UNIVERSITATIS PALACKIANAE OLOMUCENSIS
FACULTAS RERUM NATURALIUM

Department of Mathematical Analysis
and Applications of Mathematics

ODAM
2009

Editor: Horymír Netuka
Technical Editor: Miloslav Závodný

OBSAH — CONTENTS


Univ. Palacki. Olomuc., Fac. rer. nat.,
Dept of Math. Anal. and Appl. of Math.
ODAM (2009) 4–22

On Determining the Moore–Penrose Inverse*

Radek KUČERA [1], Jitka MACHALOVÁ [2]

[1] *Department of Mathematics and Descriptive Geometry*
VŠB–Technical University of Ostrava, 17. listopadu 15
CZ-708 33 Ostrava-Poruba, Czech Republic
e-mail: radek.kucera@vsb.cz

[2] *Department of Mathematical Analysis and Application of Mathematics*
Faculty of Science, Palacký University
tř. 17. listopadu 12, 771 46 Olomouc, Czech Republic
e-mail: machalova@inf.upol.cz

Abstract

The paper deals with generalized inverses to a class of large-scale matrices arising from the TFETI domain decomposition method. It is shown that the Moore-Penrose inverse may be obtained by the product of an arbitrary generalized inverse with orthogonal projectors. Applying an eigenvalue analysis, the scalability of the TFETI method is proved for model problems.

Key words: Generalized inverse, Moore–Penrose inverse, orthogonal projectors, TFETI domain decomposition method, condition number.

2000 Mathematics Subject Classification: 15A09, 15A12, 65F35, 65N22

*Supported by the grant 101/08/0574 of the Grant Agency of the Czech Republic and by the Research Project MSM 6198910027 of the Czech Ministry of Education.

1 Introduction

The FETI (Finite Element Tearing and Interconnecting) method [8] belongs to the most efficient domain decomposition techniques for the numerical solution of boundary value problems described by elliptic partial differential equations (PDEs). The algebraic problem arising from the FETI method consists of the *saddle-point system*, i.e., the problem for finding $(\bar{u}, \bar{\lambda}) \in \mathbb{R}^n \times \mathbb{R}^m$ satisfying:

$$\begin{pmatrix} A & B^\top \\ B & 0 \end{pmatrix} \begin{pmatrix} u \\ \lambda \end{pmatrix} = \begin{pmatrix} f \\ g \end{pmatrix}, \tag{1.1}$$

where $A \in \mathbb{R}^{n\times n}$ is the stiffness matrix, $B \in \mathbb{R}^{m\times n}$ is the "gluing" matrix, $f \in \mathbb{R}^n$, and $g \in \mathbb{R}^m$. Let us note that $\bar{\lambda}$ is the vector of *Lagrange multipliers* enforcing the continuity of the PDE solution.

There are two main benefits of the FETI approach. Firstly, the stiffness matrix has a block diagonal structure that enables us to handle the blocks in parallel. Secondly, the condition number of all blocks and, consequently of the whole stiffness matrix, may be independent on the size of the discrete problem (under additional assumptions on the finite element partitions). It is well-known that convergence of conjugate gradient type methods is determined by the condition number of the matrix [10]. Therefore the number of iterations needed to get a solution to (1.1) with a given accuracy may be independent on the size of the discrete problem, as well. This property is known as the *scalability* of the method.

The classical FETI algorithm [8] consists of eliminating the first unknown u from (1.1) after that the resulting (dual) linear system in terms of λ is solved iteratively. As the diagonal blocks in A may be singular, the elimination requires to use a generalized inverse to A and a basis of the null-space of A. One of reasons for developing variants of the FETI method was the effort to overcome difficulties in computing with generalized inverses and in identifying null-spaces. The variant FETI-DP [7, 13] modifies the original FETI method so that A is non-singular. Then the null-space is trivial and the inverse to A exists. The opposite strategy gave rise to the TFETI (Total FETI) method [4] in which also the Dirichlet boundary conditions of the PDE problem are enforced by Lagrange multipliers. In this case, the null-space is as large as possible, since all diagonal blocks in A are stiffness matrices to the original PDEs with the *pure* homogeneous Neumann boundary conditions. The advantage is that the null-space basis is known à-priori and, in addition, it may be assembled by mechanical arguments with negligible cost.

The main goal of our paper is to show how to handle in the TFETI method with the generalized inverse to A. As stiffness matrices to elliptic PDE problems are symmetric, positive definite, the generalized Cholesky factorization can be applied [10]. Then, by inverting the regular part of the Cholesky factor, one can easily get a generalized inverse to A. Unfortunately, the factorization procedure is sensitive to round-off errors, as the zero pivots must be recognized. Therefore

Farhat and Gerardin [6] proposed to finish the Cholesky factorization by the singular value decomposition (SVD) when it is appeared a pivot that is suspected of zero. This technique leads to the robust algorithm. The other development of this idea was done by Dostál et. al. [5] whose proposed to detect actively small principal submatrices of A that are decomposed by the SVD while the (non-singular) rest of A is treated by the Cholesky factorization again. Their numerical experiments confirmed that the resulting generalized inverse may be sufficiently close to the Moore–Penrose (MP) one. It should be noted that the MP inverse is the best generalized inverse for the TFETI method. The reason is that it minimizes (among all generalized inverses) the norm of the computed vector [3] that keeps the FETI algorithm as stable as possible. This fact is highly important when the saddle-point system (1.1) is large. Our results will show that the (exact) MP inverse can be obtained from an arbitrary (stable computable) generalized inverse modifying it by the orthogonal projector on the range-space of A. Since the orthogonal projector is available in the TFETI method due to the knowledge of the null-space basis, the MP inverse may be easily implemented. Our idea is closely related to Pyle's algorithm [18] which, however, assembles the MP inverse projecting the generalized inverse computed by the orthogonal factorization of the matrix.

The paper is organized as follows. Section 2 deals with a generalized inverse to an arbitrary (rectangular) matrix. We derive the three-condition characterization of the MP inverse based on identifying its null-space and range-space. Then we prove how to get the MP inverse from an arbitrary generalized inverse using orthogonal projectors. In Section 3 we discuss the MP inverse for solving saddle-point systems. The eigenvalue analysis shows that the condition number of the dual projected equation is bounded by the condition number of A. Applying this result in Section 4 we prove the scalability of the TFETI method for simple model problems in one and two space dimensions (1D and 2D).

2 Generalized inverses and projectors

In this section, we prove three conditions determining the Moore–Penrose (MP) inverse to a rectangular matrix. Then we show how to adapt an arbitrary generalized inverse into the MP one using orthogonal projectors. Before getting these results we start with preliminaries.

Let $\mathbb{R}^{m\times n}$ be the set of $m \times n$ real matrices and let $A \in \mathbb{R}^{m\times n}$. The symbols $\operatorname{Ker} A$ and $\operatorname{Im} A$ stand for the kernel (null-space) and the image (range-space) of A, respectively. The rank of A is defined by $r(A) := \dim \operatorname{Im} A$ and it is known that $r(A) = r(A^\top)$, where $A^\top$ is the transpose to A. Finally, let I denote the (square) identity matrix.

By a *generalized inverse* to A we call such $X \in \mathbb{R}^{n\times m}$ that satisfies the following equation:

$$A = AXA. \tag{2.1}$$

Let us note that there is a generalized inverse for any A but it is not uniquely determined by (2.1). The MP inverse to A, denoted here by $A^\dagger$, is a particular generalized inverse that is unique for any A. There are various definitions of $A^\dagger$. Here we will define $A^\dagger$ by the singular value decomposition (SVD) of A.

Theorem 2.1 *Let $A \in \mathbb{R}^{m\times n}$ be of rank $r = r(A)$. There are orthogonal $U \in \mathbb{R}^{m\times m}$, $V \in \mathbb{R}^{n\times n}$ and diagonal $\Sigma \in \mathbb{R}^{m\times n}$ with non-negative entries so that*

$$A = U\Sigma V^\top. \tag{2.2}$$

Moreover, the diagonal entries of Σ can be sorted in the decreasing order so that

$$\Sigma = \begin{pmatrix} \widehat{\Sigma} & 0 \\ 0 & 0 \end{pmatrix}, \tag{2.3}$$

where $\widehat{\Sigma} \in \mathbb{R}^{r\times r}$ is the non-singular part of Σ.

Proof. See [1]. □

By the SVD of A we understand U, V, and Σ satisfying (2.2) and (2.3). Let us note that the SVD is uniquely determined by (2.2) and (2.3) for any A. In addition we split U and V accordingly to (2.3), i.e., $U_1 \in \mathbb{R}^{m\times r}$, $U_2 \in \mathbb{R}^{m\times m-r}$, $V_1 \in \mathbb{R}^{n\times r}$, and $V_2 \in \mathbb{R}^{n\times n-r}$ are such that U, V take the form $U = (U_1, U_2)$, $V = (V_1, V_2)$, respectively. Then

$$A = (U_1, U_2) \begin{pmatrix} \widehat{\Sigma} & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} V_1^\top \\ V_2^\top \end{pmatrix} \tag{2.4}$$

and

$$A^\top = (V_1, V_2) \begin{pmatrix} \widehat{\Sigma} & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} U_1^\top \\ U_2^\top \end{pmatrix}. \tag{2.5}$$

The following lemma shows the orthogonal decompositions of $\mathbb{R}^m$ and $\mathbb{R}^n$ defined by A and $A^\top$, respectively.

Lemma 2.1 *It holds:*

$$\mathbb{R}^m = \operatorname{Im} A \oplus \operatorname{Ker} A^\top \quad \textit{and} \quad \operatorname{Im} A \perp \operatorname{Ker} A^\top; \tag{2.6}$$

$$\mathbb{R}^n = \operatorname{Im} A^\top \oplus \operatorname{Ker} A \quad \textit{and} \quad \operatorname{Im} A^\top \perp \operatorname{Ker} A. \tag{2.7}$$

Proof. The columns of U from the SVD of A are basis in $\mathbb{R}^m$. Further (2.4) and (2.5) imply that the columns of U_1 and U_2 are the orthogonal bases in $\operatorname{Im} A$ and $\operatorname{Ker} A^\top$, respectively. The relations (2.6) follow from $U = (U_1, U_2)$. The proof of (2.7) is analogous. □

Now we give the definition of the MP inverse based on the SVD.

Definition 2.1 *Let $A \in \mathbb{R}^{m\times n}$ be given. By the MP inverse to A we call $A^\dagger$ defined by*

$$A^\dagger := (V_1, V_2) \begin{pmatrix} \widehat{\Sigma}^{-1} & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} U_1^\top \\ U_2^\top \end{pmatrix}, \tag{2.8}$$

where U, V, and $\widehat{\Sigma}$ are determined by the SVD of A.

It is easy to verify that the MP inverse given by (2.8) is the generalized inverse, i.e., (2.1) is satisfied for $X = A^\dagger$. The following lemma follows immediately.

Lemma 2.2 *Let $A^\dagger$ be the MP inverse to A. Then*

$$\operatorname{Ker} A^\dagger = \operatorname{Ker} A^\top, \qquad \operatorname{Im} A^\dagger = \operatorname{Im} A^\top. \tag{2.9}$$

Proof. Compare (2.8) and (2.5). □

Remark 2.1 Let us note that Moore's definition [15, 11] of the MP inverse, more or less forgotten, tell us that the MP inverse $A^\dagger$ is fully determined by the following three conditions:

$$A = AA^\dagger A, \quad \operatorname{Im} A^\dagger \subseteq \operatorname{Im} A^\top, \quad \operatorname{Im}(A^\dagger)^\top \subseteq \operatorname{Im} A.$$

We will prove, in spite of Moor's definition, that the MP-invese is fully determined by (2.9). We will need several auxiliary results. First of all we recall the well-known conditions used by Penrose [16] to define the MP inverse.

Lemma 2.3 *Let $A \in \mathbb{R}^{m\times n}$ be given. Then X is the MP inverse to A, i.e. $X = A^\dagger$, iff*

$$A = AXA, \quad XA = (XA)^\top, \quad AX = (AX)^\top, \quad X = XAX. \tag{2.10}$$

Proof. There is a unique X determined by (2.10); see [1]. One can verify that $X = A^\dagger$ given by (2.8) satisfies all equations (2.10). □

Lemma 2.4 *Let P be a square matrix. If $\operatorname{Im} P \perp \operatorname{Im}(I-P)$, then P is symmetric.*

Proof. The orthogonality of $\operatorname{Im} P$ and $\operatorname{Im}(I-P)$ is equivalent to $P^\top(I-P) = 0$ so that $P^\top = P^\top P = (P^\top P)^\top = (P^\top)^\top = P$. □

Any square matrix P satisfying $P^2 = P$ is called the *projector* on $\operatorname{Im} P$. Moreover, if $\operatorname{Im} P \perp \operatorname{Im}(I-P)$, then the projector P is called *orthogonal*.

Lemma 2.5 *Let P be a projector. Then P is orthogonal iff it is symmetric.*

Proof: As Lemma 2.4 holds, it remains to prove that the symmetry implies the orthogonality. It is $P^\top(I-P) = P^\top - P^\top P = P - P^2 = 0$. □

Lemma 2.6 *Let X be a generalized inverse to A. Then*
(i) $Y := AX$ is the projector on $\operatorname{Im} A$*;*
(ii) $I - Y^\top$ is the projector on $\operatorname{Ker} A^\top$*;*
(iii) $Z := (XA)^\top$ is the projector on $\operatorname{Im} A^\top$*;*
(iv) $I - Z^\top$ is the projector on $\operatorname{Ker} A$*.*

Proof: As $Y^2 = AXAX = AX = Y$, we see that Y is the projector on $\operatorname{Im} Y \subseteq \operatorname{Im} A$. For $x \in \operatorname{Im} A$, $x = Ay$, we obtain $Yx = AXAy = Ay = x$ so that $\operatorname{Im} Y = \operatorname{Im} A$ and (i) holds. Further, $(I - Y^\top)^2 = I - 2Y^\top + (Y^\top)^2 = I - Y^\top$ implies that $I - Y^\top$ is the projector on $\operatorname{Im}(I - Y^\top)$. Moreover, $A^\top(I - Y^\top) = A^\top - A^\top X^\top A^\top = 0$ yields $\operatorname{Im}(I - Y^\top) \subseteq \operatorname{Ker} A^\top$. For $x \in \operatorname{Ker} A^\top$, we have $(I - Y^\top)x = x - X^\top A^\top x = x$ so that $\operatorname{Im}(I - Y^\top) = \operatorname{Ker} A^\top$ and therefore (ii) is true. The proof of (iii) and (iv) is analogous. □

Corollary 2.1 *Let X be a generalized inverse to A. It holds:*
(a) $\operatorname{Im} AX = \operatorname{Im} A$*;*
(b) $\operatorname{Im}(I - (AX)^\top) = \operatorname{Ker} A^\top$*;*
(c) $\operatorname{Im}(XA)^\top = \operatorname{Im} A^\top$*;*
(d) $\operatorname{Im}(I - XA) = \operatorname{Ker} A$*;*
(e) $\operatorname{Ker}(AX)^\top = \operatorname{Ker} A^\top$*;*
(f) $\operatorname{Ker}(I - AX) = \operatorname{Im} A$*;*
(g) $\operatorname{Ker} XA = \operatorname{Ker} A$*;*
(h) $\operatorname{Ker}(I - (XA)^\top) = \operatorname{Im} A^\top$*.*

Proof. The statements (a)-(d) follow from Lemma 2.6 and (e)-(h) are the equalities of the respective orthogonal complements. □

Now we prove the three conditions determining the MP inverse.

Theorem 2.2 *Let $A \in \mathbb{R}^{m\times n}$ be given. Then X is the MP inverse to A, i.e. $X = A^\dagger$, iff*
(i) $A = AXA$,
(ii) $\operatorname{Im} X = \operatorname{Im} A^\top$*,*
(iii) $\operatorname{Ker} X = \operatorname{Ker} A^\top$*.*

Proof. As Lemma 2.2 holds, it remains to prove the converse implication "⇐". We will verify $(2.10)_1$-$(2.10)_4$. The first condition $(2.10)_1$ holds since it is (i). The trivial inclusion $\operatorname{Im} X \supseteq \operatorname{Im} XA$ and (ii) imply $\operatorname{Im} A^\top \supseteq \operatorname{Im} XA$. By Corollary 2.1(c) we obtain $r(A^\top) = r((XA)^\top) = r(XA)$ so that $\operatorname{Im} A^\top = \operatorname{Im} XA$. Using (2.7) and Corollary 2.1(d) we arrive at

$$\operatorname{Im} XA = \operatorname{Im} A^\top \perp \operatorname{Ker} A \ = \operatorname{Im}(I - XA).$$

Therefore Lemma 2.4 implies $XA = (XA)^\top$ that is $(2.10)_2$. To prove $(2.10)_3$ we start with (iii) and Corollary 2.1(e) that give $\operatorname{Ker} X = \operatorname{Ker}(AX)^\top$. Passing to the

orthogonal complements we get $\operatorname{Im} X^\top = \operatorname{Im} AX$ that yields $r(X^\top) = r(AX) = r((AX)^\top)$. This equality together with the obvious inclusion $\operatorname{Im} X^\top \supseteq \operatorname{Im}(AX)^\top$ lead to $\operatorname{Im} X^\top = \operatorname{Im}(AX)^\top$. By (2.7), again (iii), and Corollary 2.1(b) we get

$$\operatorname{Im}(AX)^\top = \operatorname{Im} X^\top \perp \operatorname{Ker} X = \operatorname{Ker} A^\top = \operatorname{Im}(I - (AX)^\top).$$

Now Lemma 2.4 implies $AX = (AX)^\top$ that is $(2.10)_3$. Moreover we obtain

$$\operatorname{Im} X^\top \perp \operatorname{Im}(I - AX)$$

or, equivalently, $X(I - AX) = 0$ that proves $(2.10)_4$. □

Example 2.1 Let us consider the (symmetric) matrix

$$A = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$$

with $\operatorname{Im} A$ and $\operatorname{Ker} A$ determined by the vectors $(1,1)^\top$ and $(1,-1)^\top$, respectively; see Figure 2.1. Let us define $S_X := \sum_{ij} x_{ij}$ for $X = (x_{ij}) \in \mathbb{R}^{2\times 2}$. It is readily

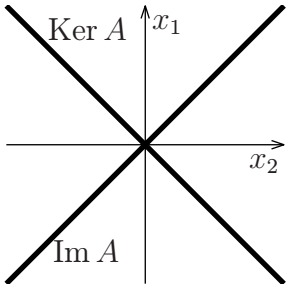


Figure 2.1: Given A.

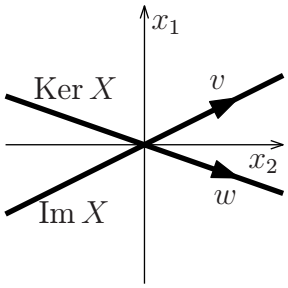


Figure 2.2: Arbitrary generalized inverse X.

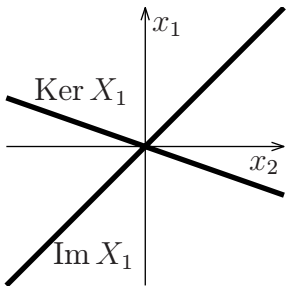


Figure 2.3: Generalized inverse X_1 with arbitrary $\operatorname{Ker} X_1$.

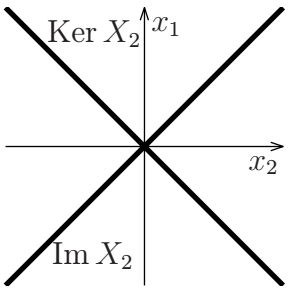


Figure 2.4: MP inverse $X_2 = A^\dagger$.

seen that $AXA = A$ is equivalent with $S_X = 1$. Therefore the generalized inverse to A can be obtained from each $M \in \mathbb{R}^{2\times 2}$ such that $S_M \neq 0$ by

$$X := S_M^{-1} M. \tag{2.11}$$

We will show how to modify an arbitrary generalized inverse X into the MP one. Let us introduce M with the image and the kernel generated by the nonzero vectors $v = (v_1, v_2)^\top$ and $w = (w_1, w_2)$, respectively, i.e.,

$$M := \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} (w_2, -w_1).$$

Let us define the generalized inverse X by (2.11); see Figure 2.2. Note that

$$S_M = (v_1 + v_2)(w_2 - w_1) \tag{2.12}$$

and $S_M \neq 0$ implies $v_1 \neq -v_2$ and $w_1 \neq w_2$ or, in other words, $\operatorname{Im} X \neq \operatorname{Ker} A$ and $\operatorname{Ker} X \neq \operatorname{Im} A$ for any generalized inverse X. The orthogonal projector on $\operatorname{Im} A$ reads as follows:

$$P_A = \begin{pmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{pmatrix}.$$

Now

$$X_1 := P_A X = S_M^{-1} \begin{pmatrix} (v_1 + v_2)/2 \\ (v_1 + v_2)/2 \end{pmatrix} (w_2, -w_1)$$

is the generalized inverse with $\operatorname{Im} X_1 = \operatorname{Im} A$ and $\operatorname{Ker} X_1 = \operatorname{Ker} X$; see Figure 2.3. Further

$$X_2 := X_1 P_A = S_M^{-1} \begin{pmatrix} (v_1 + v_2)/2 \\ (v_1 + v_2)/2 \end{pmatrix} ((w_2 - w_1)/2, (w_2 - w_1)/2)$$

is the generalized inverse with $\operatorname{Im} X_2 = \operatorname{Im} A$ and $\operatorname{Ker} X_2 = \operatorname{Ker} A$; see Figure 2.4. Therefore $X_2 = A^\dagger$ by Theorem 2.2 and, moreover due to (2.12), it follows

$$A^\dagger = \begin{pmatrix} 1/2 \\ 1/2 \end{pmatrix} (1/2, 1/2) = \begin{pmatrix} 1/4 & 1/4 \\ 1/4 & 1/4 \end{pmatrix}.$$

The observations of the example hold in general.

Theorem 2.3 *Let $A \in \mathbb{R}^{m\times n}$ be given. Let X be an arbitrary generalized inverse to A and let P_A and $P_{A^\top}$ be the orthogonal projectors on* $\operatorname{Im} A$ *and* $\operatorname{Im} A^\top$*, respectively. Then*

$$A^\dagger = P_{A^\top} X P_A. \tag{2.13}$$

Proof. Notice that P_A, $P_{A^\top}$ are symmetric by Lemma 2.5. We obtain $P_A A = A$ and $AP_{A^\top} = (P_{A^\top} A^\top)^\top = (A^\top)^\top = A$. Now we will verify that $A^\dagger$ defined by (2.13) satisfies $(2.10)_1$-$(2.10)_4$. The first equality $(2.10)_1$ is straightforward since

$$AA^\dagger A \ = AP_{A^\top} X P_A A \ = AXA = A. \tag{2.14}$$

To prove $(2.10)_2$ we take arbitrary $x, y \in \mathbb{R}^n$ and consider respective $z_x, z_y \in \mathbb{R}^m$ so that $P_{A^\top} x = A^\top z_x$, $P_{A^\top} y = A^\top z_y$. Then

$$\begin{aligned} x^\top A^\dagger A\ y &= x^\top P_{A^\top} X P_A A\ y \ = \ x^\top P_{A^\top} X A\ y \ = \ z_x^\top AXA\ y \ = \ z_x^\top A\ y \ = \ x^\top P_{A^\top} y \\ &= x^\top A^\top z_y \ = \ x^\top A^\top X^\top A^\top z_y \ = \ x^\top A^\top X^\top P_{A^\top} y \ = \ x^\top (P_{A^\top} XA)^\top y \\ &= x^\top (P_{A^\top} X P_A A)^\top y \ = \ x^\top (A^\dagger A)^\top y \end{aligned}$$

yields $(2.10)_2$. The proof of $(2.10)_3$ is analogous. Finally, let us consider $x \in \mathbb{R}^n$, $y \in \mathbb{R}^m$, and $z_y \in \mathbb{R}^n$ so that $P_A y = A z_y$. We derive

$$\begin{aligned} x^\top A^\dagger A\ A^\dagger y &= x^\top A^\dagger A\ P_{A^\top} X P_A y \ = \ x^\top A^\dagger A\ P_{A^\top} X A z_y \ = \ x^\top A^\dagger AXA z_y \\ &= x^\top A^\dagger A\ z_y \ = \ x^\top A^\dagger P_A y \ = \ x^\top P_{A^\top} X P_A P_A y \\ &= x^\top A^\dagger y \end{aligned}$$

that proves $(2.10)_4$. □

Remark 2.2 The orthogonal projectors in (2.13) can be expressed by the SVD (2.4) as $P_A = I - V_2 V_2^\top$ and $P_{A^\top} = I - U_2 U_2^\top$. It is easily seen that V_2 and U_2 in P_A and $P_{A^\top}$ may be replaced by an arbitrary matrix whose columns form the orthogonal bases in $\operatorname{Ker} A$ and $\operatorname{Ker} A^\top$, respectively. In the next section, we will assume that the knowledge of such bases is an à-priori information about our problem.

Remark 2.3 If $m = n$ and A is symmetric, then (2.13) simplifies into $A^\dagger = P_A X P_A$. The necessary and sufficient conditions characterizing the MP inverse take the form:

$$A = AXA, \quad \operatorname{Im} X = \operatorname{Im} A, \quad X \text{ is symmetric},$$

or

$$A = AXA, \quad \operatorname{Ker} X = \operatorname{Ker} A, \quad X \text{ is symmetric}.$$

3 MP inverse in saddle-point systems

This section deals with solving saddle-point linear systems with singular diagonal blocks by the method combing the Schur complement reduction with orthogonal projectors. This solution strategy is an algebraic background for various

variants of the FETI method [8, 4]. In terms of the standard saddle-point terminology [2] it is the combination of the range-space method and the null-space method applied to the primal and the dual saddle-point system, respectively. We shall see that the use of the MP inverse based on (2.13) is natural and it simplifies both the implementation as well as the analysis.

First of all we introduce notation. Let $\mathbb{V} \subseteq \mathbb{R}^q$ be a subspace. The kernel and the image of any matrix $M \in \mathbb{R}^{p\times q}$ on $\mathbb{V}$ will be denoted by $\mathrm{Ker}(M|\mathbb{V})$ and $\mathrm{Im}(M|\mathbb{V})$, respectively. If M is symmetric, positive semi-definite (with $p = q$) on $\mathbb{V}$, we will denote the largest eigenvalue on $\mathbb{V}$ by $\sigma_{\max}(M|\mathbb{V})$ and the smallest eigenvalue on $\mathbb{V}$ by $\sigma_{\min}(M|\mathbb{V})$. The spectral condition number of M on $\mathbb{V}$ is defined by

$$\kappa(M|\mathbb{V}) := \frac{\sigma_{\max}(M|\mathbb{V})}{\sigma_{\min}(M|\mathbb{V})}.$$

Moreover, when $\mathbb{V} = \mathbb{R}^q$, we write as before $\mathrm{Ker}\, M = \mathrm{Ker}(M|\mathbb{V})$, $\mathrm{Im}\, M = \mathrm{Im}(M|\mathbb{V})$, and $\sigma_{\min}(M) = \sigma_{\min}(M|\mathbb{V})$, $\sigma_{\max}(M) = \sigma_{\max}(M|\mathbb{V})$, $\kappa(M) := \kappa(M|\mathbb{V})$.

Let us note that $0 < \sigma_{\min}(M|\,\mathrm{Im}\, M)$, $\sigma_{\max}(M|\,\mathrm{Im}\, M) = \sigma_{\max}(M)$, and $\kappa(M|\,\mathrm{Im}\, M) < +\infty$, if M is the non-zero matrix.

3.1 Algorithm

We shall consider the problem for finding $(\bar{u}, \bar{\lambda}) \in \mathbb{R}^n \times \mathbb{R}^m$ satisfying:

$$\mathcal{A}\begin{pmatrix} u \\ \lambda \end{pmatrix} = \begin{pmatrix} f \\ g \end{pmatrix} \tag{3.1}$$

with the saddle-point matrix

$$\mathcal{A} := \begin{pmatrix} A & B^\top \\ B & 0 \end{pmatrix},$$

where $A \in \mathbb{R}^{n\times n}$ is symmetric, positive semi-definite, and singular, $B \in \mathbb{R}^{m\times n}$, $f \in \mathbb{R}^n$, and $g \in \mathbb{R}^m$. We suppose that (3.1) is uniquely solvable that is guaranteed by the following necessary and sufficient conditions [12]:

$$\mathrm{Ker}\, B^\top = \{0\}, \tag{3.2}$$

$$\mathrm{Ker}\, A \cap \mathrm{Ker}\, B = \{0\}. \tag{3.3}$$

Notice that (3.2) is the condition on the full row-rank of B. Moreover, we assume that an orthonormal basis of $\mathrm{Ker}\, A$ is known à-priori and that its vectors are columns of $R \in \mathbb{R}^{n\times l}$, $l = n - r(A)$. Then $P_A = I - RR^\top$ is the orthogonal projector on $\mathrm{Im}\, A$ and the MP inverse to A is given by Theorem 2.3 as

$$A^\dagger = (I - RR^\top)X(I - RR^\top), \tag{3.4}$$

where X is an arbitrary generalized inverse to A. Under our assumptions X is easily available by a variant of the Cholesky factorization [5].

The first equation in (3.1) is satisfied iff

$$f - B^\top \bar{\lambda} \in \operatorname{Im} A \tag{3.5}$$

and

$$\bar{u} = A^\dagger (f - B^\top \bar{\lambda}) + R\bar{\alpha} \tag{3.6}$$

for an appropriate $\bar{\alpha} \in \mathbb{R}^l$. Let us note that $A^\dagger(f - B^\top \bar{\lambda}) \in \operatorname{Im} A$ and $R\bar{\alpha} \in \operatorname{Ker} A$. Since $\operatorname{Im} A$ is the orthogonal complement of $\operatorname{Ker} A$, $\bar{\alpha}$ is determined uniquely by (3.6) and, moreover, (3.5) can be equivalently written as

$$R^\top (f - B^\top \bar{\lambda}) = 0. \tag{3.7}$$

Further substituting (3.6) into the second equation in (3.1) we arrive at

$$-BA^\dagger B^\top \bar{\lambda} + BR\bar{\alpha} = g - BA^\dagger f. \tag{3.8}$$

Summarizing (3.8) and (3.7) we find that the pair $(\bar{\lambda}, \bar{\alpha}) \in \mathbb{R}^m \times \mathbb{R}^l$ satisfies:

$$\mathcal{S} \begin{pmatrix} \lambda \\ \alpha \end{pmatrix} = \begin{pmatrix} d \\ e \end{pmatrix}, \tag{3.9}$$

where

$$\mathcal{S} := \begin{pmatrix} BA^\dagger B^\top & -BR \\ -R^\top B^\top & 0 \end{pmatrix}$$

is the (negative) *Schur complement* of A in $\mathcal{A}$, $d := BA^\dagger f - g$, and $e := -R^\top f$. As both $\mathcal{S}$ and $\mathcal{A}$ are simultaneously invertible [12], we can compute first $(\bar{\lambda}, \bar{\alpha})$ by solving (3.9) and then we obtain $\bar{u}$ from (3.6). Let us note that (3.9) has formally the same saddle-point structure as that of (3.1), however, its size is considerably smaller.

Before discussing the solution method for (3.9) we introduce new notation

$$F := BA^\dagger B^\top, \quad G := -R^\top B^\top$$

which renders (3.9) into

$$\begin{pmatrix} F & G^\top \\ G & 0 \end{pmatrix} \begin{pmatrix} \lambda \\ \alpha \end{pmatrix} = \begin{pmatrix} d \\ e \end{pmatrix}. \tag{3.10}$$

Now we shall split (3.10) using the orthogonal projector P_G on $\operatorname{Ker} G$. As (3.3) implies that G is of full row-rank, we can identify P_G with the following matrix:

$$P_G := I - G^\top (GG^\top)^{-1} G.$$

Applying P_G on the first equation in (3.10) we obtain that $\bar{\lambda}$ satisfies:

$$P_G F\lambda = P_G d, \quad G\lambda = e. \tag{3.11}$$

In order to arrange (3.11) as one equation on the vector space $\operatorname{Ker} G$ we decompose the solution $\bar{\lambda}$ into $\bar{\lambda}_{\mathrm{Im}} \in \operatorname{Im} G^\top$ and $\bar{\lambda}_{\mathrm{Ker}} \in \operatorname{Ker} G$ as

$$\bar{\lambda} = \bar{\lambda}_{\mathrm{Im}} + \bar{\lambda}_{\mathrm{Ker}}. \tag{3.12}$$

Since $\bar{\lambda}_{\mathrm{Im}}$ is easily available via

$$\bar{\lambda}_{\mathrm{Im}} = G^\top (GG^\top)^{-1} e,$$

it remains to show how to get $\bar{\lambda}_{\mathrm{Ker}}$. Substituting (3.12) into (3.11) we can see that $\bar{\lambda}_{\mathrm{Ker}}$ satisfies:

$$P_G F\lambda_{\mathrm{Ker}} = P_G(d - F\bar{\lambda}_{\mathrm{Im}}), \quad \lambda_{\mathrm{Ker}} \in \operatorname{Ker} G. \tag{3.13}$$

Let us note that this equation is uniquely solvable, as $P_G F : \operatorname{Ker} G \mapsto \operatorname{Ker} G$ is invertible iff $\mathcal{A}$ is invertible [12]. Finally note that, if $\bar{\lambda}$ is known, the solution component $\bar{\alpha}$ is given by

$$\bar{\alpha} = (GG^\top)^{-1} G(d - F\bar{\lambda}). \tag{3.14}$$

3.2 Eigenvalue analysis

The key point of the presented algorithm is the equation (3.13). Its solution can be computed by the projected variant of the conjugate gradient method [8]. As the rate of convergence of this method is determined by the condition number [10], we shall analyze bounds on the eigenvalues of $P_G F$ on $\operatorname{Ker} G$.

First note that $P_G F$ is symmetric on $\operatorname{Ker} G$:

$$\mu^\top P_G F\nu = \mu^\top P_G F P_G \nu = \nu^\top P_G^\top F^\top P_G^\top \mu = \nu^\top P_G F\mu \quad \forall \mu, \nu \in \operatorname{Ker} G.$$

It is also easy to see that $P_G F$ is positive semi-definite on $\operatorname{Ker} G$:

$$\mu^\top P_G F\mu = \mu^\top P_G B A^\dagger B^\top P_G \mu \geq 0 \quad \forall \mu \in \operatorname{Ker} G.$$

As $P_G F$ is invertible on $\operatorname{Ker} G$, one can deduce that it is positive definite on $\operatorname{Ker} G$. Bellow we will prove an positive lower bound for the smallest eigenvalue of $P_G F$ on $\operatorname{Ker} G$. To this end, we will assume that there are constants $0 < c_{A,1} < c_{A,2}$ such that

$$c_{A,1} \leq \sigma_{\min}(A|\operatorname{Im} A) \quad \text{and} \quad \sigma_{\max}(A) \leq c_{A,2}. \tag{3.15}$$

Moreover, as the matrix $BB^\top$ is positive definite due to (3.2), there are constants $0 < c_{B,1} < c_{B,2}$ such that

$$c_{B,1} \leq \sigma_{\min}(BB^\top) \quad \text{and} \quad \sigma_{\max}(BB^\top) \leq c_{B,2}. \tag{3.16}$$

We obtain immediately the following result.

Lemma 3.1 *Let $A^\dagger$ be the MP inverse to symmetric, positive definite A. Then*

$$c_{A,2}^{-1} \le \sigma_{\min}(A^\dagger|\operatorname{Im} A), \quad \sigma_{\max}(A^\dagger) \le c_{A,1}^{-1}. \tag{3.17}$$

Proof: It follows from the definition (2.8) of $A^\dagger$, since the non-zero eigenvalues of A are the diagonal entries of $\widehat{\Sigma}$ in the SVD (2.4), and $\operatorname{Im} A = \operatorname{Im} A^\dagger$. □

Now we shall prove the main result of this section.

Theorem 3.1 *Let $P_G F$ be the operator of* (3.13)*. Then*

$$c_{A,2}^{-1} c_{B,1} \le \sigma_{\min}(P_G F|\operatorname{Ker} G), \quad \sigma_{\max}(P_G F|\operatorname{Ker} G) \le c_{A,1}^{-1} c_{B,2}, \tag{3.18}$$

and

$$\kappa(P_G F|\operatorname{Ker} G) \le \frac{c_{B,2}}{c_{B,1}} \cdot \frac{c_{A,2}}{c_{A,1}}. \tag{3.19}$$

Proof: As the proofs of both bounds (3.18) are analogous, we confine ourself to the lower bound:

$$\sigma_{\min}(P_G F|\operatorname{Ker} G) = \min_{\substack{\mu\in\operatorname{Ker} G\\ \mu\neq 0}} \frac{\mu^\top P_G F \mu}{\mu^\top \mu} = \min_{\substack{R^\top B^\top \mu = 0\\ \mu\neq 0}} \frac{\mu^\top B A^\dagger B^\top \mu}{\mu^\top \mu}$$

$$= \min_{\substack{R^\top v = 0\\ v = B^\top \mu\\ \mu\neq 0}} \frac{v^\top A^\dagger v}{v^\top v} \cdot \frac{\mu^\top B B^\top \mu}{\mu^\top \mu} \ge \min_{\substack{v\in\operatorname{Im} A\cap\operatorname{Im} B^\top\\ v\neq 0}} \frac{v^\top A^\dagger v}{v^\top v} \cdot \min_{\substack{\mu\in\operatorname{Ker} G\\ \mu\neq 0}} \frac{\mu^\top B B^\top \mu}{\mu^\top \mu}.$$

Further using (3.16),

$$\min_{\substack{\mu\in\operatorname{Ker} G\\ \mu\neq 0}} \frac{\mu^\top B B^\top \mu}{\mu^\top \mu} = \sigma_{\min}(BB^\top|\operatorname{Ker} G) \ge \sigma_{\min}(BB^\top) \ge c_{B,1}$$

and, by Lemma 3.1,

$$\min_{\substack{v\in\operatorname{Im} A\cap\operatorname{Im} B^\top\\ v\neq 0}} \frac{v^\top A^\dagger v}{v^\top v} \ge \min_{\substack{v\in\operatorname{Im} A\\ v\neq 0}} \frac{v^\top A^\dagger v}{v^\top v} = \sigma_{\min}(A^\dagger|\operatorname{Im} A) \ge c_{A,2}^{-1}.$$

Therefore

$$\sigma_{\min}(P_G F|\operatorname{Ker} G) \ge c_{A,2}^{-1} c_{B,1}$$

that is the lower bound. The inequality (3.19) follows immediately from (3.18). □

Remark 3.1 Let the MP inverse $A^\dagger$ in F be replaced by an arbitrary generalized inverse X satisfying solely $AXA = A$. Then Theorem 3.1 remains valid. To prove this result it is enough to show that the eigenvalue bounds (3.17) does not depend on the choice of a generalized inverse:

$$\sigma_{\min}(X|\operatorname{Im} A) = \min_{\substack{v\in\operatorname{Im} A\\ v\neq 0}} \frac{v^\top X v}{v^\top v} = \min_{\substack{w=w_0+w_1\\ w_0\in\operatorname{Ker} A, w_1\in\operatorname{Im} A\\ w_1\neq 0}} \frac{w^\top AXAw}{w^\top A^2 w} = \min_{\substack{w_1\in\operatorname{Im} A\\ w_1\neq 0}} \frac{w_1^\top A\ w_1}{w_1^\top A^2 w_1}.$$

Therefore

$$\sigma_{\min}(X|\operatorname{Im} A) = \sigma_{\min}(A^\dagger|\operatorname{Im} A) \geq c_{A,2}^{-1}$$

and analogously for the largest eigenvalue.

Remark 3.2 When B is not full-row rank matrix, then $P_G F$ is singular on $\operatorname{Ker} G$. In particular, there is $\mu_0 \in \operatorname{Ker} B^\top$, $\mu_0 \neq 0$, $B^\top \mu_0 = 0$, and $G\mu_0 = R^\top B^\top \mu_0 = 0$, for which

$$\sigma_{\min}(P_G F|\operatorname{Ker} G) = \frac{\mu_0^\top B A^\dagger B^\top \mu_0}{\mu_0^\top \mu_0} = 0.$$

4 Application in the TFETI method

Applying previous results to the saddle-point osystem (3.1) arising from the TFETI method [4] we will prove that the condition number of $P_G F$ on $\operatorname{Ker} G$ may not depend on the size of the problem. First we mention main principles of the TFETI method.

The FETI as well as the TFETI methods belong to the class of non-overlapping domain decomposition methods proposed for the parallel solution of boundary value problems described by elliptic PDEs on a bounded domain $\Omega \subset \mathbb{R}^d$, $d = 1, 2, 3$. Let L denote the diameter of Ω. The idea consists of decomposing Ω into sub-domains Ω_k, $k = 1, \ldots, s$, so that $\overline{\Omega} = \bigcup_{k=1}^s \overline{\Omega}_k$ and $\Omega_k \cap \Omega_l = \emptyset$, $k \neq l$, and considering the PDEs independently on each Ω_k. Therefore the corresponding stiffness matrix A exhibits the block diagonal structure

$$A = diag\,(A_1, \ldots, A_s) \tag{4.1}$$

with $A_k \in \mathbb{R}^{N_k \times N_k}$ being symmetric positive semi-definite. Let us note that the number of sub-domains s is typically proportional to $(L/H)^d$, where H is the diameter of the largest Ω_k, while the size N_k of A_k corresponds to $(H/h)^d$, where h is the element norm of the finite element approximation. The sub-domain interconnectivity is enforced by the second equation in (3.1), in which

$$B = (B_1, \ldots, B_s) \tag{4.2}$$

with $B_k \in \mathbb{R}^{m \times N_k}$, where m is proportional to $\sum_{k=1}^s N_k^{(d-1)/d}$. Let us note that finite element nodes shared by more than two sub-domains generate usually linearly dependent rows in B. In agreement with (3.2) we will assume that this

redundancy is eliminated from B and that the resulting full-row rank matrix is denoted by B again.

The TFETI method enforces also the Dirichlet boundary conditions through the matrix B. The main advantage of this strategy is the fact that for each Ω_k we generate the corresponding block A_k of A as the stiffness matrix to the original PDEs with the pure homogeneous Neumann conditions on the boundary of Ω_k. Consequently, all blocks A_k exhibit the same kernel dimension, say l, and their kernel basis may be identified by a mechanical interpretation of the PDEs [4]. In particular, we can assemble the basis for $\operatorname{Ker} A$ in the matrix $R \in \mathbb{R}^{n\times sl}$, $n = \sum_{k=1}^{s} N_k$, with the following block diagonal structure:

$$R = diag(R_1, \ldots, R_s), \tag{4.3}$$

where $R_k \in \mathbb{R}^{N_k\times l}$ may be obtained without any computation. As (3.4) requires orthogonality of R, we shall assume that columns of R_k are orthogonalized and that the resulting orthogonal matrix is denoted by R_k again. Let us note that the orthogonalization procedure (if it is necessary) is cheap due to the special structure of R_k.

4.1 Model problem in 1D

Let $\Omega = (0, L)$, $L > 0$. Let us consider the following problem:

$$-u'' = b \ \text{ in } \Omega, \quad u(0) = 0, \quad u'(L) = 0, \tag{4.4}$$

where $b \in C(\Omega)$. Let all sub-domains Ω_k of Ω be of the same lengths $H = L/s$ so that $\Omega_k = ((k-1)H, kH)$, $k = 1, \ldots, s$. On each Ω_k we consider an equidistant partition with N nodes so that $h = H/(N-1)$. Note that the decomposition parameter H and the discretization parameter h satisfy:

$$N = 1 + \frac{H}{h}. \tag{4.5}$$

The approximation of (4.4) based on the TFETI method with the linear finite elements leads to the blocks in (4.1) given by $A_k = A(h, N) \in \mathbb{R}^{N\times N}$, where

$$A(h,N) = \frac{1}{h}\begin{pmatrix} 1 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 1 \end{pmatrix}. \tag{4.6}$$

In each block $B_k \in \mathbb{R}^{s\times N}$ of (4.2) there are at most two non-zero entries, i.e. "1" in the first position of the kth row and "-1" in the last position of the $(k+1)$th

row (note that B_s contains only one "-1" at the beginning of the last row). Finally, all blocks $R_k \in \mathbb{R}^{N\times 1}$ of (4.3) read as follows:

$$R_k = \frac{1}{\sqrt{N}}(1,\dots,1)^\top. \tag{4.7}$$

Lemma 4.1 *The eigenvalues $\sigma_k = \sigma_k(h,N)$ of $A(h,N)$ read as follows:*

$$\sigma_k = \frac{1}{h}(2-2\cos\theta_k), \quad \theta_k = \frac{k\pi}{N}, \quad k = 0,1,\dots,N-1.$$

Proof: Using the standard trigonometric formulas one can verify that $v_k = (\cos(j+\frac{1}{2})\theta_k)_{j=0}^{N-1}$ is the eigenvector corresponding to σ_k; see [17]. □

Theorem 4.1 *Let $N \geq 4$. Let $P_G F$ be the operator of* (3.13) *given by* (4.1)-(4.3) *arising from the TFETI method applied to* (4.4)*. It holds:*

$$\frac{h}{4} \leq \sigma_{\min}(P_G F|\operatorname{Ker} G),$$

$$\sigma_{\max}(P_G F|\operatorname{Ker} G) \leq \frac{24}{11\pi^2}\cdot\frac{(h+H)^2}{h},$$

$$\kappa(P_G F|\operatorname{Ker} G) \leq \frac{96}{11\pi^2}\left(1+\frac{H}{h}\right)^2.$$

Proof: We estimate the constants from (3.15) and (3.16). As $BB^\top = diag(1,2,\dots,\dots,2) \in \mathbb{R}^{s\times s}$, we obtain immediately $c_{B,1} = 1$ and $c_{B,2} = 2$. As all diagonal blocks of A are $A(h,N)$, it follows from Lemma 4.1 that

$$\sigma_{\min}(A|\operatorname{Im} A) = \sigma_1 \quad \text{and} \quad \sigma_{\max}(A) = \sigma_{N-1}.$$

Therefore we can take $c_{A,2} = 4/h$. To estimate $c_{A,1}$ we use the Taylor expansion for $\cos\theta_1$:

$$\sigma_1 = \frac{2}{h}(\theta_1^2/2! - \theta_1^4/4! + \theta_1^6/6! - \theta_1^8/8! - \dots).$$

As $N \geq 4$, we have $\theta_1 \leq \pi/4 < 1$ so that $\theta_1^{2j}/(2j)! - \theta_1^{2(j+1)}/(2(j+1))! \geq 0$ and therefore

$$\sigma_1 \geq \frac{2}{h}(\theta_1^2/2! - \theta_1^4/4!).$$

Further $\theta_1^2 \geq \theta_1^4$ implies

$$\sigma_1 \geq \frac{11}{12h}\theta_1^2 = c_{A,1}.$$

Substituting $\theta_1 = \pi/N$ and (4.5) we arrive at

$$c_{A,1} = \frac{11\pi^2 h}{12(h+H)^2}.$$

The rest of the proof consists of using Theorem 3.1. □

Remark 4.1 For $N = 3$, we can take $c_{A,1} = \sigma_1 = 1/h$.

4.2 Scalar model problem in 2D

Let $\Omega = (0, L_x) \times (0, L_y)$, $L_x, L_y > 0$. Let us consider the Poisson problem:

$$-\Delta u \ = b \ \text{ in } \Omega, \quad u \ = 0 \ \text{ on } \gamma_d, \quad \frac{\partial u}{\partial \nu} = 0 \ \text{ on } \gamma_n, \tag{4.8}$$

where $\gamma_d = \{0\} \times (0, L_y)$, $\gamma_n = \partial\Omega \setminus \overline{\gamma}_d$, $b \in C(\Omega)$, and ν denotes the unit outer normal vector to the boundary $\partial\Omega$. Let all sub-domains of Ω be rectangles $\Omega_k = \Omega_{k_x} \times \Omega_{k_y}$, where $\Omega_{k_z} = ((k_z - 1)H_z, k_z H_z)$, $k_z = 1, \ldots, s_z$, $H_z = L_z/s_z$ for $z = x, y$. The number of Ω_k is $s = s_x s_y$ and the correspondence between k_x, k_y and k is given by $k = k_x + (k_y - 1)s_x$. Let us construct equidistant partitions of the sides of Ω_k with the same stepsizes $h_x = H_x/(N_x - 1)$ and $h_y = H_y/(N_y - 1)$ for all k. Thus, each Ω_k is partitioned by $N_k = N_x N_y$ nodes into $(N_x - 1)(N_y - 1)$ rectangles. Finally, we assume that each of the rectangles is cut by its diagonal into two triangles. On this triangulation of Ω_k we define the finite-element space of continuous piecewise linear functions that is used to approximate the solution to (4.8) by the TFETI method. Let us note that $h = (h_x^2 + h_y^2)^{1/2}$ and $H = (H_x^2 + H_y^2)^{1/2}$. The blocks A_k in (4.1) are given by

$$A_k = A_x \otimes I_y + I_x \otimes A_y, \tag{4.9}$$

where $A_z = A(h_z, N_z) \in \mathbb{R}^{N_z \times N_z}$ is defined by (4.6), $I_z \in \mathbb{R}^{N_z \times N_z}$ is the identity, $z = x, y$, and $\otimes$ stands for the Kronecker tensor product of matrices. The non zero entries of blocks B_k in (4.2) are "1" and "-1" on the positions corresponding to the nodes lying on the boundaries $\partial\Omega_k$ (the signs reflect an orientation of the outer normal vector). Recall that we assume full-row rank B without redundant rows. In order to simplify the next presentation we assume that, in addition, the rows of B are orthogonal (due to an orthogonalization procedure with negligible cost). Finally, note that the blocks $R_k \in \mathbb{R}^{N \times 1}$ in (4.3) are given by (4.7) again.

Theorem 4.2 *Let $N_x = N_y \geq 4$ and $H_x = H_y$. Let $P_G F$ be the operator of* (3.13) *given by* (4.1)-(4.3) *arising from the TFETI method applied to* (4.8). *It holds:*

$$\frac{\sqrt{2}h}{16} \leq \sigma_{\min}(P_G F|\operatorname{Ker} G),$$

$$\sigma_{\max}(P_G F|\operatorname{Ker} G) \leq \frac{6\sqrt{2}}{11\pi^2} \cdot \frac{(h+H)^2}{h},$$

$$\kappa(P_G F|\operatorname{Ker} G) \leq \frac{96}{11\pi^2}\left(1 + \frac{H}{h}\right)^2.$$

Proof: The proof is analogous to Theorem 4.1. Now $c_{B,1} = c_{B,2} = 1$, as B is orthogonal. It is well-known that eigenvalues of any Kronecker product matrix are given by products of eigenvalues of particular matrices [10]. Therefore

(4.9) implies that all eigenvalues $\sigma(A_k)$ of A_k read as $\sigma(A_k) = \sigma(A_x) + \sigma(A_y)$, where $\sigma(A_x)$ and $\sigma(A_y)$ are eigenvalues of $A_x = A(h_x, N_x)$ and $A_y = A(h_y, N_y)$, respectively. As all diagonal blocks in A are the same matrices A_k, we obtain

$$\sigma_{\min}(A|\operatorname{Im} A) = \min\{\sigma_{1,x}, \sigma_{1,y}\}, \qquad \sigma_{\max}(A) = \sigma_{N_x-1,x} + \sigma_{N_y-1,y},$$

where $\sigma_{j,z} = \sigma_j(h_z, N_z)$, $j = 0, 1, \ldots, N_z$, for $z = x, y$ are given by Lemma 4.1. When $h_x = h_y = 2^{-\frac{1}{2}}h$ and $H_x = H_y = 2^{-\frac{1}{2}}H$, we derive as in the proof of Theorem 4.1 that

$$\sigma_{\max}(A) \le \frac{4}{h_x} + \frac{4}{h_y} = \frac{8\sqrt{2}}{h} = c_{A,2},$$

$$\sigma_{\min}(A|\operatorname{Im} A) = \sigma_{1,x} \ge \frac{11\pi^2 h_x}{12(h_x + H_x)^2} = \frac{11\sqrt{2}\pi^2 h}{12(h + H)^2} = c_{A,1}.$$

The rest consists of using Theorem 3.1. □

References

[1] Ben-Israel, A., Greville, T.: *Generalized inverses: theory and applications.* Springer, New York, 2003 (2nd ed.).

[2] Benzi, M., Golub, G. H., Liesen, J.: *Numerical solution of saddle point systems.* Acta Numerica **14** (2005), 1–137.

[3] Campbell, S. L., Meyer, C. D.: *Generalized inverses of linear transformations.* In: SIAM Series: Classics in Applied Mathematics 56, SIAM, Philadelphia, 2009.

[4] Dostál, Z., Horák, D., Kučera, R.: *Total FETI—an easier implementable variant of the FETI method for numerical solution of elliptic PDE.* Communications in Numerical Methods in Engineering **22**, 12 (2006), 1155–1162.

[5] Dostál, Z., Kozubek, T., Markopoulos, A., Menšík, M.: *Cholesky factorization of a positive semidefinite matrix with known rigid modes.* SIAM Journal for Matrix Analysis and Applications (2009) (submitted).

[6] Farhat, C., Gerardin, M.: *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 **41**, 4 (1998), 675-–696.

[7] Farhat, C., Lesoinne, M., LeTallec, P., Pierson, K., Rixen. D.: *FETI-DP: a dual-primal unified FETI method. I. A faster alternative to the two-level FETI method.* Internat. J. Numer. Methods Engrg. **50** (2001), 1523–1544.

[8] Farhat, C., Mandel, J., Roux, F. X.: *Optimal convergence properties of the FETI domain decomposition method.* Comput. Methods Appl. Mech. Engrg. **115** (1994), 367–388.

[9] Fragakis, Y., Papadrakakis, M.: *A unified framework for formulating domain decomposition methods in structural mechanics.* Technical Report, National Technical University of Athens, Greecee, March 2002.

[10] Golub, G. H., Van Loan, C. F.: *Matrix computation.* The Johns Hopkins University Press, Baltimore, 1996 (3th ed.).

[11] Gan, G.: *On the relation between Moore's and Penrose's conditions.* IJMMS **30**, 8 (2002), 505—509.

[12] Haslinger, J., Kozubek, T., Kučera, R., Peichl, G.: *Projected Schur complement method for solving non-symmetric saddle-point systems arising from fictitious domain approach.* Numerical Linear Algebra with Applications **14**, 9 (2007), 713–739.

[13] Klawonn, A., Widlund, O. B., Dryja, M.: *Dual-primal FETI methods for three-dimensional elliptic problems with heterogeneous coefficients.* SIAM J. Numer. Anal. **40** (2002), 159–179.

[14] Kouachi, S.: *Eigenvalues and eigenvectors of several tridiagonal matrices.* (1999).

[15] Moore, E. H.: *On the reciprocal of the general algebraic matrix.* Bull. Amer. Math Soc. **21** (1920), 394—395.

[16] Penrose, R.: *A generalized inverse for matrices.* Math. Proc. Cambridge Philos. Soc. **51** (1955), 406—413.

[17] Strang, G.: *The discrete cosine transform.* SIAM Review **41** (1999), 135—147.

[18] Shinozaki, N., Sibuya, M., Tanabe, K.: *Numerical algorithms for the Moore-Penrose inverse of a matrix: Direct methods.* Ann. Inst. Statist. Math. **24**, 1 (1972), 193–203.


Univ. Palacki. Olomuc., Fac. rer. nat.,
Dept of Math. Anal. and Appl. of Math.
ODAM (2009) 23–40

Metody vnitřních bodů pro řešení kontaktních úloh

Jitka MACHALOVÁ [1], Pavel ŽENČÁK [2], Radek KUČERA [3]

Katedra matematické analýzy a aplikací matematiky
Přírodovědecká fakulta, Univerzita Palackého
tř. 17. listopadu 12, 771 46 Olomouc, Česká republika
[1] *e-mail: machalova@inf.upol.cz*
[2] *e-mail: zencak@inf.upol.cz*

[3] *Katedra matematiky a deskriptivní geometrie*
VŠB–TU Ostrava, 17. listopadu 15
CZ-708 33 Ostrava-Poruba, Česká republika
e-mail: radek.kucera@vsb.cz

Abstrakt

V tomto článku se zabýváme aplikací metod vnitřních bodů na speciální úlohu nelineárního programování vznikající při řešení kontaktní úlohy s daným třením. Jako referenční metoda, se kterou metody vnitřních bodů porovnáváme, je použita metoda aktivních množin, konkrétně QPC algoritmus popsaný v [2] a [3]. Z metod vnitřních bodů se zaměříme zejména na metodu sledování cesty, Mehrotrovu metodu prediktor–korektor a na metodu vycházející z obecnějších metod pro nelineární úlohy. Metody porovnáváme z hlediska výpočetního času a počtu iterací v Matlabu.

Key words: Primárně-duální metody vnitřních bodů, metoda sledování cesty, Mehrotrova metoda prediktor–korektor, lineární elasticita, kontaktní úloha s daným třením.

2000 Mathematics Subject Classification: 90C25, 90C51, 49M29

1 Úvod

V článku se zabýváme použitím metod vnitřních bodů (ozn. IPM z anglického *interiror point methods*) pro řešení optimalizační úlohy, která vznikne vytvořením duální úlohy k diskretizované úloze kontaktní úlohy s daným třením. Nejdříve si zformulujeme úlohu kvadratického programování s jednoduchými lineárními

omezeními a se speciálními kvadratickými omezeními a popíšeme si pro ni tři různé metody vnitřních bodů vycházející z primárně-duální formulace. Konkrétně se jedná o upravenou metodu sledování cesty, Mehrotrovu metodu typu prediktor-korektor a zjednodušenou variantu metody určené pro nelineární úlohy, která vychází z článku [1].

V další části velmi stručně popíšeme kontaktní úlohu s daným třením, její diskretizaci metodou konečných prvků a duální úlohu k diskretizované úloze, která vede právě na výše popsanou úlohu kvadratického programování s kvadratickými omezeními. Podrobnější popis je možné najít v článcích [2] a [3]. V poslední části se zabýváme aplikací metod vnitřních bodů pro řešení této duální úlohy, při které je třeba vyřešit problémy vyplývající z toho, že nemáme přímo k dispozici Hessovu matici minimalizované funkce, pouze víme, že je možné ji vypočítat přenásobením inverzní matice k matici tuhosti z primární formulace kontaktní úlohy vhodnými maticemi. V práci ukazujeme, že tento postup je sice možné provést, ale pro větší úlohy není výpočetně efektivní. Proto používáme k řešení lineárních soustav, které je třeba řešit v jednotlivých iteracích metod vnitřních bodů, metodu konjugovaných gradientů s předpodmíněním. Používáme několik různých předpodmínění, které jsou založeny na konstrukci předpodmiňovací matice pomocí Schurova komplementu. Součástí jsou numerické testy všech uvedených způsobů výpočtu, které ukazují, že metody vnitřních bodů jsou zcela porovnatelné s algoritmy založenými na metodě aktivních množin.

2 IPM a kvadratické programování s kvadratickými omezeními

V této části se budeme zabývat metodami vnitřních bodů pro řešení úlohy kvadratického programování s kvadratickými omezeními ve speciálním tvaru.

2.1 Úloha a formulace nutných a postačujících podmínek minima

Nechť je dáno $m \in \mathbb{N}$. Hledáme řešení úlohy:

$$\text{minimalizovat } \frac{1}{2}\mathbf{x}^{\mathsf{T}}\mathbf{A}\mathbf{x} - \mathbf{x}^{\mathsf{T}}\mathbf{b} \tag{1a}$$

$$\text{za podmínek } x_{2,i}^2 + x_{3,i}^2 \leq g_i^2 \qquad \text{pro } i \in \{1, 2, \ldots, m\} \tag{1b}$$

$$x_{1,i} \geq l_i \qquad \text{pro } i \in \{1, 2, \ldots, m\} \tag{1c}$$

kde

$$\mathbf{x_j} = (x_{j,1}, x_{j,2}, \ldots, x_{j,m})^{\mathsf{T}} \in \mathbb{R}^m \text{ pro } j \in \{1, 2, 3\}$$
$$\mathbf{x} = \left(\mathbf{x}_1^{\mathsf{T}}, \mathbf{x}_2^{\mathsf{T}}, \mathbf{x}_3^{\mathsf{T}}\right)^{\mathsf{T}} \in \mathbb{R}^{3m}.$$

Nyní si zformulujeme nutné a postačující podmínky minima této úlohy. Označme

$$\begin{aligned}\mathbf{l} &= (l_1, l_2, \ldots, l_m)^\mathsf{T} \\ \mathbf{e} &= (1, 1, \ldots, 1)^\mathsf{T} \in \mathbb{R}^m \\ \mathbf{G} &= diag\,(\mathbf{g}) \\ \mathbf{X}_2 &= diag\,(\mathbf{x}_2) \\ \mathbf{X}_3 &= diag\,(\mathbf{x}_3)\end{aligned}$$

Lagrangeova funkce je pak definována předpisem

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) := \frac{1}{2}\mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x} - \mathbf{x}^\mathsf{T}\mathbf{b} + \boldsymbol{\mu}^\mathsf{T}\left(\mathbf{X}_2^2 + \mathbf{X}_3^2 - \mathbf{G}^2\right)\mathbf{e} + \boldsymbol{\lambda}^\mathsf{T}(\mathbf{l} - \mathbf{x}_1) \tag{2}$$

kde $\boldsymbol{\lambda} \in \mathbb{R}^m$ a $\boldsymbol{\mu} \in \mathbb{R}^m$ jsou nezáporné vektory Lagrangeových multiplikátorů. Jelikož se jedná o úlohu konvexního programování, tak nutnými a postačujícími podmínkami existence minima jsou *Karush–Kuhn–Tuckerovy (KKT) podmínky*, které lze zapsat ve tvaru

$$\frac{\partial \mathcal{L}}{\partial \mathbf{x}}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \mathbf{0}, \tag{3}$$

$$\frac{\partial \mathcal{L}}{\partial \lambda}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) + \mathbf{s} = \mathbf{0}, \tag{4}$$

$$\boldsymbol{\lambda}^\mathsf{T}\mathbf{s} = \mathbf{0}, \tag{5}$$

$$\frac{\partial \mathcal{L}}{\partial \mu}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) + \mathbf{d} = \mathbf{0}, \tag{6}$$

$$\boldsymbol{\mu}^\mathsf{T}\mathbf{d} = \mathbf{0}, \tag{7}$$

$$\boldsymbol{\lambda} \geq \mathbf{0},\ \mathbf{s} \geq \mathbf{0},\ \boldsymbol{\mu} \geq \mathbf{0},\ \mathbf{d} \geq \mathbf{0}, \tag{8}$$

kde $\mathbf{s} \in \mathbb{R}^m$ a $\mathbf{d} \in \mathbb{R}^m$ jsou vektory doplňkových proměnných.

Tuto úlohu můžeme zapsat i v jiné podobě. Označme

$$\begin{aligned}\mathbf{S} &= diag\,(\mathbf{d}) \\ \boldsymbol{\Lambda} &= diag\,(\boldsymbol{\lambda}) \\ \mathbf{D} &= diag\,(\mathbf{d}) \\ \mathbf{M} &= diag\,(\boldsymbol{\mu})\end{aligned}$$

a nechť matice $\mathbf{A}$ a vektor $\mathbf{b}$ jsou rozděleny do bloků $\mathbf{A_{i,j}} \in \mathbb{R}^{m\times m}$ a $\mathbf{b_i} \in \mathbb{R}^m$ pro $i, j \in \{1, 2, 3\}$ v souladu s rozdělením vektoru $\mathbf{x}$, tj.

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} & \mathbf{A}_{13} \\ \mathbf{A}_{21} & \mathbf{A}_{22} & \mathbf{A}_{23} \\ \mathbf{A}_{31} & \mathbf{A}_{32} & \mathbf{A}_{33} \end{pmatrix}, \qquad \mathbf{b} = \begin{pmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \\ \mathbf{b}_3 \end{pmatrix}.$$

Označíme-li navíc

$$\mathbf{y} = (\boldsymbol{\lambda}^\mathsf{T}, \mathbf{s}^\mathsf{T}, \boldsymbol{\mu}^\mathsf{T}, \mathbf{d}^\mathsf{T})^\mathsf{T}$$

a

$$F(\mathbf{x}, \mathbf{y}) = \begin{pmatrix} \mathbf{A}_{11}\mathbf{x}_1 + \mathbf{A}_{12}\mathbf{x}_2 + \mathbf{A}_{13}\mathbf{x}_3 - \boldsymbol{\lambda} - \mathbf{b}_1 \\ \mathbf{A}_{21}\mathbf{x}_1 + (\mathbf{A}_{22} + 2\mathbf{M})\mathbf{x}_2 + \mathbf{A}_{23}\mathbf{x}_3 - \mathbf{b}_2 \\ \mathbf{A}_{31}\mathbf{x}_1 + \mathbf{A}_{32}\mathbf{x}_2 + (\mathbf{A}_{33} + 2\mathbf{M})\mathbf{x}_3 - \mathbf{b}_3 \\ -\mathbf{x}_1 + \mathbf{s} + \mathbf{l} \\ \boldsymbol{\Lambda}\mathbf{S}\mathbf{e} \\ (\mathbf{X}_2^2 + \mathbf{X}_3^2 - \mathbf{G}^2)\,\mathbf{e} + \mathbf{d} \\ \mathbf{M}\mathbf{D}\mathbf{e}_2 \end{pmatrix},$$

pak soustava KKT podmínek odpovídá soustavě nelineárních rovnic

$$F(\mathbf{x}, \mathbf{y}) = \mathbf{0} \quad \text{s podmínkami } \mathbf{y} \geq \mathbf{0} \tag{9}$$

2.2 Newtonova metoda v IPM

Základem naší implementace metod vnitřních bodů jsou postupy pro úlohu lineárního programování uvedené v [4] a [5]. Podstatou této implementace metod vnitřních bodů je řešení soustavy (9) upravenou Newtonovou metodou zachovávající navíc podmínky nezápornosti.

Algoritmus 1 Řešení KKT Newtonovou metodou
Je dáno $\mathbf{x}^{(0)} \in \mathbb{R}^{3m}$, $\mathbf{y}^{(0)} \in \mathbb{R}^{4m}_+$, $\delta \in (0,1)$ *a* $\epsilon \geq 0$. *Položíme* $k := 0$.

1. *Vyřešíme*

$$J(\mathbf{x}^{(k)}, \mathbf{y}^{(k)}) \begin{pmatrix} \Delta\mathbf{x}^{(k+1)} \\ \Delta\mathbf{y}^{(k+1)} \end{pmatrix} = -F(\mathbf{x}^{(k)}, \mathbf{y}^{(k)}) \tag{10}$$

2. *Vypočítáme*

$$\alpha^{(k)} = \min_{\Delta y_i^{(k+1)} < 0} \{1, -\delta y_i^{(k)} / \Delta y_i^{(k+1)}\},$$

   kde $\delta \in (0,1)$ *zajišťuje splnění* $\mathbf{y}^{(k+1)} > \mathbf{0}$ *(např.* $\delta = 0.999$*).*

3. *Položíme* $\mathbf{x}^{(k+1)} := \mathbf{x}^{(k)} + \alpha^{(k)}\,\Delta\mathbf{x}^{(k+1)}$ *a* $\mathbf{y}^{(k+1)} := \mathbf{y}^{(k)} + \alpha^{(k)}\,\Delta\mathbf{y}^{(k+1)}$.

4. *Je-li* $\|(\Delta\mathbf{x}^{(k+1)}, \Delta\mathbf{y}^{(k+1)})\| \leq \epsilon$ *vrátíme* $(\mathbf{x}^*, \mathbf{y}^*) = (\mathbf{x}^{(k+1)}, \mathbf{y}^{(k+1)})$ *a skončíme, jinak položíme* $k := k + 1$ *a vrátíme se ke kroku 1.*

Řešení úlohy tímto postupem je ovšem často zdlouhavé, neboť může vyžadovat velký počet iterací, proto jsme pro praktickou implementaci zvolili dvě metody, které tento základní postup vylepšují. Jsou to metoda sledování cesty a Mehrotrova metoda.

2.3 Metoda sledování cesty

Odvození metody vychází z tzv. *centrální cesty* (ozn. $\mathcal{C}$), což je množina bodů $(\mathbf{x}^\tau, \mathbf{y}^\tau)$ řešících pro každou hodnotu parametru τ, $\tau > 0$ následující úlohu

$$F(\mathbf{x}^\tau, \mathbf{y}^\tau) = (\mathbf{0}^\mathsf{T}, \mathbf{0}^\mathsf{T}, \mathbf{0}^\mathsf{T}, \mathbf{0}^\mathsf{T}, \tau\mathbf{e}^\mathsf{T}, \mathbf{0}^\mathsf{T}, \tau\mathbf{e}^\mathsf{T})^\mathsf{T}, \qquad \mathbf{y} > \mathbf{0}$$

která vznikla drobnou změnou KKT podmínek. Poznamenejme, že parametr τ odpovídá tzv. bariérovému parametru z bariérových metod s logaritmickou bariérou. V naší implementaci jsme tento koncept mírně upravili zavedením dvou nezávislých parametrů τ_l resp. τ_q pro lineární resp. kvadratické podmínky, tj. řešíme úlohu

$$F(\mathbf{x}^\tau, \mathbf{y}^\tau) = \mathbf{c}_s(\tau_l, \tau_q), \qquad \mathbf{y} > \mathbf{0} \tag{11}$$

kde $\mathbf{c}_s(\tau_l, \tau_q) = (\mathbf{0}^\mathsf{T}, \mathbf{0}^\mathsf{T}, \mathbf{0}^\mathsf{T}, \mathbf{0}^\mathsf{T}, \tau_l\mathbf{e}^\mathsf{T}, \mathbf{0}^\mathsf{T}, \tau_q\mathbf{e}^\mathsf{T})^\mathsf{T}$

Výhodou tohoto postupu je , že hledaný směr je kompromis mezi Newtonovým směrem a tzv. centrujícím směrem. Pokud totiž hledáme pouze v Newtonově směru, tak rychle dojdeme k hranici oblasti (daleko od řešení) a krok v dalších iteracích je velmi malý. Oproti klasické bariérové metodě s logaritmickou bariérou, ve které se hledají přímo body na centrální cestě, je výhodou , že hledáme pouze body určitým způsobem „blízko“ centrální cesty, což je výpočetně méně náročné. Hodnoty parametrů τ_l resp. τ_q počítáme jako součin tzv. míry duality $\beta_l^{(k)}$ resp. $\beta_q^{(k)}$ a centrujícího parametru σ_l resp. σ_q. Míry duality, které vyjadřují vzdálenost aktuální iterace od centrální cesty, počítáme podle vztahů

$$\beta_l^{(k)} = \frac{(\boldsymbol{\lambda}^{(k)})^\mathsf{T}\mathbf{s}^{(k)}}{m}, \qquad \beta_q^{(k)} = \frac{(\boldsymbol{\mu}^{(k)})^\mathsf{T}\mathbf{d}^{(k)}}{m}.$$

Centrující parametry mohou nabývat hodnot z intervalu $[0, 1]$, přičemž krajní hodnoty odpovídají volbě Newtonova směru resp. centrujícího směru.V naší implementaci pro jejich stanovení užíváme heuristické pravidlo ze softwaru LOQO

$$\sigma_l = 0.1\left(\min\left\{0.05\frac{1-\xi_l}{\xi_l}, 2\right\}\right)^3 \quad \text{a} \quad \sigma_q = 0.1\left(\min\left\{0.05\frac{1-\xi_q}{\xi_q}, 2\right\}\right)^3,$$

kde

$$\xi_l = \frac{\min_{1\le i\le m}\lambda_i^{(k)}s_i^{(k)}}{\beta_l^{(k)}} \quad \text{a} \quad \xi_q = \frac{\min_{1\le i\le m}\mu_i^{(k)}d_i^{(k)}}{\beta_q^{(k)}}.$$

Velikost centrujícího parametru tak závisí na odchylce individuálních podmínek komplementarity od jejich průměru (tj. od míry duality).

Algoritmus 2 Metoda sledování cesty
Je dáno $\mathbf{x}^{(0)} \in \mathbb{R}^{3m}$, $\mathbf{y}^{(0)} \in \mathbb{R}^{4m}_+$, $\delta \in (0, 1)$ *a* $\epsilon \ge 0$. *Položíme* $k := 0$.

1. *Vypočítáme* $\beta_l^{(k)} = {\lambda^{(k)}}^\top s^{(k)}/m$, $\beta_q^{(k)} = {\mu^{(k)}}^\top d^{(k)}/m$ *a* $\tau_l^{(k)} = \sigma_l \beta_l^{(k)}$, $\tau_q^{(k)} = \sigma_q \beta_q^{(k)}$.

2. *Vyřešíme*

$$J(\mathbf{x}^{(k)}, \mathbf{y}^{(k)}) \begin{pmatrix} \Delta\mathbf{x}^{(k+1)} \\ \Delta\mathbf{y}^{(k+1)} \end{pmatrix} = -F(\mathbf{x}^{(k)}, \mathbf{y}^{(k)}) + \mathbf{c}_s(\tau_l^{(k)}, \tau_q^{(k)}) \tag{12}$$

3. *Vypočítáme*

$$\alpha^{(k)} = \min_{\Delta y_i^{(k+1)} < 0} \{1, -\delta y_i^{(k)} / \Delta y_i^{(k+1)}\},$$

kde $\delta \in (0,1)$ *zajišťuje splnění* $\mathbf{y}^{(k+1)} > \mathbf{0}$ *(např.* $\delta = 0.999$*).*

4. *Položíme* $\mathbf{x}^{(k+1)} := \mathbf{x}^{(k)} + \alpha^{(k)}\, \Delta\mathbf{x}^{(k+1)}$ *a* $\mathbf{y}^{(k+1)} := \mathbf{y}^{(k)} + \alpha^{(k)}\, \Delta\mathbf{y}^{(k+1)}$.

5. *Je-li* $\|(\Delta\mathbf{x}^{(k+1)}, \Delta\mathbf{y}^{(k+1)})\| \leq \epsilon$ *vrátíme* $(\mathbf{x}^*, \mathbf{y}^*) = (\mathbf{x}^{(k+1)}, \mathbf{y}^{(k+1)})$ *a skončíme, jinak položíme* $k := k+1$ *a pokračujeme krokem 1.*

2.4 Mehrotrova metoda typu prediktor–korektor

Jednou z nejčastěji prakticky používaných metod je Mehrotrova metoda typu prediktor korektor. V prediktoru počítáme pouze Newtonův směr, který je lineární aproximací aktuální trajektorie k řešení KKT podmínek. Pak vypočítáme chybu této lineární aproximace a použijeme ji ke korekci a tím využijeme i kvadratickou informaci o této trajektorii. V každé iteraci tak musíme vypočítat řešení dvou soustav lineárních rovnic se stejnou maticí.

Algoritmus 3 Mehrotrova metoda – prediktor
Je dáno $\mathbf{x}^{(0)} \in \mathbb{R}^{3m}$, $\mathbf{y}^{(0)} \in \mathbb{R}^{4m}_+$, $\delta \in (0,1)$ *a* $\epsilon \geq 0$. *Položíme* $k := 0$.

(P1) Vyřešíme

$$J(\mathbf{x}^{(k)}, \mathbf{y}^{(k)}) \begin{pmatrix} \Delta\mathbf{x}^{P} \\ \Delta\mathbf{y}^{P} \end{pmatrix} = -F(\mathbf{x}^{(k)}, \mathbf{y}^{(k)}) \tag{13}$$

(P2) Vypočítáme

$$\alpha^P = \min_{\Delta y_i^P < 0} \{1, -\delta y_i^{(k)} / \Delta y_i^P\}$$

(P3) Předpovíme míru duality pro krok z prediktoru:

$$\beta_l^P = \frac{\left(\boldsymbol{\lambda}^{(k)} + \alpha^P \Delta\boldsymbol{\lambda}^P\right)^{\mathsf{T}} \left(\mathbf{s}^{(k)} + \alpha^P \Delta\mathbf{s}^P\right)}{m}$$

$$\beta_q^P = \frac{\left(\boldsymbol{\mu}^{(k)} + \alpha^P \Delta\boldsymbol{\mu}^P\right)^{\mathsf{T}} \left(\mathbf{d}^{(k)} + \alpha^P \Delta\mathbf{d}^P\right)}{m}$$

(P4) Předpověď použijeme pro výpočet centrujícího parametru:

$$\sigma_l = \left(\beta_l^P/\beta_l^{(k)}\right)^3, \qquad \sigma_q = \left(\beta_q^P/\beta_q^{(k)}\right)^3$$

Mehrotrova metoda – korektor

(K1) Vyřešíme

$$J(\mathbf{x}^{(k)},\mathbf{y}^{(k)})\begin{pmatrix}\Delta\mathbf{x}^{(k+1)}\\ \Delta\mathbf{y}^{(k+1)}\end{pmatrix} = -F(\mathbf{x}^{(k)},\mathbf{y}^{(k)}) + \mathbf{c}_s(\tau_l^{(k)},\tau_q^{(k)}) - \\ -(\mathbf{0}^\mathsf{T},\mathbf{0}^\mathsf{T},\mathbf{0}^\mathsf{T},\mathbf{0}^\mathsf{T},\mathbf{e}^\mathsf{T}\Delta\boldsymbol{\Lambda}^P\Delta\boldsymbol{S}^P,\mathbf{0}^\mathsf{T},\mathbf{e}^\mathsf{T}\Delta\boldsymbol{M}^P\Delta\boldsymbol{D}^P)^\mathsf{T}$$

(K2) Vypočítáme

$$\alpha^{(k)} = \min_{\Delta y_i^{(k+1)}<0}\{1, -\delta y_i^{(k)}/\Delta y_i^{(k+1)}\}$$

(K3) Položíme $\mathbf{x}^{(k+1)} := \mathbf{x}^{(k)} + \alpha^{(k)}\,\Delta\mathbf{x}^{(k+1)}$ *a* $\mathbf{y}^{(k+1)} := \mathbf{y}^{(k)} + \alpha^{(k)}\,\Delta\mathbf{y}^{(k+1)}$.

(K4) Je-li $\|(\Delta\mathbf{x}^{(k+1)},\Delta\mathbf{y}^{(k+1)})\| \le \epsilon$ *vrátíme* $(\mathbf{x}^*,\mathbf{y}^*) = (\mathbf{x}^{(k+1)},\mathbf{y}^{(k+1)})$ *a skončíme, jinak položíme* $k := k+1$ *a vrátíme se ke kroku (P1).*

2.5 Metoda založená na formulaci pro nelineární úlohy

Předchozí metody vycházejí z formulace pro úlohy lineárního programování, naše úloha je ovšem obecnější. Obě metody fungují spolehlivě, ale důkazy konvergence a zkoumání výpočetní náročnosti jsou založeny na linearitě a tak je pro náš případ není možné zopakovat. Proto jsme se rozhodli vyzkoušet i variantu IPM založenou na metodě určené pro nelineární optimalizační úlohy, kde by situace měla být snazší. Vycházíme přitom z článku [1]. Vlastnosti naší úlohy nám umožnily postup zjednodušit.

Algoritmus 4 Zjednodušená verze algoritmu
Je dáno $\mathbf{x}^{(0)} \in \mathbb{R}^{3m}$, $\mathbf{y}^{(0)} \in \mathbb{R}^{4m}_+$, $\delta,\kappa \in (0,1)$ *a* $\epsilon \ge 0$. *Položíme* $k := 0$.

1. *Vypočítáme* $\nu = \max_{1\le i\le 7m}|F(\mathbf{x}^{(k)},\mathbf{y}^{(k)})|$ *a* $\tau^{(k)} = \min\{\kappa\nu,\nu^2\}$.
2. *Vyřešíme*

$$J(\mathbf{x}^{(k)},\mathbf{y}^{(k)})\begin{pmatrix}\Delta\mathbf{x}^{(k+1)}\\ \Delta\mathbf{y}^{(k+1)}\end{pmatrix} = -F(\mathbf{x}^{(k)},\mathbf{y}^{(k)}) + \mathbf{c}_s(\tau^{(k)},\tau^{(k)}) \qquad (14)$$

3. *Vypočítáme*

$$\alpha_p^{(k)} = \min\left\{1, \min_{\Delta s_i^{(k+1)}<0}\left\{\frac{-\delta s_i^{(k)}}{\Delta s_i^{(k+1)}}\right\}, \min_{\Delta d_j^{(k+1)}<0}\left\{\frac{-\delta d_j^{(k)}}{\Delta d_j^{(k+1)}}\right\}\right\},$$

$$\alpha_d^{(k)} = \min\left\{1, \min_{\Delta\lambda_i^{(k+1)}<0}\left\{\frac{-\delta\lambda_i^{(k)}}{\Delta\lambda_j^{(k+1)}}\right\}, \min_{\Delta\mu_j^{(k+1)}<0}\left\{\frac{-\delta\mu_j^{(k)}}{\Delta\mu_i^{(k+1)}}\right\}\right\}$$

4. Položíme

$$\begin{aligned}\left(\mathbf{x}^{(k+1)}, \mathbf{s}^{(k+1)}, \mathbf{d}^{(k+1)}\right) &= \left(\mathbf{x}^{(k)}, \mathbf{s}^{(k)}, \mathbf{d}^{(k)}\right) \\ &\quad +\alpha_p^{(k)} \left(\Delta\mathbf{x}^{(k+1)}, \Delta\mathbf{s}^{(k+1)}, \Delta\mathbf{d}^{(k+1)}\right), \\ \left(\boldsymbol{\lambda}^{(k+1)}, \boldsymbol{\mu}^{(k+1)}\right) &= \left(\boldsymbol{\lambda}^{(k)}, \boldsymbol{\mu}^{(k)}\right) + \alpha_d^{(k)} \left(\Delta\boldsymbol{\lambda}^{(k+1)}, \Delta\boldsymbol{\mu}^{(k+1)}\right)\end{aligned}$$

5. Je-li $\|(\Delta\mathbf{x}^{(k+1)}, \Delta\mathbf{y}^{(k+1)})\| \leq \epsilon$ *vrátíme* $(\mathbf{x}^*, \mathbf{y}^*) = (\mathbf{x}^{(k+1)}, \mathbf{y}^{(k+1)})$ *a skončíme, jinak položíme* $k := k + 1$ *a vrátíme se ke kroku 1.*

3 Formulace kontaktní úlohy lineární elasticity s daným třením

V této části si ukážeme formulaci úlohy lineární elasticity s daným třením, její diskretizaci a formulaci duální úlohy, která vede na minimalizační úlohu, kterou jsme se zabývali v předchozí části. Uvedeme zde pouze některé výsledky, celý postup vychází z článků [2] a [3], ve kterých je také uveden algoritmus (nazvaný QPC) pro řešení tohoto minimalizačního problému. Tento algoritmus je založen na použití metody aktivní množiny, projekce gradientu a metody konjugovaných gradientů, v dalším nám slouží jako jakýsi referenční algoritmus pro porovnávání účinnosti metod vnitřních bodů.

3.1 Kontaktní úloha lineární elasticity s daným třením

Nejprve si stručně popíšeme úlohu, kterou řešíme. Nechť $\Omega \in \mathbb{R}^3$ je oblast s Lipschitzovskou hranicí takovou, že $\delta\Omega = \bar{\Gamma}_u \cup \bar{\Gamma}_p \cup \bar{\Gamma}_c$, kde na Γ_u je předepsáno nulové posunutí, na Γ_p je dáno plošné zatížení $p \in (L^2(\Gamma_p))^3$ a na části Γ_c je jednostranná podpora s počáteční vzdáleností $d \in L^\infty(\Gamma_c)$. Dále předpokládáme, že na těleso působí objemové síly $f \in (L^2(\bar{\Omega}))^3$.

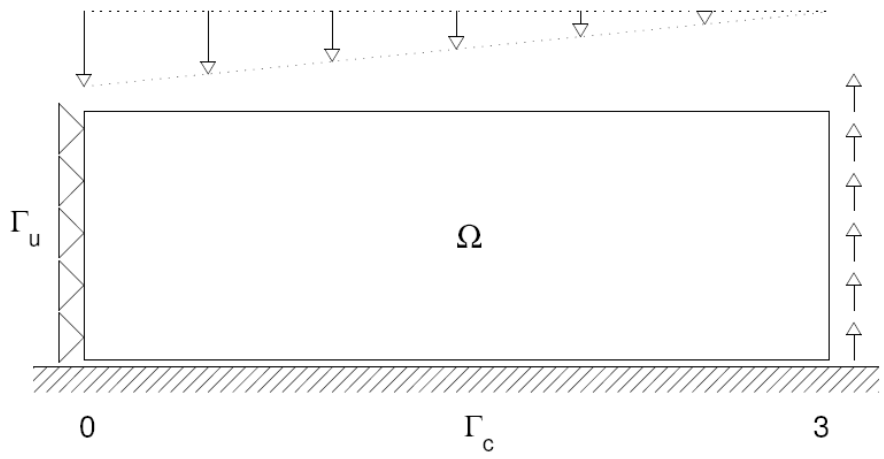
Je dán prostor virtuálních posunutí

$$V = \{v \in (H^1(\Omega))^3 \mid v = (0,0,0) \text{ na } \Gamma_u\}$$

a konvexní množina kinematicky přípustných posunutí

$$K = \{v \in V \mid v_\nu := v \cdot \nu \leq d \text{ na } \Gamma_c\},$$

kde $\nu \in (L^\infty(\Gamma_c))^3$



3.2 Diskretizace úlohy

Diskretizujeme-li úlohu metodou konečných prvků dostaneme následující optimalizační úlohu.
Najít $\mathbf{u} \in \mathcal{K}$ takové, že

$$\mathcal{J}(\mathbf{u}) = \min_{\mathbf{v}\in\mathcal{K}} \mathcal{J}(\mathbf{v}),$$

kde

$$\mathcal{J}(\mathbf{v}) = \frac{1}{2}\mathbf{v}^{\mathsf{T}}\mathbf{K}\mathbf{v} - \mathbf{v}^{\mathsf{T}}\mathbf{f} + \mathbf{g}^{\mathsf{T}}\|\mathbf{T}\mathbf{v}\|_{vect}$$

a

$$\mathcal{K} = \{\mathbf{v} \in \mathbb{R}^n \mid \mathbf{N}\,\mathbf{v} \leq \mathbf{d}\},$$

přičemž symbol $\|\mathbf{T}\mathbf{v}\|_{vect}$ je definován jako

$$\|\mathbf{T}\mathbf{v}\|_{vect} := (\|(\mathbf{T}\mathbf{v})_1\|_{\mathbb{R}^2}, \|(\mathbf{T}\mathbf{v})_2\|_{\mathbb{R}^2}, \ldots, \|(\mathbf{T}\mathbf{v})_m\|_{\mathbb{R}^2})^{\mathsf{T}} \in \mathbb{R}^m.$$

3.3 Duální úloha

Předchozí optimalizační úlohu nebudeme řešit přímo, místo ní budeme řešit její duální úlohu, která vede na minimalizaci funkcionálu

$$D(\mathbf{x}) = \frac{1}{2}\mathbf{x}^{\mathsf{T}}\mathbf{A}\mathbf{x} - \mathbf{x}^{\mathsf{T}}\mathbf{b}$$

na množině

$$\left\{\mathbf{x} = (\mathbf{x}_1^{\mathsf{T}}, \mathbf{x}_2^{\mathsf{T}}, \mathbf{x}_3^{\mathsf{T}})^{\mathsf{T}} \in \mathbb{R}^m \mid \mathbf{x}_1 \geq \mathbf{l},\ \|(x_{2_i}, x_{3_i})\|^2 \leq g_i^2,\ i = 1, \ldots, m\right\},$$

kde

$$\mathbf{A} = \mathbf{B}\mathbf{K}^{-1}\mathbf{B}^{\mathsf{T}}, \quad \mathbf{B} = \begin{pmatrix} \mathbf{N} \\ \mathbf{T}_1 \\ \mathbf{T}_2 \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} \mathbf{N}\mathbf{K}^{-1} - \mathbf{d} \\ \mathbf{T}_1\mathbf{K}^{-1} \\ \mathbf{T}_2\mathbf{K}^{-1} \end{pmatrix}.$$

4 IPM v úloze lineární elasticity s daným třením

Je zřejmé, že duální úloha odvozená v předchozí části, je úloha kvadratického programování s kvadratickými omezeními ve tvaru (1). K jejímu řešení tedy můžeme použít metody vnitřních bodů, které jsme popsali ve druhé kapitole článku. Pro efektivitu programu bude ovšem důležité, jak naložíme s tím, že matice $\mathbf{A}$ je dána vzorcem $\mathbf{A} = \mathbf{B}\mathbf{K}^{-1}\mathbf{B}^{\mathsf{T}}$. Ukážeme si zde několik možných postupů.

4.1 Použití eliminačních metod IPM s přímým výpočtem matice A

Nejprve zkusíme aplikovat iterační metody přímo, tj. nejprve matici $\mathbf{A}$ vypočítáme z předpisu a poté aplikujeme metody vnitřních bodů s použitím přímých řešičů. K tomu, aby byl výpočet matice $\mathbf{A}$ efektivní, nebudeme používat přímo vzorec $\mathbf{A} = \mathbf{B}\mathbf{K}^{-1}\mathbf{B}^{\mathsf{T}}$použijeme Choleského rozklad matice $\mathbf{K} = \mathbf{L}\mathbf{L}^{\mathsf{T}}$. Užitím $\mathbf{A} = \mathbf{B}\left(\mathbf{L}\mathbf{L}^{\mathsf{T}}\right)^{-1}\mathbf{B}^{\mathsf{T}}$ pak postupujeme takto:

1. Najdeme řešení m soustav rovnic s dolní trojúhelníkovou matici

$$\mathbf{L}\mathbf{Y} = \mathbf{B}^{\mathsf{T}}$$

2. Najdeme řešení m soustav rovnic s horní trojúhelníkovou matici

$$\mathbf{L}^{\mathsf{T}}\mathbf{Z} = \mathbf{Y}$$

3. Nakonec vypočítáme

$$\mathbf{A} = \mathbf{B}\mathbf{Z}$$

Tento postup jsme otestovali s použitím prvních dvou algoritmů vnitřních bodů a porovnali s metodou QPC.

Z tabulky 1 je zřejmé, že obě metody vnitřních bodů jsou ve všech případech rychlejší než QPC. Při bližším pohledu ovšem zjistíme, že je to způsobeno jen vyšší efektivitou Matlabu při simultánním řešení více soustav lineárních rovnic se stejnou maticí, neboť je jasné, že výpočet matice $\mathbf{A}$ v metodách vnitřních bodů vyžaduje řešit m soustav lineárních rovnic s maticí $\mathbf{K}$, kdežto QPC jich pro větší úlohy vyžaduje mnohem méně (viz. sloupec Av). Stejně tak podíl času potřebného pro výpočet matice $\mathbf{A}$ na celkovém výpočetním času metod vnitřních bodů výrazně roste a je tak zřejmé, že IPM nejsou tak dobře škálovatelné jako QPC.

4.2 IPM bez přímého výpočtu matice A užitím iteračních metod s předpodmíněním plnou maticí

Jak jsme si ukázali, tak pro využití přímých řešičů je překážkou to, že je nutné nejprve vypočítat matici $\mathbf{A}$, což je časově náročné. Proto jsme se zabývali použitím iteračních metod (přesněji metody konjugovaných gradientů) pro výpočet řešení lineárních soustav, které vznikají v jednotlivých iteracích metod vnitřních bodů. Pokud se podíváme na dříve popsané metody vnitřních bodů, tak ve všech řešíme lineární soustavy s Jakobiho maticí $J(\mathbf{x}, \mathbf{y})$. Tuto matici je možné snadno převést eliminací vektorů doplňkových proměnných $\mathbf{s}$ a $\boldsymbol{d}$ na soustavu s tzv. *rozšířenou maticí* ve tvaru

$$\mathbf{J}_R = \mathbf{J}_R(\boldsymbol{x}, \boldsymbol{w}) = \left(\begin{array}{ccc|cc} \mathbf{A}_{11} & \mathbf{A}_{12} & \mathbf{A}_{13} & -\mathbf{I} & \mathbf{0} \\ \mathbf{A}_{21} & \mathbf{A}_{22}+2\mathbf{M} & \mathbf{A}_{23} & \mathbf{0} & 2\mathbf{X}_2 \\ \mathbf{A}_{31} & \mathbf{A}_{32} & \mathbf{A}_{33}+2\mathbf{M} & \mathbf{0} & 2\mathbf{X}_3 \\ \hline -\mathbf{I} & \mathbf{0} & \mathbf{0} & -\boldsymbol{\Lambda}^{-1}\mathbf{S} & \mathbf{0} \\ \mathbf{0} & 2\mathbf{X}_2 & 2\mathbf{X}_3 & \mathbf{0} & -\mathbf{M}^{-1}\mathbf{D} \end{array}\right)$$

kde $\mathbf{w} = (\boldsymbol{\lambda}, \boldsymbol{\mu})$). Tuto matici blokově píšeme následovně

$$\mathbf{J}_R = \begin{pmatrix} \mathbf{A_R} & \mathbf{B_R} \\ \mathbf{B_R^T} & \mathbf{D_R} \end{pmatrix}$$

n	**m**	**QPC**		**Sledování cesty**			**Mehrotrova met.**		
		Čas	**Av**	**Čas**	**Čas A**	%	**Čas**	**Čas A**	%
162	54	0.29	203	0.07	0.03	45	0.12	0.03	26
900	180	2.08	311	0.68	0.34	50	1.07	0.34	31
2646	378	12.91	347	5.85	3.46	59	7.00	3.26	47
5832	648	53.4	384	27.1	18.1	67	27.0	15.8	59
10890	990	126.2	408	79.7	58.5	73	90.0	60.5	67
18252	1404	361.9	493	246.2	192.5	78	274.0	184	67
28350	1890	809.4	478	620.5	493.0	79	677.6	493.5	73

Tabulka 1: Porovnání metod vnitřních bodů užívajících přímý řešič s algoritmem QPC pro různé diskretizace kontaktní úlohy. Číslo n je počet primárních a m je počet duálních proměnných, ve sloupci **Čas** je celkový čas výpočtu, ve sloupci **Čas A** je čas potřebný na výpočet matice $\mathbf{A}$, ve sloupci **%** je uvedeno kolik procent z celkového výpočetního času výpočet matice $\mathbf{A}$ zabere a ve sloupci **Av** je počet iterací algoritmu QPC, tj. kolik soustav lineárních rovnic s maticí $\mathbf{K}$ je třeba řešit.

Dále si označme

$$\bar{\mathbf{J}}_R = \left(\begin{array}{ccc|cc} \mathbf{0} & \mathbf{0} & \mathbf{0} & -\mathbf{I} & \mathbf{0} \\ \mathbf{0} & 2\mathbf{M} & \mathbf{0} & \mathbf{0} & 2\mathbf{X}_2 \\ \mathbf{0} & \mathbf{0} & 2\mathbf{M} & \mathbf{0} & 2\mathbf{X}_3 \\ \hline -\mathbf{I} & \mathbf{0} & \mathbf{0} & -\mathbf{\Lambda}^{-1}\mathbf{S} & \mathbf{0} \\ \mathbf{0} & 2\mathbf{X}_2 & 2\mathbf{X}_3 & \mathbf{0} & -\mathbf{M}^{-1}\mathbf{D} \end{array}\right)$$

matici, která vznikne z matice $\mathbf{J}_R$ odečtením matice $\mathbf{A}$ od bloku $\mathbf{A_R}$. Při řešení soustavy s maticí $\mathbf{J}_R$ použitím metody konjugovaných gradientů je třeba pouze násobit matici $\mathbf{J}_R$ vektorem $(\mathbf{x}, \mathbf{w})$. Toto násobení se realizuje po částech a to následovně:

1. přímo vypočítáme součin matice $\bar{\mathbf{J}}_R$ s vektorem $(\mathbf{x}, \mathbf{w})$
2. součin matice $\mathbf{A}$ s vektorem $\mathbf{w}$ vypočítáme užitím $\mathbf{Aw} = \mathbf{BK}^{-1}\mathbf{B}^{\mathsf{T}}\mathbf{w}$, kde se místo násobení maticí $\mathbf{K}^{-1}$ řeší soustava lineárních rovnic s maticí $\mathbf{K}$ pomocí její Choleského faktorizace vypočítané dříve.
3. oba výsledky „sečteme“

Iterační metody bez předpodmínění fungují pro matice produkované IPM velmi špatně a pomalu. Proto používáme metody s předpodmíněním, přičemž k předpodmínění používáme matici $\widetilde{\mathbf{J}}_\mathbf{R} \approx \mathbf{J}_R^{-1}$. Nyní postupně ukážeme, jak tuto aproximaci najít.

Vyjádření matice $\mathbf{J}_R^{-1}$ užitím Schurova komplementu

Nejprve se podívejme, co platí pro matici $\mathbf{J}_R^{-1}$. Vyjdeme z blokového zápisu matice $\mathbf{J}_R$. Matici $\mathbf{J}_R^{-1}$ pak můžeme užitím Schurova komplementu podmatice $\mathbf{A_R}$ vyjádřit následovně

$$\mathbf{J}_R^{-1} = \begin{pmatrix} \mathbf{A_R}^{-1} - \mathbf{A_R}^{-1}\mathbf{B_R}\mathbf{F} & -\mathbf{A_R}^{-1}\mathbf{B_R}\mathbf{H}^{-1} \\ \mathbf{F} & \mathbf{H}^{-1} \end{pmatrix},$$

kde $\mathbf{H} = \mathbf{D_R} - \mathbf{B_R}^{\mathsf{T}}\mathbf{A_R}^{-1}\mathbf{B_R}$ a $\mathbf{F} = -\mathbf{H}^{-1}\mathbf{B_R}^{\mathsf{T}}\mathbf{A_R}^{-1}$. Matice $\mathbf{A}$ a $\mathbf{A_R}$ ani jejich inverze nemáme k dispozici. Proto budeme počítat pouze s aproximacemi jejich inverzí.

Výpočet $\widetilde{\mathbf{A}} \approx \mathbf{A}^{-1}$

Nejdříve se podívejme na to, co platí pro matici $\mathbf{A}$. Vyjdeme z toho, že pro matici $\mathbf{B}$ platí

$$\mathbf{BP} = (\mathbf{B}_1, \mathbf{0}),$$

kde matice $\mathbf{B}_1$ je obecně čtvercová matice s plnou hodností (v našem případě dokonce diagonální) a $\mathbf{P}$ je permutační matice.

Matici $\mathbf{P}^\mathsf{T}\mathbf{K}\mathbf{P}$ pak rozdělíme na bloky označené následovně:

$$\begin{pmatrix} \mathbf{K}_{11} & \mathbf{K}_{12} \\ \mathbf{K}_{21} & \mathbf{K}_{22} \end{pmatrix}$$

Užitím Schurova komplementu podmatice $\mathbf{K}_{22}$ dostáváme ze vztahu $\mathbf{A} = \mathbf{B}\mathbf{K}^{-1}\mathbf{B}^\mathsf{T}$ vyjádření

$$\mathbf{A} = \mathbf{B}_1 \left(\mathbf{K}_{11} - \mathbf{K}_{12}\, \mathbf{K}_{22}^{-1}\, \mathbf{K}_{21}\right)^{-1} \mathbf{B}_1^\mathsf{T}.$$

Z předchozího vztahu pro matici $\mathbf{A}^{-1}$ odvodíme

$$\mathbf{A}^{-1} = \mathbf{B}_1^{-1} \left(\mathbf{K}_{11} - \mathbf{K}_{12}\mathbf{K}_{22}^{-1}\, \mathbf{K}_{21}\right) \mathbf{B}_1^{-\mathsf{T}}. \tag{15}$$

Výpočet matice $\mathbf{A}^{-1}$ pomocí tohoto vzorce je výpočetně příliš náročný, proto jej neprovádíme a využijeme jej pouze jako základ pro výpočet aproximace $\widetilde{\mathbf{A}}$ matice $\mathbf{A}$. Aproximaci provádíme dvěma způsoby popsanými následujícími vzorci

$$\widetilde{\mathbf{A}} = \mathbf{B}_1^{-1} \left(\mathbf{K}_{11} - \mathbf{K}_{12}\ \left(\mathbf{diag}\ \mathbf{K}_{22}\right)^{-1} \mathbf{K}_{21}\right) \mathbf{B}_1^{-\mathsf{T}}, \tag{16a}$$

$$\widetilde{\mathbf{A}} = \mathbf{B}_1^{-1} \left(\mathbf{K}_{11} - \bar{\mathbf{K}}_{12}\ \left(\mathbf{U}_{11} - \bar{\mathbf{U}}_{12}\, \bar{\mathbf{V}}^{-1} \bar{\mathbf{U}}_{21}\right)^{-1} \bar{\mathbf{K}}_{21}\right) \mathbf{B}_1^{-\mathsf{T}} \tag{16b}$$

, kde $\bar{\mathbf{V}} = \mathbf{V}_{11} - \mathbf{V}_{12}\ \left(\mathbf{diag}\ \mathbf{V}_{22}\right)^{-1} \mathbf{V}_{21}$ a kde $\mathbf{P}_1$ resp. $\mathbf{P}_2$ je permutační matice taková, že platí $\mathbf{K}_{12}\mathbf{P}_1 = (\bar{\mathbf{K}}_{12}, \mathbf{0})$, což určuje rozdělení matice $\mathbf{P}_1^\mathsf{T}\mathbf{K}_{22}\mathbf{P}_1$ na bloky

$$\mathbf{P}_1^\mathsf{T}\mathbf{K}_{22}\mathbf{P}_1 = \begin{pmatrix} \mathbf{U}_{11} & \mathbf{U}_{12} \\ \mathbf{U}_{21} & \mathbf{U}_{22} \end{pmatrix}$$

a $\mathbf{P}_2$ je permutační matice taková, že platí $\mathbf{U}_{12}\mathbf{P}_2 = (\bar{\mathbf{U}}_{12}, \mathbf{0})$, čímž je určeno rozdělení bloků matice $\mathbf{P}_2^\mathsf{T}\mathbf{U}_{22}\mathbf{P}_2$ následovně:

$$\mathbf{P}_2^\mathsf{T}\mathbf{U}_{22}\mathbf{P}_2 = \begin{pmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} \\ \mathbf{V}_{21} & \mathbf{V}_{22} \end{pmatrix}$$

Postup provádění předpodmínění

Nyní už máme popsány všechny potřebné části výpočtu předpodmiňovací matice a můžeme tak celý postup shrnout následovně:

1. Na počátku výpočtu vypočítáme matici $\widetilde{\mathbf{A}} \approx \mathbf{A}^{-1}$ jedním z výše popsaných vzorců

2. V každé iteraci IPM (tj. ve vnější iteraci) vypočítáme aproximaci $\widetilde{\mathbf{A}}_\mathbf{R}$ matice $\mathbf{A}_{\mathbf{J}_R}^{-1}$ řešením soustavy

$$\left(\mathbf{I} + \widetilde{\mathbf{A}}\left(\mathbf{A}_\mathbf{R} - \mathbf{A}\right)\right) \widetilde{\mathbf{A}}_{\mathbf{J}_R} = \widetilde{\mathbf{A}} \tag{17}$$

   vzniklé užitím vztahu pro rozdíl dvou inverzních matic.

3. V každé iteraci IPM vypočítáme také matici $\mathbf{H} = \mathbf{D_R} - \mathbf{B_R}^{\mathsf{T}} \widetilde{\mathbf{A}}_\mathbf{R} \mathbf{B_R}$

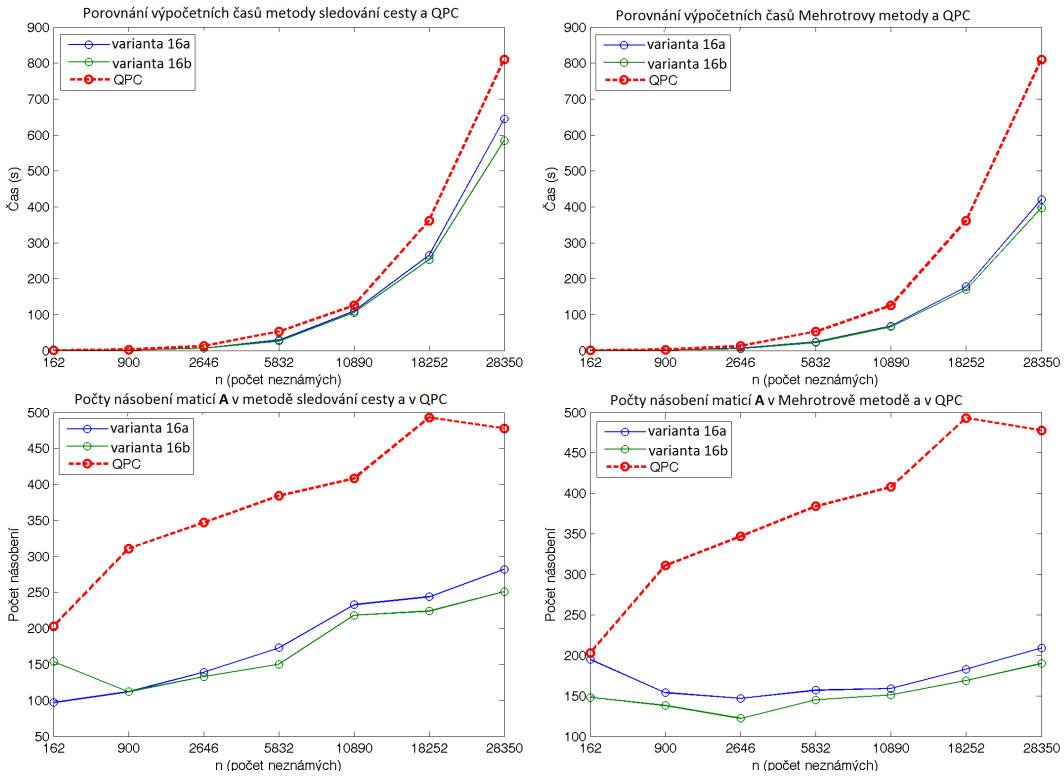
4. V každé iteraci iterační metody pro řešení soustav lineárních rovnic (tj. vnitřní iterace) provádíme předpodmínění maticí $\widetilde{\mathbf{J}}_R$ danou předpisem

$$\widetilde{\mathbf{J}}_R = \begin{pmatrix} \widetilde{\mathbf{A}}_\mathbf{R} - \widetilde{\mathbf{A}}_\mathbf{R}\mathbf{B_R}\mathbf{F} & -\widetilde{\mathbf{A}}_\mathbf{R}\mathbf{B_R}\mathbf{H}^{-1} \\ \mathbf{F} & \mathbf{H}^{-1} \end{pmatrix} \tag{18}$$

kde $\mathbf{F} = -\mathbf{H}^{-1}\mathbf{B_R}^{\mathsf{T}}\widetilde{\mathbf{A}}_\mathbf{R}$.

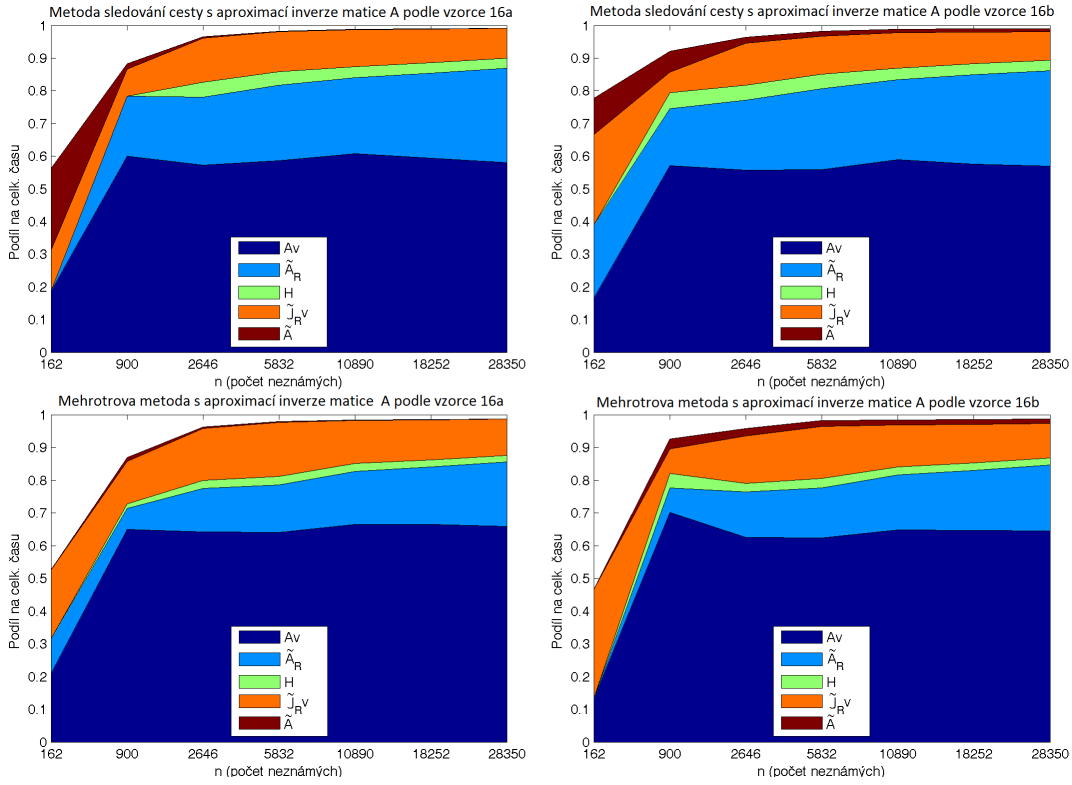
Numerické výsledky

Nyní můžeme přistoupit k porovnání metod vnitřních bodů, které jako vnitřní řešič používají metodu konjugovaných gradientů s předpodmíněním, s metodou QPC.



Obrázek 1: Porovnání metody sledování cesty (vlevo) a Mehrotrovy metody (vpravo) s metodou QPC (červeně) z hlediska výpočetního času (nahoře) a počtu násobení maticí $\mathbf{A}$ (dole) při užití metody konjugovaných gradientů (dole). Každé násobení maticí $\mathbf{A}$ odpovídá jednomu výpočtu řešení soustavy s maticí $\mathbf{K}$. Modře je vyznačena varianta metody, ve které je výpočet matice $\widetilde{\mathbf{A}}$ v předpodmínění prováděn podle vzorce (16a), zeleně je označen výpočet podle vzorce (16b).

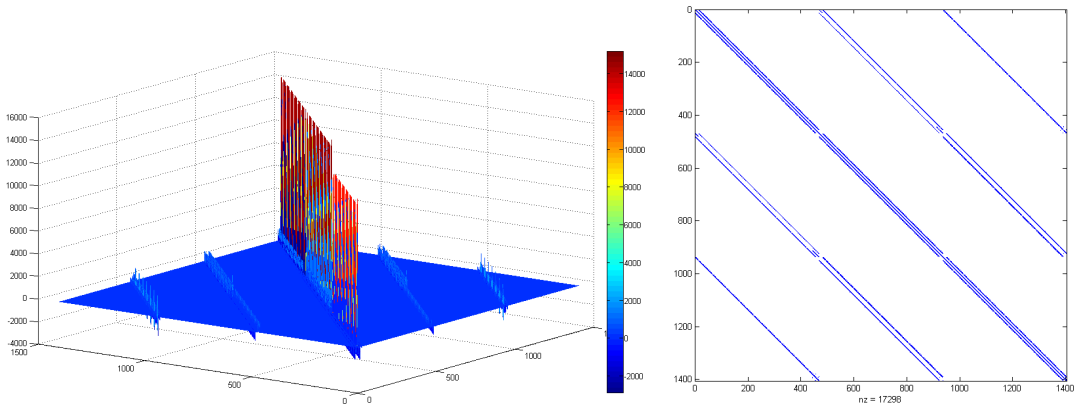
Z numerických výsledků plyne, že metody vnitřních bodů jsou pro testované úlohy srovnatelné resp. mírně rychlejší než QPC (viz. obr. 1), a tentokrát to není žádný klam způsobený Matlabem. Nicméně podíl výpočtu (obr. 2) předpodmínění na celkovém čase mírně narůstá, byť velmi pomalu, proto by pro (výrazně) větší úlohy mohlo docházet ke zpomalení metody oproti algoritmu QPC. Navíc je třeba v paměti ukládat plné matice, byť relativně malých rozměrů, což zvyšuje nároky na paměť.



Obrázek 2: Podíly jednotlivých částí výpočtu na celkovém výpočetním čase. Metoda sledování cesty je na obou horních obrázcích, Mehrotrova metoda na obou dolních. Vpravo jsou výsledky numerických testů pro výpočet matice $\widetilde{\mathbf{A}}$ podle vzorce (16a), vlevo jsou výsledky výpočtu užívajícího vzorec (16b). Tmavě modrou barvou je zvýrazněn podíl násobení matice $\mathbf{A}$ vektorem na celkovém čase, světle modře podíl výpočtu aproximace $\widetilde{\mathbf{A}}_{\mathbf{R}}$, světle zeleně podíl výpočtu matice $\mathbf{H}$, oranžově podíl násobení matice $\widetilde{\mathbf{J}}_R$ vektorem a hnědě podíl doby výpočtu matice $\widetilde{\mathbf{A}}$ na celkovém čase.

4.3 IPM bez přímého výpočtu matice A užitím iteračních metod s předpodmíněním řídkou maticí

Podívejme se nyní, zda by nebylo možné provádět předpodmínění tak, abychom nemuseli používat plnou matici k reprezentaci aproximace matice $\mathbf{A}^{-1}$. Tato matice je sice plná, nicméně většina prvků má vzhledem k největšímu prvku matice zanedbatelnou velikost, jak ukazuje obrázek 3.



Obrázek 3: Grafické znázornění velikosti prvků matice $\mathbf{A}^{-1}$ pro $n = 18252$ a $m = 1404$ (vlevo), a struktura nenulových prvků v matici, která vznikne z matice $\mathbf{A}^{-1}$v případě, že vymažeme všechny prvky, které jsou (v absolutní hodnotě) menší než 2,5 % největšího prvku (vpravo).

To nás vedlo k návrhu tří způsobů ukládání aproximace matice $\mathbf{A}^{-1}$ a to následovně:

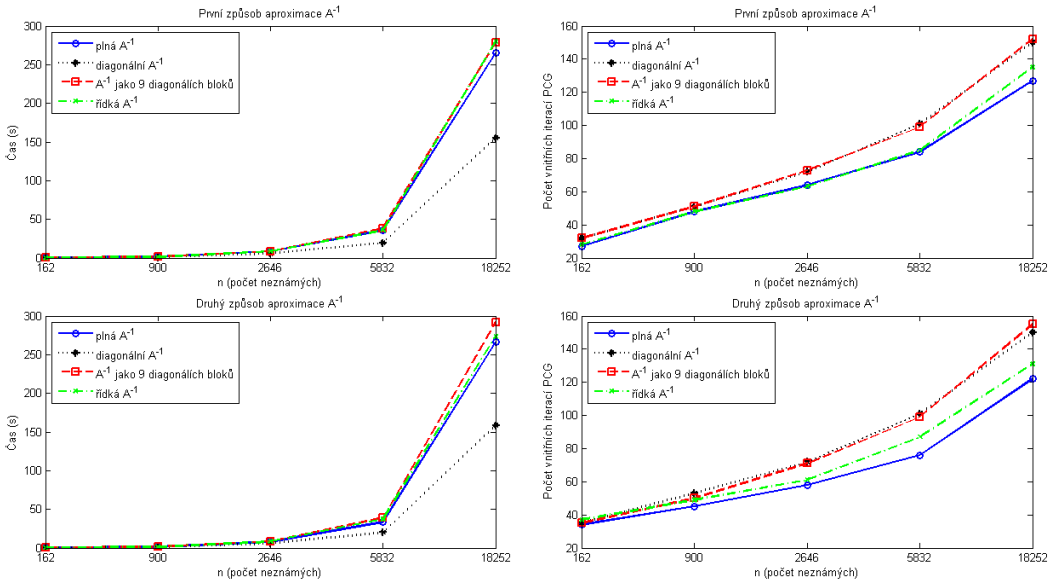
1. maticí diagonálních prvků (v literatuře známo jako indefinitní předpodmínění),

2. blokovou maticí s 9 bloky (3x3) tvořenými diagonálními maticemi,

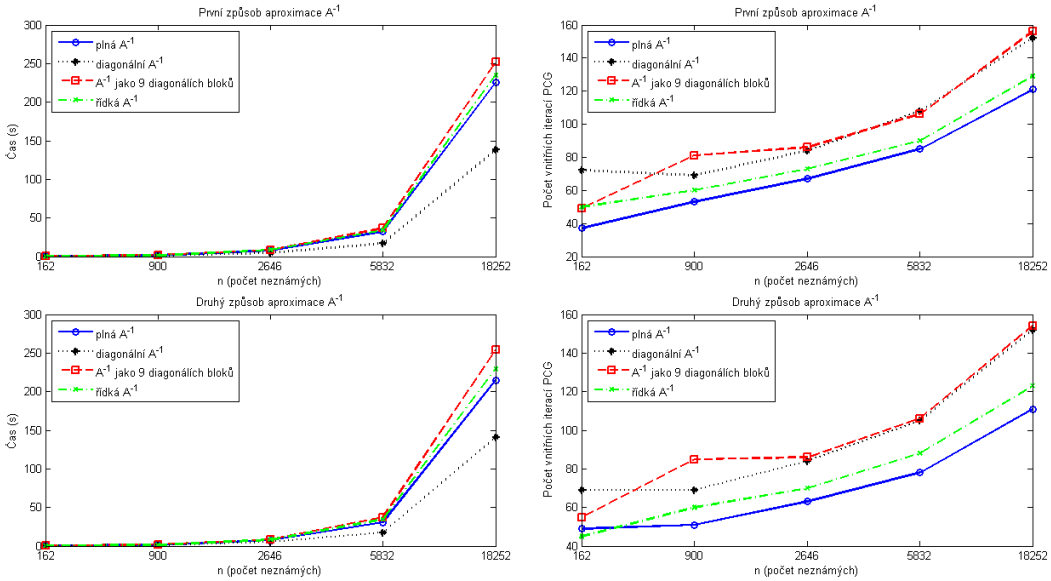
3. řídkou maticí uchovávající pouze největší prvky, které jsou v absolutní hodnotě větší než 2,5 % největšího prvku.

Numerické výsledky

Při numerických testech jsme porovnávali výpočet užitím těchto tří způsobů ukládání matice $\widetilde{\mathbf{A}}$ s výpočtem ukládajícím celou matici $\widetilde{\mathbf{A}}$ (červeně). Výsledky výpočtu s diagonální maticí jsou vyznačeny černě, výsledky s blokovou maticí zeleně a výsledky řídkou maticí zeleně. Porovnání jsme provedli pro metodu sledování cesty, Mehrotrovu metodu i pro algoritmus 4. Na obrázcích 4–6 jsou vždy



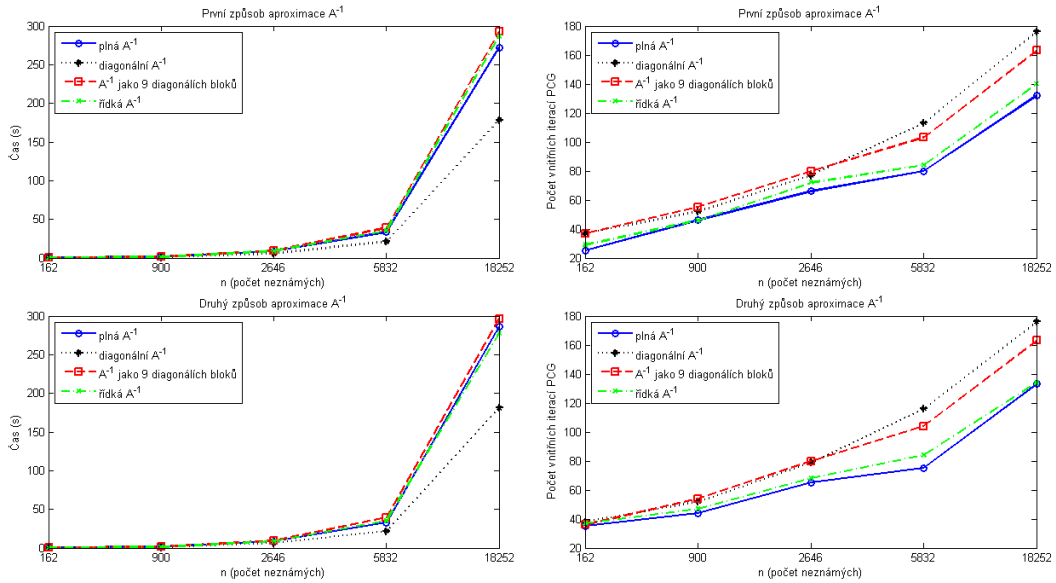
Obrázek 4: Porovnání různých předpodmínění pro metodu sledování cesty



Obrázek 5: Porovnání různých předpodmínění pro Mehrotrovu metodu

nahoře výsledky pro pro výpočet matice $\widetilde{\mathbf{A}}$ podle vzorce (16a) a dole jsou výsledky výpočtu užívajícího vzorec (16b). Vlevo je porovnání výpočetních časů a vpravo počtu vnitřních iterací.

Z grafů znázorňujících výsledky numerických testů je zřejmé, že u všech tří metod vnitřních bodů dochází při použití řídkých předpodmiňovacích matic k mírnému nárůstu počtu vnitřních iterací oproti variantě algoritmu využívající k předpodmínění plné matice. U indefinitního předpodmínění využívajícího pouze dia-



Obrázek 6: Porovnání různých předpodmínění pro algoritmus 4.

gonální matici je to ovšem vyváženo nižší výpočetní náročností předpodmínění, čímž dokonce dochází ke zrychlení algoritmu.

Literatura

[1] Griva, I., Shano, D. F., Vanderbei, R. J., Beson, H. Y.: *Global convergence of a primal-dual interior-point method for nonlinear programming.* Optimization Online, 2004.

[2] Haslinger, J, Dostál, Z., Kučera, R.: *An algorithm for the numerical realization of 3D contact problems with Coulomb frictions.* Journal of Computational and Applied Mathematics **164–165** (2004), 387–408.

[3] Kučera, R.: *Convergence rate of an optimal algorithm for minimizing quadratic functions with separable convex constraints.* 2007.

[4] Nocedal, J., Wright, S. J.: *Numerical Optimization.* Springer-Verlag, New York, 1999.

[5] Wright, S. J.: *Primal-dual interior-point methods.* SIAM, 1997.

Univ. Palacki. Olomuc., Fac. rer. nat.,
Dept of Math. Anal. and Appl. of Math.
ODAM (2009) 41–51

Nové možnosti v řešení úlohy ohybu nosníku s podložím

Horymír NETUKA

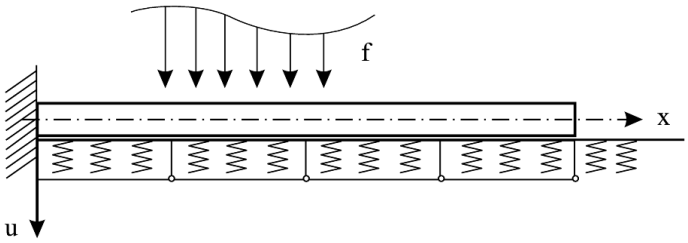
Katedra matematické analýzy a aplikací matematiky
Přírodovědecká fakulta, Univerzita Palackého,
Tř. 17. listopadu 1192/12, 771 46 Olomouc, Česká republika
e-mail: netuka@inf.upol.cz

Abstrakt

Příspěvek se zabývá tématem, které bylo prezentováno na předcházejících dvou ročnících ODAMu, a to nosníkem na pružném podloží. S ohledem na tuto skutečnost zde není podrobněji diskutován ani výchozí matematický model ani některé další detaily, které se týkají např. použití metody konečných prvků pro uvažovaný problém nebo úloh komplementarity a metod jejich řešení. Pozornost je soustředěna na zcela novou formulaci dané problematiky a možnosti jejího řešení, které odtud vyplývají.

1 Fyzikální situace

Zamyslíme-li se nad výchozí situací, která vzniká při sestavování matematicko-fyzikálního modelu nosníku na podloží, zjistíme, že k cíli vedou v zásadě dvě cesty. Ty se liší podle toho, zda je ve výsledném modelu uvažována



Obrázek 1: Nosník na Winklerově podloží

- 1 fyzikální entita, tj. nosník, a přes *odezvovou funkci* zahrnujeme vlivy okolí, tj. podloží,
- 2 fyzikální entity, tj. nosník i podloží, což vede na *kontaktní úlohu*.

Druhou možností se zde zabývat nebudeme, neboť tento způsob řešení naší relativně jednoduché úlohy je poněkud příliš komplikovaný.

2 Model s odezvovou funkcí

Standardně používaná metodika spočívá v použití odezvové funkce. Výhodou tohoto postupu je i to, že platí elementární rovnost

$$1 \text{ fyzikální entita } = 1 \text{ matematická rovnice}$$

V našem případě bude za předpokladu, že uvažujeme *Euler–Bernoulliův model nosníku*, taková matematická rovnice vypadat následovně

$$EI\, u^{IV}(x) + r(x) = f(x) \qquad \forall x \in (0, L), \tag{1}$$

kde $r(x)$ je příslušná odezvová funkce. Podrobně i z obecného hlediska je tématika odezvových funkcí zpracována v [7].

Jak je vcelku dobře známo, první model podloží je spojen se jménem Emila Winklera.

3 Emil Winkler – život a dílo

Emil WINKLER (německý inženýr) *18. 4. 1835 Falkenberg, †27. 8. 1888 Berlín

Navštěvoval stavební řemeslnickou školu v Holzmindenu, načež zaměstnán byl prakticky a po absolvování vysoké školy technické v Drážďanech přijat byl do služeb saského ředitelství vodních staveb. R. 1863 se habilitoval jako soukromý docent na polytechnice v Drážďanech, r. 1865 stal se profesorem stavitelství inženýrského na pražské a r. 1868 na vídeňské polytechnice. R. 1877 povolán na berlínskou stavební akademii.

Nejvýznamnější publikace:

- Die Lehre von der Elastizität und Festigkeit. Prag, 1867.
- Vorträge über Eisenbahnbau gehalten am königlich-Böhmischen polytechnischen Landesinstitut in Prag. Prag, 1869.
- Vorträge über Brückenbau. Wien, 1870.
- Neue Theorie des Erddruckes nebst einer Geschichte der Theorie des Erddruckes und der hierüber angestellten Versuche. Wien, 1872.
- Wahl der zulässigen Inanspruchnahme der Eisenkonstruktionen. Wien, 1877.

Zdroj: Ottův slovník naučný.

4 Pasternakovo podloží

Obrázek 2: Winklerovo podloží a pružné podloží

Jak je zřejmé z uvedeného obrázku, Winklerovo podloží nesplňuje zcela představy o pružném chování podloží. Trvalo poměrně dosti dlouhou dobu, než byly vypracovány modely, které se tomuto pojetí více přiblížily. Jedním z nejznámějších a nejpoužívanějších se stal model podloží, který publikoval ruský inženýr Petr Leonťjevič Pasternak (1885–1963) ve své práci

- `Osnovy novogo metoda rasčeta fundamentov na uprugom osnovanii pri pomošči dvuch koefficientov posteli`. Gosudarstvennoe izdatelstvo literatury po stroitelstvu i architekture, Moskva–Leningrad, 1954.

V tomto případě bude model podloží *dvouparametrický*, tj. závisející na dvou parametrech k_F a k_S které reprezentují tuhost podloží a tuhost smykových vrstev. Odezvová funkce pak bude vyhlížet pro klasické Winkler–Pasternakovo podloží takto

$$r(x) = k_F\, u(x) - k_S\, u''(x) \qquad x \in (0, L), \tag{2}$$

přičemž pro $k_S = 0$ máme čisté Winklerovo podloží.

Pod pojmem klasické podloží zde rozumíme *oboustranné*, tj. pevně spojené s nosníkem. Předmětem našeho zájmu je však tzv. *jednostranné* podloží, které s nosníkem spojeno není. Takové podloží lze charakterizovat následující odezvovou funkcí

$$r(x) = k_F\, u^+(x) - k_S\, (u^+)''(x) \qquad x \in (0, L), \tag{3}$$

kde používáme označení $u^+(x) = \max\{0, u(x)\}$.

5 Variační formulace

Nyní přistoupíme k variační formulaci uvažovaného jednostranného problému: nalézt $u \in V$ tak, že

$$J(u) = \min_{v \in V} J(v) \tag{4}$$

kde

$$J(v) = \frac{1}{2}\, EI \int_0^L (v'')^2 dx + \frac{1}{2}\, k_F \int_0^L (v^+)^2 dx + \\ + \frac{1}{2}\, k_S \int_0^L ((v^+)')^2 dx - \int_0^L f v\, dx \tag{5}$$

a V je prostor kinematicky přípustných průhybů, jehož konkrétní podoba závisí na zadaných okrajových podmínkách a platí $H_0^2((0,L)) \subseteq V \subseteq H^2((0,L))$.

Nebudeme zde podrobně diskutovat tuto úlohu a odkazujeme se na dosavadní práce z této oblasti, viz [6], [8], [9], [10], které se však týkají pouze jednostranného Winklerova podloží. Z pohledu autora je zde ale několik neuspokojivých skutečností:

1. Winklerovo podloží je aproximováno pomocí nosníkových prvků, což zjevně nesouhlasí s fyzikální realitou, viz obr. 2;

2. jelikož mají nosník i podloží společnou konečně-prvkovou síť, musí být za účelem získání kvalitního řešení použito velké množství nosníkových prvků, což se jeví z hlediska chování nosníku zbytečné.

Naším záměrem bude proto nalézt novou formulaci uvažované problematiky a tím i nové možnosti jejího řešení.

6 Kanonická injekce

Zavedeme do naší úlohy novou proměnnou $q \in Q$, přičemž bude $Q = H^1((0,L))$ pro Pasternakův model a $Q = L^2((0,L))$ pro Winklerův model podloží. Tato nová proměnná bude svázaná s původní proměnnou $v \in V$ *lineárním* vztahem

$$Bv = q, \tag{6}$$

kde B je obecně lineární spojitý operátor z V do Q. V našem případě konkrétně položíme

$$B = id, \tag{7}$$

takže B je tzv. *kanonická injekce* (anglicky *canonical injection*) z V do Q.

Na okraj poznamenejme, že volba $B = \Delta$ vede na *smíšenou metodu konečných prvků* (viz např. [1]), což však sleduje zcela jiné cíle, než náš postup. Pomocí něho totiž provedeme dekompozici uvažovaného problému.

7 Dekompozice

Namísto úlohy (4) budeme nyní uvažovat úlohu ekvivalentní, kterou obdržíme pomocí vztahů (6) a (7), přičemž od sebe oddělíme proměnné nosníku a podloží:

nalézt $[u,p] \in W = \{[v,q] \in V \times Q \colon v = q\}$ tak, že

$$\mathcal{J}(u,p) = \min_{[v,q] \in W} \mathcal{J}(v,q) \tag{8}$$

kde

$$\mathcal{J}(v,q) = \frac{1}{2}\,EI \int_0^L (v'')^2 dx \; + \frac{1}{2}\,k_F \int_0^L (q^+)^2 dx \; + \\ + \frac{1}{2}\,k_S \int_0^L ((q^+)')^2 dx \; - \int_0^L f v \, dx \tag{9}$$

Podstatné zde je, že přecházíme od nepodmíněné minimalizace k minimalizaci s podmínkou. To je sice složitější úloha, avšak nyní můžeme využít lagrangiány a sedlo-bodové techniky.

8 Lagrangián a sedlo-bodová formulace

Položme $V \subseteq H^2((0,L))$, $Q = H^1((0,L))$, přičemž pro $k_S = 0$ ale bude $Q = L^2((0,L))$, a $\Lambda = L^2((0,L))$. Definujme pro naši úlohu (8) *lagrangián* $\mathcal{L}$ na $V \times Q \times \Lambda$ následovně

$$\mathcal{L}(v,q,\mu) = \frac{1}{2}\,EI \int_0^L (v'')^2 dx + \frac{1}{2}\,k_F \int_0^L (q^+)^2 dx \; + \\ + \frac{1}{2}\,k_S \int_0^L ((q^+)')^2 dx - \int_0^L f v \, dx + \int_0^L \mu(v \; - q)\, dx \tag{10}$$

Dále definujme úlohu

nalézt sedlový bod $[u,p,\lambda] \in V \times Q \times \Lambda$ lagrangiánu $\mathcal{L}$ tak, že

$$\mathcal{L}(u,p,\mu) \le \mathcal{L}(u,p,\lambda) \le \mathcal{L}(v,q,\lambda) \qquad \forall v \in V, q \in Q, \mu \in \Lambda \tag{11}$$

což lze zformulovat i takto

nalézt sedlový bod $[u,p,\lambda] \in V \times Q \times \Lambda$ tak, že

$$\mathcal{L}(u,p,\lambda) = \inf_{[v,q]\in V\times Q} \sup_{\mu\in\Lambda} \mathcal{L}(v,q,\mu) = \sup_{\mu\in\Lambda} \inf_{[v,q]\in V\times Q} \mathcal{L}(v,q,\mu) \tag{12}$$

Možnosti řešení sedlo-bodové úlohy jsou v zásadě dvojí

1. přes soustavu nelineárních rovnic,

2. pomocí metody rozšířených lagrangiánů.

9 Soustava rovnic

Z (10) a (12) odvodíme následující soustavu v nekonečně-dimenzionálním prostoru

$$EI \int_0^L u'' v'' \, dx + \int_0^L \lambda v \, dx = \int_0^L f v \, dx \qquad \forall v \in V \tag{13}$$

$$k_F \int_0^L p^+ q \, dx + k_S \int_0^L (p^+)' q' \, dx - \int_0^L \lambda q \, dx = 0 \qquad \forall q \in Q \tag{14}$$

$$\int_0^L u \, \mu \, dx - \int_0^L p \mu \, dx = 0 \qquad \forall \mu \in \Lambda \tag{15}$$

Z jednotlivých rovnic soustavy (13)-(15) pak snadno obdržíme následující vztahy

$$u = p \qquad \text{sk. vš. v } L^2((0,L)), \tag{16}$$

pro Winklerův model

$$\lambda = k_F \, p^+ \qquad \text{sk. vš. v } L^2((0,L)) \tag{17}$$

a pro Pasternakův model *formálně* obdržíme

$$\lambda = k_F p^+ - k_S \, (p^+)'' \qquad \text{sk. vš. v } L^2((0,L)). \tag{18}$$

Soustavu (13)–(15) převedeme do konečné dimenze obvyklým postupem. Nechť V_h, Q_h and Λ_h jsou *konečnědimenzionální* podprostory prostorů V, Q a Λ. Nyní definujme úlohu

nalézt *diskrétní řešení* $u_h \in V_h, p_h \in Q_h$ a $\lambda_h \in \Lambda_h$ tak, že

$$EI \int_0^L u_h'' v_h'' \, dx + \int_0^L \lambda_h v_h \, dx = \int_0^L f_h v_h \, dx \quad \forall v_h \in V_h \tag{19}$$

$$k_F \int_0^L p_h^+ q_h \, dx + k_S \int_0^L (p_h^+)' q_h' \, dx - \int_0^L \lambda_h q_h \, dx = 0 \quad \forall q_h \in Q_h \tag{20}$$

$$\int_0^L u_h \mu_h \, dx - \int_0^L p_h \mu_h \, dx = 0 \quad \forall \mu_h \in \Lambda_h \tag{21}$$

Převedení do maticového tvaru nyní vyžaduje

1. zvolení vhodných prostorů konečných prvků,
2. splnění Babuška–Brezziho podmínek.
   Ty mají pro sedlo-bodovou úlohu

$$a(u_h, v_h) + b(v_h, p_h) = \langle f, v_h \rangle \qquad \forall v_h \in V_h \tag{22}$$

$$b(u_h, q_h) = \langle g, q_h \rangle \qquad \forall q_h \in Q_h \tag{23}$$

následující tvar (viz např. [1])

$$\inf_{u_h \in V_{h,0}} \sup_{v_h \in V_{h,0}} \frac{a(u_h, v_h)}{\|u_h\| \, \|v_h\|} = \alpha_h > 0 \tag{24}$$

$$\inf_{q_h \in Q_h} \sup_{v_h \in V_h} \frac{b(v_h, q_h)}{\|u_h\| \, \|q_h\|} = \beta_h > 0 \tag{25}$$

kde $V_{h,0} = \{v_h \in V_h \colon b(v_h, q_h) = 0 \quad \forall q_h \in Q_h\}$.

Předně budeme za účelem řešení vytvářet dvě konečně-prvkové sítě. První bude spojená s nosníkem, druhá s podložím a multiplikátory (to je dáno vztahy (17) resp. (18)). Tato druhá síť bude vnořena do sítě první a bude tedy mít počet prvků rovný celočíselnému násobku počtu prvků první sítě. Prostor V_h pak vygenerujeme pomocí standardních nosníkových prvků, viz např. [8]. Prostor Q_h zvolíme v závislosti na tom, zda se jedná o Winklerův nebo Pasternakův model podloží. V prvním případě zvolíme konstantní prvky, v druhém lineární. V případě prostoru Λ_h použijeme rovněž konstantní prvky. Při takovémto postupu lze ukázat splnění Babuška–Brezziho podmínek.

Soustava (19)-(21) má pak maticový tvar

$$\begin{aligned} \boldsymbol{K}_B \, \boldsymbol{u} \qquad\qquad + \boldsymbol{M}\boldsymbol{\lambda} &= \boldsymbol{f} \\ \boldsymbol{K}_F \, \boldsymbol{p}^+ - \boldsymbol{N}\boldsymbol{\lambda} &= \boldsymbol{o} \\ \boldsymbol{M}^T \boldsymbol{u} - \boldsymbol{N}^T \boldsymbol{p} \qquad\quad &= \boldsymbol{o} \end{aligned} \tag{26}$$

Zde značí $\boldsymbol{K}_B$ standardní matici tuhosti nosníku, $\boldsymbol{K}_F$ matici tuhosti podloží, $\boldsymbol{M}$ a $\boldsymbol{N}$ jsou matice vazeb mezi proměnnými, které jsou rozdílné s ohledem na to, že používáme dvě různé sítě. Přitom matice $\boldsymbol{K}_F$ obecně závisí na průběhu funkce p_h^+ a tato funkce je *nediferencovatelná*.

Odtud dostáváme následující možnosti řešení soustavy (26)

- nehladká Newtonova metoda,
- převod na úlohu komplementarity.

První možností se zde zabývat nebude, podrobněji se podíváme na druhou alternativu.

10 Úlohy komplementarity

Nejprve se seznámíme se základním vymezením úloh komplementarity. *Úlohou nelineární komplementarity* (NCP) rozumíme úlohu

pro danou vektorovou funkci $\boldsymbol{F}(\boldsymbol{z})$ nalézt vektor $\boldsymbol{z}$ tak, že

$$\begin{aligned} \boldsymbol{z}^T \, \boldsymbol{F}(\boldsymbol{z}) &= 0 \\ \boldsymbol{z} \, , \boldsymbol{F}(\boldsymbol{z}) &\geq 0 \end{aligned}$$

Je-li $\boldsymbol{F}$ afinní, tj. $\boldsymbol{F}(\boldsymbol{z}) = \boldsymbol{R}\boldsymbol{z} + \boldsymbol{q}$, kde $\boldsymbol{R}$ je daná čtvercová matice a $\boldsymbol{q}$ daný vektor, pak dostáváme *úlohu lineární komplementarity* (LCP)

nalézt vektory $\boldsymbol{w}$ a $\boldsymbol{z}$ tak, že

$$\begin{aligned} \boldsymbol{w} &= \boldsymbol{R}\boldsymbol{z} + \boldsymbol{q} \\ \boldsymbol{w}^T\boldsymbol{z} &= 0 \\ \boldsymbol{w}, \boldsymbol{z} &\geq 0 \end{aligned}$$

Jestliže se v úloze komplementarity vyskytují proměnné, pro které neplatí komplementární vztahy, pak zpravidla hovoříme o *úloze smíšené komplementarity*, kterou v obecném případě značíme (MiCP) a v případě afinní funkce (MLCP), podrobněji např. viz [3].

11 Převod na úlohu (MiCP) a její řešení

Soustavu (26) uvedeme na tvar odpovídající nějaké úloze komplementarity. Za tím účelem provedeme rozklad proměnné p_h v (19)-(21) na její kladnou a zápornou část a totéž indukovaně i pro jejího vektorového reprezentanta $\boldsymbol{p}$ v (26):

$$p_h = p_h^+ - p_h^- \quad \rightarrow \quad \boldsymbol{p} = \boldsymbol{p}^+ - \boldsymbol{p}^- \tag{27}$$

$$p_h^+, p_h^- \geq 0 \quad \rightarrow \quad \boldsymbol{p}^+, \boldsymbol{p}^- \geq 0 \tag{28}$$

Naše soustava (26) evidentně náleží do kategorie smíšené komplementarity a po úpravě vypadá nyní takto

$$\boldsymbol{K}_B\,\boldsymbol{u} + \boldsymbol{M}\boldsymbol{\lambda} = \boldsymbol{f} \tag{29}$$

$$\boldsymbol{K}_F\,\boldsymbol{p}^+ - \boldsymbol{N}\boldsymbol{\lambda} = \boldsymbol{o} \tag{30}$$

$$\boldsymbol{M}^T\boldsymbol{u} - \boldsymbol{N}^T\boldsymbol{p}^+ + \boldsymbol{N}^T\boldsymbol{p}^- = \boldsymbol{o} \tag{31}$$

$$\boldsymbol{p}^{+T}\boldsymbol{p}^- = 0 \tag{32}$$

$$\boldsymbol{p}^+, \boldsymbol{p}^- \geq 0 \tag{33}$$

Pro *Winklerovo podloží* máme $k_S = 0$ a prvky matice $\boldsymbol{K}_F$ dokážeme explicitně stanovit, neboť díky konstantním prvkům je matice diagonální a jakoby poskládána z řady pružin. Proto je výsledná úloha typu (MLCP). Tento typ úloh dokážeme bez větších problémů řešit např. pomocí Lemkeho metody nebo primárně-duální metody vnitřních bodů, viz např. [9].

Pro *Pasternakovo podloží* je $k_S \neq 0$ a výsledná úloha je typu (MiCP), neboť prvky matice $\boldsymbol{K}_F$ nelze explicitně určit. Můžeme pak např. použít metodu sekvenciální linearizace, která se autorovi osvědčila a která pro úlohu typu (NCP) postupuje podle následujícího obecného schématu:

k-tý krok: $\boldsymbol{z}^k$ je dáno a $\boldsymbol{z}^{k+1}$ získáme řešením (MLCP)

$$\begin{aligned} \boldsymbol{z}^T\left(\boldsymbol{F}(\boldsymbol{z}^k) + \boldsymbol{J}_F(\boldsymbol{z}^k)\,(\boldsymbol{z} - \boldsymbol{z}^k)\right) &= 0 \\ \boldsymbol{z} \geq 0,\ \boldsymbol{F}(\boldsymbol{z}^k) + \boldsymbol{J}_F(\boldsymbol{z}^k)\,(\boldsymbol{z} - \boldsymbol{z}^k) &\geq 0 \end{aligned}$$

přitom $\boldsymbol{J}_F$ značí jakobián funkce $\boldsymbol{F}$, tj. $(\boldsymbol{J}_F)_{ij} = \partial F_i/\partial x_j$.

Konvergenci iteračního procesu však nelze za obecných předpokladů zaručit.

12 Rozšířené lagrangiány

Velmi zajímavou alternativu k předchozímu více méně obvyklému postupu řešení pomocí soustavy nelineárních rovnic pro sedlový bod představuje použití metody rozšířených lagrangiánů. *Rozšířený lagrangián* $\mathcal{L}_r$ pro libovolné $r > 0$ je pro naši úlohu definován výrazem

$$\mathcal{L}_r(v, q, \mu) = \mathcal{L}(v, q, \mu) + \frac{r}{2} \int_0^L (v \ - q)^2 \, dx \qquad v \in V, q \in Q, \mu \in \Lambda \tag{34}$$

Dále definujme druhou sedlo-bodovou úlohu následovně

nalézt sedlový bod $[u, p, \lambda] \in V \times Q \times \Lambda$ funkcionálu $\mathcal{L}_r$ tak, že

$$\mathcal{L}_r(u, p, \mu) \ \leq \ \mathcal{L}_r(u, p, \lambda) \leq \mathcal{L}_r(v, q, \lambda) \quad \forall v \in V, q \in Q, \mu \in \Lambda \tag{35}$$

Použití rozšířeného lagrangiánu je založeno na následujícím tvrzení, jehož důkaz lze nalézt v [4] nebo [5].

Nechť $[u, p, \lambda]$ sedlový bod lagrangiánu $\mathcal{L}$ na $V \times Q \times \Lambda$. Potom $[u, p, \lambda]$ je sedlovým bodem rozšířeného lagrangiánu $\mathcal{L}_r$ pro každé $r > 0$ a naopak. Navíc u je řešením původní úlohy (4) a platí $p = u$.

Lze tudíž obě sedlo-bodové úlohy (11) a (35) zaměnit, což je představuje z výpočetního hlediska významnou výhodu. K určení sedlového bodu funkcionálu $\mathcal{L}_r$ použijeme následující variantu *Uzawova algoritmu* (horní index u proměnných zde představuje iteraci):

ALGORITMUS 1
$\lambda^0 \in \Lambda$ dáno
pak pro $n = 0, 1, \ldots$
určíme u^n, p^n řešením úlohy
nalézt $[u^n, p^n] \in V \times Q$ tak, že
$\mathcal{L}_r(u^n, p^n, \lambda^n) \leq \mathcal{L}_r(v, q, \lambda^n) \qquad \forall v \in V, \forall q \in Q$
určíme λ^{n+1} následovně
$\lambda^{n+1} = \lambda^n + \rho(u^n - p^n) \qquad \rho > 0$

Evidentně nejobtížnějším krokem tohoto postupu je vyřešení minimalizační úlohy v proměnných u a p pro $\mathcal{L}_r$ při pevném λ^n. Místo toho můžeme použít blokovou relaxaci a nahradit tento algoritmus následujícím, který vyžaduje minimalizaci vždy jen v jediné proměnné.

ALGORITMUS 2
$p^0 \in Q$, $\lambda^1 \in \Lambda$ dány
pak pro $n = 1, 2, \ldots$
určíme u^n, p^n řešením úloh

nalézt $u^n \in V$ tak, že
$\mathcal{L}_r(u^n, p^{n-1}, \lambda^n) \leq \mathcal{L}_r(v, p^{n-1}, \lambda^n) \qquad \forall v \in V$
nalézt $p^n \in Q$ tak, že
$\mathcal{L}_r(u^n, p^n, \lambda^n) \leq \mathcal{L}_r(u^n, q, \lambda^n) \qquad \forall q \in Q$
určíme λ^{n+1} následovně
$\lambda^{n+1} = \lambda^n + \rho(u^n - p^n) \qquad \rho > 0$

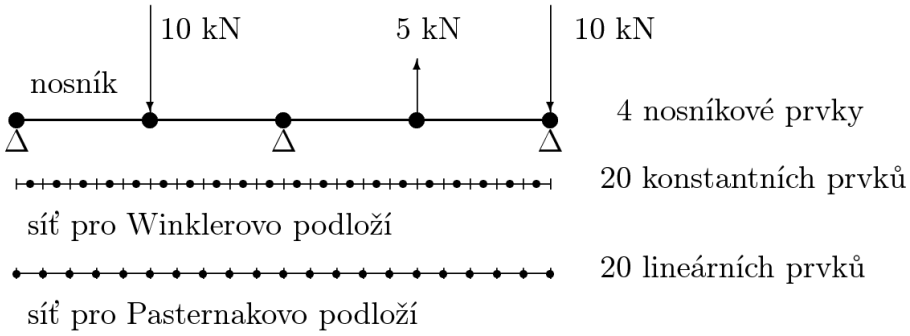
Pokud jde o konvergenci obou algoritmů, lze konstatovat, že

- algoritmus 1 konverguje pro $0 < \rho < 2r$
- algoritmus 2 konverguje pro $0 < \rho < ((1 + \sqrt{5})/2)r$

Důkazy lze nalézt v [4] nebo [5]. Dobrou volbou je zpravidla hodnota $\rho = r$.

13 Příklad

Nyní budeme ilustrovat výše vyloženou teorii a metody řešení na poměrně jednoduchém příkladu. Mějme dán nosník délky $L = 4\,m$ spočívající na podloží a třech podporách v bodech $x = 0\,m$, $x = 2\,m$ a $x = 4\,m$. Nosník a podloží jsou zadány pomocí následujících hodnot:
$EI = 2 \times 10^7\, Nm^2$,
$k_F = 2 \times 10^7\, Nm^{-3}$,
$k_S = 3 \times 10^7\, Nm^{-1}$.



Obrázek 3: Náčrt zadání příkladu

Na nosník působí tři osamělé síly zadané v bodech $x = 1\,m$, $x = 3\,m$ a $x = 4\,m$, jak je to patrné z uvedeného obrázku. Zde je také vidět ukázka, jak lze konstruovat konečně-prvkové sítě pro nosník a podloží (kulaté body označují uzly jednotlivých sítí).

V následující tabulce vidíme výsledky pro maximální a minimální průhyby pro několik různých sítí. Tyto hodnoty musí být přenásobeny hodnotou $10^{-5}\,m$. V případě Winklerova podloží je pochopitelně $k_S = 0$.

Z výsledků lze vypozorovat, že konvergence je v případě Winklerova podloží rychlejší než u Pasternakova. To ovšem není nijak překvapivý závěr, neboť Pasternakův model je komplexnější než původní Winklerův.

počet prvků		Winklerovo podloží				Pasternakovo podloží			
		klasické		jednostranné		klasické		jednostranné	
nosník	podl.	u max	u min	u max	u min	u max	u min	u max	u min
4	40	6.237	-4.525	6.433	-5.036	4.230	-2.872	4.657	-4.307
4	100	6.236	-4.524	6.431	-5.035	4.233	-2.873	4.660	-4.307
4	400	6.236	-4.524	6.431	-5.035	4.233	-2.874	4.661	-4.307
8	40	6.238	-4.526	6.433	-5.036	4.238	-2.878	4.664	-4.312
8	104	6.237	-4.525	6.432	-5.036	4.240	-2.880	4.667	-4.311
8	400	6.237	-4.525	6.432	-5.035	4.241	-2.880	4.667	-4.311

Literatura

[1] Brezzi, F., Fortin, M.: *Mixed and hybrid finite element methods.* Springer, Berlin, 1991.

[2] Ekeland, I., Témam, R.: *Convex analysis and variational problems.* SIAM, Philadelphia, 1999.

[3] Facchinei, F., Pang, J.-S.: *Finite-dimensional variational inequalities and complementarity problems. Volume I and II.* Springer, New York, 2003.

[4] Fortin, M., Glowinski, R.: *Augmented Lagrangian methods: Applications to the numerical solution of boundary-value problems.* North-Holland, Amsterdam, 1983.

[5] Glowinski, R.: *Numerical methods for nonlinear variational problems.* Springer, Berlin–Heidelberg, 1984.

[6] Horák, J. V., Netuka, H.: *Matematický model třídy nelineárních podloží Winklerovského typu: I. Spojitý případ.* Proceedings of 21st conference with international participation Computational Mechanics 2005, Hrad Nečtiny, November 7–9, 2005, published by UWB in Pilsen, 2005, 235–242.

[7] Horák, J. V., Netuka, H.: *Mathematical model of pseudointeractive set: 1D body on nonlinear subsoil. I. Theoretical aspects.* Engineering Mechanics **14**, 5 (2007), 311–325.

[8] Netuka, H., Horák, J. V.: *Matematický model třídy nelineárních podloží Winklerovského typu: II. Diskrétní případ.* Proceedings of 21st conference with international participation Computational Mechanics 2005, Hrad Nečtiny, November 7–9, 2005, published by UWB in Pilsen, 2005, 431–438.

[9] Netuka, H., Horák, J. V.: *Soustava nosník–pružiny–podloží po dvou letech.* Proceedings ODAM 2007, Dept. Math. Anal. and Appl. Math., Fac. Sci., Palacky Univ., Olomouc (2007), 18–42.

[10] Sysala, S.: *Unilateral elastic subsoil of Winkler's type: Semi-coercive beam problem.* Applications of Mathematics **53**, 4 (2008), 347–379.


Univ. Palacki. Olomuc., Fac. rer. nat.,
Dept of Math. Anal. and Appl. of Math.
ODAM (2009) 52–67

Interakce dvou elastických těles: Algoritmizace úlohy

Ivona SVOBODOVÁ

VŠB-TU Ostrava, kMDg

Abstrakt

V příspěvku pracujeme nehladkým funkcionálem $\mathcal{P}_\psi$, který popisuje potenciální energii elastického tělesa, jehož deformace je ovlivněna odezvou jiného tělesa nebo prostředí. Tím prvním je tenká mezikruhová deska s Neumannovými okrajovými podmínkami, na kterou působí předem známá zatížení. Druhým tělesem je pak nadloží, podloží nebo překážka, jejichž vliv je popsán nediferencovatelným operátorem ψ. Cílem příspěvku je prezentace a porovnání dvou algoritmů pro výpočet diskrétního řešení. První je důsledkem aplikace nehladké Newtonovy metody, druhý vyplývá z použití metody postupných aproximací.

1 Úvod

Problematika úloh kontaktu dvou elastických těles patří k častým otázkám plynoucím například z inženýrské praxe. V literatuře je této obtížné problematice věnováho mnoho prostoru, viz například [1] a reference tam uvedené. Ke zjednodušení dospějeme, pokud pozornost zaměříme jen na jedno z těles. Řešená úloha se redukuje na jednu rovnici rovnováhy, odpovídající tomuto tělesu, rozšířenou o speciální člen, který popisuje vliv odezvy tělesa zbývajícího.

V tomto příspěvku je tělesem, na které se zaměříme, elastická mezikruhová rotačně symetrická deska

$$Q = \left\{ (x, y, z) \in \mathbb{R}^3;\ a < \sqrt{x^2 + y^2} < b,\ -\frac{\mathsf{h}}{2} < z < \frac{\mathsf{h}}{2} \right\},$$

kde a, b, h jsou dané kladné konstanty, $a < b$ a $\mathsf{h} \ll 1$. Odpovídající matematický model je odvozen na základě lineární teorie elasticity. Vzhledem k rotační symetrii jsou veškeré formulace uvedeny ve tvaru po transformaci do cylidrických souřadnic

$$\begin{aligned} r &= \sqrt{x^2 + y^2}, \\ \varphi &= \operatorname{arctg}\left(\tfrac{y}{x}\right), \\ z &= z\ . \end{aligned}$$

Funkce průhybu desky splňuje rovnici rovnováhy, která je obyčejnou diferenciální rovnicí (ODR) čtvrtého řádu, a Neumannovy okrajové podmínky. Variačnímu řešení odpovídá argument minima potenciálu celkové energie na vhodném prostoru funkcí. Vzhledem ke geometrii úlohy je tento prostor váhovým Sobolevovým prostorem [3].

Výchozím bodem pro popis vlivu druhého tělesa je tzv. Winklerův model podloží a jeho zobecněná jednostranná varianta. Speciálním členem, jímž je rovnice rovnováhy rozšířena, je operátor

$$\psi(u) = \sum_{i=1}^{m} k_{N_i}(u - L_{+i})^{+} - \sum_{j=1}^{n} k_{P_j}(u + L_{-j})^{-},$$

kde funkce k_{N_i} popisují průběh odezvy i-té vrstvy, která je od desky vzdálená L_{+i} ve směru kladné poloosy z. Podobně lze interpretovat k_{P_j} a L_{-j}. Zahrneme-li operátor ψ do rovnice rovnováhy mezikruhové desky, stane se funkcionál celkové energie systému nehladkým. Proto bude zapotřebí uvést podmínky pro existenci a jednoznačnost variačního řešení. Klasická i variační formulace úlohy včetně podmínek existence variačního řešení je shrnuta v kapitole 2.

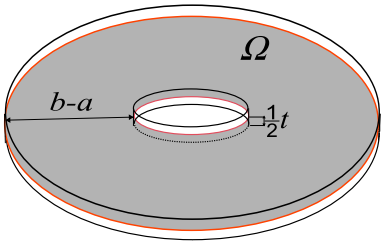
Energetický funkcionál navíc zůstáná nehladkým i v algebraické reprezentaci, která vyplne z diskretizace úlohy ve smyslu konečných prvků. Cílem tohoto příspěvku je proto hledání vhodné výpočtové metody. V kapitole 3 představíme dva algoritmy. První je založen na nehladké verzi Newtonovy metody, která dosahuje superlineární konvergence [6]. Nevýhodou je ale požadavek na dostatečně blízkou počáteční aproximaci. Druhý algoritmus odvozený z metody postupných aproximací se jeví jako globálně konvergentní, nicméně přesný důkaz zatím zůstává otevřenou otázkou. Numerické výsledky získané oběma metodami vzájemně porovnáme v kapitole 4.

2 Formulace úlohy

Deformaci desky Q charakterizujeme vektorovým polem U, který popisuje výsledné posunutí. Z Kirchhoffovy teorie pro tenké desky (tloušťka $\mathsf{h} \ll 1$) plyne, že složky U jsou ve tvaru

$$U_r = -z\frac{\partial}{\partial r}w(r,\varphi), \qquad U_\varphi = -z\tfrac{1}{r}\frac{\partial}{\partial \varphi}w(r,\varphi), \qquad U_z = w(r,\varphi).$$

pro $(r, \varphi, z) \in (a, b) \times (-\pi, \pi) \times (-\mathsf{h}/2, \mathsf{h}/2)$. Nechť Ω je tak zvaná střednicová



Obrázek 1: Střednicová plocha desky Ω.

plocha desky, $\Omega = \left\{(x, y) \in \mathbb{R}^2 \ ; \ a < \sqrt{x^2 + y^2} < b\right\}$, viz Obrázek 1. Transformaci oblasti Ω a její hranice $\partial\Omega$ do cylindrických souřadnic označme Ω_c a $\partial\Omega_c$, tj. $\Omega_c = (a, b) \times (-\pi, \pi)$ a $\partial\Omega_c = \{a, b\} \times (-\pi, \pi)$. Rovnice rovnováhy tenké mezikruhové desky je ve tvaru

$$D_0\mathsf{h}^2 \triangle_c^2 w(r, \varphi) + \psi(w)(r, \varphi) = f(r, \varphi) \qquad (r, \varphi) \in \Omega_c,$$

kde $D_0\mathsf{h}^2 \in \mathbb{R}^1$, $\triangle_c$ je Laplaceův operátor transformovaný do cylidrických souřadnic a zobrazení ψ popisuje vliv druhého tělesa (okolního prostředí desky). Neumannovy okrajové podmínky jsou dané rovnostmi

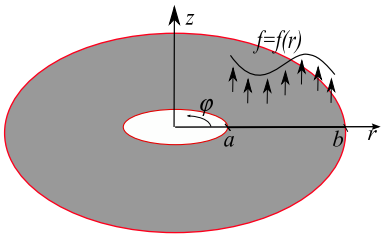
$$\begin{aligned} \mathcal{M}_n(w) &= m_n, \\ \mathcal{T}_n(w) + \frac{1}{r}\frac{\partial}{\partial s}\left[r\mathcal{M}_{ns}(w)\right] &= p_n, \end{aligned} \qquad \text{na } \partial\Omega_c$$

pro hraniční operátory $\mathcal{T}_n$, $\mathcal{M}_n$ a $\mathcal{M}_{ns}$ mající význam normálových sil, ohybových a kroutících momentů.

Neboť celý systém těles i působících sil je rotačně symetrický, složky vektorového pole U se zjednoduší na

$$U_r = -zu'(r), \qquad U_\varphi = 0, \qquad U_z = u(r),$$

kde $(r, z) \in (a, b) \times (-\mathsf{h}/2, \mathsf{h}/2)$, $u = u(r)$ je tzv. *funkce průhybu* a $(\ )' := \frac{\partial}{\partial r}(\ )$.



Obrázek 2: Řez (a, b) střednicovou plochou desky Ω.

2.1 Klasická a variační formulace úlohy

Klasické řešení $u \in C^4\big((a,b)\big)$ úlohy ohybu tenké mezikruhové rotačně symetrické desky splňuje okrajovou úlohou složenou z *rovnice rovnováhy*

$$D_0\mathsf{h}^2 \triangle_c^2 u\ (r) + \psi(u)(r) = f(r) \qquad r \in (a,b) \tag{1}$$

a Neumannových okrajových podmínek

$$\mathcal{M}u(r) = m_r \quad \text{a} \quad \mathcal{T}u(r) = p_r \qquad \text{pro } r \in \{a,b\}, \tag{2}$$

kde $\triangle_c^2 u(r) = \frac{1}{r}[r[\frac{1}{r}[r\cdot u'(r)]']']'$, konstanta D_0 závisí na elastických materálových koeficientech nazvaných Youngův modul pružnosti $E > 0$, a Poissonovo číslo $\sigma \in (0, \frac{1}{2})$, $D_0 = \frac{E}{12(1-\sigma^2)}$. Funkce $f \in C^0\big((a,b)\big)$ popisuje dané objemové síly, které působí kolmo k Ω a které jsou konstantní na každé kružnici se středem v počátku souřadnic, viz Obrázek 2. Hodnoty m_a, m_b, p_a a p_b v okrajových podmínkách (2) jsou dané a operátory $\mathcal{T}$ a $\mathcal{M}$ reprezentující *příčné síly* a *ohybové momenty* na hranici jsou definovány

$$\mathcal{T}u\ := D_0\mathsf{h}^2\ (r\,u''' + u'' - \tfrac{1}{r}u') \quad \text{a} \quad \mathcal{M}u\ := D_0\mathsf{h}^2\ (r\,u'' + \sigma\,u').$$

Operátor ψ je obecně dán rovností

$$\psi(u) = \sum_{i=1}^{m} k_{N_i}(u - L_{+i})^+ - \sum_{j=1}^{n} k_{P_j}(u + L_{-j})^- \qquad m,n \in \mathbb{N}. \tag{3}$$

kde $k_{N_i}, k_{P_j} \in C^0\big((a,b)\big)$ i $L_{+i}, L_{-j} \in \mathbb{R}^1$ jsou nezáporné. Kladná a záporná část funkce v jsou definovány pořadě

$$v^+ := \tfrac{1}{2}(|v| + v) \quad \text{a} \quad v^- := \tfrac{1}{2}(|v| - v).$$

Poznámka 1 (fyzikální smysl operátoru ψ)
V matematickém modelu (1)–(2) je protřednictvím ψ zahrnut vliv předem neznámé aktivní části elastického prostředí. Operátor (3) popisuje prostředí s m vrstvami nad deskou a n vrstvami pod ní. Vzdálenost mezi i-tou vrstvou a deskou je L_{+i} v případě horního prostředí a L_{-i} je vzdáleností dolní vrstvy. Funkce $k_{N_i} = k_{N_i}(r)$ popisuje průběh odezvy i-té horní vrstvy a $k_{P_j} = k_{P_j}(r)$ charakterizuje odezvu j-té dolní. V případě, že dojde k průhybu desky v rozmezí $L_{-(k+1)}$ až L_{-k}, pak je výsledná deformace ovlivněna první až k-tou dolní vrstvou. Celkový průhyb se zmenší o reakci $\sum_{i=1}^{k} L_{-i}\,k_{P_i}(r)$.

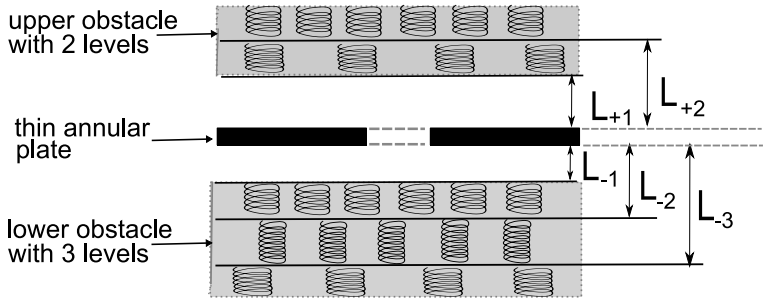
Pokud konstanty L_{+i} seřadíme tak, že $L_{+1} = \min_i L_{+i}$, pak L_{+1} je vzdáleností horního prostředí od desky. V případě, že $L_{+1} > 0$, jde o horní překážku. Pokud $L_{+1} = 0$ mluvíme o nadloží. Analogickou úvahu lze provést i pro konstanty L_{-j}, viz Obrázek 3.

Zvolený tvar operátoru ψ umožňuje popsat například tyto fyzikální případy.

1. Vliv lineárního prostředí o tuhosti k, $\psi(u) = ku$, jestliže položíme $k_{N_1} = k_{P_1} = k$ na (a,b), $L_{+1} = L_{-1} = 0$, $k_{N_i} \equiv 0$ pro $i = 2,3,\dots m$ a $k_{P_j} \equiv 0$ pro $j = 2,3,\dots n$;

2. Vliv jednostranného Winklerova podloží o tuhosti k_{P_1}, $\psi(u) = -k_{P_1}u^-$, položíme-li $L_{-1} = 0$, $k_{N_i} \equiv 0$ pro $i = 1,2,3,\dots m$ a $k_{P_j} \equiv 0$ pro $j = 2,3,\dots n$;

3. Vliv horní elastické překážky o tuhosti k_{N_1} a vzdálenosti od desky L_{+1}, $\psi(u) = k_{N_1}(u - L_{+1})^+$, jestliže dosadíme $L_{+1} > 0$, $k_{N_i} \equiv 0$ pro $i = 2,3,\dots m$ a $k_{P_j} \equiv 0$ pro $j = 1,2,3,\dots n$.



Obrázek 3: Příklad mezikruhové desky s horní dvou-vrstvou a dolní tří-vrstvou překážkou.

Ke zobecnění klasické formulace je zapotřebí zavést tzv. prostor funkcí s konečnou energií. Uvažujme funkci ρ kladnou s.v. na (a,b) a označme $L^2_{\varrho(r)}\big((a,b)\big)$ váhový Lebesgueův prostor se skalárním součinem $(u,v)_{\varrho(r)} := \int_a^b u(r)\,v(r)\,\varrho(r)dr$ indukujícím normu $|\cdot|_{\varrho(r)}$. *Prostorem funkcí s konečnou energií* je váhový Sobolevův prostor

$$H^2\big((a,b);r,\tfrac{1}{r},r\big) := \{v = v(r) \mid v, v'' \in L^2_r\big((a,b)\big) \ \wedge \ v' \in L^2_{\frac{1}{r}}\big((a,b)\big)\}.$$

se skalárním součinem $((u,v))_{[r,\frac{1}{r},r]} := (u,v)_r + (u',v')_{\frac{1}{r}} + (u'',v'')_r$ a indukovanou normou $\|\cdot\|_{[r,\frac{1}{r},r]}$. Blíže viz [3].

Lineární prostor $H^2\big((a,b);r,\frac{1}{r},r\big)$ není pouze prostorem funkcí s konečnou energií ale i *prostorem virtuálních posunutí*, který odpovídá zobecněné formulaci úlohy (1), (2). Tedy (v návaznosti na ustálené značení) položme

$$V \ = H^2\big((a,b);r,\tfrac{1}{r},r\big).$$

O veličinách v definici (3) operátoru ψ předpokládejme, že

$$\left.\begin{array}{rrrl} k_{N_i} \in L^\infty\big((a,b)\big), & k_{N_i} \geq 0 & \text{s.v. na } (a,b), \\ L_{+i} \in \mathbb{R}^1, & L_{+i} \geq 0 & \text{pro } i = 1,2,\dots m, \\ k_{P_j} \in L^\infty\big((a,b)\big), & k_{P_j} \geq 0 & \text{s.v. na } (a,b), \\ L_{-j} \in \mathbb{R}^1, & L_{-j} \geq 0 & \text{pro } j = 1,2,\dots n, \end{array}\right\} \tag{4a}$$

kde $m, n \in \mathbb{N}$, a navíc

$$-L_{-n} \leq \cdots \leq -L_{-2} \leq -L_{-1} \leq 0 \leq L_{+1} \leq L_{+2} \leq \cdots \leq L_{+m}. \tag{4b}$$

Předpoklady (4a) a (4b) jsou důsledky fyzikální interpretace ψ, viz Poznámku 1.

Pro funkce $w, v \in V$ definujme zobrazení

$$\left.\begin{aligned} a_0(w,v) &:= D_0 \mathsf{h}^2 \left((w', v')_{\frac{1}{r}} + (w'', v'')_r + \sigma(w'', v')_1 + \sigma(w', v'')_1 \right), \\ a_\psi(w,v) &:= (\psi(w), v)_r, \\ \mathcal{F}(w) &:= (f, w)_r - \langle p_r, w \rangle_{\{a,b\};r} + \langle m_r, w' \rangle_{\{a,b\};r}, \end{aligned}\right\} \tag{5}$$

pro danou funkci $f \in L^2_r\big((a,b)\big)$ a hodnoty $p_a, p_b, m_a, m_b \in \mathbb{R}^1$, které pocházejí z dualit

$$\langle p_r, w \rangle_{\{a,b\};r} = p_b\, w(b)\, b - p_a\, w(a)\, a \quad \text{a} \quad \langle m_r, w' \rangle_{\{a,b\};r} = m_b\, w'(b)\, b - m_a\, w'(a)\, a.$$

Funkcionál $\mathcal{P}_\psi \colon V \mapsto \mathbb{R}^1$,

$$\mathcal{P}_\psi(v) := \tfrac{1}{2} a_0(v,v) + \tfrac{1}{2} a_\psi(v,v) - \mathcal{F}(v)\,, \tag{6}$$

je *energetickým potenciálem* studované úlohy. Funkci $u \in V$ nazveme *variačním řešením*, jestliže

$$u \ = \arg\min_{v \in V} \mathcal{P}_\psi(v). \tag{7}$$

Potenciál $\mathcal{P}_\psi$ je konvexní a diferencovatelný na V. Proto u z (7) lze ekvivalentně charakterizovat Eulerovou extremální podmínkou

$$a_0(u,v) + a_\psi(u,v) = \mathcal{F}(v) \qquad \text{pro všechna } v \ \in V\ . \tag{8}$$

2.2 Existence variačního řešení

V této podkapitole shrneme tvrzení o existenci řešení variační úlohy (7). Pokud nebude napsáno jinak, pak v následujících větách předpokládáme, že potenciál $\mathcal{P}_\psi$ je definován (6), (5) a veličiny v operátoru ψ splňují obecné předpoklady (4a) a (4b).

Věta 1 *Uvažujme variační úlohu* (7) *pro* ψ *ve tvaru*

$$\psi(u) = \sum_{i=1}^{m} k_{N_i} (u - L_{+i})^+,$$

kde $k_{N_1} > 0$ *s.v. na* (a,b).

(i) Jestliže existuje variační řešení úlohy (7)*, potom* $\mathcal{F}(1) \geq 0$.

(ii) Jestliže $\mathcal{F}(1) > 0$, potom existuje jediné variační řešení úlohy (7).

(iii) Jestliže $\mathcal{F}(1) = 0$, potom existuje variační řešení úlohy (7), které je jednoznačné až na aditivní konstantu z intervalu $(-\infty, L_{+1})$

Věta 2 *Uvažujme variační úlohu (7) pro ψ ve tvaru*

$$\psi(u) = -\sum_{j=1}^{n} k_{P_j}(u + L_{-j})^{-},$$

kde $k_{P_1} > 0$ s.v. na (a, b).

(i) Jestliže existuje variační řešení úlohy (7), potom $\mathcal{F}(1) \leq 0$.

(ii) Jestliže $\mathcal{F}(1) < 0$, potom existujejediné variační řešení úlohy (7).

(iii) Jestliže $\mathcal{F}(1) = 0$, potom existuje variační řešení úlohy (7), které je jednoznačné až na aditivní konstantu z intervalu $(-L_{-1}, +\infty)$.

Věta 3 *Uvažujme variační úlohu (7) pro ψ ve tvaru (3). Jestliže*

$$\mathcal{F}(1) \neq 0\,,$$

potom existuje jediné variační řešení úlohy (7).

Věta 4 *Uvažujme variační úlohu (7) pro ψ ve tvaru (3), kde $k_{N_1} > 0$ a $k_{P_1} > 0$ s.v. na (a, b). Předpokládejme, že*

$$\mathcal{F}(1) = 0.$$

(i) Jestliže $L_{+1} \neq 0$ nebo $L_{-1} \neq 0$, potom existuje variační řešení úlohy (7), které je jednoznačné až na aditivní konstantu z intervalu $(-L_{-1}, L_{+1})$.

(ii) Jestliže $L_{+1} = L_{-1} = 0$, potom existuje jediné variační řešení úlohy (7).

Prezentace důkazů není cílem tohoto příspěvku. Proto jen shrňme, že základem důkazů o existenci řešení je direktní ortogonální rozklad Hilbertova prostoru V na konvexní kužel K a příslušný negativní polární kužel $K^{\ominus}$. Na $K^{\ominus}$ je zobrazení $a_0 + a_\psi$ z (8) ekvivalentním skalárním součinem na V, a tedy existence variačního řešení je zajištěna. Na $V \setminus K$ jsou zapotřebí dodatečné podmínky, které si existenci řešení vynutí. Jak je patrné z uvedených vět, jsou tyto požadavky kladeny na data úlohy obsažená ve funkcionálu $\mathcal{F}$. Důkazy o jednoznačnosti respektive jednoznačnosti až na aditivní konstantu lze provést sporem.

3 Diskretizace a algebraická formulace

Uvažujme množinu uzlů dělení intervalu $\langle a, b\rangle$

$$\mathcal{N}_h := \{r_i \,:\, i = 1, \dots N+1\},$$

kde

$$a = r_1 < r_2 < \cdots < r_N < r_{N+1} = b \tag{9}$$

pro $N \in \mathbb{N}$. Parametr h je normou diskretizace, $h = \max_i\{r_{i+1} - r_i\}$. Dělení $\mathcal{N}_h$ odpovídá konečně prvkový prostor

$$V_h := \{v_h \in C^1\big((a,b)\big) \,:\, v_h|_{\langle r_i, r_{i+1}\rangle} \text{ kubický polynom pro } i = 1, \dots N\}, \tag{10}$$

jehož dimenze je $2N + 2$. Bázi V_h označíme $\{\varphi_k\}_{k=1}^{2N+2}$. Každému prvku $v_h \in V_h$ lze tedy přiřadit vektor $v = (v_k)_{k=1}^{2N+2}$, jehož složky jsou souřadnice funkce v_h vzhledem k $\{\varphi_k\}_{k=1}^{2N+2}$, kde $v_{2i-1} = v_h(r_i)$, $v_{2i} = v_h'(r_i)$ pro všechna $i = 1 \dots N+1$.

Hledané *diskrétní řešení* $u_h \in V_h$ splňuje rovnici

$$a_0(u_h, \varphi_k) + a_\psi(u_h, \varphi_k) = \mathcal{F}(\varphi_k) \tag{11}$$

pro všechna $k = 1, 2, \dots 2N + 2$.

Zobrazení a_ψ z (11) lze zapsat jako váhový skalární součin

$$a_\psi(u_h, \varphi_k) = (\psi(u_h), \varphi_k)_r. \tag{12}$$

Všimněme si, že $\psi(u_h)$ není prvkem prostoru V_h. Proto je zapotřebí nalézt vhodnou algebraickou reprezentaci (12) a potažmo celé soustavy (11). Uvedeme dva možné přístupy vedoucí na odlišné výpočtové algoritmy.

3.1 Algoritmus založený na nehladké Newtonově metodě

Zobrazení (12) upravíme tak, abychom nehladkou Newtonovu metodu (nNM) mohli použít. Aproximujeme-li obdélníkovou kvadraturní formulí integrál

$$a_\psi(u_h, \varphi_k) = \int_{a_k}^{b_k} \psi\big(u_h(r)\big)\, \varphi_k(r)\, r\, dr \qquad \text{pro } k = 1, 2, \dots 2N+2\,,$$

kde interval $(a_k, b_k) \subset (a, b)$, $a_k, b_k \in \mathcal{N}_h$, je nosičem φ_k, dostaneme zobrazení a_ψ^h, kde

$$a_\psi^h(u_h, \varphi_k) = \begin{cases} \xi_k(\zeta_k - \xi_k)\psi\big(u_h(\zeta_k - \xi_k)\big)\varphi_k(\zeta_k - \xi_k) & \text{pro } k = 1, 2, \\ \xi_k\zeta_k\psi\big(u_h(\zeta_k)\big)\varphi_k(\zeta_k) & \text{pro } k = 3, 4, \dots 2N-1, 2N, \\ \xi_k(\zeta_k + \xi_k)\psi\big(u_h(\zeta_k + \xi_k)\big)\varphi_k(\zeta_k + \xi_k) & \text{pro } k = 2N+1, 2N+2 \end{cases}$$

pro $\zeta_k := \frac{b_k + a_k}{2}$ a $\xi_k := \frac{b_k - a_k}{2}$. Neboť ζ_k a $\zeta_k \pm \xi_k$ jsou prvky $\mathcal{N}_h$, hodnoty $u_h(\zeta_k)$, $u_h(\zeta_k + \xi_k)$ a $u_h(\zeta_k - \xi_k)$ odpovídají konkrétním složkám vektoru $u = (u_k)_{k=1}^{2N+2}$.

Navíc funkční hodnoty bázových funkcí φ_k jsou známé přesně. Proto lze a^h_ψ zjednodušit na

$$a^h_\psi(u_h, \varphi_k) = \begin{cases} \xi_1 a\psi(u_1) & \text{pro } k = 1, \\ \xi_k \zeta_k \psi(u_k) & \text{pro } k = 3, 5, 7, \ldots 2N-1, \\ \xi_{2N+1} b\psi(u_{2N+1}) & \text{pro } k = 2N+1, \\ 0 & \text{pro } k \text{ sudé}\,. \end{cases} \tag{13}$$

Ze srovnání (12) a (13) plyne, že jsme pro u_h a tedy pro vektor $u = (u_k)_{k=1}^{2N+2}$ odvodili konečně–dimenzionální aproximaci nelineárního zobrazení (3). Pro k-tou složku $\psi(u_h)_k$ totiž platí, že

$$\psi(u_h)_k \approx \sum_{i=1}^{m} k_{N_i}(u_k - L_{+i})^+ - \sum_{j=1}^{n} k_{P_j}(u_k + L_{-j})^-. \tag{14}$$

Tímto jsme pro diskrétní úlohu (11) nalezli první algebraickou reprezentaci, ve které

$$\begin{gathered} \text{hledáme } u \in \mathbb{R}^{2N+2}, \quad \text{tak, že} \\ K\ u + \sum_{i=1}^{m} B_{+i}(u - L_{+i})^+ - \sum_{j=1}^{n} B_{-j}(u + L_{-j})^- = f, \end{gathered} \tag{15}$$

kde matice tuhosti $K \in \mathbb{R}^{2N+2\times 2N+2}$ je standardně sestavená vzhledem k zobrazení a_0, vektor $f \in \mathbb{R}^{2N+2}$ odpovídá objemovým zatížením a diagonální matice B_{+i}, B_{-j} mají podle (13) prvky složené z hodnot ξ_k, ζ_k, a, b a hodnot funkcí k_{N_i} a k_{P_j}. Pro vektor $u = (u_k)_{k=1}^{2N+2}$ a konstanty $L_{+i}, L_{-j} \in \mathbb{R}^+$ chápeme výrazy $(u - L_{+i})^+$ a $(u + L_{-j})^-$ po složkách.

K řešení (15) použijeme nNM, jejíž hlavní myšlenkou je použití tak zvaných *vyhlazovacích funkcí* v Newtonovských iteračních rovnicích namísto klasických derivací, viz [6].

Definice 3.1 Nechť X a Y jsou Banachovy prostory, $\mathcal{O} \subseteq X$ je otevřená množina a $L(X, Y)$ je množinou všch omezených lineárních zobrazení z X do Y. Zobrazení $F\colon \mathcal{O} \mapsto Y$ se nazývá

- *vyhladitelně diferencovatelné v bodě* $u \in \mathcal{O}$, jestliže existuje zobrazení $F^o\colon \mathcal{O} \mapsto L(X, Y)$ tak, že třída $\{F^o(u+v),\ v \text{ dostatečně malé}\}$ je stejnoměrně omezená v operátorové normě a

  $$\lim_{v\to 0} \frac{1}{\|v\|_X} \|F(u+v) - F(v) - F^o(u+v)\, v\|_Y = 0.$$

  Zobrazení F^o se nazývá *vyhlazovací funkce (slanting function) pro* F *v bodě* u;

- *vyhladitelně diferencovatelná na* $\mathcal{O}$, jestliže existuje $F^o\colon \mathcal{O} \mapsto L(X, Y)$, které je vyhlazovací funkcí pro F ve všech bodech $u \in \mathcal{O}$. Říkáme, že F^o je *vyhlazovací funkcí pro* F *na* $\mathcal{O}$.

Věta 5 ([6]) *Nechť funkce F je vyhladitelně diferencovatelná na $\mathcal{O}$ s vyhlazovací funkcí F^o. Nechť $u^* \in \mathcal{O}$ řeší rovnici*

$$F(u) = 0.$$

Jestliže $F^o(u)$ není na $\mathcal{O}$ singulární a $\{\|F^o(u)^{-1}\| \ : \ u \in \mathcal{O}\}$ je ohraničená, potom nNM iterace

$$u^{(k+1)} = u^{(k)} - F^o(u^{(k)})^{-1} F(u^{(k)})$$

konvergují superlineárně[1] *k u^* pro dostatečně malé $\|u^{(0)} - u^*\|_X$.*

Příklad: Nechť $F(u) = u^+$, $u \in \mathbb{R}^1$. Potom vyhlazovací funkcí je multifunkce

$$F^o(u) := \begin{cases} 1 & u > 0, \\ \delta & u = 0, \ \delta \in \mathbb{R}^1 \text{ libovolné}, \\ 0 & u < 0. \end{cases}$$

Vraťme se k připravené úloze (15) a definujme zobrazení

$$F(u) := Ku + \sum_{i=1}^{m} B_{+i}(u - L_{+i})^+ - \sum_{j=1}^{n} B_{-j}(u + L_{-j})^- - f.$$

Řešení (15) splňuje rovnici $F(u) = 0$. Vyhlazovací funkcí F^o pro F je zobrazení

$$F^o(u) := K + \sum_{i=1}^{m} B_{+i} D\big(\mathcal{A}_{+i}(u)\big) - \sum_{j=1}^{n} B_{-j} D\big(\mathcal{A}_{-j}(u)\big),$$

kde diagonální matice $D\big(\mathcal{A}_{+i}(u)\big)$, $D\big(\mathcal{A}_{-j}(u)\big)$ odpovídají aktivním množinám $\mathcal{A}_{+i}(u)$, $\mathcal{A}_{-j}(u)$ definovaným

$$\mathcal{A}_{+i}(u)_k = \begin{cases} 1 & u_k > L_{+i}, \\ 0 & \text{jinak}, \end{cases} \qquad \mathcal{A}_{-j}(u)_k = \begin{cases} 1 & u_k < -L_{-j}, \\ 0 & \text{jinak} \end{cases}$$

pro $k = 1, \ldots 2N + 2$.

Rovnive nNM pro výpočet $(k+1)$-ní iterace je tedy tvaru

$$\left(K \ + \sum_{i=1}^{m} B_{+i} D\big(\mathcal{A}_{+i}(u^{(k)})\big) - \sum_{j=1}^{n} B_{-j} D\big(\mathcal{A}_{-j}(u^{(k)})\big) \right) u^{(k+1)} = f. \qquad (16)$$

Nyní tedy můžeme formulovat první výpočtový algoritmus využívající techniku aktivních množin.

[1] Superlineární konvergencí $u^{(k)}$ k u^* rozumíme, že $\lim_{k\to\infty} \frac{\|u^{(k+1)} - u^*\|_X}{\|u^{(k)} - u^*\|_X} = 0$.

Algoritmus nNM:

1. Sestav matice K, B_{+i}, B_{-j} a vektor f.
2. Volba počátečního odhadu $u^{(0)}$.
3. Pro k-tou iteraci $u^{(k)}$ proveď:
   (a) sestav aktivní množiny $\mathcal{A}_{+i}(u^{(k)})$ a $\mathcal{A}_{-j}(u^{(k)})$,
   (b) vyřeš (16),
   (c) vyhodnoť ukončovací kritérium $\frac{\|u^{(k+1)}-u^{(k)}\|}{\|u^{(k)}\|} \leq \epsilon$.

Poznámka 2 Cílem algoritmu nNM je nalézt vhodnou kombinaci „0“ a „1“ v aktivních množinách $\mathcal{A}_{+i}$ a $\mathcal{A}_{-j}$. Každá tato množina má $N+1$ prvků a proto algoritmus nNM teoreticky nejvýše po 2^{N+1} krocích skončí. A právě podmínku na maximální počet kroků můžeme použít jako další ukončovací kritérium. Zajistíme tak, že se výpočet „nezacyklí“ mezi několika málo stavy. To může například nastat vlivem chyby zaokrouhlení, kdy v některém z uzlů dělení $\mathcal{N}_h$ zdánlivě nevyhovuje ani stav „1“ ani stav „0“, tj. odpovídající iterační řešení nesplňují kritérium ukončení (tzv. zig-zag effect).

3.2 Algoritmus založený na metodě postupných aproximací

Základní metodou pro druhý algoritmus je iterační způsob k hledání pevného bodu daného zobrazení, tzv. metoda postupných aproximací (MPA).

Označme $X = \psi(V_h)$, $X \subset C^0\big((a,b)\big)$ a $\mathcal{K}^{\pm}$ množinu všech zobrazení z V_h to X takových, že

$$v_h \mapsto \sum_i k_{1,i}(v_h + L_i)^+ + \sum_j k_{2,j}(v_h + L_j)^-$$

kde $L_i, L_j \in \mathbb{R}^1$, $k_{1,i}, k_{2,j} \in L^\infty\big((a,b)\big)$ pro všechny indexy i, j.

Tvrzení 1 *Ke každému zobrazení $\psi \in \mathcal{K}^{\pm}$ existuje dvojice zobrazení ψ_0 a ℓ, kde $\psi_0 \in \mathcal{K}^{\pm}$ a ℓ je lineární zobrazení z V_h do V_h tak, že je splněna následující rovnice*

$$\psi + \psi_0 = \ell \quad \text{na } V_h.$$

Příklad: Položíme-li $\psi(v_h) = k(v_h + L)^+$, potom zobrazení ψ_0 a ℓ z Tvrzení 1 jsou tvaru

$$\psi_0(v_h) = -k(v_h + L)^- \quad \text{a} \quad \ell(v_h) = k(v_h + L).$$

Použijeme-li Tvrzení 1 pro úpravu rovnice (11), pak pro všechny indexy k platí, že

$$a_0(u_h, \varphi_k) + (\ell(u_h), \varphi_k)_r = \mathcal{F}(\varphi_k) + (\psi_0(u_h), \varphi_k)_r.$$

Označme $\mathcal{S} : V_h \to V_h$ jako operátor, který každému $w_h \in V_h$ přiřadí řešení v_h soustavy rovnic

$$a_0(v_h, \varphi_k) + (\ell(v_h), \varphi_k)_r = \mathcal{F}(\varphi_k) + (\psi_0(w_h), \varphi_k)_r \qquad \text{pro všechna } k. \tag{17}$$

Potom diskrétní řešení $u_h \in V_h$ úlohy (11) je pevným bodem operátoru $\mathcal{S}$. Druhou algebraickou formulaci (11) získáme, jestliže použijeme MPA na (17). Odpovádající iterační rovnice je ve tvaru

$$(K \ + L)u^{(n+1)} = f + f_{\psi_0}^{(n)}, \tag{18}$$

kde K a f jsou stejné jako v (11), prvky matice $L \in \mathbb{R}^{(2N+2)\times(2N+2)}$ jsou dány $L_{lk} = (\ell(\varphi_l), \varphi_k)_r$ a vektoru $f_{\psi_0}^{(n)} \in \mathbb{R}^{2N+2}$ jsou $f_{\psi_0,k}^{(n)} = (\psi_0(u^{(n)}), \varphi_k)_r$ pro všechny indexy $k, l = 1, \ldots, 2N + 2$.

Algoritmus MPA:

1. Sestav matice K, L a vektor f.
2. Volba počátečního odhadu $u^{(0)}$.
3. Pro k-tou iteraci $u^{(k)}$ proveď:
   (a) sestav vektor $f_{\psi_0}^{(n)}$,
   (b) vyřeš (18),
   (c) vyhodnoť ukončovací kritérium (stejné jako v algoritmu nNM).

Poznámka 3 Matice $K + L$ z rovnice (18) je pozitivně definitní. Tato vlastnost zaručuje existenci iteračního řešení pro libovolný počáteční odhad $u^{(0)}$ oproti metodě nNM, viz věta 5.

Konvergenci algoritmu MPA budeme testovat numericky a výsledná řešení porovnáme s výsledky získané metodou nNM.

4 Numerické experimenty

Uvedeme dva příklady. V prvním předepisujeme stabilní okrajové podmínky s nulovým posunutím (`w(a)=w(b)=0`), které samy o sobě zajišťují existenci i jednoznačnost řešení. Youngův modul a Poissonovo číslo (`E = 2.140e+011` a `sigma = 2.900e-001`) odpovídají oceli, což je materiál, jehož tvarové vlastnosti odpovídají lineární teorii elasticity.

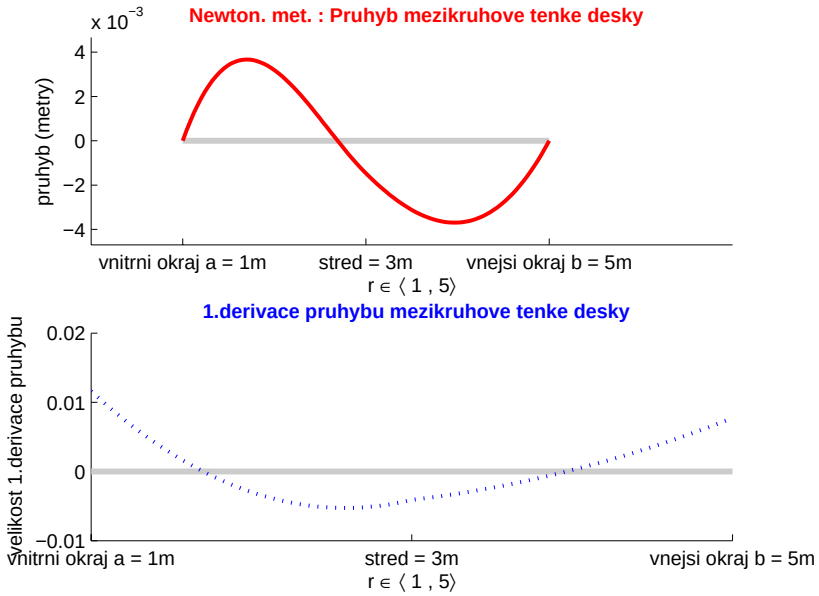
I. Vstupní data první úlohy

```
      Rozmery desky
           tloustka               h = 1.000e-002 metru
           vnitrni polomer        a = 1.000e+000 metru
           vnejsi polomer         b = 5.000e+000 metru
      Ciselne charakteristiky materialu
           Younguv modul          E = 2.140e+011 N/m^2
           Poissonovo cislo    sigma = 2.900e-001
      Zatizeni ve smeru osy z~je
                          na celem intervalu ( 1 , 5 )
                                f(r) = 0  N/m^2
      Deska je umistena v~pruznem prostredi, ktere se sklada
z~-  nadlozi po cele delce s~odporem k_n = 5.000e+004 N/m^2
  Okrajove podminky
  --->   pokles podpor   v~r=a o~velikosti  w(a) = 0.000e+000 metru
  --->   pokles podpor   v~r=b o~velikosti  w(b) = 0.000e+000 metru
  --->   moment        pro r=a ve tvaru    Mw(a) = 5.500e+004
  --->   moment        pro r=b ve tvaru    Mw(b) = 9.000e+004
```

V následující tabulce je uvedeno porovnání výsledků získaných metodami nNM a MPA pro měnící se počet dělení N.

	počet iterací		velikost rel. rezidua		maximální
N	nNM	MPA	nNM	MPA	odchylka výsledků
2	1	4	5.27154e-017	1.64098e-004	1.577543e-005
10	2	4	1.19103e-017	1.15469e-007	5.089773e-007
20	2	4	1.55926e-017	2.74973e-009	6.666227e-008
50	2	4	2.18885e-017	9.40658e-012	8.322865e-009
200	2	4	3.04560e-017	3.75928e-014	1.112272e-009

Druhý příklad počítá s deskou stejných rozměrů, ale materiálové konstanty (`E = 1.000e+007` a `sigma = 4.999e-001`) jsou zvoleny tak, že výsledné chování je poměrně elastičtější než u oceli. Například be se mohlo jednat o pryž. Ačkoli reálné chování těles vyrobených z tohoto materiálu nepopisuje zvolená lineární teorie,



Obrázek 4: Výsledný průhyb odpovídající zadání první úlohy.

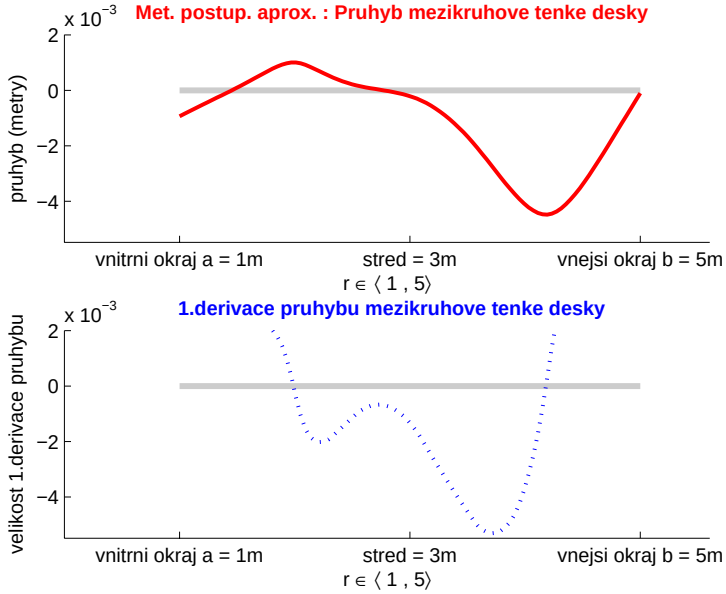
uvádíme jej pouze z důvodu „viditelnosti“ výsledného průhybu v grafu. Okrajové podmínky předepisujeme nestabilní (`Tw(a)=Mw(a)=Tw(b)=Mw(b)=0`), tedy homogenní tvar (2).

II. Vstupní data druhé úlohy

```
      Rozmery desky
           tloustka                h = 1.000e-002 metru
           vnitrni polomer         a = 1.000e+000 metru
           vnejsi polomer          b = 5.000e+000 metru
      Ciselne charakteristiky materialu
           Younguv modul           E = 1.000e+007 N/m^2
           Poissonovo cislo    sigma = 4.999e-001
      Zatizeni ve smeru osy z~je
                        v~bode r1 = 2
                                f1(r) = 30   N/m^2
                        v~bode r2 = 4.2
                                f2(r) = -20  N/m^2
      Deska je umistena v~pruznem prostredi, ktere se sklada
z~-  nadlozi po cele delce s~odporem k_n = 5.000e+004 N/m^2
  Okrajove podminky
  --->   pricna sila     pro r=a ve tvaru       Tw(a) = 0
  --->   moment          pro r=a ve tvaru       Mw(a) = 0
  --->   pricna sila     pro r=b ve tvaru       Tw(b) = 0
  --->   moment          pro r=b ve tvaru       Mw(b) = 0
  Neekvidistantni adaptivni deleni na 20 casti.
```

Následuje postupný výčet iterací pro MPA. V případě metody nNM došlo k divergenci vlivem nedostatečně blízké počáteční aproximace řešení.

```
Postupne Aproximace : Iteracni reseni vysledne soustavy.
 ------------------------------------------------------------
  | iterace  |       rel.chyba       |   el. rezidua     |
 ------------------------------------------------------------
  |       1  |      5.48506e-001     |   1.04308e-005    |
  |       2  |      3.79994e-001     |   6.62548e-006    |
  |       3  |      2.76282e-001     |   4.82157e-006    |
  |       4  |      2.13875e-001     |   3.78108e-006    |
  |       5  |      1.73273e-001     |   3.10336e-006    |
  |       6  |      1.44979e-001     |   2.62575e-006    |
  |       7  |      1.24082e-001     |   2.26828e-006    |
  |       8  |      1.08368e-001     |   2.02529e-006    |
  |       9  |      9.61472e-002     |   1.83887e-006    |
  |      10  |      8.63444e-002     |   1.68602e-006    |
 ------------------------------------------------------------
Pro rovnici A*w^(n+1) = f + B(k_p*w^(n)[+]-k_n*w^(n)[-]) jsou vzorce
 rel.chyba = |w - w0| / |w0|
 rel.reziduum =
    |A*w+Bplus*w[+]-Bminus*w[-] - f| / ((|A|+|Bplus|+|Bminus|)*|w|+|f|)
 pouzita je sloupcova norma |w| := max(abs(w)) a |A| := max(sum(abs(A')))
 matice Bplus a Bminus jsou sestaveny podle aktivni mnoziny reseni
 (viz Newt. metoda)
```



Obrázek 5: Výsledný průhyb odpovídající zadání druhé úlohy.

Literatura

[1] Ciarlet, P. G.: *Mathematical Elasticity, Volume II: Theory of Plates.* Elsevier, Amsterdam, 1997

[2] Horák, J., Svobodová, I.: *Modelování mezikruhové desky s podložím.* In: 13th confer. Modern Mathematical Methods in Engineering, Dolní Lomná, 2004.

[3] Kufner, A.: *Weighted Sobolev Spaces.* Teubner, Leipzig, 1980.

[4] Gallagher, R. H.: *Finite Element Analysis: Fundamentals.* Prentice Hall, Englewood Cliffs, New Jersey, 1975.

[5] Salač, P.: *Optimal Design of an Elastic Circular Plate na a Unilateral Elastic Foundation. II: Approximate Problems.* ZAMM · Z. Angew. Math. Mech. **82**, 1 (2002), 33–42.

[6] Chen, X., Nashed, Z. Qi, L.: *Smoothing Methods a Semismooth Methods pro Nondifferentiable Operator Equations.* SIAM Journal na Numerical Analysis **38**, 4 (2001), 1200–1216.

[7] Nečas, J., Hlaváček, I.: *Mathematical Theory of Elastic and Elastico-Plastic Bodies: An Introduction.* Elsevier, Amsterdam–Oxford–New York, 1981.

[8] Kufner, A., Fučík, S.: *Nelineární diferenciální rovnice.* NTL, Praha, 1978.

[9] Aubin, J.-P.: *Applied functional analysis.* Wiley, New York, 1999.


Univ. Palacki. Olomuc., Fac. rer. nat.,
Dept of Math. Anal. and Appl. of Math.
ODAM (2009) 68–84

Optimalizace nosníku na jednostranném podloží: Existence řešení

Roman ŠIMEČEK

Katedra matematické analýzy a aplikací matematiky
Přírodovědecká fakulta, Univerzita Palackého,
Tř. 17. listopadu 1192/12, 771 46 Olomouc, Česká republika
e-mail: simecekr@seznam.cz

Abstrakt

V příspěvku se zabýváme úlohou optimalizace nosníku spočívajícího na jednostranném podloží. Matematický model nosníku je Euler–Bernoulliův a model podloží je Winklerův. Stavovou úlohu představuje nelineární semikoercivní problém čtvrtého řádu se smíšenými okrajovými podmínkami. Předmětem optimalizace je tloušťka nosníku a koeficient tuhosti podloží. Cílem práce je formulovat podmínky existence řešení stavové úlohy a dokázat existenci řešení úlohy optimalizace.

1 Úvod

V této práci se budeme zabývat analýzou řešitelnosti problému optimalizace nosníku spočívajícího na jednostranném podloží. Jedná se o úlohu, která se vyskytuje v praktických inženýrských a mechanických aplikacích. Vzpomenout můžeme například železniční dopravu, stavebnictví a další.

Jako matematicko-fyzikální model nosníku je zvolen Euler–Bernoulliův. Tento model vychází z matematické teorie pružnosti a za předpokladu splnění všech jeho podmínek (rozměry nosníku, orientace zatížení a další) je reprezentován obyčejnou diferenciální rovnicí 4. řádu s okrajovými podmínkami. Využívá se dimenzionální redukce a celkově jde o 1-D úlohu (viz [11]). Pro modelování kontaktu

nosníku s podložím nebudeme jako v klasických kontaktních úlohách uvažovat podloží jako další samostatné elastické těleso. Vliv podloží do modelu zahrneme přidáním tzv. odezvové funkce závislé na koeficientu tuhosti podloží $q(x)$, průhybu $u(x)$ a případně jeho derivacích. V našem případě budeme uvažovat jednoparametrické jednostranné podloží Winklerova typu. Jeho odezvová funkce má tvar $q(x)u^+(x)$ (viz [8],[13]). Z literatury je obecně známa varianta oboustranného podloží s odezvovou funkcí $q(x)u(x)$, které má nespornou výhodu, že výsledný matematický model je lineární (viz [8],[12]). Často však tento model ne úplně dobře aproximuje reálnou situaci a to zejména pokud je nosník na podloží pouze položen a není s ním pevně spojen. V takovém případě je nutné použít variantu jednostranného podloží a výsledný matematický model se tak stává nelineární. Lze také použít dvouparametrický model podloží. Nejznámější variantou je dvouparametrický Pasternakův model podloží s odezvovou funkcí $q(x)u(x) - k(x)u''(x)$, kde druhý parametr $k(x)$ souvisí s účinkem příčných (smykových) sil v podloží. Speciálním případem je pak dokonale tuhé podloží (překážka) a výsledný model pak vede na variační nerovnici 1. druhu (viz [5],[14],[8]).

Vlivem zvolených okrajových podmínek se stavová úloha stává semi-koercivní. K zaručení existence a jednoznačnosti jejího řešení je třeba formulovat dodatečné podmínky na zatížení. Tím vyloučíme přípustná tuhá posunutí s nulovou energií a v podstatě tak zaručíme existenci nenulové kontaktní zóny mezi nosníkem a podložím.

Předmětem optimalizace bude tloušťka nosníku $t(x)$ a také koeficient tuhosti podloží $q(x)$. Z literatury je znám případ optimalizace tloušťky nosníku (viz [4],[5]). V této práci přidáváme do optimalizace ještě koeficient tuhosti podloží, což může být přínosné zejména v případech, kdy je možno tuhost podloží nějakým způsobem ovlivňovat.

Základním krokem v důkazu existence řešení optimalizační úlohy je omezenost řešení stavové úlohy vzhledem k návrhové proměnné. V koercivních úlohách je tato vlastnost zajištěna Fridrichsovou nebo zobecněnou Fridrichsovou nerovností na prostoru kinematicky přípustných posunutí. V semi-koercivním případě však nemůžeme tyto nerovnosti přímo použít. Nejprve provede rozklad prostoru přípustných posunutí na vhodné podprostory a na nich pak užijeme nerovnosti Poincarého typu.

2 Některé pomocné věty

V této kapitole uvedeme některé výsledky z oblasti prostorů funkcí, minimalizace konvexních funkcionálů a konvexní analýzy. V této práci používáme Lebesqueovy $L^p(\Omega), p = 1, 2, \infty$, Sobolevovy prostory $H^k(\Omega), k = 1, 2$, a prostory spojitých funkcí $C^k(\bar{\Omega}), k = 1, 2, 3, 4$, kde Ω je otevřený, omezený a neprázdný interval v $\mathbb{R}^1$. Normu prostoru $L^2(\Omega)$ budeme značit $||\cdot||_{2,\Omega}$. Standartní normu prostoru $H^2(\Omega)$ budeme značit $||\cdot||_{k,2,\Omega}$ a seminormy $|\cdot|_{k,2,\Omega},\ k = 0, 1, 2$. V pří-

padě, že bude zřejmé o jaký interval se jedná, budeme symbol Ω vynechávat. Dále budeme používat prostory polynomů k-tého stupně označené jako $\mathbf{P}_k$. Bližší informace o prostorech funkcí najdeme například v [10]. V následujících lemmatech uvedeme některé vlastnosti kladné části funkce z $H^2(\Omega)$.

Lemma 1 *Nechť $u \in H^2(\Omega)$, pak platí $u^+ \in H^1(\Omega)$,*

$$\left|u^+(x) - v^+(x)\right| \leq |u(x) - v(x)| \ \forall x \in \Omega, \ \forall u, v \in C(\bar{\Omega}).$$

Navíc platí, že $\|u^+\|_{1,2,\Omega} \leq \|u\|_{1,2,\Omega} \ \forall v \in H^2(\Omega)$.

Důkaz předchozího lemmatu nalezneme v [13]. Následující pomocné věty uvedeme bez důkazu.

Lemma 2 *Nechť $u_n, u \in H^k(\Omega), \ \Omega \subset \mathbb{R}^n$. Dále nechť $u_n \rightharpoonup u$ v $H^k(\Omega)$, pak platí $u_n^+ \rightharpoonup u^+$ v $H^k(\Omega)$.*

Lemma 3 *Nechť $s, t \in \mathbb{R}$, pak platí*

$$\lim_{\epsilon \to 0} \frac{[(s + \epsilon t)^+]^2 - [s^+]}{\epsilon} = 2s^+ t.$$

Pro účely důkazu jednoznačné řešitelnosti stavové úlohy dokážeme modifikovanou variantu známé Poincarého nerovnosti (viz [10]).

Věta 1 (Nerovnost Poincarého typu) *Nechť Ω je neprázdný interval v $\mathbb{R}^1$. Dále nechť je definován prostor $\mathbb{V} = \{v \in H^2(\Omega) : v'(0) = 0\}$, pak existují konstanty k_1, k_2 závislé na intervalu Ω tak, že platí*

$$\|v(x)\|^2_{2,2,\Omega} \leq k_1 |v(x)|^2_{2,2,\Omega} + k_2 \left(v(x), 1\right)^2_{2,\Omega} \quad \forall v \in \mathbb{V}. \tag{1}$$

Důkaz Pro všechny funkce z prostoru $H^1(\Omega)$ platí dobře známá Poincarého nerovnost.

$$||v(x)||^2_{2,\Omega} \leq c_1 |v(x)|^2_{1,2,\Omega} + c_2 \left(v(x), 1\right)^2_{2,\Omega} \quad \forall v \in H^1(\Omega).$$

K oběma stranám nerovnosti přičteme výraz $|v(x)|^2_{1,2,\Omega}$. Dostáváme

$$||v(x)||^2_{2,\Omega} + |v(x)|^2_{1,2,\Omega} \leq (c_1 + 1)|v(x)|^2_{1,2,\Omega} + c_2 \left(v(x), 1\right)^2_{2,\Omega} \quad \forall v \in H^1(\Omega).$$

Pro funkce z $\mathbb{V} = \{v \in H^2(\Omega) : v'(0) = 0\}$ platí následující nerovnost

$$|v(x)|^2_{1,2,\Omega} \leq c_3 |v(x)|^2_{2,2,\Omega} + c_4 \left(v'(0)\right)^2 \quad \forall v \in \mathbb{V}.$$

Pak tedy platí

$$||v(x)||^2_{2,\Omega} + |v(x)|^2_{1,2,\Omega} \leq c_5 |v(x)|^2_{2,2,\Omega} + c_2 \left(v(x), 1\right)^2_{2,\Omega} \quad \forall v \in \mathbb{V}.$$

Přičtením výrazu $|v(x)|^2_{2,2,\Omega}$ k oběma stranám nerovnice dostáváme požadovanou nerovnost. □

Nyní uvedeme dvě varianty dobře známé základní věty variačního počtu. Důkazy nalezneme například v [3].

Věta 2 *Nechť V je reflexivní Banachův prostor a $\mathcal{J}\colon V \to \mathbb{R}$ je slabě zdola polospojitý funkcionál na V a nechť platí*

$$\lim_{\substack{\|u\|_V \to \infty \\ u \in U}} \mathcal{J}(u) = +\infty, \tag{2}$$

Potom existuje alespoň jeden bod $u^* \in V$ takový, že

$$\mathcal{J}(u^*) = \inf_{u \in V} \mathcal{J}(u). \tag{3}$$

Je-li $\mathcal{J}$ navíc ryze konvexní na V, pak je toto řešení jednoznačné.

Věta 3 *Nechť V je reflexivní Banachův prostor, $U \subseteq V$ je neprázdná, konvexní a uzavřená podmnožina prostoru V, $\mathcal{J} : V \to \mathbb{R}$ je slabě zdola polospojitý funkcionál na U a nechť platí*

$$\lim_{\substack{\|u\|_V \to \infty \\ u \in U}} \mathcal{J}(u) = +\infty, \tag{4}$$

Potom existuje alespoň jeden bod $u^* \in U$ takový, že

$$\mathcal{J}(u^*) = \inf_{u \in U} \mathcal{J}(u). \tag{5}$$

Je-li $\mathcal{J}$ navíc ryze konvexní na U, pak je toto řešení jednoznačné.

V dalším uvedeme jedno z kriterií konvexity G-diferencovatelného funkcionálu. Důkaz nalezneme v [2].

Lemma 4 *Nechť $\mathcal{J}$ je funkcionál definovaný na konvexní otevřené podmnožině M normovaného lineárního prostoru X. Potom jestliže je $\mathcal{J}$ G-diferencovatelný na M a platí*

$$D\mathcal{J}(u; u - v) - D\mathcal{J}(v; u - v) \geq 0 \quad \forall u, v \in M,$$

pak je $\mathcal{J}$ konvexní na M.

Na závěr prezentujeme dvě věty z konvexní analýzy na jejichž základě provedeme rozklad prostoru přípustných posunutí na kužel tuhých posunutí a jeho negativní polární kužel (viz [1]).

Věta 4 (O nejlepší aproximaci) *Nechť $\mathbf{H}$ je Hilbertův prostor. Dále nechť $\mathcal{C} \subset \mathbf{H}$ je uzavřená konvexní podmnožina. Potom $\forall x \in \mathbf{H}$ $\exists!$ $y \in \mathcal{C}$ tak, že platí*

$$\begin{gathered} \|x - y\|_{\mathbf{H}} = \min_{z \in \mathcal{C}} \|x - z\|_{\mathbf{H}} \\ \Updownarrow \\ (x - y, x - z)_{\mathbf{H}} \leq 0 \quad \forall z \in \mathcal{C}. \end{gathered}$$

Prvek $y = P_{\mathcal{C}}(x)$ se nazývá nejlepší aproximací prvku x na $\mathcal{C}$ a $P_{\mathcal{C}} : \mathbf{H} \to \mathcal{C}$ je projektor nejlepší aproximace na podmnožinu $\mathcal{C}$.

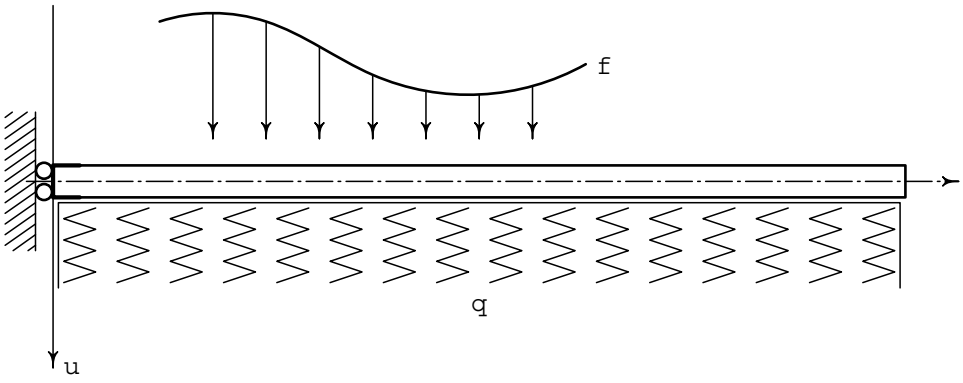
Věta 5 *Nechť* $\mathbf{H}$ *je Hilbertův prostor. Dále nechť* $\mathcal{C} \subset \mathbf{H}$ *je uzavřený konvexní kužel s vrcholem v* 0*. Potom* $\forall x \in \mathbf{H}\ \exists!\ \{y, z\} \in \mathcal{C} \times \mathcal{C}^{\ominus}$ *tak, že platí*

$$x = y \oplus z, \quad (y, z)_{\mathbf{H}} = 0.$$

Prvky $y = P_{\mathcal{C}}(x)$, $z = P_{\mathcal{C}^{\ominus}}(x)$ *jsou nejlepší aproximace prvku* x *na* $\mathcal{C}$,$\mathcal{C}^{\ominus}$*. Zobrazení* $P_{\mathcal{C}}, P_{\mathcal{C}^{\ominus}}$ *jsou operátory nejlepší aproximace na kuželu* $\mathcal{C}$ *a jeho negativním polárním kuželu* $\mathcal{C}^{\ominus}$*.*

3 Matematický model stavové úlohy

Budeme uvažovat elastický nosník délky l, který je umístěn v intervalu $\Omega = (0, l)$. Na levém konci je upevněn tak, že se může pohybovat ve vertikálním směru, ale nemůže se natáčet. V bodě $x = 0$ je tedy předepsána stabilní okrajová podmínka $u'(0) = 0$. Nosník po celé své délce spočívá na jednostranném elastickém podloží Winklerova typu. Podloží je aktivní pouze v případě že se nosník prohýbá směrem do podloží. Použijeme Euler-Bernoulliův model ohybu nosníku, který je za určitých předpokladů (na rozměry nosníku, orientace zatížení a další) důsledkem dimenzionální redukce v obecném problému elasticity, viz [11]. Výsledný model je pak jednodimenzionální. Vliv podloží zahrneme do modelu užitím tzv. odezvové funkce závislé na tuhosti podloží $q(x)$ a průhybu $u(x)$, popřípadě jeho derivacích. Příčný průřez nosníku je obdélníkový a tloušťka nosníku je daná funkcí $t(x)$.



Obrázek 1: Schéma nosníku

Klasická formulace úlohy ohybu nosníku má podobu nelineární obyčejné diferenciální rovnice 4.řádu se smíšenými okrajovými podmínkami:

Nalézt funkci $u(x) \in C^4(\Omega) \cap C^3(\bar{\Omega})$ tak, aby

$$\begin{cases} (\beta(x)t^3(x)u''(x))'' + q(x)u^+(x) = f(x) & \forall x \in \Omega \\ u'(0) = u''(0) = u''(l) = u'''(0) = 0, \end{cases}$$

kde t, q a f jsou funkce představující tloušťku nosníku, koeficient tuhosti podloží a intezitu zatížení. Funkce u reprezentuje hledaný průhyb nosníku a u^+ je její

kladná část

$$u^+(x) = \frac{u(x) + |u(x)|}{2}, \ x \in \Omega.$$

Funkce β má následující podobu

$$\beta(x) = \frac{8}{12} b(x) E(x),$$

kde E je Youngův modul pružnosti materiálu a b je funkce reprezentující šířku nosníku.

4 Množina U_{ad}

Předmětem optimalizace bude tloušťka nosníku $t(x)$ a tuhost podloží $q(x)$. Narozdíl od klasické úlohy tvarové optimalizace se optimalizační proměnné v tomto případě objevují jako koeficienty diferenciálního operátoru stavové rovnice, zatímco oblast, přes kterou integrujeme, zůstává neměnná. Množinu přípustných návrhových proměnných pro tloušťku $t(x)$ definujeme jako

$$U_{ad}^t = \Bigg\{ t \in C(\Omega) \ : \ 0 < T_0 \le t(x) \le T_1 \ \ \forall x \in \Omega, \\ \int_\Omega t(x) \, \mathrm{d}x = T_2, \ \ |t'(x)| \le T_3 \ \ \forall x \in \Omega \Bigg\}.$$

Předpokládáme, že tloušťku nosníku lze vyjádřit spojitou ohraničenou funkcí. Konstanty T_0, T_1, T_2 a T_3 je nutno volit tak, aby byla množina U_{ad}^t neprázdná. Konvergenci v množině U_{ad}^t zavedeme jako stejnoměrnou konvergenci spojitých funkcí v intervalu Ω.

$$\left[\ t_n \longrightarrow t \text{ v } U_{ad}^t \ \right] \Longleftrightarrow \left[\ t_n(x) \rightrightarrows t(x) \text{ v } \Omega \ \right]. \tag{6}$$

Funkce z U_{ad}^t jsou stejně omezené a díky podmínce $|t'(x)| \le T_3$ i stejnoměrně spojité na Ω. Potom podle Arzela - Ascoliho věty (viz [5], [10]) je U_{ad}^t se zavedenou konvergencí kompaktní podmnožinou v $C(\Omega)$.

Obdobným způsobem zavedeme množinu přípustných návrhových proměnných pro tuhost podloží $q(x)$.

$$U_{ad}^q = \Bigg\{ q \in L^2(\Omega) \ : \ Q_0 \le q(x) \le Q_1 \ \ \forall x \in \Omega \Bigg\}.$$

Optimální tuhost podloží budeme vybírat z funkcí, které jsou ohraničené v intevalu Ω. Konstanty Q_0, Q_1 volíme tak, že platí $U_{ad}^q \neq \emptyset$. Konvergenci v U_{ad}^q definujeme jako slabou konvergenci v Lebesgueově prostoru $L^2(\Omega)$.

$$\left[\ q_n \longrightarrow q \text{ v } U_{ad}^q \ \right] \Longleftrightarrow \left[\ q_n(x) \rightharpoonup q(x) \text{ v } L^2(\Omega) \ \right]. \tag{7}$$

Z uvedených předpokladů plyne užitím Eberlajn-Šmuljanovy věty (viz [3]), že U_{ad}^q je slabě kompaktní podmožinou v $L^2(\Omega)$. Pro praktické výpočty můžeme uvažovat navíc podobné podmínky ve tvaru integrálu a omezení derivace jako u tloušťky. Nicméně pro matematickou analýzu problému stačí pouze podmínka uvedená v definici U_{ad}^q.

Výsledná množina všech přípustných návrhových proměnných má tvar kartézského součinu množin U_{ad}^t a U_{ad}^q. Uspořádané dvojice prvků $[t, q]$ z U_{ad} budeme souhrnně označovat jako $\mathbf{e}$.

$$U_{ad} = U_{ad}^t \times U_{ad}^q, \ \left[\ \mathbf{e} \in U_{ad} \ \right] \Longleftrightarrow \left[\ \mathbf{e} = [t, q], \ t \in U_{ad}^t \ \wedge \ t \in U_{ad}^q \ \right].$$

Konvergenci v množině U_{ad} definujeme v souladu s konvergencemi pro jednotlivé množiny U_{ad}^t a U_{ad}^q.

$$\left[\ \mathbf{e}_n \longrightarrow \mathbf{e} \text{ v } U_{ad} \ \right] \Longleftrightarrow \left[\ (t_n \longrightarrow t \text{ v } U_{ad}^t) \ \wedge \ (q_n \longrightarrow q \text{ v } U_{ad}^q) \ \right].$$

Z kompaktnosti množin U_{ad}^t a U_{ad}^q a z vlastností kartézského součinu plyne i kompaktnost množiny U_{ad} vzhledem k právě definované konvergenci. Kompaktnost množiny přípustných návrhových proměnných je jednou ze základních podmínek v důkazu existence řešení úloh tvarové optimalizace.

5 Variační formulace stavové úlohy

Nyní obraťme svou pozornost zpět ke stavové úloze. Klasickou formulaci stavové úlohy můžeme použít pouze za předpokladu dostatečné hladkosti vstupních dat. V praktických aplikacích však takovou hladkost v mnoha případech nemůžeme předem zaručit. Definujme tedy tzv. variační formulaci, která vychází z principu minima potenciální energie a umožní nám zobecnit předpoklady na hladkost vstupních dat a jako prostor přípustných posunutí uvažovat podprostor Sobolevova prostoru $H^2(\Omega)$. Nechť $\mathbf{e} \in U_{ad}$ je libovolná pevně zvolená uspořádaná dvojice z U_{ad}, $\beta \in L^\infty(\Omega)$ a existuje konstanta β_0 tak, že $0 < \beta_0 \leq \beta(x)$ s.v. v Ω. Dále definujme prostor přípustných kinematických posunutí zachycující stabilní okrajovou podmínku:

$$\mathbb{V} = \{v \in H^2(\Omega) : v'(0) = 0\}.$$

Variační formulace stavové úlohy má pro pevné $\mathbf{e} \in U_{ad}$ následující tvar:

$$\begin{cases} \text{Nalézt } u^* \in \mathbb{V} \text{ tak, že} \\ E_{\mathbf{e}}(u^*) \leq E_{\mathbf{e}}(v) \ \ \forall v \in \mathbb{V}, \end{cases} \qquad (P(\mathbf{e}))$$

kde $E_{\mathbf{e}}$ je funkcionál celkové potenciální energie nosníku, daný předpisem

$$\begin{aligned} E_{\mathbf{e}}(v) &:= \frac{1}{2}\left(a_{\mathbf{e}}(v,v) + b_{\mathbf{e}}(v^+,v^+)\right) - F(v), \\ a_{\mathbf{e}}(u,v) &:= \int_\Omega \beta(x)t^3(x)u''(x)v''(x)\,\mathrm{d}x, \\ b_{\mathbf{e}}(u,v) &:= \int_\Omega q(x)u(x)v(x)\,\mathrm{d}x, \\ F(v) &:= \int_\Omega f(x)v(x)\,\mathrm{d}x - \sum_{i,j}(F_i v(x_i) - M_j v'(x_j)). \end{aligned}$$

Formy $a_{\mathbf{e}} : H^2 \times H^2 \to \mathbb{R}$ a $b_{\mathbf{e}} : H^1 \times H^1 \to \mathbb{R}$ reprezentují práci vnitřních sil resp. práci podloží. Hodnoty F_i, M_j představují zobecněné síly v bodech $x_i, x_j \in \bar{\Omega}$.

Z lemmatu 1 plyne, že $v^+ \in H^1(\Omega)$ a funkcionál $E_{\mathbf{e}}$ je dobře definován. V dalším blíže ukážeme vlastnosti forem $a_{\mathbf{e}}, b_{\mathbf{e}}$ a také funkcionálu $E_{\mathbf{e}}$. První důležitou vlastností je *omezenost*.

$$\begin{aligned} |a_{\mathbf{e}}(u,v)| &= \left|\int_\Omega \beta t^3 u''v''\,\mathrm{d}x\right| \le \int_\Omega |\beta t^3 u''v''|\,\mathrm{d}x \le \int_\Omega |\beta||t^3||u''||v''|\,\mathrm{d}x \le \\ &\le MT_1^3(|u''|,|v''|)_{2,\Omega} \le MT_1^3 \|u\|_2 \|v\|_2 \le \\ &\le MT_1^3 \|u\|_{2,2} \|v\|_{2,2} \qquad \forall u,v \in H^2(\Omega),\ \forall \mathbf{e} \in U_{ad}. \end{aligned}$$

$$\begin{aligned} \left|b_{\mathbf{e}}(u^+,v)\right| &= \left|\int_\Omega qu^+v\,\mathrm{d}x\right| \le \int_\Omega |qu^+v|\,\mathrm{d}x \le \int_\Omega |q||u^+||v|\,\mathrm{d}x \le \\ &\le Q_1(|u^+|,|v|)_{2,\Omega} \le Q_1||u^+||_2 \|v\|_2 \le \\ &\le Q_1 \|u\|_{2,2} \|v\|_{2,2} \qquad \forall u,v \in H^2(\Omega),\ \forall \mathbf{e} \in U_{ad}. \end{aligned}$$

V důkazu omezenosti jsme využili vlastností množiny U_{ad}, faktu $\beta \in L^\infty(\Omega)$ a Cauchy-Schwarzovy nerovnosti (viz [10]). Je zřejmé, že formy $a_{\mathbf{e}}$ a $b_{\mathbf{e}}$ jsou bilineární. Opět užitím Cauchy-Schwarzovy nerovnosti a vlastností derivace dokážeme, že F je spojitý lineární funkcionál.

6 Existence řešení stavové úlohy

V dalším se již budeme věnovat důkazu existence řešení stavové úlohy $(P(\mathbf{e}))$. Využijeme přitom základní věty variačního počtu (věty 2 a 3). V těchto větách vystupuje slabá polospojitost zdola daného funkcionálu jako jeden ze základních předpokladů. Postačující podmínkou slabé polospojitosti zdola funkcionálu je jeho *konvexita* a *G-diferencovatelnost*. V dalším ukážeme, že funkcionál $E_{\mathbf{e}}$

splňuje tyto dvě vlastnosti. Nejprve se zaměříme na G-diferencovatelnost. Při výpočtu derivace funkcionálu $E_{\mathbf{e}}$ využijeme lemma 3. Gateaux derivace formy $a_{\mathbf{e}}$ a funkcionálu F jsou z literatury obecně známé, zaměříme se tedy na G-diferencovatelnost formy $b_{\mathbf{e}}$.

$$\begin{aligned}
&\frac{1}{2}\lim_{\epsilon\to 0}\frac{b_{\mathbf{e}}((u+\epsilon v)^+,(u+\epsilon v)^+)-b_{\mathbf{e}}(u^+,u^+)}{\epsilon}=\\
&=\frac{1}{2}\lim_{\epsilon\to 0}\frac{\int\limits_\Omega q(x)[(u+\epsilon v)^+]^2\,\mathrm{d}x-\int\limits_\Omega q(x)[u^+]^2\,\mathrm{d}x}{\epsilon}=\\
&=\frac{1}{2}\int\limits_\Omega q(x)\lim_{\epsilon\to 0}\frac{[(u+\epsilon v)^+]^2-[u^+]^2}{\epsilon}\,\mathrm{d}x=\\
&=\frac{1}{2}\int\limits_\Omega q(x)2u^+v\,\mathrm{d}x=b_{\mathbf{e}}(u^+,v)\quad\forall u,v\in H^2(\Omega),\ \forall\mathbf{e}\in U_{ad}.
\end{aligned}$$

Funkcionál $DE_{\mathbf{e}}(u;\cdot)$ je spojitý a lineární na $H^2(\Omega)$ pro libovolné $u\in H^2(\Omega)$ a platí

$$DE_{\mathbf{e}}(u;v)=a_{\mathbf{e}}(u,v)+b_{\mathbf{e}}(u^+,v)-F(v)\quad\forall u,v\in H^2(\Omega),\ \forall\mathbf{e}\in U_{ad}.$$

Potom můžeme říci, že $E_{\mathbf{e}}$ je Gateauxovsky diferencovatelný na $H^2(\Omega)$. Této vlastnosti využijeme i pro důkaz jeho konvexity. Postupovat budeme v souladu s lemmatem 4, které je zobecněním lemmatu z klasické analýzy říkajícího, že pokud je derivace reálné funkce neklesající, pak je funkce konvexní. V našem případě platí

$$\begin{aligned}
&DE_{\mathbf{e}}(u;u-v)-DE_{\mathbf{e}}(v;u-v)=\\
&\quad=a_{\mathbf{e}}(u-v,u-v)+b_{\mathbf{e}}(u^+-v^+,u-v)=\\
&\quad=\int\limits_\Omega\beta t^3(u''-v'')^2\,\mathrm{d}x+\int\limits_\Omega q(u^+-v^+)(u-v)\,\mathrm{d}x\geq\\
&\quad=\int\limits_\Omega\beta t^3(u''-v'')^2\,\mathrm{d}x+\int\limits_\Omega q(u^+-v^+)^2\,\mathrm{d}x\geq\\
&\quad\geq\beta_0T_0^3\,|u-v|_{2,2}^2+Q_0||u^+-v^+||_2^2\geq 0\quad\forall u,v\in H^2(\Omega),\ \forall\mathbf{e}\in U_{ad}.
\end{aligned}$$

Podle lemmatu 4 je tak dokázána konvexita funkcionálu $E_{\mathbf{e}}$, která se ve spojení s G-diferencovatelností stává postačující podmínkou slabé polospojitosti zdola na $H^2(\Omega)$ (viz [3]).

V dalším se zaměříme na poslední důležitou vlastnost funkcionálu $E_{\mathbf{e}}$, *koercivitu*. Stavová úloha obsahuje malá tuhá posunutí, která jsou přípustná a podloží je pro ně neaktivní. Pro taková posunutí se koná nulová práce. Abychom z úlohy

tato tuhá posunutí vyloučili, zavedeme nutné a postačující podmínky koercivity resp. existence minima a provedeme vhodný rozklad prostoru $\mathbb{V}$.

Obecně je třeba stavovou úlohu zjednodušit eliminováním všech tuhých posunutí $\mathcal{R} \equiv \mathbf{P}_1$. V našem případě označme $\mathcal{K}$ konvexní kužel přípustných tuhých posunutí definovaný jako

$$\mathcal{K} = \mathcal{R} \cap \mathbb{V} \equiv \mathbf{P}_0.$$

Dále definujeme jeho podkužel $\mathcal{K}_\mathbb{V} \subset \mathcal{K}$ přípustných tuhých posunutí, pro která se nekoná žádná práce.

$$\mathcal{K}_\mathbb{V} = \{v \in \mathcal{K} : a_\mathbf{e}(v, v) + b_\mathbf{e}(v^+, v) = 0\} = \{p \in \mathbf{P}_0 : p \leq 0\}.$$

Věty 4 a 5 nám dávají možnost jednoznačné dekompozice prostoru $\mathbb{V}$ na ortogonální součet kužele $\mathcal{K}_\mathbb{V}$ a jeho negativního polární kužele. Negativní polární kužel $\mathcal{K}_\mathbb{V}^\ominus$ je vzhledem ke skalárnímu součinu na $H^2(\Omega)$ definován jako

$$\mathcal{K}_\mathbb{V}^\ominus = \{v \in \mathbb{V} : (v, p)_{2,2} \leq 0 \ \ \forall p \in \mathcal{K}_\mathbb{V}\} = \{v \in \mathbb{V} : (v, 1)_2 \geq 0\}.$$

Pak můžeme provést ortogonální rozklad prostoru $\mathbb{V}$ na direktní součet

$$\mathbb{V} = \mathcal{K}_\mathbb{V} \oplus \mathcal{K}_\mathbb{V}^\ominus.$$

Pro každý průhyb $v \in \mathbb{V}$ pak existuje jednoznačná dekompozice ve tvaru

$$v = p \oplus \bar{v}, \ \ p \in \mathcal{K}_\mathbb{V} \ \wedge \ \bar{v} \ \in \mathcal{K}_\mathbb{V}^\ominus, \ \ (p, \bar{v})_{2,2} = p(\bar{v}, 1)_2 = 0. \tag{8}$$

S přihlédnutím k definici kuželů $\mathcal{K}_\mathbb{V}$ a $\mathcal{K}_\mathbb{V}^\ominus$ a k vlastnostem rozkladu (8) je zřejmé, že musí nastat právě jedna z následujících možností

$$\begin{aligned} &(\mathbf{i}) \qquad p = 0 \ \wedge \ (\bar{v}, 1)_2 \geq 0, \\ &(\mathbf{ii}) \qquad p \leq 0 \ \wedge \ (\bar{v}, 1)_2 = 0. \end{aligned}$$

Věta 6 (Nutná a postačující podmínka existence řešení) *Nechť $\beta \in L^\infty(\Omega)$, $\mathbf{e} = [t, q] \in U_{ad}$. Stavová úloha $(P(\mathbf{e}))$ má alespoň jedno řešení právě tehdy, když*

$$F(1) \geq 0 \tag{9}$$

Důkaz

1. Nutnost "⇒"
   Nechť $w \in \mathbb{V}$ je řešení úlohy $(P(\mathbf{e}))$, pak

   $$a_\mathbf{e}(w, v) + b_\mathbf{e}(w^+, v) = F(v) \quad \forall v \in \mathbb{V} \tag{10}$$

   Rovnice (10) musí platit $\forall p \in \mathcal{K}_\mathbb{V}$.

   $$0 \geq b_\mathbf{e}(w^+, p) = a_\mathbf{e}(w, p) + b_\mathbf{e}(w^+, p) = F(p) = pF(1) \quad \forall p \in \mathcal{K}_\mathbb{V}.$$

   Z vlastností formy $b_\mathbf{e}$ je zřejmé, že podmínka (9) platí.

2. Postačitelnost "⇐"
Nechť platí podmínka (9). Využijeme rozkladu $v = p \oplus \bar{v}$ a funkcionál $E_{\mathbf{e}}$ rozepíšeme jako

$$\begin{aligned} 2E_{\mathbf{e}}(v) &= 2E_{\mathbf{e}}(p+\bar{v}) = a_{\mathbf{e}}(\bar{v},\bar{v}) + b_{\mathbf{e}}(v^+,v^+) - 2F(p) - 2F(\bar{v}) \geq \\ &\geq \beta_0 T_0^3 \,|\bar{v}|_{2,2}^2 + Q_0||(p+\bar{v})^+||_2^2 + 2|p|F(1) - 2F(\bar{v}). \end{aligned}$$

V případě **(i)** platí $p = 0$. Pak tedy $v \equiv \bar{v}$ a $(\bar{v},1)_2 \geq 0$. Z vlastností funkce $\bar{v}^+$ plyne

$$0 \leq (\bar{v},1)_2^2 \leq (\bar{v}^+,1)_2^2 \leq l||\bar{v}^+||_2^2 \tag{11}$$

Užitím nerovností (11),(1) a podmínky (9) dostáváme

$$\begin{aligned} 2E_{\mathbf{e}}(v) &= 2E_{\mathbf{e}}(\bar{v}) = a_{\mathbf{e}}(\bar{v},\bar{v}) + b_{\mathbf{e}}(v^+,v^+) - 2F(p) - 2F(\bar{v}) \geq \\ &\geq \beta_0 T_0^3 \,|\bar{v}|_{2,2}^2 + Q_0||\bar{v}^+||_2^2 - 2F(\bar{v}) \geq \\ &\geq \beta_0 T_0^3 \,|\bar{v}|_{2,2}^2 + \frac{Q_0}{l}(\bar{v},1)_2^2 - 2F(\bar{v}) \geq \\ &\geq \|v\|_{2,2}\,(c_5\,\|v\|_{2,2} - 2\,\|f\|^*) \end{aligned}$$

Odtud už plyne koercivita funkcionálu $E_{\mathbf{e}}$ pro každé $\mathbf{e} \in U_{ad}$.

V případě **(ii)** platí $(\bar{v},1)_2 = 0$ a $p \leq 0$. Potom

$$\begin{aligned} 2E_{\mathbf{e}}(v) &= 2E_{\mathbf{e}}(c+\bar{v}) = a_{\mathbf{e}}(\bar{v},\bar{v}) + b_{\mathbf{e}}(v^+,v^+) - 2F(p) - 2F(\bar{v}) \geq \\ &\geq \beta_0 T_0^3 \,|\bar{v}|_{2,2}^2 + Q_0||(p+\bar{v})^+||_2^2 + 2|p|F(1) - 2F(\bar{v}) \geq \\ &\geq \beta_0 T_0^3 \,|\bar{v}|_{2,2}^2 + |p|F(1) - F(\bar{v}) \geq \\ &\geq \beta_0 T_0^3 \,|\bar{v}|_{2,2}^2 + (\bar{v},1)_2^2 + 2|p|F(1) - 2F(\bar{v}) \geq \\ &\geq c_6\,\|\bar{v}\|_{2,2}^2 + 2|p|F(1) - 2\,\|f\|^*\,\|\bar{v}\|_{2,2} \end{aligned}$$

Protože rozklad $\mathcal{K}_{\mathbb{V}} \oplus \mathcal{K}_{\mathbb{V}}^{\ominus}$ je ortogonální, platí $\|v\|_{2,2}^2 = \|\bar{v}\|_{2,2}^2 + |p|^2$. Jestliže $\|v\|_{2,2} \to \infty$, potom to samé musí platit pro alespoň jednu část funkce v. Z posledního odhadu funkcionálu $E_{\mathbf{e}}$ a z podmínky $F(1) \geq 0$ pak plyne jeho koercivita pro každé $\mathbf{e} \in U_{ad}$.

Máme tedy splněny všechny předpoklady věty 2 a existuje tak alespoň jedno řešení úlohy $(P(\mathbf{e}))$ na $\mathbb{V}$. □

Poznámka 1 Z předchozího důkazu a z omezenosti forem $a_{\mathbf{e}}, b_{\mathbf{e}}$ je zřejmé, že existují konstanty $c_1, c_2 > 0$, tak že

$$c_1||v||_{2,2}^2 \leq a_{\mathbf{e}}(v,v) + b_{\mathbf{e}}(v^+,v) \leq c_2||v||_{2,2}^2 \quad \forall v \in \mathcal{K}_{\mathbb{V}}^{\ominus},\ \forall \mathbf{e} \in U_{ad} \tag{12}$$

Součet forem $(a_{\mathbf{e}}(v,v) + b_{\mathbf{e}}(v^+,v))^{1/2}$ je tedy pro každé $\mathbf{e} \in U_{ad}$ ekvivalentní normou na negativním polárním kuželu $\mathcal{K}_{\mathbb{V}}^{\ominus}$.

Věta 7 (Nutná a postačující podmínka existence a jednoznačnosti řešení) *Nechť* $\beta \in L^\infty(\Omega)$, $\mathbf{e} = [t, q] \in U_{ad}$. *Stavová úloha* $(P(\mathbf{e}))$ *má právě jedno řešení právě tehdy, když*

$$F(1) > 0 \tag{13}$$

Důkaz

1. Nutnost "$\Rightarrow$"
   První část důkazu provedeme sporem. Předpokládejme, že existuje řešení úlohy $(P(\mathbf{e}))$ a je jednoznačné, označme ho jako w. Dále předpokládejme, že podmínka (13) neplatí. Pak z předchozí věty plyne, že musí platit rovnost

   $$F(1) = 0.$$

   Podle Eulerovy nutné podmínky (viz [3]) platí

   $$a_{\mathbf{e}}(w, v) + b_{\mathbf{e}}(w^+, v) = F(v) \quad \forall v \in \mathbb{V}$$

   a tedy

   $$\begin{aligned} a_{\mathbf{e}}(w, p) + b_{\mathbf{e}}(w^+, p) = F(p) = pF(1) \quad \forall p \in \mathcal{K}_{\mathbb{V}},\ p \neq 0, \\ b_{\mathbf{e}}(w^+, p) = 0 \quad \forall p \in \mathcal{K}_{\mathbb{V}},\ p \neq 0, \end{aligned}$$

   Odtud plyne, že $w^+ = 0,\ w \leq 0$ a $w + p < 0\ \forall p \in \mathcal{K}_{\mathbb{V}},\ p \neq 0$. Z vlastnosti $b_{\mathbf{e}}((w+p)^+, v) = 0\ \forall p \in \mathcal{K}_{\mathbb{V}},\ p \neq 0$ je zřejmé, že $w + p$ je řešením úlohy $(P(\mathbf{e}))$. To je ale ve sporu s předpokladem jednoznačnosti řešení. Musí tedy platit podmínka (13).

2. Postačitelnost "$\Leftarrow$"
   Nechť platí podmínka (13). Dále nechť $p \in \mathcal{K}_{\mathbb{V}},\ p \neq 0$ je řešením stavové úlohy $(P(\mathbf{e}))$. Pak

   $$\begin{aligned} a_{\mathbf{e}}(p, v) + b_{\mathbf{e}}(p^+, v) = F(v) \quad \forall v \in \mathbb{V}, \\ a_{\mathbf{e}}(p, p) + b_{\mathbf{e}}(p^+, p) = F(p) \quad \forall p \in \mathcal{K}_{\mathbb{V}},\ p \neq 0, \\ 0 = F(1). \end{aligned}$$

   Což je spor s podmínkou (13), která tedy zaručí, že žádné tuhé posunutí z $\mathcal{K}_{\mathbb{V}}$ nemůže být řešením úlohy $(P(\mathbf{e}))$. Za platnosti dané podmínky tedy stačí zkoumat vlastnosti funkcionálu $E_{\mathbf{e}}$ pouze na uzavřeném konvexním kuželu $\mathcal{K}_{\mathbb{V}}^{\ominus}$. Z poznámky 1 plyne ryzí konvexita funkcionálu $E_{\mathbf{e}}$ na $\mathcal{K}_{\mathbb{V}}^{\ominus}$ a podle věty 3 existuje právě jedno řešení úlohy $(P(\mathbf{e}))$. □

7 Úloha tvarové optimalizace

Máme tedy zformulovanou stavovou úlohu a uvedli jsme podmínky, za kterých je jednoznačně řešitelná pro každé $\mathbf{e} \in U_{ad}$. Pokračovat budeme zavedením cenového funkcionálu a formulací celkové úlohy optimalizace. Na závěr uvedeme důkaz existence řešení dané optimalizační úlohy.

Cenový funkcionál definujeme nejprve obecně jako zobrazení

$$\mathcal{J} : U_{ad} \times \mathbb{V} \to \mathbb{R} \tag{14}$$

Úloha optimalizace nosníku má potom následující podobu

$$\begin{cases} \text{Nalézt } \mathbf{e}^* \in U_{ad} \text{ tak, že} \\ \mathcal{J}(\mathbf{e}^*, u(\mathbf{e}^*)) \leq \mathcal{J}(\mathbf{e}, u(\mathbf{e})) \ \ \forall \mathbf{e} \in U_{ad}, \end{cases} \tag{$\mathcal{P}$}$$

kde $u(\mathbf{e}) \in \mathbb{V}$ je řešení stavové úlohy $(P(\mathbf{e}))$.

Poznámka 2 Schéma úlohy $(\mathcal{P})$ má následující podobu

$$\mathbf{e} \longmapsto u(\mathbf{e}) \longmapsto \mathcal{J}(\mathbf{e}, u(\mathbf{e})). \tag{15}$$

Obecně to s sebou přináší obtíže při numerické realizaci takových úloh. Cenový funkcionál vzniklý složením dvou zobrazení ve výsledku nemusí být konvexní a diferencovatelný, zejména v případě problémů se stavovou variační nerovnicí. Abychom tedy zvolili vhodný postup numerické realizace, je nutná podrobná matematická analýza problému. Další nepříjemností je nutnost řešení stavové úlohy v každém kroku vnějšího optimalizačního algoritmu. Toto nám značně prodlužuje výpočet a klade větší nároky na používanou výpočetní techniku.

Poznámka 3 Nyní uveďme některé příklady cenových funkcionálů.

$$\mathcal{J}_1(\mathbf{e}, u(\mathbf{e})) = \int_\Omega f(x)u(x)\,\mathrm{d}x, \quad \mathcal{J}_2(\mathbf{e}, u(\mathbf{e})) = \int_\Omega |u(x)|\,\mathrm{d}x,$$

$$\mathcal{J}_3(\mathbf{e}, u(\mathbf{e})) = \int_\Omega u^2(x)\,\mathrm{d}x, \quad \mathcal{J}_4(\mathbf{e}, u(\mathbf{e})) = \int_\Omega (u'(x))^2\,\mathrm{d}x.$$

Uvedené funkcionály nejsou explicitně závislé na návrhové proměnné $\mathbf{e}$. Obecně je však cenový funkcionál závislý jak na návrhové proměnné, tak na řešení stavové úlohy.

V dalším budeme předpokládat, že pro cenový funkcionál $\mathcal{J}$ platí

$$\liminf_{n\to\infty} \mathcal{J}(\mathbf{e}_n, u_n) \geq \mathcal{J}(\mathbf{e}, u)$$

pro každé $\mathbf{e}, \mathbf{e}_n \in U_{ad},\ \mathbf{e}_n \to \mathbf{e}$ a pro každé $u, u_n \in \mathbb{V},\ u_n \to u$. Funkcionál je tedy zdola polospojitý na $U_{ad} \times \mathbb{V}$.

Nyní přistoupíme ke klíčovému bodu celé analýzy řešitelnosti úlohy $(\mathcal{P})$. Tím bude spojitá závislost řešení stavové úlohy na návrhové proměnné. Jde v podstatě o analýzu první části schématu (15).

Věta 8 (O spojité závislosti) *Nechť* $\mathbf{e}_n, \mathbf{e} \in U_{ad},\ \mathbf{e}_n \to \mathbf{e}$. *Dále nechť* $u_n(\mathbf{e}_n) \in \mathbb{V}$ *jsou řešení stavové úlohy* $(P(\mathbf{e}_n))$ *a platí podmínka* $L(1) > 0$. *Pak existuje funkce* $u \in \mathbb{V}$ *tak, že*

$$u_n \to u \ \text{ ve } \mathbb{V}$$

a navíc $u = u(\mathbf{e})$ *je řešením stavové úlohy* $(P(\mathbf{e}))$.

Důkaz Nechť $\mathbf{e}_n, \mathbf{e} \in U_{ad},\ \mathbf{e}_n \to \mathbf{e}$. Z definice konvergence v U_{ad} plyne, že $t_n(x) \rightrightarrows t(x)$ v Ω a $q_n(x) \rightharpoonup q(x)$ v $L^2(\Omega)$. Dále nechť u_n řeší stavovou úlohu $(P(\mathbf{e}_n))$. Pak podle Eulerovy nutné podmínky minima (viz [3]) platí

$$a_{\mathbf{e}}(u_n, v) + b_{\mathbf{e}}(u_n^+, v) = F(v) \quad \forall v \in \mathbb{V}. \tag{16}$$

Volbou $v = u_n$ dostáváme

$$a_{\mathbf{e}}(u_n, u_n) + b_{\mathbf{e}}(u_n^+, u_n) = F(u_n). \tag{17}$$

Podmínka (13) a poznámka 1 zajistí platnost následujícího vztahu

$$c_1 ||u_n||_{2,2}^2 \le a_{\mathbf{e}}(u_n, u_n) + b_{\mathbf{e}}(u_n^+, u_n) = F(u_n) \le ||f||_2 ||u_n||_{2,2}.$$
$$||u_n||_{2,2} \le c.$$

Konstanta c je nezávislá na indexu n. Posloupnost $\{u_n\}$ je omezená a existuje vybraná podposloupost $\{u_{n_j}\}$ spolu s funkcí $u \in \mathbb{V}$ tak, že

$$u_{n_j} \rightharpoonup u \quad \text{v } H^2(\Omega). \tag{18}$$

Nyní je třeba dokázat, že u řeší úlohu $(P(\mathbf{e}))$. Zapíšeme stavovou úlohu pro $\mathbf{e}_{n_j}$, u_{n_j} a provedeme limitní přechod.

$$\int_\Omega \beta t_{n_j}^3 u''_{n_j} v'' \,\mathrm{d}x + \int_\Omega q_{n_j} u_{n_j}^+ v \,\mathrm{d}x = \int_\Omega f v \,\mathrm{d}x \quad \forall v \in \mathbb{V}. \tag{19}$$

Nejprve budeme zkoumat první člen rovnosti (19). Využijeme omezenosti $\{u_n\}$, slabé konvergence (18) a definice konvergence v U_{ad}.

$$\lim_{n_j\to\infty} \int_\Omega \beta t_{n_j}^3 u''_{n_j} v'' \,\mathrm{d}x = \int_\Omega \lim_{n_j\to\infty} [\beta t_{n_j}^3 u''_{n_j} v'' \pm \beta t^3 u''_{n_j} v''] \,\mathrm{d}x =$$
$$= \int_\Omega \lim_{n_j\to\infty} \beta [t_{n_j}^3 - t^3] u''_{n_j} v'' \,\mathrm{d}x + \int_\Omega \lim_{n_j\to\infty} \beta t^3 u''_{n_j} v'' \,\mathrm{d}x =$$
$$= \int_\Omega \beta t^3 u'' v'' \,\mathrm{d}x.$$

V analýze druhého člene využijeme lemma 2, omezenosti $\{u_n\}$, slabé konvergence (18) a definice konvergence v U_{ad}. Z omezenosti $\{u_n\}$ plyne i omezenost $\{u_n^+\}$,

neboť $||u_n^+||_{1,2} \leq ||u_n||_{2,2} \leq c$. Protože vnoření $H^2(\Omega) \subset H^1(\Omega)$ je kompaktní, platí

$$(u_n \rightharpoonup u \quad \text{v } H^2(\Omega)) \Rightarrow (u_n \to u \quad \text{v } L^2(\Omega)). \tag{20}$$

Provedením limitního přechodu v druhém členu dostáváme

$$\begin{aligned}
&\lim_{n_j\to\infty} \int_\Omega q_{n_j} u_{n_j}^+ v \,\mathrm{d}x = \int_\Omega \lim_{n_j\to\infty} q_{n_j} u_{n_j}^+ v \pm q u_{n_j}^+ v \,\mathrm{d}x = \\
&= \int_\Omega \lim_{n_j\to\infty} [q_{n_j} - q] u_{n_j}^+ v \,\mathrm{d}x + \int_\Omega \lim_{n_j\to\infty} q u_{n_j}^+ v \,\mathrm{d}x = \\
&= \int_\Omega q u^+ v \,\mathrm{d}x.
\end{aligned}$$

Celkově tedy platí

$$\int_\Omega \beta t^3 u'' v'' \,\mathrm{d}x + \int_\Omega q u^+ v \,\mathrm{d}x = \int_\Omega f v \,\mathrm{d}x \quad \forall v \in \mathbb{V}.$$

Limitní funkce $u \in \mathbb{V}$ je řešením stavové úlohy pro $\mathbf{e} = [t, q]$. Protože je stavová úloha za daných podmínek jednoznačně řešitelná, k funkci u nekonverguje pouze vybraná posloupnost $\{u_{n_j}\}$, ale celá $\{u_n\}$. Posledním bodem je důkaz silné konvergence. Víme, že posloupnost $\{u_n\}$ konverguje k u slabě, dokážeme-li konvergenci "v normě", bude tím dokázána i silná konvergence $u_n \to u$ ve $\mathbb{V}$. Z poznámky 1 víme, že $|||v||| = (a_{\mathbf{e}}(v,v) + b_{\mathbf{e}}(v^+, v))^{1/2}$ je tedy pro každé $\mathbf{e} \in U_{ad}$ ekvivalentní normou na $\mathcal{K}_{\mathbb{V}}^{\ominus}$. Stačí dokázat konvergenci posloupnosti $\{u_n\}$ v normě $|||\cdot|||$.

$$\begin{aligned}
\lim_{n\to\infty} |||u_n||| &= \lim_{n\to\infty} \int_\Omega \beta t^3 (u_n'')^2 + q (u_n^+)^2 \,\mathrm{d}x = \\
&= \lim_{n\to\infty} \int_\Omega \beta [t^3 - t_n^3](u_n'')^2 + [q - q_n](u_n^+)^2 \,\mathrm{d}x + \\
&+ \lim_{n\to\infty} \int_\Omega \beta t_n^3 (u_n'')^2 + q_n (u_n^+)^2 \,\mathrm{d}x = \lim_{n\to\infty} \int_\Omega f u_n \,\mathrm{d}x = \\
&= \int_\Omega f u \,\mathrm{d}x = \int_\Omega \beta t^3 (u'')^2 + q (u^+)^2 \,\mathrm{d}x = |||u|||.
\end{aligned}$$

Využili jsme opět omezenosti posloupností $\{u_n\}$ a $\{u_n^+\}$, vlastností konvergence v U_{ad} a definice stavové úlohy pro $\mathbf{e}, \mathbf{e}_n$. Tvrzení věty je tímto dokázáno. □

Shrnutím všech dosavadních poznatků o problému dostáváme následující existenční větu.

Věta 9 *Nechť platí podmínka $F(1) > 0$, potom existuje alespoň jedno řešení optimalizační úlohy $(\mathcal{P})$.*

Důkaz Označme $\lambda = \inf_{\mathbf{e}\in U_{ad}} \mathcal{J}(\mathbf{e}, u(\mathbf{e}))$. Z vlastnosti infima plyne existence minimizační posloupnosti $\{\mathbf{e}_n\} \subset U_{ad}^t$ takové, že

$$\lambda = \lim_{n\to\infty} \mathcal{J}(\mathbf{e}_n, u_n(\mathbf{e}_n)).$$

Z kompaktnosti U_{ad} plyne existence vybrané posloupnosti $\{\mathbf{e}_{n_j}\}$ a prvku $\mathbf{e}^*$ tak, že $\mathbf{e}_{n_j} \to \mathbf{e}^*$ v U_{ad}. Z věty 8 o spojité závislosti je zřejmé, že $u_n(\mathbf{e}_n) \to u^*(\mathbf{e}^*)$, kde $u_n(\mathbf{e}_n), u^*(\mathbf{e}^*)$ řeší stavovou úlohu $(P(\mathbf{e}_n))$, resp. $(P(\mathbf{e}^*))$. Díky polospojitosti cenového funkcionálu platí

$$\lambda = \lim_{n\to\infty} \mathcal{J}(\mathbf{e}_n, u_n(\mathbf{e}_n)) \geq \mathcal{J}(\mathbf{e}^*, u^*(\mathbf{e}^*)).$$

Pak $\mathcal{J}(\mathbf{e}^*, u^*(\mathbf{e}^*)) = \min_{\mathbf{e}\in U_{ad}} \mathcal{J}(\mathbf{e}, u(\mathbf{e}))$, čímž je tvrzení věty dokázáno. □

8 Závěr

V práci jsme se zabývali analýzou řešitelnosti úlohy optimalizace nosníku spočívajícího na jednostranném podloží. Použili jsme Euler–Bernoulliův model nosníku s Winklerovým podložím, které bylo do modelu zahrnuto použitím tzv. odezvové funkce. Z matematického úhlu pohledu byla stavová úloha nelineárním semi-koercivním problémem čtvrtého řádu se smíšenými okrajovými podmínkami. Předmětem optimalizace byla tloušťka nosníku a koeficient tuhosti podloží.

Byly stanoveny podmínky existence a jednoznačnosti řešení stavové úlohy. Použili jsme přístup založený na dekompozici prostoru přípustných posunutí na kužel tuhých posunutí a jeho negativní polární kužel. Vzhledem k semikoercivitě stavové úlohy nemůžeme použít k důkazu existence řešení použít standartní Fridrichsovu nebo Poincarého nerovnost. Pro tyto účely jsme pro náš speciální případ dokázali nerovnost Poincarého typu, kterou lze na zmíněných podprostorech použít. Stanovili jsme také podmínky existence řešení celkové optimalizační úlohy. Uvedli jsme i několik možných variant cenových funkcionálů splňujících potřebné vlastnosti.

Na tuto práci lze navázat i několika zobecněními. Místo Winklerova jednoparametrického modelu podloží můžeme použít například dvouparametrický Pasternakův model s odezvovou funkcí $q(x)u(x) - k(x)u''(x)$. Speciálním případem je pak dokonale tuhé podloží (překážka) a výsledný model pak vede na variační nerovnici 1. druhu. Máme také několik dalších možností volby okrajových podmínek, pro které zůstává stavová úloha semikoercivní (viz [13]). V optimalizační části můžeme uvažovat přidání dalšího optimalizačního parametru, například Youngova modulu pružnosti $E(x)$ nebo šířky $b(x)$. Můžeme také uvažovat nosník s jiným než obdélníkovým průřezem.

Na tento příspěvek bychom chtěli navázat v další práci, která by se měla věnovat aproximaci problému a jeho algebraické podobě. Dále bychom rádi navrhli postup numerické realizace problému a jeho algoritmického zpracování.

Literatura

[1] Aubin, J. P.: *Applied Functional Analysis.* J. Wiley and Sons, New York, 1979

[2] Ekeland, I., Temam, R.: *Convex Analysis and Variational Problems.* North Holland, 1976

[3] Fučík, S., Kufner, A.: *Nelineární diferenciální rovnice.* Praha, SNTL, 1978

[4] Haslinger, J., Mäkinen, R. A. E.: *Introduction to Shape Optimization: Theory, Approximaion and Computation.* SIAM, Philadelphia, 2003

[5] Haslinger, J., Neittaanmaki, P.: *Finite Element Aproximation for Optimal Shape, Material and Topology Design.* Wiley, Chichester, 1997 (second edition).

[6] Haslinger, J.: *A note on contact shape optimization with semicoercive state problems.* Applications of Mathematics **47**, 5 (2002), 397–410.

[7] Horák, J. V., Fibinger, P.: *O rešitelnosti semikoercivních úloh ohybu desek – nerovnice.* ODAM, Olomouc, 2001

[8] Horák, J. V., Netuka, H.: *Mathematical model of pseudointeractive set: 1D body on nonlinear subsoil: I. Theoretical aspects.* Engineering Mechanics **14** (2007).

[9] Chleboun, J.: *Optimal design of an elastic beam on an elastic basis.* Applications of Mathematics **31**, 2 (1986), 118–140.

[10] Kufner, A., John, O, Fučík, S.: *Function Spaces*, Academia, Praha, 1977

[11] Nečas, H., Hlaváček, I.: *Úvod do matematické teorie pružných a pružně plastických těles.* Praha, SNTL, 1981

[12] Netuka, H., Horák, J. V.: *Soustava nosník–pružiny–podloží po dvou letech.* ODAM, Olomouc, 2007

[13] Sysala, S.: *Unilateral subsoil of Winkler's type: Semi-coercive beam problem.* Applications of Mathematics, Praha.

[14] Šimeček, R.: *Sizing Optimization of an Elastic Beam with a Rigid Obstacle: Numerical Realization.* SVOC, Olomouc, 2009.