).

The discrete Green's function (DGF) has proved to be a very usefull tool for investigation of the DMP for higher-order finite element methods. For $y \in \Omega$, the DGF $G_{hp,y} \in V_{hp}$ is defined as the unique solution of the problem

$$\int_\Omega \left(a w'_{hp} G'_{hp,y} + \kappa^2 w_{hp} G_{hp,y}\right) \,\mathrm{d}x = w_{hp}(y) \quad \forall w_{hp} \in V_{hp}.$$

We denote $G_{hp}(x,y) = G_{hp,y}(x)$. The following properties are important and easy to prove:

(i) $u_{hp}(y) = \displaystyle\int_\Omega G_{hp}(x,y) f(x) \,\mathrm{d}x$

(ii) If $\varphi_1, \varphi_2, \ldots, \varphi_N$ is a basis of V_{hp} and if $\mathbb{A} \in \mathbb{R}^{N\times N}$ is the stiffness matrix with entries $\mathbb{A}_{ij} = \int_\Omega (a\varphi'_i\varphi'_j + \kappa^2\varphi_i\varphi_j) \,\mathrm{d}x$, $i,j = 1,2,\ldots,N$, then

$$G_{hp}(x,y) = \sum_{i=1}^N \sum_{j=1}^N (\mathbb{A}^{-1})_{ij} \varphi_i(x) \varphi_j(y).$$

Property (i) immediately implies that the DMP (3.1) is satisfied if and only if $G_{hp}(x,y) \geq 0$ for all $(x,y) \in \Omega^2$. The nonnegativity of G_{hp} in Ω^2 can be investigated directly using the explicit formula from property (ii).

3.1 DMP for Poisson Problem

For Poisson problem, i.e., in case $a = 1$ and $\kappa = 0$, there exist an explicit formula for the entries of the inverse of the stiffness matrix $(\mathbb{A}^{-1})_{ij}$ for linear finite elements. This formula together

with other special properties of the 1D Poisson problem enables to localize the investigation of the nonnegativity of G_{hp} on the reference element $\widehat{K} = [-1, 1]$. Hence, for given polynomial degree p_K of an element $K \in \mathcal{T}_{hp}$ we construct a reference DGF $\widehat{G}_{hp}(\xi, \eta)$ for $(\xi, \eta) \in \widehat{K}^2$. The reference DGF depends also on the position of K in Ω and on the size of K. The special structure of $\widehat{G}_{hp}(\xi, \eta)$ enables to guarantee its nonnegativity independently on the position of K in Ω provided certain polynomial of degree $p_K - 1$ in both ξ and η is nonnegative for $(\xi, \eta) \in \widehat{K}^2$. Nonnegativity of this polynomial can be investigated analytically for $p_K \leq 3$ and numerically for higher polynomial degrees, see Section 4 below. Afterall, we show that discretization (2.2) satisfies the DMP if $H^K_{\mathrm{rel}} = h_K/(b_\Omega - a_\Omega) \leq 9/10$ for all $K \in \mathcal{T}_{hp}$, where h_K stands for the length of K. The detailed analysis can be found in [4].

3.2 DMP for Diffusion-Reaction Problem

The special properties of the Poisson problem, however, are not available for the diffusion-reaction problem. Therefore, we use a concept based on the discrete minimum energy extensions ψ_i of the standard (Courante) lowest-order basis functions φ_i with respect to all remaining higher-order basis functions. This easily enables to infer the following conditions which guarantee the DMP:
(a) the discrete minimum energy extensions ψ_i are nonnegative in Ω,
(b) the off-diagonal entries of the stiffness matrix assembled from ψ_i are nonpositive,
(c) the DGF restricted to the square K^2 is nonnegative for all elements $K \in \mathcal{T}_{hp}$.

Verification of these conditions is, unfortunately, quite demanding, but feasible for the 1D diffusion-reaction problem. The analysis of condition (c) for given polynomial degree p_K of K relies on the nonnegativity of certain polynomial. Nonnegativity of this polynomial was verified for elements of degree up to 10 using the technique of interval arithmetic, see Section 4 below. The following theorem, see [7], shows the weakest and simplest condition we obtained.

Theorem. *Let $\mathcal{T}_{hp}$ be a finite element mesh in an interval $\Omega = (a_\Omega, b_\Omega)$. Let the polynomial degrees p_K of the elements $K \in \mathcal{T}_{hp}$ do not exceed 10. Denote by h_K and $H^K_{\mathrm{rel}} = h_K/(b_\Omega - a_\Omega)$ the length and the relative length of the element $K \in \mathcal{T}_{hp}$ and by κ^2_K and a_K the constant values of the coefficients κ^2 and a on the element K. If*

$$\frac{\kappa^2_K h^2_K/a_K}{\kappa^2_K h^2_K/a_K + \gamma^3} \leq H^K_{\mathrm{rel}} \leq 1/3 \quad \text{for all } K \in \mathcal{T}_{hp},$$

where $\gamma^3 \approx 5.608797$, then the discretization (2.2) satisfies the DMP.

Let us remark that the value γ^3 comes from the analysis of the cubic elements and leads to the most strict condition for all the considered polynomial degrees. Further, we remark that our computations indicate the validity the above theorem for arbitrary distribution of polynomial degrees. Practically, however, we checked it for polynomials of degree at most 10.

4 Nonnegativity of Multivariate Polynomials

The DMP results for both Poisson and diffusion-reaction problems are based on verification of nonnegativity of certain multivariate polynomials on a rectangular domain. The rectangular domain can be easily transformed to the entire Euclidean space. Clearly, a polynomial is

nonnegative if it can be written as a sum of squares of another polynomials. Unfortunately, there exists nonnegative polynomials which cannot be written as a sum of squares of another polynomials. An example is the Motzkin form $f(x, y, z) = z^6 + x^4y^2 + x^2y^4 - 3x^2y^2z^2$.

The 17th of the famous 24 Hilbert's problems is to prove that any nonnegative polynomial can be written as a sum of squares of rational functions. This was proved in 1927 by Emil Artin. There exist (NP-hard) algorithms for finding these sums of squares. However, these algorithms are complicated and difficult to use.

An interesting and easy to implement approach is the usage of interval arithmetic. In the interval arithmetic the arithmetic operations are defined on intervals. If I and J are two intervals and if $*$ is an arithmetic operation then the interval $R = I * J$ is guaranteed to contain all possible results $\{r = a * b$, where $a \in I,\ b \in J\}$.

The idea how to verify nonnegativity of a function f on an interval I is to use the interval arithmetic and compute an interval $R = f(I)$ containing all possible outputs of a function f on an interval I. If R is nonnegative (contains nonnegative numbers only) then nonnegativity of f in I is verified. If not, we split I into two (or more) subintervals and repeat the process for all these subintervals. If this algorithm terminates after a finite number of steps, the nonnegativity of f in I is verified.

Acknowledgement: The support of Grant No. 102/07/0496 of the Czech Science Foundation, Grant No. IAA100760702 of the Grant Agency of the Academy of Sciences of the Czech Republic, and institutional research plan No. AV0Z10190503 of the Academy of Sciences of the Czech Republic is gratefully acknowledged.

References

[1] W. Höhn, H.-D. Mittelmann: *Some remarks on the discrete maximum-principle for finite elements of higher order*, Computing 27 (1981) 145–154.

[2] R.S. Varga: *Matrix iterative analysis*, Prentice-Hall, Englewood Cliffs, 1962.

[3] R.S. Varga: *On discrete maximum principle*, J. SIAM Numer. Anal. 3 (1966) 355–359.

[4] T. Vejchodský, P. Šolín: *Discrete Maximum Principle for Higher-Order Finite Elements in 1D*, Math. Comp. 76 (2007) 1833–1846.

[5] T. Vejchodský, P. Šolín: *Discrete Maximum Principle for Poisson Equation with Mixed Boundary Conditions Solved by hp-FEM*, submitted to Advances in Applied Mathematics and Mechanics, 2008.

[6] T. Vejchodský, P. Šolín: *Discrete Maximum Principle for a 1D Problem with Piecewise-Constant Coefficients Solved by hp-FEM*, J. Numer. Math., 15 (2007) 233–243.

[7] T. Vejchodský: *Higher-order discrete maximum principle for 1D diffusion-reaction problems*, submitted to Appl. Numer. Math., 2008.

[8] E.G. Yanik: *Sufficient conditions for a discrete maximum principle for high-order collocation methods*, Comput. Math. Appl. 17 (1989), 1431–1434.

Parallel MatSol library for solution of contact problems and contact shape optimization problems

V. Vondrák, T. Kozubek, A. Markopoulos, T. Brzobohatý

VSB-Technical University of Ostrava

1 Introduction

During last several years, our research team in the Dept. of Applied Mathematics, VŠB-Technical University of Ostrava has been focused to development of scalable algorithms for contact problems and contact shape optimization problems. These algorithms are based on FETI domain decomposition methods which are well known by its parallel and numerical scalability. Our algorithms were originally implemented in C++ library called OOSol (Object Oriented SOLvers) [2]. This library benefits from their modularity and extensibility, nevertheless the slow development of the code caused by unavailability of advanced auxiliary algorithms necessary for debugging of complex algorithms, shows to be its biggest disadvantage. Therefore we started to implement simultaneously all these algorithms into new library that is developed in Mathworks Matlab environment [3] which is equipped with many of these helpful functions. We call this library MatSol (MATlab SOLvers) [1]. Several years ago the Mathworks company introduced Matlab Distributed Computing Engine which allowed to run Matlab functions also on parallel computers. Hence, the MatSol got full functionality of OOSol library including parallel algorithms and recently represents our primary testing and developing library. In our paper, we would like to present functionality of the MatSol library to the solution of realistic contact problems with millions of degrees of freedom showing parallel and numerical scalability of the implemented Total FETI method. The interface between MatSol and ANSYS or COMSOL [4, 5] will be presented as well. This feature allows very simply plug the MatSol library into the commercial finite element packages. Some comparisons of the commercial solvers and MatSol will be shown. The efficiency of the solving algorithms will be also presented on such complex problem as contact shape optimization problems.

2 Structure of MatSol library

In 2007, authors of the paper established development of a new library MatSol for domain decomposition based solution of problems in mechanics. Today structure of the library with the typical solution process flow is described in Figure 1. The solution process starts from the model which is either already in model database or it is converted to the model database from standard commercial and non-commercial preprocessors like ANSA, ANSYS, COMSOL, PMD [4, 5, 6] etc. The list of preprocessing tools is not limited and any of new one can be simply plugged into the library creating proper database convertor. Preprocessing part continues in MatSol depending on solved problem. User can solve deterministic or stochastic problems, static or transient analysis, optimization problems, problems in linear and non-linear elasticity, contact problems. For assemble of numerical model we are using finite or boundary element methods. As the domain decomposition techniques the FETI or BETI methods are implemented. The solution process could be run either in sequential or parallel mode. The solution algorithms are implemented in such a way, that the code is the same for both sequential and parallel mode.

The parallel mode is run using Matlab Parallel Computing Server and Parallel Computing Toolbox. MatSol library includes also tools for postprocessing of results and advanced tools for postplotting of the problems. The results of the problem are then through the model database converted to the modelling tools for further postprocessing.

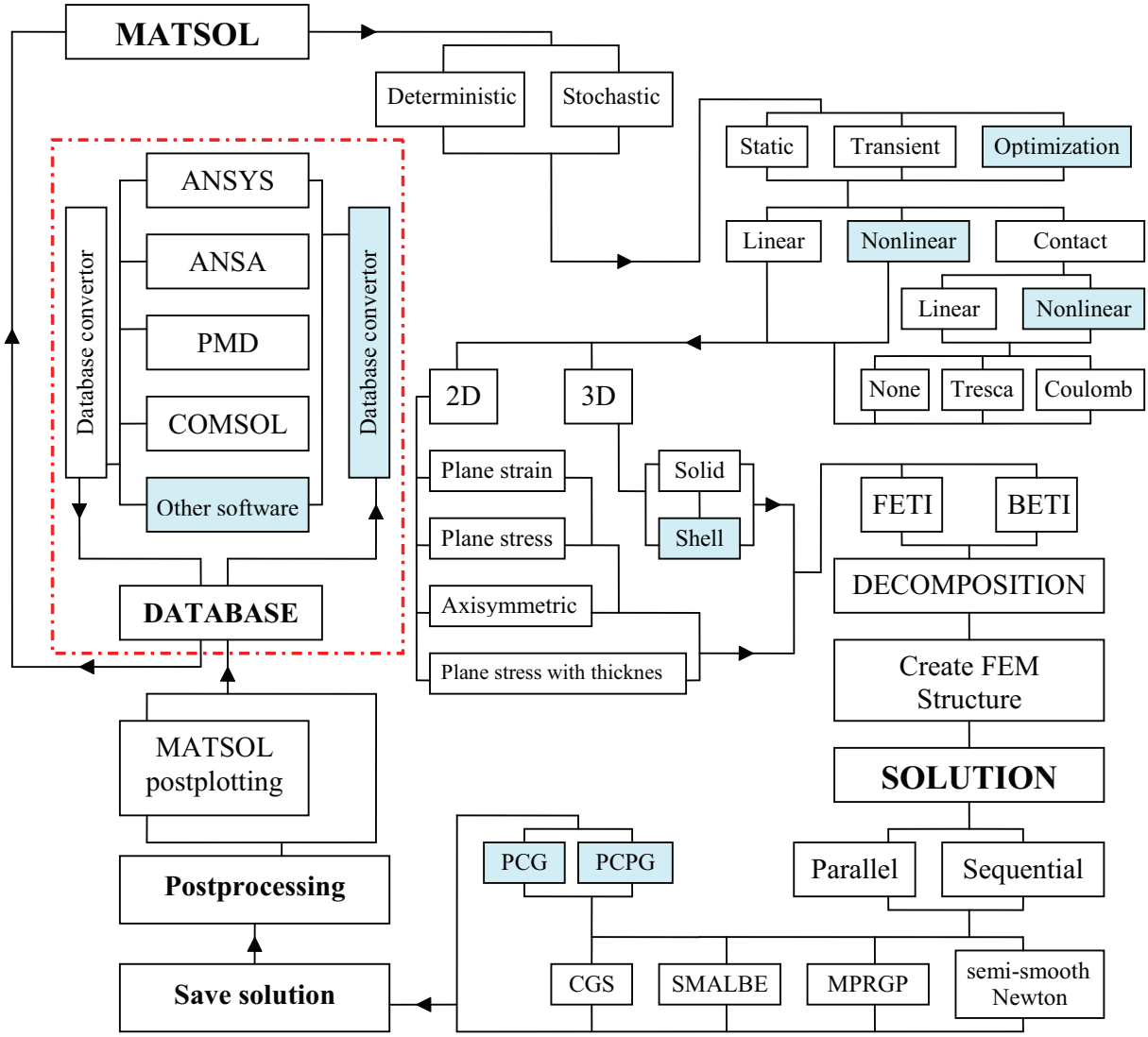


Figure 1: Structure of MatSol library

Described structure of MatSol library allows to override standard solvers in commercial and non-commercial finite element packages and substitute them by these ones which are implemented in MatSol. This gives a very useful alternative to users of commercial packages and great tool for algorithm developers to test the new algorithms on the realistic problems.

3 Solved problems

In this section, we shall present typical problems solved using MatSol library and efficiency of implemented algorithms. All problems were solved on the computational cluster HP BLc7000 with 9 nodes. Each node is equipped with 2 dual core AMD Opteron processors and 8GB RAM and interconnected by infiniband network. On this cluster we have installed 24 licences

of Matlab parallel computing server.

The first example we would like to present is a part of gear box as depicted in Figure 2. It is typical mechanical engineering application in the linear elasticity. The finite element model was discretized by 0.5 mil. nodes, i.e. 1.5 mil. degrees of freedom (DOFs). The model was decomposed into 10, 20, 40 and 80 subdomains and performance of solving algorithms you can see in Figure 5.

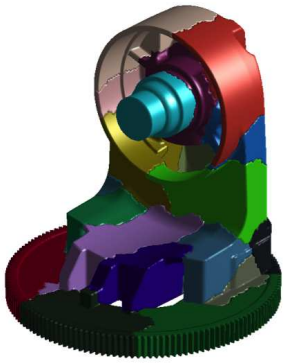


Figure 2: Part of a gear box

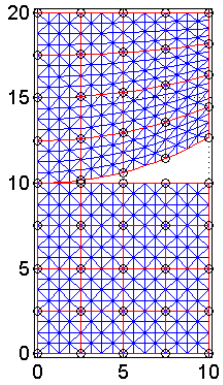


Figure 3: 2D Hertz problem

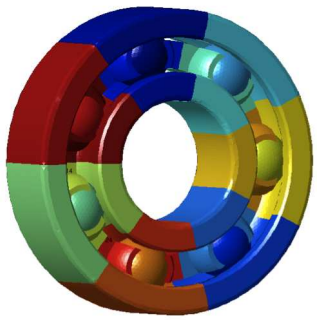


Figure 4: Ball bearing

The second problem is classical benchmark of contact mechanics. It is 2D Hertz problem with floating upper body, see Figure 3. The model was decomposed into $2^k, k = 1, 2, ..., 9$ subdomains, each discretized by 100×100 nodes. The largest problem solved 10,240.000 unknown DOFs! Summarized number of iterations and solution times are collected in Table 1. More realistic contact problem is depicted in Figure 4. The ball bearing is assembled from 10 totally independent and free parts in mutual contact. We have solved 2 discretization models. First one with 300 thousands DOFs decomposed into 28 subdomains and second one with 1.5 millions DOFs decomposed into 63 subdomains. Solution time was 2380s for the smaller problem using sequential code, resp. 339s in case of parallel code. The solution of the larger model needed 6.5 hours in case of parallel code. Unfortunately the sequential code we couldn't use because needed computer with at least 48GB of RAM.

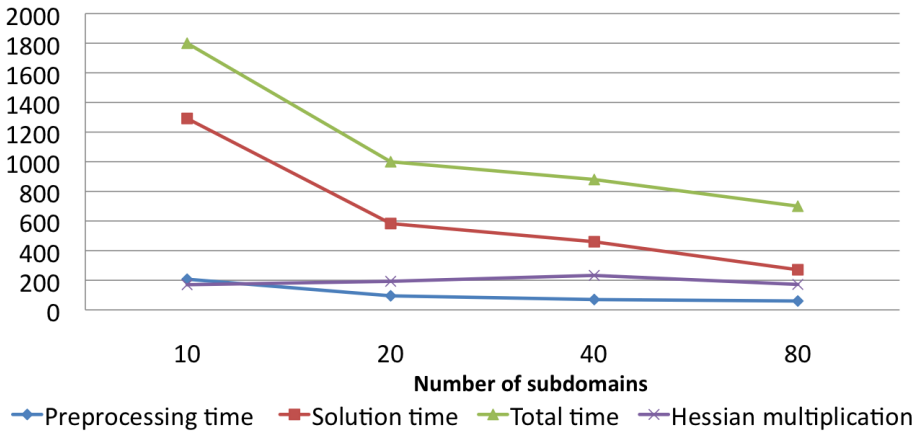


Figure 5: Scalability of Total FETI MatSol solver - a gear box problem

We would like to demonstrate MatSol contact shape optimization capabilities on Hertz problem which is depicted in Figure 6. The shape of the bottom body was parameterized with 16 design variables. The compliance was used as the shape optimization objective function with

Table 1: Performance of MatSol parallel library - 2D Hertz problem

#Subdomains	2	4	8	16	32	64	128	256	512
Primal variables	40k	80k	160k	320k	640k	1280k	2560k	5120k	10240k
Dual variables	600	1200	2400	5200	11200	23200	48000	97600	198400
Hessian multiplications	45	65	52	60	88	91	127	109	134
CG steps	28	42	38	46	41	33	23	28	44
Preprocessing time (s)	6	6	6	6	12	18	40	119	149
Solver time (s)	3	4	10	18	34	45	117	223	458
Total Time (s)	10	12	25	40	78	130	290	660	1300

constraints on feasible design. Optimized design was obtained after 120 design iterations and the parallel solution of the one design step was six times faster than the standard sequential code. Comparison of the initial and optimized stress distribution is in Figures 6 and 7.

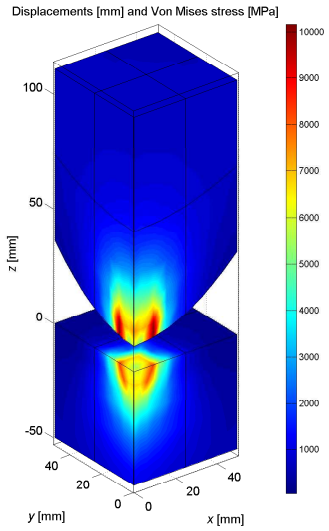


Figure 6: Initial design stress distribution

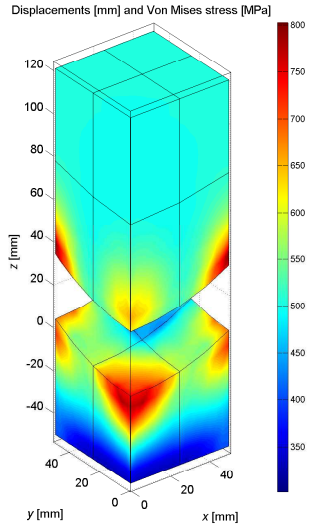


Figure 7: Optimized design stress distribution

Acknowledgement: This work has been supported by the grant GAČR 101/08/0574 and by project of Ministry of Education of the Czech Republic MSM6198910027 [7].

References

[1] MatSol library. *http://www.am.vsb.cz/matsol*.

[2] OOSol library. *http://www.am.vsb.cz/oosol*.

[3] MATLAB - The Language of Technical Computing. *http://www.mathworks.com*

[4] ANSYS. *http://www.ansys.com*.

[5] COMSOL Muliphysics. *http://www.comsol.com*.

[6] PMD Manuals. *http://www.it.cas.cz/manual/pmd*.

[7] COMSIO - Computationally Intensive Simulations and Optimizations. *http://comsio.vsb.cz*.

Winter school lectures

V. Dolejší, M. Feistauer:
Discontinuos Galerkin methods and applications to compressible flow

Z. Dostál:
Duality for QP problems with semidefinite Hessian and contact problems

J. Haslinger:
Structural optimization

J. Chleboun:
What is the role of the worst scenario method in solving problems with uncertain input data ?

J. Kruis:
Uncertainty in engineering problems described by fuzzy sets

T. Kozubek:
A numerical solution of elliptic boundary value problems with uncertain data and geometry

D. Novák, M. Vořechovský:
Small-sample simulační metody typu Monte Carlo

M. Rozložník:
Numerical stability of symmetric indefinite solvers: direct methods

S. Ratschan:
Interval computation: Why ? When ? How ?

IT for Innovations

Discontinuous Galerkin Methods and Applications to Compressible Flow

Part 2. DGFEM for Evolution Convection-Diffusion Problems

*

Miloslav Feistauer
Charles University Prague
Faculty of Mathematics and Physics

In cooperation with V. Dolejší, V. Kučera, V. Sobotíková and K. Švadlenka

*Presented at Winter School on Numerical Analysis SNA'09, 2.2. - 6.2. 2009, Ostrava

Goal: to work out a sufficiently accurate, robust and theoretically based method for the numerical solution of compressible flow with a wide range of Mach numbers and Reynolds numbers

Difficulties:
nonlinear convection dominating over diffusion $\Longrightarrow$
– boundary layers, wakes for large Reynolds numbers
– shock waves, contact discontinuities for large Mach numbers
– instabilities caused by acoustic effects for low Mach numbers

One of promising, efficient methods for the solution of compressible flow is the discontinuous Galerkin finite element method (DGFEM) using piecewise polynomial approximation of a sought solution without any requirement on the continuity between neighbouring elements.

In this paper we shall be concerned with the analysis of the DGFEM for the solution of a nonlinear nonstationary convection-diffusion equation, which is a simple prototype of the compressible Navier-Stokes system.

Continuous model problem

Let us consider the problem to find $u : Q_T = \Omega \times (0,T) \to \mathbb{R}$ such that

$$\text{a)}\ \frac{\partial u}{\partial t} + \sum_{s=1}^{d} \frac{\partial f_s(u)}{\partial x_s} = \varepsilon \Delta u + g \quad \text{in } Q_T, \tag{1}$$

$$\text{b)}\ u|_{\Gamma_D \times (0,T)} = u_D, \quad \text{c)}\ \varepsilon \frac{\partial u}{\partial n}|_{\Gamma_N \times (0,T)} = g_N,$$

$$\text{d)}\ u(x,0) = u^0(x),\ x \in \Omega.$$

We assume that $\Omega \subset \mathbb{R}^d,\ d = 2,3$, is a bounded polygonal (if $d = 2$) or polyhedral (if $d = 3$) domain with Lipschitz-continuous boundary $\partial\Omega = \Gamma_D \cup \Gamma_N$, $\Gamma_D \cap \Gamma_N = \emptyset$ and $T > 0$. The diffusion coefficient $\varepsilon > 0$ is a given constant, $g : Q_T \to \mathbb{R}$, $u_D : \Gamma_D \times (0,T) \to \mathbb{R}$, $g_N : \Gamma_N \times (0,T) \to \mathbb{R}$, and $u^0 : \Omega \to \mathbb{R}$ are given functions, $f_s \in C^1(\mathbb{R})$, $s = 1,\dots,d$, are prescribed fluxes.

DG space semidiscretization

Let $\mathcal{T}_h$ ($h > 0$) be a *partition* of the closure $\overline{\Omega}$ of the domain Ω into a finite number of closed triangles ($d = 2$) or tetrahedra ($d = 3$) K with mutually disjoint interiors such that

$$\overline{\Omega} = \bigcup_{K \in \mathcal{T}_h} K. \tag{2}$$

We call $\mathcal{T}_h$ a *triangulation* of Ω and do not require the standard conforming properties from the finite element method.

$h_K =$ diam (K), $\quad h =$ max$_{K \in \mathcal{T}_h} h_K$, $\quad \rho_K =$ largest ball inscribed into K

Let $K, K' \in \mathcal{T}_h$. We say that K and K' are *neighbours*, if the set $\partial K \cap \partial K'$ has positive $(d-1)$-dimensional measure. We say that $\Gamma \subset K$ is a *face* of K, if it is a maximal connected open subset either of $\partial K \cap \partial K'$, where K' is a neighbour of K, or of $\partial K \cap \partial\Omega$.

$\mathcal{F}_h$= the system of all faces of all elements $K \in \mathcal{T}_h$,
the set of all innner faces:

$$\mathcal{F}_h^I = \{\Gamma \in \mathcal{F}_h;\ \Gamma \subset \Omega\}, \tag{3}$$

the set of all "Dirichlet" boundary faces:

$$\mathcal{F}_h^D = \{\Gamma \in \mathcal{F}_h;\ \Gamma \subset \partial\Omega_D\}, \tag{4}$$

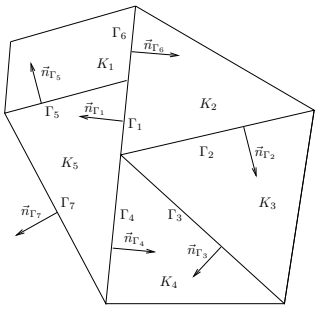
the set of all "Neumann" boundary faces:

$$\mathcal{F}_h^N = \{\Gamma \in \mathcal{F}_h,\ \Gamma \subset \partial\Omega_N\}. \tag{5}$$

Obviously, $\mathcal{F}_h = \mathcal{F}_h^I \cup \mathcal{F}_h^D \cup \mathcal{F}_h^N$. For a shorter notation we put

$$\mathcal{F}_h^{ID} = \mathcal{F}_h^I \cup \mathcal{F}_h^D, \qquad \mathcal{F}_h^{DN} = \mathcal{F}_h^D \cup \mathcal{F}_h^N. \tag{6}$$

For each $\Gamma \in \mathcal{F}_h$ we define a *unit normal vector* n_Γ. We assume that for $\Gamma \in \mathcal{F}_h^{DN}$ the normal n_Γ has the same orientation as the outer normal to $\partial\Omega$. For each face $\Gamma \in \mathcal{F}_h^I$ the orientation of n_Γ is arbitrary but fixed. See Figure .



Elements with hanging nodes

$d(\Gamma)$ = diameter of $\Gamma \in \mathcal{F}_h$.

Broken Sobolev spaces

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

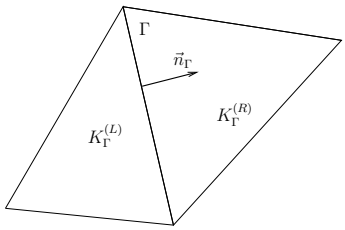
$$H^k(\Omega, \mathcal{T}_h) = \{v; v|_K \in H^k(K) \ \forall K \in \mathcal{T}_h\} \tag{7}$$

with the norm

$$\|v\|_{H^k(\Omega,\mathcal{T}_h)} = \left(\sum_{K\in\mathcal{T}_h} \|v\|^2_{H^k(K)} \right)^{1/2} \tag{8}$$

and the seminorm

$$|v|_{H^k(\Omega,\mathcal{T}_h)} = \left(\sum_{K\in\mathcal{T}_h} |v|^2_{H^k(K)} \right)^{1/2}. \tag{9}$$



Neighbouring elements

For each $\Gamma \in \mathcal{F}_h^I$ there exist two neighbouring elements $K_\Gamma^{(L)}, K_\Gamma^{(R)} \in \mathcal{T}_h$ such that $\Gamma \subset \partial K_\Gamma^p \cap \partial K_\Gamma^n$. We use a convention that $K_\Gamma^{(R)}$ lies in the direction of n_Γ and $K_\Gamma^{(L)}$ lies in the opposite direction to n_Γ, see Figure . ($K_\Gamma^{(L)}$, $K_\Gamma^{(R)}$ are *neighbours*.)

For $v \in H^1(\Omega, \mathcal{T}_h)$ and $\Gamma \in \mathcal{F}_h^I$, we introduce the following notation:

$$\begin{aligned} v|_\Gamma^{(L)} &= \text{the trace of } v|_{K_\Gamma^{(L)}} \text{ on } \Gamma, \\ v|_\Gamma^{(R)} &= \text{the trace of } v|_{K_\Gamma^{(R)}} \text{ on } \Gamma, \\ \langle v \rangle_\Gamma &= \frac{1}{2}\left(v|_\Gamma^{(L)} + v|_\Gamma^{(R)} \right), \\ [v]_\Gamma &= v|_\Gamma^{(L)} - v|_\Gamma^{(R)}. \end{aligned} \tag{10}$$

The value $[v]_\Gamma$ depends on the orientation of n_Γ, but the value $[v]_\Gamma n_\Gamma$ is independent of this orientation.

For $\Gamma \in \mathcal{F}_h^{DN}$ there exists element $K_\Gamma^{(L)} \in \mathcal{T}_h$ such that $\Gamma \subset K_p \cap \partial\Omega$. For $v \in H^1(\Omega, \mathcal{T}_h)$, we set

$$v|_\Gamma^{(L)} = \text{the trace of } v|_{K_\Gamma^{(L)}} \text{ on } \Gamma, \tag{11}$$

For $\Gamma \in \mathcal{F}_h^{DN}$ by $v|_\Gamma^{(R)}$ we formally denote the exterior trace of v on Γ given either by a Dirichlet boundary condition or by an extrapolation from the interior of Ω.

The approximate solution – sought in the space of discontinuous piecewise polynomial functions

$$S_h = S_h^{p,-1} = \{v; v|_K \in P^p(K) \ \forall K \in \mathcal{T}_h\},$$

$p > 0$ – integer, $P^p(K)$ – the space of all polynomials on K of degree at most p.

Derivation of the discrete problem

Assume that u – sufficiently regular exact solution
– multiply equation (1), a) by any $\varphi \in H^2(\Omega, \mathcal{T}_h)$
– integrate over $K \in \mathcal{T}_h$
– apply Green's theorem
– sum over all $K \in \mathcal{T}_h$

After some manipulation we obtain the identity

$$\begin{aligned} &\int_\Omega \frac{\partial u}{\partial t} \varphi \,\mathrm{d}x \\ &+ \sum_{K\in\mathcal{T}_h} \sum_{\substack{\Gamma\in\mathcal{F}_h \\ \Gamma\subset\partial K}} \int_\Gamma \sum_{s=1}^d f_s(u)\,(n_{\partial K})_s\,\varphi|_\Gamma \,\mathrm{d}S \\ &- \sum_{K\in\mathcal{T}_h} \int_K \sum_{s=1}^d f_s(u)\,\frac{\partial \varphi}{\partial x_s}\,\mathrm{d}x \\ &+ \sum_{K\in\mathcal{T}_h} \int_K \varepsilon\,\nabla u\cdot\nabla\varphi \,\mathrm{d}x \\ &- \sum_{\Gamma\in\mathcal{F}_h^I} \int_\Gamma \varepsilon \langle \nabla u\rangle \cdot \boldsymbol{n}_\Gamma [\varphi] \,\mathrm{d}S \\ &- \sum_{\Gamma\in\mathcal{F}_h^D} \int_\Gamma \varepsilon\,\nabla u\cdot \boldsymbol{n}_\Gamma\,\varphi \,\mathrm{d}S \\ = &\int_\Omega g\,\varphi\,\mathrm{d}x + \sum_{\Gamma\in\mathcal{F}_h^N} \int_\Gamma \varepsilon\,\nabla\,u\cdot\boldsymbol{n}_\Gamma\,\varphi\,\mathrm{d}S. \end{aligned} \tag{12}$$

To the left-hand side of (12) we add now the terms

$$-\theta \sum_{\Gamma\in\mathcal{F}_h^I} \int_\Gamma \varepsilon \langle \nabla\varphi\rangle \cdot \boldsymbol{n}_\Gamma [u]\,\mathrm{d}S \quad (=0). \tag{13}$$

Further, to the left-hand side and the right-hand side of (12) we add the terms

$$-\theta \sum_{\Gamma\in\mathcal{F}_h^D} \int_\Gamma \varepsilon\, \nabla\varphi \cdot \boldsymbol{n}_\Gamma\, u\,\mathrm{d}S \tag{14}$$

and

$$-\theta \sum_{\Gamma\in\mathcal{F}_h^D} \int_\Gamma \varepsilon\, \nabla\varphi \cdot \boldsymbol{n}_\Gamma\, u_D\,\mathrm{d}S,$$

respectively, which are identical due to the Dirichlet condition
We consider tho following possibilities:

$\theta = -1$ nonsymmetric discretization of diffusion terms (NIPG) (15)

$\theta = 1$ symmetric discretization of diffusion terms (SIPG)
$\theta = 0$ incomplete discretization of diffusion terms (IIPG)

In view of the Neumann condition, we replace the second term on the right-hand side of (12) by

$$\sum_{\Gamma\in\mathcal{F}_h^N} \int_\Gamma g_N\,\varphi\,\mathrm{d}S. \tag{16}$$

Because of the stabilization of the scheme we introduce the *interior penalty*

$$\varepsilon \sum_{\Gamma\in\mathcal{F}_h^I} \int_\Gamma \sigma [u]\,[\varphi]\,\mathrm{d}S \quad (=0) \tag{17}$$

and the *boundary penalty*

$$\varepsilon \sum_{\Gamma\in\mathcal{F}_h^D} \int_\Gamma \sigma\, u\,\varphi\,\mathrm{d}S = \varepsilon \sum_{\Gamma\in\mathcal{F}_h^D} \int_\Gamma \sigma u_D \varphi \mathrm{d}S, \tag{18}$$

where σ is a suitable *weight*.

On the basis of above considerations we introduce the following forms defined for $u,\,\varphi \in H^2(\Omega,\mathcal{T}_h)$:
$(\cdot,\cdot)$ – $L^2(\Omega)$-scalar product,

$$\begin{aligned} a_h(u,\varphi) \;=\; & \sum_{K\in\mathcal{T}_h} \int_K \varepsilon\, \nabla u\cdot\nabla\varphi\,\mathrm{d}x \\ & - \sum_{\Gamma\in\mathcal{F}_h^I} \int_\Gamma \varepsilon \langle\nabla u\rangle\cdot\boldsymbol{n}_\Gamma [\varphi]\,\mathrm{d}S \\ & -\theta \sum_{\Gamma\in\mathcal{F}_h^I} \int_\Gamma \varepsilon \langle\nabla\varphi\rangle\cdot\boldsymbol{n}_\Gamma [u]\,\mathrm{d}S \\ & - \sum_{\Gamma\in\mathcal{F}_h^D} \int_\Gamma \varepsilon\, \nabla u\cdot\boldsymbol{n}_\Gamma\,\varphi\,\mathrm{d}S \\ & -\theta \sum_{\Gamma\in\mathcal{F}_h^D} \int_\Gamma \varepsilon\, \nabla\varphi\cdot\boldsymbol{n}_\Gamma\, u\,\mathrm{d}S \end{aligned} \tag{19}$$

diffusion form
$\theta = -1$ nonsymmetric discretization of diffusion terms (NIPG)
$\theta = 1$ symmetric discretization of diffusion terms (SIPG)
$\theta = 0$ incomplete discretization of diffusion terms (IIPG)

$$J_h^\sigma(u,\varphi) \;=\; \sum_{\Gamma\in\mathcal{F}_h^I} \int_\Gamma \sigma [u]\,[\varphi]\,\mathrm{d}S + \sum_{\Gamma\in\mathcal{F}_h^D} \int_\Gamma \sigma\, u\,\varphi\,\mathrm{d}S \tag{20}$$

interior and boundary penalty

$$\begin{aligned} \ell_h(\varphi)(t) \;=\; & \int_\Omega g(t)\,\varphi\,\mathrm{d}x + \sum_{\Gamma\in\mathcal{F}_h^N} \int_\Gamma g_N(t)\,\varphi\,\mathrm{d}S \\ & -\theta \sum_{\Gamma\in\mathcal{F}_h^D} \int_\Gamma \varepsilon\, \nabla\varphi\cdot\boldsymbol{n}_\Gamma\, u_D(t)\,\mathrm{d}S \\ & + \varepsilon \sum_{\Gamma\in\mathcal{F}_h^D} \int_\Gamma \sigma\, u_D(t)\,\varphi\,\mathrm{d}S \end{aligned} \tag{21}$$

right-hand side form

Finally, the convective terms are approximated with the aid of a *numerical flux* $H = H(u,v,\boldsymbol{n})$ by the form

$$\begin{aligned} b_h(u,\varphi) = & -\sum_{K\in\mathcal{T}_h} \int_K \sum_{s=1}^d f_s(u) \frac{\partial\varphi}{\partial x_s}\,\mathrm{d}x \\ & + \sum_{\Gamma\in\mathcal{F}_h^I} \int_\Gamma H\left(u|_\Gamma^{(L)}, u|_\Gamma^{(R)}, \boldsymbol{n}_\Gamma\right) [\varphi]|_\Gamma\,\mathrm{d}S \\ & + \sum_{\Gamma\in\mathcal{F}_h^{DN}} \int_\Gamma H\left(u|_\Gamma^{(L)}, u|_\Gamma^{(R)}, \boldsymbol{n}_\Gamma\right) \varphi|_\Gamma^{(L)}\,\mathrm{d}S \end{aligned} \tag{22}$$

convective form
H – **numerical flux**
Definition of the boundary state $u|_\Gamma^{(R)}$ **for** $\Gamma\subset\partial\Omega : u|_\Gamma^{(R)} := u|_\Gamma^{(L)}$ **(extrapolation)**

Assumptions (H):

1. $H(u,v,\boldsymbol{n})$ **is defined in** $\mathbb{R}^2\times B_1$, **where** $B_1 = \{\boldsymbol{n}\in\mathbb{R}^d; |\boldsymbol{n}| = 1\}$, **and** *Lipschitz-continuous* **with respect to** $u,\,v$:

$$|H(u,v,\boldsymbol{n}) - H(u^*,v^*,\boldsymbol{n})| \le C_L(|u-u^*| + |v-v^*|),$$
$$u,\,v,\,u^*,\,v^*\in\mathbb{R},\;\; \boldsymbol{n}\in B_1.$$

2. $H(u,v,\boldsymbol{n})$ **is** *consistent*:

$$H(u,u,\boldsymbol{n}) = \sum_{s=1}^d f_s(u)\,n_s, \quad u\in\mathbb{R},\;\; \boldsymbol{n} = (n_1,\ldots,n_d)\in B_1.$$

3. $H(u,v,\boldsymbol{n})$ **is** *conservative*:

$$H(u,v,\boldsymbol{n}) = -H(v,u,-\boldsymbol{n}), \quad u,\,v\in\mathbb{R},\;\; \boldsymbol{n}\in B_1.$$

The exact sufficiently regular solution u satisfies the identity

$$\left(\frac{\partial u(t)}{\partial t},\varphi_h\right)+b_h(u(t),\varphi_h)+a_h(u(t),\varphi_h)+\varepsilon J_h^\sigma(u(t),\varphi_h) =\ell_h(\varphi_h)(t) \quad \text{for all } \varphi_h\in S_h \text{ and for a.a. } t\in(0,T).$$

Discrete problem

We say that u_h is a DGFE approximate solution of the convection-diffusion problem (1), if

$$\begin{aligned}
&\text{a)} \quad u_h\in C^1([0,T];S_h), \qquad (23)\\
&\text{b)} \quad (\frac{\partial u_h(t)}{\partial t},\varphi_h)+a_h(u_h(t),\varphi_h)+b_h(u_h(t),\varphi_h)+J_h^\sigma(u_h(t),\varphi_h)\\
&\qquad =\ell_h(\varphi_h)(t) \qquad \forall\,\varphi_h\in S_h,\ \forall\, t\in(0,T),\\
&\text{c)} \quad u_h(0)=u_h^0=S_h\text{–approximation of } u^0.
\end{aligned}$$

The discrete problem is equivalent to a large system of nonlinear ordinary differential equations.

In practical computations: suitable *time discretization* is applied, e.g.
– Euler forward or backward scheme,
– Runge–Kutta methods,
– discontinuous Galerkin time discretization

The forward Euler and Runge-Kutta schemes are *conditionally stable* – time step is strongly restricted by the *CFL-stability condition*.

Suitable: *semi-implicit scheme* - leads to a linear algebraic system on each time level

Integrals are evaluated with the aid of *numerical integration*.

Error analysis

Assumptions

– Assumptions (H)
– The weak solution u of problem (1) is regular, namely

$$\frac{\partial u}{\partial t}\in L^2(0,T;H^{p+1}(\Omega)). \qquad (24)$$

Then

$$\begin{aligned}
&\frac{d}{dt}(u(t),\varphi_h)+a_h(u(t),\varphi_h)+\varepsilon J_h^\sigma(u(t),\varphi_h) \qquad (25)\\
&\quad +b_h(u(t),\varphi_h)=\ell_h(\varphi_h)(t),\\
&\quad \forall\varphi_h\in S_h,\ \text{for a.e. } t\in(0,T).
\end{aligned}$$

– $\{\mathcal{T}_h\}_{h\in(0,h_0)}$, $h_0>0$, - regular system of triangulations of the domain Ω: there exists $C_T>0$ such that

$$\frac{h_K}{\rho_K}\le C_T \quad \forall K\in\mathcal{T}_h \quad \forall h\in(0,h_0). \qquad (26)$$

Some auxiliary results

Multiplicative trace inequality:
There exists a constant $C_M>0$ independent of v, h and K such that

$$\begin{aligned}
&\|v\|^2_{L^2(\partial K)} \qquad (27)\\
&\le C_M\left(\|v\|_{L^2(K)}\,|v|_{H^1(K)}+h_K^{-1}\|v\|^2_{L^2(K)}\right),\\
&\quad K\in\mathcal{T}_h,\ v\in H^1(K),\ h\in(0,h_0).
\end{aligned}$$

Inverse inequality:
There exists a constant $C_I>0$ independent of v, h, and K such that

$$|v|_{H^1(K)}\le C_I h_K^{-1}\|v\|_{L^2(K)}, \quad v\in P^p(K),\ K\in\mathcal{T}_h,\ h\in(0,h_0). \qquad (28)$$

S_h-interpolation:
For $v\in L^2(\Omega)$ we denote by $\Pi_h v$ the $L^2(\Omega)$-projection of v on S_h:

$$\Pi_h v\in S_h, \quad (\Pi_h v-v,\varphi_h)=0 \qquad \forall\,\varphi_h\in S_h. \qquad (29)$$

Properties of the operator Π_h:
There exists a constant $C_A>0$ independent of h,K,v such that

$$\begin{aligned}
&\|\Pi_h v-v\|_{L^2(K)}\le C_A h_K^{k+1}|v|_{H^{k+1}(K)}, \qquad (30)\\
&|\Pi_h v-v|_{H^1(K)}\le C_A h_K^{k}|v|_{H^{k+1}(K)},\\
&|\Pi_h v-v|_{H^2(K)}\le C_A h_K^{k-1}|v|_{H^{k+1}(K)},
\end{aligned}$$

for all $v\in H^{k+1}(K)$, $K\in\mathcal{T}_h$ and $h\in(0,h_0)$, where $k\in[1,p]$ is an integer.

If u and u_h denote the exact and approximate solutions, then we set $\eta(t)=\Pi_h u(t)-u(t), \xi(t)=u_h(t)-\Pi_h u(t)(\in S_h)$ for a.e. $t\in(0,T)$.

Truncation error in the convection form: If $\partial\Omega_D=\partial\Omega, \partial\Omega_N=\emptyset$, then

$$\begin{aligned}
&|b_h(u,\xi)-b_h(u_h,\xi)| \qquad (31)\\
&\le C\left(|\xi|^2_{H^1(\Omega,\mathcal{T}_h)}+J_h^\sigma(\xi,\xi)\right)^{1/2}\left(h^{p+1}|u|_{H^{p+1}(\Omega)}+\|\xi\|_{L^2(\Omega)}\right).
\end{aligned}$$

If $\partial\Omega_N\neq\emptyset$, then

$$\begin{aligned}
&|b_h(u,\xi)-b_h(u_h,\xi)| \qquad (32)\\
&\le C\left(|\xi|^2_{H^1(\Omega,\mathcal{T}_h)}+J_h^\sigma(\xi,\xi)\right)^{1/2}\left(h^{p+1/2}|u|_{H^{p+1}(\Omega)}+\|\xi\|_{L^2(\Omega)}\right).
\end{aligned}$$

Coercivity:
An important step in the analysis of error estimates is the *coercivity* of the form

$$A_h(u,v)=a_h(u,v)+\varepsilon J_h^\sigma(u,v), \qquad (33)$$

which reads

$$\begin{aligned}
&A_h(\varphi_h,\varphi_h)\ge\frac{\varepsilon}{2}\left(|\varphi_h|^2_{H^1(\Omega,\mathcal{T}_h)}+J_h^\sigma(\varphi_h,\varphi_h)\right), \qquad (34)\\
&\quad \varphi\in S_h,\ h\in(0,h_0).
\end{aligned}$$

We shall discuss the validity of estimate (34) in various situations.

(I) Conforming mesh $\mathcal{T}_h$

Let the mesh $\mathcal{T}_h$ have the standard properties from the finite element method:
if $K, K' \in \mathcal{T}_h$, $K \neq K'$, then $K \cap K' = \emptyset$ or $K \cap K'$ is a common vertex or $K \cap K'$ is a common edge (or $K \cap K'$ is a common face in the case $d = 3$) of K and K'.
In this case we set

$$\sigma|_\Gamma = \frac{C_W}{d(\Gamma)}, \quad \Gamma \in \mathcal{F}_h. \tag{35}$$

Then the coercivity inequality (34) holds under the following choice of the constant C_W:

$$C_W > 0 \text{ (e. g. } C_W = 1\text{) for NIPG version}, \tag{36}$$
$$C_W \geq 4C_M(1 + C_I) \quad \text{for SIPG version}, \tag{37}$$
$$C_W \geq C_M(1 + C_I) \quad \text{for IIPG version}, \tag{38}$$

where C_M and C_I are constants from (27) and (28), respectively.

(II) Nonconforming mesh $\mathcal{T}_h$

In this case $\mathcal{T}_h$ is formed by closed triangles with mutually disjoint interiors with hanging nodes in general. Then the coercivity inequality (34) is guaranteed under conditions (36) – (38). However, in this case it is necessary to assume that

$$h_K \leq C_D\, d(\Gamma), \quad \Gamma \in \mathcal{F}_h, \Gamma \subset \partial K, \tag{39}$$

in order to prove the estimate

$$J_h^\sigma(\eta, \eta) \leq C h^p |u|_{H^{p+1}(\Omega)}. \tag{40}$$

(III) Nonconforming mesh $\mathcal{T}_h$ without assumption (39)

It is obvious that condition (39) is rather restrictive in some cases. In order to avoid it, we change the definition of the weight σ:

$$\sigma|_\Gamma = \frac{2C_W}{h_{K_\Gamma^{(L)}} + h_{K_\Gamma^{(R)}}}, \quad \Gamma \in \mathcal{F}_h^I, \tag{41}$$
$$\sigma|_\Gamma = \frac{C_W}{h_{K_\Gamma^{(L)}}}, \quad \Gamma \in \mathcal{F}_h^D.$$

Due to theoretical analysis, it is necessary to introduce the assumption of a "quasiuniformity" of the mesh:

$$h_{K_\Gamma^{(L)}} \leq C_N\, h_{K_\Gamma^{(R)}}, \quad \Gamma \in \mathcal{F}_h^I. \tag{42}$$

(Hence, $C_N \geq 1$.)
Then the coercivity inequality (34) holds under the following choice of C_W:

$$C_W > 0 \text{ (e. g. } C_W = 1\text{) for NIPG version}, \tag{43}$$
$$C_W \geq 2C_M(1 + C_I)\,(1 + C_N) \quad \text{for SIPG version}, \tag{44}$$
$$C_W \geq C_M(1 + C_I)\,(1 + C_N) \quad \text{for IIPG version}. \tag{45}$$

Proof of the coercivity inequality (34) in the case (III) for SIPG version:
Using the definition of the forms a_h and J_h^σ and the Cauchy and Young's inequalities, we find that for any $\delta > 0$ we have

$$a_h(\varphi_h, \varphi_h) \geq \varepsilon |\varphi_h|^2_{H^1(\Omega, \mathcal{T}_h)} - \varepsilon\omega - \varepsilon \frac{\delta}{C_W} J_h^\sigma(\varphi_h, \varphi_h),$$

where

$$\omega = \frac{1}{\delta} \sum_{\Gamma \in \mathcal{F}_h^I} \int_\Gamma \frac{h_{K_\Gamma^{(L)}} + h_{K_\Gamma^{(R)}}}{2} |\langle \nabla \varphi_h \rangle|^2 \, \mathrm{d}S + \sum_{\Gamma \in \mathcal{F}_h^D} \int_\Gamma h_{K_\Gamma^{(L)}} |\nabla \varphi_h|^2 \mathrm{d}S.$$

In view of (42),

$$\omega \leq \frac{1}{\delta} \frac{1 + C_N}{2} \sum_{K \in \mathcal{T}_h} h_K \int_{\partial K} |\nabla \varphi_h|^2 \mathrm{d}S.$$

Now, the application of (27) and (28) yields the estimate

$$\omega \leq \frac{1}{2\delta} C_M(1 + C_I)\,(1 + C_N)\, |\varphi_h|^2_{H^1(\Omega, \mathcal{T}_h)}.$$

If we set $\delta := C_M(1 + C_I)\,(1 + C_N)$ and use assumption (44), we immediately arrive at (34).

In the IIPG case we can proceed similarly.

Error estimates
Assumptions:
– (H),
– regularity of u,
– regularity of the mesh,
– $u_h^0 = \Pi_h u^0$,
– σ, $d(\Gamma), h_K$ and C_W satisfy assumptions from the cases (I) or (II) or (III).
Then the error $e_h = u - u_h$ satisfies the estimate

$$\max_{t \in [0,T]} \|e_h(t)\|^2_{L^2(\Omega)} \tag{46}$$
$$+ \frac{\varepsilon}{2} \int_0^t \left(|e_h(\vartheta)|^2_{H^1(\Omega, \mathcal{T}_h)} + J_h^\sigma(e_h(\vartheta), e_h(\vartheta)) \right) d\vartheta$$
$$\leq C\, h^{2p}, \quad h \in (0, h_0), \tag{47}$$

with a constant $C > 0$ independent of h.

Sketch of the proof

Let us subtract the relations valid for the exact and approximate solutions, set $\varphi_h = \xi_h$ and use the coercivity inequality:

$$\begin{aligned} &\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|\xi(t)\|^2_{L^2(\Omega)} + \frac{\varepsilon}{2}|\xi(t)|^2_{H^1(\Omega,\mathcal{T}_h)} + \frac{\varepsilon}{2}J^\sigma_h(\xi(t),\xi(t)) \quad (48)\\ \le\ & b_h(u(t),\xi(t)) - b_h(u_h(t),\xi(t)) - \left(\frac{\partial\eta(t)}{\partial t},\xi(t)\right)\\ & -a_h(\eta(t),\xi(t)) - \varepsilon J^\sigma_h(\eta(t),\xi(t)) \quad \text{for a.a. } (0,T). \end{aligned}$$

Now we estimate individual terms in (48):

$$\begin{aligned} &\frac{\mathrm{d}}{\mathrm{d}t}\|\xi\|^2_{L^2(\Omega)} + \varepsilon|\xi|^2_{H^1(\Omega,\mathcal{T}_h)} + \varepsilon J^\sigma_h(\xi,\xi) \quad (49)\\ \le\ & C\Big\{\left(J^\sigma_h(\xi,\xi)^{1/2} + |\xi|_{H^1(\Omega,\mathcal{T}_h)}\right)\left(\|\xi\|_{L^2(\Omega)} + h^p|u|_{H^{p+1}(\Omega)}\right)\\ & + h^{p+1}\,|\partial u/\partial t|_{H^{p+1}(\Omega)}\,\|\xi\|_{L^2(\Omega)}\\ & + \varepsilon\, h^{p+1/2}|u|_{H^{p+1}(\Omega)}\left(J^\sigma_h(\xi,\xi)^{1/2} + |\xi|_{H^1(\Omega,\mathcal{T}_h)}\right)\Big\} \end{aligned}$$

Now we apply Young's inequality:

$$\begin{aligned} &\frac{\mathrm{d}}{\mathrm{d}t}\|\xi\|^2_{L^2(\Omega)} + \varepsilon|\xi|^2_{H^1(\Omega,\mathcal{T}_h)} + \varepsilon J^\sigma_h(\xi,\xi) \quad (50)\\ \le\ & \frac{\varepsilon}{2}\left(J^\sigma_h(\xi,\xi) + |\xi|^2_{H^1(\Omega,\mathcal{T}_h)}\right) + C\left\{\left(1+\frac{1}{\varepsilon}\right)\|\xi\|^2_{L^2(\Omega)}\right.\\ & \left.+\frac{1}{\varepsilon}\left((\varepsilon^2h^{2p} + h^{2p+1})\,|u|^2_{H^{p+1}(\Omega)}\right) + h^{2p+2}\,|\partial u/\partial t|^2_{H^{p+1}(\Omega)}\right\}\\ & \text{a. e. in } (0,T). \end{aligned}$$

The integration of (50) from 0 to $t \in [0,T]$ and the relation $\xi(0) = u^0_h - \Pi_h u^0 = 0$ yield

$$\begin{aligned} &\|\xi(t)\|^2_{L^2(\Omega)} + \frac{\varepsilon}{2}\int_0^t\left(|\xi(\vartheta)|^2_{H^1(\Omega,\mathcal{T}_h)} + J^\sigma_h(\xi(\vartheta),\xi(\vartheta))\right)\mathrm{d}\vartheta \quad (51)\\ \le\ & C\left\{\left(1+\frac{1}{\varepsilon}\right)\int_0^t\|\xi(\vartheta)\|^2_{L^2(\Omega)}\,\mathrm{d}\vartheta + \frac{1}{\varepsilon}h^{2p}\int_0^t\left((\varepsilon^2+h)\,|u(\vartheta)|^2_{H^{p+1}(\Omega)}\right)\mathrm{d}\vartheta\right.\\ & \left.+\,h^{2p+2}\int_0^t|\partial u(\vartheta)/\partial t|^2_{H^{p+1}(\Omega)}\,\mathrm{d}\vartheta\right\}, \quad t\in[0,T]. \end{aligned}$$

Using Gronwall's lemma, we get

$$\begin{aligned} &\|\xi(t)\|^2_{L^2(\Omega)} + \frac{\varepsilon}{2}\int_0^t\left(|\xi(\vartheta)|^2_{H^1(\Omega,\mathcal{T}_h)} + J^\sigma_h(\xi(\vartheta),\xi(\vartheta))\right)\mathrm{d}\vartheta \quad (52)\\ \le\ & C\left((\varepsilon + h/\varepsilon)\|u\|^2_{L^2(0,T;H^{p+1}(\Omega))} + h^2\|\partial u/\partial t\|^2_{L^2(0,T;H^{p+1}(\Omega))}\right)\\ & \times h^{2p}\exp\left(\tilde{C}\frac{1+\varepsilon}{\varepsilon}t\right), \quad t\in[0,T], \end{aligned}$$

(C and $\tilde{C}$ are constants independent of t, h, ε, u).

Now, since $e_h = \xi + \eta$ and thus,

$$\begin{aligned} \|e_h\|^2_{L^2(\Omega)} &\le 2\left(\|\xi\|^2_{L^2(\Omega)} + \|\eta\|^2_{L^2(\Omega)}\right), \quad (53)\\ |e_h|^2_{H^1(\Omega,\mathcal{T}_h)} &\le 2\left(|\xi|^2_{H^1(\Omega,\mathcal{T}_h)} + |\eta|^2_{H^1(\Omega,\mathcal{T}_h)}\right),\\ J^\sigma_h(e_h,e_h) &\le 2\left(J^\sigma_h(\xi,\xi) + J^\sigma_h(\eta,\eta)\right), \end{aligned}$$

we use the above result, estimate the terms with η and obtain the sought error estimate.

Optimal error estimates

The error estimate (46) is optimal in the $L^2(H^1)$-norm, but suboptimal in the $L^\infty(L^2)$-norm.
We carried out the analysis of the $L^\infty(L^2)$-optimal error estimate under the following assumptions.

Assumptions (B):
– the discrete diffusion form a_h is symmetric (i.e. we consider the SIPG version),
– the polygonal domain Ω is convex,
– the exact solution u satisfies the regularity condition,
– conditions (H) are satisfied,
– $u^0_h = \Pi_h u^0$,
– $\Gamma_D = \partial\Omega$ and $\Gamma_N = \emptyset$.

The application of the Aubin-Nitsche technique based on the use of the elliptic dual problem considered for each $z \in L^2(\Omega)$:

$$-\Delta\psi = z \quad \text{in } \Omega, \qquad \psi|_{\partial\Omega} = 0. \quad (54)$$

Then the weak solution $\psi \in H^2(\Omega)$ and there exists a constant $C > 0$, independent of z, such that

$$\|\psi\|_{H^2(\Omega)} \le C\|z\|_{L^2(\Omega)}. \quad (55)$$

For each $h \in (0, h_0)$ and $t \in [0,T]$ we define the function $u^*_h(t)$ as the "A_h-projection" of $u(t)$ on S_h, i. e. a function satifying the conditions

$$u^*_h(t) \in S_h, \qquad A_h(u^*_h(t),\varphi_h) = A_h(u(t),\varphi_h) \quad \forall\,\varphi_h \in S_h, \quad (56)$$

and set $\chi = u - u^*_h$.
Using the elliptic dual problem (54), we proved the existence of a constant $C > 0$ such that

$$\|\chi\|_{L^2(\Omega)} \le Ch^{p+1}|u|_{H^{p+1}(\Omega)}, \quad (57)$$

$$\|\chi_t\|_{L^2(\Omega)} \le Ch^{p+1}|u_t|_{H^{p+1}(\Omega)}, \ h \in (0,h_0). \quad (58)$$

This, the estimate of the truncation error in the form b_h (31), multiple application of Young's inequality and Gronwall's lemma represent important tools for obtaining the $L^\infty(L^2)$-error estimate:

Theorem. Let assumptions (B) be fulfilled. Then the error $e_h = u - u_h$ satisfies the estimate

$$\|e_h\|_{L^\infty(0,T;L^2(\Omega))} \le Ch^{p+1}, \quad (59)$$

with a constant $C > 0$ independent of h.

Remark *The constant C in the error estimates is of the order $O(\exp(\tilde{C}T/\varepsilon)$, which blows up for $\varepsilon \to 0+$.*
= a consequence of the application of necessary tools for overcoming the nonlinear convective terms, namely Young's inequality and Gronwall's lemma.

Improved estimates for a linear model convection-diffusion-reaction problem

Find $u : Q_T = \Omega \times (0,T) \to I\!R$ **such that**

$$\begin{aligned} \frac{\partial u}{\partial t} + \boldsymbol{v}\cdot\nabla u - \varepsilon\triangle u + cu &= g \quad \text{in } Q_T,\\ u &= u_D \quad \text{on } \Gamma_D \times (0,T),\\ \varepsilon\frac{\partial u}{\partial \boldsymbol{n}} &= u_N \quad \text{on } \Gamma_N \times (0,T),\\ u(x,0) &= u^0(x), \quad x \in \Omega. \end{aligned}$$

Γ_D = **inlet, where** $\boldsymbol{v}\cdot\boldsymbol{n} < 0$

In the case $\varepsilon = 0$ **we put** $u_N = 0$ **and ignore the Neumann condition;** Γ_D = **inlet:** $\boldsymbol{v}\cdot\boldsymbol{n} < 0$

Assumptions on data (A)

a) some regularity of $g, u^0, u_D, u_N, \boldsymbol{v}, c$

b) $c - \frac{1}{2}\mathrm{div}\boldsymbol{v} \ge \gamma_0 > 0$ **in** Q_T **with a constant** γ_0**,**

c) $\varepsilon \ge 0$**.**

M.F. & K. Švadlenka: Error estimate

$$\begin{aligned} &\max_{t\in[0,T]} \|e_h(t)\|^2_{L^2(\Omega)}\\ + \quad &\frac{\varepsilon}{2}\int_0^T \left(|e_h(\vartheta)|^2_{H^1(\Omega,\mathcal{T}_h)} + J_h^\sigma(e_h(\vartheta), e_h(\vartheta))\right) \mathrm{d}\vartheta\\ &\le C(\varepsilon + h)h^{2p}, \end{aligned}$$

with C **independent of** $\varepsilon \to 0+$**.**

Is this estimate optimal??

Example

2D linear hyperbolic equation

$$\frac{\partial u}{\partial t} + v_1\frac{\partial u}{\partial x_1} + v_2\frac{\partial u}{\partial x_2} + cu = g \quad \text{in } \Omega \times (0,T),$$

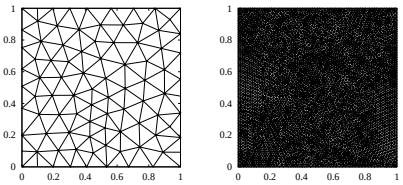
with $\Omega = (0,1)^2, v_1 = 0.3, v_2 = 0.4$ **and** $c = 0.5$**, equipped with initial condition and boundary condition.**

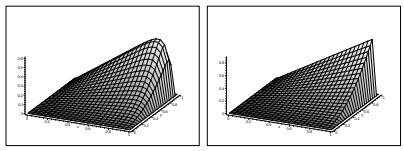
g, u^0 **– defined so that the exact solution has the form**

$$u(x_1,x_2,t) = \left(1 - e^{-t}\right)\left(x_1x_2^2 - x_2^2 e^{2^{\frac{x_1-1}{\nu}}} - x_1 e^{3^{\frac{x_2-1}{\nu}}} + e^{\frac{2x_1+3x_2-5}{\nu}}\right),$$

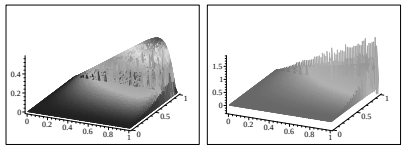
Linear elements applied on a sequence of meshes

The meshes $\mathcal{T}_{h_1}$ **and** $\mathcal{T}_{h_7}$**.**





The solution and approximate solution for $\nu = 0.1$ **(left) and** $\nu = 0.01$ **(right) at t=10**



Computational errors in L^2**-norm and the experimental order of convergence**

l	$\mathcal{T}_h$	h_l	$\nu = 0.1$ e_{h_l}	$\nu = 0.1$ α_l	$\nu = 0.01$ e_{h_l}	$\nu = 0.01$ α_l
1	125	0.173	0.0257	–	0.400	–
2	250	0.128	0.0158	1.61	0.272	1.28
3	500	0.090	0.0068	2.40	0.136	1.97
4	1000	0.064	0.0048	1.01	0.098	0.96
5	2000	0.045	0.0020	2.53	0.044	2.27
6	4000	0.032	0.0014	1.00	0.033	0.84
7	8000	0.023	0.0006	2.67	0.014	2.47
global order of accuracy $\bar{\alpha}$				1.85		1.66

Conclusion

Further results:

– the effect of numerical integration (M.F., V. Sobotíková)

– analysis of nonlinear diffusion depending on the sought solution (M.F., V. Kučera) and on the gradient of the solution (V. Dolejší),

– analysis of the hp-version of the DGFEM (V. Dolejší)

– analysis of BDF DG schemes (V. Dolejší, M. Vlasák)

– DGFEM in space and time (M.F., K. Švadlenka, J. Hájek, J. Česenek, V. Dolejší, M. Vlasák)

– DGFEM is rather robust and efficient technique for the numerical solution of convection-diffusion problems and compressible flow

– developed method allows to solve compressible flow with a wide range of Mach numbers

Discontinuous Galerkin Methods and Applications to Compressible Flow

Part 4. Examples of Some Further Applications of the DGFEM to Compressible Flow

Standard numerical methods have difficulties with the solution of low Mach number flows
$\Longrightarrow$ various modifications of the Euler (Navier-Stokes) equations are introduced (e.g. R. Klein, C.-D. Munz,...) allowing the solution of low Mach number flows

M.F., V. Dolejší, V. Kučera: DG unconditionally stable scheme for the solution of compressible flow using conservative variables – allowing the solution of flow with all positive Mach numbers
Main ingredients:

– semi-implicit time stepping based on
  homogeneity of fluxes
  Vijayasundaram numerical flux
– characteristic treatment of the boundary conditions
– isoparametric elements at curved boundaries

– limiting of order of accuracy in order to avoid the Gibbs phenomenon:

a)Define the discontinuity indicator $g^k(i)$ proposed by M.F., Dolejší and Schwab: Math. Comput. Simul. (2003):

$$g^k(K) = \int_{\partial K} [\rho_h^k]^2 \,\mathrm{d}S/(h_K|K|^{3/4}), \quad K \in \mathcal{T}_h. \tag{60}$$

b)Define the discrete indicator

$$G^k(K) = 0 \quad \text{if } g^k(K) < 1, \quad G^k(K) = 1 \quad \text{if } g^k(K) \geq 1, \quad K_i \in \mathcal{T}_h. \tag{61}$$

c)To the left-hand side of of the scheme we add the artificial viscosity form

$$\beta_h(\boldsymbol{w}_h^k, \boldsymbol{w}_h^{k+1}, \boldsymbol{\varphi}) = \nu_1 \sum_{K\in\mathcal{T}_h} h_K G^k(K) \int_K \nabla \boldsymbol{w}_h^{k+1} \cdot \nabla \boldsymbol{\varphi} \,\mathrm{d}\boldsymbol{x} \tag{62}$$

d)Augment the left-hand side of the scheme by adding the form

$$J_h(\boldsymbol{w}_h^k, \boldsymbol{w}_h^{k+1}, \boldsymbol{\varphi}) = \nu_2 \sum_{\Gamma\in\mathcal{F}_h^I} \frac{1}{2}(G^k(K_\Gamma^{(L)}) + G^k(K_\Gamma^{(R)})) \int_\Gamma [\boldsymbol{w}_h^{k+1}] \cdot [\boldsymbol{\varphi}] \,\mathrm{d}\mathcal{S}, \tag{63}$$

$\nu_1, \nu_2 \approx 1$.

Examples

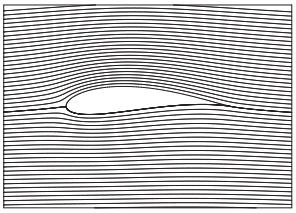
quadratic triangular elements

1) Inviscid flow

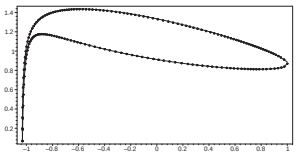
a) Low Mach number flow at incompressible limit

Stationary flow past a Joukowski profile
constant far field quantities $\Longrightarrow$ the flow is irrotational and homoentropic
complex function method: exact solution of incompressible inviscid irrotational flow satisfying the Kutta–Joukowski trailing condition, provided the velocity circulation around the profile, related to the magnitude of the far field velocity, $\gamma_{\text{ref}} = 0.7158$
Compressible flow: $M_\infty = 10^{-4}$, $\#\mathcal{T}_h = 5418$

The maximum density variation in compressible flow $\rho_{\max} - \rho_{\min} = 1.04 \cdot 10^{-8}$.
Computed velocity circulation related to the magnitude of the far field velocity: $\gamma_{\text{refcomp}} = 0.7205$, $\Longrightarrow$ the relative error 0.66%



Compressible low Mach flow past a Joukowski profile, approximate solution, streamlines

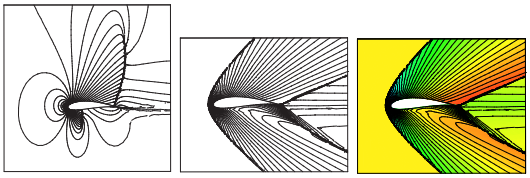


Velocity distribution along the profile: ∘ ∘ ∘ – exact solution of incompressible flow, ——— – approximate solution of compressible low Mach flow

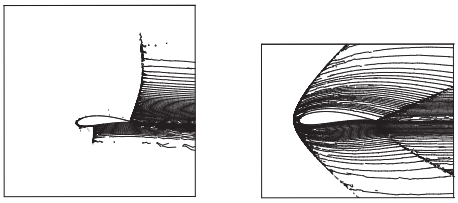
b) Transonic and hypersonic flow with shock waves past the Joukowski profile

with far field Mach number $M_\infty = 0.8$ and $M_\infty = 2.0$, respectively

The maximum density variation: $\rho_{max} - \rho_{min} = 0.94$ for $M_\infty = 0.8$ and $\rho_{max} - \rho_{min} = 2.61$ for $M_\infty = 2.0$



Mach number isolines of the flow past a Joukowski profile with $M_\infty = 0.8$ (left) and $M_\infty = 2.0$ (right)



Entropy isolines of the flow past a Joukowski profile with $M_\infty = 0.8$ (left) and $M_\infty = 2.0$ (right)

2) Viscous compressible flow

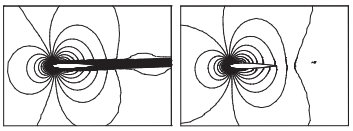
a) Stationary viscous flow past NACA0012 profile

$\theta = 0$ – IIPG

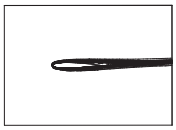
far-field Mach number $M = 0.5$

angle of attack $\alpha = 2°$

Reynolds number $Re = 5000$



NACA0012 $\alpha = 2°$ viscous flow, Mach number isolines (left), pressure isolines (right)



entropy isolines.

b) Non-stationary viscous flow past NACA0012 profile

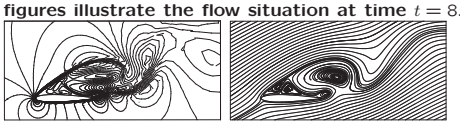
far-field flow has Mach number $M = 0.5$

angle of attack $\alpha = 25°$

Reynolds number $Re = 5000$

possible to observe an unsteady vortex shedding from the airfoil

figures illustrate the flow situation at time $t = 8.5$

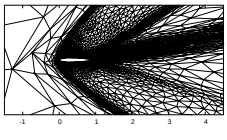


NACA0012 $\alpha = 25°$ viscous flow, Mach number isolines (left), streamlines (right)

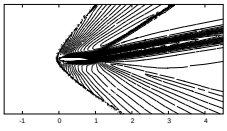
NACA0012 $\alpha = 25°$ viscous flow, entropy isolines

c) Hypersonic viscous flow

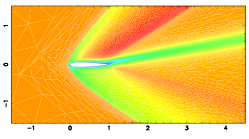
Flow past NACA0012 profile:
Far field Mach number $M_\infty = 2, \alpha = 10^\circ$
Reynolds number = 1000



Mesh for viscous flow - constructed by ANGENER - V. Dolejší



Mach number isolines for viscous flow



Distribution of the Mach number for viscous flow

3) Nonstationary inviscid flow in time-dependent domains

- part of simulation of fluid-structure interaction

DG combined with the ALE technique

(M. F., V. Kučera, J. Prokopová)

Importance of the simulation of fluid and structure interaction:

– **design of airplanes (investigation of wings and tails vibrations)**

– **design of steam turbomachines (vibrations of blades)**

– **car industry (in order to avoid noise)**

– **civil engineering (interaction of a strong wind with structures - TV towers, cooling towers, bridges etc.)**

– **medicine (creation of voice)**

In all these examples: flow of gases, i.e. compressible flow for low Mach numbers often incompressible model used sometimes the compressibility plays an important role

Continuous problem

Consider inviscid compressible flow in a bounded domain $\Omega_t \subset I\!R^2$ **depending on time** $t \in [0, T]$**. Let the boundary of** Ω_t **consist of three different parts:** $\partial\Omega_t = \Gamma_I \cup \Gamma_O \cup \Gamma_{W_t}$

Γ_I **- inlet**
Γ_O **- outlet**
Γ_{W_t} **- impermeable walls that may move in dependence on time.**

Euler equations written in the conservative form:

$$\frac{\partial \mathbf{w}}{\partial t} + \sum_{s=1}^{2} \frac{\partial \boldsymbol{f}_s(\mathbf{w})}{\partial x_s} = 0, \text{ in } \Omega_t, \ t \in (0, T), \tag{64}$$

$$\mathbf{w} = (\rho, \rho v_1, \rho v_2, E)^{\mathrm{T}} \in I\!R^4,$$

$$\boldsymbol{f}_i(w) = (\rho v_i, \rho v_1 v_i + \delta_{1i} p, \rho v_2 v_i + \delta_{2i} p, (E + p) v_i)^{\mathrm{T}}.$$

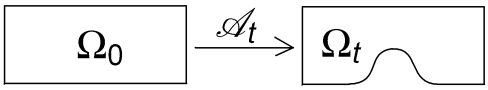
$$p = (\gamma - 1)\,(E - \rho |\mathrm{v}|^2/2). \tag{65}$$

Notation: ρ **- fluid density,** p **- pressure**
$\mathrm{v} = (v_1, v_2)$ **- velocity vector,** E **- total energy,** $\gamma > 1$ **- Poisson adiabatic constant**
Initial condition: $\mathbf{w}(x, 0) = \mathbf{w}^0(x), \ x \in \Omega_0$
Boundary conditions: based on the solution of a local linearized Riemann problem

ALE formulation

The dependence of the domain on time is taken into account with the aid of a regular **ALE mapping** from a reference domain Ω_0 onto the current configuration Ω_t:

$$\mathcal{A}_t : \overline{\Omega}_0 \to \overline{\Omega}_t, \text{ i.e. } \mathcal{A}_t : X \in \overline{\Omega}_0 \mapsto x = x(X,t) \in \overline{\Omega}_t. \tag{66}$$



The ALE mapping $\mathcal{A}_t$.

Domain velocity:

$$\tilde{z}(X,t) = \frac{\partial}{\partial t}\mathcal{A}_t(X), t \in [0,T], X \in \Omega_0, \tag{67}$$
$$z(x,t) = \tilde{z}(\mathcal{A}_t^{-1}(x), t),\ t \in [0,T],\ x \in \overline{\Omega}_t$$

ALE derivative of a function $f = f(x,t)$ defined for $x \in \Omega_t,\ t \in [0,T]$:

$$\frac{D^A}{Dt} f(x,t) = \frac{\partial \tilde{f}}{\partial t}(X,t)|_{X=\mathcal{A}_t^{-1}(x)}, \tag{68}$$

where

$$\tilde{f}(X,t) = f(\mathcal{A}_t(X), t),\ X \in \Omega_0.$$

It is possible to show that

$$\frac{D^A f}{Dt} = \frac{\partial f}{\partial t} + \boldsymbol{z} \cdot \operatorname{grad} f = \frac{\partial f}{\partial t} + \operatorname{div}(\boldsymbol{z} f) - f \operatorname{div} \boldsymbol{z}. \tag{69}$$

$\Longrightarrow$ **ALE formulation** of the Euler equations:

$$\frac{D^A \mathbf{w}}{Dt} + \sum_{s=1}^{2} \frac{\partial \boldsymbol{g}_s(\mathbf{w})}{\partial x_s} + \mathbf{w} \operatorname{div} \boldsymbol{z} = 0,$$

$\boldsymbol{g}_s, s = 1, 2,$ - **ALE modified inviscid fluxes**:

$$\boldsymbol{g}_s(w) := \boldsymbol{f}_s(w) - z_s \mathbf{w}. \tag{70}$$

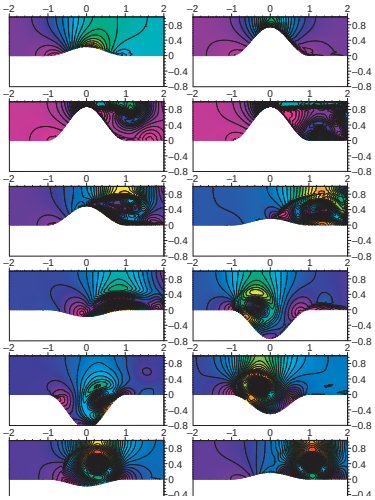
Example

Consider compressible flow in a channel with the initial rectangular shape $\Omega_0 = [-2,2] \times [0,1]$, where the lower wall of the channel is moving in the interval $X_1 \in (-1,1)$:

$$0.45 \sin(0.4t)\,(\cos(\pi X_1) + 1),\ X_1 \in (-1,1). \tag{71}$$

This movement is interpolated to the whole domain resulting in the ALE mapping $\mathcal{A}_t$.

Figure 1: velocity isolines at different time instants during one period



The solution contains a **vortex formation**, when the lower wall starts to descend, convected through the domain. Moreover, we see that a **contact discontinuity** is developed, when the channel becomes narrow.

Conclusion

– DGFEM = a robust and accurate method for the solution of compressible flow
– combination with ALE method allows the solution of flow problems in time dependent domains

Further goals:

– include viscosity
– coupling with structure models
– applications to complex FSI problems

Duality for QP problems with semidefinite Hessian and contact problems

Z. Dostál

VSB-Technical University Ostrava

1 Introduction

The duality theory of convex programming turned out to be an important tool in the development of scalable algorithms for the numerical solution of elliptic partial differential equations, such as the Lamé equations describing equilibrium of an elastic body subject to prescribed traction [14], as it enables to reduce the original problem to the problem with more favorable structure. It seems that it is even more important for the solution contact problems of elasticity, as it reduces the inequality constraints, which describe the non-penetration conditions (frictionless problems) [13] or stick-slip conditions (Tresca friction) [7], to the bound constraints, which can be treated much more efficiently by the recently proposed algorithms [6, 16]. Moreover, it can turn the more difficult semicoercive contact problems into much simpler strictly convex quadratic problems.

To exploit the latter feature effectively, it is necessary to have the duality theory for convex quadratic problems which admits cost functions with symmetric positive semidefinite (SPS) Hessian. In spite of its obvious importance, the author have not found a convenient reference to the duality theory concerning the primal problem

$$\min_{\boldsymbol{x}\in\Omega_{IE}} f(\boldsymbol{x}), \quad \Omega_{IE} = \{\boldsymbol{x}\in\mathbb{R}^n:\ [\boldsymbol{B}\boldsymbol{x}]_{\mathcal{I}} \le \boldsymbol{c}_{\mathcal{I}},\ [\boldsymbol{B}\boldsymbol{x}]_{\mathcal{E}} = \boldsymbol{c}_{\mathcal{E}}\}, \tag{1.1}$$

where f is a quadratic function with the symmetric Hessian $\boldsymbol{A}\in\mathbb{R}^{n\times n}$ and the linear term defined by $\boldsymbol{b}\in\mathbb{R}^n$, $\boldsymbol{B} = [\boldsymbol{b}_1,\dots,\boldsymbol{b}_m]^T \in \mathbb{R}^{m\times n}$ is a matrix with possibly dependent rows, $\boldsymbol{c} = [c_i]\in\mathbb{R}^m$, and $\mathcal{I}$, $\mathcal{E}$ are disjoint sets of indices which decompose $\{1,\dots,m\}$. The point of this lecture is to fill in this gap and to indicate applications in development of effective solvers for contact problems.

2 Constrained dual problem

First observe that if $\boldsymbol{A}$ is only positive semidefinite and $\boldsymbol{b}\ne\boldsymbol{o}$, then the cost function f need not be bounded from below. Thus $-\infty$ can be in the range of the dual function Θ. We resolve this problem by keeping Θ quadratic at the cost of introducing equality constraints. The basic results read as follows.

Theorem 5 *Let matrices $\boldsymbol{A},\boldsymbol{B}$, vectors $\boldsymbol{b},\boldsymbol{c}$, and index sets $\mathcal{I},\mathcal{E}$ be those from the definition of problem (1.1) with $\boldsymbol{A}$ positive semidefinite and $\Omega_{IE}\ne\emptyset$. Let $\boldsymbol{R}\in\mathbb{R}^{n\times d}$ be a full rank matrix such that*

$$\mathrm{Im}\boldsymbol{R} = \mathrm{Ker}\boldsymbol{A},$$

let $\boldsymbol{A}^+$ denote a symmetric positive semidefinite generalized inverse of $\boldsymbol{A}$, and let

$$\Theta(\boldsymbol{\lambda}) = -\frac{1}{2}\boldsymbol{\lambda}^T\boldsymbol{B}\boldsymbol{A}^+\boldsymbol{B}^T\boldsymbol{\lambda} + \boldsymbol{\lambda}^T(\boldsymbol{B}\boldsymbol{A}^+\boldsymbol{b}-\boldsymbol{c}) - \frac{1}{2}\boldsymbol{b}^T\boldsymbol{A}^+\boldsymbol{b}. \tag{2.1}$$

Then the following statements hold:
(i) If $(\overline{\boldsymbol{x}}, \overline{\boldsymbol{\lambda}})$ *is a KKT pair for (1.1), then* $\overline{\boldsymbol{\lambda}}$ *is a solution of*

$$\max_{\boldsymbol{\lambda} \in \Omega_{BE}} \Theta(\boldsymbol{\lambda}), \quad \Omega_{BE} = \{\boldsymbol{\lambda} \in \mathbb{R}^m : \ \boldsymbol{\lambda}_{\mathcal{I}} \geq \boldsymbol{o}, \ \boldsymbol{R}^T \boldsymbol{B}^T \boldsymbol{\lambda} = \boldsymbol{R}^T \boldsymbol{b}\}. \tag{2.2}$$

Moreover, there is $\overline{\boldsymbol{\alpha}} \in \mathbb{R}^d$ *such that* $(\overline{\boldsymbol{\lambda}}, \overline{\boldsymbol{\alpha}})$ *is a KKT pair for problem (2.2) and*

$$\overline{\boldsymbol{x}} = \boldsymbol{A}^+(\boldsymbol{b} - \boldsymbol{B}^T \overline{\boldsymbol{\lambda}}) + \boldsymbol{R}\overline{\boldsymbol{\alpha}}. \tag{2.3}$$

(ii) If $(\overline{\boldsymbol{\lambda}}, \overline{\boldsymbol{\alpha}})$ *is a KKT pair for problem (2.2), then* $\overline{\boldsymbol{x}}$ *defined by (2.3) is a solution of the equality and inequality constrained problem (1.1).*
(iii) If $(\overline{\boldsymbol{x}}, \overline{\boldsymbol{\lambda}})$ *is a KKT pair for problem (1.1), then*

$$f(\overline{\boldsymbol{x}}) = \Theta(\overline{\boldsymbol{\lambda}}). \tag{2.4}$$

For the proof see [6].

2.0.1 Uniqueness of a KKT pair

We shall supply our basic result on duality with the results concerning the uniqueness of the solution for the *constrained dual problem*

$$\min_{\boldsymbol{\lambda} \in \Omega_{BE}} \theta(\boldsymbol{\lambda}), \quad \Omega_{BE} = \{\boldsymbol{\lambda} \in \mathbb{R}^m : \ \boldsymbol{\lambda}_{\mathcal{I}} \geq \boldsymbol{o}, \ \boldsymbol{R}^T \boldsymbol{B}^T \boldsymbol{\lambda} = \boldsymbol{R}^T \boldsymbol{b}\}, \tag{2.5}$$

where θ is defined by

$$\theta(\boldsymbol{\lambda}) = -\Theta(\boldsymbol{\lambda}) - \frac{1}{2}\boldsymbol{b}^T \boldsymbol{A}^+ \boldsymbol{b} = \frac{1}{2}\boldsymbol{\lambda}^T \boldsymbol{B}\boldsymbol{A}^+ \boldsymbol{B}^T \boldsymbol{\lambda} - \boldsymbol{\lambda}^T(\boldsymbol{B}\boldsymbol{A}^+ \boldsymbol{b} - \boldsymbol{c}). \tag{2.6}$$

Theorem 6 *Let the matrices* $\boldsymbol{A}, \boldsymbol{B}$*, the vectors* $\boldsymbol{b}, \boldsymbol{c}$*, and the index sets* $\mathcal{I}, \mathcal{E}$ *be those from the definition of problem (1.1) with* $\boldsymbol{A}$ *positive semidefinite,* $\Omega_{IE} \neq \emptyset$*, and* $\Omega_{BE} \neq \emptyset$*. Let* $\boldsymbol{R} \in \mathbb{R}^{n \times d}$ *be a full rank matrix such that*

$$\mathrm{Im}\boldsymbol{R} = \mathrm{Ker}\boldsymbol{A}.$$

Then the following statements hold:

(i) If $\boldsymbol{B}^T$ *and* $\boldsymbol{B}\boldsymbol{R}$ *are full column rank matrices, then there is a unique solution* $\widehat{\boldsymbol{\lambda}}$ *of problem (2.5).*

(ii) If $\widehat{\boldsymbol{\lambda}}$ *is a unique solution of the constrained dual problem (2.5),*

$$\mathcal{A} = \{i : \ [\boldsymbol{\lambda}]_i > 0\} \cup \mathcal{E},$$

and $\boldsymbol{B}_{\mathcal{A}*}\boldsymbol{R}$ *is a full column rank matrix, then there is a unique triple* $(\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{\lambda}}, \widehat{\boldsymbol{\alpha}})$ *such that* $(\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{\lambda}})$ *solves the primal problem (1.1) and* $(\widehat{\boldsymbol{\lambda}}, \widehat{\boldsymbol{\alpha}})$ *solves the constrained dual problem (2.5). If* $\widehat{\boldsymbol{\lambda}}$ *is known, then*

$$\widehat{\boldsymbol{\alpha}} = (\boldsymbol{R}^T \boldsymbol{B}_{\mathcal{A}*}^T \boldsymbol{B}_{\mathcal{A}*} \boldsymbol{R})^{-1} \boldsymbol{R}^T \boldsymbol{B}_{\mathcal{A}*}^T \left(\boldsymbol{B}_{\mathcal{A}*} \boldsymbol{A}^+ \boldsymbol{B}^T \widehat{\boldsymbol{\lambda}} - (\boldsymbol{B}_{\mathcal{A}*} \boldsymbol{A}^+ \boldsymbol{b} - \boldsymbol{c}_{\mathcal{A}}) \right) \tag{2.7}$$

and

$$\widehat{\boldsymbol{x}} = \boldsymbol{A}^+(\boldsymbol{b} - \boldsymbol{B}^T \widehat{\boldsymbol{\lambda}}) + \boldsymbol{R}\widehat{\boldsymbol{\alpha}}. \tag{2.8}$$

(iii) If $\boldsymbol{B}^T$ *and* $\boldsymbol{B}_{\mathcal{E}*}\boldsymbol{R}$ *are full column rank matrices, then there is a unique triple* $(\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{\lambda}}, \widehat{\boldsymbol{\alpha}})$ *such that* $(\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{\lambda}})$ *solves the primal problem (1.1) and* $(\widehat{\boldsymbol{\lambda}}, \widehat{\boldsymbol{\alpha}})$ *solves the constrained dual problem (2.5).*

For the proof see [6]. The mechanical illustration of the above theorem is in Figure 1 and Figure 2.

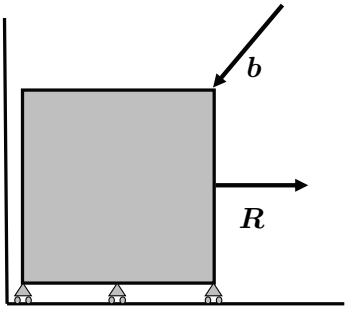


Figure 1: Unique displacement

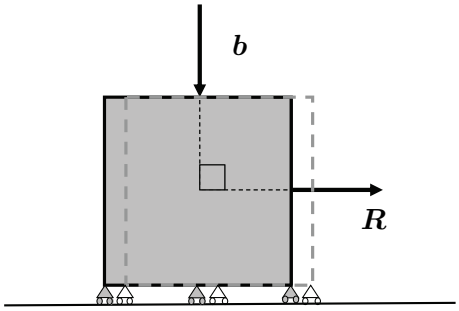


Figure 2: Nonunique displacement

3 Optimal solution of contact problems

The above results are, together with the Total FETI [9] and the results in development of optimal quadratic programming algorithms [6, 4, 12], the key ingredients in the development of scalable algorithms for the solution of contact problems of elasticity discretized either by the boundary element method [2, 17] or the finite element method [11]. After resolving some long standing problems, such as convergence of the algorithms for longer steps [5] or stable and cheap evaluation of the generalized inverse [10], the algorithms were implemented into our MATLAB code MatSol [15] and used to the parallel solution of large academic problems (with more then 10 millions of nodal variables) [11] and difficult real world problems, such as analysis of ball bearings in Figure 3 (see [3]).

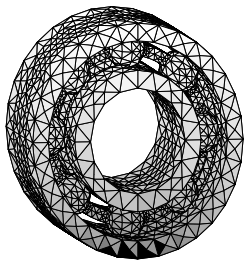


Figure 3: Ball bearings

4 Conclusion

We have presented some recent results of duality theory and indicated their role in development of optimal algorithms for contact problems. Current research includes parallel implementation of the algorithms in C, implementation of preconditioners, and adaptation of our algorithms to the solution of more complex problems [12].

Acknowledgement: This research is supported by the project MSM6198910027 provided by the Ministry of Education of Czech Republic and the project GAČR 201/07/0294 provided by the Grant Agency of Czech Republic.

References

[1] D.P. Bertsekas: Nonlinear Optimization. Athena Scientific, Belmont (1999)

[2] J. Bouchala, Z. Dostál, M.Sadowská: Theoretically Supported Scalable BETI Method for Variational Inequalities. Computing **82**, 53–75 (2008)

[3] T. Brzobohatý, Z. Dostál, T. Kozubek, A. Markopoulos, V. Vondrák: Scalable TFETI algorithm for the solution of semicoercive multibody contact problems of elasticity. In preparation.

[4] Z. Dostál: Inexact semi-monotonic augmented Lagrangians with optimal feasibility convergence for quadratic programming with simple bounds and equality constraints. SIAM J. Numer. Anal. **43**, 1, 96-115 (2005)

[5] Z. Dostál: On the decrease of a quadratic function along the projected–gradient path. ETNA **31**, 25–59 (2008)

[6] Z. Dostál: Optimal Quadratic Programming Algorithms, with Applications to Variational Inequalities, Springer, New York, 2009.

[7] Z. Dostál, J. Haslinger, R. Kučera: Implementation of fixed point method for duality based solution of contact problems with friction. J. Comput. Appl. Math. **140**, 1–2, 245–256 (2002)

[8] Z. Dostál, D. Horák: Theoretically supported scalable FETI for numerical solution of variational inequalities, SIAM Journal on Numerical Analysis 45, 2 (2007) 500513.

[9] Z. Dostál, D. Horák, R. Kučera: Total FETI - an easier implementable variant of the FETI method for numerical solution of elliptic PDE, Communications in Numerical Methods in Engineering 22 (2006) 1155-1162.

[10] Z. Dostál, T. Kozubek, A. Markopoulos, M. Menšík: Cholesky factorization of a positive semidefinite matrix with known null space, submitted.

[11] Z. Dostál, T. Kozubek, V. Vondrák: Scalable TFETI algorithm for the solution of coercive multibody contact problems of elasticity. Submitted.

[12] Z. Dostál, R. Kučera: An Optimal Algorithm for Minimization of Quadratic Functions with Bounded Spectrum subject to Separable Constraints, in preparation.

[13] Z. Dostál, V. Vondrák, D. Horák, C. Farhat, P. Avery: Scalable FETI algorithms for frictionless contact problems. In *Domain Methods in Science and Engineering XVII*. Langer U et al. (eds). Springer, Berlin, Lecture Notes in Computational Science and Engineering (LNCSE) 2008; **60**:263–270.

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

[15] T. Kozubek , A. Markopoulos , T. Brzobohatý , R. Kučera , V. Vondrák: MatSol - MATLAB efficient solvers for problems in engineering. http://www.am.vsb.cz/matsol.

[16] R. Kučera, Convergence rate of an optimization algorithm for minimizing quadratic functions with separable convex constraints, SIAM J. Optim., 19 (2008), pp. 846–862.

[17] Sadowská, M.: Scalable Total BETI for 2D and 3D Contact Problems. Ph.D. Thesis, FEECS VŠB-Technical University of Ostrava (2008)

①

Structural optimization

a) sizing optimization;
b) shape optimization;
c) topology optimization

Abstract setting of a class of shape optimization problems

$\mathcal{O}$ ... family of admissible domains
$\mathcal{O} \subset \tilde{\mathcal{O}}$

$(\mathcal{P}(\Omega))$ $\Omega \longmapsto u(\Omega) \in V(\Omega)$ ... solution of a state problem (PDE, inequality, ..)

$\mathcal{G} = \{(\Omega, u) \mid \Omega \in \mathcal{O},\ u \text{ solves } (\mathcal{P}(\Omega))\}$

$J : (\Omega, y) \longmapsto J(\Omega, y) \in \mathbb{R}^1$ ... cost functional

②

$$(\mathbb{P}) \quad \begin{cases} \text{Find } (\Omega^*, u^*) \in \mathcal{G} \text{ s.t.} \\ J(\Omega^*, u^*) = \min\limits_{\mathcal{G}} J(\Omega, u) \end{cases}$$

Existence analysis

Convergence in $\tilde{\mathcal{O}}$

$\{\Omega_n\},\ \Omega_n, \Omega \in \tilde{\mathcal{O}}$ $\qquad \Omega_n \xrightarrow{\tilde{\mathcal{O}}} \Omega,\ n \to \infty$

Requirement: $\Omega_n \xrightarrow{\tilde{\mathcal{O}}} \Omega \Rightarrow \Omega_{n_k} \xrightarrow{\tilde{\mathcal{O}}} \Omega$ for any $\{\Omega_{n_k}\} \subset \{\Omega_n\}$

Example: $\Omega_n \to \Omega \iff \partial\Omega_n \rightrightarrows \partial\Omega,\ n \to \infty$.

Convergence in $\{V(\Omega), \Omega \in \tilde{\mathcal{O}}\}$

$\{y_n\},\ y_n \in V(\Omega_n),\ y \in V(\Omega)$ $\qquad y_n \rightsquigarrow y$

Requirement: $y_n \rightsquigarrow y \Rightarrow y_{n_k} \rightsquigarrow y$ for any $\{y_{n_k}\} \subset \{y_n\}$.

③

Assumptions

(A1) $\mathcal{G}$ is compact:

for any $\{(\Omega_n, u_n)\}$, $(\Omega_n, u_n) \in \mathcal{G}$ $\exists \{(\Omega_{n_k}, u_{n_k})\} \subset \{(\Omega_n, u_n)\}$ $\exists (\Omega, u) \in \mathcal{G}$ s.t.

$\Omega_{n_k} \xrightarrow{\tilde{\mathcal{O}}} \Omega$, $u_n \rightsquigarrow u$

(A2) J is lower semicontinuous:

$\Omega_n \xrightarrow{\tilde{\mathcal{O}}} \Omega$, $y_n \rightsquigarrow y \Rightarrow \liminf\limits_{n\to\infty} J(\Omega_n, y_n) \geq J(\Omega, y)$

$\Omega_n, \Omega \in \tilde{\mathcal{O}}$, $y_n \in V(\Omega_n)$, $y \in V(\Omega)$

Theorem Let (A1) and (A2) be satisfied. Then $(\mathbb{P})$ has a solution

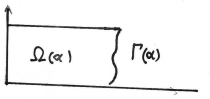
Verification of (A1)

(i) $\mathcal{O}$ is compact with respect to $\xrightarrow{\tilde{\mathcal{O}}}$;

(ii) $\Omega_n \xrightarrow{\tilde{\mathcal{O}}} \Omega$ and $(\Omega_n, u_n) \in \mathcal{G} \Rightarrow \exists \{u_{n_k}\} \subset \{u_n\}$ s.t. $u_{n_k} \rightsquigarrow u$ and $(\Omega, u) \in \mathcal{G}$

④

Applications



$\mathcal{O} = \{\Omega(\alpha) \mid \alpha \in \mathcal{U}_{ad}\}$
$\tilde{\mathcal{O}} = \{\Omega(\alpha) \mid \alpha \in \tilde{\mathcal{U}}_{ad}\}$

$\tilde{\mathcal{U}}_{ad} = \{\alpha \in C^{0,1}([0,1]) \mid 0 < \alpha_{min} \leq \alpha \leq \alpha_{max} \text{ in } [0,1],\ |\alpha'| \leq L_0 \text{ a.e. in } (0,1)\}$

$\mathcal{U}_{ad} = \{\alpha \in \tilde{\mathcal{U}}_{ad} \mid \text{meas } \Omega(\alpha) = \gamma > 0 \text{ given}\}$

Cost functional

$$J(\alpha, y) = \frac{1}{2} \int_{\Omega(\alpha)} (y - z_d)^2 \, dx, \qquad z_d \in L^2_{loc}(\mathbb{R}^2)$$

Example 1 (Dirichlet problem)

$$\begin{cases} -\Delta u(\alpha) = f & \text{in } \Omega(\alpha) \\ u(\alpha) = 0 & \text{on } \partial\Omega(\alpha) \end{cases}, \quad f \in L^2_{loc}(\mathbb{R}^2)$$

$$(\mathcal{P}(\alpha)) \quad \begin{cases} \text{Find } u(\alpha) \in H^1_0(\Omega(\alpha)) \text{ s.t.} \\ \int\limits_{\Omega(\alpha)} \nabla u(\alpha) \cdot \nabla v \, dx = \int\limits_{\Omega(\alpha)} f v \, dx \quad \forall v \in H^1_0(\Omega(\alpha)) \end{cases}$$

(5)

$$(\mathbb{P}) \quad \begin{cases} \text{Find } \alpha^* \in \mathcal{U}_{ad} \text{ s.t.} \\ J(\alpha^*, u(\alpha^*)) \le J(\alpha, u(\alpha)) \quad \forall \alpha \in \mathcal{U}_{ad}, \end{cases}$$

where $u(\alpha)$ solves $(\mathcal{P}(\alpha))$.

Convergence in $\tilde{O}$: $\Omega_n \xrightarrow{\tilde{\sigma}} \Omega \iff \alpha_n \to \alpha$ in $[0,1]$;

Convergence in $\{H_0^1(\Omega(\alpha)), \alpha \in \tilde{\mathcal{U}}_{ad}\}$:

$\hat{\Omega} \supset \Omega(\alpha) \quad \forall \alpha \in \tilde{\mathcal{U}}_{ad}$

$$y \longmapsto \tilde{y} = \begin{cases} y & \text{in } \Omega(\alpha), \\ 0 & \text{in } \hat{\Omega} \setminus \bar{\Omega}(\alpha) \end{cases}, \quad y \in H_0^1(\Omega(\alpha))$$

$$y_n \rightsquigarrow y \iff \begin{cases} \tilde{y}_n \to \tilde{y} \text{ in } H_0^1(\hat{\Omega}) \\ \text{or} \\ \tilde{y}_n \to \tilde{y} \text{ in } H_0^1(\hat{\Omega}) \end{cases}$$

$y_n \in H_0^1(\Omega(\alpha_n))$, $y \in H_0^1(\Omega(\alpha))$

Theorem Problem $(\mathbb{P})$ has a solution.

(6)

Lemma (verification of (A1)). Let $\alpha_n \to \alpha$ in $[0,1]$, $\alpha_n, \alpha \in \tilde{\mathcal{U}}_{ad}$ and $u_n := u(\alpha_n)$ be a solution of $(\mathcal{P}(\alpha_n))$, $n \to \infty$. Then

$$\tilde{u}_n \to \tilde{u} \text{ in } H_0^1(\hat{\Omega})$$

and $u(\alpha) := \tilde{u}|_{\Omega(\alpha)}$ solves $(\mathcal{P}(\alpha))$.

Proof: (a) boundedness of $\{\tilde{u}_n\}$:

$$c\|\tilde{u}_n\|_{1,\hat{\Omega}}^2 \le |\tilde{u}_n|_{1,\hat{\Omega}}^2 = \int_{\Omega_n} |\nabla u_n|^2 dx = \int_{\Omega_n} f u_n \, dx \le \|f\|_{0,\hat{\Omega}} \|\tilde{u}_n\|_{1,\hat{\Omega}}$$

↑ Friedrich's inequality

$\exists \{\tilde{u}_{n_k}\} \subset \{\tilde{u}_n\}$: $\tilde{u}_{n_k} \to \tilde{u} \in H_0^1(\hat{\Omega})$

(b) $\tilde{u}|_{\Omega(\alpha)}$ solves $(\mathcal{P}(\alpha))$

$\tilde{u} \equiv 0$ in $\hat{\Omega} \setminus \Omega(\alpha) \Rightarrow \tilde{u}|_{\Omega(\alpha)} \in H_0^1(\Omega(\alpha))$

Let $\varphi \in C_0^\infty(\Omega(\alpha))$ be given.

Then $\tilde{\varphi}|_{\Omega_n} \in C_0^\infty(\Omega_n)$ for n large enough.

(7)

$$\int_{\hat{\Omega}} \nabla \tilde{u}_{n_k} \cdot \nabla \tilde{\varphi} \, dx = \int_{\hat{\Omega}} f \tilde{\varphi} \, dx$$

$$\downarrow$$

$$\int_{\hat{\Omega}} \nabla \tilde{u} \cdot \nabla \tilde{\varphi} \, dx = \int_{\hat{\Omega}} f \tilde{\varphi} \, dx$$

$$\Updownarrow$$

$$\int_{\Omega(\alpha)} \nabla u(\alpha) \cdot \nabla \varphi \, dx = \int_{\Omega(\alpha)} f \varphi \, dx \quad \forall \varphi \in C_0^\infty(\Omega(\alpha))$$

$\Rightarrow u(\alpha)$ solves $(\mathcal{P}(\alpha))$

Strong convergence : $|\tilde{u}_{n_k}|_{1,\hat{\Omega}} \to |\tilde{u}|_{1,\hat{\Omega}}$

∎

Lemma (verification of (A2)). Let $\alpha_n \to \alpha$ in $[0,1]$, $\alpha_n, \alpha \in \tilde{\mathcal{U}}_{ad}$ and $\tilde{y}_n \to \tilde{y}$ in $H_0^1(\hat{\Omega})$, where $y_n \in H_0^1(\Omega_n)$, $y \in H_0^1(\Omega(\alpha))$. Then

$$J(\alpha_n, y_n) \to J(\alpha, y), \quad n \to \infty.$$

Proof: $J(\alpha_n, y_n) = \frac{1}{2} \int_{\hat{\Omega}} \chi_n |\tilde{y}_n - z_d|^2 dx \to \frac{1}{2} \int_{\hat{\Omega}} \chi |\tilde{y} - z_d|^2 dx$

(8)

where χ_n, χ are the characteristic functions of $\Omega(\alpha_n)$, $\Omega(\alpha)$, respectively

Example 2 (Neumann problem)

$$\begin{cases} -\Delta u(\alpha) + u(\alpha) = f \text{ in } \Omega(\alpha), \ f \in L^2_{bc}(\mathbb{R}^2) \\ \frac{\partial u(\alpha)}{\partial \nu} = 0 \quad \text{on } \partial\Omega(\alpha) \end{cases}$$

$$(\mathcal{P}(\alpha)) \quad \begin{cases} \text{Find } u(\alpha) \in H^1(\Omega(\alpha)) \text{ s.t.} \\ \int_{\Omega(\alpha)} (\nabla u(\alpha) \cdot \nabla v + u(\alpha) v) dx = \int_{\Omega(\alpha)} f v \, dx \quad \forall v \in H^1(\Omega(\alpha)) \end{cases}$$

Convergence in $\{H^1(\Omega(\alpha)), \alpha \in \tilde{\mathcal{U}}_{ad}\}$

$\tilde{y} = E_\alpha y \qquad E_\alpha \in \mathcal{L}(H^1(\Omega(\alpha)), H^1(\hat{\Omega}))$

$$y_n \rightsquigarrow y \iff \begin{cases} \tilde{y}_n \to \tilde{y} \text{ in } H^1(\hat{\Omega}) \\ \text{a} \\ \tilde{y}_n \to \tilde{y} \text{ in } H^1(\hat{\Omega}) \end{cases}$$

Question : $\|E_\alpha\| \le \tilde{c} \qquad ? \tilde{c} = \tilde{c}(\alpha) ?$

(9)

Can $\tilde{c}$ be estimated independently of $\alpha \in \tilde{\mathcal{U}}_{ad}$?

Yes, since $\Omega \in \tilde{\mathcal{O}}$ satisfies the uniform cone property $\Rightarrow$ uniform extension property

$$\|\tilde{u}_m\|_{1,\hat{\Omega}} \le \tilde{c}\,\|u_m\|_{1,\Omega_m} \le c \quad \forall m \in \mathbb{N}$$

$\Rightarrow \exists\, \{\tilde{u}_{m_\ell}\} \subset \{\tilde{u}_m\}$: $\tilde{u}_{m_\ell} \to \tilde{u}$ in $H^1(\hat{\Omega})$

and $u(\alpha) := \tilde{u}|_{\Omega(\alpha)}$ solves $(\mathcal{P}(\alpha))$.

$$\int_{\hat{\Omega}} \chi_{m_\ell} \nabla \tilde{u}_{m_\ell} \cdot \nabla \varphi \, dx = \int_{\hat{\Omega}} \chi_{m_\ell} f \varphi \, dx \qquad \forall \varphi \in H^1(\hat{\Omega}).$$

$$\downarrow$$

$$\int_{\hat{\Omega}} \chi \nabla \tilde{u} \cdot \nabla \varphi \, dx = \int_{\hat{\Omega}} \chi f \varphi \, dx \qquad \forall \varphi \in H^1(\hat{\Omega}).$$

Example 3 (linear elasticity problem)

$$V(\alpha) = \{ v = (v_1, v_2) \in (H^1(\Omega(\alpha)))^2 \mid v_i = 0 \text{ on } \partial\Omega(\alpha) \setminus \Gamma(\alpha),\ i = 1,2 \}.$$

(10)

$$(\mathcal{P}(\alpha)) \quad \left\{ \begin{array}{l} \text{Find } u(\alpha) \in V(\alpha) \text{ s.t.} \\ \displaystyle\int_{\Omega(\alpha)} C_{ijkl}\, \varepsilon_{ij}(u(\alpha))\, \varepsilon_{kl}(v)\, dx = \langle L, v \rangle_\alpha \quad \forall v \in V(\alpha) \end{array} \right.$$

$$\exists\, \gamma > 0 : \quad C_{ijkl}(x)\, \xi_{ij}\, \xi_{kl} \ge \gamma\, \xi_{ij}\, \xi_{ij} \qquad \text{a.e. in } \hat{\Omega}$$

$$\forall \xi_{ij} = \xi_{ji}$$

$$C_{ijkl} \in L^\infty(\hat{\Omega})$$

$v := u(\alpha)$

$$\tilde{\gamma} \|u_m\|^2_{1,\Omega_m} \le \int_{\Omega(\alpha_m)} C_{ijkl}\, \varepsilon_{ij}(v(\alpha_m))\, \varepsilon_{kl}(v(\alpha_m))\, dx$$

↑ Korn's inequality

Question : $\tilde{\gamma} = \tilde{\gamma}(\alpha)$?, $\alpha \in \tilde{\mathcal{U}}_{ad}$

Can $\tilde{\gamma}$ be estimated independently of $\alpha \in \tilde{\mathcal{U}}_{ad}$?

Yes if Ω possesses the uniform cone property.

Example 4 (unilateral problems)

$$\left\{ \begin{array}{l} -\Delta u(\alpha) = f \quad \text{in } \Omega(\alpha), \quad f \in L^2_{loc}(\mathbb{R}^2) \\ u(\alpha) = 0 \quad \text{on } \partial\Omega(\alpha) \setminus \overline{\Gamma}(\alpha) \\ u(\alpha) \ge 0,\ \dfrac{\partial u(\alpha)}{\partial \nu} \ge 0,\ u(\alpha)\dfrac{\partial u(\alpha)}{\partial \nu} = 0 \quad \text{on } \Gamma(\alpha) \end{array} \right.$$

(11)

$$K(\alpha) = \{ v \in H^1(\Omega(\alpha)) \mid v = 0 \text{ on } \partial\Omega(\alpha) \setminus \overline{\Gamma}(\alpha),\ v \ge 0 \text{ on } \Gamma(\alpha) \}$$

$$(\mathcal{P}(\alpha)) \quad \left\{ \begin{array}{l} \text{Find } u(\alpha) \in k(\alpha) \text{ s.t.} \\ \displaystyle\int_{\Omega(\alpha)} \nabla u(\alpha) \cdot \nabla (v - u(\alpha))\, dx \ge \int_{\Omega(\alpha)} f(v - u)\, dx \quad \forall v \in k(\alpha) \end{array} \right.$$

$v = 0, 2u$ into $(\mathcal{P}(\alpha)) \Rightarrow$

$$\int_{\Omega(\alpha)} |\nabla u(\alpha)|^2 dx = \int_{\Omega(\alpha)} f u(\alpha)\, dx$$

$$\Rightarrow \quad \|\tilde{u}(\alpha)\|_{1,\hat{\Omega}} \le c \qquad \forall \alpha \in \mathcal{U}_{ad},$$

where $\tilde{u}(\alpha) = E_\alpha(u(\alpha)) \in H^1(\hat{\Omega})$.

Let $\alpha_m \rightrightarrows \alpha$ in $[0,1]$, $\alpha_m, \alpha \in \tilde{\mathcal{U}}_{ad}$. Then there exists $\{\tilde{u}_{m_\ell}\} \subset \{\tilde{u}_m\}$ such that

$$\tilde{u}_{m_\ell} \to \tilde{u} \quad \text{in } H^1(\hat{\Omega}).$$

Does $u(\alpha) := \tilde{u}|_{\Omega(\alpha)}$ solve $(\mathcal{P}(\alpha))$?

Yes! Sketch of the proof

(12)

(i) $u(\alpha) \in k(\alpha) \iff u(\alpha) \in H^1(\Omega(\alpha))$, $u(\alpha) = 0$ on $\partial\Omega(\alpha) \setminus \overline{\Gamma}(\alpha)$ and $u(\alpha) \ge 0$ on $\Gamma(\alpha)$

Lemma Let $\alpha_m \rightrightarrows \alpha$ in $[0,1]$, $\alpha_m, \alpha \in \tilde{\mathcal{U}}_{ad}$ and $y_m \to y$ in $H^1(\hat{\Omega})$. Then

$$y_m|_{\Gamma(\alpha_m)} \circ \alpha_m \to y|_{\Gamma(\alpha)} \circ \alpha \quad \text{in } L^2((0,1))$$

Let $v \in k(\alpha)$ be given and $v^* \in H^1_0(\hat{\Omega})$ be such that $v^*|_{\Omega(\alpha)} = v$. Then there exists a sequence $\{v_j\}$, $v_j \in H^1_0(\hat{\Omega})$ such that

$$v_j \to v^* \quad \text{in } H^1_0(\hat{\Omega})$$

$$v_j|_{\Omega(\alpha_m)} \in k(\alpha_m) \quad \text{for } m \text{ large enough}$$

Construction of $\{v_j\}$

$v^* \ge 0$ on $\partial\hat{\Omega} \cup \Gamma(\alpha) \Rightarrow \exists\, \varphi \in H^1_0(\hat{\Omega})$, $\varphi \ge 0$ in $\hat{\Omega}$ such that $\varphi = v^*$ on $\partial\hat{\Omega} \cup \Gamma(\alpha)$

(13)

$v^* = \varphi + w$, $w|_{\Omega(\alpha)} \in H_0^1(\Omega(\alpha))$

$w|_{\tilde{\Sigma}(\alpha)} \in H_0^1(\Sigma(\alpha))$, $\tilde{\Sigma}(\alpha) = \hat{\Omega} \setminus \overline{\Omega}(\alpha)$

$\exists \{w_j\}$, $w_j|_{\Omega(\alpha)} \in C_0^\infty(\Omega(\alpha))$, $w_j|_{\tilde{\Sigma}(\alpha)} \in C_0^\infty(\Sigma(\alpha))$

such that $w_j \to w$ in $H_0^1(\hat{\Omega})$.

Set $v_j := \varphi + w_j$. Then $\{v_j\}$ has the required property.

Let $j \in N$ be fixed and $n \in N$ large enough.

$$\int_{\hat{\Omega}} \chi_n \nabla \tilde{u}_n \cdot \nabla (v_j - \nabla \tilde{u}_n)\, dx \geq \int_{\hat{\Omega}} \chi_n f (v_j - \tilde{u}_n)\, dx$$

$\downarrow n \to \infty$ $\qquad$ $\downarrow n \to \infty$

$$\int_{\hat{\Omega}} \chi \nabla \tilde{u} \cdot \nabla (v_j - \nabla \tilde{u})\, dx \geq \int_{\hat{\Omega}} \chi f (v_j - \tilde{u})\, dx$$

$\downarrow j \to \infty$ $\qquad$ $\downarrow j \to \infty$

$$\boxed{\int_{\hat{\Omega}} \chi \nabla \tilde{u} \cdot (\nabla v - \nabla \tilde{u})\, dx \geq \int_{\hat{\Omega}} \chi f (v - \tilde{u})\, dx}$$

(14)

Approximation of (P)

$\mathcal{O} \sim \mathcal{O}_\varepsilon \subset \tilde{\mathcal{O}}$ $\quad \forall \varepsilon \to 0+$;

$\Omega_\varepsilon \in \mathcal{O}_\varepsilon$ ... discrete design domain ;

(Boundaries realized by splines, e.g.)

$\Omega_\varepsilon \longmapsto \Omega_{\varepsilon h}$ ... Computational domain ;

$\Omega_\varepsilon \longleftrightarrow \Omega_{\varepsilon h}$

$\mathcal{O}_{\varepsilon h} = \{\Omega_{\varepsilon h}\}$, $\mathcal{O}_{\varepsilon h} \subset \tilde{\mathcal{O}}$

$(\mathbb{P}(\Omega_{\varepsilon h}))_h$ $\Omega_{\varepsilon h} \longmapsto u_h(\Omega_{\varepsilon h}) \in V_h(\Omega_{\varepsilon h})$ ... discretization of the state problem

$(\mathbb{P})_{\varepsilon h}$ $\quad$ Find $\Omega^*_{\varepsilon h} \in \mathcal{O}_{\varepsilon h}$ such that

$J(\Omega^*_{\varepsilon h}, u_h(\Omega^*_{\varepsilon h})) \leq J(\Omega_{\varepsilon h}, u_h(\Omega_{\varepsilon h}))$

$\forall\, \Omega_{\varepsilon h} \in \mathcal{O}_{\varepsilon h}$

$\Omega^*_{\varepsilon h}$ ... optimal computational domain ;

Ω^*_ε ... optimal discrete design domain

(15)

Convergence analysis

(B1) $\forall\, \Omega \in \mathcal{O}$ $\exists \{\Omega_\varepsilon\}$, $\Omega_\varepsilon \in \mathcal{O}_\varepsilon$:

$\Omega_\varepsilon \xrightarrow{\tilde{\sigma}} \Omega$, $\Omega_{\varepsilon h} \xrightarrow{\tilde{\sigma}} \Omega$;

(B2) for any $\{\Omega_\varepsilon\}$, $\Omega_\varepsilon \in \mathcal{O}_\varepsilon$ $\exists \{\Omega_{\varepsilon_j}\} \subset \{\Omega_\varepsilon\}$:

$\Omega_{\varepsilon_j} \xrightarrow{\tilde{\sigma}} \Omega \in \mathcal{O}$, $\Omega_{\varepsilon_j h_j} \xrightarrow{\tilde{\sigma}} \Omega$,

$u_{h_j}(\Omega_{\varepsilon_j h_j}) \rightharpoonup u(\Omega)$;

(B3) $\left.\begin{array}{l} \Omega_{\varepsilon h} \xrightarrow{\tilde{\sigma}} \Omega \,,\ \Omega_{\varepsilon h} \in \mathcal{O}_{\varepsilon h} \,,\ \Omega \in \mathcal{O} \\ u_h(\Omega_{\varepsilon h}) \rightharpoonup u(\Omega) \end{array}\right\} \Rightarrow$

$\Rightarrow \quad J(\Omega_{\varepsilon h}, u_h(\Omega_{\varepsilon h})) \to J(\Omega, u(\Omega))$.

Theorem. Let (B1) - (B3) be satisfied. Then for any sequence $\{(\Omega^*_\varepsilon, u_h(\Omega^*_\varepsilon))\}$ there $\exists \{(\Omega^*_{\varepsilon_j}, u_{h_j}(\Omega^*_{\varepsilon_j}))\}$ s.t.

$\Omega^*_{\varepsilon_j} \xrightarrow{\tilde{\sigma}} \Omega^*$

$u_{h_j}(\Omega^*_{\varepsilon_j}) \rightharpoonup u(\Omega^*)$, $j \to \infty$

and $(\Omega^*, u(\Omega^*))$ is an optimal pair of $(\mathbb{P})$.

(16)

Applications

$s_\varepsilon \in \mathcal{U}^{ad}_\varepsilon$

s_ε ... piecewise second order Bézier curve ;

$s_{\varepsilon h}$ ... linear Lagrange interpolation of s_ε ;

$\Omega(s_{\varepsilon h})$ .. computational domain ;

$\mathcal{T}(h, s_\varepsilon)$ ... triangulation of $\overline{\Omega}(s_{\varepsilon h})$;

$\overline{\Omega}(s_{\varepsilon h}) \longmapsto V_h(s_{\varepsilon h})$.

Assumptions on $\{\mathcal{T}(h, s_\varepsilon)\}$:

(a) for $h, \varepsilon > 0$ fixed , $\{\mathcal{T}(h, s_\varepsilon)\}$, $s_\varepsilon \in \mathcal{U}^{ad}_\varepsilon$ consists of topologically equivalent triangulations ;

(b) for $h, \varepsilon \to 0+$, $\{\mathcal{T}(h, s_\varepsilon)\}$ is uniformly regular with respect to h, ε and $s_\varepsilon \in \mathcal{U}^{ad}_\varepsilon$.

(17)

Example $\begin{cases} -\Delta u(x) = f & \text{in } \Omega(x) \\ u(x) = 0 & \text{on } \partial\Omega(x) \end{cases}$

$V_h(S_{xh}) = \{ v_h \in C(\overline{\Omega}(S_{xh})) \mid v_h|_T \in P_1(T) \ \forall T \in \mathcal{T}(h, S_{xh}),$
$v_h = 0 \text{ on } \partial\Omega(S_{xh}) \}$

$(\mathcal{P}(S_{xh}))_h$ $\begin{cases} \text{Find } u_h := u_h(S_{xh}) \in V_h(S_{xh}): \\ \int_{\Omega(S_{xh})} \nabla u_h \cdot \nabla v_h \, dx = \int_{\Omega(S_{xh})} f v_h \, dx \quad \forall v_h \in V_h(S_{xh}) \end{cases}$

$(\mathbb{P})_{xh}$ $\begin{cases} \text{Find } S_x^* \in \mathcal{U}_{ad}^x \text{ s.t.} \\ J(S_x^*, u_h^*) \le J(S_x, u_h) \quad \forall S_x \in \mathcal{U}_{ad}^x, \end{cases}$

where $u_h := u_h(S_{xh})$ solves $(\mathcal{P}(S_{xh}))_h$.

Algebraic form of $(\mathbb{P})_{xh}$

$\Omega_x \longleftrightarrow \vec{x}$ ... vector of discrete design variables

$\mathcal{O}_x \longleftrightarrow \mathcal{U} \subset \mathbb{R}^d$ (the space dimension d is the same for all $\Omega_x \in \mathcal{O}_x$).

(18)

$(\mathcal{P}(\vec{x}))$ $A(\vec{x})\vec{x}(\vec{x}) = \vec{f}(\vec{x})$

Topological equivalency of $\mathcal{T}(h, S_x)$, $S_x \in \mathcal{U}_{ad}^x \longrightarrow$
$\vec{x} \mapsto A(\vec{x})$, $\vec{x} \mapsto \vec{f}(\vec{x})$ are continuous

$(\vec{\mathbb{P}})$ $\begin{cases} \text{Find } \vec{x}^* \in \mathcal{U} \text{ s.t.} \\ J(\vec{x}^*, \vec{x}(\vec{x}^*)) \le J(\vec{x}, \vec{x}(\vec{x})) \quad \forall \vec{x} \in \mathcal{U}, \end{cases}$

where $\vec{x}(\vec{x})$ solves $(\mathcal{P}(\vec{x}))$ and $J(\vec{x}, \vec{x}) \longleftrightarrow J(S_x, u_h)$

$(\vec{\mathbb{P}})$ ... a non-linear mathematical programming problem.

Convergence analysis

Lemma (verification of B2) Let $S_x \rightrightarrows \alpha$ in $[0,1]$, $x \to 0+$, $S_x \in \mathcal{U}_{ad}^x$, $\alpha \in \mathcal{U}_{ad}$ and $u_h := u_h(S_x)$ be a solution of $(\mathcal{P}(S_{xh}))_h$. Then

$\tilde{u}_h \to u$ in $H_0^1(\hat{\Omega})$

and $u(x) := u|_{\Omega(\alpha)}$ solves $(\mathcal{P}(\alpha))$.

(19)

Proof: (i) $\{\tilde{u}_h\}$ is bounded in $H_0^1(\hat{\Omega})$

$\Rightarrow \exists \{\tilde{u}_{h_j}\} \subset \{\tilde{u}_h\}$ and $u \in H_0^1(\hat{\Omega})$:

$\tilde{u}_{h_j} \to u$ in $H_0^1(\hat{\Omega})$;

(ii) ? $u|_{\Omega(\alpha)}$ solves $(\mathcal{P}(\alpha))$? $u \equiv 0$ in $\hat{\Omega} \setminus \overline{\Omega(\alpha)}$

$\Rightarrow u|_{\Omega(\alpha)} \in H_0^1(\Omega(\alpha))$

Let $\varphi \in C_0^\infty(\Omega(\alpha))$ and φ_h be the piecewise linear Lagrange interpolate of $\tilde{\varphi}|_{\Omega(S_{xh})}$ on $\mathcal{T}(h, S_x)$. Then

- $\varphi_h \in V_h(S_{xh})$ for x, h small enough;
- $\|\tilde{\varphi}_h - \tilde{\varphi}\|_{1,\infty,\hat{\Omega}} = \|\varphi_h - \tilde{\varphi}\|_{1,\infty,\Omega_{xh}} \le ch \|\tilde{\varphi}\|_{C^2(\hat{\Omega})} \to 0+;$

$\int_{\hat{\Omega}} \chi_{S_{x_j h_j}} \nabla u_{h_j} \cdot \nabla \varphi_{h_j} \, dx = \int_{\hat{\Omega}} \chi_{S_{x_j}} f \varphi_{h_j} \, dx$

$\downarrow j \to \infty$

$\int_{\hat{\Omega}} \chi \nabla u \cdot \nabla \tilde{\varphi} \, dx = \int_{\hat{\Omega}} \chi f \tilde{\varphi} \, dx$

(20)

Theorem (convergence result) Let $\{(S_x^*, u_h^*)\}$ be a sequence of optimal pairs of $(\mathbb{P})_{xh}$, $x, h \to 0+$. Then one can find a subsequence $\{(S_{x_j}^*, u_{h_j}^*)\} \subset \{(S_x^*, u_h^*)\}$ such that

$(*)$ $\begin{cases} S_{x_j}^* \rightrightarrows \alpha^* & \text{in } [0,1] \\ \tilde{u}_{h_j}^* \to u^* & \text{in } H_0^1(\hat{\Omega}), \ j \to \infty \end{cases}$

and $(\alpha^*, u^*|_{\Omega(\alpha^*)})$ is an optimal pair of $(\mathbb{P})$.

Any accumulation point of $\{(S_x^*, u_h^*)\}$ in the sense of $(*)$ possesses this property.

Sizing optimization

V... a real Hilbert space, V'... dual of V;

U... a Banach space;

$\mathcal{U}_{ad} \subset U$ ... a compact subset of U (admissible controls)

$\forall e \in \mathcal{U}_{ad}$ $e \mapsto a_e : V \times V \longmapsto \mathbb{R}^1$

The system $\{a_e\}$, $e \in \mathcal{U}_{ad}$ satisfies:

(21)

(A1) $\exists M>0: |a_e(y,v)| \le M\|y\|\|v\| \quad \forall y,v\in V, \forall e\in \mathcal{U}_{ad}$;

(A2) $\exists \alpha>0: a_e(v,v) \ge \alpha\|v\|^2 \quad \forall v\in V, \forall e\in \mathcal{U}_{ad}$;

State problem:

$e\in \mathcal{U}_{ad}$, $f\in V'$

$(\mathcal{P}(e))$ $\begin{cases} \text{Find } u(e)\in V \text{ s.t.} \\ a_e(u(e),v) = \langle f, v\rangle \quad \forall v\in V \end{cases}$

Cost functional:

$J: \mathcal{U}_{ad}\times V \longmapsto \mathbb{R}^1$

Abstract sizing optimization problem

$(\mathbb{P})$ $\begin{cases} \text{Find } e^*\in \mathcal{U}_{ad} \text{ s.t.} \\ J(e^*,u(e^*)) \le J(e,u(e)) \quad \forall e\in \mathcal{U}_{ad}, \end{cases}$

where $u(e)\in V$ solves $(\mathcal{P}(e))$.

Existence analysis for $(\mathbb{P})$

(A3) $e_n \to e$ in U, $e_n, e\in \mathcal{U}_{ad} \Rightarrow$

$\sup_{\substack{\|y\|\le 1 \\ \|v\|\le 1}} |a_{e_n}(y,v) - a_e(y,v)| \to 0$

(22)

(A4) $y_n \to y$ in V, $e_n \to e$ in U, $e_n, e\in \mathcal{U}_{ad} \Rightarrow$

$\liminf_{n\to\infty} J(e_n, y_n) \ge J(e,y)$.

Equivalent expression of (A3):

$A(e)\in \mathcal{L}(V,V')$: $\langle A(e)y, v\rangle = a_e(y,v) \quad \forall y,v\in V$.

Then (A3) $\Leftrightarrow e_n\to e$ in $U \Rightarrow A(e_n)\to A(e)$ in $\mathcal{L}(V,V')$.

Lemma (Continuous dependence of solutions on e)

Let $e_n\to e$ in U and $u_n := u(e_n)$ be a solution of $(\mathcal{P}(e_n))$, $n\to\infty$. Then

$u_n \to u(e)$ in V, $n\to\infty$

and $u(e)$ solves $(\mathcal{P}(e))$.

Proof: (A1)+(A2) $\Rightarrow$ $\{u_n\}$ is bounded in V $\Rightarrow$

$\exists \{u_{n_j}\}\subset\{u_n\}$: $u_{n_j} \to u$ in V

$a_{e_{n_j}}(u_{n_j}, v) = \langle f, v\rangle \quad \forall v\in V$.

Letting $j\to\infty$ and using (A3) we obtain:

(23)

$a_e(u,v) = \langle f, v\rangle \quad \forall v\in V \Rightarrow$

$\Rightarrow u := u(e)$ solves $(\mathcal{P}(e))$.

Strong convergence:

$\alpha\|u_n - u\|^2 \le \langle A(e_n)(u_n - u), u_n - u\rangle =$

$= \langle A(e_n)u_n, u_n - u\rangle - \langle A(e_n)u, u_n - u\rangle$

$= \langle f, u_n - u\rangle - \langle A(e_n)u, u_n - u\rangle \to 0, \quad n\to\infty.$

Theorem (existence) Let $\mathcal{U}_{ad}\subset U$ be a compact subset of U and (A1)-(A4) be satisfied. Then $(\mathbb{P})$ has a solution.

Applications

Example (thickness optimization of a beam).

$(\mathcal{P}(e))$ $\begin{cases} (\beta e^3 w'')''(x) = f(x), \quad x\in(0,\ell), \ f\in L^2((0,\ell)), \\ w(0) = w'(0) = w(\ell) = w'(\ell) = 0 \end{cases}$

$\mathcal{U}_{ad} = \{ e\in C^{0,1}([0,\ell]) \mid 0 < e_{min} \le e \le e_{max} \text{ in } [0,\ell],$

$|e'(x)|\le L_0 \text{ a.e. in } (0,\ell),$

$\int_0^\ell e(t)\,dx = g \}.$

(24)

$J(y) = \int_0^\ell fy\,dx$ ... compliance

$(\mathbb{P})$ $\begin{cases} \text{Find } e^*\in \mathcal{U}_{ad} \text{ s.t.} \\ J(u(e^*)) \le J(u(e)) \quad \forall e\in \mathcal{U}_{ad}, \end{cases}$

where $u(e)$ solves $(\mathcal{P}(e))$.

Theorem $(\mathbb{P})$ has a solution.

Proof: $\mathcal{U}_{ad}$ is a compact subset of $U = C([0,\ell])$,

$a_e(y,v) = \int_0^\ell \beta e^3 y''v''\,dx, \quad y,v\in H_0^2((0,\ell))$.

System $\{a_e\}$, $e\in\mathcal{U}_{ad}$ satisfies (A1)-(A3) and J is continuous.

Discretization and convergence analysis.

$\{V_h\}$, $V_h\subset V$, $\dim V_h = n(h)\to\infty$, $h\to 0+$;

$U\subseteq \tilde{U}$, $\{U_h\}$, $U_h\subset\tilde{U}$, $\dim U_h = m(h)\to\infty$, $h\to 0+$;

$\mathcal{U}_h^{ad}\subset U_h$ ... compact subsets of U_h.

The system $\{a_e\}$ is defined also for $e\in \bigcup_{h>0}\mathcal{U}_h^{ad}$.

(25)

Let $e_h \in \mathcal{U}_h^{ad}$ be given.

$(\mathcal{P}(e_h))_h$ $\begin{cases} \text{Find } u_h(e_h) \in V_h \text{ s.t.} \\ a_{e_h(u_h(e_h)),v_h}) = \langle f, v_h \rangle \quad \forall v_h \in V_h \end{cases}$

$(\mathcal{P})_h$ $\begin{cases} \text{Find } e_h^* \in \mathcal{U}_h^{ad} \text{ s.t.} \\ J(e_h^*, u_h(e_h^*)) \le J(e_h, u_h(e_h)) \quad \forall e_h \in \mathcal{U}_h^{ad}, \end{cases}$

where $u_h(e_h) \in V_h$ solves $(\mathcal{P}(e))_h$.

Convergence analysis

$(A1)_h$ $\exists \tilde{M} > 0 : |a_{e_h}(v,w)| \le M \|v\| \|w\| \quad \forall v, w \in V, e_h \in \bigcup_{h>0} \mathcal{U}_h^{ad}$;

$(A2)_h$ $\exists \tilde{\alpha} > 0 : a_{e_h}(v,v) \ge \tilde{\alpha} \|v\|^2 \quad \forall v \in V;\ \forall e_h \in \bigcup_{h>0} \mathcal{U}_h^{ad}$;

$(A3)_h$ $e_h \to e$ in $\tilde{U}$, $e_h \in \mathcal{U}_h^{ad}$, $e \in \mathcal{U}_{ad} \Rightarrow A(e_h) \to A(e)$ in $\mathcal{L}(V, V')$;

$(A4)_h$ $\forall v \in V \quad \exists \{v_h\}, v_h \in V_h : \ v_h \to v$ in V;

$(A5)_h$ $\forall e \in \mathcal{U}_{ad} \quad \exists \{e_h\}, e_h \in \mathcal{U}_h^{ad} : \ e_h \to e$ in $\tilde{U}$;

$(A6)_h$ for any $\{e_h\}$, $e_h \in \mathcal{U}_h^{ad}$ $\exists \{e_{h_j}\} \subset \{e_h\}$ and $e \in \mathcal{U}_{ad}$: $e_{h_j} \to e$ in $\tilde{U}$.

(26)

$(A7)_h$ $e_h \to e$ in U, $y_h \to y$ in V, $e_h \in \mathcal{U}_h^{ad}$, $e \in \mathcal{U}_{ad}$, $y_h \in V_h$, $y \in V \Rightarrow \lim_{h \to 0+} J(e_h, y_h) = J(e, y)$

Lemma Let $(A1)_h$–$(A4)_h$ be satisfied and $\{e_h\}$, $e_h \in \mathcal{U}_h^{ad}$ be such that $e_h \to e$ in $\tilde{U}$, $h \to 0+$. Then

$$u_h(e_h) \to u(e) \text{ in } V, \ h \to 0+,$$

and $u(e)$ solves $(\mathcal{P}(e))$.

Theorem Let $(A1)_h$–$(A7)_h$ be satisfied. Then for any sequence $\{(e_h^*, u_h(e_h^*))\}$ of optimal pairs of $(\mathcal{P})_h$, $h \to 0+$ $\exists \{(e_{h_j}^*, u_{h_j}(e_{h_j}^*))\} \subset \{(e_h^*, u_h(e_h^*))\}$ s.t.

$$(*) \qquad \begin{cases} e_{h_j}^* \to e^* \text{ in } \tilde{U}, \\ u_{h_j}(e_{h_j}^*) \to u(e^*) \text{ in } V, \ j \to \infty \end{cases}$$

and $(e^*, u(e^*))$ is an optimal pair of $(\mathcal{P})$. Any accumulation point of $\{(e_h^*, u_h(e_h^*))\}$ in the sense of $(*)$ possesses this property.

(27)

Applications

$\Delta_h : 0 = a_0 < a_1 < \dots < a_{d(h)} = l$ .. equidistant partition of $[0,l]$

$V_h = \{v_h \in C^1([0,l]) \mid v_h|_{\overline{a_{i-1}, a_i}} \in P_3\} \cap H_0^2((0,l))$

$\mathcal{U}_h^{ad} = \{e_h \in L^\infty((0,l)) \mid e_h|_{\overline{a_{i-1}, a_i}} \in P_0, \ e_{min} \le e_h \le e_{max} \text{ in } [0,l], \ \int_0^l e_h \, dx = \mu, \ |e_h^{i+1} - e_h^i| \le L_0 h\}$, $e_h^i \equiv e_h|_{[a_{i-1}, a_i]}$

For $e_h \in \mathcal{U}_h^{ad}$ define:

$(\mathcal{P}(e_h))_h$ $\begin{cases} \text{Find } u_h(e_h) \in V_h \text{ s.t.} \\ \int_0^l \beta e_h^3 u_h'(e_h) v_h' \, dx = \int_0^l f v_h \, dx \quad \forall v_h \in V_h \end{cases}$

$(\mathcal{P})_h$ $\begin{cases} \text{Find } e_h^* \in \mathcal{U}_h^{ad} \text{ s.t.} \\ J(u_h(e_h^*)) \le J(u_h(e_h)) \quad \forall e_h \in \mathcal{U}_h^{ad} \end{cases}$

All the assumptions $(A1)_h$–$(A7)_h$ are satisfied.

What Is the Role of the Worst Scenario Method in Solving Problems with Uncertain Input Data ?

J. Chleboun

Czech Technical University in Prague

1 Introduction

The worst scenario method is inspired by one of the leading principles of safe design: to be on the safe side even if the design and, more adequately for the purposes of our lecture, its mathematical (computational) model are burdened with uncertainty. The amount of uncertainty in the model behavior has to be analyzed to exclude or admit a possible violation of the safe side policy.

In other words, if input data of a mathematical model is uncertain, then model output data is uncertain too. To evaluate the uncertainty of outputs, their extremal values that can appear due to the uncertain inputs have to be identified, which is usually done through identifying the particular inputs that are responsible for the extremal output values. This is the key idea of the worst (case) scenario method.

Although such an idea is not new, its applications to ODE- or PDE-driven problems does not seem to be common, especially if the uncertainty is not limited to scalar parameters but also burdens the functions that appear in differential equations as input data.

By knowing the extremes that bounds the behavior of a mathematical model, an analyst can be more confident in making decisions. In practice, however, the knowledge of mere extremes may not be particularly important because the inputs are often weighted, but the worst scenario method does not take the weights into consideration. The most notable methods that deal with weighted uncertainty are stochastic methods. Although coupling the worst scenario idea with stochastic approaches is possible, we will not elaborate on it here. Instead, we will focus on two less common weighting approaches and we will show that to analyze the propagation of weighted uncertainty from model inputs to model outputs, we have to resort to the worst scenario method as a tool for obtaining the weight of outputs.

2 Mathematical Framework

Let us consider the following abstract problem (state problem): Find $u(a) \in V$ such that

$$A(a; u(a)) = f, \tag{2.1}$$

where A is a differential operator dependent on a parameter a (consequently, the solution $u(a)$ is also a-dependent, as indicated by the notation), f stands for a right-hand side function, and V is the relevant space of functions. Instead of (2.1), one can imagine an a-dependent elliptic boundary value problem characterized by A, an operator, and $V \subset H$ where H is the relevant Sobolev space.

The parameter a belongs to $\mathcal{U}_{\mathsf{ad}}$, the set of admissible parameters. This set represents the amount and the character of uncertainty that accompanies a. It is assumed that problem (2.1) is uniquely solvable for any $\in \mathcal{U}_{\mathsf{ad}}$.

Let the state solution $u(a)$ be evaluated through a functional $\Phi(a, u(a))$. By virtue of the uniqueness of $u(a)$, we can define

$$\Psi(a) = \Phi(a, u(a)), \tag{2.2}$$

the criterion-functional (also called the quantity of interest) that defines a direct link between a particular value of the uncertain parameter and the feature of the state solution that is represented through Φ. Again, one can imagine, for example, an a-dependent elasticity problem whose solution (a displacement field) is "processed" by Ψ to deliver numeral information the analyst is interested in.

In the worst scenario method, we are searching for $a_0 \in \mathcal{U}_{\mathsf{ad}}$ such that

$$a_0 = \arg\min_{a \in \mathcal{U}_{\mathsf{ad}}} \Psi(a). \tag{2.3}$$

A slight modification of (2.3) leads to the other extreme

$$a^0 = \arg\max_{a \in \mathcal{U}_{\mathsf{ad}}} \Psi(a). \tag{2.4}$$

The compactness of $\mathcal{U}_{\mathsf{ad}}$ and the continuity of Ψ are sufficient for obtaining a_0 and a^0; a more detailed analysis can be found in [4]. It is assumed that the image of $\mathcal{U}_{\mathsf{ad}}$ under the map Ψ is an interval.

An approximation of (2.3) and (2.4) is necessary to numerically solve the respective problems. To this end, $\mathcal{U}_{\mathsf{ad}}$ is approximated by $\mathcal{U}_{\mathsf{ad}}^N$, a set identifiable with a compact subset of $\mathbb{R}^N$. If $\mathcal{U}_{\mathsf{ad}}$ comprises functions (which is the case we focus on), their finite-dimensional approximation is necessary. Also, the state problem is approximated by a proper method; take for instance the finite element method, the boundary element method, etc. As a consequence, problems

$$a_{0,N} = \arg\min_{a_N \in \mathcal{U}_{\mathsf{ad}}^N} \Psi_h(a_N) \text{ and } a^{0,N} = \arg\max_{a_N \in \mathcal{U}_{\mathsf{ad}}^N} \Psi_h(a_N) \tag{2.5}$$

are solved instead of (2.3)-(2.4). In (2.5), $\Psi_h(a_N) = \Phi(a_N, u_h(a_N))$ and u_h is the approximate state solution.

The relevant convergence issues are addressed in [4] and [3].

3 Weighting the Inputs

Let us sketch two non-stochastic approaches to weighting the input values.

In fuzzy set theory, a membership function μ is defined to indicate the weight of the elements of $\mathcal{U}_{\mathsf{ad}}$, $\mu : \mathcal{U}_{\mathsf{ad}} \to [0, 1]$; see [1], [6]. The goal of the uncertainty propagation analysis is to infer μ_Ψ, the membership function of

$$I_\Psi = [\Psi(a_0),\ \Psi(a^0)],$$

the interval of the quantity of interest induced by $\mathcal{U}_{\mathsf{ad}}$. It turns out, that μ_Ψ can be obtained through solving a sequence of (2.3)- and (2.4)-like problems where $\mathcal{U}_{\mathsf{ad}}$ is replaced by

$${}^\alpha\mathcal{U}_{\mathsf{ad}} = \{a \in \mathcal{U}_{\mathsf{ad}} :\ \mu(a) \geq \alpha\} \quad \alpha \in (0, 1].$$

to obtain related scenarios ${}^\alpha a_0$ and ${}^\alpha a^0$ as well as intervals

$${}^\alpha I_\Psi = [\Psi({}^\alpha a_0),\ \Psi({}^\alpha a^0)].$$

Then

$$\mu_\Psi(x) = \max\{\alpha \in [0,1] : \ x \in {}^{\alpha}I_\Psi\}.$$

Inspired by the Dempster-Shafer theory [2], [5], the other approach assumes a finite family of admissible sets that are weighted in such a way that the sum of the weights equals one; these sets are called focal elements, see [1]. For any other set of input values (that is, unweighted), two values, Bel and Pl, are calculated from the focal elements and their weights. These values give a lower and an upper bound on the likelihood of the set (in other words, they indicate the relevance of the set to the information included in the focal elements). The goal is to identify the focal elements in the values of the quantity of interest. It is obvious that the output focal elements are the images of the input focal elements under the map Ψ and that to obtain them, problems like (2.3) and (2.4) have to be solved. The output focal elements and their weights allow for calculating Bel and Pl of sets of possible output values (i.e., quantity of interest values).

4 Conclusion

The worst scenario method can be used as such without coupling with other methods. However, it seems to be more useful as a part of a method that weights data. In that case, the uncertainty analysis delivers more information but asks for the repeated solving of the worst scenario problems, which is computationally demanding. Moreover, the method leads to solving global optimization problems which is also a challenging task.

Acknowledgement: This work was supported by the Ministry of Education, Youth, and Sports of the Czech Republic through contract MSM 6840770003 and through grant No. IAA100190803 from the Grant Agency of AS CR.

References

[1] A. Bernardini: *What are the random and fuzzy sets and how to use them for uncertainty modelling in engineering systems?* In: I. Elishakoff (ed.): Whys nad Hows in Uncertainty Modelling. Probability, Fuzziness, and Anti-Optimization, CISM Courses and Lectures No. 388, Springer-Verlag, Wien, 1999.

[2] A. P. Dempster: *Upper and lower probabilities induced by a multivalued mapping.* Ann. Math. Stat. 38 (1967), pp. 325–339.

[3] P. Harasim: *On the worst scenario method: a modified convergence theorem and its application to an uncertain differential equation.* Appl. Math. 53 (2008), pp. 583–598.

[4] I. Hlaváček, J. Chleboun, I. Babuška: *Uncertain Input Data Problems and the Worst Scenario Method.* Elsevier, Amsterdam, 2004.

[5] G. Shafer: *A Mathematical Theory of Evidece.* Princeton University Press, Princeton, 1976.

[6] H. J. Zimmermann: *Fuzzy Set Theory — and its Applications*, fourth ed. Kluwer Academic Publishers, Boston, 2001.

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

Uncertainty in Engineering Problems Described by Fuzzy Sets

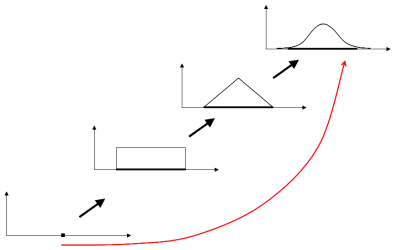
Jaroslav Kruis, Petr Štemberk

Outline

- Uncertainties
- Probability and Mathematical Statistic
- Fuzzy Sets
- Selected problems
- Parallelization

Uncertainties

uncertainty	x	random variable
theory of fuzzy sets	x	theory of probability



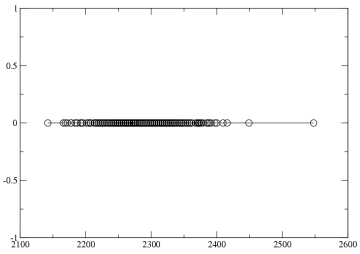
Uncertainties in Mechanics

- geometry (sizes of beams, thickness of plates)
- material parameters (Young's modulus of elasticity, density, yield stress, fracture energy)
- load (magnitude, orientation, time changes)

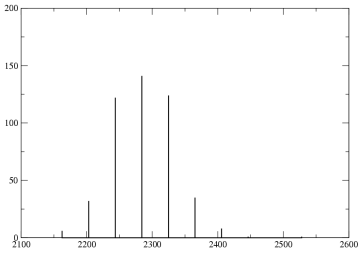
Measured Data

density of concrete measured in concreting plant during one year (kg/m^3)

number of values	470	465
minimum value	2142,22	2165,59
average value	2285,78	2284,63
maximum value	2547,73	2400,00
standard deviation	48,19	45,03
median	2284,44	2281,48
mode	2332,81	2388,79

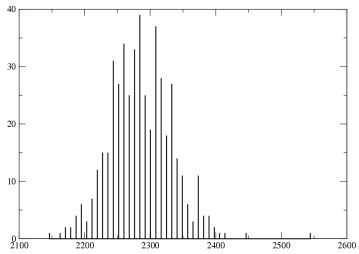


CTU Uncertainties and Fuzzy Sets J. Kruis



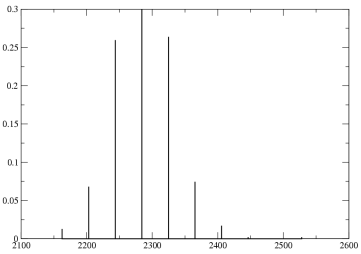
2. - 6. 2. 2009 SNA09, Institute of Geonics ASCR Ostrava

CTU Uncertainties and Fuzzy Sets J. Kruis



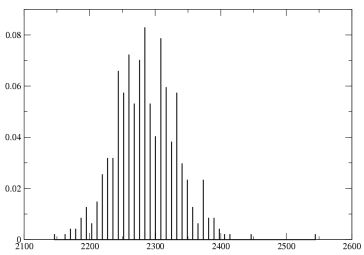
2. - 6. 2. 2009 SNA09, Institute of Geonics ASCR Ostrava

CTU Uncertainties and Fuzzy Sets J. Kruis



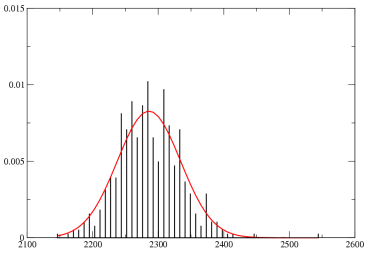
2. - 6. 2. 2009 SNA09, Institute of Geonics ASCR Ostrava

CTU Uncertainties and Fuzzy Sets J. Kruis



2. - 6. 2. 2009 SNA09, Institute of Geonics ASCR Ostrava

CTU Uncertainties and Fuzzy Sets J. Kruis



2. - 6. 2. 2009 SNA09, Institute of Geonics ASCR Ostrava

χ^2 goodness-of-fit test

i	$n\Delta\Phi(x_i)$		$n\Delta\Phi(x_i)$
1	5,983789		5,730882
2	36,141110		17,215111
3	107,544974		39,470382
4	157,667482		69,072631
5	113,882642		92,260082
6	40,526241		94,057806
7	7,105236		73,189540
8	0,613739	< 5	43,468671
9	0,026119	< 5	19,705020
10	0,000520	< 5	6,645456

removed element	number of removed elements	number of wrong intervals	χ^2_0 value
none	0	3	1927,877841
2547,73	1	2	3,488365
2449,04	2	2	6,789602
2415,87	3	1	13,933338
2142,22	4	1	13,265698
2409,52	5	0	7,507428

number of degrees of freedom $\nu = 10 - 2 - 1 = 7$

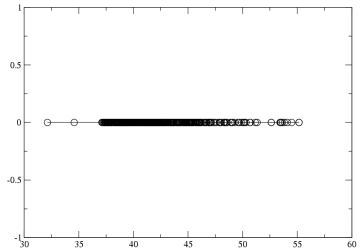
(10 intervals, 2 parameters of distribution),

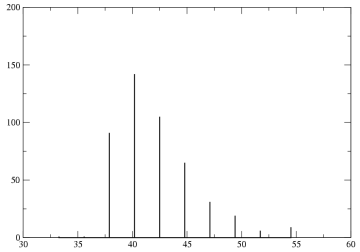
p_2	0,1	0,05	0,01	0,005	0,001
$\chi^2_{p_2}$	12,2	14,07	18,48	20,28	24,32

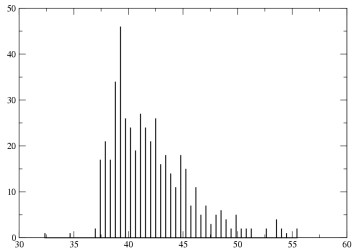
number of removed elements	χ^2_0 value	critical value $< \chi^2_{p_2}$	significance level
5	7,507428	12,2	0,1

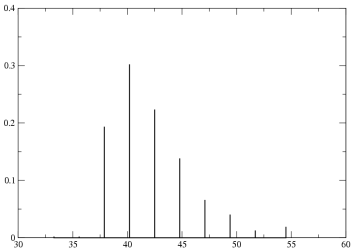
Strength of concrete during one year (MPa).

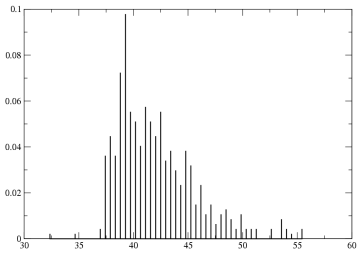
number of values	470
minimum value	32,13
average value	42,074
maximum value	55,16
standard deviation	3,6466
median	41,38
mode	42,49











Classical Set Theory - Crisp Sets

Cantor:

A set is a combination of particular, well-distinguishable objects, which are called elements, to an ensemble.

An element x belongs to a set A: $x \in A$.

A set is an ensemble of elements with property $V(x)$, the set is denoted by $\{x; V(x)\}$.

membership function expresses the degree of truth of the statement $x \in A$.

$\mu_A(x) = 1 \Leftrightarrow x \in A$

$\mu_A(x) = 0 \Leftrightarrow x \notin A$

The functional values of μ_A are 0 or 1.

The classical set theory is too “strict”.

L.A. Zadeh: Outline of a New Approach to the Analysis of Complex Systems and Decision Processes. IEEE Trans. Syst. Man. Cybern., 1, 1973.

Fuzzy Sets

Introduction of the fuzzy sets:

L.A. Zadeh: Fuzzy Sets. Information and Control, 8, p. 338-353, 1965.

Fuzzy Set: 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.

L is a set of real numbers, in most cases it is an interval $\langle 0; 1 \rangle$

$\mu_A(x) = 0$ 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
$\mathrm{supp}\ A = \{x; \mu_A(x) > 0\}$.

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 $\ker A \neq \emptyset$.

α-cut of fuzzy set A, where $\alpha \in L$, is a crisp set
$A_\alpha = \{x; \mu_A(x) \geq \alpha\}$.

α-level of fuzzy set A, where $\alpha \in L$, is a crisp set
$A^\alpha = \{x; \mu_A(x) = \alpha\}$.

If $\alpha_1, \alpha_2 \in L$ and $\alpha_1 \leq \alpha_2$, then $A_{\alpha_2} \subseteq A_{\alpha_1}$

Height of the fuzzy set A is

$$\mathrm{hgt}\, A = \max_{x \in U} \mu_A(x)$$

If $\ker A \neq \emptyset$, then $\mathrm{hgt}\, A = 1$.

The union of the fuzzy sets A and B is a set $C = A \cup B$ with membership function

$$\mu_C(x) = \max_{x \in U} \{\mu_A(x); \mu_B(x)\}\ .$$

The intersection of the fuzzy sets A and B is a set $C = A \cap B$ with membership function

$$\mu_C(x) = \min_{x \in U} \{\mu_A(x); \mu_B(x)\}\ .$$

The complement of the fuzzy set A is a set $\bar{A} = U - A$ with membership function

$$\mu_{\bar{A}}(x) = 1 - \mu_A(x)\ .$$

A fuzzy set $A \subset U$ is called convex if all α-cuts are convex sets, i.e.

for all $x, y \in A$ and a number $0 \leq \lambda \leq 1$ the following relationship $\lambda x + (1 - \lambda) y \in A$ holds.

A fuzzy set $A \subset U$ is convex if for all $x, y \in U$ and a number $0 \leq \lambda \leq 1$ means $\mu_A(\lambda x + (1 - \lambda) y) \geq \min\{\mu_A(x); \mu_A(y)\}$

Extension Principle

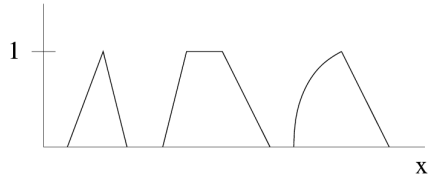
The extension principle represents the mathematical basis for the mapping of fuzzy sets into a result space.

Extension Principle. Let U and V be fundamental sets, $f : U \to V$ be a mapping and $A \subset U$ be a subset of the fundamental set U. The mapping f leads to the fuzzy set $f(A) \subset V$ with the membership function

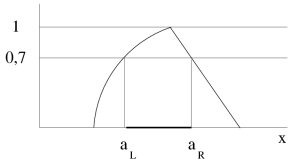
$$\mu_{f(A)}(y) = \begin{cases} \max\limits_{x \in f^{-1}(y)} \mu_A(x) \\ 0, \quad \text{if } : f^{-1}(y) = \emptyset \end{cases}$$

for all $y \in V$.

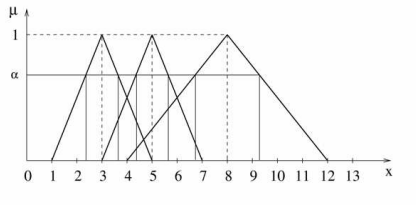
Fuzzy Numbers



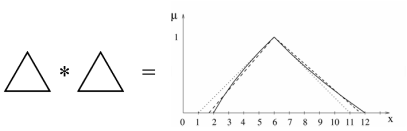
α-cut method



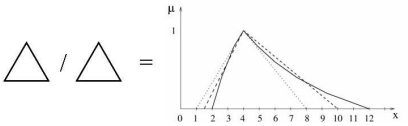
Addition of Fuzzy Numbers



Multiplication of Fuzzy Numbers



Division of Fuzzy Numbers



Vibration influenced by uncertainties

Undamped vibration of a single degree of freedom is assumed for simplicity

equation of motion

$$m\frac{\mathrm{d}^2w(t)}{\mathrm{d}t^2} + kw(t) = 0$$

$m > 0$ denotes the weight of mass and $k > 0$ denotes the stiffness of spring

natural circular frequency

$$\omega_0 = \sqrt{\frac{k}{m}}$$

equation of motion

$$\frac{\mathrm{d}^2w(t)}{\mathrm{d}t^2} + \omega_0^2 w(t) = 0$$

solution

$$w(t) = w_A \sin(\omega_o t + \phi)$$

w_A denotes the amplitude of vibration, ϕ denotes the phase angle

initial conditions

$$w(0) = d\,, \qquad \frac{\mathrm{d}\,w(0)}{\mathrm{d}t} = v$$

d denotes the initial displacement and v denotes the initial velocity.

Case n. 1.

the weight, stiffness and initial velocity are crisp numbers while the initial displacement is a fuzzy number

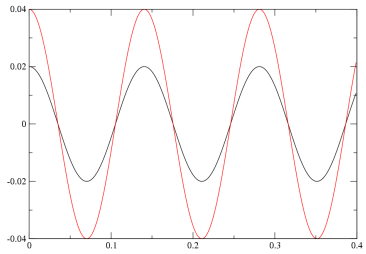
$$m > 0\,, \quad k > 0\,, \quad v = 0\,, \quad d_f = [d^-, d^+]$$

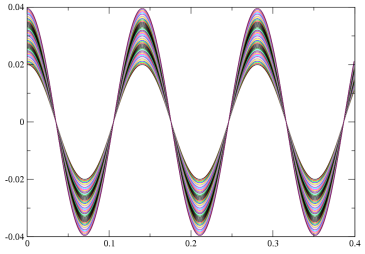
solution of the equation of motion has the form

$$w(t) = d_f \sin(\omega_0 t + \frac{\pi}{2})$$

numerical example

$$\begin{aligned} m &= 20 \\ k &= 40000 \\ d^- &= 0.02 \quad d^+ = 0.04 \\ v &= 0 \\ \phi &= \frac{\pi}{2} \end{aligned}$$





Case n. 2.

the weight, stiffness and initial displacement are crisp numbers while the initial velocity is a fuzzy number

$$m > 0\,, \quad k > 0\,, \quad d = 0\,, \quad v_f = [v^-, v^+]$$

solution of the equation of motion has the form

$$w(t) = \frac{v_f}{\omega_0} \sin \omega_0 t$$

Case n. 3.

the weight, initial displacement and initial velocity are crisp numbers while the stiffness is a fuzzy number

$$m > 0\,, \quad d \neq 0\,, \quad v = 0\,, \quad k_f = [k^-, k^+]$$

natural circular frequency is a fuzzy number in the form

$$\omega_{0f} = \sqrt{k_f/m} \quad \Rightarrow \quad \omega_0^- = \sqrt{k^-/m}, \quad \omega_0^+ = \sqrt{k^+/m}$$

solution of the equation of motion has the form

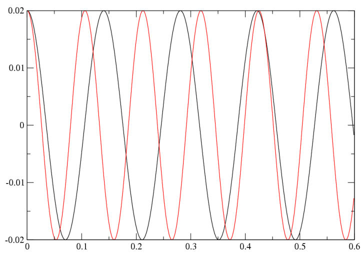
$$w(t) = d \sin(\omega_{0f} t + \frac{\pi}{2})$$

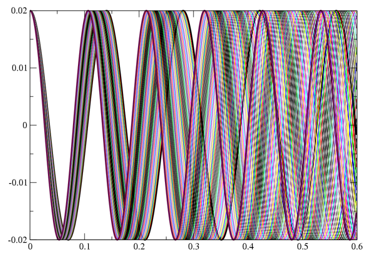
$$d\sin(\omega_0^-(t_f+T)+\frac{\pi}{2}) = d\sin(\omega_0^+ t_f+\frac{\pi}{2})$$

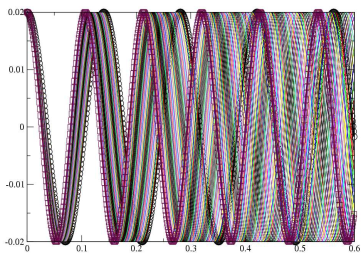
$$2\pi = (\omega_0^+ - \omega_0^-)t_f \quad \Rightarrow \quad t_f = \frac{2\pi}{\omega_0^+ - \omega_0^-}$$

numerical example

$$\begin{aligned} m &= 20 \\ k^- &= 40000 \quad k^+ = 70000 \\ d &= 0.02 \\ v &= 0 \\ \phi &= \frac{\pi}{2} \\ t_f &= 0.435141 \end{aligned}$$







Vibration of Plane Frame Structure

Description of the plane frame structure

height	16 m
width	10 m
columns	0,5 x 0,5 m
beams	0,5 x 0,5 m
Young modulus of elasticity	30 GPa $\pm$ 10%
density of concrete	2500 kg/m^3 $\pm$ 10%

equation of motion of free vibration

$$(\boldsymbol{K} - \omega_0^2 \boldsymbol{M})\boldsymbol{v} = \boldsymbol{0}$$

$\boldsymbol{K}$ stiffness matrix

$\boldsymbol{M}$ mass matrix

ω_0 natural circular frequency

$\boldsymbol{v}$ eigenvector (mode shape)

subspace iteration with Gram-Schmidt orthonormalization

Response Surface Function

$\tilde{\boldsymbol{X}}$ space of input data

$\tilde{\boldsymbol{Y}}$ space of output data

$\boldsymbol{X}$ m-dimensional space of input data

$\boldsymbol{Y}$ n-dimensional space of output data

response of system

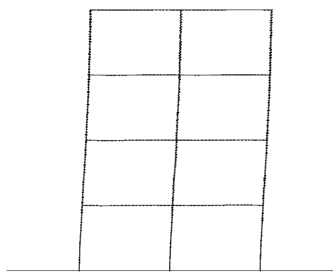
$$\tilde{\boldsymbol{y}} = \tilde{\mathcal{F}}(\tilde{\boldsymbol{x}})$$

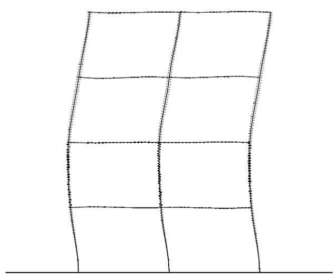
$$\boldsymbol{y} = \tilde{\boldsymbol{F}}(\boldsymbol{x})$$

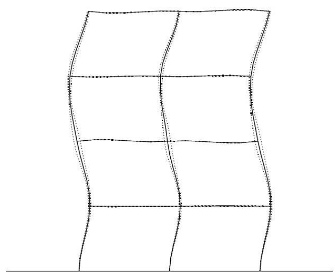
$$y_k^{[l]} = \tilde{\boldsymbol{F}}_k(\boldsymbol{x}^{[l]})$$

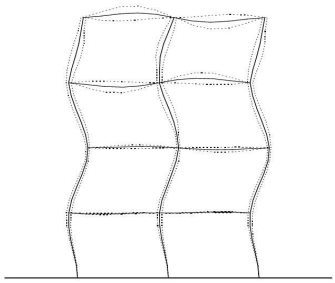
$$f^{(k)} = a^{(k)} + \sum_{i=1}^{i=m} b_i^{(k)} x_i + \sum_{i=1}^{i=m} \sum_{j=1}^{j=m} c_{ij}^{(k)} x_i x_j$$

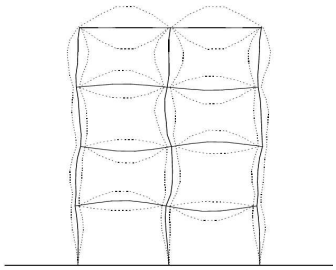
$$F^{(k)}(a^{(k)}, b_i^{(k)}, c_{ij}^{(k)}) = \sum_{l=1}^{l=s} (f^{(k)}(\boldsymbol{x}^{[l]}) - y_k^{[l]})^2$$

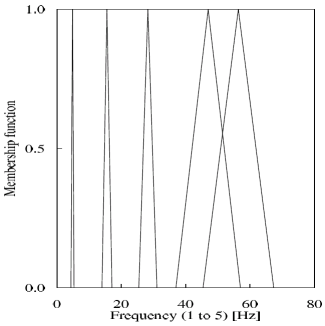


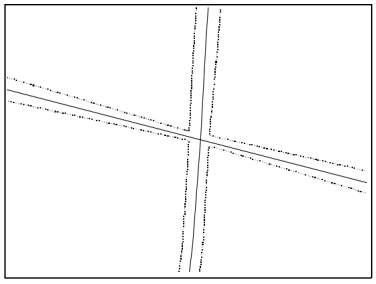


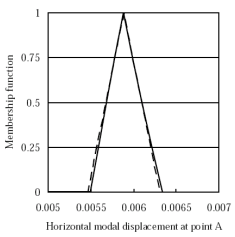


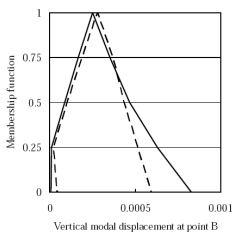












Vibration of Plane Frame Structure Caused by Earthquake

equation of motion

$$\boldsymbol{M}\ddot{\boldsymbol{d}} + \boldsymbol{C}\dot{\boldsymbol{d}} + \boldsymbol{K}\boldsymbol{d} = \boldsymbol{f}$$

expansion into eigenvectors

$$\boldsymbol{d} = \boldsymbol{V}\boldsymbol{u}$$

equation of motion in the case of earthquake loading

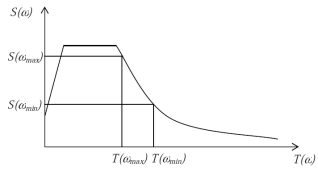
$$\boldsymbol{V}^T\boldsymbol{M}\boldsymbol{V}\ddot{\boldsymbol{u}} + \boldsymbol{V}^T\boldsymbol{C}\boldsymbol{V}\dot{\boldsymbol{u}} + \boldsymbol{V}^T\boldsymbol{K}\boldsymbol{V}\boldsymbol{u} = -\boldsymbol{V}^T\boldsymbol{M}\boldsymbol{s}\ddot{d}_g$$

system of equations

$$\boldsymbol{I}\ddot{\boldsymbol{u}} + \boldsymbol{D}\dot{\boldsymbol{u}} + \boldsymbol{\Omega}_0^2\boldsymbol{u} = -\boldsymbol{h}$$

$$\ddot{u}_i + D_i\dot{u}_i + \omega_{0,i}^2 u_i = h_i$$

$$y_i = \frac{S(\omega_{0,i})}{\omega_{0,i}^2}$$

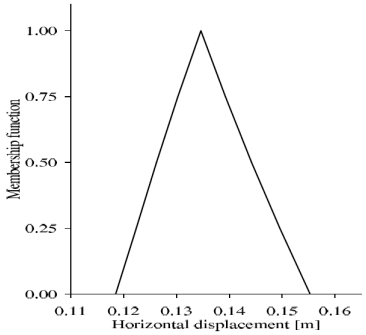


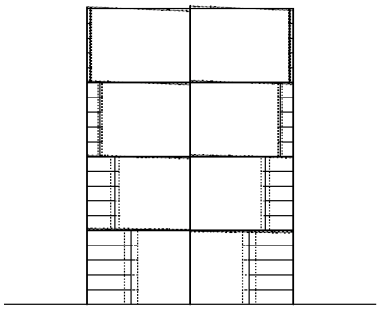
displacements

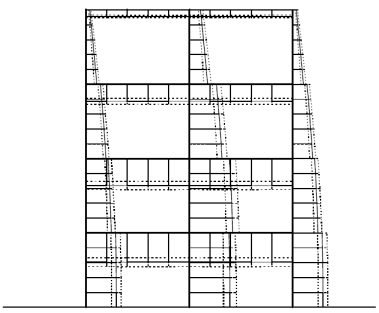
$$\boldsymbol{d} = \sum_{i=1}^{i=n} \left(\boldsymbol{v}_i^T \boldsymbol{M}\boldsymbol{s} y_i\right) \boldsymbol{v}_i$$

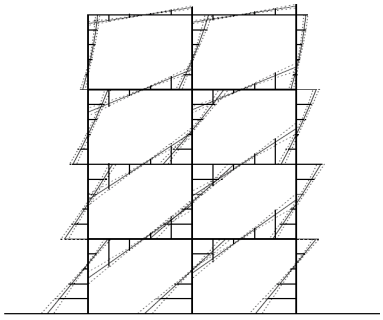
forces and moments

$$\boldsymbol{f} = (\boldsymbol{K} - \omega_0^2 \boldsymbol{M})\boldsymbol{d}$$



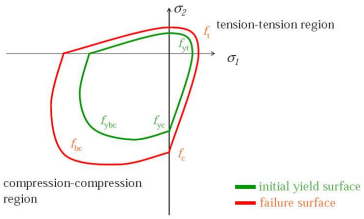






Fuzzification of Chen Model of Plasticity

Loading surfaces of concrete in biaxial-stress plane:



Initial yield surface in compression-compression region:

$$f_0^c(\sigma) = J_2 + \frac{A_0}{3} I_1 - \tau_0^2 = 0$$

Initial yield surface in compression-compression region:

$$f_u^c(\sigma) = J_2 + \frac{A_u}{3} I_1 - \tau_u^2 = 0$$

Formulae for material constants (obtained by simple tests):

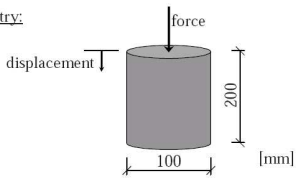
$$A_0 = \frac{f_{ybc}^2 - f_{yc}^2}{2 f_{ybc} - f_{yc}}$$

$$\tau_0^2 = \frac{f_{yc} f_{ybc} (2 f_{yc} - f_{ybc})}{3(2 f_{ybc} - f_{yc})}$$

$$A_u = \frac{f_{bc}^2 - f_c^2}{2 f_{bc} - f_c}$$

$$\tau_u^2 = \frac{f_{yc} f_{ybc} (2 f_{yc} - f_{ybc})}{3(2 f_{ybc} - f_{yc})}$$

Geometry:

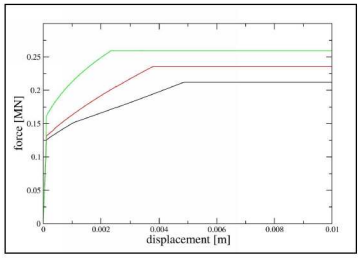


Material strength characteristics in [MPa]:

f_c	f_{yc}	f_t	f_{yt}	f_{bc}	f_{ybc}
30	18	2.7	1.6	34.8	21

← vary by ± 10 %

Force-displacement diagram:



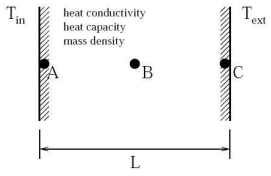
Heat Transfer

nonstationary heat transfer

$$k\frac{\partial^2 T}{\partial x^2} + z = \rho c \frac{\partial T}{\partial t}$$

conductivity	$1.67\ J/(msK)$
heat capacity	$840\ J/(kgK)$
mass density	$2400\ kg/m^3$

Table 1: Material characteristics



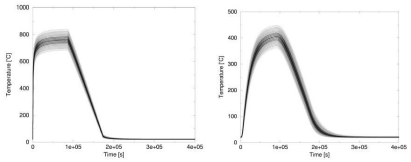


Figure 2: Temperature in A and B

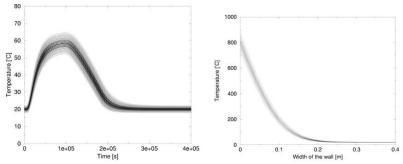


Figure 3: Temperature in C and temperature in wall after 1 hour

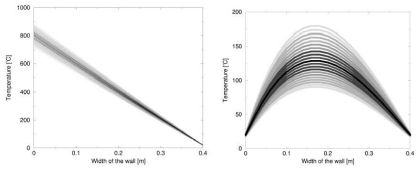


Figure 4: Temperature in wall after 1 day and 2 days

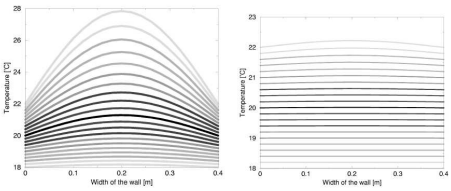


Figure 5: Temperature in wall after 3 days and 4 days

Parallelization

- α-cut method leads to a large number of samples, e.g., in the case of vibration of plane frame $3^{2\times4} = 6561$ and $5^{2\times4} = 390625$ samples were used
- independent samples
- easy parallelization

CTU Uncertainties and Fuzzy Sets J. Kruis

	speedup	ideal speedup
singleprocessor computation	333 s	
parallel computation on 6 processors	58 s	55,5 s
parallel computation on 21 processors	18 s	15,8 s

2. - 6. 2. 2009 SNA09, Institute of Geonics ASCR Ostrava

A numerical solution of elliptic boundary value problems with uncertain data and geometry

T. Kozubek

VSB - Technical University of Ostrava

1 Introduction

An efficient method for the numerical solution of elliptic PDEs in domains depending on random variables has been introduced in [1]. The key feature is the combination of a fictitious domain approach and a polynomial chaos expansion. The PDE is solved in a larger, fixed domain (the fictitious domain), with the original boundary condition enforced via a Lagrange multiplier acting on a random manifold inside the new domain. A (generalized) Wiener expansion is invoked to convert such a stochastic problem into a deterministic one, depending on an extra set of real variables (the stochastic variables). Discretization is accomplished by standard mixed finite elements in the physical variables and a Galerkin projection method with numerical integration (which coincides with a collocation scheme) in the stochastic variables. A stability and convergence analysis of the method, as well as numerical results, are provided in [1]. The convergence is "spectral" in the polynomial chaos order, in any subdomain which does not contain the random boundaries.

2 Setting of the problem

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

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

$$\left\{ \begin{array}{rcl} -\triangle u(\,\cdot\,,\omega) & = & f \ \text{ in } D(\omega), \\ u(\,\cdot\,,\omega) & = & 0 \ \text{ on } \Gamma(\omega), \end{array} \right. \qquad \big(\mathcal{P}(\omega)\big)$$

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

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

3 The stochastic FD formulation

The stochastic FD formulation reads as follows: Find $\hat{u}(\,\cdot\,,\omega) \in H_0^1(\hat{D})$ and $\lambda(\,\cdot\,,\omega) \in M(\omega) := H^{-1/2}(\Gamma(\omega))$ such that, a.s. in Ω,

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

We assume that, a.s., $\Gamma(\omega)$ is obtained from a reference $C^{1,1}$ or polygonal boundary Γ_0 as the image of a piecewise smooth invertible mapping $\gamma_0(\omega)$. More precisely, we assume that $\Gamma(\omega) = \gamma_0(\omega)(\Gamma_0)$, where $\gamma_0(\omega)$ belongs to $C^{1,p}(\Gamma_0)$ (the space of all continuous and piecewise continuously differentiable mappings $\gamma : \Gamma_0 \to \mathbb{R}^2$) and its inverse $\gamma_0(\omega)^{-1}$ exists and belongs to $C^{1,p}(\Gamma(\omega))$. The function $\gamma_0 : \Omega \to C^{1,p}(\Gamma_0)$ is assumed to be a random variable belonging to $L^\infty(\Omega, dP; C^{1,p}(\Gamma_0))$, i.e., γ_0 is a jointly measurable function on the Borel sets of $\Gamma_0 \times \Omega$ for which there exists a constant $g_0 > 0$ such that $\|\gamma_0(\omega)\|_{C^{1,p}(\Gamma_0)} \leq g_0$ a.s. in Ω; the same occurs for the inverse mapping, i.e., $\|\gamma_0(\omega)^{-1}\|_{C^{1,p}(\Gamma(\omega))} \leq g_0$ a.s. in Ω.

Let $\mathbb{E}[X] = \int_\Omega X(\omega)\, dP(\omega)$ be the expected value of a real-valued random variable X. Let $L^2(\Omega, dP) = \{X : \Omega \to \mathbb{R} \,|\, X \text{ is a random variable such that } \mathbb{E}\left[X^2\right] < +\infty\}$ be the space of second order random variables over the probability space (Ω, F, P). We denote by $L^2(\Omega, dP; H_0^1(\hat{D}))$ the space of the random variables $v : \Omega \to H_0^1(\hat{D})$ (i.e., $v : \hat{D} \times \Omega \to \mathbb{R}$ is jointly measurable and $v(\,\cdot\,,\omega) \in H_0^1(\hat{D})$ a.s. in Ω) with finite second order moment $\mathbb{E}\left[\|v\|^2_{H_0^1(\hat{D})}\right] = \int_{\hat{D}} \mathbb{E}\left[|\nabla v|^2\right] d\mathbf{x} < +\infty$. The definition of the space $L^2(\Omega, dP; H^{-1/2}(\Gamma_0))$ is similar. Finally, the space $L^2(\Omega, dP; H^{-1/2}(\Gamma))$ is defined as follows: $\mu \in L^2(\Omega, dP; H^{-1/2}(\Gamma))$ means that $\mu_0 \in L^2(\Omega, dP; H^{-1/2}(\Gamma_0))$, where $\mu_0(\omega) \in H^{-1/2}(\Gamma_0)$ is defined a.s. in Ω by the conditions $\langle \mu_0, v_0 \rangle_{\Gamma_0} = \langle \mu, v_0 \circ \gamma_0^{-1} \rangle_{\Gamma(\omega)}$ for all $v_0 \in H^{1/2}(\Gamma_0)$.

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

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

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

4 (Wiener) polynomial chaos

This section is devoted to recalling some basic facts about polynomial chaos (see, e.g., [2]), as well as to setting the notation.

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

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

where δ_{nm} is the Kronecker symbol. Next, denote by $\mathbf{y} = (y_k)_{k \geq 1} \in \mathbb{R}^{\mathbb{N}_0}$ any infinite sequence of real variables, and by $\boldsymbol{\nu} = (\nu_k)_{k \geq 1} \in \mathbb{N}^{\mathbb{N}_0}$ any infinite sequence of integers which is "finite",

i.e., such that $\nu_k > 0$ only for a finite number of indices; let $|\boldsymbol{\nu}| = \sum_{k\geq 1} \nu_k$. Define the multidimensional Hermite polynomials of order $|\boldsymbol{\nu}|$ as $H_{\boldsymbol{\nu}}(\mathbf{y}) = \prod_{k=1}^{\infty} H_{\nu_k}(y_k)$; note that the definition is meaningful since $H_0(y) \equiv 1$, hence, $H_{\boldsymbol{\nu}}(\mathbf{y})$ actually depends only on a finite number of components of $\mathbf{y}$. These polynomials are mutually orthonormal, in the following sense:

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

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

Theorem 7 *Let $\Phi \in L^2(\Omega, dP)$ and let $\Phi_{\boldsymbol{\nu}} = \mathbb{E}[\Phi \mathcal{H}_{\boldsymbol{\nu}}]$ for any finite $\boldsymbol{\nu}$. Then,*

$$\Phi = \sum_{\boldsymbol{\nu} \text{ finite}} \Phi_{\boldsymbol{\nu}} \mathcal{H}_{\boldsymbol{\nu}} \qquad \text{in } L^2(\Omega, dP).$$

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

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

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

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

So far, we have focussed on Gaussian random variables. Similar representations can be given for second order random variables over other probabilistic spaces admitting a density function. The system of orthonormal polynomials which gives rise to a *generalized polynomial chaos*, similar to the Wiener chaos, is determined by the density function; for instance, the uniform density obviously leads to the Legendre polynomials. We refer to [2] for more details.

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

$$\Phi(\omega) = \varphi(\mathbf{Y}_K(\omega)), \qquad \mathbf{Y}_K(\omega) := (Y_1(\omega), \ldots, Y_K(\omega)), \tag{4.1}$$

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

5 The deterministic formulation of $\left(\hat{\mathcal{P}}^S\right)$

We go back to the stochastic formulation $\left(\hat{\mathcal{P}}^S\right)$. We assume that the boundary $\Gamma(\omega)$ of $D(\omega)$ depends on ω via K mutually independent real random variables $Y_1(\omega), \ldots, Y_K(\omega)$ with zero mean and unit variance with respect to a density function ρ defined on some interval $I \subseteq \mathbb{R}$. Let $\mathbf{Y}_K(\omega)$ and ϱ be defined as above. Since we assumed in Section 3 that $\Gamma(\omega) = \gamma_0(\omega)(\Gamma_0)$, equation (4.1) easily yields $\gamma_0(\omega) = \gamma_0^*(\mathbf{Y}_K(\omega))$, where $\gamma_0^* = \gamma_0^*(\mathbf{y})$ is a family of $C^{1,p}(\Gamma_0)$-mappings defined in $\mathbf{I} = I^K$, with inverses $\gamma_0^*(\mathbf{y})^{-1}$ in $C^{1,p}(\Gamma^*(\mathbf{y}))$. Thus, $\Gamma^*(\mathbf{y}) = \gamma_0^*(\mathbf{y})(\Gamma_0)$ is a parametrization of the set of the admissible boundaries of the stochastic domains $D(\omega)$.

Since $\hat{u}$ and λ depend on ω only through $\Gamma(\omega)$, the Doob-Dynkin lemma assures that this dependence takes place via $\mathbf{Y}_K(\omega)$, i.e., we have $\hat{u}(\,\cdot\,,\omega) = \hat{u}^*(\,\cdot\,,\mathbf{Y}_K(\omega))$ and $\lambda(\,\cdot\,,\omega) = \lambda^*(\,\cdot\,,\mathbf{Y}_K(\omega))$, where $\hat{u}^*(\,\cdot\,,\mathbf{y}) \in H_0^1(\hat{D})$ and $\lambda^*(\,\cdot\,,\mathbf{y}) \in H^{-1/2}(\Gamma^*(\mathbf{y}))$, a.e. in $\mathbf{I}$. Condition $\hat{u} \in L^2(\Omega, dP; H_0^1(\hat{D}))$ is then equivalent to $\hat{u}^* \in L^2_\varrho(\mathbf{I}; H_0^1(\hat{D}))$; similarly, $\lambda \in L^2(\Omega, dP; H^{-1/2}(\Gamma))$ is equivalent to $\lambda^* \in L^2_\varrho(\mathbf{I}; H^{-1/2}(\Gamma^*))$ (with obvious meaning of the notation).

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

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

6 Discretization of the deterministic formulation

Discretization is accomplished by standard mixed finite elements in the physical variables and a Galerkin projection method with numerical integration (which coincides with a collocation scheme) in the stochastic variables. Thus instead of solving very large algebraic saddle-point system resulting from the discretization of $\left(\hat{\mathcal{P}}^D\right)$, we will solve n original deterministic problems for n different configurations of the stochastic domain $D(\mathbf{y})$, where n is the number of Gauss (collocation) points $\mathbf{y_q}$. We can simply parallelize all computations. For more details see [1], where a stability and convergence analysis of the method have been presented. We showed that, in any subdomain that does not contain the random boundaries, the convergence is "spectral" in the polynomial chaos order. Solution of a few problems will be presented during the talk.

Acknowledgement: This work has been supported by the grant ASCR IAA100750802 and by the Research Project MSM6198910027 of the Czech Ministry of Education.

References

[1] C. Canuto, T. Kozubek: *A fictitious domain approach to the numerical solution of PDEs in stochastic domains*, Numerische Mathematik 107, 2007, 257–293.

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

NUMERICAL STABILITY OF SYMMETRIC INDEFINITE SOLVERS: DIRECT METHODS

Miro Rozložník

Institute of Computer Science
Czech Academy of Sciences, Prague

joint work with Gil Shklarski and Sivan Toledo

Seminar of Numerical Analysis SNA09, Ostrava

February 2-6, 2009

Outline

Solution of a symmetric indefinite system of linear equations

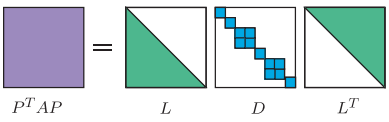
$$Ax = b$$

A is symmetric (indefinite)

$$P^T A P P^T x = P^T b$$

$$LDL^T P^T x = P^T b \qquad LTL^T P^T x = P^T b$$

Block Bunch-Kaufmann factorization



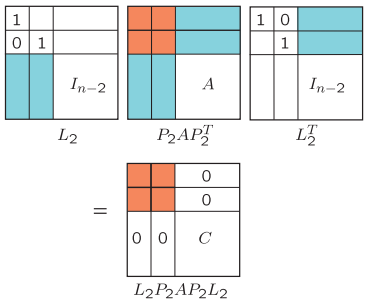
A is symmetric (definite or indefinite)

L is **unit** lower triangular

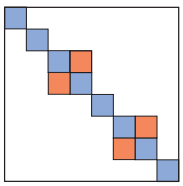
D is **symmetric block diagonal** with 1x1,2x2 blocks

P is a **permutation** matrix

Bunch-Kaufmann factorization



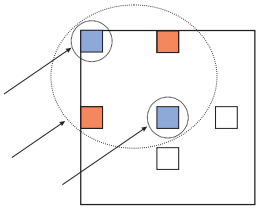
Bunch-Kaufmann factorization



$$\underbrace{L_{n-1}P_{n-1}\ldots L_2 P_2 L_1 P_1 A P_1^T L_1^T P_2^T L_2^T \ldots P_{n-1}^T L_{n-1}^T}_{D}$$

Bunch-Kaufmann pivoting strategy

- complete pivoting $O(n^3)$ comparisons Bunch, Parlett
- partial pivoting $O(n^2)$ comparisons implemented in LINPACK, LAPACK



Triangular tridiagonalization

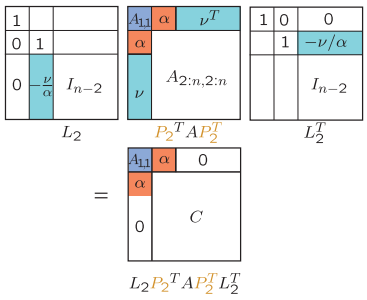
A is symmetric (definite or indefinite)

L is **unit** lower triangular

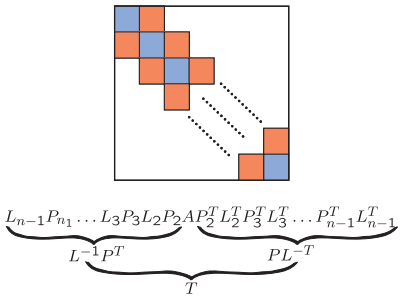
T is **symmetric** tridiagonal

P is a **permutation** matrix

Parlett - Reid reduction



Parlett - Reid reduction

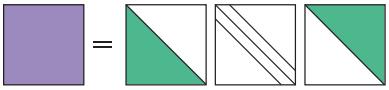


Parlett - Reid reduction

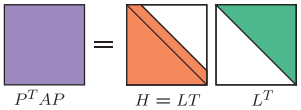
- The reduced matrix remains symmetric during reduction, the updates are performed on a half of the matrix
- Complexity: at each step two rank-one updates on half a matrix $2(n-1)^2$; $O(n)$ other operations; total $2/3n^3 + O(n^2)$

→ Aasen's factorization

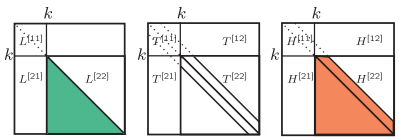
Parlett - Reid reduction



Aasen's factorization



Notation



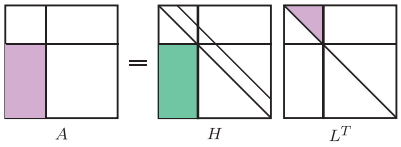
Parlett - Reid reduction

works on $L^{[22]}T^{[22]}(L^{[22]})^T$

Aasen's factorization

* for $k > 1$ $L^{[22]}T^{[22]}(L^{[22]})^T \neq H^{[22]}(L^{[22]})^T$
* update of $A^{[22]}$
  compute $(k+1)$-th column L and k-th column of T and H
* pivoting strategy

Aasen's factorization – Phase 1

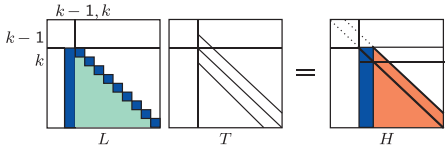


Compute i-th column of $H^{[21]}$ from i-th column of $A^{[21]}$ and previous columns of $H^{[21]}$ and $L^{[11]}$

- known

Aasen's factorization – Phase 2

Second phase – extract the first column of $L^{[22]}T^{[22]}$

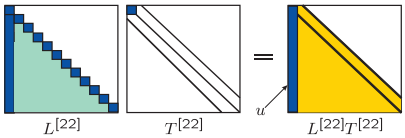


$$u \leftarrow H^{[22]}_{1,1:\text{last}} - L_{1:\text{last},k-1}T_{k,k-1}$$

- known

Aasen's factorization – Phase 3

Third phase – extract $T^{[22]}_{1,1}$

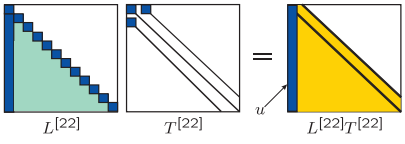


$$T^{[22]}_{1,1} = u_1$$

- known

Aasen's factorization – Phase 4

Fourth phase – extract $T^{[22]}_{1,2}$ and $L^{[22]}_{2:\text{last},2}$



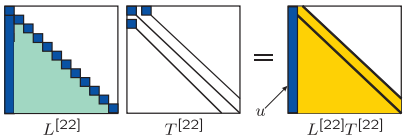
$$u_{2:\text{last}} = T^{[22]}_{1,1}\, L^{[22]}_{2:\text{last},1} + T^{[22]}_{1,2}\, L^{[22]}_{2:\text{last},2}$$

$$T^{[22]}_{1,2}\, L^{[22]}_{2:\text{last},2} = u_{2:\text{last}} - T_{1,1}\, L^{[22]}_{2:\text{last},1}$$

- known

Aasen's factorization – pivoting strategy

Fourth phase – extract $T^{[22]}_{1,2}$ and $L^{[22]}_{2:\text{last},2}$

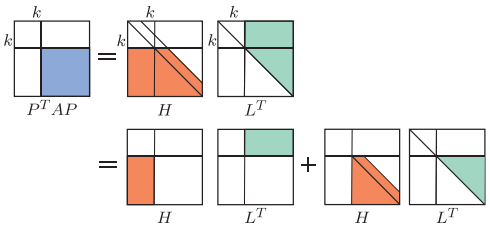


$$u_{2:\text{last}} = T^{[22]}_{1,1} L^{[22]}_{2:\text{last},1} + T^{[22]}_{1,2} L^{[22]}_{2:\text{last},2}$$

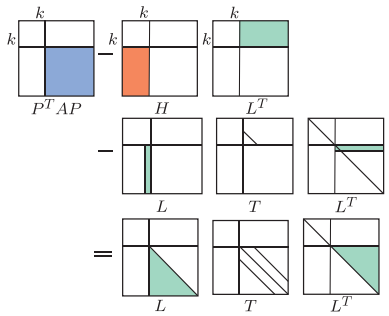
$$T^{[22]}_{1,2} L^{[22]}_{2:\text{last},2} = u_{2:\text{last}} - T_{1,1} L^{[22]}_{2:\text{last},1}$$

First switch variable $(i+1)$ with the largest index in

Partitioned factorization



Partitioned factorization



Numerical stability – main result

$$A + \Delta A = \overline{L}\,\overline{T}\,\overline{L}^T$$

$$|\Delta A| \le c_3(n,k)\, u\, |\overline{L}|\,|\overline{T}|\,|\overline{L}|^T$$

$$c_3(n,k) = c_1\left(n + \left\lfloor \frac{n}{k} \right\rfloor + 2\right), \quad c_3(n,1) = c_1(n+3)$$

Basic assumptions on BLAS

$$X \in \mathbb{R}^{m,k}, \quad Y \in \mathbb{R}^{k,n}, \quad Z = XY \in \mathbb{R}^{m,n}, \quad \overline{Z} = fl(XY)$$

conventional BLAS:

$$|\overline{Z} - Z| \le c_1(k) u |X|\,|Y| \quad c_1(k) = \frac{k}{1-ku}$$

Strassen:

$$\|\overline{Z} - Z\| \le c_3(m,n,k,p) u \|X\|\,\|Y\|$$

Numerical stability – Proof 1

$$\begin{aligned}
A^{[11]} + \Delta A^{[11]} &= \overline{L}^{[11]}\overline{T}^{[11]}\left(\overline{L}^{[11]}\right)^T \\
\left|\Delta A^{[11]}\right| &\le c_3(k,1) u \left|\overline{L}^{[11]}\right| \left|\overline{T}^{[11]}\right| \left|\overline{L}^{[11]}\right|^T \\
A^{[21]} + \Delta A^{[21]} &= \overline{H}^{[21]}\left(\overline{L}^{[11]}\right)^T \\
\left|\Delta A^{[21]}\right| &\le c_1(k) u \left|\overline{H}^{[21]}\right| \left|\overline{L}^{[11]}\right|^T \\
\overline{H}^{[21]} + \Delta H^{[21]} &= \overline{L}^{[21]}\overline{T}^{[11]} + \overline{L}^{[22]}_{:,1}\overline{T}^{[21]}_{1,k} \\
\left|\Delta H^{[21]}\right| &\le c_1(3)\left(\left|\overline{L}^{[21]}\right| \left|\overline{T}^{[11]}\right| + \left|\overline{L}^{[22]}_{:,1}\right| \left|\overline{T}^{[21]}_{1,k}\right|\right)
\end{aligned}$$

Numerical stability – Proof 2

$$\overline{C}^{[22]} + \Delta C^{[22]} = A^{[22]} - \overline{H}^{[21]} \left(\overline{L}^{[21]}\right)^T - \overline{L}^{[21]}_{:,k} \overline{T}^{[21]}_{1,k} \left(\overline{L}^{[22]}_{:,1}\right)^T$$

$$\left|\Delta C^{[22]}\right| \leq c_1(k+1)\, u \left(\left|\overline{H}^{[21]}\right| \left|\overline{L}^{[21]}\right|^T + \left|\overline{L}^{[21]}_{:,k}\right| \left|\overline{T}^{[21]}_{1,k}\right| \left|\overline{L}^{[22]}_{:,1}\right|^T \right)$$

$$\overline{C}^{[22]} + \Delta\overline{C}^{[22]} = \overline{L}^{[22]} \overline{T}^{[22]} \left(\overline{L}^{[22]}\right)^T$$

$$\left|\Delta\overline{C}^{[22]}\right| \leq c_3(n-k,k)u \left|\overline{L}^{[22]}\right| \left|\overline{T}^{[22]}\right| \left|\overline{L}^{[22]}\right|^T$$

Solution of a linear system

Assuming $c_4(n)uk_\infty\left(\overline{T}\right) < 1$

$$\left(A + \widehat{\Delta A}\right)\overline{x} = b + \widehat{\Delta b}$$

$$\left\|\widehat{\Delta A}\right\|_\infty \leq c_5(n,k)u \left\|\overline{T}\right\|_\infty, \left\|\widehat{\Delta b}\right\|_\infty \leq c_5(n,k)u \left\|\overline{T}\right\|_\infty \|\overline{x}\|_\infty$$

growth factor $\rho_n = \dfrac{\max_{i,j}\left|\overline{T}_{i,j}\right|}{\max_{i,j}\left|A_{i,j}\right|}$

$$\max\left\{ \frac{\left\|\widehat{\Delta A}\right\|_\infty}{\|A\|_\infty}, \frac{\left\|\widehat{\Delta b}\right\|_\infty}{\|A\|_\infty \|\overline{x}\|_\infty} \right\} \leq c_5(n,k)nu\rho_n$$

Bunch Kaufmann factorization – numerical stability

$$P(A + \Delta A)P^T = \overline{L}\,\overline{D}\,\overline{L}^T$$

$$|\Delta A| \leq c_6(n)u \left(|A| + \left|\overline{L}\right| \left|\overline{D}\right| \left|\overline{L}\right|^T \right)$$

Solution of a linear systems

Assuming that $c_7(n)uk\left(\overline{D}\right) < 1$

$$\left(A + \widehat{\Delta A}\right)\overline{x} = b$$

$$\left|\widehat{\Delta A}\right| \leq c_6(n)u \left(|A| + \left|\overline{L}\right| \left|\overline{D}\right| \left|\overline{L}\right|^T \right)$$

growth factor $\rho_n = \dfrac{\max_{i,j,k}\left|\overline{a}^{(k)}_{ij}\right|}{\max_{i,j}|a_{ij}|}$

$$\frac{\left\|\widehat{\Delta A}\right\|_\infty}{\|A\|_\infty} \leq c_6(n)nu\rho n$$

Parallel implementation

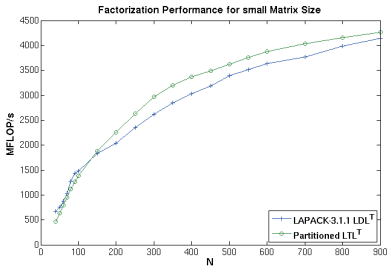
LAPACK uses blocked Bunch-Kaufmann factorization (Dongarra, Anderson)

Cache-efficient partitioned triangular tridiagonalization (Shklarski, Toledo – submitted to ACM TOMS)

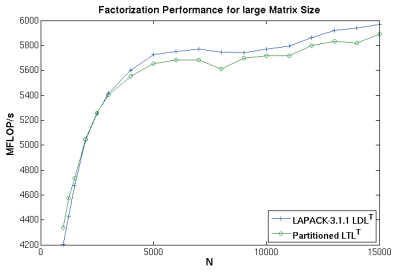
Numerical examples

- Machine
  - Core 2 Duo 6400, 2.13Ghz, 4GB of main memory
  - Linux x86_64
- Partitioned LTL^T
  - C implementation, GCC
  - Batch size = 64
  - Fused LTL^T factor and QR of T
- Partitioned LDL^T
  - LAPACK 3.1.1 (Bunch-Kaufman)
  - Block size = 64
- BLAS: GOTO BLAS 1.12, confined to a single core
- Matrices:
  - Symmetric matrices, elements uniformly distributed in $(-1,1)$

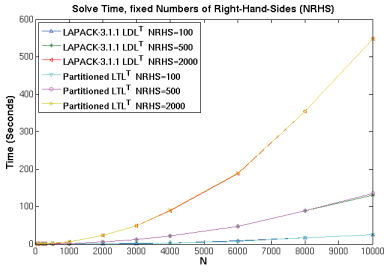
Numerical examples



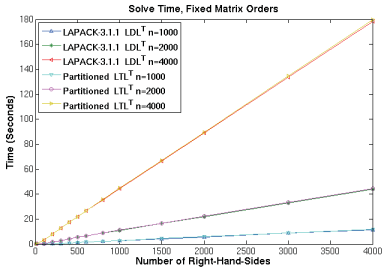
Numerical examples



Numerical examples



Numerical examples



Conclusions

LDL^T	LTL^T
✌ Reveals inertia	✌ Does not reveal inertia
✌ Easy to solve with D	✌ Slightly harder to solve with T
✌ Bunch Kauffman Pivoting	✌ Simple Pivoting
✌ $L_{i,j}$ can grow	✌ Bounded $L_{i,j}$
✌ Bounded D	✌ T can grow

THANK YOU FOR YOUR ATTENTION

R, G. Shklarski, S. Toledo: Partitioned triangular tridiagonalization, submitted to ACM Transactions on Mathematical Software

C and Matlab Codes at

`http://www.tcu.ac.il/~stoledo/research.html`

IT4Innovations
Centrum excelence

www.it4i.cz

The IT4Innovations project aims to create a unique structure with both national and international significance, focused on key areas of science and research such as the development of the information society, the development of embedded systems, innovative medicine and nanotechnologies – and of course information technologies themselves.

Moreover, the IT4Innovations project represents an exceptional synergy of scientific, research and development capacities in **computer science and computational mathematics**, with the goal of stimulating the development of a wide range of modern and progressive technologies.

The project IT4Innovations is currently under development for the expected call in the EU Operational Programme Research and Development for Innovations. The project has the following structure:

IT4People

- **IT4Disaster Management:** IT for modelling and management of crisis situations
- **IT4Traffic Management:** IT for monitoring and intelligent management of traffic
- **IT4Economy:** IT for financial simulations and agile logistical computations

SC4Simulations

- **SC4Industry**: supercomputer simulations for solving industrial problems
- **SC4NaturalSciences**: supercomputer modelling and simulation in natural sciences
- **SC4Nanotechnologies**: modelling with supercomputers in nanotechnologies

EC4Innovations

- **EC4Industry:** embedded systems in industrial applications (control systems, service life monitoring, etc.)
- **EC4Mechatronics:** development of systems based on the interdisciplinary combination of mechanical, electronic and IT systems
- **EC4InnovativeMedicine:** development of embedded systems for medical applications

Theory4IT

- **IT4Softcomputing:** research of mathematical principles and methods of processing knowledge burdened with uncertainties and their use in the development of methods applicable in decision-making, management, complex systems design etc**.**
- **IT4Bioinformatics:** research of algorithms inspired by biological models (evolutional and genetic algorithms, ant colony theory, neuron networks, etc.)
- **IT4Formal Methods:** modern methods used in software engineering
- **IT4Knowledge (Information Technology for Knowledge): research of knowledge mining and the development of special data structures for storage of extensive collections of weakly structured data.**
- **IT4Multiagent:** research of strategy and cooperation in multiagent systems

Title: Seminar on Numerical Analysis & Winter School
Proceedings of the conference SNA'09, Ostrava, February 2 – 6, 2009

Editors: Radim Blaheta, Jiří Starý

Published by: Institute of Geonics AS CR Ostrava

Printed by: Publishing Centre of VSB – Technical University Ostrava

First edition

Ostrava, 2009

80 copies

ISBN 978-80-86407-60-9