UNIVERSITATIS PALACKIANAE OLOMUCENSIS
FACULTAS RERUM NATURALIUM

Department of Mathematical Analysis
and Applications of Mathematics

ODAM
1999

PREPRINT SERIES, 31 (1999)

Editors: Jiří V. Horák & Miloslav Závodný

Katedra matematické analýzy a aplikací matematiky
Přírodovědecká fakulta Univerzity Palackého v Olomouci
Tomkova 40, 779 00 Olomouc — tel.: (42 68) 412 604, 412 210

Olomoucké dny aplikované matematiky 1999

Vážené kolegyně a kolegové,

předkládáme Vám poprvé sborník přednášek z konference Olomoucké dny aplikované matematiky 1999, která se již tradičně konala na konci letního semestru ve dnech 17.–18. června 1999 v budovách PřF UP Olomouc, tentokráte se zaměřením především na aplikace v úlohách mechaniky kontinua.

Chceme tak založit novou tradici v prezentaci a archivaci příspěvků. V tomto sborníku jsou však publikovány jen ty příspěvky, jež byly dodány včas v oznámeném termínu a v požadované formě (TEX).

Publikované texty neprošly žádnými korekturami, a tak za jejich odbornou i jazykovou stránku nesou odpovědnost výhradně jejich autoři.

Děkujeme všem aktivním účastníkům za jejich přednášky a autorům písemných textů za jejich včasné odeslání editorům. Organizátory konference potěšila zejména aktivní účast studentů doktorandského studia katedry MAaAM, kteří referovali o výsledcích své výzkumné činnosti.

Na shledání v Olomouci na rozšířeném programu ODAM 2001 v roce 2001 se těší a srdečně zve organizační výbor.

Jiří Horák, Miloslav Závodný

Jiří V. Horák
Department of Mathematical Analysis and Applications of Mathematics, Faculty of Science, Palacký University, Tomkova 40, 779 00 Olomouc, Czech Republic
horakj@risc.upol.cz

Miloslav Závodný
Department of Mathematical Analysis and Applications of Mathematics, Faculty of Science, Palacký University, Tomkova 40, 779 00 Olomouc, Czech Republic
zavodny@risc.upol.cz

CONTENTS


Univ. Palacki. Olomuc., Fac. rer. nat.,
Dept of Math. Anal. and Appl. of Math.
ODAM (1999) 5–20

Optimal interpolation with quadratic splines on simple grid *

Jiří KOBZA

Department of Mathematical Analysis and Applications of Mathematics,
Faculty of Science, Palacký University,
Tomkova 40, 779 00 Olomouc, Czech Republic
e-mail: kobza@risc.upol.cz

Abstrakt

Quadratic interpolatory splines on simple knotset depend on one free parameter. We can find optimal value of such a parameter to reach minimum of proper norm or another quantitative criteria. It gives to the user some possibility to choose such a variant of computing algorithm (to prescribe free parameter, criterium) which corresponds to his needs. Some frequently used criteria are considered in details and demonstrated on examples.

Key words: Splines, quadratic splines, optimal spline interpolation.

1991 Mathematics Subject Classification: 41A15, 65D05

*Supported by grant No. 201/96/0665 of The Grant Agency of Czech Republic.

1 Problem statement

Let us have given vectors $\mathbf{x}$ of real monotone knotset and prescribed function values $\mathbf{s}$ in knots

$$\mathbf{x} = \{x_i,\ i = 0(1)n+1\}, \quad \mathbf{s} = \{s_i,\ i = 0(1)n+1\} \tag{1}$$

with knot stepsizes and slopes

$$h_i = x_{i+1} - x_i\,, \quad p_i = (s_{i+1} - s_i)/h_i\,,\ i = 0(1)n. \tag{2}$$

It is well known (see [2], [3]), that quadratic spline (from C^1—with the defect one) which interpolates in knots x_i the function values s_i has one free parameter—usually taken as value of the first or second derivative in some knot. When we take the values s_i, s_{i+1} as two local parameters for the local spline representation on the interval $[x_i, x_{i+1}]$, then the third local parameter (e.g. first or second derivative in knot) can be computed by recursion from *continuity conditions (CC)* $s'(x_i - 0) = s'(x_i + 0)$ in knots x_i, expressed in local parameters used (see examples below). So the choice of free parameter influences the behavior of the spline over the whole interval and according to the properties of (CC) the errors are propagated without damping (see [3]). When we have no idea about the proper value of the free parameter corresponding to the problem under search, we may ask for a procedure for finding some optimal value of such a parameter according to the criterion which the user can choose more easily. Such criteria may represent some norms of first or second derivative (for continuous functions or vectors of their values in knots), some measure of the curvature. In the most used cases the problem can be then formulated as some quadratic programming problem. We can even find explicit solution for optimal values in some problems. It is also possible to use the simple approach which computes the vector of optimal parameters with pseudoinverse matrix. Applying results from difference equations to continuity conditions we can obtain also some another type of algorithms for solving such special types of quadratic programming problems. Let us mention, that the quadratic splines interpolating function values have not extremal property on some wider class of functions (as have linear or natural cubic splines) and so we search for the optimizer in the class of quadratic splines on simple grid only (the points of interpolation coincide with spline knots). Some comparison of results is given on examples in the last section.

2 Minimal norm of first derivative

When we use the local representation of quadratic spline

$$s(x) = (1-u^2)s_i + u^2 s_{i+1} + h_i u(1-u)m_i \tag{3}$$

with local variable $u = (x - x_i)/h_i$, local parameters s_i, s_{i+1} and $m_i = s'(x_i)$, then the continuity conditions can be written as recurrence

$$m_i + m_{i+1} = 2p_i\,,\ i = 0(1)n, \quad p_i = (s_{i+1} - s_i)/h_i\,. \tag{4}$$

When the values $s_i,\ i = 0(1)n+1$ are given, then values $m_i,\ i = 0(1)n+1$ depend on one free paramater. Let us choose as free parameter the value m_0; then remaining values m_i can be computed from relations (4) and we obtain by induction (with $Q_0 = 0$)

$$m_i = (-1)^i m_0 + 2Q_i \quad \text{with} \quad Q_i = \sum_{j=0}^{i-1}(-1)^{i+j+1}p_j\,; \quad \text{then} \quad \frac{dm_i}{dm_0} = (-1)^i. \tag{5}$$

The L_2-norm of the first derivative can be assumed as the measure of the speed changes in the continuous process described by the data given. In the class of continuous interpolating functions (more generally in W_2^1) this norm is minimized by interpolating polygon (see [3], [5]). Let us find the optimizer in the class of quadratic splines. Using Simpson's rule of numerical integration, we can write (exactly for quadratic splines)

$$\begin{aligned} J_1(s) = \|s'(x)\|_2^2 &= \int_{x_0}^{x_{n+1}} [s'(x)]^2 dx = \frac{1}{6}\sum_{i=0}^{n} h_i\Big[m_i^2 + 4\Big(\frac{m_i + m_{i+1}}{2}\Big)^2 + m_{i+1}^2\Big] \\ &= \frac{1}{3}\sum_{i=0}^{n} h_i(m_i^2 + m_i m_{i+1} + m_{i+1}^2) = \sum_{i=0}^{n} h_i p_i^2 + \frac{1}{3}\sum_{i=0}^{n} h_i(p_i - m_i)^2 \end{aligned} \tag{6}$$

where $m_{i+1} = 2p_i - m_i$ was substituted from (4). Our problem has no trivial solution with $p_i = m_i$ (CC) not fulfilled] and so we can find here the lower bound for $J_1(s) = J_1(m_0)$ only.

2.1 Explicit solution for optimal value m_0

Using (4) and (5) we can obtain for the terms in the last part of (6) the following expression

$$p_i - m_i = (-1)^{i+1}m_0 + P_i\,, \quad P_i = p_i - 2Q_i\,. \tag{7}$$

For the value of $J_1(s)$ as function of free parameter m_0 we then obtain

$$J_1(m_0) = \sum_{i=0}^{n} h_i p_i^2 + \frac{1}{3}\sum_{i=0}^{n} h_i P_i^2 + \frac{2}{3}m_0\sum_{i=0}^{n}(-1)^{i+1}h_i P_i + \frac{1}{3}m_0^2\sum_{i=0}^{n} h_i\,. \tag{8}$$

The necessary condition for minimum with respect to parameter m_0, $dJ_1/dm_0 = 0$, results then in the explicit expression for optimal value m_0

$$m_0 = \frac{1}{x_{n+1} - x_0}\sum_{i=0}^{n}(-1)^i h_i P_i = \frac{1}{x_{n+1} - x_0}\sum_{i=0}^{n}(-1)^{i+1}(p_i - 2Q_i). \tag{9}$$

The remaining local parameters m_i, $i = 1(1)n+1$ can be then computed from recurrences (4) without some error growth or damping.

When we want to minimize l_2-norm of the vector $\mathbf{m}$,

$$J_{1d}(s) = J_{1d}(m_0) = \sum_{i=0}^{n+1} m_i^2\,, \quad \text{then} \quad \frac{dJ_{1d}}{dm_0} = 2\sum_{i=0}^{n+1}(-1)^i m_i\,, \tag{10}$$

and the necessary condition of minima and (7) give the optimal value

$$m_0 = \frac{2}{n+2}\sum_{i=0}^{n+1}(-1)^i Q_i = \frac{2}{n+2}\sum_{j=0}^{n}(-1)^{j+1}(n+1-j)p_j\,. \tag{11}$$

2.2 Optimal vector m from linear system

The necessary condition for *minimum of* $J_1(m_0)$ can be written with the use of (5) as

$$0 = \frac{dJ_1}{dm_0} = \sum_{i=0}^{n}\frac{dJ_1}{dm_i}\frac{dm_i}{dm_0} = \sum_{i=0}^{n}(-1)^i\frac{dJ_1}{dm_i} = \sum_{i=0}^{n+1} c_i m_i\,, \tag{12}$$

where $c_0 = h_0$; $c_i = (-1)^i(h_{i-1}+h_i)$, $i = 1(1)n$; $c_{n+1} = (-1)^{n+1}h_n$.

We can now complete the system of equations (4) with this condition and obtain the system of equations with simple structure for computing all optimal values m_i :

$$m_i + m_{i+1} = 2p_i\,, \quad i = 0(1)n; \tag{13}$$

$$\sum_{i=0}^{n+1} c_i m_i = 0 \quad \text{or equivalently} \quad \sum_{i=0}^{n}(-1)^i h_i m_i = \sum_{i=0}^{n}(-1)^i h_i p_i\,.$$

When we are interested in the vector $\mathbf{m}$ of the first derivatives of the spline *with minimal l_2-norm*, we can compute such a vector from underdetermined system of equations

$$m_i + m_{i+1} = 2p_i\,, \quad i = 0(1)n \tag{14}$$

using pseudoinverse matrix (e.g. function *pinv* in MATLAB)—it is known (see [1]) that the solution of such system $\mathbf{Ax} = \mathbf{b}$ with minimal l_2-norm can be computed as $\mathbf{x} = pinv(\mathbf{A})\cdot\mathbf{b}$. Such solution we can compute also with solvers for the regular systems from the system (14) completed by condition following from (10)

$$\sum_{i=0}^{n+1}(-1)^i m_i = 0. \tag{15}$$

Theorem 1 *Let us have given the vectors of spline knots* $\mathbf{x}$ *and function values* $\mathbf{s}$. *Then there exists unique quadratic spline interpolating data* $(\mathbf{x}, \mathbf{s})$ *with minimal value of* $J_1(s)$, *determined by the initial value* m_0 *given in (9). The vector of optimal values of the first derivative can be computed from (14). For the spline with minimal value of* $J_{1d}(s)$ *these optimal parameters are given in (11) and as the solution of the system (14)–(15).*

3 Minimal norm of the second derivative

We can use the second derivative $M_i = s''(x_i + 0)$ as the third local parametr in spline local representation

$$s(x) = s(x_i + h_i u) = (1-u)s_i + us_{i+1} + \frac{1}{2}h_i^2 u(u-1)M_i\,. \tag{16}$$

The continuity conditions for the spline first derivatives in knots can be now written as recurrence for second derivatives (see e.g. [3], [4])

$$h_i M_i + h_{i+1} M_{i+1} = 2(p_{i+1} - p_i), \quad i = 0(1)n-1. \tag{17}$$

We can easily find the values of local parameters M_i which give *minimum to the l_2-norm of the vector* $\mathbf{M}$—we simply use the pseudoinverse for the solution of underdetermined system of recurrences (17). The first derivatives we can compute then as $m_i = p_i - h_i M_i/2$.

It is well-known (see e.g. [2], [3]) that natural cubic splines minimize the L_2-norm of the second derivative on class of W_2^2 interpolants. We shall find such a minimizer in the class of quadratic splines only. Because the quadratic spline second derivative is piecewise constant, we have

$$J_2(s) = \|s''(x)\|_2^2 = \int_{x_0}^{x_{n+1}} [s''(x)]^2 dx = \sum_{i=0}^{n} h_i M_i^2\,. \tag{18}$$

3.1 Optimal values of the second derivative

When we use the initial value of the second derivative M_0 as a free parameter, we obtain by recursion and differentiation

$$h_i M_i = S_i + (-1)^i h_0 M_0\,, \qquad \frac{dM_i}{dM_0} = (-1)^i h_0/h_i\,, \tag{19}$$

with $S_0 = 0$, Q_i from (5) and

$$S_i = 2p_i + 2(-1)^i p_0 + 4\sum_{j=1}^{i-1}(-1)^{j+i} p_j = 2[p_i - (-1)^i p_0] - 4Q_i\,. \tag{20}$$

The necessary condition for minimum of $J_2(s) = J_2(M_0)$ reads now

$$0 = \frac{dJ_2}{dM_0} = 2\sum_{i=0}^{n} h_i M_i \frac{dM_i}{dM_0} = 2\sum_{i=0}^{n}\Big[S_i + (-1)^i h_0 M_0\Big](-1)^i h_0/h_i\,. \tag{21}$$

We can easily find now *the explicit expression for optimal value* M_0 as

$$h_0 M_0 = \Big(\sum_{i=0}^{n}\frac{1}{h_i}\Big)^{-1}\sum_{i=0}^{n}(-1)^{i+1} S_i/h_i\,. \tag{22}$$

Remaining values of M_i can be then computed from recurrences (17), values of the first derivative as $m_i = p_i - h_i M_i/2$.

When we complete the system of continuity conditions (17) with the equation which we obtain inserting relation (19) into necessary condition for minimum of (18), the resulting system with simple structure we can write as

$$\begin{aligned} h_i M_i + h_{i+1} M_{i+1} &= 2(p_{i+1} - p_i), \quad i = 0(1)n-1 \\ \sum_{i=0}^{n} (-1)^i M_i &= 0 \ . \end{aligned} \tag{23}$$

We can now solve this regular system of equations and we obtain the whole *vector of optimal values* M_i for the optimal spline we search.

3.2 Optimal values of the first derivative

The problem discussed in this section can be solved also with the first derivative m_0 as free parameter. When we use the relation $M_i = (m_{i+1} - m_i)/h_i$, we can express the minimized functional in terms of parameters m_i as

$$J_2(s) = \sum_{i=0}^{n} h_i M_i^2 = \sum_{i=0}^{n} \frac{1}{h_i}(m_i^2 - 2m_i m_{i+1} + m_{i+1}^2) = 4\sum_{i=0}^{n} \frac{1}{h_i}(m_i^2 - 2p_i m_i). \tag{24}$$

When we substitute (5) into necessary condition for minimum

$$\frac{dJ_2}{dm_0} = \sum_{i=0}^{n} \frac{dJ_2}{dm_i}\frac{dm_i}{dm_0} = 8\sum_{i=0}^{n} \frac{1}{h_i}(m_i - p_i)(-1)^i = 0, \tag{25}$$

we obtain *the explicit expression for optimal value* m_0:

$$m_0 \sum_{i=0}^{n} \frac{1}{h_i} = \sum_{i=0}^{n} (-1)^i \frac{1}{h_i}(p_i - 2Q_i). \tag{26}$$

The remaining values of m_i we then can compute from recurrences (4).

Another possibility is to complete the system (4) with the condition (25) and to solve the resulting linear system

$$\begin{aligned} m_i + m_{i+1} &= 2p_i \, , \quad i = 0(1)n \\ \sum_{i=0}^{n} (-1)^i \frac{1}{h_i} m_i &= \sum_{i=0}^{n} (-1)^i \frac{1}{h_i} p_i \ . \end{aligned} \tag{27}$$

When we need *to minimize the* l_2*-norm of* $\mathbf{M}$, then the functional

$$J_{2d}(m_0) = \sum_{i=0}^{n} M_i^2 = 4\sum_{i=0}^{n} \frac{1}{h_i^2}(p_i - m_i)^2 \tag{28}$$

is minimized for the initial value of free parameter m_0 with

$$m_0 \sum_{i=0}^{n} \frac{1}{h_i^2} = \sum_{i=0}^{n} (-1)^i \frac{1}{h_i^2}(p_i - 2Q_i). \tag{29}$$

The remaining values of m_i we can compute from (4).

It is also possible *to compute the whole vector of optimal values* from the system of equations (4) completed with the equation (compare with (27))

$$\sum_{i=0}^{n} (-1)^i \frac{1}{h_i^2} m_i = \sum_{i=0}^{n} \frac{1}{h_i^2} p_i \,. \tag{30}$$

We can resume the results of this section in the following Theorem.

Theorem 2 *Given the spline knotset* $\mathbf{x}$ *and prescribed function values* $\mathbf{s}$*, there exists unique quadratic spline minimizing the functional* $J_2(s)$*. The optimal value of the corresponding free parameter* M_0 *is given in (22). We can compute the vector of optimal values* $\mathbf{M}$ *also from linear system (23) or as pseudoinverse solution of the system (17). When we choose the free parameter* m_0*, its optimal value is given in (26); alternatively we can compute the optimal vector* $\mathbf{m}$ *from linear system (27). The optimal value of* m_0 *for spline with minimimal value of* $J_{1d}(s)$ *(*l_2*-norm of* $\mathbf{M}$*) is given in (29). The whole vector* $\mathbf{m}$ *of optimal values we can compute from the system (4) completed with (30).*

Remark Let us mention the generally different values of free parameter m_0 which minimize the norms of first or second derivative.

4 Minimal norm of vector [m, M]

We can use also local parameters s_i, m_i, M_i in the Taylor's local spline representation

$$s(x) = s_i + m_i(x - x_i) + \frac{1}{2} M_i (x - x_i)^2. \tag{31}$$

The continuity conditions for the function and first derivative values can be now written as

$$\begin{aligned} s_i + h_i m_i + \frac{1}{2} h_i^2 M_i &= s_{i+1} \,, \\ m_i + h_i M_i &= m_{i+1} \,, \quad i = 0(1)n. \end{aligned} \tag{32}$$

Denoting again $p_i = (s_{i+1} - s_i)/h_i$, we can write relations (32) as the underdetermined system of $2n + 2$ linear equations with simple matrix structure for $2n + 3$ unknown parameters

$$\begin{aligned} m_i + \frac{1}{2} h_i M_i &= p_i \,, \\ m_i - m_{i+1} + h_i M_i &= 0 \,, \quad i = 0(1)n \end{aligned} \tag{33}$$

and solve it with pseudoinverse matrix for *parameters* **m, M** *with minimal* l_2*-norm of the vector* [**m, M**].

We can apply similar approaches as in foregoing to discuss the optimal value of the parameter M_0. By induction and Q_i given in (5) we obtain

$$\begin{aligned} h_i M_i &= 2[p_i + (-1)^i p_0] - 4Q_i + (-1)^i h_0 M_0\,, \\ m_i &= (-1)^{i+1}[p_0 + h_0 M_0] + 2Q_i\,. \end{aligned} \tag{34}$$

From it we obtain

$$\frac{dm_i}{dM_0} = (-1)^{i+1} h_0\,, \qquad \frac{dM_i}{dM_0} = (-1)^i \frac{h_0}{h_i}\,. \tag{35}$$

For l_2-norm of the vector [**m, M**]

$$J_{3d}(s) = \sum_{i=0}^{n+1} m_i^2 + \sum_{i=0}^{n} M_i^2 \quad \text{the condition} \quad \frac{dJ_{3d}}{dM_0} = 0 \tag{36}$$

results in the following relation between parameters m_i, M_i

$$\sum_{i=0}^{n+1} (-1)^i h_i m_i + \sum_{i=0}^{n} (-1)^{i+1} M_i = 0. \tag{37}$$

When we complete the relations (33) with equation (37), we obtain *system of linear equations for computing optimal values of all parameters* m_i, M_i.

We can obtain also *the explicit expression for optimal value of the parameter* M_0. Substitution of (34) into (37) leads to the relation

$$\begin{aligned} h_0 M_0 (n + 2 + H_2) = &-p_0(n + 2 + 2H_2) + 2\sum_{i=0}^{n} (-1)^{i+1} \frac{1}{h_i^2} p_i \\ &- 2\sum_{i=0}^{n+1} (-1)^i Q_i + 4\sum_{i=0}^{n} (-1)^{i+1} \frac{1}{h_i^2} Q_i \end{aligned} \tag{38}$$

with

$$H_2 = \sum_{i=0}^{n} \frac{1}{h_i^2}$$

and $m_i,\ M_i$ computed from (32).

The results of this section we can summarize in the following Theorem.

Theorem 3 *There always exists unique quadratric spline interpolating in given knots* **x** *prescribed values* **s** *with minimal* l_2*-norm of the vector* $[\mathbf{m}, \mathbf{M}]$*. These optimal parameters can be computed from the system (32) using pseudoinverse, or as the solution of the system (33) completed with (37). The optimal value of free parameter* M_0 *can be computed also from (38).*

5 Minimal norm of s(x)

The norms of the second or the first derivative (or of the curvature—see e.g. [6]) have immediate geometric or mechanic meaning and are frequently used in curve approximation. The norm of **s** is determined by the data given and is generally different from the L_2-norm of $s(x)$. When there is some need we can minimize also this norm using foregoing approaches. Using the local representation (3), we have

$$\begin{aligned} J_0(s) &= \|s(x)\|_2^2 = \int_{x_0}^{x_n} s^2(x)dx \\ &= \frac{1}{30}\sum_{i=0}^{n} h_i\Big[16s_i^2 + 6s_{i+1}^2 + 8s_is_{i+1} + h_im_i(7s_i + 3s_{i+1}) + h_i^2m_i^2\Big]. \end{aligned} \tag{39}$$

When we use the relation (22) to express the dependency of $J_0(s)$ on parameter m_0, the necessary condition for minimum can be written now as

$$0 = \frac{dJ_0}{dm_0} = \frac{1}{30}\sum_{i=0}^{n}(-1)^i h_i^2(7s_i + 3s_{i+1}) + \frac{2}{30}\sum_{i=0}^{n}(-1)^i h_i^3 m_i\,. \tag{40}$$

Substituting now from (5), we obtain for *the optimal value of the first derivative* the expression

$$m_0\sum_{i=0}^{n}\frac{1}{h_i^3} = \sum_{i=0}^{n}(-1)^{i+1}h_i^2\Big[7s_i + 3s_{i+1} + 2h_iQ_i\Big]. \tag{41}$$

The remainig values of m_i we can compute from recurrences (4). The whole *vector of optimal values* **m** we can compute also from the system of linear equations (4) completed with the condition of minima—we obtain the system with more easily computed elements

$$\begin{aligned} m_i + m_{i+1} &= 2p_i\,, \quad i = 0(1)n, \\ 2\sum_{i=0}^{n}(-1)^i h_i^3 m_i &= \sum_{i=0}^{n}(-1)^{i+1}h_i^2(7s_i + 3s_{i+1})\,. \end{aligned} \tag{42}$$

Theorem 4 *There exists the unique quadratic spline interpolating on the given knotset* **x** *the prescribed function values* **s** *with minimal value of the functional* $J_0(s)$*. The optimal value of its free parameter* m_0 *can be computed explicitly from (41), the whole vector of optimal parameters* **m** *alternatively from the system of equations (42).*

6 Derivative values interpolation problem

In case that in knots x_i the derivative values m_i are prescribed, there exists interpolating quadratic spline, which depends on one free parameter—let us choose it

as s_0. The continuity conditions for function values in knots we obtain from (4) (or using trapezoidal rule of numerical integration) as

$$s_{i+1} - s_i = \frac{1}{2} h_i (m_i + m_{i+1}) \,, \quad i = 0(1)n \,. \tag{43}$$

Applying here summation and then differentiation we obtain

$$s_k = s_0 + \sum_{i=0}^{k-1} \frac{1}{2} h_i (m_i + m_{i+1}) = s_0 + S_k \,; \qquad \frac{ds_k}{ds_0} = 1. \tag{44}$$

6.1 Minimal norm of vector s

We can compute *the whole vector of spline function values with minimal l_2-norm* using pseudoinverse for the solution of underdetermined system of equations (43).

Another possibility is to apply (44) and standart analytical approach to

$$J_{0d} = \sum_{i=0}^{n+1} s_i^2 \,; \qquad \frac{dJ_{0d}}{ds_0} = 2 \sum_{i=0}^{n+1} s_i \frac{ds_i}{ds_0} = 2 \sum_{i=0}^{n+1} s_i = 0. \tag{45}$$

The *optimal values* s_i we can obtain now also from the system of linear equations (43) completed with (45):

$$-s_i + s_{i+1} = \frac{1}{2} h_i (m_i + m_{i+1}); \; i = 0(1)n, \quad \sum_{i=0}^{n+1} s_i = 0. \tag{46}$$

We can obtain *the explicit expression for the optimal value of the free parameter* s_0 using summation in (44) and (45):

$$s_0 = -\frac{1}{n+2} \sum_{i=1}^{n+1} S_i \,, \quad S_i = \sum_{j=0}^{i-1} \frac{1}{2} h_j (m_j + m_{j+1}). \tag{47}$$

6.2 Minimal L_2-norm of s(x)

The L_2-norm of quadratic interpolatory spline with local parameters s_i, m_i we have presented in (39). With respect to the second relation in (44) we can express the condition of minima $dJ_0/ds_0 = 0$ after some calculations as

$$2h_0 s_0 + (h_0 + 2h_1) s_1 + (h_1 + 2h_2) s_2 + \cdots + (h_{n-1} + 2h_n) s_n + h_n s_{n+1} = -\frac{1}{2} \sum_{i=0}^{n} h_i^2 m_i \,. \tag{48}$$

When we complete the continuity conditions (43) with this condition, we obtain the system of linear equations with simple structure for computing optimal values s_i.

When we substitute (44) into (48) we obtain for *the optimal value of the free parameter* s_0 the expression

$$3(x_{n+1}-x_0)s_0 = -\sum_{i=1}^{n}(h_{i-1}+2h_i)S_i - h_nS_{n+1} - \frac{1}{2}\sum_{i=0}^{n}h_i^2 m_i\,. \tag{49}$$

Theorem 5 *The optimal value of the free parameter* s_0 *of the quadratic spline interpolating in the spline knots* $\mathbf{x}$ *the derivative values* $\mathbf{m}$ *is given by the formula (47) for minimum ot the* l_2*-norm and by the formula (49) for the minimum of the* L_2*-norm of that spline. The whole vector of optimal values* s_i *can be computed from recurrence (43), or from systems of equations (46), or (43) completed by (48).*

7 Another computing algorithms

7.1 Least squares approach

The spline continuity conditions—recurrences (4), (17) or (43) can be presented as the first order difference equation. The particular solution of such equation with any given starting value can be easily computed by forward or backward recursions. For the difference equation

$$a_i y_{i+1} = b_i y_i + c_i\,, \quad a_i \neq 0, \quad i = 0(1)n \tag{50}$$

let us denote $\mathbf{u}$ the particular solution of homogeneous equation with $u_0 = 1$ and $\mathbf{v}$ the particular solution of nonhomogeneous equation with $v_0 = 0$. Then the particular solutin of nonhomogeneous equation (49) with initial value y_0 we can write as $\mathbf{y} = y_0\mathbf{u} + \mathbf{v}$. (In case of constant coefficients $a_i = a$, $b_i = b$, $b/a > 1$ we have to compute backwards from stability reasons.) In the norm defined by some scalar product *the solution with minimal norm* is determined by initial value (see e.g. [5]).

$$y_0 = -(\mathbf{v},\mathbf{u})/(\mathbf{u},\mathbf{u}). \tag{51}$$

The value of $J_1(s)$ in (6) we can write as quadratic form

$$J_1(s) = \frac{1}{3}\sum_{i=0}^{n}(m_i^2 + m_i m_{i+1} + m_{i+1}^2) = \frac{1}{3}\mathbf{m^T R m} \tag{52}$$

with positive definite symmetric tridiagonal matrix $\mathbf{R}$,

$$diag(\mathbf{R}) = [h_0, h_0+h_1, \ldots, h_{n-1}+h_n, h_n]\,, \tag{53}$$

and subdiagonal given as

$$\frac{1}{2}[h_0, h_1, \ldots, h_n]\,.$$

When we denote

$$(\mathbf{y},\mathbf{y})_R = \frac{1}{3}\mathbf{y^T R y}$$

the scalar product corresponding to matrix $\mathbf{R}$, then we can apply (51) for computation of corresponding optimal value m_0 in this case.

Similarly we can present the functionals $J_{1d}(s), J_2(s), J_{2d}(s), J_{0d}(s)$ as scalar products defined by (symmetric positive definite) identity matrices $\mathbf{I}$ or diagonal matrix $\mathbf{D} = \mathrm{diag}[h_i]$. The optimal values of corresponding free parameters can be computed again folowing the formula (51). The remaining values can be then computed from corresponding recurrence formula. We summarize the described approach in the following algorithm.

Algorithm s2optlsq: Let us have given data **x, s** *and functional minimized.*

1. *Choose corresponding recurrences (continuity conditions).*
2. *Find particular solutions* **u, v** *of homogeneous and nonhomogeneous difference equation (computed with forwards or backwards recurrences).*
3. *Compute components of corresponding matrix* $\mathbf{R}$*.*
4. *Compute optimal initial value (51) of local parameter chosen.*
5. *Compute all local parameters needed (from recursions).*

7.2 Quadratic programming approach

The functional $J_0(s)$ defined in (39) we can write in matrix form as

$$J_0(s) = C + \mathbf{q^T m} + \mathbf{m^T D_1 m} \tag{54}$$

with constant C, vector $\mathbf{q}$ and diagonal matrix $\mathbf{D_1}$ which can be recognized from (39). The problem

$$C + \mathbf{b^T m} + \mathbf{m^T D_1 m} \longrightarrow \ min\ ; \quad m_i + m_{i+1} = 2p_i\ ,\ i = 0(1)n \tag{55}$$

can be solved by standard algorithms of quadratic programming with equality constrains.

Similarly can be all problems mentioned in this article stated as simple quadratic programming problems with equality constrains.

8 Examples

Some of algorithms mentioned above were implemented in MATLAB function **[m,M] =s2xs(x,s,par,k,iv)**, which for given input data

```
-  x  ... vector of spline knots ( = points ofinterpolation)
-  s  ... vector of prescribed function values in knots
-  par ...control parameter for choice from variants
     par==1 ... [k,iv]=[knot index, s'(x(k)) prescribed]
     par==2 ... [k,iv]=[knot index, s''(x(k)) prescribed]
     par==3 ... [k,iv]=[], min l2-norm(m)   computed
     par==4 ... [k,iv]=[], min L2-norm(s')  computed
     par==5 ... [k,iv]=[], min L2-norm(s'') computed
     par==6 ... [k,iv]=[], min L2-norm(s)   computed
     par==7 ... [k,iv]=[], min l2-norm(M)   computed
     par==8 ... [k,iv]=[], min l2-norm([m,M]) computed
```

computes the vectors **m, M** of local parameters of the interpolatory quadratic spline. This function was used in the following examples.

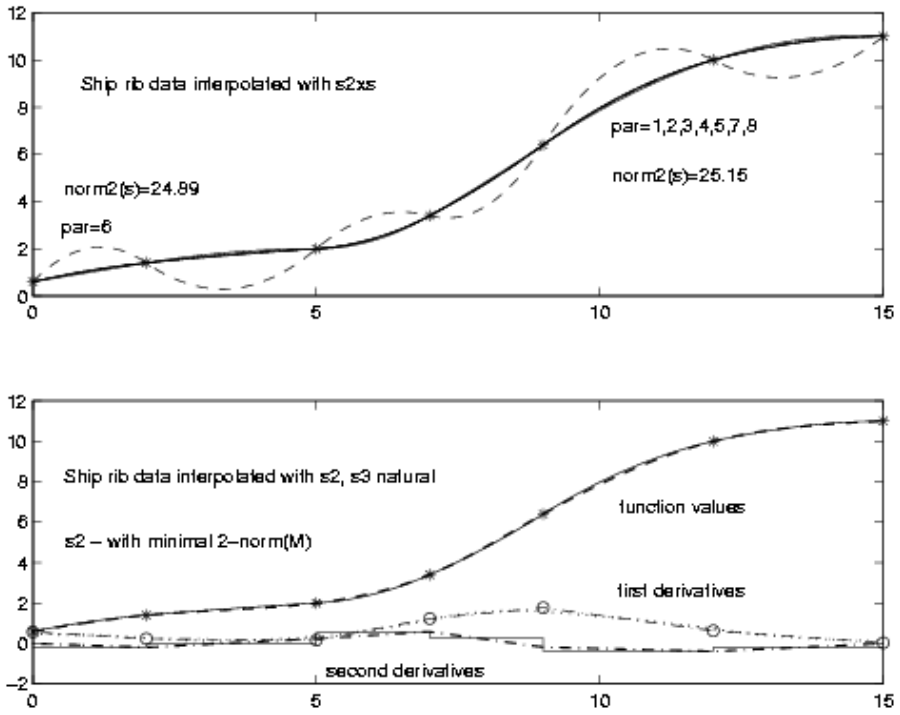


Fig. 1 a, b

Example 1 Let us have monotone ship rib data [6]

```
x=[0 2 5 7 9 12 15];  s=[0.6 1.4 2 3.4 6.4 10 11].
```

The results of interpolating this data with **s2xs** for values of control parameters

par=1 ($k = 1, m_1 = 0.4$), par=2 ($k = 1, M_1 = -0.1$), par=3,4,5,7,8

are very near to the result of interpolation with natural cubic spline (function **csape** from *Spline Toolbox* used) and give the splines with L_2-norm about 25.15. Quite different is the result for *par=6* with L_2-norm about 24.9—a lot smaller (Fig. 1a).

We can also see on Fig. 1b plots of the first and second derivatives of natural cubic and quadratic spline ($par = 7$) here with nice correspondence.

Example 2 For the frequently tested type of "staircase" monotone data

```
x = [0    1   2  2.5  2.7  2.9  3.4   4   5   6],
s = [0.3 0.4 0.7 1.5  3.5  5.5  6.2  6.5 6.7 6.8]
```

the results of interpolation with natural cubic spline (monotonicity not preserved in the neighborhood of rapid changes of the first derivative) is shown in Fig. 2. The oscilations we can see on plots corresponding to interpolatory quadratic splines corresponding to $p = 3$, $p = 5$, $p = 7$. Much better result we obtain when we minimize the discrete approximation of L_2-norm of the curvature with quadratic spline (first derivative approximated from data)—see the full line on Fig. 2.

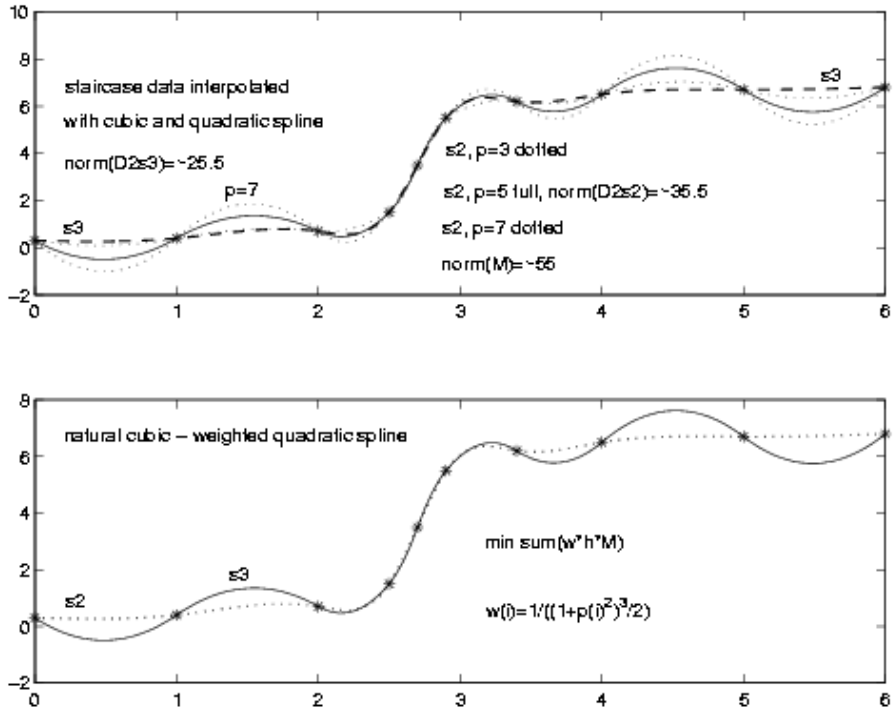


Fig. 2 a, b

Example 3 For another monotone data on equidistant knotset $\mathbf{x} = 0{:}1{:}20$,

```
s=[1 2 5 6 7 12 15 22 24 25 35 36 37 45 47 48 55 56 58 59 60]
```

still the natural cubic spline preserves better monotonicity than quadratic spline—see Fig. 3, where also plots of first and second derivatives are given and rounded values of some norms.

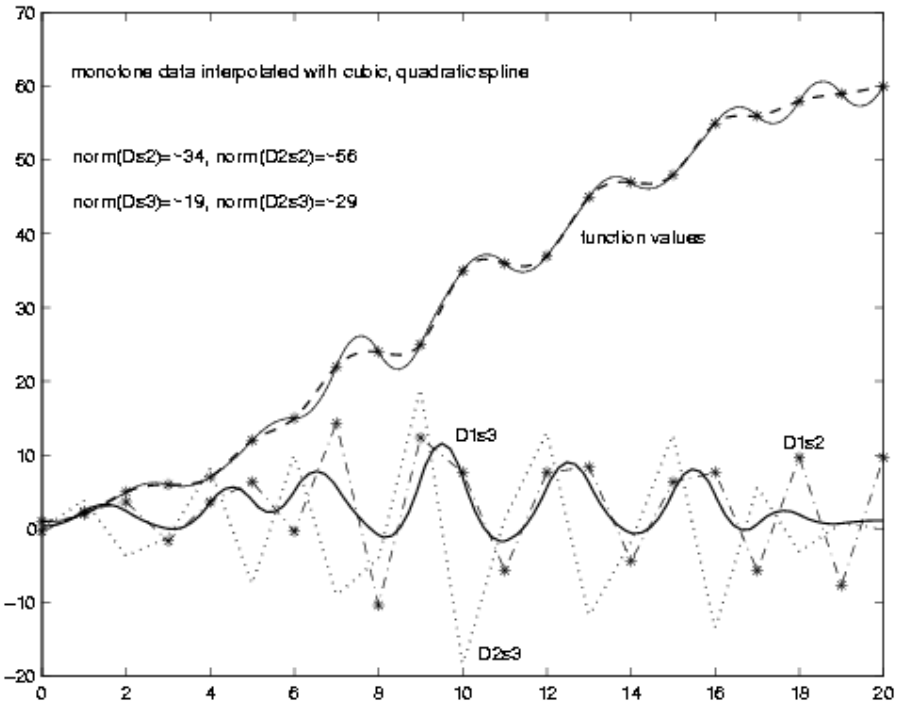


Fig. 3

Example 4 Results of interpolation on more general data on equidistant knotset $\mathbf{x} = 0{:}1{:}20$,

```
s=[15 11 3 5 0 -2 -7 -1 6 10 12 16 19 17 13 12 8 6 4 1 0]
```

are plotted on Fig. 4—we can again see the better results with natural cubic spline (L_2-norms of the first, second derivative equal about 44, 29) than with quadratic spline (norms 46, 112), due substantially from oscilations near boundaries. We obtain again better results for quadratic spline using mentioned discrete approximation of the norm of curvature (equal about 29 for cubic, 30 for quadratic spline—but the oscilations on the left still have place for quadratic spline).

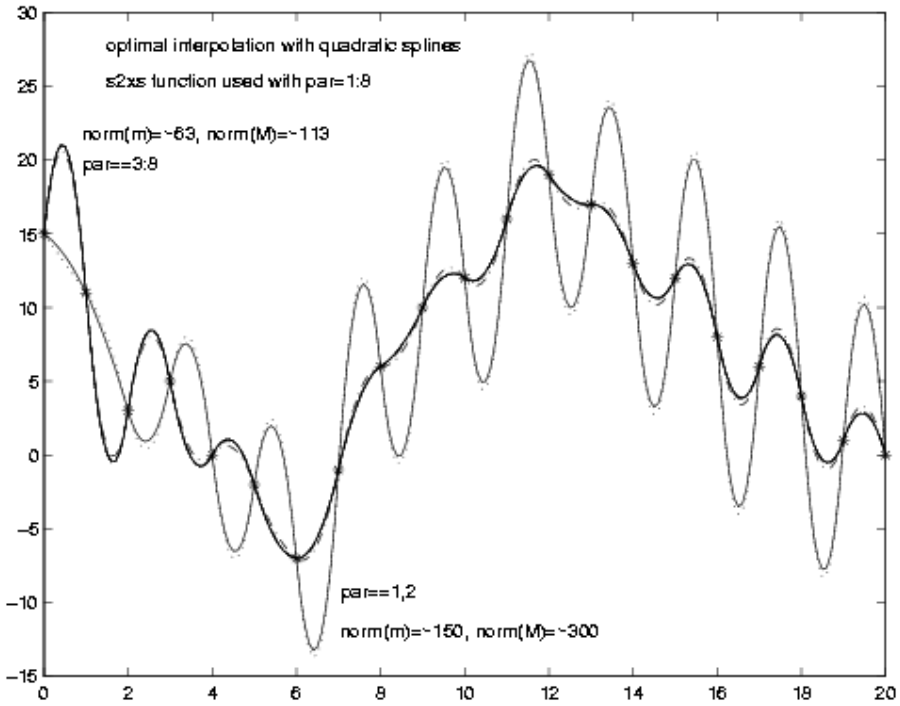


Fig. 4

Reference

[1] Bjorck, A.: *Numerical Methods for Least Squares Problems.* SIAM, PA 1996.

[2] De Boor, C.: *A Practical Guide to Splines.* Springer, 1978.

[3] Kobza, J.: *Splajny.* UP Publ., Olomouc, 1993 (textbook in Czech).

[4] Kobza, J.: *Quadratic and Quartic Splines in MATLAB.* Folia Fac. Sci. Nat. Univ. Masaryk. Brunensis, Mathematica **5** (1997), 47–65.

[5] Kobza, J.: *Optimal polygonal interpolation.* Preprint Series, Dept. MAAM, Fac. Sci. UP Olomouc, 32/1998.

[6] Maess, B., Maess, G, *Interpolating quadratic splines with norm-minimal curvature.* Rostock. Math. Kolloq. **26** (1984), 83–88.


Univ. Palacki. Olomuc., Fac. rer. nat.,
Dept of Math. Anal. and Appl. of Math.
ODAM (1999) 21–33

Poznámky k aplikacím funkcionální analýzy v problémech s hyperelastickými materiály

Milan KONEČNÝ

Department of Mathematics, University of Ostrava,
Brafová 7, Czech Republic,
e-mail: konecny@osu.cz

Klíčová slova: Nelineární funkcionální analýza, variační počet, optimalizace, slabá spojitost, SC-lokální slabá spojitost, SC-rozkladatelný funkcionál, nelineární elasticita, hyperelasticita.

1991 Mathematics Subject Classification: 47H30

1 Úvod

Co je cílem toho příspěvku? Článek je víceúčelový. Chtěl by jsem připomenout osobnost Doc. RNDr. Svatopluka Fučíka, CSc, který zemřel právě před dvaceti lety a zanechal po sobě jistý odkaz, který se snažím plnit. Dále by jsem se chtěl zamyslet nad způsoby tvorby matematiky a jistými aplikacemi. A samozřejmě by jsem chtěl ukázat některé výsledky z nelineární funkcionální analýzy a jejich aplikace.

2 Úvahy o matematice

V této části článku se trochu zamyslím nad řešením matematických problémů. V následující sekci se zabývám větou, která mne k tomu motivovala.

2.1 O poslední Svatoplukově větě, jejim důkazu, zobecnění a jeho odkazu

Doc. RNDr. Svatopluk Fučík, CSc byla jedna z nejvýznamějších osobností v oblasti české matematiky v 20 století. Je to neoddiskutovatelné. Pracoval v oblasti funkcionální analýzy a aplikací v teorii obyčejných a parciálních diferenciálních rovnic. Napsal přes 50 článků a podílel se na pěti knihách. Je to neuvěřitelné, protože vše stačil udělat do svých nedožitých 35 let. Navíc byl vynikající pedagog, a samozřejmě i organizátor. Nevím jak se klasifikují žáci, tím myslím, kdy se člověk může považovat za žáka, ale věřím, že jeho žák jsem i když možná ten nejhorší. Ale asi jsem zase ten nejžakovatější. Proč? Jistě neznámějším a nejúspěšnějším žákem doc. RNDr. Svatopluka Fučíka, CSc. je prof. RNDr. Pavel Drábek, DrSc, ale ten u něho neabsolvoval, žádnou přednášku a také jim nebyl zkoušen. On totiž je tak dobrý, že ho nemuseli ani zkoušet. Což se mi nepodařilo. Mne zkoušet museli. Ale já byl první diplomant doc. RNDr. Svatopluka Fučíka, CSc., také jsem byl jeho pomvěd, dělal jsem pod ním rigorózum a málem jsem u něho dělal aspiranturu. Plán aspiranský jsem měl. Bohužel strana a vláda to v té době nedovolila a já tím pádem nastoupil krkolomnou cestu za tituly. A na té cestě poznávám jaký byl doc. RNDr. Svatopluk Fučík, CSc vynikající pedagog a školitel. Jistě kdyby žil, by měl jednu z nejvýznamnělších matematických škol.

A teď k té jeho poslední větě. Svatopluk, jak mu i my jeho žáci někdy říkáme, i když jsme si s nim nikdy netykali, měl snahu o vazbu na praxi. Sice trochu idealistickou, ale to bylo jeho mládím a také dobou. Počítače ještě pořádně nepracovaly. Přesto v hloubi duše určitě moc pošilhával i po konkrétních aplikacích. V dubnu 1979 byl seminář v Příhrazech, byl jsem tam dokonce i já. Na seminář mě pustili omylem, v té době jsem působil ještě na VŠB, ale už to odeznívalo. Chtěl jsem odejít do praxe. Když se řekne praxe pro matematika, tak to znamená výzkumný ústav a tam jsem měl také namířeno. Na semináři mi Svatoplukovi kolegové řekli, že to s ním je už špatné, ale že ho mohu jít navštívit do nemocnice, která je na Karlově náměstí.Ležel tam na intenzivce. Když jsem se vracel ze semináře, tak jsem přes víkend zůstal v Praze a šel jsem navštívit pana docenta. Moc to nechci popisovat. Byl to pro mne jeden z nejsilnějších zážitků života. V ne zrovna v nejlepším rozpoložení jsem mu řekl: „Pane docente rozhodl jsem se že půjdu do praxe.“ A ted přišla ta poslední Svatoplukova věta: „*Jo v praxi to je těžký!!!!*“ Rozloučili jsme se a já jsem rozrušen odešel. Asi za dva týdny jsem se dozvěděl, že doc. RNDr. Svatopluk Fučík, CSc. zemřel. Odešel jsem skutečně do výzkumáku a později jsem dokonce pracoval v konstrukci nákladních automobilů. A tam jsem našel ten důkaz toho, že v praxi je to těžké. Skutečně se to potvrdilo. A jaké je zobecnění této věty? To už jsem udělal já: „*Jo v praxi to je těžký, ale hezký!*“ A jak je to s tím odkazem. Založil jsem dva semináře, Svatoplukovo centrum, Svatoplukovo gymnázium, Matematickou soutěž pro žáky devítiletek a Matematickou soutěž pro vysokoškoláky. Ze Svatoplukova centra v Brušperku se stalo kulturně vzdělavací centrum, které svými aktivitami má dosah po celé republice.

Vše je celkem uspěšné, až na moji osobní vědeckou činnost, tam ten odkaz moc neplním a je to prozatím velmi slabé. Ale snad se to už napravuje.

2.2 Jak se dělá matematika — dva druhy matematiků

Je více přístupu k matematice. Někteří tvrdí, že matematika se dělá v uzavřeném bílém pokoji. No moc se mi to nezdá. Ti druzí zase vychází z praxe, tedy pozorováním realného světa. Jsou asi ještě jiné způsoby. Vše by se mělo nějak kombinovat. A asi to tak je, jak z následujícího uvidíte. Já jsem v poslední době spíše orientován na motivaci z praxe. I když od motivace z praxe přejít až k teoretickým výsledkům je těžké a někdy závěrečné formulace jsou značně pozměněné od původních formulací! Však to sami uvidíte.

3 Od technického problému, až k teoretickým výsledkům

V následující části článku by jsem chtěl ukázat vývoj jistých abstraktních výsledků, které vyšly z praxe. Zpočátku je velmi obecná úloha, formulovaná z praxe, která se pozvolna zjednodušuje, aby jsme ji mohli vyřešit. Nakonec po řadě zjednodušení dostaneme nějaké výsledky, které ale přestanou mít fyzikální význam a vazbu na původní úlohu.

3.1 Problém praxe

Jedno ze zajímavých období mého života, bylo období kdy jsem byl v konstrukci nákladních automobilů v Tatře Kopřivnici. Na autě bylo mnoho problémů, které většinou směřovaly k jistému druhu tvarové-materiálové optimalizaci. V té době se mi dostala do rukou zajímavá součást „Gumový blok“. Položil jsem si otázku: Jaké parametry a jaký tvar by měl mít „Gumový blok“, aby se při zatíženích nedotkl krabice, ve kterou byla součást ohraničena, pro danou třídu zatížení? Jistě — je to *úloha materiálově-tvarové optimalizace vzhledem k třídě zatížení s přepážkou*. Tedy hledám materiál a tvar pryžové součásti, aby při zatížení posuvy nebyly moc veliké a aby to pružilo.

3.2 Zjednodušení problému

V rámci řešitelnosti problému úlohu zjednodušíme. Nebudeme řešit problém materiálově tvarové optimalizace s přepážkou, nýbrž budeme řešit problém — optimalizace vzhledem k zatížení. Tedy budeme hledat zatížení, při kterém jsou největší posuvy a zjistíme zda dané posuvy jsou tak veliké, že se dané těleso dotkne krabice.

3.3 Fyzikální formulace problému

V následujícím naformulujeme výše uvedený problém pomocí pojmů matematické analýzy a parciálních diferenciálních rovnic. V prvé části je formulován problém a v druhé jsou specifikované pojmy a předpoklady za kterých má daná formulace význam.

Problém POF — optimalizace vzhledem k silovým podmínkám

Nechť $\mathbf{M_0} \subset \mathbf{L}_\mathrm{p}(\Omega)$ představuje množinu vnitřních sil a $\mathbf{M_1} \subset \mathbf{C}(\Gamma_1)$ množinu vnějších sil.

Najdi vnitřní silu $\mathbf{f_0} \in \mathbf{M_0} \subset \mathbf{L}_\mathrm{p}(\Omega)$ a vnější sílu $\mathbf{g_0} \in \mathbf{M_1} \subset \mathbf{C}(\Gamma_1)$ tak, že

$$||\mathbf{u}_{\mathbf{f_0},\mathbf{g_0}}|| = \sup ||\mathbf{u}_{\mathbf{f},\mathbf{g}}||, \quad \mathbf{f} \in \mathbf{M_0} \text{ a } \mathbf{g} \in \mathbf{M_1},$$

kde $\mathbf{u}_{\mathbf{f},\mathbf{g}}$ jsou posuvy tělesa pro dané pro silové podmínky $\mathbf{f}, \mathbf{g}$, tedy řeší níže uvedenou parciální diferenciální rovnici.

$$\begin{cases} -\mathbf{div}\{(\mathbf{I} + \nabla \mathbf{u}_{\mathbf{f},\mathbf{g}})\check{\mathbf{\Sigma}}(\mathbf{E}(\mathbf{u}_{\mathbf{f},\mathbf{g}}))\} = \mathbf{f} & \text{v } \mathbf{L}_\mathrm{p}(\Omega) \\ \mathbf{u}_{\mathbf{f},\mathbf{g}} = \mathbf{u}_0 \quad \text{na } \Gamma_0 \\ (\mathbf{I} + \nabla \mathbf{u}_{\mathbf{f},\mathbf{g}})\check{\mathbf{\Sigma}}(\mathbf{E}(\mathbf{u}_{\mathbf{f},\mathbf{g}}))\mathbf{n} = \mathbf{g} \quad \text{na } \Gamma_1 \end{cases}$$

kde $\mathbf{f} \in \mathbf{L}_\mathrm{p}(\Omega)$, $\mathbf{u}_0 \in \mathbf{C}(\Gamma_0)$, $\mathbf{g} \in \mathbf{C}(\Gamma_1)$.

3.4 Další zjednodušení

Takto formulovanou úlohu neumíme vyřešit, alespoň doposud. Budeme muset provést další zjednodušení a to tak aby jsme mohli dokázat existenci řešení problému. Co to bude mít za důsledek? Začneme ztrácet fyzikální význam a začneme se dostávat do bílého pokoje. Ale s tím, že myšlenky a problém jsme si ale přinesli z venku. Aby jsme mohli použít větu o lokálním difeomorfismu musíme uvažovat speciální případ, kdy $\Gamma_0 = \partial\Omega, \mathrm{u}_0 = 0$ — tedy případ, v kterém těleso je zatíženo pouze vnitřními silami a posuvy jsou nulové na hranici *POVFO-problém*. Právě tímto problémem se budeme v následujícím zabývat. Takto definovaný problém je ovšem vzdálen fyzikální realitě, přesto jeho řešení ukazuje jisté postupy, které mohou být dále rozvíjeny.

3.5 Předpoklady fyzikální formulace problému

Abychom mohli korektně definovat výše uvedený problém musíme předpokládat následující:

Předpoklady MFPP

Značení i předpoklady jsou převzány z [1].

1. Nechť $\mathrm{p} > 3, \Omega \subset \mathbf{R}^3, \partial\Omega \in C^2$, $\mathbf{S^3}$ — je prostor symetrických matic 3×3. Označme $\mathbf{V}^{\mathrm{p}} = \mathbf{W}^{2,\mathrm{p}}(\Omega) \bigcap \mathbf{W}_0^{1,\mathrm{p}}(\Omega) = (W^{2,p}(\Omega))^3 \bigcap (W_0^{1,p}(\Omega))^3$, $\mathbf{L}_{\mathrm{p}}(\Omega) = (L_{\mathrm{p}}(\Omega))^3$, $\mathbf{C}(\overline{\Omega}) = (C(\overline{\Omega}))^{3x3}$, $\mathbf{L}_1(\Omega) = (L_1(\Omega))^3$.
Na prostorech se používají standartní normy viz [1]

2. Při modelování problému budeme používat

   a) Green–Saint–Venant tenzor deformace $\mathbf{E}$, pro který platí

   $$\mathbf{E}(\mathbf{u}) = \frac{\mathbf{1}}{\mathbf{2}}(\nabla\mathbf{u}^\top + \nabla\mathbf{u} + \nabla\mathbf{u}^\top \cdot \nabla\mathbf{u}), \tag{1}$$

   kde $\mathbf{u}$ je vektor posunutí.

   b) Druhý Piola–Kirchhoff tenzor napětí $\mathbf{\Sigma}$.

   c) Platnost rovnice rovnováhy, která má následující tvar:

   $$-\mathbf{div}\{(\mathbf{I} + \nabla\mathbf{u})\check{\mathbf{\Sigma}}(\mathbf{E}(\mathbf{u}))\} = \mathbf{f} \quad \text{na } \Omega \tag{2}$$

3. Vlastnosti tělesa

   a) Těleso je reprezentováno omezenou oblastí $\Omega \subset \mathbf{R}_3$, s hranicí $\partial\Omega = \Gamma$, $\partial\Omega \in C^{2+k}$, kde $k \geq 0$.

   b) Těleso Ω je z homogenního, izotropního materiálu.

   c) Funkce odezvy

   $$\mathbf{\Sigma} = \check{\mathbf{\Sigma}}(\mathbf{E}), \tag{3}$$

   která realizuje vztah mezi Green–Saint–Venant tenzorem deformace $\mathbf{E}$ a druhým Piola–Kirchhoff tenzorem napětí $\mathbf{\Sigma}$, má následující vlastnosti:

   i) Zobrazuje

   $$\check{\Sigma} : \mathbf{W}(\mathbf{0}) \subset \mathbf{S}^3 \to \mathbf{S}^3,$$

   kde $\mathbf{W}(\mathbf{0})$ je otevřené okolí v $\mathbf{S}^3$.

   ii) Vztah mezi tenzorem deformace a tenzorem napětí je charakterizován rovnicí

   $$\mathbf{\Sigma} = \check{\mathbf{\Sigma}}(\mathbf{E}) = \lambda(tr\mathbf{E})I \ + 2\mu\mathbf{E} + O(\|\mathbf{E}\|^2_{\mathbf{S}^3}), \tag{4}$$

   kde $\lambda > 0$, $\mu > 0$ jsou tzv. *Lamého konstanty.*

   iii) Těleso je v tzv. *přirozeném stavu*, což funkce odezvy modeluje podmínkou

   $$\mathbf{0} = \check{\Sigma}(\mathbf{E}(\mathbf{0})). \tag{5}$$

iv) Funkce odezvy je minimálně dvakrát spojitě diferencovatelná

$$\check{\mathbf{\Sigma}} \in C^2. \tag{6}$$

4. Těleso je zatíženo silou $\mathbf{f} \in \mathbf{L}_{\mathrm{p}}(\Omega)$

5. Vlastnosti okrajových podmínek

   a) Geometrické podmínky reprezentované předepsanými posuvy $\mathbf{u}_0$ na $\Gamma_0 \subset \partial\Omega$.

   b) Fyzikální podmínky — reprezentované předepsaným napětím $\mathbf{g}$ na $\Gamma_1 \subset \partial\Omega$,

   kde platí $\Gamma_0 \cap \Gamma_1 = 0$ a $\mathbf{u}_0$, $\mathbf{g}$ jsou dostatečně hladké.

Poznámky:

1. Předpoklady MFPP — předpoklady Matematické formulace problému pružnosti.

2. Předpoklad hladkosti hranice $\partial\Omega \in C^2$ je podstatný pro regularitu linearizovaného problému.

3. Existence $\mathbf{W}(\mathbf{0})$ a jeho tvar viz věta (1.8-3) [1] a Dodatek: Lemma 2.

4. Charakter funkce odezvy pro homogenní izotropní materiály mezi tenzorem deformace a tenzorem napětí (4) je opodstatněn větou (3.8-1) [1], viz Dodatek: Lemma 3, kde je ve tvaru

$$\mathbf{\Sigma} = \check{\mathbf{\Sigma}}(\mathbf{E}) = \lambda(tr\mathbf{E})I \ + 2\mu\mathbf{E} + o(\|\mathbf{E}\|_{\mathbf{S}^3}), \tag{7}$$

   který je slabší než (4).

5. Existují jednoduché vztahy mezi Lamého konstantami λ, μ a Poissonovým číslem E a Youngovým modulem pružnosti ν, které se v praxi užívají častěji.

6. Předpoklad přirozeného stavu (5) fyzikálně říká, že těleso v nezatíženém stavu má nulové napětí. A realizuje nám podmínku nulového řešení pro nulovou pravou stranu pro operátor problému

$$\mathbf{A}(\mathbf{0}) = \mathbf{0}. \tag{8}$$

7. Podmínka diferencovatelnosti (6) je podstatná pro diferencovatelnost operátoru elasticity.

8. Charakter funkce odezvy (4) a jeji diferencovatelnost (6) implikují přirozenou podmínku (5).

9. Důvodem používání Green–Saint–Venant tenzoru a nelineárního vztahu k tenzoru napětí je potřeba zpřesnit matematický model pružnosti, a tedy řešit matematicky nelineární problém.

10. Problém je nelineární ze dvou důvodů

    i) použitím Green–Saint–Venant tenzoru deformace, který má nelineární charakter, obsahuje součiny derivací,

    ii) funkčním vztahem mezi Green–Saint–Venant tenzorem deformace a druhým Piola–Kirchhoffovým tenzorem napětí, který je realizován nelineární funkcí odezvy.

11. Rovnost pro rovnici rovnováhy je myšlena ve smyslu distribucí. Rovnost pro okrajové podmínky je myšlena ve smyslu stop. Pod pojmem derivace myslíme distributivní derivaci.

12. Tato formulace je slabší, než formulace „klasická“, která je definována v prostorech spojitých funkcí a silnější než tzv „slabá“ formulace, která vznikne z formulace „distributivní“ pomocí Greenovy věty a někdy bývá nesprávně nazývaná „principem virtuálních prací“.

13. Musí se dokázat korektnost zadání tohoto problému. To znamená, že diferenciální operátor, zobrazuje prostor $\mathbf{W}^{2,\mathrm{p}}(\Omega)$ do prostoru $\mathbf{L}_{\mathrm{p}}(\Omega)$

14. Důvodem pro nerovnost $p > 3$ je, aby součin dvou funkcí z $W^{2,p}(\Omega)$ se nacházel opětně v $W^{2,p}(\Omega)$, tedy aby daný prostor byl Banachová algebra viz [1].

3.6 Formulace problému pomocí pojmů funkcionální analýzy

Nyní se budem zabývat abstraktním popisem výše uvedených problémů pomocí pojmů funkcionální analýzy.

V následujícím budeme značit X, Y, X_0, Z normované prostory.

Problém POF — optimalizace vzhledem k silovým podmínkám

Je ho možno charakterizovat z hlediska funkcionální analýzy jako hledání extrémů funkcionálu na množině, která je charakterizována dvěmi vazbovými podmínkami. Nechť je dán funkcionál

$$J : Dom[J] \subset X \to \mathbf{R}$$

a vazbové podmínky

$$Au = f, f \in M_0 \subset Y, A : Dom[A] \subset X \to Y$$

$$Tu = g, g \in M_1 \subset Z, T : Dom[T] \subset X \to Z,$$

Problémem je najít $f_0 \in M_0$, $g_0 \in M_1$, $f_1 \in M_0$, $g_1 \in M_1$ tak, že

$$J(u_{f_0,g_0}) = \inf_{Au_{f,g}=f, Tu=g} J(u_{f,g}), f \in M_0, g \in M_1,$$
$$J(u_{f_1,g_1}) = \sup_{Au_{f,g}=f, Tu=g} J(u_{f,g}), f \in M_0, g \in M_1,$$

kde $M_0 \subset Y, M_1 \subset Z$.

Poznámky:

1. Operátor A představuje rovnici rovnováhy, tedy vztah mezi zatížením vnitřními silami a posuvy.
2. Operátor T realizuje okrajové podmínky — silové i deformační.
3. Problém by se také dal zapsat jako problém s třemi podmínkami. Operátor, který realizuje okrajové podmínky se vlastně zkládá ze dvou operátorů.
4. Takto formulovaný problém prozatím nedovedeme vyřešit. Bude předmětem dalšího zkoumaní.

Abychom docílili alespoň jistých teoretických výsledků, budeme se zabývat následujícím problémem.

Problém POVFO — optimalizace vzhledem vnitřním silám, za předpokladu nulových posuvů na hranici

Tento problém je možno charakterizovat, jako hledání extrémů na množině, která je charakterizována pouze jednou vazbovou podmínkou. Podmínka nulových posuvů je zahrnuta v definici prostoru posuvů.

Nechť je dán funkcionál

$$J : Dom[J] \subset X_0 \to \mathbf{R}$$

a vazbova podmínka

$$A : Dom[A] \subset X_0 \to Y,$$

Problémem je nalezení $f_0 \in M$, $f_1 \in M$ tak, že

$$J(u_{f_0}) = \inf_{Au_f=f} J(u_f), f \in M,$$
$$J(u_{f_1}) = \sup_{Au_f=f} J(u_f), f \in M,$$

kde $M \subset Y$.

Problém POVFO je vlastně variantou následujícího ryze teoretického formulovaného problému:

Problém OPS — optimalizace vzhledem k pravým stranám

Nechť $A : Dom[A] \subset X \to Y$ je operátor a funkcionál $J : Dom[J] \subset X \to \mathrm{R}$ a $M \subset Y$.

Najdi $f_0 \in M$, $f_1 \in M$ tak, že

$$J(u_{f_0}) = \inf_{Au_f = f} J(u_f), f \in M,$$
$$J(u_{f_1}) = \sup_{Au_f = f} J(u_f), f \in M.$$

Poznámky:

1. Pro $f \in M$ může nastat, že rovnice $Au = f$ nemá řešení, má konečný počet řešení a má nekonečně mnoho řešení.
2. *POF-problém* je obecnější než *OPS-problém* navíc se dá říci v jistém smyslu, že *POVFO-problém* je variantou *OPS-problému*.

Lema 1 *Nechť je dán* OPS-problém *a platí, že existuje inverzní operátor* $A^{-1} : M \subset Y \to X$, *potom* OPS-problém *je eqvivalentní s následujícím problémem:*

Najdi $f_0 \in M$, $f_1 \in M$ *tak, že*

$$J(A^{-1}(f_0)) = \inf_{f \in M} J(A^{-1}(f))$$
$$J(A^{-1}(f_1)) = \sup_{f \in M} J(A^{-1}(f))$$

kde $M \subset Y$, *tedy nalezení extrémů funkcionálu* $I = J \circ A^{-1}$ *na* $M \subset Y$.

Důkaz: Jednoduchá úvaha. □

4 Abstraktní existenční výsledek

V následující sekci se budeme zabývat abstraktním výsledkem pro existenci minima a maxima funkcionálu na podmnožinách Banachových prostorů. Základem bude věta o existenci extrému na slabě sequenciálně kompaktních množinách v Banachových prostorech slabě sequenciálně spojitých funkcionalů. Zavedeme pojmy *SC-rozkladatelného funkcionálu* a *SC-slabé lokálni spojitosti v bodě*. Hlavním výsledkem je věta o existenci extrémů pro SC-rozkladatelné funkcionály na uzavřené kouli v reflexivním Banachově prostoru. Tuto větu budeme aplikovat v další sekci na problémy s hyperelastickými materiály.

V úvodu zavedeme několik značení a definic. Nechť X, Y, Z jsou Banachovy prostory X^* spojitý duál a dále označme $B_X^r(x_\circ) = \{x \in X, \|x - x_\circ\|_X < r$ a $\overline{B_X^r(x_\circ)}$ je uzavěr k $B_X^r(x_\circ)$ v X. Nechť $A : Dom[A] \subset X \to Y$ operátor zobrazující X do Y s definičním oborem $Dom[A]$. Silnou konvergenci budeme značit v $X(\|u_n - u\|_X \to 0)$ jako $u_n \to u$ v X, slabá konvergence v $X(\langle b, u_n - u\rangle \to 0, \forall b \in X^*)$ jako $u_n \rightharpoonup u$ v X.

Definice 1 (spojitosti definované pomoci konvergence)
Operátor $A : Dom[A] \subset X \to Y$ je

— *spojitý na* $M \subset Dom[A]$

$$\forall(\{x_n\}_{n=1}^{\infty} \cup \{x\} \subset M)(x_n \to x \text{ v } X \Rightarrow A\ x_n \to Ax \text{ v } Y)$$

— *zesíleně spojitý na* $M \subset Dom[A]$

$$\forall(\{x_n\}_{n=1}^{\infty} \cup \{x\} \subset M)(x_n \rightharpoonup x \text{ v } X \Rightarrow A\ x_n \to Ax \text{ v } Y)$$

— *slabě spojitý na* $M \subset Dom[A]$

$$\forall(\{x_n\}_{n=1}^{\infty} \cup \{x\} \subset M)(x_n \rightharpoonup x \text{ v } X \Rightarrow A\ x_n \rightharpoonup Ax \text{ v } Y)$$

— *SC lokálně slabě spojitý v bodě* $x_\circ \in X$ Existují $r, s > 0$ tak, že

$$A : \overline{B_X^r(x_\circ)} \subset X \to \overline{B_Y^s(Ax_\circ)} \subset Y$$

splňují

$$\forall(\{x_n\}_{n=1}^{\infty} \cup \{x\} \subset \overline{B_X^r(x_\circ)}) \quad (x_n \rightharpoonup x \text{ v } X \Rightarrow Ax_n \rightharpoonup Ax \text{ v } Y)$$

Poznámky:V Banachových prostorech slabá spojitost a slabá sequencionální spojitost je eqvivalentní.

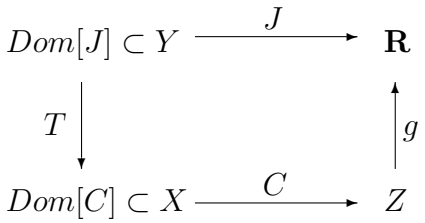
Definice 2 (SC-rozkladatelný funkcionál) Nechť $(Y, ||\ ||_Y)$ je Banachův prostor a J je funkcionál definovaný na $Dom[J] \subset Y$. Řekneme, že J je *SC-rozkladatelný* (má SC-rozklad), pokud platí:
Existují Banachovy prostory $(X, ||\ ||_X)$ a $(Z, ||\ ||_Z)$ a T, C, g
$T : Dom[T] \subset Y \to X$, $Dom[I] \subset Dom[T]$
$C : Dom[C] \subset X \to Z$
$g : Z \to \mathbf{R}$
s následujícími vlastnostmi:
T — slabě spojité zobrazení $Rang[T] \subset Dom[C]$
C — zesíleně spojité zobrazení
g — spojitý funkcionál
a platí rozklad:

$$J = g \circ C \circ T \text{ na } Dom[J]$$

a tedy následující graf je komutativní.

Věta 1 *Každý SC-rozkladatelný funkcionál je slabě spojitý.*

Důkaz: Plyne z definice rozkladu. □



Věta 2 (Existence extrémů SC-rozkladatelného funkcionálu) *Nechť* $(Y, ||\ ||_Y)$ *je Banachův prostor. Nechť* $J : Dom[J] \subset Y \to \mathbf{R}$ *má SC-rozklad a nechť* $M \subset Dom[J]$ *je slabě kompaktní množina v* $(Y, ||\ ||_Y)$*. Potom existuje* $y_0, y_1 \in M$ *tak, že platí:*

$$\inf_{y\in M} J(y) = J(y_0), \qquad \sup_{y\in M} J(y) = J(y_1).$$

Důkaz: Z toho, že J je SC-rozkladatelný plyne, že J je slabě spojitý. Jelikož navíc M je slabě kompaktní můžeme aplikovat větu o existenci extrémů pro slabě spojité funkcionály na slabě kompaktních množinách. Z toho plyne výše uvedené tvrzení. □

Věta 3 (ENMSV — Extrémy normy na množině s vazbou) *Nechť* $(X, ||\ ||_X)$ *a* $(Y, ||\ ||_Y)$ *jsou Banachovy prostory a* $(Y, ||\ ||_Y)$ *je navic reflexivní a nechť existuje* $(Z, ||\ ||_Z)$ *tak, že* $X \hookrightarrow\hookrightarrow Z$ *a nechť* $A : X \to Y$*. Nechť platí*

$A(0) = 0$

A^{-1} *je SC-lokálně slabě spojitý v* 0_Y*.*

Potom existují $y_0, y_1 \in Y$ *a* $s > 0$ *tak, že platí*

$$||x_0||_Z = \inf_{Ax=y, y\in M} ||x||_Z, Ax_0 = y_0$$

$$||x_1||_Z = \sup_{Ax=y, y\in M} ||x||_Z, Ax_1 = y_1$$

kde $M = \overline{B_Y^s(0)}$

Důkaz:

Víme, že $M = \overline{B_Y^s(0)}$ je slabě kompaktní v $(Y, ||\ ||_Y)$, jelikož se jedná o reflexivní Banachův prostor. Označme $\nu : Z \to \mathbf{R}$, $\nu(y) = ||y||_Z$. Je evidentní, že $J = \nu \circ id \circ A^{-1}$ je SC-rozklad a tedy existují y_0, y_1 podle věty. □

4.1 Aplikace v hyperelasticitě

V následující dokážeme existenci řešení modifikovaného problému, kde operátor hyperelasticity odpovídá problému DPO, případu s nulovými posuvy na hranici a nulovým vnějším zatížením. Pro důkaz existence řešení, je podstatné že, inverzní operátor k operátoru odpovidájící problému DPO je SC-lokálně slabě spojitý.

Věta 4 (SC-slabá lokální spojitost inverzního operátoru k operátoru hyperelasticity v nule) *Nechť jsou splněny předpoklady a operátor hyperelasticity je definován na prostoru* $\mathbf{V}^{\mathrm{p}} = \mathbf{W}^{2,\mathrm{p}}(\Omega) \bigcap \mathbf{W}_0^{1,\mathrm{p}}(\Omega)$. *Potom platí, že inverzní operátor je SC-lokálně slabě spojitý v* 0_Y.

Důkaz: Podrobně je proveden důkaz v [9]. V prvém kroku se pomocí věty o lokálním difeomorfismu dokáže, že existuje lokálně spojitý inverzní operátor o kterém se v druhém kroku dokáže, že je slabě spojitý.

Důkaz lokální spojitosti inverzního operátoru plyne z věty o lokálním difeomorfismu a mužete ho najít [1, 6]. Z předpokladu dvakrát spojité diferencovatelnosti funkce odezvy $\hat{\Sigma}$, plyne že nelineární operátor $A : Dom[A] \subset \mathbf{V}^{\mathrm{p}} \to \mathbf{L}_{\mathrm{p}}(\Omega)$ a A je spojitě diferencovatelný. Podstatné je, že Sobolevův prostor $W^{1,p}(\Omega)$ je algebra pro $p > 3$ a spojitě diferencovatelné zobrazení $\hat{\Sigma}$ zobrazuje $\mathbf{W}^{\mathbf{1},\mathbf{p}}(\Omega)$ do $\mathbf{W}^{\mathbf{1},\mathbf{p}}(\Omega)$. Předpoklad, že referenční konfigurace je v přirozeném stavu implikuje, že $A(0_x) = 0_y$. Rovnice $A'(0_x)u = f$ je lineární problém elasticity, který implikuje že $A'(0_x)$ je isomorfismus mezi prostory $\mathbf{V}^{\mathrm{p}}$ and $\mathbf{L}_{\mathrm{p}}(\Omega)$. Všechny předpoklady věty o lokálním difeomorfismu jsou splněny, tedy platí, že existuje spojitý lokální inverzní operátor k operátoru hyperelasticity.

Slabá spojitost se dokáže využitím kompaktního vnoření $W^{2,p}(\Omega)$ do $W^{1,p}(\Omega)$ a $W^{1,p}(\Omega)$ do $C(\overline{\Omega})$ a pomocí zeslabené spojitosti zobecněné rovnice, která navíc splňuje podmínku regularity. Podrobně viz [9]. □

Věta 5 (SC-rozkladatelnost funkcionálu) *Funkcionál* $J_0 : M \subset \mathbf{L}_{\mathrm{p}}(\Omega) \to \mathrm{R}$ *definovaný předpisem* $J_0 : f \mapsto ||u_f||$, *kde* $Au_f = f$ *je SC-rozkladatelný a má následující rozklad:*

$$J_0 = \nu \circ id \circ A^{-1},$$

kde

$A^{-1} : M \subset \mathbf{L}_{\mathrm{p}}(\Omega) \to \mathbf{V}^{\mathrm{p}}$ *je lokální inverzní operátor k operátoru hyperelasticity*

$id : \mathbf{V}^{\mathrm{p}} \to \mathbf{C}(\overline{\Omega}), \mathbf{u} \mapsto \mathbf{u}$ *vnoření do spojitých funkcí*

$\nu : \mathbf{C}(\overline{\Omega}) \to \mathrm{R}, u \mapsto ||u||$ *zobrazení normy*

Důkaz: Důkaz plyne z věty o SC-lokální slabé spojitosti inverzního operátoru hyperelasticity, Rellich–Kondraševovy věty o vnoření a spojitosti normy. □

Věta 6 (Problém POVFO pro hyperelastické materiály) *Nechť jsou splněny předpoklady MFPP, potom existuje* $s_0 > 0$ *tak, že problém POVFO má řešení podmnožinu* $M = \overline{B_Y^s(0_Y)}\ Y$,*kde* $0 < s \leq s_0$.

Důkaz: Důkaz plyne s věty o SC-rozkladatelnosti funkcionálu a věty ENMSV. □

5 Závěr

Získali jsme výsledek o existenci řešení, nevíme ale nic o jednoznačnosti a kde se extrémy nacházejí. Dá se očekávat, že minimum bude jednoznačné, že maximum bude ležet na hranici množiny M. Také je otázkou jak se dané extrémy budou hledat. Touto problematikou se budou zabývat následující práce. Hlavním nedostatkem je ovšem, že problém nemodeluje původní technickou problematiku, a velmi špatně se fyzikálně interpretuje. Na závěr by jsem chtěl poděkovat učastníkům semináře ODAM 99, kteří byli ochotni mne poslouchat, Ing. Jiří Horákovi, CSc za morální podporu a MUDr. Lence Karpetové, která mi vytvořila zajímavé podmínky pro tvorbu toho referátu.

Reference

[1] Ciarlet, P. G.: Mathematical Elasticity, Vol I: Three Dimensional Elasticity. Studies in Mathematics and its Applications 20 (1988), North Holland, Amsterdam.

[2] Franců, J.: *Monotone operators.* Appl. Math., Praha, (1990).

[3] Franců, J.: *Weakly continuous operators. Aplication to differential equations.* Applications of mathematics, ČSAV Praha, (1994).

[4] Giaquinta, M., Modica, G. and Souček, J.: *Cartesian Currents in the Calculus of Variations I.* Springer, Berlin, 1998.

[5] Giaquinta, M., Modica, G. and Souček, J.: *Cartesian Currents in the Calculus of Variations II.* Springer, Berlin, 1998.

[6] Konečný, M.: *Remarks to nonliner elasticity.* (to appear).

[7] Konečný, M. *Remarks to abstract optimization and aplications.* (to appear).

[8] Konečný, M. *Optimal control in hyperelasticity.* (to appear).

[9] Konečný, M. *Remarks to weakly continuous inverse operators in reflexive Banach spaces and an application in hyperelasticity.* Acta Mathematica et Informatica Univ. Ostraviensis, Universitas Ostraviensis, Ostrava, 1999.

[10] Nečas, J.: *Introduction to the Theory of Nonlinear Elliptic Equations*, TEUBNER-TEXTE zur Mathematik, B. G. Teubner Verlagsgesellschaft, Leipzig, 1983.

[11] Zeidler, E.: *Nonlinear Functional Analysis and its Applications*, Springer-Verlag, New York, 1986.

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

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


Univ. Palacki. Olomuc., Fac. rer. nat.,
Dept of Math. Anal. and Appl. of Math.
ODAM (1999) 35–43

Newtonova úloha s dvojí nejistotou

Vítězslav KRIŠTOF

Department of Mathematical Analysis and Applications of Mathematics,
Faculty of Science, Palacký University,
Tomkova 40, 779 00 Olomouc, Czech Republic
e-mail: kristof@risc.upol.cz

1 Úvod

Hodně technických, ekonomických, přírodních a společenských problémů dokážeme v dnešní době vcelku dobře matematicky modelovat. Pokud řešíme nějaký problém a vycházejí nám výsledky, které se liší od reálného života, potom chyba může být buď ve špatně zvoleném matematickém modelu, nebo v nepřesně zadaných vstupních datech, jako jsou koeficienty rovnic, okrajové podmínky a jiné. Bohužel ne vždy dokážeme přesně stanovit tyto vstupní hodnoty, a proto jsou někdy odhadovány s jistou přesností. Tedy vstupní data jsou nejistá a leží v určitém intervalu. Tento interval obvykle nazýváme množinou přípustných dat.

Úlohy s nejistými vstupními daty se dají řešit buď pomocí pravděpodobnosti (stochasticky), nebo tzv. *metodou spolehlivého řešení*, která vznikla poměrně nedávno, viz. [4], [5], [6] a další články autorů Hlaváčka a Chlebouna.

Metoda spolehlivého řešení, nebo též „metoda nejhoršího scénáře“, hledá nejhorší možné řešení, přičemž vždy zůstává na tzv. „bezpečné straně“, tj. v množině přípustných řešení. Další výhodou této metody je, že se v ní využívá teorie i algoritmů optimálního návrhu, což je vcelku dobře matematicky prozkoumaná oblast.

V této práci se budeme zabývat řešením eliptické rovnice s nelineární Newtonovou okrajovou podmínkou. Dále budeme předpokládat, že naše „nelinearita“ má polynomiální charakter. Takovýto problém se může vyskytovat např. při elektrolýze, kde výsledné řešení popisuje koncentraci elektrolytu. Nelineární hraniční

podmínka určuje tok přes hranici a koeficient κ můžeme chápat jako konstantu úměrnosti. Parametr α označuje, jak moc brání koncentrace elektrolytu průtoku přes hranici.

2 Formulace problému

Nechť $\Omega \subset R^2$ je omezená oblast s hranicí $\Gamma = \partial\Omega$, která je lipschitzovsky spojitá. Uzávěr oblasti Ω budeme označovat $\overline{\Omega}$. Naším úkolem bude nalézt $u : \overline{\Omega} \to R$ takové, že

$$-\triangle u = f \quad \text{v } \Omega, \tag{1}$$

$$\frac{\partial u}{\partial n} + \kappa |u|^\alpha u = \varphi \quad \text{na } \partial\Omega, \tag{2}$$

kde $f : \Omega \to R$, φ: $\partial\Omega \to R$ jsou dané funkce. Předpokládejme dále, že koeficienty α a κ jsou nejisté a nacházejí se v určitém intervalu. Kartézský součin intervalů budeme nazývat množina přípustných dat a značit

$$U_{ad} = \{(\alpha, \kappa) \in R \times R : 0 < \alpha_1 \le \alpha \le \alpha_2 \ ; \ \alpha_1 < \alpha_2; \\ 0 < \kappa_1 \le \kappa \le \kappa_2 \ ; \ \kappa_1 < \kappa_2 \ \}.$$

Klasické řešení můžeme definovat jako funkci $u \in C^2(\Omega) \cap C^1(\overline{\Omega})$ splňující (1) a (2). Prostor $C^2(\Omega) \cap C^1(\overline{\Omega})$ je vcelku dosti omezující požadavek na řešení, a proto místo klasického řešení budeme hledat tzv. *slabé řešení*. Budeme pracovat na Lebesgueových a Sobolevových prostorech $L^2(\Omega)$, $L^2(\partial\Omega)$ a $W^{1,2}(\Omega) \equiv H^1(\Omega)$.

Předpokládejme tedy, že $f \in L^2(\Omega)$ a $\varphi \in L^2(\partial\Omega)$. Pomocí Greenovy věty dostáváme slabou formulaci problému (1) a (2) :

$$\int_\Omega \nabla u \nabla v dx + \kappa \int_{\partial\Omega} |u|^\alpha uv dS = \int_\Omega fv dx + \int_{\partial\Omega} \varphi v dS.$$

Pokud si označíme

$$\mathsf{a}(\alpha, \kappa; u, v) = \int_\Omega \nabla u \nabla v dx + \kappa \int_{\partial\Omega} |u|^\alpha uv dS \tag{3}$$

$$\mathsf{L}(v) = \int_\Omega fv dx + \int_{\partial\Omega} \varphi v dS, \tag{4}$$

potom můžeme *slabým řešením* problému (1), (2), kde $(\alpha, \kappa) \in U_{ad}$ pevné, nazvat funkci $u : \Omega \to R$ takovou, že

$$\begin{aligned} &u \in H^1(\Omega) \\ &\mathsf{a}(\alpha, \kappa; u, v) = \mathsf{L}(v) \quad \forall v \in H^1(\Omega). \end{aligned} \tag{5}$$

Lema 1: Funkcionál L a zobrazení $v \in H^1(\Omega) \to \mathsf{a}(\alpha, \kappa; u, v)$ jsou spojité lineární formy na $H^1(\Omega)$ pro každé $u \in H^1(\Omega)$ a každé $(\alpha, \kappa) \in U_{ad}$.

Důkaz je založen na větě o stopách.

V důsledku tohoto tvrzení existuje zobrazení $\mathsf{A}(\alpha,\kappa) : H^1(\Omega) \to (H^1(\Omega))^*$ a funkcionál $\mathsf{b} \in (H^1(\Omega))^*$ pro něž platí

$$\begin{aligned}&\langle \mathsf{A}(\alpha,\kappa)u, v\rangle = \mathsf{a}(\alpha,\kappa;u,v),\\ &\langle \mathsf{b}, v\rangle = \mathsf{L}(v),\\ &\forall u, v \in H^1(\Omega).\end{aligned}$$

Tedy rovnost (5) můžeme přepsat pomocí operátorové rovnice na

$$\mathsf{A}(\alpha,\kappa)u = \mathsf{b}. \tag{6}$$

3 Řešení operátorové rovnice

Poznámka:
1) $(H^1(\Omega))^*$ značí duální prostor k $H^1(\Omega)$ a $\langle\,\cdot\,,\,\cdot\,\rangle$ dualitu mezi $(H^1(\Omega))^*$ a $H^1(\Omega)$.
2) Normu na prostoru $H^1(\Omega)$ budeme zapisovat $||\,\cdot\,||_1$ a symbol $|\,\cdot\,|_1$ bude znamenat seminormu na $H^1(\Omega)$.
3) Označení $u_n \to u$ znamená silnou konvergenci (v normě), zatímco slabou konvergenci zapisujeme $u_n \rightharpoonup u$.

Platí následující věta:
Věta 1: Buď V reflexivní separabilní Banachův prostor a A: $V \to V^*$ operátor

- koercivní tj.

$$\lim_{||u||_V \to \infty} \frac{\langle \mathsf{A}u, u\rangle}{||u||_V} = \infty$$

- spojitý na podprostorech konečné dimenze,
- ohraničený tj. existuje funkce $M : R^+ \to R^+$ taková, že

$$||\mathsf{A}u||_{V^*} \le M(||u||_V) \qquad \forall\ u \in V$$

- a splňuje tzv. podmínku $(M)_0$:

$$\left[\begin{array}{l} u_n \rightharpoonup u\ ,\ \mathsf{A}u_n \rightharpoonup \mathsf{b} \\ \langle \mathsf{A}u_n, u_n\rangle \to \langle \mathsf{b}, u\rangle \end{array}\right] \Longrightarrow \quad \mathsf{A}u = \mathsf{b}.$$

Potom A je surjektivní t.j. rovnice (6) má řešení pro každé $\mathsf{b} \in V^*$. Dále A^{-1} jako mnohoznačné zobrazení je ohraničené.

Pokud je A navíc ještě ryze monotónní t.j.

$$\langle \mathsf{A}u - \mathsf{A}v, u - v\rangle > 0 \qquad \forall u, v \in V,\ u \neq v,$$

potom množina $\mathsf{A}^{-1}b$ je jednobodová a tedy úloha (6) má právě jedno řešení.
Důkaz. Viz [2].

Lema 2: Operátor $\mathsf{A}(\alpha,\kappa)$ z rovnice (6) je pro $\forall(\alpha,\kappa)\in U_{ad}$ ryze monotónní.
Důkaz. K důkazu ryzí monotonie operátoru $\mathsf{A}(\alpha,\kappa)$ nám stačí ověřit ryzí monotonii bilineární formy $\mathsf{a}(\alpha,\kappa;u,v)$, tj.

$$\mathsf{a}(\alpha,\kappa;u,u-v)-\mathsf{a}(\alpha,\kappa;v,u-v)>0,\quad u,v\in H^1(\Omega),\ u\neq v.$$

Jelikož funkce $\psi(t)=|t|^\alpha t$ je rostoucí v R dostáváme, že

$$\mathsf{a}(\alpha,\kappa;u,u-v)-\mathsf{a}(\alpha,\kappa;v,u-v)=\int_\Omega|\nabla(u-v)|^2dx+\kappa\int_{\partial\Omega}[|u|^\alpha u-|v|^\alpha v](u-v)dS\geq 0.$$

Předpokládejme nyní, že $\mathsf{a}(\alpha,\kappa;u,u-v)-\mathsf{a}(\alpha,\kappa;v,u-v)=0$. Potom je hned vidět, že $|u-v|_1=0$ a $u-v=0$ skoro všude na $\partial\Omega$, a tedy $u-v=$ konst. v Ω. Nakonec z věty o stopách dostáváme rovnost $u-v=0$, čímž je tvrzení lemmatu dokázáno.

Lema 3: Pro $\forall(\alpha,\kappa)\in U_{ad}$ je operátor $\mathsf{A}(\alpha,\kappa)$ z rovnice (6) koercivní.
Důkaz. Aby $\mathsf{A}(\alpha,\kappa)$ byl koercivní operátor, stačí dokázat, že

$$\mathsf{a}(\alpha,\kappa;u,u)=\langle\mathsf{A}(\alpha,\kappa)u,u\rangle\geq c||u||_1^2\quad\text{ pro }\ ||u||_1\geq 1$$

Nechť tedy $u\in H^1(\Omega)$ a $||u||_1\geq 1$. Položíme-li $w=\frac{u}{||u||_1}\in H^1(\Omega)$, potom $||w||_1=1$. Z (3) dostáváme

$$\mathsf{a}(\alpha,\kappa;u,u)=\int_\Omega(\nabla u)^2dx+\kappa\int_{\partial\Omega}|u|^{\alpha+2}dS=|u|_1^2+\kappa||u||^{\alpha+2}_{L^{\alpha+2}(\partial\Omega)}.$$

Z nerovnosti (viz. [1])

$$|v|_1^2+||v||^q_{L^q(\partial\Omega)}\geq\tilde{c}\quad\forall v\in H^1(\Omega),||v||_1=1,q\geq 1$$

plyne pro $q:=\alpha+2$ a v:= w, že

$$\tilde{c}\leq\frac{|u|_1^2}{||u||_1^2}+\frac{||u||^{\alpha+2}_{L^{\alpha+2}(\partial\Omega)}}{||u||_1^{\alpha+2}}$$

$$\tilde{c}||u||_1^2\leq|u|_1^2+\frac{||u||^{\alpha+2}_{L^{\alpha+2}(\partial\Omega)}}{||u||_1^\alpha}\leq|u|_1^2+||u||^{\alpha+2}_{L^{\alpha+2}(\partial\Omega)}.$$

A jelikož

$$\mathsf{a}(\alpha,\kappa;u,u)\geq\min(1,\kappa)(|u|_1^2+||u||^{\alpha+2}_{L^{\alpha+2}(\partial\Omega)}),$$

platí

$$\mathsf{a}(\alpha,\kappa;u,u)\geq\tilde{c}\min(1,\kappa)||u||_1^2\geq\tilde{c}\min(1,\kappa_1)||u||_1^2.\tag{7}$$

Lema 4: Operátor $\mathsf{A}(\alpha,\kappa)$ z rovnice (6) je pro $\forall(\alpha,\kappa)\in U_{ad}$ ohraničený a spojitý.
Důkaz spojitosti plyne z následující nerovnosti

$$|\mathsf{a}(\alpha,\kappa;u,v)-\mathsf{a}(\alpha,\kappa;w,v)|\leq c(1+||u||_1^{\alpha_2}+||w||_1^{\alpha_2})||u-w||_1||v||_1\quad\forall u,v,w\in H^1(\Omega)\tag{8}$$

a ohraničenost dostaneme, pokud do nerovnosti (8) dosadíme w=0.
Důkaz nerovnosti (8) si rozdělíme do 2 částí. Nejprve budeme odhadovat výraz

$$I_0 = \int_\Omega |\nabla(u-w)\nabla v|dx$$

a potom výraz

$$I_1 = \int_{\partial\Omega} ||u|^\alpha u - |w|^\alpha w|\,|v|dS.$$

Výraz I_0 odhadneme užitím Hölderovy nerovnosti jako

$$0 \le I_0 \le ||u-w||_1||v||_1.$$

Pro odhad výrazu I_1 si zadefinujeme funkci

$$\varphi(t) = |r+t(s-r)|^\alpha(r+t(s-r)),\ t\in[0,1],\ r,s\in R.$$

Platí rovnost $\varphi^{'}(t) = (\alpha+1)(s-r)|r+t(s-r)|^\alpha$ a také, že

$$|s|^\alpha s - |r|^\alpha r = \varphi(1)-\varphi(0) = \int_0^1 \varphi^{'}(t)dt = (\alpha+1)(s-r)\int_0^1 |r+t(s-r)|^\alpha dt.$$

Dále ještě užijeme odhad $|r+t(s-r)|^\alpha \le |r|^\alpha + |s|^\alpha$ pro $t\in[0,1]$, který plyne z neexistence lokálního extrému funkce $|r+t(s-r)|^\alpha$ na otevřeném intervalu (0,1). Celkově dostáváme, že

$$0 \le I_1 \le \kappa(\alpha+1)\int_{\partial\Omega}|u-w|(|u|^\alpha+|w|^\alpha)|v|dS.$$

Nyní užitím Hölderovy nerovnosti tvaru

$$\int_{\partial\Omega}|a_1a_2a_3|dS \le \prod_{i=1}^{3}(\int_{\partial\Omega}|a_i|^{p_i}dS)^{1/p_i},$$

kde $p_i > 1$, i = 1,2,3; $\frac{1}{p_1}+\frac{1}{p_2}+\frac{1}{p_3}=1$ a $a_i \in L^{p_i}(\partial\Omega)$ získáme odhad

$$0 \le I_1 \le \kappa_2(\alpha_2+1)||u-w||_{L^{p_1}(\partial\Omega)}(||u||^{\alpha_2}_{L^{\alpha_2 p_2}(\partial\Omega)} + ||w||^{\alpha_2}_{L^{\alpha_2 p_2}(\partial\Omega)})||v||_{L^{p_3}(\partial\Omega)},$$

který v důsledku kompaktního vnoření $H^1(\Omega)$ do $L^p(\partial\Omega)$ pro $\forall p\in[1,\infty)$ (viz. [9]) můžeme přepsat na

$$0 \le I_1 \le \kappa_2(\alpha_2+1)||u-w||_1(||u||^{\alpha_2}_1 + ||w||^{\alpha_2}_1)||v||_1.$$

Sečtením odhadů I_0 a I_1 je již nerovnost (8) dokázána, neboť

$$|\mathsf{a}(\alpha,\kappa;u,v) - \mathsf{a}(\alpha,\kappa;w,v)| \le I_0 + I_1.$$

Věta 2: Pro každé $(\alpha,\kappa)\in U_{ad}$ existuje právě jedno řešení rovnice (6) a tedy i problému (1), (2).

Důkaz. Použijeme větu 1, kde za prostor V volíme $H^1(\Omega)$. Ze spojitosti $\mathsf{A}(\alpha,\kappa)$ plzne spojitost na konečně dimenzionálních podprostorech a dále také hemispojitost operátoru $\mathsf{A}(\alpha,\kappa)$, tj.

$$t_n \in R, \quad t_n \to 0 \Rightarrow \mathsf{A}(\alpha,\kappa)(u+t_n v) \rightharpoonup \mathsf{A}(\alpha,\kappa)(u).$$

A nakonec (viz. [2]) z hemispojitosti a monotonie operátoru $\mathsf{A}(\alpha,\kappa)$ plyne podmínka $(M)_0$, čímž jsou splněny všechny předpoklady věty 1, a tudíž rovnice (6) má právě jedno řešení.

4 Metoda spolehlivého řešení

Nyní si zvolíme kritérium pro naši úlohu. Budeme hledat maximální koncentraci v zadané oblasti Ω, proto kritériem může být např. střední hodnota na vybraných podoblastech.

Pokud si zadefinujeme

$$\psi_j(u) = \frac{1}{|G_j|}\int_{G_j} u dx, \qquad G_j \subset \Omega; \quad j = 1,\dots,\delta,$$
$$\psi_j(u) = \frac{1}{|G_j|}\int_{G_j} u dS, \qquad G_j \subset \partial\Omega; \quad j = \delta+1,\dots,N,$$

tak naše kritérium bude mít tvar

$$\Phi(u) = \max_{j=1,\dots,N}(\psi_j(u)).$$

Naším cílem potom bude řešit *Maximalizační problém:*
Nalézt

$$(\alpha^0,\kappa^0) = \arg\max_{(\alpha,\kappa)\in U_{ad}}[\max_{j=1,\dots,N}\psi_j(u(\alpha,\kappa))] \tag{9}$$

kde $u(\alpha,\kappa)$ je slabým řešením problému (1) a (2).

Lema 5: Existují kladné konstanty c_0, c_1 a c_2, které nezávisejí na α a κ, takové, že platí následující nerovnosti:

$$c_0||v||_1^2 \le \mathsf{a}(\alpha,\kappa;v,v) \tag{10}$$
$$|\mathsf{a}(\alpha,\kappa;u,v)| \le c_1(1+||u||_1^{\alpha_2})||u||_1||v||_1 \tag{11}$$
$$|\mathsf{L}(v)| \le c_2||v||_1 \tag{12}$$

pro $\forall u, v \in H^1(\Omega)$.
Důkaz. První nerovnost plyne z koercivity operátoru $\mathsf{A}(\alpha,\kappa)$. Druhou nerovnost dostaneme, pokud do vztahu (8) dosadíme za w=0. Poslední nerovnost vyplývá ze spojitosti L a vztahu (4), neboť pro lineární operátory jsou pojmy ohraničenost a spojitost ekvivalentní.

Lema 6: Nechť $(\alpha_n, \kappa_n) \in U_{ad}$, $\alpha_n \to \alpha$ v R, $\kappa_n \to \kappa$ v R pro $n \to \infty$. Potom $(\alpha, \kappa) \in U_{ad}$ a $u(\alpha_n, \kappa_n) \rightharpoonup u(\alpha, \kappa)$ (slabě) v $H^1(\Omega)$.
Důkaz. Z uzavřenosti množiny U_{ad} vyplývá, že $(\alpha, \kappa) \in U_{ad}$. Dále nechť $\bar{u} \in H^1(\Omega)$ je libovolný, ale pevně zvolený prvek. Pomocí nerovností (10), (11), (12) a vztahu (5) dostáváme

$$\begin{aligned} c_0||u(\alpha_n, \kappa_n)||_1^2 &\leq \mathsf{a}(\alpha_n, \kappa_n; u(\alpha_n, \kappa_n), u(\alpha_n, \kappa_n)) = \mathsf{L}(u(\alpha_n, \kappa_n)) \leq \\ &\leq c_2||u(\alpha_n, \kappa_n)||_1 \end{aligned}$$

Z této nerovnosti plyne, že

$$||u(\alpha_n, \kappa_n)||_1 \leq c \qquad \forall n$$

a tudíž existuje podposloupnost $\{u(\alpha_m, \kappa_m)\} \subset H^1(\Omega)$ a funkce $u \in H^1(\Omega)$ tak, že

$$u(\alpha_m, \kappa_m) \rightharpoonup u \quad v\ H^1(\Omega).$$

Nyní již zbývá dokázat jen rovnost $u = u(\alpha, \kappa)$. Nechť $v \in H^1(\Omega)$ libovolné. Naše úloha má tvar

$$\mathsf{a}(\alpha_m, \kappa_m; u(\alpha_m, \kappa_m), v) = \mathsf{L}(v).$$

Limitním přechodem pro $m \to \infty$ dostaneme:

$$\begin{aligned} \lim_{m\to\infty} \int_\Omega \nabla u(\alpha_m, \kappa_m) \nabla v dx &= \int_\Omega \nabla u \nabla v dx, \\ \lim_{m\to\infty} \kappa_m \int_{\partial\Omega} |u(\alpha_m, \kappa_m)|^{\alpha_m} u(\alpha_m, \kappa_m) v dS &= \kappa \int_{\partial\Omega} |u|^\alpha u v dS, \end{aligned}$$

neboť ze slabé konvergence v prostoru $H^1(\Omega)$ plyne, že $\nabla u(\alpha_m, \kappa_m) \rightharpoonup \nabla u$ v $L^2(\Omega)$ a podle věty o kompaktním vnoření do prostoru stop (viz. [9]) platí, že $u(\alpha_m, \kappa_m) \to u$ v $L^q(\partial\Omega)$ pro $\forall q \geq 1$. Dále jestliže si oynačíme $\gamma(\alpha, \kappa, u(\alpha, \kappa)) = \kappa|u|^\alpha$, pak existuje kladná konstanta $\bar{c}$ tak, že diferenciál 1. řádu fce γ můžeme odhadnout jako

$$d\gamma = |u|^\alpha \cdot \triangle\kappa + \kappa|u|^\alpha \ln|u| \cdot \triangle\alpha + \kappa\alpha|u|^{\alpha-1} \cdot \triangle(|u|) \leq \bar{c}$$

a tedy můžeme psát

$$\begin{aligned} &|\int_{\partial\Omega} \gamma(\alpha_m, \kappa_m, u(\alpha_m, \kappa_m)) v - \gamma(\alpha, \kappa, u(\alpha, \kappa)) v dS| \leq \\ &\leq \int_{\partial\Omega} |\gamma(\alpha_m, \kappa_m, u(\alpha_m, \kappa_m)) - \gamma(\alpha, \kappa, u(\alpha, \kappa))||v| dS \leq \\ &\leq (\int_{\partial\Omega} |\gamma(\alpha_m, \kappa_m, u(\alpha_m, \kappa_m)) - \gamma(\alpha, \kappa, u(\alpha, \kappa))|^2 dS)^{\frac{1}{2}} (\int_{\partial\Omega} |v|^2 dS)^{\frac{1}{2}} \to 0. \end{aligned}$$

Tím dostáváme, že

$$\lim_{m\to\infty} \mathsf{a}(\alpha_m, \kappa_m; u(\alpha_m, \kappa_m), v) = \mathsf{a}(\alpha, \kappa; u, v)$$

a tedy $u(\alpha_n, \kappa_n) \rightharpoonup u(\alpha, \kappa)$ v $H^1(\Omega)$.

Věta 3: Existuje alespoň jedno řešení problému (9).
Důkaz. Z věty 2 víme, že pro $\forall(\alpha, \kappa) \in U_{ad}$ existuje právě jediné řešení $u(\alpha, \kappa)$, a tedy můžeme definovat funkcionál

$$J(\alpha, \kappa) = \max_{j=1,\dots,N} \psi_j(u(\alpha, \kappa)).$$

Nechť $\{\alpha_n, \kappa_n\}$ je posloupnost prvků z U_{ad} taková, že

$$\lim_{n\to\infty} J(\alpha_n, \kappa_n) = \sup_{(\alpha,\kappa)\in U_{ad}} J(\alpha, \kappa).$$

Z kompaktnosti U_{ad} v R existuje vybraná podposloupnost $\{\alpha_m, \kappa_m\} \subset \{\alpha_n, \kappa_n\}$, která je konvergentní a jejíž limita leží v U_{ad}, tj.

$$(\alpha_m, \kappa_m) \to (\alpha^0, \kappa^0) \ \text{ v R } \ \text{a } \ (\alpha^0, \kappa^0) \in U_{ad}.$$

Pomocí lemma 6 dostáváme slabou konvergenci

$$u(\alpha_m, \kappa_m) \rightharpoonup u(\alpha^0, \kappa^0) \ \ v \ \ H^1(\Omega).$$

Z Rellichovy věty (viz. [9]) nyní plyne silná konvergence

$$u(\alpha_m, \kappa_m) \to u(\alpha^0, \kappa^0) \ \ v \ \ L^2(\Omega),$$

což v našem případě znamená, že

$$\lim_{m\to\infty} \psi_j(u(\alpha_m, \kappa_m)) = \psi_j(u(\alpha^0, \kappa^0)) \ \text{ pro } j \le \delta.$$

Pro $j > \delta$ využijeme kompaktnosti operátoru stop. Dále využitím záměny limity a maximalizace platí rovnost

$$\begin{aligned}\lim_{m\to\infty} J(\alpha_m, \kappa_m) &= \lim_{m\to\infty} \max_{j=1,\dots,N} \psi_j(u(\alpha_m, \kappa_m)) = \max_{j=1,\dots,N} \lim_{m\to\infty} \psi_j(u(\alpha_m, \kappa_m)) = \\ &= \max_{j=1,\dots,N} \psi_j(u(\alpha^0, \kappa^0)) = J(\alpha^0, \kappa^0),\end{aligned}$$

a tedy

$$J(\alpha^0, \kappa^0) = \sup_{(\alpha,\kappa)\in U_{ad}} J(\alpha, \kappa).$$

Tím je dokázáno, že (α^0, κ^0) je řešením problému (9).

Reference

[1] M. Feistauer, K. Najzar: *Finite element approximation of a problem with a nonlinear Newton boundary condition.* Numer. Math. **78** (1998), 403–425.

[2] J. Franců: *Úvod do teorie monotónních operátorů.* Brno 1987.

[3] J. Franců: *Solvability of operator equations. Survey directed to differential equations.* Brno 1995.

[4] I. Hlaváček: *Reliable solutions of elliptic boundary value problems with respact to uncertain data.* Nonlinear Analysis, Theory, Meth. & Appls. **30** (1997), 3879–3890.

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

[6] J. Chleboun: *Reliable solution for 1D quasilinear elliptic equation with uncertain coefficients.* Zasláno do J. Math. Anal. Appl.

[7] D. Kinderlehrer, G. Stampacchia: *An introduction to variational inequalities and their applications.* Academic Press, New York 1980.

[8] A. Kufner, O. John, S. Fučík: *Function spaces.* Academia, Prague, 1977.

[9] J. Nečas: *Les méthodes directes en théorie des équations elliptiques.* Academia, Masson, Paris, 1967.

[10] J. Nečas: *Introduction to the theory of nonlinear elliptic equations.* Teubner-Texte zur Math. **52**, Leipzig, 1983.


Univ. Palacki. Olomuc., Fac. rer. nat.,
Dept of Math. Anal. and Appl. of Math.
ODAM (1999) 45–58

Problém rušivých parametrů při zakládání stavby

Pavla KUNDEROVÁ

Department of Mathematical Analysis and Applications of Mathematics,
Faculty of Science, Palacký University,
Tomkova 40, 779 00 Olomouc, Czech Republic
e-mail: kunderov@risc.upol.cz

Abstrakt

V článku je ukázán příklad užití statistického univariátního modelu s rušivými parametry.

1991 Mathematics Subject Classification: 62J05

1 Úvod, označení

Při zakládání velké stavby je nezbytně nutno velmi spolehlivě určit okamžik, kdy se podloží po provedených rozsáhlých zemních úprav ustálí do té míry, že je možno pokračovat ve stavbě bez rizika jejího následného poškození.

Na staveništi se sleduje k bodů, jejichž výška se opakovaně měří v okamžicích $t_1, t_2, \ldots, t_m$. Je třeba zvolit model, který popisuje klesání podloží ve zvolených bodech a na základě opakovaných měření určit odhady parametrů tohoto modelu.

Při volbě modelu musíme rozhodnout, zda budeme pokládat klesání podloží za stejnoměrné na celém pozemku nebo za nestejnoměrné.

Vzhledem k tomu, že na ploše $500\,m \times 500\,m$ se měří např. 30 kontrolních bodů, je důležité rozhodnutí, které parametry budeme pokládat za užitečné a které za rušivé. Při vhodné volbě je možno podstatně zkrátit výpočty.

Výsledek měření i-tého bodu v j-té epoše lze popsat takto

$$\eta_i(t_j) = \beta_i(t_j) + \varepsilon_{ij} = \beta_i - \kappa_1^i(1 - e^{-\kappa_2^i t_j}) + \varepsilon_{ij}, \quad i = 1, \ldots, k,\ j = 1, \ldots, m.$$

β_i značí výšku i-tého bodu v čase t_0, funkce $\kappa_1^i(1-e^{-\kappa_2^i t})$ popisuje pohyb zemních vrstev v i-tém bodě. Jestliže neznámé parametry $\kappa_1^i > 0$, $\kappa_2^i > 0$, pokládáme za odlišné v jednotlivých sledovaných bodech, předpokládáme nestejné geologické složení podloží stavby, tj. nestejnoměrný pokles.

Cílem měření je určit odhady parametrů $\beta_1, \ldots, \beta_k$ a parametrů κ_1^i, κ_2^i, $i = 1, \ldots, k$. Pro kvalitní provedení stavby je nejen nezbytně nutná správná volba modelu, ale je také nutné, aby variabilita odhadů neznámých parametrů byla nízká. Velký rozptyl odhadů by mohl zcela znehodnotit dosažené výsledky.

Inženýr, který stavbu vede, musí vědět, kdy je možno pokračovat ve stavbě, tj. kdy bude pro všechny sledované body další pokles podloží zanedbatelný. To znamená, že je nutno určit pro každý sledovaný bod takové t_i, pro které funkce modelující v tomto bodě pokles dosáhne určité hranice, tj. zjistit, pro které t_i platí

$$\kappa_1^i(1 - e^{-\kappa_2^i t_i}) = C\kappa_1^i, \quad i = 1, \ldots, k,$$

kde $0 < C < 1$ je vhodná konstanta dostatečně blízká 1. Ve stavbě můžeme pokračovat v čase $t = max\{t_1, \ldots, t_k\}$.

Nechť R^n označuje prostor všech n-rozměrných reálných vektorů, $\boldsymbol{u}_p$ a $\boldsymbol{A}_{m,n}$ označuje reálný sloupcový p-rozměrný vektor a reálnou matici rozměru $m \times n$. Symboly $\boldsymbol{A}', \mathscr{R}(\boldsymbol{A}), \mathscr{N}(\boldsymbol{A}), r(\boldsymbol{A})$ označují transpozici, prostor vytvořený nad sloupci matice $\boldsymbol{A}$, nulový prostor a hodnost matice $\boldsymbol{A}$. Symbol $\boldsymbol{I}$ označuje jednotkovou matici. $\boldsymbol{A}^-$ označuje libovolnou pseudoinverzní (g-inverzní) matici k matici $\boldsymbol{A}$ (splňující $\boldsymbol{A}\boldsymbol{A}^-\boldsymbol{A} = \boldsymbol{A}$).

$\boldsymbol{P}_A$ resp. $\boldsymbol{Q}_A$ označuje ortogonální projektor na $\mathscr{R}(\boldsymbol{A})$ resp. na $\mathscr{R}^\perp(\boldsymbol{A}) = \mathscr{N}(\boldsymbol{A}')$, $\boldsymbol{A}^\perp$ označuje libovolnou matici, pro kterou $\mathscr{R}^\perp(\boldsymbol{A}) = \mathscr{R}(\boldsymbol{A}^\perp)$.

Je-li $\mathscr{R}(\boldsymbol{A}) \subset \mathscr{R}(\boldsymbol{S})$, $\boldsymbol{S}$ pozitivně semidefinitní, označuje symbol $\boldsymbol{P}_A^{S^-}$ projektor projektující vektory z prostoru $\mathscr{R}(\boldsymbol{S})$ do prostoru $\mathscr{R}(\boldsymbol{A})$ podél $\mathscr{R}(\boldsymbol{S}\boldsymbol{A}^\perp)$. Všechny takové projektory tvoří třídu matic $\boldsymbol{A}(\boldsymbol{A}'\boldsymbol{S}^-\boldsymbol{A})^-\boldsymbol{A}'\boldsymbol{S}^- + \boldsymbol{F}(\boldsymbol{I} - \boldsymbol{S}\boldsymbol{S}^-)$, kde $\boldsymbol{F}$ je libovolná matice příslušného rozměru, viz [4], (2.14). $\boldsymbol{Q}_A^{S^-} = \boldsymbol{I} - \boldsymbol{P}_A^{S^-}$.

2 Řešení problému

Protože uvažovaný model není lineární v parametrech, je nutno jej linearizovat pomocí prvních dvou členů Taylorova rozvoje funkce $\kappa_1^i(1 - e^{-\kappa_2^i t})$ ve vhodném bodě $(\kappa_{1,0}^i, \kappa_{2,0}^i)$, $\kappa_{1,0}^i > 0, \kappa_{2,0}^i > 0$. Dostaneme model

$$\eta_i(t_j) = \beta_i - [\kappa_{1,0}^i(1-e^{-\kappa_{2,0}^i t_j}) + (1-e^{-\kappa_{2,0}^i t_j})(\kappa_1^i - \kappa_{1,0}^i) + \kappa_{1,0}^i t_j e^{-\kappa_{2,0}^i t_j}(\kappa_2^i - \kappa_{2,0}^i)] + \varepsilon_{ij},$$

$$i = 1, \dots, k,\ j = 1, \dots, m.$$

Označme

$$Y_i^{(j)} = \eta_i(t_j) + \kappa_{1,0}^i(1 - e^{-\kappa_{2,0}^i t_j}),\ \varphi_1^i(t) = -(1 - e^{-\kappa_{2,0}^i t}),\ \varphi_2^i(t) = -\kappa_{1,0}^i t e^{-\kappa_{2,0}^i t},$$

$$\delta\kappa_1^i = \kappa_1^i - \kappa_{1,0}^i, \quad \delta\kappa_2^i = \kappa_2^i - \kappa_{2,0}^i,\ i = 1, \dots, k,\ j = 1, \dots, m.$$

Tedy

$$\boldsymbol{Y}_i^{(j)} = \beta_i + \varphi_1^i(t_j)\delta\kappa_1^i + \varphi_2^i(t_j)\delta\kappa_2^i + \varepsilon_{ij},\ i = 1, \dots, k,\ j = 1, \dots, m.$$

Uvažujme vektor pozorování $\boldsymbol{Y} = (\boldsymbol{Y}_1^{(1)}, \dots, \boldsymbol{Y}_1^{(m)}, \dots, \boldsymbol{Y}_k^{(1)}, \dots, \boldsymbol{Y}_k^{(m)})'$.

Náš model ze zakládání stavby lze přepsat ve tvaru

$$\boldsymbol{Y} = (\boldsymbol{W}, \boldsymbol{Z}) \begin{pmatrix} \vartheta \\ \beta \end{pmatrix} + \varepsilon \tag{1}$$

kde

$$\boldsymbol{W} = \begin{pmatrix} \varphi_1^1(t_1) & \varphi_2^1(t_1) & 0 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ \varphi_1^1(t_m) & \varphi_2^1(t_m) & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & \varphi_1^2(t_1) & \varphi_2^2(t_1) & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & & 0 & 0 \\ 0 & 0 & \varphi_1^2(t_m) & \varphi_2^2(t_m) & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & \dots & \varphi_1^k(t_1) & \varphi_2^k(t_1) \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & \varphi_1^k(t_m) & \varphi_2^k(t_m) \end{pmatrix},$$

$$\boldsymbol{Z} = \begin{pmatrix} 1 & 0 & \dots & 0 \\ \vdots & \vdots & & 0 \\ 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & & 0 \\ 0 & 1 & \vdots & 0 \\ \dots & & & \dots \\ 0 & 0 & \dots & 1 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix}, \quad \vartheta = \begin{pmatrix} \delta\kappa_1^1 \\ \delta\kappa_2^1 \\ \delta\kappa_1^2 \\ \delta\kappa_2^2 \\ \vdots \\ \delta\kappa_1^k \\ \delta\kappa_2^k \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}.$$

Tedy matice $\boldsymbol{W}$ typu $mk \times 2k$ a matice $\boldsymbol{Z}$ typu $mk \times k$ jsou známé, vektor ϑ o $2k$ složkách je vektor užitečných parametrů, β je vektor rušivých parametrů.

Předpokládáme, že platí

$$E(\boldsymbol{Y}) = (\boldsymbol{W}, \boldsymbol{Z}) \begin{pmatrix} \vartheta \\ \beta \end{pmatrix}, \quad var(Y) = \boldsymbol{V},$$

a že $\boldsymbol{V}$ je taková varianční matice, pro kterou je splněno

$$\mathscr{R}(\boldsymbol{W}, \boldsymbol{Z}) \ \subset \ \mathscr{R}(\boldsymbol{V}). \tag{2}$$

Z tohoto předpokladu plyne, že $\boldsymbol{Y} \in \mathscr{R}(\boldsymbol{V})$ skoro jistě. Uvažujeme ve shodě s článkem [4] tzv. velký model (se všemi parametry)

$$\mathscr{M}_a(\boldsymbol{V}) = [\boldsymbol{Y}, \ \boldsymbol{W}\vartheta + \boldsymbol{Z}\beta, \ \boldsymbol{V}], \tag{3}$$

a tzv. malý model (bez rušivých parametrů)

$$\mathscr{M}(\boldsymbol{V}) = [\boldsymbol{Y}, \ \boldsymbol{W}\vartheta, \ \boldsymbol{V}]. \tag{4}$$

Označení Označme obdobně jako ve [4] symbolem $\mathscr{E}_a$ třídu všech lineárních funkcí $p'\vartheta$ vektoru užitečných parametrů, které jsou nevychýleně odhadnutelné v modelu $\mathscr{M}_a(\boldsymbol{V})$, tj. třídu těch lineárních funkcí $p'\vartheta$ pro které existuje odhadová funkce $l'\boldsymbol{Y}$ taková, že $E[l'\boldsymbol{Y}] = p'\vartheta$, $\forall\vartheta$, $\forall\beta$. Index a bude v dalším textu vždy označovat, že uvažujeme model se všemi parametry.

Obdobně symbol $\mathscr{E}$ bude označovat třídu takových funkcí $\boldsymbol{p}'\vartheta$, které jsou nevychýleně odhadnutelné v modelu $\mathscr{M}(\boldsymbol{V})$.
Platí (viz [4], vztahy (2.1), (2.2))

$$\mathscr{E} = \{\boldsymbol{p}'\vartheta : \boldsymbol{p} \in \mathscr{R}(\boldsymbol{W}')\}, \tag{5}$$

$$\mathscr{E}_a = \{\boldsymbol{p}'\vartheta : \boldsymbol{p} \in \mathscr{R}(\boldsymbol{W}'\boldsymbol{Q}_Z)\}. \tag{6}$$

Poznámka 1
V dalším textu budeme symbolem $\hat{\vartheta}_a$ resp. $\hat{\vartheta}$ označovat V^- -LS odhad parametru ϑ počítaný v modelu $\mathscr{M}_a(\boldsymbol{V})$ resp. v modelu $\mathscr{M}(\boldsymbol{V})$, (viz [2], str.161).

Podle předpokladu (2), $\widehat{\boldsymbol{p}'\vartheta}_a$ resp. $\widehat{\boldsymbol{p}'\vartheta}$ je BLUE funkce $\boldsymbol{p}'\vartheta \in \mathscr{E}_a$, resp. $\boldsymbol{p}'\vartheta \in \mathscr{E}$, (viz [2], věta 5.3.2, str. 162).

Věta 1

$$\widehat{\boldsymbol{p}'\vartheta} = \boldsymbol{p}'(\boldsymbol{W}'\boldsymbol{V}^-\boldsymbol{W})^-\boldsymbol{W}'\boldsymbol{V}^-\boldsymbol{Y}, \quad \text{je-li} \quad \boldsymbol{p}'\vartheta \in \mathscr{E}, \tag{7}$$

$$\widehat{\boldsymbol{p}'\vartheta}_a = \boldsymbol{p}'(\boldsymbol{W}'\boldsymbol{V}^-\boldsymbol{Q}_Z^{V^-}\boldsymbol{W})^-\boldsymbol{W}'\boldsymbol{V}^-\boldsymbol{Q}_Z^{V^-}\boldsymbol{Y}, \quad \text{je-li} \quad \boldsymbol{p}'\vartheta \in \mathscr{E}_a, \tag{8}$$

$$var[\widehat{\boldsymbol{p}'\vartheta}] = \boldsymbol{p}'(\boldsymbol{W}'\boldsymbol{V}^-\boldsymbol{W})^-\boldsymbol{p}, \quad \text{je-li} \quad \boldsymbol{p}'\vartheta \in \mathscr{E}, \tag{9}$$

$$var[\widehat{\boldsymbol{p}'\vartheta}_a] = \boldsymbol{p}'(\boldsymbol{W}'\boldsymbol{V}^-\boldsymbol{Q}_Z^{V^-}\boldsymbol{W})^-\boldsymbol{p}, \quad \text{je-li} \quad \boldsymbol{p}'\vartheta \in \mathscr{E}_a. \tag{10}$$

Tyto výrazy jsou invariantní na volbu g-inversních matic.

Důkaz v modelu $\mathscr{M}_a$ máme

$$\begin{pmatrix} \hat{\vartheta}_a \\ \hat{\beta}_a \end{pmatrix} = \left[(\boldsymbol{W},\boldsymbol{Z})'\boldsymbol{V}^-(\boldsymbol{W},\boldsymbol{Z})\right]^- \begin{pmatrix} \boldsymbol{W}' \\ \boldsymbol{Z}' \end{pmatrix} \boldsymbol{V}^-\boldsymbol{Y}$$

$$= \begin{bmatrix} \boldsymbol{W}'\boldsymbol{V}^-\boldsymbol{W}, & \boldsymbol{W}'\boldsymbol{V}^-\boldsymbol{Z} \\ \boldsymbol{Z}'\boldsymbol{V}^-\boldsymbol{W}, & \boldsymbol{Z}'\boldsymbol{V}^-\boldsymbol{Z} \end{bmatrix}^- \begin{pmatrix} \boldsymbol{W}'\boldsymbol{V}^- \\ \boldsymbol{Z}'\boldsymbol{V}^- \end{pmatrix} \boldsymbol{Y}. \tag{11}$$

Odhad získaný dosazením tohoto výrazu do nevychýleně odhadnutelné funkce je určen jednoznačně.

Užitím následující Rohdeho vzorce pro g-inverzi blokové p.s.d. matice (viz [3], Lemma 13, str. 68)

$$\begin{pmatrix} \boldsymbol{A}, & \boldsymbol{B} \\ \boldsymbol{B}', & \boldsymbol{C} \end{pmatrix}^-$$

$$= \begin{pmatrix} \boldsymbol{A}^- + \boldsymbol{A}^-\boldsymbol{B}(\boldsymbol{C}-\boldsymbol{B}'\boldsymbol{A}^-\boldsymbol{B})^-\boldsymbol{B}'\boldsymbol{A}^-, & -\boldsymbol{A}^-\boldsymbol{B}(\boldsymbol{C}-\boldsymbol{B}'\boldsymbol{A}^-\boldsymbol{B})^- \\ -(\boldsymbol{C}-\boldsymbol{B}'\boldsymbol{A}^-\boldsymbol{B})^-\boldsymbol{B}'\boldsymbol{A}^-, & (\boldsymbol{C}-\boldsymbol{B}'\boldsymbol{A}^-\boldsymbol{B})^- \end{pmatrix}$$

$$= \begin{pmatrix} (\boldsymbol{A}-\boldsymbol{B}\boldsymbol{C}^-\boldsymbol{B}')^-, & -(\boldsymbol{A}-\boldsymbol{B}\boldsymbol{C}^-\boldsymbol{B}')^-\boldsymbol{B}\boldsymbol{C}^- \\ -\boldsymbol{C}\boldsymbol{B}'(\boldsymbol{A}-\boldsymbol{B}\boldsymbol{C}^-\boldsymbol{B}')^-, & \boldsymbol{C}^- + \boldsymbol{C}^-\boldsymbol{B}'(\boldsymbol{A}-\boldsymbol{B}\boldsymbol{C}^-\boldsymbol{B}')^-\boldsymbol{B}\boldsymbol{C}^- \end{pmatrix},$$

dostaneme první řádek $\boldsymbol{A}_{11}, \boldsymbol{A}_{12}$ g-inverzní matice k blokové matici ve výrazu (11):

$$\boldsymbol{A}_{11} = [(\boldsymbol{W}'\boldsymbol{V}^-\boldsymbol{W} - \boldsymbol{W}'\boldsymbol{V}^-\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{V}^-\boldsymbol{Z})^-\boldsymbol{Z}'\boldsymbol{V}^-\boldsymbol{W}]^- = (\boldsymbol{W}'\boldsymbol{V}^-\boldsymbol{Q}_Z^{V^-}\boldsymbol{W})^-.$$

Při úpravě jsme užili následující ekvivalenci

$$A\boldsymbol{B}^-\boldsymbol{C} \quad \text{nezávisí na volbě} \quad \boldsymbol{B}^- \quad \Leftrightarrow \ \mathscr{R}(\boldsymbol{A}') \subset \mathscr{R}(\boldsymbol{B}') \ \& \ \mathscr{R}(\boldsymbol{C}) \subset \mathscr{R}(\boldsymbol{B}),$$

(viz [3], Lemma 8, str.65). Z této ekvivalence vyplývá, že předpoklad (2) zaručuje nezávislost výrazů $\boldsymbol{Z}'\boldsymbol{V}^-\boldsymbol{Z}$, $\boldsymbol{Z}'\boldsymbol{V}^-\boldsymbol{W}$, $\boldsymbol{Z}'\boldsymbol{V}^-\boldsymbol{Y}$ na volbě matice $\boldsymbol{V}^-$. Zvolíme-li $\boldsymbol{V}^-$ p.d., máme $\boldsymbol{Q}_Z^{V^-} = \boldsymbol{I} - \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{V}^-\boldsymbol{Z})^-\boldsymbol{Z}'\boldsymbol{V}^-$.

$$\boldsymbol{A}_{12} = -(\boldsymbol{W}'\boldsymbol{V}^-\boldsymbol{Q}_Z^{V^-}\boldsymbol{W})^-\boldsymbol{W}'\boldsymbol{V}^-\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{V}^-\boldsymbol{Z})^-.$$

Tedy

$$\hat{\vartheta}_a = [(\boldsymbol{W}'\boldsymbol{V}^-\boldsymbol{Q}_Z^{V^-}\boldsymbol{W})^-\boldsymbol{W}'\boldsymbol{V}^- - (\boldsymbol{W}'\boldsymbol{V}^-\boldsymbol{Q}_Z^{V^-}\boldsymbol{W})^-\boldsymbol{W}'\boldsymbol{V}^-\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{V}^-\boldsymbol{Z})^-\boldsymbol{Z}'\boldsymbol{V}^-]\boldsymbol{Y}$$

$$= (\boldsymbol{W}'\boldsymbol{V}^-\boldsymbol{Q}_Z^{V^-}\boldsymbol{W})^-\boldsymbol{W}'\boldsymbol{V}^-\boldsymbol{Q}_Z^{V^-}\boldsymbol{Y}.$$

Tím jsme dokázali (8).

Dále

$$var(\widehat{\boldsymbol{p}'\vartheta}_a) = \boldsymbol{p}'[\boldsymbol{W}'\boldsymbol{V}^-\boldsymbol{Q}_Z^{V^-}\boldsymbol{W}]^-\boldsymbol{W}'\boldsymbol{V}^-\boldsymbol{Q}_Z^{V^-}\boldsymbol{V}(\boldsymbol{Q}_Z^{V^-})'\boldsymbol{V}^-\boldsymbol{W}[\boldsymbol{W}'(\boldsymbol{Q}_Z^{V^-})'\boldsymbol{V}^-\boldsymbol{W}]^-\boldsymbol{p}$$

$$= \boldsymbol{p}'[\boldsymbol{W}'V^-\boldsymbol{Q}_Z^{V^-}\boldsymbol{W}]^-[\boldsymbol{W}'\boldsymbol{V}^-\boldsymbol{Q}_Z^{V^-}\boldsymbol{W}][\boldsymbol{W}'\boldsymbol{V}^-\boldsymbol{Q}_Z^{V^-}\boldsymbol{W}]^-\boldsymbol{p} = \boldsymbol{p}'[\boldsymbol{W}'\boldsymbol{V}^-\boldsymbol{Q}_Z^{V^-}\boldsymbol{W}]^-\boldsymbol{p}.$$

Výpočty se zjednodušily užitím toho, že

$$\boldsymbol{A}\boldsymbol{B}^{-}\boldsymbol{B} = \boldsymbol{A} \quad \Leftrightarrow \quad \mathscr{R}(\boldsymbol{A}') \subset \mathscr{R}(\boldsymbol{B}'),$$

(viz [3], Lemma 7, str. 65),
toho, že

$$\boldsymbol{B} = \boldsymbol{B}' \text{ p.s.d., } \mathscr{R}(\boldsymbol{A}) \subset \mathscr{R}(\boldsymbol{B}) \quad \Rightarrow \quad \mathscr{R}(\boldsymbol{A}') = \mathscr{R}(\boldsymbol{A}'\boldsymbol{B}\boldsymbol{A}),$$

(viz [3], Lemma 4, str. 64) a toho, že

$$\boldsymbol{p} \in \mathscr{R}(\boldsymbol{B}) \subset \mathscr{R}(\boldsymbol{A}) \quad \Rightarrow \quad \boldsymbol{p}'\boldsymbol{A}^{-}\boldsymbol{A}\boldsymbol{A}^{-}\boldsymbol{p} = \boldsymbol{p}'\boldsymbol{A}^{-}\boldsymbol{p}.$$

Z předpokladu (2) plyne, že $\mathscr{R}(\boldsymbol{Z}') = \mathscr{R}(\boldsymbol{Z}'\boldsymbol{V}^{-}\boldsymbol{Z})$. V důkaze předpokládáme, že $\boldsymbol{p}'\vartheta \in \mathscr{E}_a$, tj. že platí $\boldsymbol{p} \in \mathscr{R}(\boldsymbol{W}'\boldsymbol{Q}_Z) = \mathscr{R}(\boldsymbol{W}'\boldsymbol{V}^{-}\boldsymbol{Q}_Z^{V^{-}}\boldsymbol{W})$ (viz (6)). □

Označení Obdobně jako v práci [4] označme symbolem $\mathscr{E}_0(\boldsymbol{V})$ třídu těch lineárních funkcí užitečných parametrů $\boldsymbol{p}'\vartheta \in \mathscr{E}_a$, jejichž BLUE za platnosti modelu $\mathscr{M}_a(\boldsymbol{V})$ mají stejný rozptyl jako BLUE za platnosti modelu $\mathscr{M}(\boldsymbol{V})$, tj.

$$\mathscr{E}_0(\boldsymbol{V}) = \{\boldsymbol{p}'\vartheta \in \mathscr{E}_a : var[\widehat{\boldsymbol{p}'\vartheta}] = var[\widehat{\boldsymbol{p}'\vartheta}_a]\}.$$

Je dokázáno, že platí (viz [4], Theorem 3.1.)

Věta 2

$$\mathscr{E}_0(\boldsymbol{V}) = \{\boldsymbol{p}'\vartheta : \boldsymbol{p} \in \mathscr{R}[\boldsymbol{W}'\boldsymbol{V}^{-}\boldsymbol{W}\boldsymbol{Q}_{W'V^{-}Z}]\}.$$

Poznámka 2 Pro dimenzi prostoru $\mathscr{E}_0(\boldsymbol{V})$ platí (viz [4], (3.14))

$$dim\mathscr{E}_0(\boldsymbol{V}) = r(\boldsymbol{W}) - r(\boldsymbol{W}'\boldsymbol{V}^{-}\boldsymbol{Z}).$$

Věnujme se nyní *odhadům rušivých parametrů.*

Funkce $\boldsymbol{q}'\beta$ je nevychýleně odhadnutelná, jestliže

$$\boldsymbol{q} \ \in \ \mathscr{R}[(\boldsymbol{Z}'\boldsymbol{Q}_W)].$$

Obdobným postupem jako při důkazu Věty 1 lze dokázat, že pro $\boldsymbol{V}^{-}$-LS odhady rušivých parametrů v modelu (1) platí

Věta 3

$$\widehat{\boldsymbol{q}'\beta}_a = \boldsymbol{q}'(\boldsymbol{Z}'\boldsymbol{V}^{-}\boldsymbol{Q}_W^{V^{-}}\boldsymbol{Z})^{-}\boldsymbol{Z}'\boldsymbol{V}^{-}\boldsymbol{Q}_W^{V^{-}}\boldsymbol{Y}, \quad \text{je-li} \quad \boldsymbol{q} \in \mathscr{R}(\boldsymbol{Z}'\boldsymbol{Q}_W), \tag{12}$$

$$var[\widehat{\boldsymbol{q}'\beta}_a] = \boldsymbol{q}'[\boldsymbol{Z}'\boldsymbol{V}^{-}\boldsymbol{Q}_W^{V^{-}}\boldsymbol{Z}]^{-}\boldsymbol{q}, \quad \text{je-li} \quad \boldsymbol{q} \in \mathscr{R}(\boldsymbol{Z}'\boldsymbol{Q}_W), \tag{13}$$

$$var\begin{pmatrix}\hat{\vartheta}_a \\ \hat{\beta}_a\end{pmatrix} = \begin{pmatrix}\boldsymbol{A}^{-}\boldsymbol{A}\boldsymbol{A}^{-}, & \boldsymbol{A}^{-}\boldsymbol{W}'\boldsymbol{V}^{-}\boldsymbol{Q}_Z^{V^{-}}\boldsymbol{V}\boldsymbol{V}^{-}\boldsymbol{Q}_W^{V^{-}}\boldsymbol{Z}\boldsymbol{B}^{-} \\ \boldsymbol{B}^{-}\boldsymbol{Z}'\boldsymbol{V}^{-}\boldsymbol{Q}_W^{V^{-}}\boldsymbol{V}\boldsymbol{V}^{-}\boldsymbol{Q}_Z^{V^{-}}\boldsymbol{W}\boldsymbol{A}^{-}, & \boldsymbol{B}^{-}\boldsymbol{B}\boldsymbol{B}^{-}\end{pmatrix}, \tag{14}$$

kde $\boldsymbol{A} = (\boldsymbol{W}'\boldsymbol{V}^{-}\boldsymbol{Q}_Z^{V^-}\boldsymbol{W})$, $\boldsymbol{B} = (\boldsymbol{Z}'\boldsymbol{V}^{-}\boldsymbol{Q}_W^{V^-}\boldsymbol{Z})$.

Jsou-li určeny odhady parametrů $\widehat{\kappa}_2^i = \widehat{\delta\kappa_2^i} + \kappa_{2,0}^i$, $\forall i = 1, \ldots, k$, je možno vyřešit úlohu formulovanou v úvodu:
určit takové t_i, pro které $\kappa_1^i(1-e^{-\kappa_2^i t_i}) = C\kappa_1^i$, kde $C \in (0,1)$ je vhodná konstanta.

$$\hat{t_i} = -\frac{ln(1-C)}{\widehat{\kappa_2^i}} = \frac{-ln(1-C)}{\widehat{\delta\kappa_2^i} + \kappa_{2,0}^i}.$$

V takovém čase je možno pokládat podloží za „usazené“ v i-tém bodě a v čase $t = max\{t_1, \ldots, t_k\}$ je možno pokračovat ve stavbě.

Protože $\hat{t_i}$ není lineární funkcí odhadu $\widehat{\kappa_2^i}$, nejde určit rozptyl $var(\hat{t_i})$ přímo. Můžeme psát

$$\hat{t_i} = \frac{-ln(1-C)}{\widehat{\kappa_2^i}} = \frac{-ln(1-C)}{(\widehat{\kappa_2^i} - \kappa_2^i) + \kappa_2^i} = \frac{-ln(1-C)}{\kappa_2^i}\frac{1}{1 + \frac{\widehat{\kappa_2^i} - \kappa_2^i}{\kappa_2^i}} =$$

$$= -\frac{ln(1-C)}{\kappa_2^i}\left[1 - \frac{\widehat{\kappa_2^i} - \kappa_2^i}{\kappa_2^i} + \frac{(\widehat{\kappa_2^i} - \kappa_2^i)^2}{(\kappa_2^i)^2} - \cdots\right]. \tag{15}$$

Vezmeme-li první dva členy řady, dostaneme náhradu rozptylu

$$\widetilde{var}(\hat{t_i}) = \left(\frac{ln(1-C)}{\kappa_2^i}\right)^2 \frac{var(\widehat{\kappa_2^i} - \kappa_2^i)}{(\kappa_2^i)^2} = \frac{(ln(1-C))^2}{(\kappa_2^i)^4} var(\widehat{\kappa_2^i}). \tag{16}$$

V praxi je nutno ve vzorci (16) za neznámý parametr κ_2^i dosadit jeho odhad $\widehat{\kappa_2^i}$. Tak dostáváme možnost nahradit rozptyl $var(\hat{t_i}\,)$ výrazem (tzv. error propagation law, viz [1])

$$\widetilde{var}[\hat{t_i}\,] = \frac{(ln(1-C))^2}{(\widehat{\kappa_2^i})^4} var[\widehat{\kappa_2^i}]. \tag{17}$$

Předpokládáme-li $\widehat{\kappa_2^i} \sim N(\kappa_2^i, var(\widehat{\kappa_2^i}))$, získáme užitím prvních tří členů řady (15) vyjádření

$$\widetilde{var}(\hat{\boldsymbol{t}}_i) = \frac{(ln(1-C))^2}{(\kappa_2^i)^4}\left(1 + \frac{2var(\widehat{\kappa_2^i})}{(\kappa_2^i)^2}\right) var(\widehat{\kappa_2^i}). \tag{18}$$

Pro určení rozptylu odhadu $\hat{t_i}$ lze také užít výsledků uvedených v článku [1]. Víme, že

$$\hat{t_i} = \frac{-ln(1-C)}{\widehat{\delta\kappa_2^i} + \kappa_{2,0}^i} = \frac{-ln(1-C)}{\kappa_{2,0}^i}\frac{1}{\left(1 + \frac{\widehat{\delta\kappa_2^i}}{\kappa_{2,0}^i}\right)} = f(\widehat{\delta\kappa_2^i}).$$

Lze psát

$$f(\delta\kappa_2^i) = \frac{-ln(1-C)}{\kappa_{2,0}^i} \cdot \frac{1}{\left(1+\frac{\delta\kappa_2^i}{\kappa_{2,0}^i}\right)} = \frac{-ln(1-C)}{\kappa_{2,0}^i}\left[1 - \frac{\delta\kappa_2^i}{\kappa_{2,0}^i} + \frac{(\delta\kappa_2^i)^2}{(\kappa_{2,0}^i)^2} - \cdots\right],$$

pro $\delta\kappa_2^i \in \mathscr{R} = \{u : |u| < \kappa_{2,0}^i\}$, $\kappa_{2,0}^i > 0$. Uvažujme v souladu s [1] náhodnou veličinu $\varepsilon_i = \widehat{\delta\kappa_2^i} - \delta\kappa_2^i$, která má obor hodnot $\boldsymbol{S}_i \subset \mathscr{R}$ a distribuční funkci $F_i(x)$.

Podle věty 3.3. práce [1] platí

$$var(\hat{t}_i) = var(f(\widehat{\delta\kappa_2^i})) =$$

$$= \sum_{r=1}^{\infty}\sum_{j=1}^{\infty} \frac{1}{r!}\left(\frac{d^r f(x)}{dx^r}\right)_{x=\delta\kappa_2^i} \frac{1}{j!}\left(\frac{d^j f(x)}{dx^j}\right)_{x=\delta\kappa_2^i} [E\varepsilon_i^{r+j} - E\varepsilon_i^r E\varepsilon_i^j] =$$

$$= \sum_{r=1}^{\infty}\sum_{j=1}^{\infty} \frac{(-1)^{r+j}(\frac{ln(1-C)}{\kappa_{2,0}^i})^2}{(\kappa_{2,0}^i)^{r+j}\left(1+\frac{\delta\kappa_2^i}{\kappa_{2,0}^i}\right)^{r+j+2}} [E\varepsilon_i^{r+j} - E\varepsilon_i^r E\varepsilon_i^j]. \tag{19}$$

Podle stejné věty platí pro vychýlení odhadu

$$E(\hat{t}_i) - t_i = E[f(\widehat{\delta\kappa_2^i})] - f(\delta\kappa_2^i) = \sum_{j=2}^{\infty} \frac{1}{j!}\left(\frac{d^j f(x)}{dx^j}\right)_{x=\delta\kappa_2^i} E\varepsilon_i^j. \tag{20}$$

Předpokládáme-li, že $\varepsilon_i \sim N(0, \sigma_i^2)$, potom $E[(\varepsilon_i)^{2r+1}] = 0$, $\forall r = 0, 1, 2, \ldots$, $E[(\varepsilon_i)^{2r}] = 1.3.5\ldots(2r-1)\sigma_i^{2r}$, $\forall r = 1, 2, \ldots$. Podle vztahu (19)

$$var(\hat{t}_i) = \left(\frac{-ln(1-C)}{\kappa_{2,0}^i}\right)^2 \left[\frac{1}{(\kappa_{2,0}^i)^2\left(1+\frac{\delta\kappa_2^i}{\kappa_{2,0}^i}\right)^4}\sigma_i^2 + \frac{8\sigma_i^4}{(\kappa_{2,0}^i)^4\left(1+\frac{\delta\kappa_2^i}{\kappa_{2,0}^i}\right)^6} + \right.$$

$$\left. + \frac{69\sigma_i^6}{(\kappa_{2,0}^i)^6\left(1+\frac{\delta\kappa_2^i}{\kappa_{2,0}^i}\right)^8} + \frac{696\sigma_i^8}{(\kappa_{2,0}^i)^8\left(1+\frac{\delta\kappa_2^i}{\kappa_{2,0}^i}\right)^{10}} + \cdots\right]. \tag{21}$$

Pro vychýlení odhadu v tomto případě (opět jsou všechny liché momenty veličiny ε_i nulové) platí podle (20)

$$E[(\hat{t}_i)] - t_i =$$

$$= \sum_{r=1}^{\infty} \frac{1}{(2r)!}\left(\frac{d^{2r} f(x)}{dx^{2r}}\right)_{x=\delta\kappa_2^i} E(\varepsilon_i)^{2r} = \sum_{r=1}^{\infty} \frac{\frac{-ln(1-C)}{\kappa_{2,0}^i}}{(\kappa_{2,0}^i)^{2r}\left(1+\frac{\delta\kappa_2^i}{\kappa_{2,0}^i}\right)^{2r+1}} E(\varepsilon_i^{2r}) =$$

$$= -\frac{ln(1-C)}{\kappa_{2,0}^i} \times$$

$$\times \left[\frac{\sigma_i^2}{(\kappa_{2,0}^i)^2 \left(1+\frac{\delta\kappa_2^i}{\kappa_{2,0}^i}\right)^3} + \frac{3\sigma_i^4}{(\kappa_{2,0}^i)^4 \left(1+\frac{\delta\kappa_2^i}{\kappa_{2,0}^i}\right)^5} + \frac{15\sigma_i^6}{(\kappa_{2,0}^i)^6 \left(1+\frac{\delta\kappa_2^i}{\kappa_{2,0}^i}\right)^7} + \cdots \right]. \quad (22)$$

Získáme-li pro daný bod odhady $\widehat{\kappa_1^i}$, $\widehat{\kappa_2^i}$, $\hat{\beta}_i$, $i = 1, \ldots, k$, lze určit odhad funkce která popisuje výšku i-tého měřeného bodu v čase t:

$$\widehat{\psi_i(t)} = \hat{\beta}_i - \widehat{\kappa_1^i}(1 - e^{-\widehat{\kappa_2^i} t}), \quad t > 0.$$

Pro rozptyl této funkce odhadů platí

$$var[\hat{\beta}_i - \widehat{\kappa_1^i}(1 - e^{-\widehat{\kappa_2^i} t})] = \left(\frac{\vartheta\hat{\psi}_i}{\vartheta\widehat{\kappa_1^i}}, \frac{\vartheta\hat{\psi}_i}{\vartheta\widehat{\kappa_2^i}}, \frac{\vartheta\hat{\psi}_i}{\vartheta\hat{\beta}_i}\right) \left[var \begin{pmatrix} \widehat{\kappa_1^i} \\ \widehat{\kappa_2^i} \\ \hat{\beta}_i \end{pmatrix} \right] \begin{pmatrix} \frac{\vartheta\hat{\psi}_i}{\vartheta\widehat{\kappa_1^i}} \\ \frac{\vartheta\hat{\psi}_i}{\vartheta\widehat{\kappa_2^i}} \\ \frac{\vartheta\hat{\psi}_i}{\vartheta\hat{\beta}_i} \end{pmatrix}, \ i = 1, \ldots, k. \quad (23)$$

3 Příklad

Uvažujme na staveništi dva body, ve kterých provedeme měření v pěti časových okamžicích. Je to pochopitelně málo z praktického pohledu, ale na tomto jednoduchém modelu lze ilustrovat teoretické úvahy. Vzhledem k technice měření můžeme položit $\boldsymbol{V} = \sigma^2 \boldsymbol{I}$, kde $\sigma^2 = 10^{-6}\, m$.

Abychom mohli vyčíslit hodnoty odhadů, předpokládejme, že „správná“ funkce popisující klesání podloží v prvním bodě má tvar

$$\varphi_1(t) = 1 - 0.2(1 - e^{-0.3t}), \quad t > 0,$$

tj. že skutečné hodnoty parametrů jsou $\kappa_1^1 = 0.2$, $\kappa_2^1 = 0.3$. Protože nemáme k disposici reálná měření, nasimulujeme hodnoty náhodných veličin $\varphi_1^*(t_j) \sim N(\varphi_1(t_j), \sigma^2)$, $j = 1, \ldots, 5$ a vezmeme

$$\boldsymbol{Y}_1^{(j)} = \varphi^*(t_j) + \kappa_{0,1}^1(1 - e^{-\kappa_{2,0}^1 t_j}), \quad j = 1, \ldots, 5.$$

Zvolme okamžiky, ve kterých měříme: $t_1 = 0.5,\ t_2 = 1,\ t_3 = 2,\ t_4 = 4,\ t_5 = 8$, dále zvolme $\kappa_{1,0}^1 = 0.1$, $\kappa_{2,0}^1 = 0.2$, $\kappa_{1,0}^2 = 0.3$, $\kappa_{2,0}^2 = 0.4$.

Nasimulovaná data:

$$\boldsymbol{Y}_1^{(1)} = 0.9812,\ \boldsymbol{Y}_1^{(2)} = 0.9663,\ \boldsymbol{Y}_1^{(3)} = 0.9398,\ \boldsymbol{Y}_1^{(4)} = 0.9159,\ \boldsymbol{Y}_1^{(5)} = 0.8974.$$

$$(\boldsymbol{W},\boldsymbol{Z}) = \begin{pmatrix} -0.0952 & -0.0452 & 0 & 0 & 1 & 0 \\ -0.1813 & -0.0819 & 0 & 0 & 1 & 0 \\ -0.3297 & -0.1341 & 0 & 0 & 1 & 0 \\ -0.5507 & -0.1797 & 0 & 0 & 1 & 0 \\ -0.7981 & -0.1615 & 0 & 0 & 1 & 0 \\ 0 & 0 & -0.1813 & -0.1228 & 0 & 1 \\ 0 & 0 & -0.3297 & -0.2011 & 0 & 1 \\ 0 & 0 & -0.5507 & -0.2696 & 0 & 1 \\ 0 & 0 & -0.7981 & -0.2423 & 0 & 1 \\ 0 & 0 & -0.9592 & -0.0978 & 0 & 1 \end{pmatrix},$$

$$\delta = \begin{pmatrix} \delta\kappa_1^1 \\ \delta\kappa_2^1 \\ \delta\kappa_1^2 \\ \delta\kappa_2^2 \end{pmatrix},\ \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}.$$

a) Pro určení odhadů *užitečných parametrů v malém modelu* vypočteme

$$(\boldsymbol{W}'\boldsymbol{W})^-\boldsymbol{W}'$$

$$= \begin{pmatrix} 0.6300 & 1.0482 & 1.3683 & 0.6800 & -2.6007 \\ -2.6867 & -4.5490 & -6.2577 & -4.4384 & 7.0015 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0.2426 & 0.3429 & 0.2795 & -0.2514 & -1.1576 \\ -1.2723 & -1.9385 & -2.1180 & -0.5625 & 2.5908 \end{pmatrix}.$$

$$(\boldsymbol{W}'\boldsymbol{W})^- = \begin{pmatrix} 10.5942 & -36.2509 & 0 & 0 \\ -36.2509 & 135.7916 & 0 & 0 \\ 0 & 0 & 1.6578 & -4.4233 \\ 0 & 0 & -4.4233 & 16.8913 \end{pmatrix}.$$

Zvolme $p_1 = (1,0,0,0)'$, $\boldsymbol{p}_2 = (0,1,0,0)'$, $\boldsymbol{p}_3 = (0,0,1,0)$', $\boldsymbol{p}_4 = (0,0,0,1)'$. To, že $\boldsymbol{p}_j \in \mathscr{R}(\boldsymbol{W}')$, $\forall j = 1,\ldots,4$, tj. že funkce $\boldsymbol{p}_j'\vartheta, j = 1,\ldots,4$ jsou odhadnutelné v malém modelu (viz vztah (5)), ověříme pomocí identity

$$\boldsymbol{p} \in \mathscr{R}(\boldsymbol{A}) \quad \Leftrightarrow \quad \boldsymbol{A}\boldsymbol{A}^-\boldsymbol{p} = \boldsymbol{p}. \tag{24}$$

Podle (7), (9) platí

$$\widehat{\boldsymbol{p}_1'\vartheta} = \widehat{\delta\kappa_1^1}$$

$$= 0.6300\boldsymbol{Y}_1^{(1)} + 1.0482\boldsymbol{Y}_1^{(2)} + 1.3683\boldsymbol{Y}_1^{(3)} + 0.6800\boldsymbol{Y}_1^{(4)} - 2.6007\boldsymbol{Y}_1^{(5)} = 1.20592,$$

$$\widehat{\boldsymbol{p}_2'\vartheta} = \widehat{\delta\kappa_2^1}$$

$$= -2.6867\boldsymbol{Y}_1^{(1)} - 4.5490\boldsymbol{Y}_1^{(2)} - 6.2577\boldsymbol{Y}_1^{(3)} - 4.4384\boldsymbol{Y}_1^{(4)} + 7.0015\boldsymbol{Y}_1^{(5)} = -10.69499,$$

$$var[\widehat{\boldsymbol{p}_1'\vartheta}] = var[\widehat{\delta\kappa_1^1}] = 10.5942\sigma^2,$$

$$var[\widehat{\boldsymbol{p}_2'\vartheta}] = var[\widehat{\delta\kappa_2^1}] = 135.7916\sigma^2.$$

Je zřejmé, že získané odhady

$$\widehat{\kappa_1^1} = \widehat{\delta\kappa_1^1} + 0.1 = 1.30592, \quad \widehat{\kappa_2^1} = \widehat{\delta\kappa_2^1} + 0.2 = -10.49499,$$

nejsou v praxi upotřebitelné, protože příslušné funkce nelze z malého odhadu odhadovat.

b) Pro odhady *užitečných parametrů ve velkém modelu* určíme

$$(\boldsymbol{W}'\boldsymbol{Q}_Z\boldsymbol{W})^-\boldsymbol{W}'\boldsymbol{Q}_Z = \begin{pmatrix} -0.4896 & 0.4841 & 1.5410 & 1.3086 & -2.8441 \\ 8.1833 & 0.9277 & -7.9345 & -10.5406 & 9.3641 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0.9427 & 0.5644 & 0.0124 & -0.5813 & -0.9382 \\ 2.9758 & -0.5945 & -3.7393 & -2.5645 & 3.9225 \end{pmatrix},$$

$$(\boldsymbol{W}'\boldsymbol{Q}_Z\boldsymbol{W})^- = \begin{pmatrix} 12.6500 & -56.2103 & 0 & 0 \\ -56.2103 & 329.5744 & 0 & 0 \\ 0 & 0 & 2.4225 & 0.2342 \\ 0 & 0 & 0.2342 & 45.1539 \end{pmatrix}.$$

Funkce $\boldsymbol{p}_j'\vartheta, j = 1, \ldots, 4$ jsou odhadnutelné i ve velkém modelu. Podle (8) platí

$$\widehat{\boldsymbol{p}_1'\vartheta} = \widehat{\delta\kappa_1^1}$$

$$= -0.4896\boldsymbol{Y}_1^{(1)} + 0.48411\boldsymbol{Y}_1^{(2)} + 1.5410\boldsymbol{Y}_1^{(3)} + 1.3086\boldsymbol{Y}_1^{(4)} - 2.8441\boldsymbol{Y}_1^{(5)} = 0.08184,$$

$$\widehat{\kappa_1^1} = \widehat{\delta\kappa_1^1} + \kappa_{1,0}^1 = 0.08184 + 0.1 = 0.18184,$$

$$\widehat{\boldsymbol{p}_2'\vartheta} = \widehat{\delta\kappa_2^1}$$

$$= 8.1833\boldsymbol{Y}_1^{(1)} + 0.9277\boldsymbol{Y}_1^{(2)} - 7.9345\boldsymbol{Y}_1^{(3)} - 10.5406\boldsymbol{Y}_1^{(4)} + 9.3641\boldsymbol{Y}_1^{(5)} = 0.21852,$$

$$\widehat{\kappa_2^1} = \widehat{\delta\kappa_2^1} + \kappa_{2,0}^1 = 0.21852 + 0.2 = 0.41852.$$

$$var[\widehat{\boldsymbol{p}_1'\vartheta}] = var[\widehat{\delta\kappa_1^1}] = 12.6500\sigma^2,$$

$$var[\widehat{\boldsymbol{p}_2'\vartheta}] = var[\widehat{\delta\kappa_2^1}] = 329.5744\sigma^2.$$

c) *Odhady rušivých parametrů.*
Užijeme tvrzení Věty 3, vypočteme

$$(\boldsymbol{Z}'\boldsymbol{Q}_W\boldsymbol{Z})^-\boldsymbol{Z}'\boldsymbol{Q}_W = \begin{pmatrix} 0.9945 & 0.5011 & -0.1534 & -0.5583 & 0.2161 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 1.2871 & 0.4072 & -0.4912 & -0.6066 & 0.4035 \end{pmatrix},$$

diagonální prvky matice $(\boldsymbol{Z}'\boldsymbol{Q}_W\boldsymbol{Z})^-$ jsou tyto: 1.6220; 2.5945.
Zvolme $\boldsymbol{q} = (1,0)'$, $\boldsymbol{q}_2 = (0,1)'$. Parametrické funkce $\boldsymbol{q}_j'\beta, j = 1,2$, jsou odhadnutelné. Tedy podle (13),(14)

$$\widehat{\boldsymbol{q}_1'\beta} = \hat{\beta}_1 = 0.9945\boldsymbol{Y}_1^{(1)} + 0.5011\boldsymbol{Y}_1^{(2)} - 0.1534\boldsymbol{Y}_1^{(3)} - 0.5583\boldsymbol{Y}_1^{(4)} + 0.2161\boldsymbol{Y}_1^{(5)} = 0.9985,$$

$$\widehat{\boldsymbol{q}_2'\beta} = \hat{\beta}_2 = 1.2871\boldsymbol{Y}_2^{(1)} + 0.4072\boldsymbol{Y}_2^{(2)} - 0.4912\boldsymbol{Y}_2^{(3)} - 0.6066\boldsymbol{Y}_2^{(4)} + 0.4035\boldsymbol{Y}_2^{(5)},$$

$$var[\widehat{\boldsymbol{q}_1'\beta}] = var[\hat{\beta}_1] = 1.6220\sigma^2, \quad var[\widehat{\boldsymbol{q}_2'\beta}] = var[\hat{\beta}_2] = 2.5945\sigma^2.$$

Poznámka 3 Viděli jsme, že se rozptyly odhadů funkcí $\boldsymbol{p}_1'\vartheta, \boldsymbol{p}_2'\vartheta$ ve velkém a malém modelu liší, i když nejde o řádovou diferenci. To signalizuje, že tyto funkce nepatří do třídy $\mathscr{E}_0(\sigma^2\boldsymbol{I})$. Podle Věty 2

$$\mathscr{E}_0(\sigma^2\boldsymbol{I}) = \{\boldsymbol{p}'\vartheta : \boldsymbol{p} \in \mathscr{R}(\boldsymbol{W}'\boldsymbol{W}\boldsymbol{Q}_{W'Z})\},$$

kde

$$\boldsymbol{W}'\boldsymbol{W}\boldsymbol{Q}_{W'Z} = \begin{pmatrix} 0.0126 & -0.0410 & 0 & 0 \\ 0.0013 & -0.0042 & 0 & 0 \\ 0 & 0 & 0.0414 & -0.1250 \\ 0 & 0 & -0.0068 & 0.0206 \end{pmatrix}.$$

Užitím identity (24) ověříme, že rovnost

$$\boldsymbol{W}'\boldsymbol{W}\boldsymbol{Q}_{W'Z}(\boldsymbol{W}'\boldsymbol{W}\boldsymbol{Q}_{W'Z})^-\boldsymbol{p}_j = \boldsymbol{p}_j,$$

neplatí ani pro jeden index $j = 1,\ldots,4$. Např. pro $j = 1$ je na levé straně tohoto vztahu vektor $(0.9895;\ 0.1019;\ 0;\ 0)'$. To znamená, že parametrické funkce $\boldsymbol{p}_j'\vartheta$, $j = 1,\ldots,4$, skutečně nepatří do třídy $\mathscr{E}_0(\sigma^2\boldsymbol{I})$.

$r(\boldsymbol{W}'\boldsymbol{W}\boldsymbol{Q}_{W'Z}) = 2$, což je ve shodě s tvrzením v Poznámce 2.

Uvažujme nějakou funkci $\boldsymbol{p}'\vartheta$ užitečných parametrů, která patří do $\mathscr{E}_0(\sigma^2\boldsymbol{I})$. Zvolme např. $\boldsymbol{p} = (0.0126, 0.0013, 0, 0)'$ tj. první sloupec matice $\boldsymbol{W}'\boldsymbol{W}\boldsymbol{Q}_{W'Z}$. Potom $\boldsymbol{p}'\vartheta = 0.0126\delta\kappa_1^1 + 0.0013\delta\kappa_2^1$ patří do $\mathscr{E}_0(\sigma^2\boldsymbol{I})$ a je to parametrická funkce, jejíž odhad v malém i velkém modelu má stejný rozptyl. Přesvědčíme se o tom výpočtem, užijeme simulované hodnoty $\boldsymbol{Y}_1^{(i)}, i = 1,\ldots,5$. Dostaneme
a) odhad v malém modelu (viz (7),(9))

$$\widehat{\boldsymbol{p}'\vartheta} = 0.0013188, \quad var[\widehat{\boldsymbol{p}'\vartheta}] = 0.00072938\sigma^2,$$

b) odhad ve velkém modelu (viz (8), (10))

$$\widehat{\boldsymbol{p}'\vartheta} = 0.0013189, \quad var[\widehat{\boldsymbol{p}'\vartheta}] = 0.00072965\sigma^2.$$

„Skutečná“ hodnota této parametrické funkce pro $\delta\kappa_1^1 = \kappa_1^1 - \kappa_{1,0}^1 = 0.2 - 0.1 = 0.1$, $\delta\kappa_2^1 = \kappa_2^1 - \kappa_{2,0}^1 = 0.3 - 0.2 = 0.1$ je rovna 0.0013941. Dostali jsme velmi dobré hodnoty odhadů a prokázali jsme shodu $var[\widehat{\boldsymbol{p}'\vartheta}]$ v obou modelech.

Jsou-li určeny odhady $\widehat{\kappa_1^i}$, $\widehat{\kappa_2^i}$, $\hat{\beta}_i$, $i = 1, 2$, lze stanovit *odhad funkce popisující výšku i-tého bodu v čase t*:

$$\hat{\psi}_i(t) = \hat{\beta}_i - \widehat{\kappa_1^i}(1 - e^{-\widehat{\kappa_2^i t}}), \ i = 1, 2, \ t > 0.$$

Podle (23) platí pro rozptyl tohoto odhadu

$$var[\hat{\beta}_i - \widehat{\kappa_1^i}(1 - e^{-\widehat{\kappa_2^i t}})] = var(\widehat{\kappa_1^i})(1 - e^{-\widehat{\kappa_2^i} t})^2 + var(\widehat{\kappa_2^i}) t^2 e^{-2\widehat{\kappa_2^i} t} (\widehat{\kappa_1^i})^2 + var(\widehat{\beta_i}) +$$

$$+2cov(\widehat{\kappa_1^i}, \widehat{\kappa_2^i}) t\widehat{\kappa_1^i} e^{-\widehat{\kappa_2^i} t}(1 - e^{-\widehat{\kappa_2^i} t}) - 2cov(\widehat{\kappa_1^i}, \hat{\beta}_i)(1 - e^{-\widehat{\kappa_2^i} t}) - 2cov(\widehat{\kappa_2^i}, \hat{\beta}_i) t\widehat{\kappa_1^i} e^{-\widehat{\kappa_2^i} t},$$

$$i = 1, 2.$$

Příslušné rozptyly a kovariance najdeme v matici (viz (14))

$$var \begin{pmatrix} \widehat{\delta\kappa_1^1} \\ \widehat{\delta\kappa_2^1} \\ \widehat{\delta\kappa_1^2} \\ \widehat{\delta\kappa_2^2} \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \sigma^2 \begin{pmatrix} 12.6500 & -56.2103 & 0 & 0 & -1.8261 & 0 \\ -56.2103 & 329.5744 & 0 & 0 & 17.7289 & 0 \\ 0 & 0 & 2.4253 & 0.2342 & 0 & 1.4111 \\ 0 & 0 & 0.2342 & 45.1540 & 0 & 8.5632 \\ -1.8261 & 17.7289 & 0 & 0 & 1.6220 & 0 \\ 0 & 0 & 1.4111 & 8.5632 & 0 & 2.5945 \end{pmatrix}.$$

Např. pro první bod dostaneme

$$var[\hat{\beta}_1 - \widehat{\kappa_1^1}(1 - e^{-\widehat{\kappa_2^1} t})] = \sigma^2 \Big[12,6500(1 - e^{-\widehat{\kappa_2^1} t})^2 + 329,5744 t^2 e^{-2\widehat{\kappa_2^1} t} (\widehat{\kappa_1^1})^2 + 1,6220$$

$$-2 \cdot 56,2103 t\widehat{\kappa_1^1} e^{-\widehat{\kappa_2^1} t}(1 - e^{-\widehat{\kappa_2^1} t}) + 2 \cdot 1,8261(1 - e^{-\widehat{\kappa_2^1} t}) - 2 \cdot 17.7289 t\widehat{\kappa_1^1} e^{-\widehat{\kappa_2^1} t}\Big].$$

Pro ilustraci uveďme, že např. pro čas $t_5 = 8$ platí

$$\hat{\psi}_1(8) = 0.8230514, \quad var[\hat{\psi}_1(8)] = 10.4246\sigma^2, \quad \sqrt{var[\hat{\psi}_1(8)]} = 0.00322\, m.$$

Na základě výsledků uvedených na konci odstavce 2 určíme, kdy lze pokládat v prvním měřeném bodě podloží za usazené. Zvolme $C = 0.95$.

$$\hat{t_1} = -\frac{ln 0.95}{0.41852} = 7.15791,$$

tj. po osmi dnech by bylo možno pokračovat ve stavbě (v praxi musíme uvažovat situaci ve všech měřených bodech). Přesnost tohoto odhadu zjistíme určením jeho směrodatné odchylky. Podle (17)

$$\widetilde{var}[\hat{t_1}] = \frac{(ln0.05)^2}{(0.41852)^4} 329.5744\sigma^2 = 96403,8992\sigma^2,$$

$$\sqrt{\widetilde{var}[\hat{t_1}]} = 0.31048.$$

Tento výsledek upřesníme podle (18):

$$\widetilde{var}[\hat{t_1}] = 96766,6811\sigma^2, \quad \sqrt{\widetilde{var}[\hat{t_1}]} = 0.3110734.$$

Odhad směrodatné odchylky se změnil jen nepatrně. Ke stejným numerickým výsledkům dojdeme užitím prvních dvou sčítanců ve vztahu (21), kde vezmeme $\sigma_1^2 = var[\widehat{\delta\kappa_2^1}] = 329,5744\sigma^2$: $var[\hat{t_1}] = 97855.0268\sigma^2$, $\sqrt{var[\hat{t_1}]} = 0.312817$.
Pro vychýlení odhadu (viz (22), užity první dva zlomky) pro stejné σ_1^2 platí

$$E(\hat{t_1}) - t_1 = 1471.6221\sigma^2 = 0.001471.$$

Závěr: ukázalo se, že původní představa, že bychom se mohli omezit pouze na malý model není správná. K odhadu užitečných parametrů potřebujeme celou informaci, aby dosažené výsledky byly aplikovatelné v praxi.

Reference

[1] Kubáček, L.: *Nonlinear error propagation law.* Applications of mathematics **41**, 5 (1996), 329–345.

[2] Kubáček, L.: *Foundations of Estimation Theory.* Elsevier, Amsterdam, Oxford, New York, Tokyo, 1988.

[3] Kubáčková, L., Kubáček, J.: *Elimination Transformation of an Observation Vector preserving Information on the First and Second Order Parameters.* Technical Report, Institute of Geology, University of Stuttgart, No 11, (1990), 1–71.

[4] Nordstrőm, K., Fellman, J.: *Characterizations and Dispersion-Matrix Robustness of Efficiently Estimable Parametric Functionals in Linear Models with Nuisance Parameters.* Linear Algebra and its Applications **127** (1990), 341–361.

Univ. Palacki. Olomuc., Fac. rer. nat.,
Dept of Math. Anal. and Appl. of Math.
ODAM (1999) 59–84

Řešení Kuhn–Tuckerových soustav rovnic kontaktní úlohy *

Horymír NETUKA

Department of Mathematical Analysis and Applications of Mathematics,
Faculty of Science, Palacký University,
Tomkova 40, 779 00 Olomouc, Czech Republic
e-mail: netuka@risc.upol.cz

Abstrakt

Při realizaci řešení kontaktní problematiky narážíme na nutnost řešit speciální soustavy rovnic, které vznikají v podstatě při aplikaci Kuhn–Tuckerových podmínek v příslušné úloze konvexního kvadratického programování. Algoritmus jejich řešení má zásadní důležitost z hlediska efektivnosti numerické realizace celého řešení, a to zejména u velkých úloh, které vznikají v praktických aplikacích. V této práci se studují Kuhn–Tuckerovy soustavy, jejichž vedoucí bloková matice na diagonále je pozitivně semidefinitní. Rekapitulují se zde stávající postupy řešení a předkládá se nový, jenž je založený na výsledcích publikovaných v [24]. Tímto způsobem je pak možné řešit i singulární Kuhn–Tuckerovy soustavy. Na závěr jsou uvedeny ilustrační příklady. Uvažovaná problematika má význam nejen pro kontaktní úlohy, ale i v řadě problémů optimalizace nebo ve statistice (viz [21]).

*Práce byla vypracována s podporou grantu GA ČR č. 105/99/1651.

1 Úvod

1.1 Kontaktní úloha bez tření

Uvažujme kontakt tělesa reprezentovaného oblastí Ω s dokonale tuhou podložkou, přičemž neopustíme meze teorie lineární pružnosti. Označme $\mathbf{u}$ pole posunutí bodů tělesa Ω, $\boldsymbol{\varepsilon}$ tenzor malých deformací, $\boldsymbol{\tau}$ tenzor napětí, $\mathbf{T}$ vektor napětí, Γ_u bude část hranice s předepsanými nulovými posunutími, Γ_P část hranice, na níž působí povrchové síly $\mathbf{P}$, Γ_K kontaktní zóna, na níž platí podmínky nepronikání, a Γ_0 je ta část hranice, na níž jsou předepsány podmínky bilaterálního kontaktu. Zformulujeme-li tuto úlohu ve tvaru variační nerovnice

$$\mathbf{u} \in \mathbf{K}: \ a(\mathbf{u}, \mathbf{v}-\mathbf{u}) \geq L(\mathbf{v}-\mathbf{u}) \qquad \forall \mathbf{v} \in \mathbf{K}, \tag{1}$$

kde

$$\mathbf{K} = \{\mathbf{v} \in \mathbf{V} \mid v_n \leq 0 \text{ na } \Gamma_K\} \tag{2}$$

$$\mathbf{V} = \{\mathbf{v} \in (H^1(\Omega))^2 \mid v_i = 0, \ i = 1,2 \text{ na } \Gamma_u, \ v_n = 0 \text{ na } \Gamma_0\} \tag{3}$$

$$a(\mathbf{u}, \mathbf{v}) = \int_\Omega \tau_{ij}(\mathbf{u}) \varepsilon_{ij}(\mathbf{v}) \, \mathrm{d}\mathbf{x} \tag{4}$$

$$L(\mathbf{v}) = \int_\Omega F_i v_i \, \mathrm{d}\mathbf{x} + \int_{\Gamma_P} P_i v_i \, \mathrm{d}s, \tag{5}$$

resp. ve variační podobě

$$\begin{cases} \text{nalézt } \mathbf{u} \in \mathbf{K} \\ J(\mathbf{u}) = \min\limits_{\mathbf{v} \in \mathbf{K}} J(\mathbf{v}), \end{cases} \tag{6}$$

přičemž J značí kvadratický funkcionál potenciální energie

$$J(\mathbf{v}) = \frac{1}{2} a(\mathbf{v}, \mathbf{v}) - L(\mathbf{v}), \tag{7}$$

můžeme k jejímu řešení použít metodu konečných prvků. Takto získáme konečnědimensionální úlohu, jejíž formální vyjádření lze psát buďto ve variačním tvaru

$$\begin{cases} \text{nalézt } \mathbf{u}_h \in \mathbf{K}_h \text{ tak, že} \\ J(\mathbf{u}_h) = \min\limits_{\mathbf{v}_h \in \mathbf{K}_h} J(\mathbf{v}_h), \end{cases} \tag{8}$$

nebo ekvivalentně ve tvaru nerovnice

$$\begin{cases} \text{nalézt } \mathbf{u}_h \in \mathbf{K}_h \text{ tak, že} \\ a(\mathbf{u}_h, \mathbf{v}_h - \mathbf{u}_h) \geq L(\mathbf{v}_h - \mathbf{u}_h) \quad \forall \, \mathbf{v}_h \in \mathbf{K}_h, \end{cases} \tag{9}$$

kde $\mathbf{K}_h$ je jistá aproximace množiny $\mathbf{K}$ (podrobně viz např. [17] nebo [20]).

V algebraickém zápisu se pak jedná o úlohu

$$\begin{cases} \text{nalézt } \widehat{\mathbf{x}} \in \mathbb{K} \text{ tak, že} \\ f(\widehat{\mathbf{x}}) = \min\limits_{\mathbf{x} \in \mathbb{K}} f(\mathbf{x}), \end{cases} \tag{10}$$

kde

$$f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{C}\mathbf{x} - \mathbf{d}^T\mathbf{x} \tag{11}$$

je konvexní funkce a

$$\mathbb{K} = \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{A}\mathbf{x} \leq \mathbf{o}, \quad \mathbf{B}\mathbf{x} = \mathbf{o}\}. \tag{12}$$

je konvexní množina. Symetrická matice $\mathbf{C}$ je maticí tuhosti daného tělesa Ω a je obecně pouze semidefinitní, vektor $\mathbf{d}$ je vektorem zatížení tělesa Ω, $\mathbf{A}$ je matice vazeb vzniklých diskretizací podmínek nepronikání a $\mathbf{B}$ je matice vazeb vzniklých diskretizací bilaterálních podmínek. V případě tzv. rozšiřující se kontaktní zóny má množina (12) obecnější tvar

$$\mathbb{K} = \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{A}\mathbf{x} \leq \mathbf{b}, \ \mathbf{B}\mathbf{x} = \mathbf{o}\}, \tag{13}$$

když vektor $\mathbf{b}$ reprezentuje mezeru v kontaktní zóně. Je zřejmé, že (10) je úloha konvexního kvadratického programování (nikoli však ryze konvexní) o obecně velmi vysokém počtu neznámých n, jenž i o několik řádů převyšuje celkový počet vazeb m. K jejímu řešení lze využít dnes všeobecně používanou metodu aktivní množiny, jejíž kroky vyžadují efektivní řešení tzv. Kuhn–Tuckerových podmínek (podrobněji viz např. [11] a [13], popř. [16]).

1.2 Kuhn–Tuckerovy soustavy

Uvažujme symetrickou soustavu lineárních rovnic tvaru

$$\begin{pmatrix} \mathbf{C} & \mathbf{A}^T \\ \mathbf{A} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{x} \\ \boldsymbol{\lambda} \end{pmatrix} = \begin{pmatrix} \mathbf{d} \\ \mathbf{b} \end{pmatrix}, \tag{14}$$

kde $\mathbf{C} \in \mathbb{R}^{n\times n}$ je symetrická matice, $\mathbf{A} \in \mathbb{R}^{m\times n}$, $m \leq n$, $\mathbf{d} \in \mathbb{R}^n$, $\mathbf{b} \in \mathbb{R}^m$. Úloha nalézt řešení $\{\mathbf{x}, \boldsymbol{\lambda}\} \in \mathbb{R}^n \times \mathbb{R}^m$ soustavy (14) bude předmětem dalších úvah.

Takovéto soustavy vznikají použitím Kuhn–Tuckerových podmínek ([13], [11]) v úloze kvadratického programování, kdy jde o minimalizaci kvadratické funkce

$$f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{C}\mathbf{x} - \mathbf{d}^T\mathbf{x}, \quad \mathbf{x} \in \mathbb{R}^n, \tag{15}$$

za omezujících lineárních podmínek tvaru rovností

$$\mathbf{A}\mathbf{x} = \mathbf{b}. \tag{16}$$

O rovnicích vystupujících v podmínkách (16) se zpravidla předpokládá, že jsou lineárně nezávislé. V tom případě je zřejmě hodnost matice $\mathbf{A}$, kterou označíme $\mathrm{rank}(\mathbf{A})$, rovna m.

Definujme prostor sloupcových vektorů matice $\mathbf{A}$ jako

$$\mathcal{R}(\mathbf{A}) = \{\mathbf{x} \in \mathbb{R}^m \mid \exists \mathbf{y} \in \mathbb{R}^n : \ \mathbf{A}\mathbf{y} = \mathbf{x}\}$$

a nulový prostor matice $\mathbf{A}$

$$\mathcal{N}(\mathbf{A}) = \{\mathbf{x} \in \mathbb{R}^n |\ \mathbf{A}\mathbf{x} = \mathbf{o}\}.$$

Platí

$$\dim \mathcal{R}(\mathbf{A}) = \mathrm{rank}(\mathbf{A}), \qquad \dim \mathcal{N}(\mathbf{A}) = n - \mathrm{rank}(\mathbf{A}). \tag{17}$$

O matici $\mathbf{A}$ se v případě, kdy $\mathrm{rank}(\mathbf{A}) = m$, řekne, že má plnou hodnost.

Podmínky řešitelnosti jsou v publikacích z oblasti optimalizace uváděny v dvojím znění. První je založeno na výzkumech N. I. M. Goulda a má tuto podobu:

Věta 1. Nechť $\mathrm{rank}(\mathbf{A}) = m$. Označme k_- počet záporných a k_0 počet nulových vlastních čísel matice

$$\mathbf{K} = \begin{pmatrix} \mathbf{C} & \mathbf{A}^T \\ \mathbf{A} & \mathbf{0} \end{pmatrix}.$$

Potom platí

(i) soustava (14) má právě jedno řešení tehdy a jen tehdy, když $k_- = m$, $k_0 = 0$;

(ii) soustava (14) má nekonečně mnoho řešení tehdy a jen tehdy, když $k_- = m$, $k_0 > 0$ a soustava je konsistentní, tj. její pravá strana leží v $\mathcal{R}(\mathbf{K})$;

(iii) soustava (14) nemá řešení v zbývajících případech, tj. když $k_- > m$ nebo když není konsistentní.

Důkaz viz [15].

Druhé často citované tvrzení o řešitelnosti uvedeme později v odstavci o metodách řešení soustavy (14).

V dalším se budeme zajímat zejména o případ, kdy matice $\mathbf{C}$ ze soustavy (14) je pozitivně semidefinitní. Pro matici $\mathbf{C}$ tedy platí

$$\mathbf{x}^T\mathbf{C}\mathbf{x} \geq 0 \qquad \forall\, \mathbf{x} \in \mathbb{R}^n. \tag{18}$$

Soustava (14) ovšem sama zůstává i nadále nedefinitní.

Lemma 1. Pro pozitivně semidefinitní matici $\mathbf{C}$ platí $\mathbf{x}^T\mathbf{C}\mathbf{x} = 0$ právě tehdy, když $\mathbf{x} \in \mathcal{N}(\mathbf{C})$.

Důkaz Nutnou a postačující podmínkou pro určení bodu minima $\hat{\mathbf{x}}$ konvexní kvadratické funkce $\mathbf{x}^T\mathbf{C}\mathbf{x}$ je splnění rovnice $\mathbf{C}\hat{\mathbf{x}} = \mathbf{o}$, což značí, že $\hat{\mathbf{x}} \in \mathcal{N}(\mathbf{C})$. Nyní už stačí jen vzít do úvahy, že $\hat{\mathbf{x}}^T\mathbf{C}\hat{\mathbf{x}} = 0$. □

Lemma 2. Nulový prostor matice $\mathbf{K}$ je tvořen vektory $\begin{pmatrix} \mathbf{p} \\ \mathbf{q} \end{pmatrix}$, kde $\mathbf{p}$ je libovolný prvek z $\mathcal{N}(\mathbf{C}) \cap \mathcal{N}(\mathbf{A})$ a $\mathbf{q}$ libovolný prvek z $\mathcal{N}(\mathbf{A}^T)$.

Důkaz Uvedené vektory padnou do $\mathcal{N}(\mathbf{K})$, neboť

$$\mathbf{K}\begin{pmatrix} \mathbf{p} \\ \mathbf{q} \end{pmatrix} = \begin{pmatrix} \mathbf{C}\mathbf{p} + \mathbf{A}^T\mathbf{q} \\ \mathbf{A}\mathbf{p} \end{pmatrix} = \begin{pmatrix} \mathbf{o} \\ \mathbf{o} \end{pmatrix}. \tag{19}$$

Nyní ukážeme, že v $\mathcal{N}(\mathbf{K})$ se nacházejí právě jen tyto vektory.

Vezměme tedy vektor tvaru $\begin{pmatrix}\mathbf{p}\\ \mathbf{q}\end{pmatrix}$, kde $\mathbf{p}$ je jakýkoliv prvek z $\mathbb{R}^n$ neležící v $\mathcal{N}(\mathbf{C}) \cap \mathcal{N}(\mathbf{A})$ a $\mathbf{q}$ jakýkoliv prvek z $\mathbb{R}^m$ neležící v $\mathcal{N}(\mathbf{A}^T)$. Dle (19) pak musí platit

$$\mathbf{C}\mathbf{p} + \mathbf{A}^T\mathbf{q} = \mathbf{o},$$

přičemž $\mathbf{p} \in \mathcal{N}(\mathbf{A})$. Vynásobíme tuto rovnici skalárně vektorem $\mathbf{p}$, který je dle učiněného předpokladu nenulový, a dostaneme

$$\mathbf{p}^T\mathbf{C}\mathbf{p} + (\mathbf{A}\mathbf{p})^T\mathbf{q} = \mathbf{o},$$

odkud plyne dle lemmatu 1, že $\mathbf{C}\mathbf{p} = \mathbf{o}$. To však značí, že $\mathbf{p} \in \mathcal{N}(\mathbf{C})$, což dává spor. □

V dalších úvahách se omezíme na již zmiňovaný případ, kdy matice $\mathbf{A}$ má plnou hodnost. Dalším předpokladem, který učiníme, bude to, že neostrou nerovnost $n \geq m$ nahradíme ostrou nerovností $n > m$. Pokud by totiž nastal případ $n = m$, získali bychom řešení vyšetřované úlohy přímo ze soustavy (16). Takováto situace je tudíž triviální a není důvod, proč ji dále zkoumat. V dalším tedy bude platit

$$\dim \mathcal{R}(\mathbf{A}) = m, \qquad \dim \mathcal{N}(\mathbf{A}) = n - m > 0, \tag{20}$$

$$\dim \mathcal{R}(\mathbf{A}^T) = m, \qquad \dim \mathcal{N}(\mathbf{A}^T) = 0. \tag{21}$$

Vztahy (21) a lemma 2 dávají ihned následující tvrzení:

Věta 2. Nechť $\mathbf{C}$ je symetrická pozitivně semidefinitní matice a $\mathbf{A}$ je matice plné hodnosti. Matice $\mathbf{K}$ je regulární tehdy a jen tehdy, když $\mathcal{N}(\mathbf{C}) \cap \mathcal{N}(\mathbf{A}) = \{\mathbf{o}\}$.

Důsledek 1. Je-li $\mathbf{C}$ regulární a $\mathbf{A}$ má plnou hodnost, je $\mathbf{K}$ rovněž regulární a soustava (14) má právě jedno řešení.

Věta 3. Soustava rovnic (14) se symetrickou pozitivně semidefinitní maticí $\mathbf{C}$ a maticí $\mathbf{A}$ plné hodnosti má řešení právě tehdy, když platí

$$\mathbf{d} \perp (\mathcal{N}(\mathbf{C}) \cap \mathcal{N}(\mathbf{A})). \tag{22}$$

Přitom toto řešení je jediné, pokud průnik $\mathcal{N}(\mathbf{C}) \cap \mathcal{N}(\mathbf{A})$ obsahuje pouze nulový vektor.

Důkaz Jak je patrné z definice prostoru sloupcových vektorů, soustava (14) má řešení tehdy a jen tehdy, patří-li její pravá strana do prostoru $\mathcal{R}(\mathbf{K})$. To je dle základní věty lineární algebry (viz např. [27]) ekvivalentní s podmínkou její ortogonality vzhledem k prostoru $\mathcal{N}(\mathbf{K})$. Na základě lemmatu 2 a s ohledem na předpoklad (21) pak vychází

$$\mathbf{d}^T\mathbf{p} = 0 \qquad \forall\, \mathbf{p} \in (\mathcal{N}(\mathbf{C}) \cap \mathcal{N}(\mathbf{A})),$$

což je právě podmínka (22). Zbylé tvrzení je pak důsledek věty 2. □

Poznámka 1. V případě, že matice $\mathbf{K}$ je singulární, přičemž $\mathbf{A}$ má plnou hodnost a je splněna podmínka (22), lze všechna řešení soustavy (14) zřejmě zapsat ve tvaru $\{\widetilde{\mathbf{x}} + \mathbf{p},\ \boldsymbol{\lambda}\}$, kde $\widetilde{\mathbf{x}}$ je některá pevně zvolená první složka řešení a $\mathbf{p}$ je libovolný prvek z $\mathcal{N}(\mathbf{C}) \cap \mathcal{N}(\mathbf{A})$.

Důsledek 2. Nemá-li soustava rovnic (14) se symetrickou pozitivně semidefinitní maticí $\mathbf{C}$ a maticí $\mathbf{A}$ plné hodnosti řešení, značí to, že konvexní funkce (15) je zdola neohraničená.

Důkaz Tvrzení plyne ihned z toho, že neplatnost podmínky (22) implikuje $\mathbf{d} \not\perp \mathcal{N}(\mathbf{C})$ a to dle věty 1 z kap. 1 znamená, že (15) je neohraničená zdola. □

Než přikročíme k metodám řešení Kuhn–Tuckerových soustav, ukažme si několik ekvivalentních úprav pro tyto soustavy.

Úprava 1. Obdélníková soustava (16) má pro $m < n$ nekonečně mnoho řešení. Vyberme libovolně (ale pro další úvahy pevně) jedno z nich a označme ho $\widetilde{\mathbf{x}}$. Všechna ostatní řešení lze pak vyjádřit ve tvaru $\widetilde{\mathbf{x}} + \mathbf{p}$, kde $\mathbf{p} \in \mathcal{N}(\mathbf{A})$. Dosadíme tedy v (14) za $\mathbf{x}$ výraz $\widetilde{\mathbf{x}} + \mathbf{p}$ a obdržíme

$$\begin{pmatrix} \mathbf{C} & \mathbf{A}^T \\ \mathbf{A} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{p} \\ \boldsymbol{\lambda} \end{pmatrix} = \begin{pmatrix} \widetilde{\mathbf{d}} \\ \mathbf{o} \end{pmatrix}, \tag{23}$$

kde jsme označili

$$\widetilde{\mathbf{d}} = \mathbf{d} - \mathbf{C}\widetilde{\mathbf{x}}. \tag{24}$$

Pro známé $\widetilde{\mathbf{x}} \in \mathbb{R}^n$ získáme řešením této soustavy i řešení soustavy (14).

Úprava 2. Rovnici (16) vynásobme zleva maticí $\mathbf{A}^T$ takže dostaneme

$$\mathbf{A}^T\mathbf{A}\mathbf{x} = \mathbf{A}^T\mathbf{b}. \tag{25}$$

Tuto rovnici nyní vynásobíme číslem $\rho > 0$ a přičteme k první maticové rovnici soustavy (14)

$$\mathbf{C}\mathbf{x} + \mathbf{A}^T\boldsymbol{\lambda} = \mathbf{d}, \tag{26}$$

čímž získáme soustavu

$$\begin{pmatrix} \mathbf{C} + \rho\mathbf{A}^T\mathbf{A} & \mathbf{A}^T \\ \mathbf{A} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{x} \\ \boldsymbol{\lambda} \end{pmatrix} = \begin{pmatrix} \mathbf{d} + \rho\mathbf{A}^T\mathbf{b} \\ \mathbf{b} \end{pmatrix}. \tag{27}$$

Tato nová soustava má totéž řešení jako soustava původní. Poznamenejme jen, že matice $\mathbf{C} + \rho\mathbf{A}^T\mathbf{A}$ zůstává symetrická.

Úprava 3. Je-li $\mathbf{d} \in \mathcal{R}(\mathbf{C})$, lze postupovat následovně. Z množiny řešení soustavy $\mathbf{C}\mathbf{x} = \mathbf{d}$, jež je za uvedeného předpokladu neprázdná, vyberme jedno a označme $\widehat{\mathbf{x}}$. Nyní dosadíme za $\mathbf{d}$ do (26) a po označení

$$\mathbf{u} = \mathbf{x} - \widehat{\mathbf{x}}$$

dostaneme tuto soustavu rovnic

$$\begin{pmatrix} \mathbf{C} & \mathbf{A}^T \\ \mathbf{A} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{u} \\ \boldsymbol{\lambda} \end{pmatrix} = \begin{pmatrix} \mathbf{o} \\ \widehat{\mathbf{b}} \end{pmatrix}, \tag{28}$$

kde jsme položili

$$\widehat{\mathbf{b}} = \mathbf{b} - \mathbf{A}\,\widehat{\mathbf{x}}. \tag{29}$$

Pro dané $\widehat{\mathbf{x}}$ dokážeme z vypočteného řešení této soustavy ihned určit řešení soustavy (14).

Dále budeme ještě potřebovat několik pomocných tvrzení.

Lemma 3. Pro libovolnou matici $\mathbf{A} \in \mathbb{R}^{m\times n}$ je $\mathbf{A}^T\mathbf{A}$ symetrická pozitivně semidefinitní matice řádu n. Tato matice je regulární a tedy pozitivně definitní, pokud je $m \geq n$ a matice $\mathbf{A}$ má plnou hodnost.
Důkaz Symetrie je zřejmá a pozitivní semidefinitnost plyne z vyjádření:

$$\mathbf{x}^T(\mathbf{A}^T\mathbf{A})\mathbf{x} = (\mathbf{A}\mathbf{x})^T\mathbf{A}\mathbf{x} = \|\mathbf{A}\mathbf{x}\|_2^2 \geq 0 \qquad \forall\, \mathbf{x} \in \mathbb{R}^n,$$

kde $\|\ .\ \|_2$ značí eukleidovskou normu vektoru. Je-li navíc $m \geq n$ a $\mathbf{A}$ má plnou hodnost, musí být $\mathrm{rank}(\mathbf{A}) = n$ a tudíž $\dim \mathcal{N}(\mathbf{A}) = 0$, což dává zbývající tvrzení lemmatu. □

Lemma 4. Nechť matice $\mathbf{A} \in \mathbb{R}^{m\times n}$, $m < n$, má plnou hodnost. Pak

$$\mathbf{P} = \mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{A} \tag{30}$$

je matice ortogonální projekce $\mathbb{R}^n$ na $\mathcal{R}(\mathbf{A}^T)$ a

$$\mathbf{I} - \mathbf{P}$$

je matice ortogonální projekce $\mathbb{R}^n$ na $\mathcal{N}(\mathbf{A})$ a platí

$$\mathbf{P}(\mathbf{I} - \mathbf{P}) = (\mathbf{I} - \mathbf{P})\mathbf{P} = \mathbf{0}. \tag{31}$$

Důkaz viz např. [27].

Lemma 5. Nechť $\mathbf{C} \in \mathbb{R}^{n\times n}$ je symetrická pozitivně semidefinitní matice, $\mathbf{A} \in \mathbb{R}^{m\times n}$. Jestliže je $\mathcal{N}(\mathbf{C}) \cap \mathcal{N}(\mathbf{A}) = \{\mathbf{o}\}$, pak je matice $\mathbf{C} + \rho\mathbf{A}^T\mathbf{A}$ symetrická pozitivně definitní pro libovolné $\rho > 0$.
Důkaz Pro libovolné $\rho > 0$ a $\mathbf{x} \in \mathbb{R}^n$ máme

$$\mathbf{x}^T(\mathbf{C} + \rho\mathbf{A}^T\mathbf{A})\mathbf{x} = \mathbf{x}^T\mathbf{C}\mathbf{x} + \rho\mathbf{x}^T\mathbf{A}^T\mathbf{A}\mathbf{x} = \mathbf{x}^T\mathbf{C}\mathbf{x} + \rho\|\mathbf{A}\mathbf{x}\|_2^2 \geq 0.$$

Vzhledem k předpokladu $\mathcal{N}(\mathbf{C}) \cap \mathcal{N}(\mathbf{A}) = \{\mathbf{o}\}$ však tento výraz může být roven nule jen pro $\mathbf{x} = \mathbf{o}$. □

2 Přehled metod používaných k řešení Kuhn–Tuckerových soustav

2.1 Metoda nulového prostoru

Použijme úpravu 1, čímž dostaneme soustavu

$$\begin{pmatrix} \mathbf{C} & \mathbf{A}^T \\ \mathbf{A} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{p} \\ \boldsymbol{\lambda} \end{pmatrix} = \begin{pmatrix} \widetilde{\mathbf{d}} \\ \mathbf{o} \end{pmatrix}, \tag{32}$$

$$\widetilde{\mathbf{d}} = \mathbf{d} - \mathbf{C}\tilde{\mathbf{x}}. \tag{33}$$

Protože $\mathbf{p} \in \mathcal{N}(\mathbf{A})$, lze pro tento vektor použít následujícího vyjádření. Předpokládejme, že známe vektory $\mathbf{z}_1, \ldots, \mathbf{z}_{n-m} \in \mathbb{R}^n$, které tvoří bázi prostoru $\mathcal{N}(\mathbf{A})$. Sestavme z nich matici $\mathbf{Z} \in \mathbb{R}^{n\times(n-m)}$, která má zřejmě plnou hodnost a pro niž platí

$$\mathbf{AZ} = \mathbf{0}. \tag{34}$$

Vektor $\mathbf{p}$ musí být lineární kombinací členů báze, což vyjádříme pomocí neznámého vektoru koeficientů $\mathbf{x}_Z \in \mathbb{R}^{n-m}$ takto

$$\mathbf{p} = \mathbf{Z}\mathbf{x}_Z. \tag{35}$$

Dosaďme nyní tento výraz do první maticové rovnice uvažované soustavy, kterou při tom současně vynásobíme zleva maticí $\mathbf{Z}^T$. Tím dostaneme

$$\mathbf{Z}^T\mathbf{C}\mathbf{Z}\mathbf{x}_Z + \mathbf{Z}^T\mathbf{A}^T\boldsymbol{\lambda} = \mathbf{Z}^T\widetilde{\mathbf{d}}.$$

Odtud s ohledem na (34) a po dosazení za $\widetilde{\mathbf{d}}$ máme rovnici pro případné určení $\mathbf{x}_Z$

$$\mathbf{Z}^T\mathbf{C}\mathbf{Z}\mathbf{x}_Z = \mathbf{Z}^T(\mathbf{d} - \mathbf{C}\tilde{\mathbf{x}}). \tag{36}$$

Druhou neznámou $\boldsymbol{\lambda}$ pak lze získat z rovnice

$$\mathbf{A}^T\boldsymbol{\lambda} = \mathbf{d} - \mathbf{C}\tilde{\mathbf{x}} - \mathbf{C}\mathbf{Z}\mathbf{x}_Z. \tag{37}$$

Vzhledem k předpokladu plné hodnosti matice $\mathbf{A}$ lze k řešení použít pseudoinverze (viz [26]). Dostaneme

$$\boldsymbol{\lambda} = \mathbf{A}^{+T}(\mathbf{d} - \mathbf{C}\tilde{\mathbf{x}} - \mathbf{C}\mathbf{Z}\mathbf{x}_Z), \tag{38}$$

kde $\mathbf{A}^{+T}$ značí pseudoinverzní matici k matici $\mathbf{A}^T$ a platí

$$\mathbf{A}^{+T} = (\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{A}.$$

Matice $\mathbf{A}\mathbf{A}^T$ je regulární na základě lemmatu 3.

Obdobný obrat se často využívá i k určení vektoru $\widetilde{\mathbf{x}}$. Hledejme tento vektor v ortogonálním doplňku nulového prostoru, tj. v prostoru $\mathcal{R}(\mathbf{A}^T)$. Pomocí neznámého vektoru $\mathbf{x}_A \in \mathbb{R}^m$ lze pak uvažovaný vektor vyjádřit ve tvaru

$$\widetilde{\mathbf{x}} = \mathbf{A}^T\mathbf{x}_A. \tag{39}$$

Po dosazení do rovnice $\mathbf{Ax} = \mathbf{b}$ pak vychází

$$\mathbf{x}_A = (\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{b}. \tag{40}$$

Rovnice (36) se pak uvádí ve tvaru

$$\mathbf{Z}^T\mathbf{C}\mathbf{Z}\mathbf{x}_Z = \mathbf{Z}^T(\mathbf{d} - \mathbf{C}\mathbf{A}^T\mathbf{x}_A) \tag{41}$$

a hledaný vektor $\mathbf{x}$ je pak dán jako

$$\mathbf{x} = \mathbf{A}^T\mathbf{x}_A + \mathbf{Z}\mathbf{x}_Z. \tag{42}$$

Je zřejmé, že klíčovou otázkou je zde řešitelnost soustavy (41) a ta závisí zřejmě na vlastnostech matice $\mathbf{Z}^T\mathbf{C}\mathbf{Z} \in \mathbb{R}^{(n-m)\times(n-m)}$. Na to odpovídá výše zmiňovaná druhá věta o existenci řešení soustav typu (14).

Věta 4. Nechť matice $\mathbf{A}$ má plnou hodnost a nechť $\mathbf{Z} \in \mathbb{R}^{n\times(n-m)}$ je taková, že platí

$$\mathbf{A}\mathbf{Z} = \mathbf{0}, \qquad \operatorname{rank}(\mathbf{A}^T|\mathbf{Z}) = n.$$

Potom

(i) soustava (14) má právě jedno řešení tehdy a jen tehdy, když je matice $\mathbf{Z}^T\mathbf{C}\mathbf{Z}$ pozitivně definitní;

(ii) soustava (14) má nekonečně mnoho řešení tehdy a jen tehdy, když je matice $\mathbf{Z}^T\mathbf{C}\mathbf{Z}$ singulární pozitivně semidefinitní a soustava (41) je konsistentní;

(iii) soustava (14) nemá řešení v zbývajících případech, tj. když je matice $\mathbf{Z}^T\mathbf{C}\mathbf{Z}$ nedefinitní nebo když (41) není konsistentní.

Důkaz je založen na tom, že danou problematiku převedeme do oblasti optimalizační. Jak již víme, zadané soustavě (14) odpovídá úloha minimalizovat funkci (15) za podmínek (16). Soustava (14) pak představuje rovnice určující sedlový bod Lagrangeovy funkce

$$L(\mathbf{x}, \boldsymbol{\lambda}) = \frac{1}{2}\mathbf{x}^T\mathbf{C}\mathbf{x} - \mathbf{d}^T\mathbf{x} + \boldsymbol{\lambda}^T(\mathbf{A}\mathbf{x} - \mathbf{b}) \qquad \mathbf{x} \in \mathbb{R}^n, \boldsymbol{\lambda} \in \mathbb{R}^m. \tag{43}$$

Vektoru $\boldsymbol{\lambda}$ se říká vektor Lagrangeových multiplikátorů.

Výše uvedené úpravy převádějí tuto úlohu transformací (42) na úlohu nepodmíněné minimalizace kvadratické funkce

$$\widetilde{f}(\mathbf{x}_Z) = \frac{1}{2}\mathbf{x}_Z^T\mathbf{Z}^T\mathbf{C}\mathbf{Z}\mathbf{x}_Z - \mathbf{x}_Z^T\mathbf{Z}^T(\mathbf{d} - \mathbf{C}\mathbf{A}^T\mathbf{x}_A) \tag{44}$$

pro neznámou $\mathbf{x}_Z \in \mathbb{R}^{n-m}$. Rovnice pro určení bodu extrému $\nabla \widetilde{f}(\mathbf{x}_Z) = \mathbf{o}$ je pak právě rovnice (41). Vše tedy závisí na povaze funkce (44). Je-li tato funkce ryze konvexní, má rovnice (41) právě jedno řešení. To nastane v případě, že matice $\mathbf{Z}^T\mathbf{C}\mathbf{Z}$ je pozitivně definitní. Podmínka je nutná i postačující.

Je-li funkce (44) pouze konvexní, bude mít rovnice (41) nekonečně mnoho řešení. K tomu dojde tehdy, když bude matice $\mathbf{Z}^T\mathbf{C}\mathbf{Z}$ singulární a pozitivně semidefinitní. Pro řešitelnost v takovém případě ovšem ještě potřebujeme, aby soustava (41) byla konsistentní. Podmínka je opět nutná i postačující.

Ve zbývajících případech nemá funkce (44) minimum v $\mathbb{R}^{n-m}$ a rovnice (41) tudíž nebude mít řešení. □

Má-li zadaná úloha řešení, lze předchozí výsledky shrnout do následujícího algoritmu.

Algoritmus 1 *(Metoda nulového prostoru)*
Krok *0*. Nalézt bázi $\{\mathbf{z}_i\}_{i=1}^{n-m}$ prostoru $\mathcal{N}(\mathbf{A})$ a tyto vektory sestavit do matice $\mathbf{Z} \in \mathbb{R}^{n\times(n-m)}$.
Krok *1*. Postupně sestavit a vyřešit soustavy

$$\mathbf{A}\mathbf{A}^T\mathbf{x}_A = \mathbf{b}, \tag{45}$$

$$\mathbf{Z}^T\mathbf{C}\mathbf{Z}\mathbf{x}_Z = \mathbf{Z}^T(\mathbf{d} - \mathbf{C}\mathbf{A}^T\mathbf{x}_A), \tag{46}$$

$$\mathbf{A}\mathbf{A}^T\boldsymbol{\lambda} = \mathbf{A}(\mathbf{d} - \mathbf{C}\mathbf{A}^T\mathbf{x}_A - \mathbf{C}\mathbf{Z}\mathbf{x}_Z). \tag{47}$$

Krok *2*. Vypočítat

$$\mathbf{x} = \mathbf{A}^T\mathbf{x}_A + \mathbf{Z}\mathbf{x}_Z. \tag{48}$$

Protože metoda v podstatě převádí soustavu (14) řádu $n + m$ na soustavu (41) řádu $n - m$ (pomineme-li ovšem ostatní úkony obsažené v algoritmu), je její použití výhodné v případech, kdy m a n jsou velikostí blízká čísla. Použití pro úlohy velkých rozměru není výhodné, neboť matice soustavy (41) nebývá zpravidla řídká a také určení matice $\mathbf{Z}$ může činit potíže. Konečně, jisté problémy stran přesnosti, resp. numerické stability, může způsobovat i výpočet multiplikátorů $\boldsymbol{\lambda}$ pomocí pseudoinverze (viz [13]).

2.2 Metoda projekce gradientu

Také nyní použijme úpravu 1 a převedeme tak danou úlohu na soustavu

$$\begin{pmatrix} \mathbf{C} & \mathbf{A}^T \\ \mathbf{A} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{p} \\ \boldsymbol{\lambda} \end{pmatrix} = \begin{pmatrix} \widetilde{\mathbf{d}} \\ \mathbf{o} \end{pmatrix}, \tag{49}$$

$$\widetilde{\mathbf{d}} = \mathbf{d} - \mathbf{C}\widetilde{\mathbf{x}}. \tag{50}$$

Za pomoci lemmatu 4 však teď vyjádříme neznámou $\mathbf{p}$ jinak:

$$\mathbf{p} = (\mathbf{I} - \mathbf{P})\mathbf{y}, \tag{51}$$

kde $\mathbf{y} \in \mathbb{R}^n$. S ohledem na vlastnost

$$\mathbf{A}(\mathbf{I}-\mathbf{P}) = \mathbf{0}$$

nám zůstává jen první maticová rovnice

$$\mathbf{C}(\mathbf{I}-\mathbf{P})\mathbf{y} + \mathbf{A}^T\boldsymbol{\lambda} = \widetilde{\mathbf{d}}.$$

Tu nyní vynásobíme maticí $(\mathbf{I}-\mathbf{P})^T$ zleva. Obdržíme

$$(\mathbf{I}-\mathbf{P})^T\mathbf{C}(\mathbf{I}-\mathbf{P})\mathbf{y} + (\mathbf{I}-\mathbf{P})^T\mathbf{A}^T\boldsymbol{\lambda} = (\mathbf{I}-\mathbf{P})^T\widetilde{\mathbf{d}}.$$

Vezmeme-li do úvahy, že platí

$$(\mathbf{I}-\mathbf{P})^T\mathbf{A} = \mathbf{0}$$

a

$$(\mathbf{I}-\mathbf{P})^T = \mathbf{I}-\mathbf{P}$$

dostaneme po dosazení z (50) výslednou soustavu

$$(\mathbf{I}-\mathbf{P})\mathbf{C}(\mathbf{I}-\mathbf{P})\mathbf{y} = (\mathbf{I}-\mathbf{P})(\mathbf{d}-\mathbf{C}\widetilde{\mathbf{x}}). \tag{52}$$

Označíme-li ještě

$$\overline{\mathbf{C}} = (\mathbf{I}-\mathbf{P})\mathbf{C}(\mathbf{I}-\mathbf{P}) \tag{53}$$

a

$$\overline{\mathbf{d}} = (\mathbf{I}-\mathbf{P})(\mathbf{d}-\mathbf{C}\widetilde{\mathbf{x}}), \tag{54}$$

je vidět, že jsme původní problém převedli na úlohu řešit singulární semidefinitní soustavu rovnic řádu n

$$\overline{\mathbf{C}}\mathbf{y} = \overline{\mathbf{d}}. \tag{55}$$

Neznámou $\mathbf{x}$ pak vypočteme ze vztahu

$$\mathbf{x} = \widetilde{\mathbf{x}} + (\mathbf{I}-\mathbf{P})\mathbf{y} \tag{56}$$

a $\boldsymbol{\lambda}$ stejně jako u předchozí metody pomocí pseudoinverze

$$\boldsymbol{\lambda} = \mathbf{A}^{+T}(\mathbf{d}-\mathbf{C}\mathbf{x}). \tag{57}$$

Předchozí výsledky shrneme do následujícího algoritmu:

Algoritmus 2 *(Metoda projekce gradientu)*
Krok *0.* Určit $\widetilde{\mathbf{x}}$ tak, aby $\mathbf{A}\widetilde{\mathbf{x}} = \mathbf{b}$.
Sestavit matici projekce na $\mathcal{N}(\mathbf{A})$

$$\mathbf{I}-\mathbf{P} = \mathbf{I}-\mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{A}. \tag{58}$$

Krok *1.* Sestavit a vyřešit soustavu

$$\overline{\mathbf{C}}\mathbf{y} = \overline{\mathbf{d}}, \tag{59}$$

kde matice $\overline{\mathbf{C}}$ a vektor $\overline{\mathbf{d}}$ jsou dány vztahy (53) a (54).
Krok *2.* Vypočítat

$$\mathbf{x} = \tilde{\mathbf{x}} + (\mathbf{I} - \mathbf{P})\mathbf{y}, \tag{60}$$

$$\boldsymbol{\lambda} = (\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{A}(\mathbf{d} - \mathbf{C}\mathbf{x}). \tag{61}$$

Pro tuto metodu platí v zásadě totéž, co pro metodu nulového prostoru. Tedy, že její použití je výhodné v případech, kdy mezi čísly n a m nejsou příliš velké rozdíly a že výpočet multiplikátorů $\boldsymbol{\lambda}$ pomocí vztahu (61) nemusí být vždy kvalitní.

Nutnost počítat projektivní matice $\mathbf{P}$ a řešit singulární soustavy rovnic (59) způsobila, že metoda, jejíž ideu publikoval J.B. Rosen v r.1960, si nezískala oblibu. Pro úlohu kvadratického programování byla v r.1969 použita B.T. Poljakem a později v obměněné verzi publikována v [25]. V obou případech ve spojení s metodou konjugovaných gradientů.

Z naposled citovaného pramene byla metoda převzata pro řešení kontaktní úlohy v článku [16]. Zde bylo důmyslně využito toho, že v této úloze má matice $\mathbf{A}$ velmi speciální strukturu, která umožňuje vcelku snadno určit projekci $\mathbf{P}$. Protože v kontaktních úlohách je typicky $n \gg m$, byl celý postup ještě doplněn o tzv. předeliminaci. Ta byla založena na další specifické vlastnosti kontaktních úloh, v nichž se v omezujících podmínkách vyskytuje jen relativně malý počet neznámých. Soustava (14) se proto přeuspořádala tak, že v jejím blokovém zápisu

$$\begin{pmatrix} \mathbf{C}_{11} & \mathbf{C}_{21}^T & \mathbf{0}^T \\ \mathbf{C}_{21} & \mathbf{C}_{22} & \mathbf{A}^T \\ \mathbf{0} & \mathbf{A} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \boldsymbol{\lambda} \end{pmatrix} = \begin{pmatrix} \mathbf{d}_1 \\ \mathbf{d}_2 \\ \mathbf{b} \end{pmatrix} \tag{62}$$

vystupovaly v omezeních pouze neznámé $\mathbf{x}_2$. Z celé matice $\mathbf{C}$ pak po předeliminaci neznámých $\mathbf{x}_1$ zůstal jenom Schurův komplement

$$\widehat{\mathbf{C}} = \mathbf{C}_{22} - \mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{C}_{21}^T \tag{63}$$

a úloha tak dostala tvar

$$\begin{pmatrix} \widehat{\mathbf{C}} & \mathbf{A}^T \\ \mathbf{A} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{x}_2 \\ \boldsymbol{\lambda} \end{pmatrix} = \begin{pmatrix} \widehat{\mathbf{d}} \\ \mathbf{o} \end{pmatrix}, \tag{64}$$

kde

$$\widehat{\mathbf{d}} = \mathbf{d}_2 - \mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{d}_1. \tag{65}$$

Po jejím vyřešení metodou projekce gradientu se zbylé neznámé dopočítaly ze soustavy rovnic

$$\mathbf{C}_{11}\mathbf{x}_1 = \mathbf{d}_1 - \mathbf{C}_{21}^T\mathbf{x}_2. \tag{66}$$

Autor této práce pak později celý postup přepracoval za pomoci tehdejších spolupracovníků z oddělení pružnosti a pevnosti Výzkumného ústavu SIGMA (zejména ing. J. Horáka a ing. J. Petřeka) do počítačové realizace (viz [23]). Některé výsledky lze pak nalézt např. v [18]. Pro realizaci použité metody konjugovaných gradientů platí ovšem výhrady a připomínky uvedené v [24]

Řešení kontaktní problematiky pomocí metody projekce gradientu se v posledních zhruba deseti letech věnoval Z. Dostál. Publikoval zajímavé výsledky v oblasti předpodmiňování metody konjugovaných gradientů (např. [5]) a v souvislosti s metodou rozkladu oblasti (*domain decomposition*).

2.3 Metoda rozšířených lagrangiánů

Tato metoda byla v r. 1969 publikována nezávisle na sobě M. Hestenesem a M. J. D. Powellem a její pole působnosti se neomezuje pouze na námi zkoumanou problematiku. Z pohledu optimalizačního odpovídá záměně Lagrangeovy funkce (43) daného problému tzv. rozšířenou Lagrangeovou funkcí pro nějaké $\rho > 0$

$$L_\rho(\mathbf{x}, \boldsymbol{\lambda}) = \frac{1}{2}\mathbf{x}^T\mathbf{C}\mathbf{x} - \mathbf{d}^T\mathbf{x} + \boldsymbol{\lambda}^T(\mathbf{A}\mathbf{x} - \mathbf{b}) + \frac{\rho}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 \qquad \mathbf{x} \in \mathbb{R}^n, \boldsymbol{\lambda} \in \mathbb{R}^m \quad (67)$$

a následným určením jejího sedlového bodu.

Algebraicky příslušnou soustavu rovnic získáme po úpravě 2

$$\begin{pmatrix} \mathbf{C} + \rho\mathbf{A}^T\mathbf{A} & \mathbf{A}^T \\ \mathbf{A} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{x} \\ \boldsymbol{\lambda} \end{pmatrix} = \begin{pmatrix} \mathbf{d} + \rho\mathbf{A}^T\mathbf{b} \\ \mathbf{b} \end{pmatrix} \qquad \rho > 0. \quad (68)$$

Na základě lemmatu 5 víme, že, má-li naše úloha (14) právě jedno řešení, je matice $\mathbf{C} + \rho\mathbf{A}^T\mathbf{A}$ pozitivně definitní. Toho pak lze s výhodou požít ve spojení s některými iteračními metodami, zejména gradientního typu. U velkých řídkých úloh je užití metody poněkud omezené, neboť uvedená matice může být značně zaplněná. Problematika rozšířených lagrangiánů je včetně řady aplikací podrobně zpracována v pracích R. Glowinského (např. [14]). Semidefinitní případy zde však chybí.

V poslední době dosáhl v této oblasti pozoruhodných výsledků Z. Dostál (viz [6] nebo [7]) a v jeho publikacích jsou již semidefinitní případy uvažovány.

2.4 Metoda blokové eliminace

V literatuře je tato metoda uváděna pod názvem *range space method* a předpokládá, že matice $\mathbf{C}$ je regulární. V soustavě (14) se provede standardní bloková eliminace neznámé $\mathbf{x}$ pomocí vynásobení první maticové rovnice součinem $-\mathbf{A}\mathbf{C}^{-1}$ a přičtením výsledku k druhé maticové rovnici. Obdržíme maticovou rovnici, odkud lze vypočítat vektor $\boldsymbol{\lambda}$. Následně pak dokážeme určit vektor $\mathbf{x}$, jak ukazuje

Algoritmus 3 *(Metoda blokové eliminace)*
Postupně sestavit a vyřešit soustavy rovnic

$$\mathbf{A}\mathbf{C}^{-1}\mathbf{A}^T\boldsymbol{\lambda} = \mathbf{A}\mathbf{C}^{-1}\mathbf{d} - \mathbf{b}, \tag{69}$$

$$\mathbf{C}\mathbf{x} = \mathbf{d} - \mathbf{A}^T\boldsymbol{\lambda}. \tag{70}$$

Poznamenejme, že obě soustavy jsou symetrické a k jejich řešení lze tedy použít obecně např. Bunch–Parlettovu metodu, v případě (který nás právě zajímá) pozitivní definitnosti matice $\mathbf{C}$, pak Choleského metodu. Matice $\mathbf{C}^{-1}$ se pochopitelně explicitně nevytváří, ale použije se již zmíněné Bunch–Parlettovy, resp. Choleského faktorizace.

Tento postup řešení je vhodné použít tehdy, kdy předchozí dva nejsou efektivní, tedy když $n \gg m$. To ukazuje na možnost řešit takto úlohy velkého rozměru. Je ale potřeba si všimnout, že matice soustavy (69) je obecně zaplněná, což limituje efektivnost této metody na problémy, kde m není příliš veliké (maximálně několik stovek). Pro kontaktní úlohy je to však zcela vyhovující — až na to, že $\mathbf{C}$ musí být pozitivně definitní. Výhodou metody je zjevně to, že není třeba počítat matice $\mathbf{Z}$, resp. $\mathbf{P}$.

Zobecnění metody pro případ singulární matice lze provést následovně. Matici $\mathbf{C}$ je nejprve třeba uspořádat tak, aby podmatice $\mathbf{C}_{11} \in \mathbb{R}^{r\times r}$, $r = \mathrm{rank}(\mathbf{C})$, byla regulární. To pak indukuje uspořádání i zbylých komponent a dostáváme soustavu ve tvaru

$$\begin{pmatrix} \mathbf{C}_{11} & \mathbf{C}_{21}^T & \mathbf{A}_1^T \\ \mathbf{C}_{21} & \mathbf{C}_{22} & \mathbf{A}_2^T \\ \mathbf{A}_1 & \mathbf{A}_2 & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \boldsymbol{\lambda} \end{pmatrix} = \begin{pmatrix} \mathbf{d}_1 \\ \mathbf{d}_2 \\ \mathbf{b} \end{pmatrix}, \tag{71}$$

kde $\mathbf{A}_1 \in \mathbb{R}^{m\times r},\ \mathbf{d}_1 \in \mathbb{R}^r$. Definujme dále

$$\mathbf{E} = \begin{pmatrix} \mathbf{C}_{21} \\ \mathbf{A}_1 \end{pmatrix} \qquad \mathbf{F} = \begin{pmatrix} \mathbf{C}_{22} & \mathbf{A}_2^T \\ \mathbf{A}_2 & \mathbf{0} \end{pmatrix} \qquad \mathbf{h} = \begin{pmatrix} \mathbf{d}_2 \\ \mathbf{b} \end{pmatrix} \tag{72}$$

$$\mathbf{G} = \mathbf{E}\mathbf{C}_{11}^{-1}\mathbf{E}^T - \mathbf{F}. \tag{73}$$

Tím převedeme soustavu (71) na příznivý tvar s levou horní regulární podmaticí

$$\begin{pmatrix} \mathbf{C}_{11} & \mathbf{E}^T \\ \mathbf{E} & \mathbf{F} \end{pmatrix} \begin{pmatrix} \mathbf{x} \\ \mathbf{z} \end{pmatrix} \begin{pmatrix} \mathbf{d}_1 \\ \mathbf{h} \end{pmatrix}. \tag{74}$$

Platí následující tvrzení (viz [15])

Věta 5. Má-li matice $\mathbf{A}$ plnou hodnost, pak

(i) úloha má právě jedno řešení tehdy a jen tehdy, když součet počtu kladných vlastních čísel matice $\mathbf{C}_{11}$ a počtu záporných vlastních čísel matice $\mathbf{G}$ je roven n.

(ii) úloha má nekonečně mnoho řešení tehdy a jen tehdy, když matice $\mathbf{G}$ je singulární, součet počtu kladných vlastních čísel matice $\mathbf{C}_{11}$ a počtu nekladných vlastních čísel matice $\mathbf{G}$ je roven n a je-li soustava

$$\mathbf{G}\mathbf{z} = \mathbf{E}\mathbf{C}_{11}^{-1}\mathbf{d}_1 - \mathbf{h}$$

konsistentní.

(iii) úloha nemá řešení v ostatních případech.

Řešení pak lze určit pomocí následujícího algoritmu.

Algoritmus 4 *(Metoda blokové eliminace pro singulární matici* $\mathbf{C}$*)*
Krok *0*. Určit v $\mathbf{C}$ regulární podmatici $\mathbf{C}_{11}$ řádu $r = \mathrm{rank}(\mathbf{C})$ a přeuspořádat soustavu (14) na tvar (71).
Krok *1*. Postupně sestavit a vyřešit soustavy

$$\mathbf{G}\mathbf{z} = \mathbf{E}\mathbf{C}_{11}^{-1}\mathbf{d}_1 - \mathbf{h}, \tag{75}$$

$$\mathbf{C}_{11}\mathbf{x}_1 = \mathbf{d}_1 - \mathbf{E}^T\mathbf{z}, \tag{76}$$

kde matice $\mathbf{E}$, $\mathbf{F}$, $\mathbf{G}$ a vektor $\mathbf{h}$ jsou dány dle (72) a (73).
Krok *2*. Rozdělit vektor $\mathbf{z}$ na komponenty takto

$$\mathbf{z} = \begin{pmatrix} \mathbf{x}_2 \\ \boldsymbol{\lambda} \end{pmatrix} \qquad \begin{matrix} \mathbf{x}_2 \in \mathbb{R}^{n-r} \\ \boldsymbol{\lambda} \in \mathbb{R}^m \end{matrix}$$

a sestavit pak vektory řešení

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix}, \qquad \boldsymbol{\lambda}.$$

Nakonec se vrátit k původnímu uspořádání soustavy (14).

Je zřejmé, že tento postup je poměrně těžkopádný a v praxi se proto, pokud je autorovi této práce známo, nepoužívá.

2.5 Metody nevyužívající strukturu soustavy

V případech, kdy není výhodné či možné aplikovat některou z výše uvedených metod, lze rezignovat na využití speciální struktury matice soustavy (14) a použít některou z metod pro symetrické nedefinitní soustavy. Tento přístup užívají zejména specialisté na řešení velkých řídkých soustav rovnic. Lze pak vybírat mezi finitními algoritmy, které představují různé varianty Bunch–Parlettovy metody (viz např. [3], [2], [9], [8]), a mezi iteračními, kde se používají zejména metody typu konjugovaných gradientů, např. metoda GMRES (viz např. [19]) nebo QMR (viz [12]).

I když má takovýto přístup nespornou výhodu univerzálnosti, nebudeme se jím zde v dalším zabývat, neboť v této práci se pokusíme strukturu dané úlohy využít.

3 Zobecnění metody blokové eliminace

V dalších úvahách se opět soustředíme na případ symetrické pozitivně semidefinitní matice $\mathbf{C}$. Motivaci, jak postupovat dále, lze nyní nalézt v procesu regularizace singulární matice, jenž byl popsán v [24] a jenž využívá modifikovanou

Choleského metodu. Nahraďme tedy pomocí tohoto postupu matici $\mathbf{C}$ regulární maticí

$$\overline{\mathbf{C}} = \mathbf{C} + \mathbf{E}, \tag{77}$$

kde $\mathbf{E}$ je vhodná diagonální pozitivně semidefinitní matice, a řešme namísto (14) soustavu s regulární podmaticí

$$\begin{pmatrix} \overline{\mathbf{C}} & \mathbf{A}^T \\ \mathbf{A} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \overline{\mathbf{x}} \\ \overline{\boldsymbol{\lambda}} \end{pmatrix} = \begin{pmatrix} \mathbf{d} \\ \mathbf{b} \end{pmatrix}. \tag{78}$$

Toto řešení získáme pomocí algoritmu blokové eliminace. Položme dále

$$\overline{\mathbf{x}} = \mathbf{x} + \delta\mathbf{x},$$

$$\overline{\boldsymbol{\lambda}} = \boldsymbol{\lambda} + \delta\boldsymbol{\lambda}.$$

Dosazením do (78) a porovnáním obou soustav (14) a (78) dostaneme

$$\begin{pmatrix} \mathbf{C} & \mathbf{A}^T \\ \mathbf{A} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \delta\mathbf{x} \\ \delta\boldsymbol{\lambda} \end{pmatrix} = \begin{pmatrix} \mathbf{E}\overline{\mathbf{x}} \\ \mathbf{o} \end{pmatrix}. \tag{79}$$

O řešitelnosti této soustavy vypovídá následující tvrzení.

Věta 6. Nechť $\mathbf{C} \in \mathbb{R}^{n\times n}$ je symetrická pozitivně semidefinitní matice, nechť $\mathbf{E} \in \mathbb{R}^{n\times n}$ je symetrická matice taková, že matice $\mathbf{C} + \mathbf{E}$ je regulární a nechť matice $\mathbf{A} \in \mathbb{R}^{m\times n}$ má plnou hodnost. Potom platí, že má-li řešení soustava (14), má řešení i soustava (79), v níž $\overline{\mathbf{x}}$ představuje první složku řešení soustavy (78).
Důkaz Podle věty 3 je soustava (79) řešitelná, pokud

$$\mathbf{E}\overline{\mathbf{x}} \perp (\mathcal{N}(\mathbf{C}) \cap \mathcal{N}(\mathbf{A})). \tag{80}$$

Vezměme tedy libovolný vektor $\mathbf{p} \in (\mathcal{N}(\mathbf{C}) \cap \mathcal{N}(\mathbf{A}))$. Potom s ohledem na (78) platí

$$(\mathbf{E}\overline{\mathbf{x}})^T\mathbf{p} = \overline{\mathbf{x}}^T\mathbf{E}\mathbf{p} = \overline{\mathbf{x}}^T(\overline{\mathbf{C}} - \mathbf{C})\mathbf{p} = \overline{\mathbf{x}}^T\overline{\mathbf{C}}\mathbf{p} = (\overline{\mathbf{C}}\overline{\mathbf{x}})^T\mathbf{p} =$$

$$= (\mathbf{d} - \mathbf{A}^T\overline{\boldsymbol{\lambda}})^T\mathbf{p} = \mathbf{d}^T\mathbf{p} - \overline{\boldsymbol{\lambda}}^T\mathbf{A}\mathbf{p} = 0,$$

což je právě podmínka (80). □

Soustavu (79) nyní můžeme opět regularizovat a vytvořit tak proces zpřesňování řešení obdobně, jako tomu bylo v [24]. Ten je zapotřebí realizovat v dvojnásobné přesnosti. Rezidua v exaktní aritmetice jsou dána pravou stranou soustavy (79). Pro výpočet však může být vhodnější (v závislosti na požadované přesnosti výsledků) explicitní vyčíslení reziduového vektoru

$$\begin{pmatrix} \mathbf{d} - \mathbf{C}\overline{\mathbf{x}} - \mathbf{A}^T\overline{\boldsymbol{\lambda}} \\ \mathbf{b} - \mathbf{A}\overline{\mathbf{x}} \end{pmatrix}.$$

Nalezení řešení soustavy (14) trvá zpravidla jen několik málo kroků.

Poznámka 2. Regularizaci je účelné aplikovat pouze na podmatici $\mathbf{C}$, jinak dojde k změnám na celé hlavní diagonále matice $\mathbf{K}$ a připravíme se tak o výhodu využití bloku $m \times m$ nul.

Jiná avšak v podstatě ekvivalentní možnost spočívá ve vytvoření vhodného iteračního procesu. Za tímto účelem definujme rozštěpení matice $\mathbf{K}$ takto

$$\mathbf{K} = \mathbf{M} - \mathbf{N}, \tag{81}$$

kde

$$\mathbf{M} = \begin{pmatrix} \overline{\mathbf{C}} & \mathbf{A}^T \\ \mathbf{A} & \mathbf{0} \end{pmatrix}$$

a

$$\mathbf{N} = \begin{pmatrix} \mathbf{E} & \mathbf{0}^T \\ \mathbf{0} & \mathbf{0} \end{pmatrix}.$$

Z (14) pak odvodíme následující iterační proces

$$\mathbf{M} \begin{pmatrix} \mathbf{x}^{k+1} \\ \boldsymbol{\lambda}^{k+1} \end{pmatrix} = \mathbf{N} \begin{pmatrix} \mathbf{x}^{k} \\ \boldsymbol{\lambda}^{k} \end{pmatrix} + \begin{pmatrix} \mathbf{d} \\ \mathbf{b} \end{pmatrix}. \tag{82}$$

Pro takto definovaný proces lze snadno dokázat konvergenci v případě, kdy soustava (14) má jediné řešení.

Věta 7. Nechť $\mathbf{C} \in \mathbb{R}^{n\times n}$ je symetrická pozitivně semidefinitní matice, nechť $\overline{\mathbf{C}} \in \mathbb{R}^{n\times n}$ je symetrická pozitivně definitní matice definovaná vztahem (77), kde $\mathbf{E} \in \mathbb{R}^{n\times n}$ je vhodná symetrická pozitivně semidefinitní matice, a nechť matice $\mathbf{A} \in \mathbb{R}^{m\times n}$ má plnou hodnost. Je-li splněna podmínka

$$\mathcal{N}(\mathbf{C}) \cap \mathcal{N}(\mathbf{A}) = \{\mathbf{o}\}, \tag{83}$$

iterační proces (82) konverguje k řešení úlohy (14) pro libovolný počáteční odhad $\begin{pmatrix} \mathbf{x}^0 \\ \boldsymbol{\lambda}^0 \end{pmatrix}$.

Důkaz Nutnou a postačující podmínkou pro konvergenci procesu (82) je, aby spektrální poloměr matice

$$\mathbf{B} = \mathbf{M}^{-1}\mathbf{N} = \begin{pmatrix} \overline{\mathbf{C}} & \mathbf{A}^T \\ \mathbf{A} & \mathbf{0} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{E} & \mathbf{0}^T \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \tag{84}$$

byl menší než 1 (viz např. [4]). Jen poznamenejme, že matice $\mathbf{M}^{-1}$ existuje na základě důsledku 1. Uvažujme tedy zobecněný problém vlastních čísel tvaru

$$\begin{pmatrix} \mathbf{E} & \mathbf{0}^T \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \end{pmatrix} = \mu \begin{pmatrix} \overline{\mathbf{C}} & \mathbf{A}^T \\ \mathbf{A} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \end{pmatrix} \tag{85}$$

pro libovolně zvolené $\mu \neq 0$. Alespoň jedno takové číslo existuje, jinak by matice (84) byla nulová.

Druhá maticová rovnice v (85) zní

$$\mu \mathbf{A}\mathbf{u}_1 = \mathbf{o}, \tag{86}$$

odkud je zjevné, že $\mathbf{u}_1 \in \mathcal{N}(\mathbf{A})$. První maticová rovnice má tvar

$$\mathbf{E}\mathbf{u}_1 = \mu\overline{\mathbf{C}}\mathbf{u}_1 + \mu\mathbf{A}^T\mathbf{u}_2,$$

což po skalárním vynásobení (obecně komplexním) vektorem $\mathbf{u}_1$ s ohledem na (85) dává

$$\mathbf{u}_1^*\mathbf{E}\mathbf{u}_1 = \mu\mathbf{u}_1^*\overline{\mathbf{C}}\mathbf{u}_1,$$

když $\mathbf{u}_1^*$ značí transponovaný komplexně sdružený vektor k vektoru $\mathbf{u}_1$. Odtud je zřejmé, že $\mu > 0$ a rovněž $\mathbf{u}_1^*\mathbf{E}\mathbf{u}_1 > 0$. Nyní dosadíme z (77) a máme

$$(1-\mu)\mathbf{u}_1^*\mathbf{E}\mathbf{u}_1 = \mu\mathbf{u}_1^*\mathbf{C}\mathbf{u}_1. \tag{87}$$

Poněvadž $\mathbf{u}_1 \in \mathcal{N}(\mathbf{A})$, je kvůli podmínce (83) pravá strana (87) kladná. Odtud dostáváme, že $(1-\mu) > 0$. Celkem tedy je $0 \le \mu < 1$. □

Pokud vynecháme podmínku (83) garantující jednoznačnost řešení, dostaneme méně uspokojivý odhad $0 \le \mu \le 1$. V tom případě musíme k důkazu konvergence použít postup uvedený v [24].

Věta 8. Je-li soustava (1) konsistentní, pak iterační proces (82), kde $\mathbf{C} \in \mathbb{R}^{n\times n}$ je symetrická pozitivně semidefinitní matice, $\mathbf{E} \in \mathbb{R}^{n\times n}$ je symetrická pozitivně semidefinitní matice, taková, že $\overline{\mathbf{C}} = \mathbf{C} + \mathbf{E}$ je matice symetrická pozitivně definitní, a matice $\mathbf{A} \in \mathbb{R}^{m\times n}$, $m < n$, má plnou hodnost, konverguje pro libovolný počáteční vektor $\begin{pmatrix}\mathbf{x}^0\\ \boldsymbol{\lambda}^0\end{pmatrix}$ k řešení soustavy (14). Při volbě $\begin{pmatrix}\mathbf{x}^0\\ \boldsymbol{\lambda}^0\end{pmatrix} = \begin{pmatrix}\overline{\mathbf{x}}\\ \overline{\boldsymbol{\lambda}}\end{pmatrix}$ tento proces pak konverguje k řešení s minimální normou.

Důkaz Vztah (82) si nejprve upravme na tvar

$$\mathbf{u}^{k+1} = \mathbf{B}\mathbf{u}^k + \mathbf{f}, \tag{88}$$

kde $\mathbf{B}$ je matice z (84), $\mathbf{u} = \begin{pmatrix}\mathbf{x}\\ \boldsymbol{\lambda}\end{pmatrix}$ a $\mathbf{f} = \mathbf{M}^{-1}\begin{pmatrix}\mathbf{d}\\ \mathbf{b}\end{pmatrix}$. Odtud postupným dosazováním snadno obdržíme vztah

$$\mathbf{u}^{k+1} = \mathbf{B}^{k+1}\mathbf{u}^0 + \sum_{i=0}^{k}\mathbf{B}^i\mathbf{f}. \tag{89}$$

Na základě úvah provedených v důkazu věty 7 je snadné nahlédnout, že vlastní čísla matice $\mathbf{B}$ jsou reálná a platí

$$0 \le \lambda(\mathbf{B}) \le 1, \tag{90}$$

přičemž obou krajních hodnot se alespoň jednou nabyde. Potřebujeme tudíž ukázat, že existuje

$$\lim_{k\to+\infty} \mathbf{B}^k$$

(i když není $\mathbf{0}$). Matice $\mathbf{B}$ se pak nazývá *semikonvergentní* (viz [22]).

Vezměme proto do úvahy to, že matice $\mathbf{B}$ musí být podobná matici v Jordanově normálním tvaru ([10]), což lze s ohledem na předchozí úvahy zapsat takto

$$\mathbf{B} = \mathbf{P}\begin{pmatrix}\mathbf{I} & \mathbf{0}\\ \mathbf{0} & \mathbf{D}\end{pmatrix}\mathbf{P}^{-1}, \tag{91}$$

kde $\mathbf{P}$ je regulární matice řádu $n+m$, podmatice $\mathbf{I}$ má s vzhledem k (87) řád rovný $\dim\mathcal{N}(\mathbf{C})$ a podmatice $\mathbf{D}$ obsahuje zbývající vlastní čísla matice $\mathbf{B}$ a je tedy $\rho(\mathbf{D}) < 1$. Odtud máme

$$\mathbf{B}^k = \mathbf{P}\begin{pmatrix}\mathbf{I} & \mathbf{0}\\ \mathbf{0} & \mathbf{D}^k\end{pmatrix}\mathbf{P}^{-1}, \tag{92}$$

takže

$$\lim_{k\to+\infty} \mathbf{B}^k = \mathbf{P}\begin{pmatrix}\mathbf{I} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\mathbf{P}^{-1}. \tag{93}$$

Nyní lze ukázat, že iterace $\{\mathbf{u}^k\}$ konvergují. Uvažujme výraz (89) a upravme jeho pravou stranu. Dle (14) a (84) vychází

$$\mathbf{f} = \mathbf{M}^{-1}\begin{pmatrix}\mathbf{d}\\ \mathbf{b}\end{pmatrix} = \mathbf{M}^{-1}\mathbf{K}\mathbf{u} = (\mathbf{I}-\mathbf{B})\mathbf{u}, \tag{94}$$

neboť

$$\mathbf{I}-\mathbf{B} = \mathbf{I}-\mathbf{M}^{-1}\mathbf{N} = \mathbf{M}^{-1}(\mathbf{M}-\mathbf{N}) = \mathbf{M}^{-1}\mathbf{K}.$$

Odtud obdržíme vyjádření

$$\sum_{i=0}^{k}\mathbf{B}^i\mathbf{f} = \sum_{i=0}^{k}\mathbf{B}^i(\mathbf{I}-\mathbf{B})\mathbf{u} = (\mathbf{I}-\mathbf{B}^{k+1})\mathbf{u} \tag{95}$$

a celkem tedy máme

$$\mathbf{u}^{k+1} = \mathbf{B}^{k+1}\mathbf{u}^0 + (\mathbf{I}-\mathbf{B}^{k+1})\mathbf{u}. \tag{96}$$

S ohledem na (93) nyní dostáváme ihned konvergenci posloupnosti $\{\mathbf{u}^k\}$ pro $k\to+\infty$.

Označme $\mathbf{G} = \mathbf{I}-\mathbf{B}$ a připomeňme např. podle [1], že čtvercovou matici $\mathbf{X}$ řádu $n+m$ nazveme *Drazinovou pseudoinverzí* matice $\mathbf{G}$ právě tehdy, když splňuje následující vlastnosti

$$\mathbf{X}\mathbf{G}\mathbf{X} = \mathbf{X}, \tag{97}$$

$$\mathbf{G}^l\mathbf{X}\mathbf{G} = \mathbf{G}^l, \tag{98}$$

$$\mathbf{G}\mathbf{X} = \mathbf{X}\mathbf{G}, \tag{99}$$

kde l je tzv. *index* matice $\mathbf{G}$, jenž je definován jako nejmenší celé kladné číslo takové, že

$$\operatorname{rank}(\mathbf{G}^l) = \operatorname{rank}(\mathbf{G}^{l+1}).$$

Tato matice existuje, je určena jednoznačně a značíme ji $\mathbf{G}^D$.

Nyní ukážeme, že matice $\mathbf{G}$ má index roven 1. Protože podle věty o hodnosti součinu matic (viz např. [27]) je

$$\operatorname{rank}(\mathbf{G}) \geq \operatorname{rank}(\mathbf{G}^2),$$

předpokládejme, že zde platí ostrá nerovnost. Potom ovšem existuje nenulový vektor $\mathbf{w}$ takový, že $\mathbf{G}^2\mathbf{w} = \mathbf{o}$ a $\mathbf{G}\mathbf{w} \neq \mathbf{o}$. Tím jsme obdrželi nenulový vektor $\mathbf{v} = \mathbf{G}\mathbf{w}$, jenž leží v $\mathcal{N}(\mathbf{G})$. Ježto dle (81) a (84)

$$\mathbf{G} = \mathbf{I} - \mathbf{B} = \mathbf{I} - \mathbf{M}^{-1}\mathbf{N} = \mathbf{M}^{-1}(\mathbf{M} - \mathbf{N}) = \mathbf{M}^{-1}\mathbf{K}, \tag{100}$$

znamená tato skutečnost, že

$$\mathbf{M}^{-1}\mathbf{K}\mathbf{v} = \mathbf{o}$$

a tudíž $\mathbf{v} \in \mathcal{N}(\mathbf{K})$. Zároveň z definice vektoru $\mathbf{v}$ plyne, že

$$\mathbf{v} = \mathbf{M}^{-1}\mathbf{K}\mathbf{w}$$

a odtud

$$\mathbf{M}\mathbf{v} = \mathbf{K}\mathbf{w}.$$

Jestliže tento vztah vynásobíme skalárně zleva vektorem $\mathbf{v}$, obdržíme s ohledem na předchozí úvahy

$$\mathbf{v}^T\mathbf{M}\mathbf{v} = 0.$$

To je ale, s ohledem na to, že $\mathbf{v} \neq \mathbf{o}$, ve sporu s tím, že matice $\mathbf{M}$ je regulární. Čímž je tvrzení o indexu matice $\mathbf{G}$ dokázáno.

Abychom nyní ukázali, že limitou posloupnosti $\{\mathbf{u}^k\}$ pro $k \to +\infty$ je řešení soustavy (14), upravme dále výraz (93). Předně je dle (91)

$$\mathbf{G} = \mathbf{P}\begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} - \mathbf{D} \end{pmatrix}\mathbf{P}^{-1} \tag{101}$$

a odtud dostaneme ihned výraz pro Drazinovu pseudoinverzní matici k matici $\mathbf{G}$ ve tvaru (viz [1, kap. 4])

$$\mathbf{G}^D = \mathbf{P}\begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & (\mathbf{I} - \mathbf{D})^{-1} \end{pmatrix}\mathbf{P}^{-1}. \tag{102}$$

Srovnáním vztahů (93), (101) a (102) pak dojdeme k vyjádření

$$\lim_{k\to+\infty} \mathbf{B}^k = \mathbf{I} - \mathbf{G}^D\mathbf{G}. \tag{103}$$

Označíme-li

$$\lim_{k\to+\infty} \mathbf{u}^k = \tilde{\mathbf{u}},$$

dává limitní přechod pro $k \to +\infty$ v (96) s ohledem na vztah (103)

$$\tilde{\mathbf{u}} = (\mathbf{I} - \mathbf{G}^D\mathbf{G})\mathbf{u}^0 + \mathbf{G}^D\mathbf{G}\mathbf{u}. \tag{104}$$

Nyní vynásobením maticí $\mathbf{G}$ zleva dostáváme s ohledem na (98), přičemž $l = 1$,

$$\mathbf{G}\tilde{\mathbf{u}} = \mathbf{G}\mathbf{u},$$

což po dosazení z (100) a (94) dává

$$\mathbf{M}^{-1}\mathbf{K}\tilde{\mathbf{u}} = \mathbf{f}$$

a $\tilde{\mathbf{u}}$ je tedy řešením soustavy (14).

Nyní již jen zbývá ukázat, že při volbě $\mathbf{u}^0 = \overline{\mathbf{u}} = \begin{pmatrix} \overline{\mathbf{x}} \\ \overline{\boldsymbol{\lambda}} \end{pmatrix}$ konvergují iterace k řešení s minimální normou. Z výrazu (104) je ihned patrné, že pro první člen na pravé straně platí dle (98)

$$(\mathbf{I} - \mathbf{G}^D\mathbf{G})\mathbf{u}^0 \in \mathcal{N}(\mathbf{G}). \tag{105}$$

Pokud jde o druhý člen, vynásobme ho skalárně libovolně zvoleným vektorem $\mathbf{z} \in \mathcal{N}(\mathbf{G}^T)$. S ohledem na (99) je

$$\mathbf{z}^T\mathbf{G}^D\mathbf{G}\mathbf{u} = \mathbf{z}^T\mathbf{G}\mathbf{G}^D\mathbf{u} = (\mathbf{G}^T\mathbf{z})^T\mathbf{G}^D\mathbf{u} = 0.$$

Odtud plyne, že pro tento člen (upravený dle (94)) máme

$$\mathbf{G}^D\mathbf{f} \in \mathcal{R}(\mathbf{G}). \tag{106}$$

Bude-li $\mathbf{u}^0 \in \mathcal{R}(\mathbf{G})$, pak složka vektoru $\tilde{\mathbf{u}}$ ležící v $\mathcal{N}(\mathbf{G})$ bude nulová, neboť počáteční aproximaci lze vyjádřit jako $\mathbf{u}^0 = \mathbf{G}\mathbf{w}$ a dle (105), (99) a (98) je

$$(\mathbf{I} - \mathbf{G}^D\mathbf{G})\mathbf{u}^0 = (\mathbf{I} - \mathbf{G}\mathbf{G}^D)\mathbf{G}\mathbf{w} = (\mathbf{G} - \mathbf{G}\mathbf{G}^D\mathbf{G})\mathbf{w} = \mathbf{o}.$$

Obdržíme proto řešení soustavy (14) s minimální normou. Avšak vzhledem k tomu, že podle (94)

$$\overline{\mathbf{u}} = \mathbf{M}^{-1}\begin{pmatrix} \mathbf{d} \\ \mathbf{b} \end{pmatrix} = \mathbf{G}\mathbf{u},$$

vektor $\overline{\mathbf{x}}$ leží v $\mathcal{R}(\mathbf{G})$. □

4 Příklady

Nový postup řešení ilustrujme na následujících malých příkladech, Výsledky budeme zaokrouhlovat na 6 platných cifer. Převážná část prováděných operací přitom byla provedena v dvojnásobné aritmetice.

Příklad 1a. K matici

$$\mathbf{C} = \begin{pmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{pmatrix}$$

přidejme omezení

$$\mathbf{A} = (\,1 \quad -1 \quad 1\,)\,.$$

Protože je

$$\mathcal{N}(\mathbf{C}) = \mathrm{span}\{(\,1 \quad 1 \quad 1\,)^T\}$$

a

$$\mathcal{N}(\mathbf{A}) = \mathrm{span}\{(\,1 \quad 1 \quad 0\,)^T, (\,1 \quad 0 \quad -1\,)^T\},$$

je výsledná matice

$$\mathbf{K} = \begin{pmatrix} 2 & -1 & -1 & 1 \\ -1 & 2 & -1 & -1 \\ -1 & -1 & 2 & 1 \\ 1 & -1 & 1 & 0 \end{pmatrix}$$

dle věty 2 regulární. Po provedení modifikované Choleského faktorizace matice $\mathbf{C}$ obdržíme

$$\mathbf{E} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0.000181675 & 0 \\ 0 & 0 & 0.000181675 \end{pmatrix}.$$

Pro pravé strany $\mathbf{d} = (-1, 1, 1)^T$ a $\mathbf{b} = (1)$ dostaneme řešením soustavy (78)

$$\overline{\mathbf{x}} = \begin{pmatrix} 1.66598 \\ 2.99849 \\ 2.33251 \end{pmatrix}, \quad \overline{\boldsymbol{\lambda}} = 0.999031, \qquad \mathbf{r} = \begin{pmatrix} 4.4 \times 10^{-16} \\ 0.000544749 \\ 0.000423757 \\ 4.8 \times 10^{-13} \end{pmatrix}.$$

Již první krok iteračního zpřesnění dává hledané řešení

$$\overline{\mathbf{x}} + \overline{\delta\mathbf{x}} = \begin{pmatrix} 1.66667 \\ 3.00000 \\ 2.33333 \end{pmatrix}, \quad \overline{\boldsymbol{\lambda}} + \overline{\delta\boldsymbol{\lambda}} = 1.00000, \qquad \mathbf{r} = \begin{pmatrix} 2.2 \times 10^{-16} \\ 2.74793 \times 10^{-7} \\ 1.50223 \times 10^{-7} \\ -4.4 \times 10^{-16} \end{pmatrix}.$$

Příklad 1b. Zadání předchozího příkladu 1 pozměníme u matice omezení takto

$$\mathbf{A} = (\,1 \quad -1 \quad 0\,)\,.$$

Nyní je

$$\mathcal{N}(\mathbf{A}) \;=\; \mathrm{span}\{(\,1 \quad 1 \quad 0\,)^T, (\,0 \quad 0 \quad 1\,)^T\},$$

takže výsledná matice

$$\mathbf{K} = \begin{pmatrix} 2 & -1 & -1 & 1 \\ -1 & 2 & -1 & -1 \\ -1 & -1 & 2 & 0 \\ 1 & -1 & 0 & 0 \end{pmatrix}$$

je dle věty 2 singulární. Modifikovaná Choleského faktorizace matice $\mathbf{C}$ dává pochopitelně stejný výsledek jako dříve. Ježto je

$$\mathcal{N}(\mathbf{C}) \cap \mathcal{N}(\mathbf{A}) = \mathrm{span}\{(\,1 \quad 1 \quad 1\,)^T\},$$

má soustava (14) řešení např. pro pravé strany $\mathbf{d} = (1, 0, -1)^T$ a $\mathbf{b} = (2)$. Řešením soustavy (78) získáme

$$\overline{\mathbf{x}} = \begin{pmatrix} 1.75001 \\ -0.249989 \\ 0.249989 \end{pmatrix}, \quad \overline{\boldsymbol{\lambda}} = -2.50002, \qquad \mathbf{r} = \begin{pmatrix} 4.4 \times 10^{-16} \\ -0.0000454166 \\ 0.0000454166 \\ 0.000000 \end{pmatrix}.$$

Jeden krok iteračního zpřesnění pak dává

$$\overline{\mathbf{x}} + \overline{\delta\mathbf{x}} = \begin{pmatrix} 1.75000 \\ -0.250000 \\ 0.250000 \end{pmatrix}, \quad \overline{\boldsymbol{\lambda}} + \overline{\delta\boldsymbol{\lambda}} = -2.50000, \qquad \mathbf{r} = \begin{pmatrix} -4.4 \times 10^{-16} \\ -2.1 \times 10^{-9} \\ 2.1 \times 10^{-9} \\ 0.000000 \end{pmatrix}.$$

Příklad 2a. K matici

$$\mathbf{C} = \begin{pmatrix} 7 & 3 & 5 & 1 \\ 3 & 7 & 1 & 5 \\ 5 & 1 & 7 & 3 \\ 1 & 5 & 3 & 7 \end{pmatrix}$$

přidáme omezení

$$\mathbf{A} = \begin{pmatrix} 1 & -1 & 0 & 2 \\ 0 & 1 & 2 & -1 \end{pmatrix}.$$

Ježto

$$\mathcal{N}(\mathbf{C}) = \mathrm{span}\{(\,1 \quad -1 \quad -1 \quad 1\,)^T\}$$

a

$$\mathcal{N}(\mathbf{A}) = \mathrm{span}\{(\,2 \quad 2 \quad -1 \quad 0\,)^T, (\,1 \quad -1 \quad 0 \quad -1\,)^T\},$$

je výsledná matice

$$\mathbf{K} = \begin{pmatrix} 7 & 3 & 5 & 1 & 1 & 0 \\ 3 & 7 & 1 & 5 & 1 & -1 \\ 5 & 1 & 7 & 3 & 0 & 2 \\ 1 & 5 & 3 & 7 & 2 & -1 \\ 1 & -1 & 0 & 2 & 0 & 0 \\ 0 & 1 & 2 & -1 & 0 & 0 \end{pmatrix}$$

regulární. Jestliže nyní dosadíme do pravé strany soustavy (14) vektory $\mathbf{d} = (-2, -6, -3.5, 6.5)^T$ a $\mathbf{b} = (-5, 5)^T$, dá modifikovaná Choleského metoda tyto výsledky

$$\mathbf{E} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0.000423882 & 0 & 0 \\ 0 & 0 & 0.000423882 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

$$\overline{\mathbf{x}} = \begin{pmatrix} -1.99998 \\ 0.999982 \\ 1.50000 \\ -1.00000 \end{pmatrix}, \quad \overline{\boldsymbol{\lambda}} = \begin{pmatrix} 2.49994 \\ -1.00033 \end{pmatrix}, \qquad \mathbf{r} = \begin{pmatrix} 8.9 \times 10^{-16} \\ 0.000423874 \\ 0.000635822 \\ 0.00000 \\ -8.9 \times 10^{-12} \\ 3.8 \times 10^{-12} \end{pmatrix}.$$

Po jednom kroku iteračního zpřesnění obdržíme

$$\overline{\mathbf{x}} + \overline{\delta\mathbf{x}} = \begin{pmatrix} -2.00000 \\ 1.00000 \\ 1.50000 \\ -1.00000 \end{pmatrix}, \quad \overline{\boldsymbol{\lambda}} + \overline{\delta\boldsymbol{\lambda}} = \begin{pmatrix} 2.50000 \\ -1.00000 \end{pmatrix}, \qquad \mathbf{r} = \begin{pmatrix} -4.4 \times 10^{-16} \\ 7.5 \times 10^{-9} \\ 2.1 \times 10^{-10} \\ 0.00000 \\ 8.9 \times 10^{-16} \\ 8.9 \times 10^{-16} \end{pmatrix},$$

což je v mezích přesnosti výpočtu přesný výsledek.

Příklad 2b. Oproti předchozímu příkladu změníme matici omezení následovně

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 0 & -1 \\ 0 & -1 & 1 & 0 \end{pmatrix}.$$

Nyní platí, že

$$\mathcal{N}(\mathbf{A}) = \mathrm{span}\{(\,0 \quad 1 \quad 1 \quad 0\,)^T, (\,1 \quad 0 \quad 0 \quad 1\,)^T\},$$

takže máme

$$\mathcal{N}(\mathbf{C}) \cap \mathcal{N}(\mathbf{A}) = \mathrm{span}\{(\,1 \quad -1 \quad -1 \quad 1\,)^T\}.$$

Matice

$$\mathbf{K} = \begin{pmatrix} 7 & 3 & 5 & 1 & 1 & 0 \\ 3 & 7 & 1 & 5 & 0 & -1 \\ 5 & 1 & 7 & 3 & 0 & 1 \\ 1 & 5 & 3 & 7 & -1 & 0 \\ 1 & 0 & 0 & -1 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 & 0 \end{pmatrix}$$

je tudíž singulární. Soustava (14) má řešení např. pro vektory $\mathbf{d} = (7, 3.5, 0.5, -3)^T$ a $\mathbf{b} = (1, -0.5)^T$. Modifikovaná Choleského metoda určí touž matici $\mathbf{E}$, jako v předcházejícím příkladu, a následně dá tyto výsledky

$$\overline{\mathbf{x}} = \begin{pmatrix} 0.750000 \\ 0.250000 \\ -0.250000 \\ -0.250000 \end{pmatrix}, \quad \overline{\boldsymbol{\lambda}} = \begin{pmatrix} 2.50000 \\ -0.999894 \end{pmatrix}, \qquad \mathbf{r} = \begin{pmatrix} 8.9 \times 10^{-16} \\ 0.000105970 \\ -0.000105970 \\ 0.00000 \\ -2.2 \times 10^{-16} \\ -5.6 \times 10^{-17} \end{pmatrix}.$$

Po jednom kroku iteračního zpřesnění pak získáme hodnoty

$$\overline{\mathbf{x}} + \overline{\delta\mathbf{x}} = \begin{pmatrix} 0.750000 \\ 0.250000 \\ -0.250000 \\ -0.250000 \end{pmatrix}, \quad \overline{\boldsymbol{\lambda}} + \overline{\delta\boldsymbol{\lambda}} = \begin{pmatrix} 2.50000 \\ -1.00000 \end{pmatrix}, \qquad \mathbf{r} = \begin{pmatrix} 0.000000 \\ 0.000000 \\ 4.4 \times 10^{-16} \\ 0.000000 \\ 0.000000 \\ 0.000000 \end{pmatrix},$$

což je přesný výsledek.

Reference

[1] Ben-Israel, A., Greville, T. N. E.: *Generalized Inverses. Theory and Applications.* John Wiley & Sons, 1974.

[2] Bunch, J. R., Kaufman, L.: *Some stable methods for calculating inertia and solving symmetric linear systems.* Math. of Comput. **31** 1977, 163–179.

[3] Bunch, J. R., Parlett, B. N.: *Direct methods for solving symmetric indefinite systems of linear equations.* SIAM J. Numer. Anal. **8** 1971, 639–655.

[4] Ciarlet, P. G.: *Introduction to Numerical Linear Algebra and Optimization.* Cambridge Univ. Press, 1988.

[5] Dostál, Z.: *Conjugate projector preconditioning for the solution of contact problems.* Int. J. Numer. Meth. Engng. **34** 1992, 271–277.

[6] Dostál, Z., Friedlander, A., Santos, S. A.: *Augmented lagrangians with adaptive precision control for quadratic programming with equality constraints.* Zasláno k publikaci do Comp. Opt. Appl., 1998.

[7] Dostál, Z., Gomes Neto, F. A. M., Santos, S. A.: *Duality–based domain decomposition with natural coarse–space for variational inequalities.* Připravováno k publikaci, 1998.

[8] Duff, I. S.: *The Solution of Augmented Systems.* Technical report RAL 93–084, Rutherford Appleton Laboratory, 1993.

[9] Duff, I. S., Erisman, A. M., Reid, J. K.: *Direct Methods for Sparse Matrices.* Oxford Univ. Press, Oxford, 1986.

[10] Fiedler, M.: *Speciální matice a jejich použití v numerické matematice.* SNTL, Praha, 1981.

[11] Fletcher, R.: *Practical Methods of Optimization.* Second edition. John Wiley & Sons, Chichester, 1987.

[12] Freund, R. W., Nachtigal, N. M.: *A new Krylov-subspace method for symmetric indefinite linear systems.* AT&T Numerical Analysis Manuscript, Bell Labs, N.J., 1995.

[13] Gill, P. E., Murray, W., Wright, M. H.: *Practical Optimization.* Academic Press, 1981.

[14] Glowinski, R., Le Tallec, P.: *Augmented Lagrangian and Operator-Splitting Methods in Nonlinear Mechanics.* SIAM, Philadelphia, 1989.

[15] Gould, N. I. M.: *On practical conditions for the existence and uniqueness of solutions to the general equality quadratic programmig problem.* Math. Programming **32** 1985, 90–99.

[16] Haslinger, J., Tvrdý, M.: *Approximation and numerical solution of contact problems with friction.* Aplikace Matematiky **28** 1983, 55–71.

[17] Hlaváček, I., Haslinger, J., Nečas, J., Lovíšek, J.: *Numerical solution of variantional inequalities in mechanics.* Springer series in Applied Mathematical Sciences 66, Springer-Verlag, 1988.

[18] Horák, J., Netuka, H.: *Numerical realization of contact problem with friction — semicoercive case.* In: Mathematical methods in Engineering, Plzeň, 1991, 147–152.

[19] Kelley, C. T.: *Iterative Methods for Linear and Nonlinear Equations.* SIAM, Philadelphia, 1995.

[20] Kikuchi, N., Oden, J. T.: *Contact Problems in Elasticity. Study of Variational Inequalities and Finite Element Methods.* SIAM, Philadelphia, 1988.

[21] Kubáček, L.: *Foundations of Estimation Theory.* Elsevier, Amsterdam, 1988.

[22] Meyer, C. D., Plemmons, R. J.: *Convergent powers of a matrix with applications to iterative methods for singular linear systems.* SIAM J. Numer. Anal. **14** 1977, 699–705.

[23] Netuka, H.: *Řešení positivně semidefinitních soustav rovnic kontaktní úlohy.* Výzkumná zpráva č. 2438/86-DVZ/2500. SIGMA Výzkumný ústav, Olomouc, 1986.

[24] Netuka, H.: *Řešení singulárních semidefinitních soustav rovnic.* Preprint 6/1999, PřF UP, katedra MAaAM, Olomouc, duben 1999.

[25] Pšeničnyj, B. N., Danilin, Ju. M.: *Čislennyje metody v ekstremalnych zadačach.* Nauka, Moskva, 1975.

[26] Rao, C. R., Mitra, S. K.: *Generalized Inverse of Matrices and its Applications.* John Wiley & Sons, New York, 1971.

[27] Strang, G.: *Linear Algebra and Its Applications.* Academic Press, 1976.


Univ. Palacki. Olomuc., Fac. rer. nat.,
Dept of Math. Anal. and Appl. of Math.
ODAM (1999) 85–95

Convexity of histogram and convex histopolation

Pavel ŽENČÁK

Department of Mathematical Analysis and Applications of Mathematics,
Faculty of Science, Palacký University,
Tomkova 40, 779 00 Olomouc, Czech Republic
e-mail: zencak@risc.upol.cz

Abstrakt

This paper is concerned with the convexity of histogram and convex histopolation by linear and quadratic splines on refined mesh. First the definition of convexity and the necessary and sufficient criterion of histogram convexity are presented. Then it is proved that this criterion is generally global i.e. the whole histogram must be tested altogether. Algorithms based on this criterion are treated too. Finally the existence of convex quadratic spline interpolating convex histogram on twofold refined mesh is proved and algorithm for its computing is presented.

Key words: Convex histogram, construction of convex histogram interpolant, linear spline, quadratic spline.

1991 Mathematics Subject Classification: 41A15, 65D05

1 Introduction

In various applications it is often necessary to construct a smooth function that interpolates prescribed data and preserves some shape properties of them. In the last years many papers were devoted to such problems. The majority of them treats the problems of positive (see e.g. [7], [10], [14]), monotone (see e.g. [1], [2], [3], [4], [5], [6], [20]) or convex (see e.g. [3], [4], [8], [11], [13], [17]) spline interpolation of prescribed function values. Only few papers (see e.g. [9], [12], [15], [16], [18]) were devoted to problems of shape preserving interpolation of histogram. They suggest that problems of positive or monotone histopolation are only a little more complicated than equivalent problems of function values interpolation and in case of polynomial splines they can be transformed to problems of monotone or convex function values interpolation by splines with order increased by one (see e.g. [9], [12], [19]). But the convex histopolation seems to be more difficult.

2 Convexity of histogram

Let us have given histogram $G = \{g_i\}_{i=0}^n$ on the mesh

$$(\Delta x): \quad x_0 < x_1 < \ldots < x_n < x_{n+1}, \quad \text{with } h_i = x_{i+1} - x_i,\ i = 0(1)n.$$

The first question which arrises is when we can say that histogram is convex.

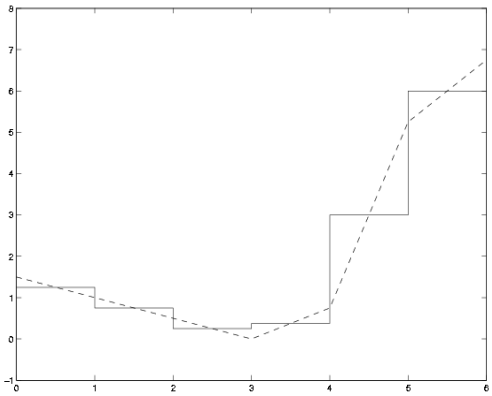


Figure 1: The nonconvexity of linear interpolatory spline on original mesh

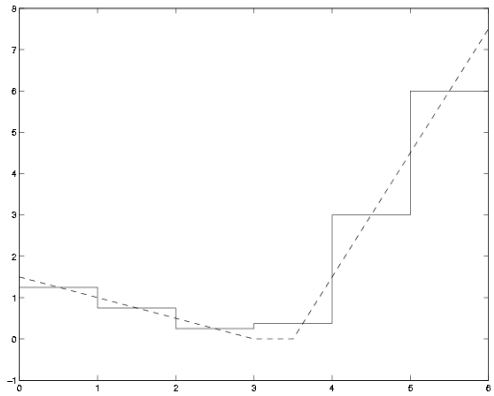


Figure 2: The convexity of linear interpolatory spline on refined mesh

Definition 2.1 We say that histogram G is convex if there exists convex continuous function f interpolating histogram G (i.e. $\int_{x_{i-1}}^{x_i} f(x)dx = h_i g_i$ for $i = 0(1)n$) on mesh (Δx).

This definition is too general for practical testing and so some more simple criterion is needed. In [15], [16] the so called *histogram in convex position* is defined as histogram which can be interpolated by convex linear spline (the

mesh is same as for histogram). But we can see (Fig. 1) that histogram $G = \{1.25, 0.75, 0.25, 0.375, 3, 6\}$ on the mesh $\{0, 1, 2, 3, 4, 5, 6\}$ is not convex according to this criterion although there exist convex linear splines on refined mesh $\{0, 1, 2, 3, 3.5, 4, 5, 6\}$ which interpolate G (see e.g. Fig. 2).

It suggests that better criterion of histogram convexity is the existence of convex linear spline on refined mesh with one added knot to any interval of original mesh. Let us denote $(\Delta^\alpha x) = \{x_i\}_{i=0}^{n+1} \cup \{x_i + \alpha_i h_i\}_{i=0}^{n}$ with $0 < \alpha_i < 1$ and let $S_{11}(\Delta^\alpha x)$ be space of linear splines on the refined mesh $(\Delta^\alpha x)$.

Theorem 2.2 (Necessary and sufficient criterion of convexity) *Histogram G is convex if and only if there exist set of numbers $\{\alpha_i\}_{i=0}^{n}$ and corresponding function $p(x) \in S_{11}(\Delta^\alpha x)$ which interpolates histogram G.*

Another question which arises is if there exists any local criterion of histogram convexity.

Theorem 2.3 *There exists no local criterion of histogram convexity i.e. there not exists $m \in N$ such that for all $n > m$ the convexity of histograms $G_j = \{g_i\}_{i=j}^{m+j}$, $j = 0(1)n - m$ implies the convexity of $G = \{g_i\}_{i=0}^{n}$.*

Remark 2.4 The proofs of previous theorems can be found in [21].

3 Algorithm of convexity testing

In this section we will show the algorithm for convexity testing based on criterion from theorem 2.2 and on idea of so called staircase algorithm (see [3], [13] and [17]). The derivation of this algorithm can be found in [21].

Let us have given histogram $G = \{g_i\}_{i=0}^{n}$ on the mesh (Δx) and let $p(x) \in S_{11}(\Delta^\alpha x)$ interpolate histogram G. Let us denote

$$s_i = p(x_i) \tag{1}$$

$$m_i^+ = p'(x_i + 0) \equiv \lim_{x \to x_i^+} p'(x) \tag{2}$$

$$m_i^- = p'(x_i - 0) \equiv \lim_{x \to x_i^-} p'(x) \tag{3}$$

$$f_m^i(s_i, s_{i+1}, m_i^+) = \frac{s_{i+1}^2 - 2(h_i m_i^+ + s_i)s_{i+1} + 2g_i h_i m_i^+ + s_i^2}{h_i(2g_i - 2s_i - h_i m_i^+)} \tag{4}$$

$$F_m^i(s_i, s_{i+1}, m_{i+1}^-) = \frac{s_i^2 - 2s_i(s_{i+1} - h_i m_{i+1}^-) + s_{i+1}^2 - 2h_i g_i m_{i+1}^-}{h_i(2s_{i+1} - 2g_i - h_i m_{i+1}^-)} \tag{5}$$

Algorithm 1 *Let the sets $W_j \subset R^2$, $j = 0(1)n + 1$ be constructed according to following rules:*

1.

$$W_0 = \left\{ \begin{array}{l} (s_0, m_0^+) : \exists \text{ convex } p_0(x) \in S_{11}(\Delta^\alpha x) \text{ interpolating } G_0 \\ \quad \text{such that } p_0(x_0) = s_0,\ p_0'(x_0 + 0) = m_0^+ \end{array} \right\}$$

2. for $j = 1(1)n$*:*

$$W_j = \left\{ \begin{array}{l} (s_j, m_j^+) : \exists \text{ convex } p_j(x) \in S_{11}(\Delta^\alpha x) \text{ interpolating } G_j \\ \quad \text{and } \exists (s_i, m_i^+) \in W_i \text{ for } i = 0(1)j-1 \text{ such that} \\ \quad p_i(x_i) = s_i,\ p_i'(x_i + 0) = m_i^+ \text{ for } i = 0(1)j \end{array} \right\}$$

3.

$$W_{n+1} = \left\{ \begin{array}{l} (s_{n+1}, m_{n+1}^-) : \exists \text{ convex } p_{n+1}(x) \in S_{11}(\Delta^\alpha x) \\ \quad \text{interpolating } G \text{and } \exists (s_i, m_i^+) \in W_i \text{ for } i = 0(1)n \\ \quad \text{such that } p_i(x_i) = s_i,\ p_i'(x_i + 0) = m_i^+ \text{ and} \\ \quad p_{n+1}(x_{n+1}) = s_{n+1},\ p_{n+1}'(x_{n+1} - 0) = m_{n+1}^- \end{array} \right\}$$

If all $W_j \neq \emptyset,\ j = 0(1)n+1$ *then there exists convex* $p(x) \in S_{11}(\Delta^\alpha x)$ *interpolating histogram* G*.*

Theorem 3.1 *The sets* W_j *from algorithm 1 can be rewritten as:*

1.

$$W_0 = \{(s_0, m_0^+) : \ m_0^+ \leq 2(g_0 - s_0)/h_0\} \tag{6}$$

2. for $j = 1(1)n$*:*

$$W_j = \left\{ \begin{array}{l} (s_j, m_j^+) : \ \exists (s_{j-1}, m_{j-1}^+) \in W_{j-1} \text{ such that} \\ \quad s_j \geq 2g_{j-1} - s_{j-1},\ m_j^+ \geq f_m^{j-1}(s_{j-1}, s_j, m_{j-1}^+), \\ \quad m_j^+ \leq 2(g_j - s_j)/h_j \end{array} \right\} \tag{7}$$

3.

$$W_{n+1} = \left\{ \begin{array}{l} (s_{n+1}, m_{n+1}^-) : \ \exists (s_n, m_n^+) \in W_n \text{ such that} \\ \quad ((s_{n+1} > 2g_n - s_n) \wedge (m_{n+1}^- \geq f_m^n(s_n, s_{n+1}, m_n^+)) \\ \quad \vee ((s_{n+1} = 2g_n - s_n) \wedge (m_{n+1}^- = 2(g_n - s_n)/h_n)) \end{array} \right\} \tag{8}$$

Let us denote

$$\begin{aligned} s_i^d &= \min\{s_i : \ \exists m_i^+ \text{ such that } (s_i, m_i^+) \in W_i\}, \\ m_i^d &= \min\{m_i^+ : \ \exists s_i \text{ such that } (s_i, m_i^+) \in W_i\}, \\ s_i^m &= \max\{s_i : \ \exists m_i^+ \text{ such that } (s_i, m_i^+) \in W_i\} \end{aligned}$$

for $i = 0(1)n$. Then algorithm 1 can be rewriten in the following form:

Algorithm 1A

1. $s_0^d := -\infty,\ s_0^m := \infty, m_0^d := -\infty$
2. $s_1^d := -\infty,\ s_0^m := (g_1h_0 + g_0h_1)/(h_0 + h_1), m_0^d := -\infty$
3. *for* $i = 2(1)n$ *do:*
   $s_i^d := 2g_{i-1} - s_{i-1}^m,\ m_i^d := 2(g_{i-1} - s_{i-1}^m)/h_{i-1}$
   if $2(g_i - s_i^d)/h_i < m_i^d$ *then goto 5*
   else if $2(g_i - s_i^d)/h_i = m_i^d$ *then* $s_i^m := s_i^d$
   else if $s_{i-1}^d = s_{i-1}^m$ *then*
   if $m_{i-1}^d = m_i^d$ *then* $s_i^m := s_i^d$
   else compute $s_i^m > s_i^d$ *such that*
   $f_m^{i-1}(s_{i-1}^d, s_i^m, m_{i-1}^d) = 2(g_i - s_i^m)/h_i$
   else $s_i^m := (g_{i-1}h_i + g_ih_{i-1})/(h_{i-1} + h_i)$
   if $s_i^m > 2g_{i-1} - s_{i-1}^d$ *then compute* $s_i^m > s_i^d$ *such*
   that $f_m^{i-1}(s_{i-1}^d, s_i^m, m_{i-1}^d) = 2(g_i - s_i^m)/h_i$
4. *i:=n+1*
5. *if* $m_{i-1}^d = 2(g_{i-1} - s_{i-1}^d)/h_{i-1}$ *then* $s_i^m := s_i^d$ *else* $s_i^m := \infty$
6. *The interval of histogram convexity is* $[x_0, x_i]$

4 Convex histopolation by linear spline

Let us have given sets $W_i \neq \emptyset$ for $i = 0(1)n + 1$ (i.e. the numbers s_i^d, s_i^m and m_i^d from algorithm 1A and functions F_m^i from (5) for $i = 0(1)n$). Then the general form of algorithm for computing convex $p(x) \in S_{11}(\Delta^\alpha x)$ interpolating histogram G can be written as:

Algorithm 2

1. *Choose some* $(s_{n+1}, m_{n+1}^-) \in W_{n+1}$
2. *for* $j = n(-1)1$ *do :*
   choose some $(s_j, m_j^+) \in W_j \cap F_m^j(s_j, s_{j+1}, m_{j+1}^-)$
   if $s_j = s_j^d$ *then* $m_j^- := m_j^d$ *else choose* m_j^- *such that* $(s_j, m_j^-) \in W_j$
3. *Choose some* $(s_0, m_0^+) \in W_0 \cap F_m^0(s_0, s_1, m_1^-)$

Some simple implementation of previous algorithm can be done by following way:

Algorithm 2A

1. $s_{n+1} := s_{n+1}^d,\ m_{n+1} := m_{n+1}^d$

2. for $j = n(-1)1$ *do :*
$\quad$ *if* $\quad 2g_j - s_{j+1} \le s_j^d \quad$ *then* $\quad s_j := s_j^d \quad$ *else* $\quad s_j := 2g_j - s_{j+1}$
$\quad m_j^+ := F_m^j(s_j, s_{j+1}, m_{j+1}^-)$
$\quad$ *if* $\quad s_j = s_j^d \quad$ *then* $\quad m_j^- := m_j^d \quad$ *else* $\quad m_j^- := m_j^+$

3. $s_0 := 2g_0 - s_1,\ m_0^+ := F_m^0(s_0, s_1, m_1^-)$

The another implementation of algorithm 2 usefull for convex C^1 interpolation (in section 5) follows:

Algorithm 2B

1. if $\quad s_{n+1}^d = s_{n+1}^m \quad$ *then* $s_{n+1} := s_{n+1}^d,\ m_{n+1}^- := m_{n+1}^d$
else $\quad s_{n+1} := s_{n+1}^d + |g_n - g_{n-1}|/8,\ m_{n+1}^- := r_{n+1}(s_{n+1})$

2. for $j = n(-1)1$ *do :*
$\quad$ *if* $\quad s_j^d = s_j^m \quad$ *then* $\quad s_j := s_j^d,\ m_j^+ := F_m^j(s_j, s_{j+1}, m_{j+1}^-),\ m_j^- := m_j^d$
$\quad$ *else* $\quad$ *compute* $s_j^p > 2g_j - s_{j+1}$ *such that* $F_m^j(s_j^p, s_{j+1}, m_{j+1}^-) = r_j(s_j^p)$
$\qquad s_j := (\max\{s_j^d, 2g_j - s_{j+1}\} + s_j^p)/2$
$\qquad m_j^+ := F_m^j(s_j, s_{j+1}, m_{j+1}^-),\ m_j^- := (m_j^+ + r_j(s_j))/2$

3. $s_0 := 2g_0 - s_1 + |2g_0 - s_1|/8,\ m_0^+ := F_m^0(s_0, s_1, m_1^-)$

Some important property of algorithm 2B is given in the following lemma.

Lemma 4.1 *If there exists convex* $p(x) \in S_{11}(\Delta^\alpha x)$ *interpolating histogram* G *such that* $p(x_i - 0) < p(x_i + 0)$ *and* $p(x_i + 0) < p(x_{i+2} - 0)$ *for some* $i = 0(1)n+1$ *then algorithm 2B find such* m_i^-, m_i^+ *and* m_{i+1}^- *that* $m_i^- < m_i^+$ *and* $m_i^+ < m_{i+1}^-$.

5 Convex C^1 histopolation by quadratic spline

Let us have given mesh (Δx) and the numbers β_i, γ_i such that $0 \le \beta_i$, $0 \le \gamma_i$, $\beta_i + \gamma_i \le 1$ for $i = 0(1)n$. Let us denote

$$(\Delta y) \ : \quad y_0 \le y_1 \le \ldots \le y_m, \quad \text{with } l_i = y_{i+1} - y_i$$

where $m = 3n + 3$ and $y_{3i} = x_i$ for $i = 0(1)n + 1$ and $y_{3i+1} = x_i + \beta_i h_i$, $y_{3i+1} = x_i + (\beta_i + \gamma_i)h_i$ for $i = 0(1)n$.

The quadratic spline $S(y)$ on the mesh (Δy) have following local representation on interval $[y_i, y_{i+1}]$ for $i = 0(1)m - 1$:

$$S(y) = p_i(6q - 6q^2) + f_i(3q^2 - 4q + 1) + f_{i+1}(3q^2 - 2q) \tag{9}$$

where $l_i p_i = \int_{y_i}^{y_{i+1}} S(y)dy$, $f_i = S(y_i)$, $f_{i+1} = S(y_{i+1})$. If $l_i > 0$ then local parameter q is given by formula $q = (y - y_i)/l_i$ else $q = 0$.

The previous representation ensures continuity of function values in the knots. To ensure the continuity of the first derivatives (in simple knots only) we must prescribe the following conditions:

$$l_{i+1}f_i + 2(l_i + l_{i+1})f_{i+1} + l_i f_{i+2} - 3l_i p_{i+1} - 3l_{i+1}p_i = 0 \quad \text{for } i = 0(1)m-2 \tag{10}$$

The necessary and sufficient conditions of convexity of $S(y)$ reduces to following inequalities:

$$f_i + f_{i+1} - 2p_i \geq 0 \quad \text{for } i = 0(1)m-1 \tag{11}$$

Theorem 5.1 *Let us have given s_i, s_{i+1}, m_i^+ and m_{i+1}^- for any $i = 0(1)n$ such that there exists convex $p(x) \in S_{11}(\Delta^\alpha x)$ satisfying $p(x_i) = s_i$, $p(x_{i+1}) = s_{i+1}$, $p(x_i + 0) = m_i^+$, $p(x_{i+1} - 0) = m_{i+1}^-$ and interpolating mean value g_i on interval $[x_i, x_{i+1}]$. Let be given $\epsilon_i^0 \geq 0$, $\epsilon_i^1 \geq 0$ and let us denote $f_{3i} = s_i$, $f_{3i+3} = s_{i+1}$, $m_{3i} = m_i^+ - \epsilon_i^0$ and $m_{3i+3} = m_{i+1}^- + \epsilon_i^1$. Then there exist numbers $\beta_i \geq 0$, $\gamma_i \geq 0$, $\beta_i + \gamma_i \leq 1$ such that quadratic spline $S(y)$ interpolates mean value g_i on interval $[y_{3i}, y_{3i+3}]$, function values f_{3i}, f_{3i+3} and the first derivatives m_{3i}, m_{3i+3} in knots y_{3i} and y_{3i+3} and $S(y)$ is convex on interval $[y_{3i}, y_{3i+3}]$.*

Proof Let $p^d(x) \in S_{11}(\Delta^\alpha x)$ satisfy $p^d(x_i) = f_{3i}$, $p^d(x_{i+1}) = f_{3i+3}$, $p^d(x_i + 0) = m_{3i}$, $p^d(x_{i+1} - 0) = m_{3i+3}$. Then $p^d(x)$ is convex on interval $[x_i, x_{i+1}]$ and its mean value g_i^d satisfay $g_i^d \leq g_i$. If $m_{3i} < m_{3i+3}$ then g_i^d is given by formula

$$g_i^d = \frac{(f_{3i+3} - f_{3i})^2 + 2h_i(f_{3i}m_{3i+3} - f_{3i+3}m_{3i}) + h_i^2 m_{3i}m_{3i+3}}{2(m_{3i+3} - m_{3i})}$$

and $p^d(x)$ have knot $x_i + \alpha_i^d h_i$ with $\alpha_i^d = (f_{3i} - f_{3i+3} + h_i m_{3i+3})/(h_i(m_{3i+3} - m_{3i}))$. Then we can compute numbers β_i, γ_i from following formulas:

$$\beta_i = \frac{g_i - (s_i + s_{i+1})/2}{g_i^d - (s_i + s_{i+1})/2}\alpha_i^d, \tag{12}$$

$$\gamma_i = 1 - \beta_i/\alpha_i^d. \tag{13}$$

If $m_{3i} = m_{3i+3}$ then we can use all numbers $\beta_i > 0$, $\gamma_i > 0$, $\beta_i + \gamma_i < 1$ Then the unknown local parameters f_{3i+1}, f_{3i+2}, m_{3i+1}, m_{3i+2}, p_{3i}, p_{3i+1} and p_{3i+2} are computed (using symbolic computing capabilities of MAPLE) as solution of the following system of continuity conditions, boundary conditions and interpolatory

condition on interval $[y_{3i}, y_{3i+3}]$:

$$\begin{pmatrix} 6 & -2 & 0 & 0 & 0 \\ 3\gamma_i & 2(\beta_i+\gamma_i) & -3\beta_i & \beta_i & 0 \\ 0 & 1-\beta_i-\gamma_i & -3(1-\beta_i-\gamma_i) & 2(1-\beta_i) & -3\gamma_i \\ 0 & 0 & 0 & 2 & -6 \\ \beta_i & 0 & \gamma_i & 0 & 1-\beta_i-\gamma_i \end{pmatrix} \begin{pmatrix} p_{3i} \\ f_{3i+1} \\ p_{3i+1} \\ f_{3i+2} \\ p_{3i+2} \end{pmatrix} \tag{14}$$
$$= \begin{pmatrix} \beta_i h_i + 4f_{3i} \\ -\gamma_i f_{3i} \\ -\gamma_i f_{3i+3} \\ (1-\beta_i-\gamma_i)h_i - 4f_{3i+3} \\ g_i \end{pmatrix}$$

Then substituting this local parameters to convexity conditions (11) we find that the resulting spline $S(y)$ is convex on $[x_i, x_{i+1}]$ for all g_i such that $g_i^d \le g_i \le (s_i + s_{i+1})/2$. □

Consequence 5.2 *Let us have the same assumtions as in the previous theorem. Then the following implications hold:*

1. *If* $m_i^+ \neq m_{i+1}^-$ *and* $\epsilon_i^0 > 0$ *or* $\epsilon_i^0 > 0$ *then* $\beta_i > 0$, $\gamma_i > 0$, $\beta_i + \gamma_i < 1$.
2. *If* $m_i^+ \neq m_{i+1}^-$ *and* $\epsilon_i^0 = 0$ *and* $\epsilon_i^0 = 0$ *then* $\beta_i = \alpha_i^d$, $\gamma_i = 0$.
3. *If* $m_i^+ = m_{i+1}^-$ *and* $\epsilon_i^0 > 0$ *or* $\epsilon_i^0 > 0$ *then* $\beta_i = 0$, $\gamma_i = 1$.

Theorem 5.3 *Let us have given convex histogram G on the mesh $\Delta(x)$ such that there exists convex $p(x) \in S_{11}(\Delta^\alpha x)$ interpolating G and satisfying $m_i^+ < m_{i+1}^-$ for $i = 0(1)n$ and $m_i^- < m_i^+$ for $i = 1(1)n$. Then there exists convex quadratic spline $S(y) \in C^1[x_0, x_n]$ which interpolates histogram G.*

Proof The statement is consequence of theorem 5.1 and consequence 5.2 where we put $\epsilon_0^0 \ge$, $\epsilon_n^1 \ge 0$ and $\epsilon_{i-1}^1 = \epsilon_i^0 = (m_i^- + m_{i+1})/2$ for $i = 1(1)n$. □

Let us have given convex histogram G on the mesh (Δx) and compute parameters s_i, m_i^- and m_i^+ for $i = 0(1)n+1$ using algorithm 2B. Then the unknown paramaters of convex quadratic spline interpolating the histogram G can be computed by following algorithm.

Algorithm 3

1. *for* $i = 0(1)n+1$ *do:* $\quad f_{3i} := s_i$
2. *if* $\quad m_{n+1}^- = m_n^+ \quad$ *then* $\quad m_{3n+3} := m_{n+1}^- \quad$ *else* $\quad m_{3n+3} := m_{n+1}^- + \epsilon_n^1$
   for $i = n(-1)1$ *do:*
   if $\quad m_{i+1}^- = m_i^+ \quad$ *then* $\quad m_{3i} := m_i^+$ *else* $\quad m_{3i} := (m_i^- + m_i^+)/2$
   if $\quad m_1^- = m_0^+ \quad$ *then* $\quad m_0 := m_0^+ \quad$ *else* $\quad m_0 := m_0^+ - \epsilon_0^0$

3. for $i = 0(1)n$ *do:*
if $m_{3i} = m_{3i+3}$ *then* $\beta_i = 1/3$ *and* $\gamma_i = 1/3$
else compute β_i *and* γ_i *using formulas (12) and (13)*

4. compute unknown parameters by solving systems (14) for $i = 0(1)n$

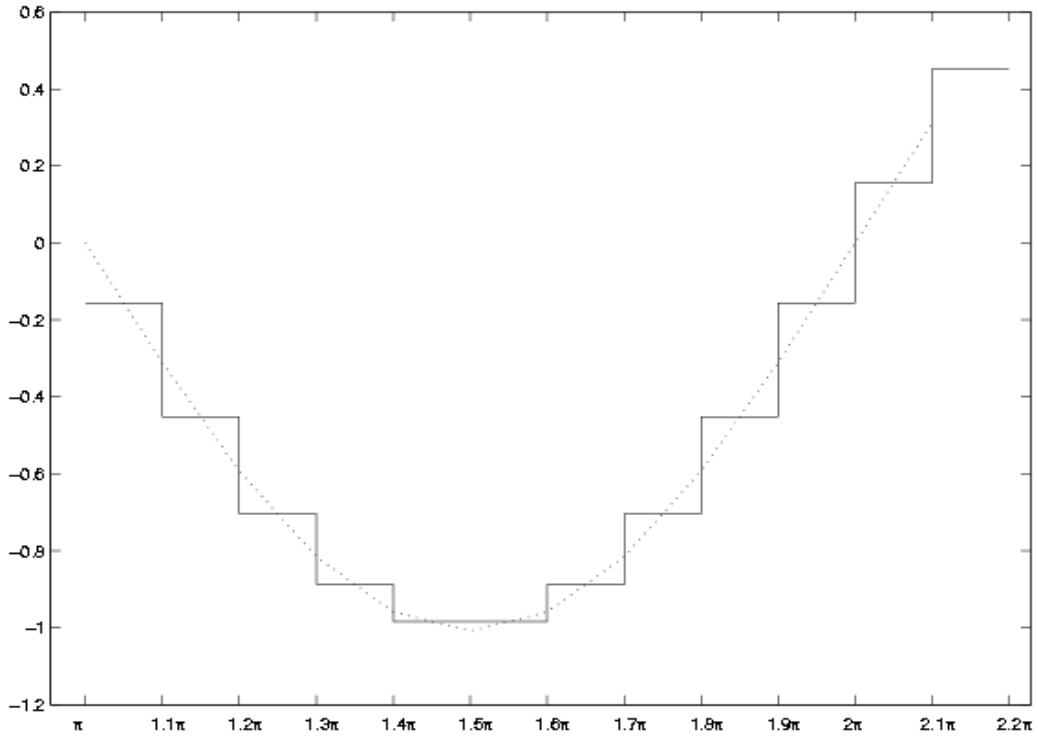


Figure 3

6 Numerical examples

Example 1 The histogram G was obtained as mean values of function $\sin(x)$ on mesh $(\Delta x) = \{\pi + i\pi/10\}_{i=0}^{12}$. The algorithm find that histograms is convex on interval $[\pi, 2\pi + \pi/10]$ in which the interval of convexity of function $\sin(x)$ is contained. The linear interpolatory spline was computed using algorithms 2A (see Fig. 3).

Example 2 The convex histogram G was obtained as mean values of function $\sin(x)$ on mesh $(\Delta x) = \{\pi + i\pi/3\}_{i=0}^{4}$. The linear interpolatory spline was computed using algorithm 2B (see Fig. 4) and quadratic interpolatory spline was computed using algorithm 3 (see Fig. 5).

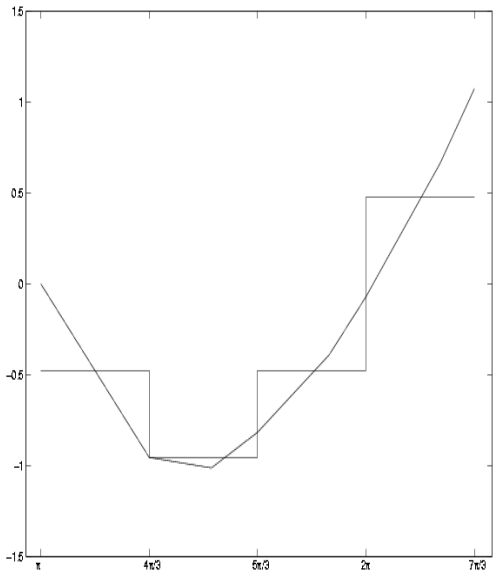


Figure 4

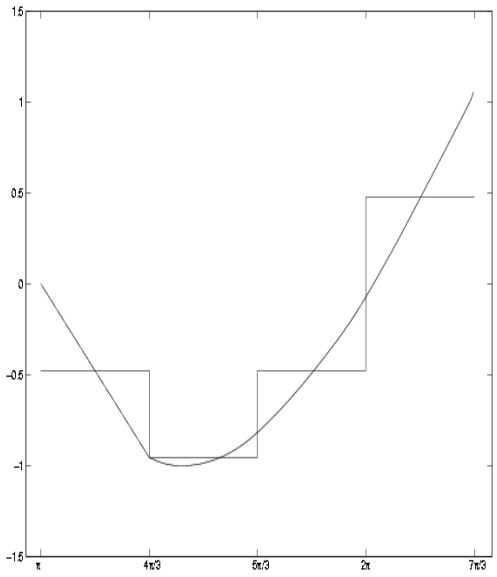


Figure 5

Reference

[1] R. K. Beatson , H. Wolkowics: *Post-processing piecewise cubics for monotonicity.* SIAM J. Numer. Anal. **26**, 2 (1989), 480–502.

[2] C. de. Boor, B. Swartz: *Piecewise monotone interpolation.* Journal of Approximation Theory **21** (1977), 411–416.

[3] P. Costantini, R. Morandi: *Monotone and convex spline interpolation.* Calcolo **21** (1984), 281–294.

[4] P. Costantini: *On monotone and convex spline interpolation.* Mathematics of Computing **46**, 173 (1986), 203–214.

[5] S. C. Eisenstst, K. R. Jackson, J. W. Lewis: *The order of monotone piecewise cubic interpolation.* SIAM J. Numer. Anal. **22**, 6 (1988), 1220–1237.

[6] F. N. Fritsch, R. E. Carlson: *Monotone piecewise cubic interpolation.* SIAM J. Numer. Anal. **17**, 2 (1980), 238–246.

[7] W. Hess, J. W. Schmidt: *Direct methods for constructing positive spline interpolation.* In: Wavelets, Images and Surface Fitting, P. J. Laurent, A. Le Méhauté and L. L. Schumaker (eds.), 1994, 287–294.

[8] W. Hess, J. W. Schmidt: *Convex C^3 interpolation with quartic splines on threefold refined grids.* Preprint TU Dresden, 1994, MATH-NM-12-1994.

[9] W. Hess, J. W. Schmidt: *Shape preserving C^3 data interpolation and C^2 histopolation with splines on threefold refined grids.* Submitted to ZAMM, 1995.

[10] A. Lahtinen: *Positive Hermite interpolation by quadratic splines.* SIAM J. Numer. Anal. **24**, 1 (1993), 223–233.

[11] B. Mulansky, J. W. Schmidt: *Constructive methods in convex interpolation using quartic splines.* Numerical Algorithms **12** (1996), 111–124.

[12] M. Sakai, R. A. Usmani: *A shape preserving area true approximation of histogram by rational splines.* BIT **28** (1988), 329–339.

[13] J. W. Schmidt, W. Hess: *Schwach verkoppelte ungleichungsysteme und konvexe Spline-Interpolation.* Elem. Math. **39** (1984), 85–95.

[14] J. W. Schmidt, W. Hess: *Positivity of cubic polynomials on intervals and positive spline interpolation.* BIT **28** (1988), 340–352.

[15] J. W. Schmidt, W. Hess, Th. Nordheim: *Shape preserving histopolation using rational quadratic splines.* Computing **44** (1990), 245–258.

[16] J. W. Schmidt, W. Hess: *Shape preserving C^2-spline histopolation.* Journal of Approximation Theory **75**, 3 (1993), 325–345.

[17] J. W. Schmidt: *Staircase algorithm and construction of convex interpolants up to the continuity C^3.* In Computers Mathematics Applications (P. Rózsa, J. W. Schmidt, B. A. Szabó (guest eds.), 1995.

[18] J. W. Schmidt: *Dual algorithms for convex approximations of histograms using cubic C^1 splines.* Numerical Analysis and Mathematical Modelling **29** (1994), 35–44.

[19] H. Spaeth: *Eindimensionale Spline-Interpolations-Algorithmen.* Oldenbourgh Verlag, 1990.

[20] Z. Yan: *Piecewise cubic curve fitting algorithm.* Math. Comp. **49**, 179 (1987), 203–213.

[21] P. Ženčák: *Some algorithm for testing convexity of histogram.* Acta Univ. Palacki. Olomuc., Fac. rer. nat. **38** (1999), 149–163.


Univ. Palacki. Olomuc., Fac. rer. nat.,
Dept of Math. Anal. and Appl. of Math.
ODAM (1999) 96–114

Poznámky k řešitelnosti jedné třídy semikoercivních 1D úloh 4. řádu *

Jiří V. Horák

Department of Mathematical Analysis and Applications of Mathematics,
Faculty of Science, Palacký University,
Tomkova 40, 779 00 Olomouc, Czech Republic
e-mail: jhorak@risc.upol.cz

Abstrakt

V předloženém příspěvku se zabýváme problematikou řešitelnosti jedné třídy okrajových úloh s diferenciálním operátorem 4. řádu pro různé typy okrajových podmínek. Důraz je kladen především na typy a kombinace neklasických okrajových podmínek majících za následek semikoercivitu výsledné úlohy, tedy i odlišný přístup k analýze její řešitelnosti. Z důvodů jednoduchosti a stručnosti se v příspěvku omezíme pouze na 1D úlohy reprezentující např. ohyb tenkých pružných těles (nosníku, deskového pásu) a pro okrajové podmínky představující různá mechanická omezení jak průběhů vertikálních průhybů a natočení, tak i jim odpovídajících reakcí v podporách (dané natáčivé či posuvné tření), případně vzájemných kombinací průhybů a natočení se zobecněnými reakcemi (jednostranné pružné natáčivé či posuvné podpory).

Modelovou úlohu lze formulovat jako problém minimalizace konvexního (často nediferencovatelného) funkcionálu nad množinou kinematických vazeb. Ekvivalentně má slabá formulace úlohy tvar eliptické variační rovnice (lineární či nelineární) nebo nerovnice (1. či 2. druhu nebo jejich kombinace), což je v tomto příspěvku upřednostněno. Jsou uvedeny vybrané typové semikoercivní případy včetně poznámek k jejich řešitelnosti:

*Práce byla vypracována s podporou grantu GA ČR č. 105/99/1651.

je indikována nutnost formulace tzv. podmínky řešitelnosti, za jejichž pomoci lze dokázat existenci (případně jednoznačnost) řešení jednotlivých semikoercivních případů (na příslušném faktorprostoru Sobolevova prostoru $H^2(\Omega)$). Tyto podmínky mají pro studovanou třídu úloh a z pohledu mechaniky kontinua (podle charakteru úlohy) různý mechanický význam: buďto jde o požadavky na velikost a orientaci výslednic zatížení nebo jde o „bilanční“ podmínky na velikosti zadávaných veličin (tření, reakcí v podporách a daného zatížení).

1 Úvod

V příspěvku je studována speciální třída jednodimensionálních okrajových úloh s diferenciálními operátory 4. řádu s cílem ukázat vliv zvoleného typu okrajových podmínek na řešitelnost odpovídající úlohy. Zvolená třída úloh reprezentuje z pohledu matematické teorie pružnosti např. ohyb deskového pásu, v jednodušším případě nosníku. Kruhové a mezikruhové desky s rotačně symetrickými daty, patřící také do 1D třídy úloh s diferenciálními operátory 4. řádu, jsou z metodických důvodů — nutno použít aparát Sobolevových váhových prostorů, zjednodušení množiny představující pohyby jako tuhého tělesa ($\mathcal{R} \subset \mathbf{P}_0$) — ponechány samostatnému studiu, viz [10]. Dále klademe v tomto příspěvku důraz na ilustraci a specifikaci jednotlivých matematických odlišností studované problematiky a to jak v odpovídajících výsledných formulacích modelových problémů, tak i ve tvaru odpovídajících podmínek řešitelnosti a následně i metodice vedení důkazů. Na tomto místě je žádoucí poznamenat, že pro speciální případy analogických úloh, ale v rámci teorie svázané termopružnosti, kdy se jednak v důsledku zvolených fyzikálních dat sdružená úloha rozpadne na oddělené a samostatné (postupné) studium pouze osových (rovinných) účinků (*stretching effect*) a pouze ohybových účinků (*bending effect*), lze následující získané výsledky použít přímo také jako jednu, ale podstatnou část řešení původní svázané úlohy (podrobnosti viz např. v [9]). V takovém případě model pro analýzu pouze ohybových účinků je sestaven stále v rámci linearizované teorie svázanné termopužnosti, ale v důsledku zvolených okrajových podmínek pro posunutí může být úloha jednak pouze semikoercivní, a jednak se může dále rozpadnout na tři jednodušší úlohy 2–hého řádu (viz např. [5]).

Z matematického hlediska budeme tedy formulovat úlohy především s neklasickými okrajovými podmínkami, tj. předepisujícími vhodné omezení hodnot neznámé funkce (posunutí) nebo jejích derivací (natočení, případně vyšších derivací reprezentujících ohybové momenty a posouvající síly). Tedy budeme studovat vliv okrajových podmínek Signoriniho typu či podmínek zachycujících vliv tření na hranici oblasti (tj. v podporách nosníku či desky), nebo jejich vzájemné sdružení prostřednictvím zadání pružných posuvných či natáčivých uložení podpor,

případně některá jejich další omezení a možné kombinace (jednostranné pružné podpory včetně vlivu tření). Samozřejmě můžeme také uvažovat smíšené okrajové podmínky, tj. různé typy v různých podporách, pro naše účely to má ale smysl pouze v kombinacích majících za výsledek semikoercivitu úlohy.

Klasické okrajové podmínky Dirichletova nebo Neumannova typu zde samostaně nebudeme analyzovat. Z metodického hlediska je takový přístup oprávněn, neboť lze ukázat, že některé typy úloh jsou limitními případy obecnějších situací (vhodnou manipulací s tuhostí pružného uložení — koeficientem v Newtonově okrajové podmínce — může Newtonova úloha s neklasickými okrajovými podmínkami přejít na Dirichletovu či Neumannovu úlohu s klasickými okrajovými podmínkami, zatímco jednostranná Newtonova úloha aproximuje např. Signoriniho úlohu, atd.).

Odpovídající matematický model v tomto příspěvku studované třídy úloh bude tedy vždy obsahovat stejnou lineární diferenciální rovnici 4. řádu a odlišnost bude pouze v typu příslušných okrajových podmínek.

Zde budeme tedy soustředit pozornost především na charakteristické aspekty jednotlivých případů studované třídy úloh, to je např. na tvar množiny kinematicky přípustných průhybových funkcí, typ variační nerovnice a charakter případné semikoercivity, tvar množiny tuhých posunutí, formulace nových, ale jak se ukáže ekvivalentních norem, vlastnosti funkcí definujících různé modely tření v kombinaci s pružnými či jednostrannými podporami, atd., zatímco samotnou mechanickou stránku problematiky (mechanický význam okrajových podmínek i podmínek řešitelnosti, atd.) ponecháme stranou — jednak byla částečně již diskutována např. v [5] a jednak je zde našim cílem především analýza řešitelnosti.

Dále, z důvodů omezenosti rozsahu příspěvku, zde nebudeme při analýze zacházet do všech podrobností (pro detailní diskusi nejdůležitějších analogických a příbuzných aspektů pro 2D úlohy matematické teorie pružnosti viz např. [3], pro 1D úlohy 4. řádu viz [5]). Pouze zdůrazníme některé významnější odlišnosti a ideje důkazové techniky pro jednotlivé případy slabých formulací modelových úloh a případné důsledky těchto odlišností, tj. např. typ lineárního prostoru či jeho konvexní podmnožiny (množiny přípustných funkcí), na kterém hledáme řešení, tvar úplné bilineární formy ve funkcionálu potenciální energie (včetně hraničních forem), výskyt různých typů nediferencovatelných členů v důsledku modelování tření, charakter úlohy — variační rovnice či nerovnice, semikoercivitu a prostor tuhých posunutí, nutnost formulace dodatečných podmínek řešitelnosti, atd., a to vše v závislosti právě na typu zvolených okrajových podmínek.

Pokud jde o problematiku řešitelnosti, má jak známo, slabá formulace (princip virtuálních prací) studované třídy úloh v případě klasických a Newtonových okrajových podmínek tvar *variační rovnice* (pro jednostranné Newtonovy okrajové podmínky, tj. uvažujeme-li pouze u^+ či Du^+, jde o nelineární rovnici) a *nerovnice* (ve všech ostatních případech, viz např. [4]). Navíc, v semikoercivních

případech, tj. když těleso (nosník, deska) není pevně uchyceno a má tedy v důsledku zvoleného typu okrajových podmínek (např. obě okrajové podmínky pro nosník mají charakter jednostranných pružných nebo tuhých podpor, nebo jednostranných či oboustranných natáčivých a posuvných podpor se třením) ponechány nějaké stupně volnosti, může dojít k jeho pohybu jako tuhého tělesa $\mathcal{R}$ (pro zde studované případy bude vždy $\mathcal{R} \subset \mathbf{P}_1$). Z matematického hlediska to znamená, že příslušná bilineární forma (reprezentující deformační práci vnitřních sil) není na odpovídajícím prostoru virtuálních posunutí eliptická. Proto k zajištění existence a jednoznačnosti řešení musíme v těchto případech navíc formulovat tzv. podmínky řešitelnosti úlohy, jež lze fyzikálně inerpretovat jako podmínky rovnováhy zatížení a reakcí v podporách, tedy buďto jako podmínky vynucující správný směr výslednice zatížení a reakcí (např. pro okrajové podmínky Signoriniho typu), nebo podmínky bilancující, např. prostřednictvím zatížení f, velikosti předepsaného natáčivého ($g_{M,0}$, $g_{M,L}$) a posuvného ($g_{T,0}$, $g_{T,L}$) tření s reakcemi $M(i)$, $i = 0, L$, $T(i)$, $i = 0, L$, odpovídajícími ohybovému momentu a posouvající síle v podporách.

K řešení uvedené problematiky použijeme standardní přístup (metodicky sledujeme postupy uvedené zejména v [3]) i aparát — variační počet a teorii prostorů integrovatelných funkcí se zobecněnými derivacemi (viz např. [1]–[4]).

2 Rovnice modelové úlohy

V tomto příspěvku budeme používat následující označení: $L \in \mathbf{R}^1$, $L > 0$ pro délku nosníku (deskového pásu), E pro Youngův modul pružnosti, J pro moment setrvačnosti (přičemž předpokládáme, že existují $E_o, J_o \in \mathbf{R}^1$ tak, že platí $E(x) \geq E_o > 0$, $J(x) \geq J_o > 0$). Pro hladkou funkci (např. pro $w \in C^3(\Omega)$) značí $\mathcal{M}(w)(x) = -E(x)J(x)D^2w(x)$ a $\mathcal{T}(w)(x) = D\mathcal{M}(w)(x)$ ohybový moment a posouvající sílu odpovídající průhybu $w = w(x)$, kde $x \in \Omega$ pro $\Omega = (0, L)$, $\partial\Omega = \{0, L\}$, dále z důvodů stručnosti označujeme $D \equiv \frac{d}{dx}$. Nakonec definujeme

$$\begin{aligned}
&\mathcal{M}^+(w)(0) = \lim_{x\to 0^+}(-E(x)J(x)D^2w(x)) \ , \\
&\mathcal{M}^-(w)(L) = \lim_{x\to L^-}(-E(x)J(x)D^2w(x)) \ , \\
&\mathcal{T}^+(w)(0) = \lim_{x\to 0^+}(D(-E(x)J(x)D^2w(x))) \ , \\
&\mathcal{T}^-(w)(L) = \lim_{x\to L^-}(D(-E(x)J(x)D^2w(x))) \ .
\end{aligned}$$

V tomto příspěvku se omezíme na analýzu úloh s lineárním diferenciálním operátorem 4. řádu $\mathcal{A} : X \to Y$, definovaným vztahem

$$(\mathcal{A}u)(x) = D^2(p(x)D^2u(x)) + q(x)u(x), \ x \in \Omega.$$

Potom pro definici klasického řešení úlohy volíme za X a Y vhodné prostory hladkých funkcí, přičemž koeficienty p, q jsou také dostatečně hladké funkce (zde např. stačí volit $X = C^4(\Omega) \cap C^3(\bar{\Omega})$, $Y = C(\Omega)$, $p \in C^2(\Omega) \cap C^1(\bar{\Omega})$, $q \in C(\Omega)$).

Z praktických důvodů (např. průřezové charakteristiky nosníku, materiálové vlastnosti či výpočtové aspekty diskretizované úlohy, atd.) budeme dále pracovat pouze se slabými formulacemi úloh, tedy předpokládáme, že pro koeficienty úlohy platí $p, q \in L_\infty(\Omega)$, zatímco pro pravou stranu (zatížení) stačí vzít $f \in L_2(\Omega)$ (samozřejmě lze volit i podstatně obecnější typ zatížení, např. $f \in [H^2(\Omega)]^*$, ale pro naše účely takovou volbou zatížení nezískáme silnější výsledky týkající se např. podmínek řešitelnosti úlohy v semikoercivních případech). Dále prostor funkcí s konečnou energií zvolíme jako $X = H^2(\Omega)$, kde $H^k(\Omega)$, k=1, 2 značí standardní Sobolevovy prostory funkcí $H^k(\Omega) = \{v \in L_2(\Omega) \mid D^i v \in L_2(\Omega), i = 1, ..., k\}$ a $[H^2(\Omega)]^*$ jejich duály, pro více podrobností, tj. přesnou definici, smysl derivací a případné hladkosti funkcí (včetně vět o vnoření), viz např. [1] nebo [2].

Protože v této práci chceme studovat především problematiku semikoercivity, omezíme se pouze na situaci, kdy platí $q \equiv 0$. Pokud by platilo $q(x) \geq q_o > 0$ pro $\forall x \in \Omega_o \subset \Omega$ a Ω_o by byla množina kladné 1D-míry, tj. $\mu(\Omega_o) > 0$, pak by šlo o úlohu nosníku uloženého na části podloží Winklerovského typu, jenž je, pro podstatně obecnější modely (nelineárního) podloží, podrobně studována v [8]. Na tomto místě připomínáme, že pokud uvažujeme pouze tzv. klasické Winklerovské podloží (tedy oboustranné) a $\mu(\Omega_o) > 0$, lze triviálně ukázat jednoznačnou řešitelnost takové úlohy (odpovídající bilineární forma reprezentující potenciální energii vnitřních sil tvoří totiž ekvivalentní skalární součin na $X = H^2(\Omega)$). Pro obecnější modely, např. pro jednostranné Winklerovské podloží nebo pro podloží s vhodným modelem tření či pro jejich kombinace může být úloha (v závislosti na zvoleném typu okrajových podmínek) opět semikoercivní, tedy i v takovém případě se znovu objeví otázka stanovení podmínek řešitelnosti.

Za uvedených předpokladů nabývá operátor $\mathcal{A} : H^2(\Omega) \to [H^2(\Omega)]^*$ (připomínáme, že platí

$$H^2(\Omega) \subset L_2(\Omega) \equiv [L_2(\Omega)]^* \subset [H^2(\Omega)]^*)$$

tvaru

$$(\mathcal{A}u)(x) = D^2(p(x)D^2u(x)), \quad u \in H^2(\Omega), \text{ s.v. } x \in \Omega,$$

kde nyní máme $p \in L_\infty(\Omega)$, kde $p(x) = E(x)J(x)$ a $x \in \Omega$. Zobecnění Greenovy věty pro naši okrajovou úlohu můžeme nyní psát v následujícím tvaru

$$\int_\Omega \mathcal{A}u(x)v(x)dx = \langle \gamma_{N,A}(u), \gamma_D(v) \rangle + a(u, v), \ u \in H^2(\Omega), \ \forall v \in H^2(\Omega),$$

kde $\langle \gamma_{N,A}(u), \gamma_D(v) \rangle$ značí dualitu mezi prostory stop funkcí a jejich duálními prostory, přičemž (viz např. [1]) $\gamma_{N,\mathcal{A}}(u) \equiv \{\mathcal{M}(u), \mathcal{T}(u)\}$ značí Neumannův operátor stop zatímco označení $\gamma_D(u) \equiv \{u, Du\}$ je užito pro standardní Dirichletův operátor stop na $H^2(\Omega)$. Poněvadž vycházíme z rovnosti $\mathcal{A}(u) = f$, dostáváme užitím předpokladu na $f \in L_2(\Omega)$ inkluzi $\mathcal{A}(u) \in L_2(\Omega)$, jež zajišťuje, jak lze ukázat, smysluplnost definice operátoru $\gamma_{N,\mathcal{A}}(u) \equiv \{\mathcal{M}(u), \mathcal{T}(u)\}$, a tedy oprávněnost předchozí formulace Greenovy věty. Připomínáme, že pro hladké funkce

z $C^3(\bar{\Omega})$ má operátor $\gamma_{N,\mathcal{A}}(u)$ význam restrikce hodnot druhých a třetích derivací funkcí $u \in C^3(\bar{\Omega})$ na $\partial\Omega$ (což mimo jiné ospravedlňuje použití stejného označení pro ohybový moment a posouvající sílu jako v předchozí definici: $\{\mathcal{M}(u), \mathcal{T}(u)\}$). Pro funkce z prostoru $H^2(\Omega)$ je obecně situace zcela odlišná, uvedený operátor restrikce hladkých funkcí je třeba vhodně prodloužit na celý $H^{1/2}(\partial\Omega) \times H^{3/2}(\partial\Omega)$ se zachováním normy, tedy definovat jako spojitý lineární funkcionál $\gamma_{N,\mathcal{A}}(u)$ nad prostorem stop funkcí právě z $H^2(\Omega)$ (podrobnosti lze nalézt opět v [1] nebo v [8]). V důsledku našeho $1D$ modelu je však k dispozici vnoření

$$H^2(\Omega) \subset C^{1,1/2}(\bar{\Omega})$$

a tedy problematika stop se zjednoduší.

Tedy pokud označíme reakce v podporách jako T a M, dostaneme pro u dostatečně hladké následující, velmi speciální, ale standardně používanou podobu předchozí Greenovy věty

$$a(u,v) = T(0)v(0) + T(L)v(L) - M(0)Dv(0) - M(L)Dv(L) + \mathcal{F}(v)\ ,$$
$$u \in H^2(\Omega),\ \forall v \in H^2(\Omega),$$

kde jsme užili standardních vztahů mezi vnitřními „silami“ a vnějšími veličinami (zobecněnými silami)

$$M(0) = -\mathcal{M}^+(u)(0),\ M(L) = \mathcal{M}^-(u)(L),$$
$$T(0) = -\mathcal{T}^+(u)(0),\ T(L) = \mathcal{T}^-(u)(L),$$

a následující definice bilineární formy $a : H^2(\Omega) \times H^2(\Omega) \to \mathbf{R}^1$ vztahem

$$a(u,v) = (EJD^2u, D^2v)_{L_2(\Omega)} = \int_\Omega E(x)J(x)D^2u(x)D^2v(x)dx \text{ pro } u,v \in H^2(\Omega)$$

a definice lineárního spojitého funkcionálu $\mathcal{F} : H^2(\Omega) \to \mathbf{R}^1$ pro $f \in L_2(\Omega)$ vztahem $\mathcal{F}(v) = \langle f, v\rangle \equiv (f,v)_o$ pro $v \in H^2(\Omega) \subset L_2(\Omega)$.

Lineární prostor virtuálních posunutí (tj. funkcí reprezentujících průhyby splňující homogenní okrajové podmínky) pro jednotlivé úlohy budeme značit $\mathcal{V}$, jeho konvexní podmnožinu kinematicky přípustných průhybových funkcí označíme $\mathcal{K}$, tedy $\mathcal{K} \subset \mathcal{V}$, přičemž vždy bude platit inkluze

$$H_o^2(\Omega) \subset \mathcal{V} \subset H^2(\Omega)\ .$$

Nakonec, pro dané hodnoty $g_{T,x}, g_{M,x} \geq 0$, definujeme spojité konvexní, ale nediferencovatelné funkcionály $j_{i,x}$ a $j_{i,x}^+$, tj. $j_{i,x}, j_{1,x}^+ : H^2(\Omega) \to \mathbf{R}^1$, $i = 1,2$, $x = 0, L$ reprezentující práci (pro zde zvolený model tzv. „daného“) tření v oboustranných, ale i jednostranných natáčivých a posuvných podporách pomocí předpisů

$$j_{1,x}(v) = g_{T,x}|v(x)|, \quad j_{1,x}^+(v) = g_{T,x}(v(x))^+\ ,\ x = 0, L,$$
$$j_{2,x}(v) = g_{M,x}|Dv(x)|, \quad j_{2,x}^+(v) = g_{M,x}(Dv(x))^+\ ,\ x = 0, L$$
$$j(v) = \sum_{i=1,x=0}^{2,L}(j_{i,x}(v) + j_{i,x}^+(v)),$$

přičemž nepožadujeme, aby současně byly všechny zadané hodnoty $g_{T,x}, g_{M,x} \geq 0$ nenulové, tedy tření může být např. zadáno jen v jedné podpoře a třeba jen pro jeden druh namáhání (např. pro natočení).

3 Okrajové podmínky

Z metodických důvodů a pro úplnost přehledu vlivu typu okrajových podmínek na tvar variační formulace úlohy i na její řešitelnost uvádíme v tomto příspěvku, i když velmi stručně, také problematiku týkající se klasických okrajových podmínek, a tam, kde je to možné i jejich souvislosti s podmínkami neklasickými.

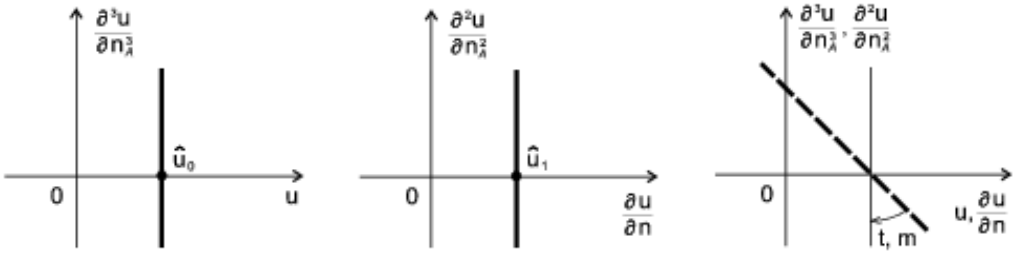
3.1 Klasické okrajové podmínky

3.1.1 *Dirichletovy okrajové podmínky*

Pro Dirichletovy, tj. stabilní okrajové podmínky nebo-li podmínky 1. druhu, reprezentující oboustranně vetknutý nosník s předepsaným poklesem a natočením podpor, předepisujeme $u(x) = \hat{u}_o(x)$, $Du(x) = \hat{u}_1(x)$ v $x = 0, L$. Potom $\mathcal{V} = H_0^2(\Omega)$ a pro zadanou funkci $\hat{u} \in H^2(\Omega)$, kde $\hat{u}$ je funkce reprezentující okrajové podmínky, tj. platí $\gamma_D(\hat{u}) \equiv \{\hat{u}, D\hat{u}\}|_{\partial\Omega} = \{\hat{u}_o, \hat{u}_1\} \in \mathbf{R}^4$, má úloha tvar následující variační rovnice, jež je jednoznačně řešitelná:

$$\exists!\ u \in H_0^2(\Omega) : u - \hat{u} \in \mathcal{V}\ ,\ a(u,v) = \mathcal{F}(v) \quad \forall v \in \mathcal{V}\ .$$

Existence i jednoznačnost řešení Dirichletovy úlohy plyne jednak ze skutečnosti, že lineární prostor kinematicky přípustných tuhých posunutí obsahuje pouze nulový prvek, tj. platí $\mathcal{R} \cap \mathcal{V} = \{0\}$, kde $\mathcal{R} = \mathbf{P}_1$, a jednak z Friedrichsovy nerovnosti.



Obr. 1: Znázornění okrajových podmínek Dirichletova typu

Z obrázku snadno vidíme, že klasické okrajové podmínky Dirichletova typu jsou speciálním případem neklasických okrajových podmínek Newtonova typu (viz dále), a to pro následující limitní přechody tuhostí odpovídajících jak posuvným tak i natáčivým podporám, tj. pro $t, m \to \infty$. Důkazy tohoto tvrzení ani podrobnosti zde nebudeme uvádíme.

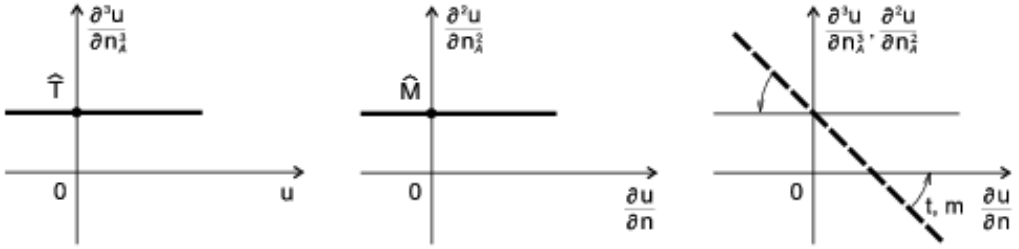
3.1.2 *Neumannovy okrajové podmínky*

V případě Neumannových nestabilních okrajových podmínek nebo-li podmínek 2. druhu je situace oproti předchozímu případu podstatně komplikovanější. Jednotlivá data úlohy f, $\hat{M}$, $\hat{T}$ nelze totiž zadat libovolně, neboť v takovém případě by nebyla zajištěna její řešitelnost: prostor virtuálních posunutí má nyní tvar $\mathcal{V} = H^2(\Omega)$ a bilineární forma a je na $\mathcal{V}$, jak lze snadno vidět, pouze semikoercivní — nulovány jsou všechny funkce z lineárního prostoru tuhých posunutí $\mathcal{R} = \mathbf{P}_1$, když nyní zřejmě platí $\mathcal{R} \cap \mathcal{V} = \mathcal{R}$.

Tedy pro nehomogenní okrajové podmínky (dané zatížení konců nosníku) zadáváme $M(x) = \hat{M}(x)$, $T(x) = \hat{T}(x)$ v $x = 0, L$ a přeformulujeme lineární funkcionál $\mathcal{F}$ na $\tilde{\mathcal{F}}(v) = \mathcal{F}(v) + \sum_{x=0}^{L}(\hat{T}(x)v(x) - \hat{M}(x)Dv(x))$.

Z důvodu semikoercivity bilineární formy musíme řešitelnost úlohy zajistit formulací tzv. podmínek řešitelnosti. Ty mají pro naši úlohu tvar rovnovážných podmínek (nulových výslednic zatížení a reakcí, pro podrobnosti viz např. [7], [8] a [9]), a při jejich splnění má vyšetřovaná úloha, ale pouze na faktorprostoru $H^2(\Omega)/\mathcal{R}$, kde $\mathcal{R} = \mathbf{P}_1$, jediné řešení

$$\exists!\ cl(u) \in H^2(\Omega)/\mathcal{R}\ :\ a(u_{\mathcal{R}}, v) = \tilde{\mathcal{F}}(v) \quad u_{\mathcal{R}} \in cl(u), \forall v \in \mathcal{V}\ .$$



Obr. 2: Znázornění okrajových podmínek Neumannova typu

Opět lze snadno nahlédnout, že také Neumannovy okrajové podmínky jsou speciálním případem podmínek Newtonových, ale nyní pro limitní přechody odpovídajících tuhostí k nule, tj. pro $t, m \to 0$.

3.1.3 *Kombinace stabilních a nestabilních okrajových podmínek*

Významné postavení, zejména v úlohách svázané termopružnosti (viz např. [5] nebo [9]), zaujímá úloha ve které zadáme speciální kombinace okrajových podmínek 1. a 2. druhu (případně analogické kombinace jejich zobecnění — viz [9] a dále). V případě tzv. „prostého podepření“ reprezentující zadání poklesu podpor se zadaným momentovým zatížením konců nosníku předepisujeme $u(x) = u_0(x)$ a $M(x) = \hat{M}(x)$ v $x = 0, L$.

Na tomto místě poznamenejme, že lze také zadat kombinace velikostí derivace (natočení) Du a zatížení $\hat{T}$ v podporách $x = 0, L$, potom definujeme lineární prostor testovacích funkcí ve tvaru $\mathcal{V} = \{v \in H^2(\Omega) \mid Dv(x) = 0, x = 0, L\}$.

V tomto případě je ale nutné formulovat opět podmínky řešitelnosti; jednoznačnost bychom mohli získat při splnění zmíněných podmínek řešitelnosti, ale pouze na faktorprostoru $H^2(\Omega)/\mathcal{X}$, kde nyní $\mathcal{X} = \mathcal{V} \cap \mathcal{R} = \mathbf{P}_0$. Úloha má tvar

$$\exists!\ cl(u) \in H^2(\Omega)/\mathcal{R}\ :\ cl(u) = cl(\hat{u})\ ,\ a(u_{\mathcal{R}}, v) = \tilde{\mathcal{F}}(v) \quad u_{\mathcal{R}} \in cl(u), \forall v \in \mathcal{V}\ ,$$

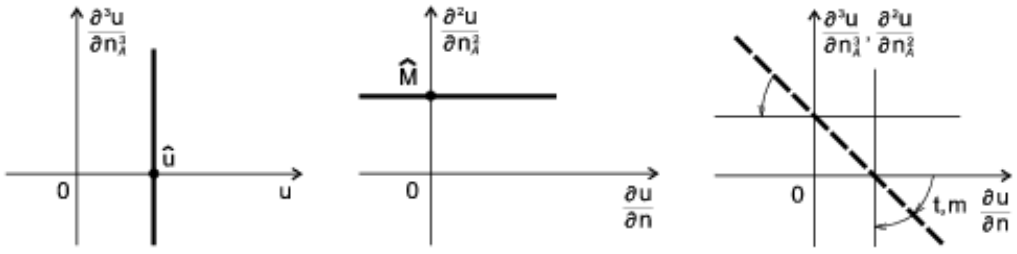
kde nyní $\hat{u} \in H^2(\Omega)$ je funkce reprezentující zadanou hodnotu natočení na hranici, tj. platí $\gamma_D(\hat{u}) \equiv \{\ .\ , D\hat{u}\}|_{\partial\Omega} = \{\ .\ , \hat{u}_1\}$ a funkcionál $\tilde{\mathcal{F}}$ je definován vztahem $\tilde{\mathcal{F}}(v) = \mathcal{F}(v) + \sum_{x=0}^{L} \hat{T}(x)v(x)$. Zdá se však, že takto zadaná úlohy má jen malý praktický význam v aplikacích, proto se s ní nebudeme dále podrobněji zabývat.

Prostor virtuálních posunutí má v prvním, podstatně významnějším případě tvar $\mathcal{V} = H^2(\Omega) \cap H_0^1(\Omega)$, příslušný lineární funkcionál reprezentující potenciální energii vnějších sil je dán vztahem $\tilde{\mathcal{F}}(v) = \mathcal{F}(v) - \sum_{x=0}^{L} \hat{M}(x)Dv(x)$ a protože lze ukázat, že bilineární forma a je na $\mathcal{V}$ koercivní, má takto formulovaná úloha jediné řešení, tj.

$$\exists!\ u \in \mathcal{V}\ :\ a(u, v) = \tilde{\mathcal{F}}(v) \quad \forall v \in \mathcal{V}\ .$$

Existence i jednoznačnost řešení úlohy s uvedenou kombinací okrajových podmínkam dostaneme z faktu, že lineární prostor kinematicky přípustných tuhých posunutí obsahuje opět pouze nulový prvek, tj. platí $\mathcal{R} \cap \mathcal{V} = \{0\}$, kde $\mathcal{R} = \mathbf{P}_1$, a pak užitím tvrzení, že druhá seminorma $|D^2u|_{H_2(\Omega)} = \|D^2u\|_{L_2(\Omega)}$ funkce ze Sobolevova prostoru $H^2(\Omega)$ je normou na jeho podprostoru $\mathcal{V}$ a to normou ekvivalentní se standardní normou definovanou v $H^2(\Omega)$.

Poznamenejme, že uvedená kombinace okrajových podmínek patří do třídy okrajových podmínek umožňujících dekompozici diferenciálního operátoru 4. řádu a následně pak užitím metody faktorizace lze eliminovat svázannost v úloze termopružnosti, viz např. [5] a [9].



Obr. 3: Znázornění kombinovaných okrajových podmínek

Z obrázku lze opět snadno vidět, že i výše uvedené kombinované okrajové podmínky jsou také speciálním případem neklasických okrajových podmínek Newtonových, ale nyní pro odlišný limitní přechod jednotlivých tuhostí odpovídajících posuvným a natáčivým podporám, tj. pro $t \to \infty$ a $m \to 0$. Důkazy tohoto tvrzení ani podrobnosti zde opět nebudeme uvádět.

3.1.4 *Smíšené okrajové podmínky*

Zatím jsme se zabývaly pouze okrajovými podmínkami stejného typu na celé hranici, to je předepsanými jak v $x = 0$, tak i v $x = L$. Je však zřejmé, že lze předepsat okrajové podmínky různého typu pro odlišné části hranice. Lze tak předepsat okrajové podmínky např. Dirichletova typu na jednom konci nosníku, tj. pro $x = 0$, a Neumannova typu na druhém konci nosníku pro $x = L$, či Dirichletova typu v $x = 0$ a prostého podepření v $x = L$, atd. Řešitelnost výsledné úlohy plyne potom buď ze zobecněné Friedrichsovy nerovnosti, nebo aplikací některého z předchozích postupů.

Opět je však třeba upozornit na některé speciální situace: např. pro kombinace okrajových podmínek Neumannova typu v $x = L$ a prostého podepření v $x = 0$ nelze data úlohy zadat libovolně, ale pro existenci a jednoznačnost řešení úlohy je nutné zajistit, aby data splňovala odpovídající podmínky řešitelnosti (v tomto zmíněném případě bude mít lineární prostor virtuálních posunutí splňujících homogenní Dirichletovy okrajové podmínky tvar $\mathcal{V} = \{v \in H^2(\Omega) \mid v(0) = 0\}$ a prostor kinematicky přípustných tuhých posunutí má potom tvar $\mathcal{R} \cap \mathcal{V} = \mathbf{P}_{1,0}$).

3.2 Neklasické okrajové podmínky: Rovnice

3.2.1 *Newtonovy okrajové podmínky*

V tomto případě jde o první případ tzv. neklasických okrajových podmínek, reprezentujících jak pružné posuvné tak i pružné natáčivé podepření konců nosníku, kdy předepisujeme

$$T(x) = -t(x)u(x) + \hat{T}(x), \quad M(x) = m(x)Du(x) + \hat{M}(x), \ x = 0, L,$$

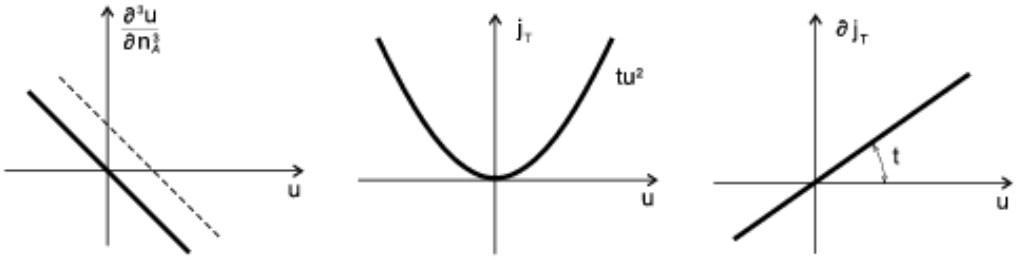
kde koeficienty $t, m \geq 0$ reprezentují tuhosti odpovídající posuvným a natáčivým podporám. Pro příslušnou variační formulaci úlohy musíme poněkud předefinovat jak předchozí bilineární formu $a : \mathcal{V} \times \mathcal{V} \to \mathbf{R}^1$ (přidáním tzv. hraničních bilineárních forem) tak i lineární formu $\mathcal{F} \in \mathcal{V}^*$ (přidáním zatížení na hranici - volných koncích), a to následujícím způsobem

$$\tilde{a}(u, v) = a(u, v) + \sum_{x=0}^{L} (t(x)u(x)v(x) + m(x)Du(x)Dv(x)),$$

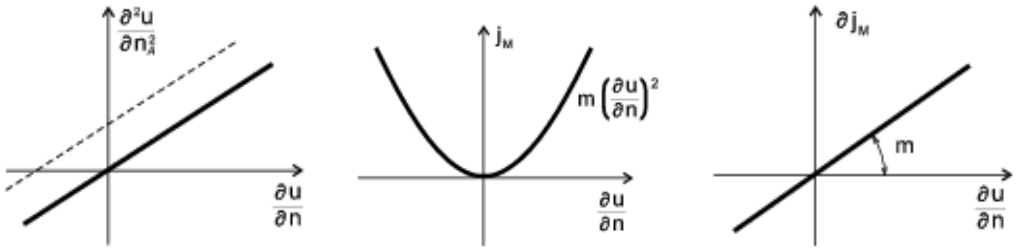
$$\tilde{\mathcal{F}}(v) = \mathcal{F}(v) + \sum_{x=0}^{L} (\hat{T}(x)v(x) - \hat{M}(x)Dv(x)).$$

Prostor virtuálních posunutí má nyní tvar $\mathcal{V} = H^2(\Omega)$ a pomocí zobecněné Friedrichsovy nerovnosti lze ukázat, že za jistých předpokladů (např. že $t(x) \geq t_o > 0$, pro $x = 0, L$) je nová forma $\tilde{a}$ na $\mathcal{V}$ koercivní. Tedy následující úloha má jediné řešení

$$\exists!\ u \in \mathcal{V} \ : \ \tilde{a}(u, v) = \tilde{\mathcal{F}}(v) \quad \forall v \in \mathcal{V}\ .$$



Obr. 4.1: Znázornění okrajových podmínek Newtonova typu pro průhyb



Obr. 4.2: Znázornění okrajových podmínek Newtonova typu pro natočení

Snadno lze vidět, že předchozí okrajové podmínky lze vhodným způsobem jednak dále kombinovat: např. dané posunutí s pružným natočením nebo pružnou posuvnou podporu s daným natočením, a jednak „smíchat“, tj. zadat různé typy podmínek na různých koncích, atp., a získat tak další typy okrajových úloh. Jejich řešitelnost však lze dokázat postupy uvedenými u předchozích typů, proto se jimi nebudeme dále podrobně zabývat.

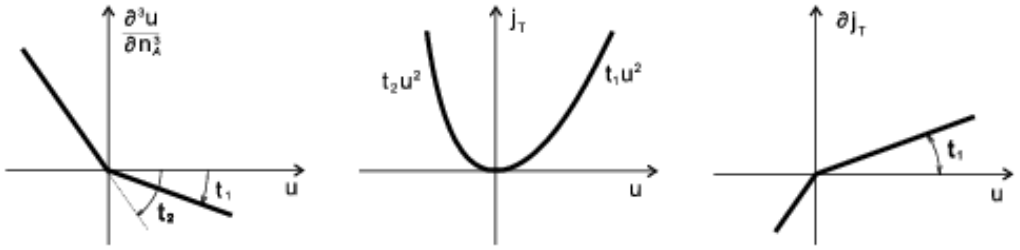
3.3 Neklasické okrajové podmínky: zobecnění

V této části uvedeme několik vzorových zobecnění Newtonových okrajových podmínek. Jde jednak o dva speciální typy okrajových podmínek (podrobnosti lze nalézt např. v [7] a také v [8]), jež dostaneme přirozeným zobecněním okrajových podmínek předchozího, Newtonova typu, a to manipulací s koeficienty reprezentující tuhosti, a jeden typ podstatně komplikovanějšího zobecnění získaný zásadní změnou předpisu pro chování podpory. V posledním případě vede totiž zobecnění okrajových podmínek na úlohu minimalizace funkcionálu obsahujícího superpotenciál ve tvaru nemonotonního nekonvexního a nediferencovatelného funkcionálu (podrobnosti lze nalézt např. v [6]).

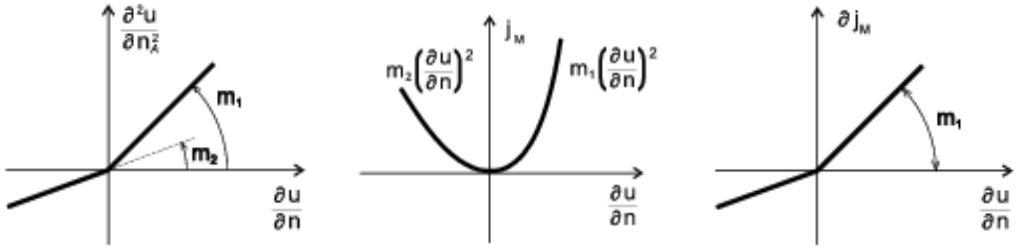
3.3.1 *Zobecněné Newtonovy okrajové podmínky*

První typ přirozeného zobecnění dostaneme, když při formulaci podmínek budeme navíc rozlišovat jednak orientaci výsledné deformace, tj. znaménko posunutí či natočení, a jednak předepíšeme různé hodnoty pro odpovídající tuhosti (odlišíme např. tahové a tlakové namáhaní pružin), tj. zadáme $t_i \geq 0$, $m_i \geq 0$,

$i = 1, 2$ a dostaneme zobecněné bilaterální Newtonovy okrajové podmínky. Podrobnou analýzu této situace z důvodů stručnosti zde vynecháme, uvádíme pouze příslušná schemata těchto zobecněných okrajových podmínek na obrázcích 4.3. pro posunutí a 4.4. pro natočení.



Obr. 4.3: Zobecnění okrajových podmínek Newtonova typu pro posunutí



Obr. 4.4: Zobecnění okrajových podmínek Newtonova typu pro natočení

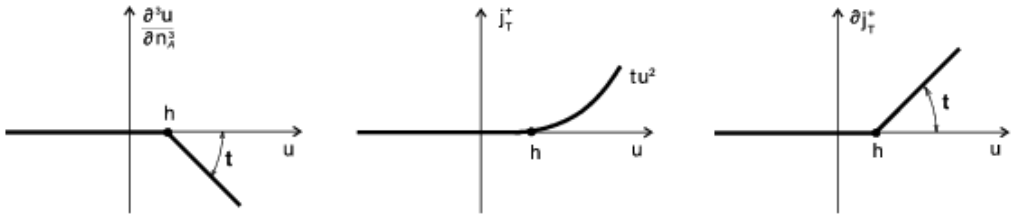
Potenciály odpovídající hraničním bilineárním formám značíme j_T (odpovídá hraniční formě $b_T(u, v) = \sum_{x=0}^{L}(t_i(x)u(x)v(x)$, zřejmě platí $\partial j_T(u, v) = b_T(u, v)$) a j_M (odpovídá hraniční formě $b_M(u, v) = \sum_{x=0}^{L}(t_i(x)Du(x)Dv(x)$, přičemž opět platí $\partial j_M(u, v) = b_M(u, v)$.

3.3.2 *Jednostranné okrajové podmínky Newtonova typu*

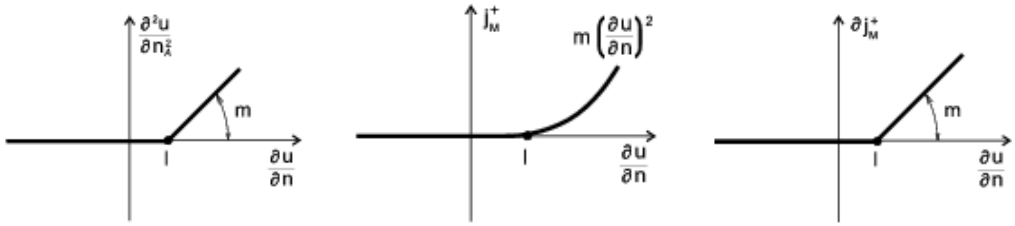
Dalším možným typem okrajových podmínek, které lze získat přímým zobecněním podmínek Newtonova typu jsou následující podmínky: v standardních předpisech pro Newtonovy podmínky stačí vázat hodnoty (včetně hodnot její derivace) hledané funkce na hranici příslušnými předpisy pouze od jisté hodnoty (např. od nuly) a pouze v jednom směru (tj. jen na jedné straně), druhým směrem ponecháme její hodnoty bez omezení (volný pohyb konce), a dostaneme tak zobecněné jednostranné Newtonovy okrajové podmínky.

Mechanický význam takových podmínek je zcela zřejmý a představuje jednostrannou posuvnou nebo natáčivou pružinu. Matematická formulace odpovídající úlohy vede na nelineární variační rovnici. Její tvar je analogický standardní Newtonově úloze, kde však v hraniční formě nyní uvažujeme pouze jednostranné hodnoty posunutí nebo natočení, tj. $u(x)^+$ a $Du(x)^+$, $x = 0, L$. Pokud jsou zadány na hranici pouze podmínky tohoto typu a nejsou kombinovány např. s Dirichletovou podmínkou, je výsledná úloha pouze semikoercivní a není možné zadávat data

úlohy zcela libovolně, ale je nutné navíc formulovat podmínky řešitelnosti svazující (vhodným způsobem zajišťujícím existenci či jednoznačnost řešení) zadávaná data úlohy. Podrobnosti i přesnou formulaci úlohy opět z důvodů stručnosti vynecháme, uvádíme pouze schemata odpovídajících podmínek a odpovídajících potenciálů.



Obr. 5.1: Znázornění okrajových podmínek pro jednostranné pružné uložení



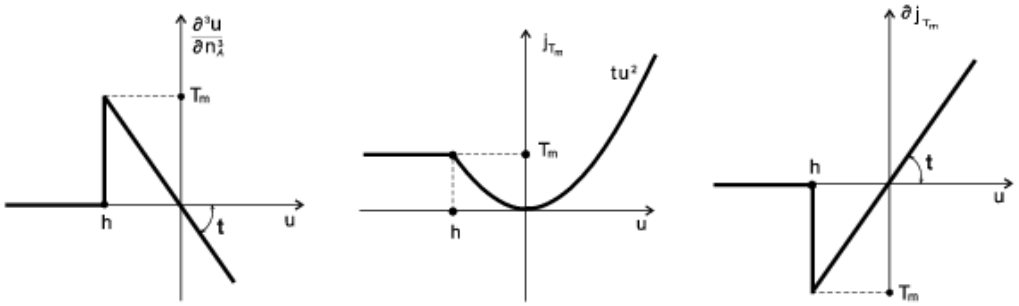
Obr. 5.2: Znázornění okrajových podmínek pro jednostranné pružné natočení

Z uvedených obrázků je snadno vidět, že jednostranné okrajové podmínky Newtonova typu lze získat z bilaterálních podmínek Newtonova typu limitním přechodem jedné z tuhostí t_i, m_i, $i = 1, 2$ k nule. Např. pro vyjasnění souvislostí mezi schematy na obrázcích 4.3 a 4.4 a na obrázcích 5.1 a 5.2 (pro homogenní případ $h = 0, l = 0$) stačí realizovat limitní přechody $t_2, m_2 \to 0$.

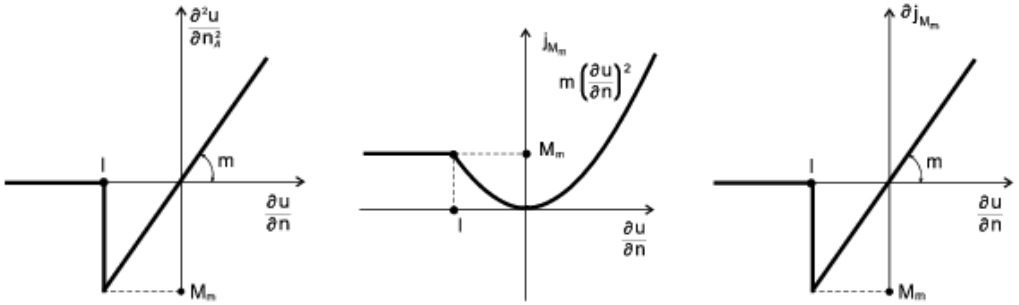
3.3.3 *Newtonovy okrajové podmínky s omezením*

Posledním, zde ilustrovaným typem zobecnění Newtonových okrajových podmínek jsou podmínky, které reprezentují pružnou posuvnou nebo pružnou natáčivou podporu, ale tentokráte jen s částečnou možností (jednostranným omezením) přenosu tahových nebo ohybových namáhání. Tato, pro technickou praxi velmi realistická okrajová podmínka má však za následek komplikace při matematické formulaci úlohy, analýze její řešitelnosti, a především při numerickém řešení diskretizované úlohy. V tomto případě je pro analýzu výsledné úlohy ve spojitém případě nutné použít aparát hemivariačních nerovnic (viz např. [6]), protože člen odpovídající této podmínce ve funkcionálu celkové potenciální energie je nediferencovatelný a nekonvexní.

Schema této okrajové podmínky a tvar příslušného superpotenciálu jsou na následujících obrázcích, další podrobnosti vynecháme, a zájemce odkazujeme na knihu [6].



Obr. 6.1: Okrajové podmínky pro pružné uložení s omezením



Obr. 6.2: Okrajové podmínky pro pružné natočení s omezením

3.4 Neklasické okrajové podmínky: Nerovnice

V této části uvedeme dva základní typy okrajových podmínek vedoucích na úlohy ve tvaru variačních nerovnic 1. a 2. druhu, případně jejich kombinací či dokonce zobecnění, pro další podrobnosti odkazujeme například na [3], [4], nebo [6].

3.4.1 *Jednostranné okrajové podmínky Signoriniho typu*

Opět jde o neklasické okrajové podmínky, kdy tentokráte ale nepředepisujeme hodnoty neznámé funkce či jejích derivací, ani jejich vzájemnou kombinaci, nýbrž jen jejich jednotlivá omezení a vzájemné sdružení (určená v podstatě aplikací Greenovy věty) na hranici oblasti. V jistém smyslu jde o zobecnění jednostranné Newtonovy okrajové podmínky (viz také obrázky 5.1 a 5.2) pro neomezený růst velikostí odpovídajích tuhostí posuvných a natáčivých pružin, tj. pro případ, kdy $t, m \to \infty$ (viz také schema na obrázku 7.). Potom pro jednostranné „tuhé“ posuvné i natáčivé uložení konců nosníku v $x = 0, L$ a v homogenním případě platí

$$T(x) \le 0,\ u(x) \le 0,\ T(x)u(x) = 0,$$

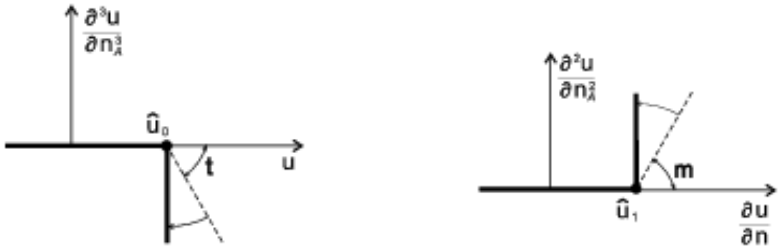
$$M(x) \ge 0,\ Du(x) \le 0,\ M(x)Du(x) = 0,\ x = 0, L.$$

Volba prostoru virtuálních posunutí závisí na volbě předepsaných omezení, tj. na stanovení, který předpis a kde se má realizovat ($x = 0, L$), a v obecném případě má tvar $\mathcal{V} = H^2(\Omega)$. Variační fomulace úlohy na konvexní množině kinematicky

přípustných průhybů $\mathcal{K}$, kde $\mathcal{K} = \{v \in \mathcal{V} \mid v(x) \leq 0, Dv(x) \leq 0, x = 0, L\}$ má tvar eliptické variační nerovnice 1. druhu, to je

$$u \in \mathcal{K}: \quad a(u, v-u) \geq \langle f, v-u \rangle \quad \text{pro } \forall v \in \mathcal{K}.$$

Pro existenci a jednoznačnost variačního řešení je i v tomto případě třeba zformulovat podmínky řešitelnosti (viz např. [3], [7] či [8]): pro zadání Signoriniho podmínek na celé hranici má množina kinematicky přípustných tuhých posunutí tvar $\mathcal{K} \cap \mathcal{R} = \mathbf{P}_1^-$, kde $\mathbf{P}_1^- = \{p \in \mathbf{P}_1 \mid p(x) \leq 0,\ x = 0, L\}$.



Obr. 7: Znázornění okrajových podmínek pro jednostranné tuhé uložení

3.4.2 *Okrajové podmínky pro dané „tuhé“ tření*

Pro zadání okrajových podmínek reprezentujících alespoň jednoduchým způsobem vliv tření budeme v této práci rozlišovat jednak tzv. „tuhé“ a jednak „pružné“ dané tření, a to podle způsobu odezvy posunutí (natočení) na velikost odpovídající reakce.

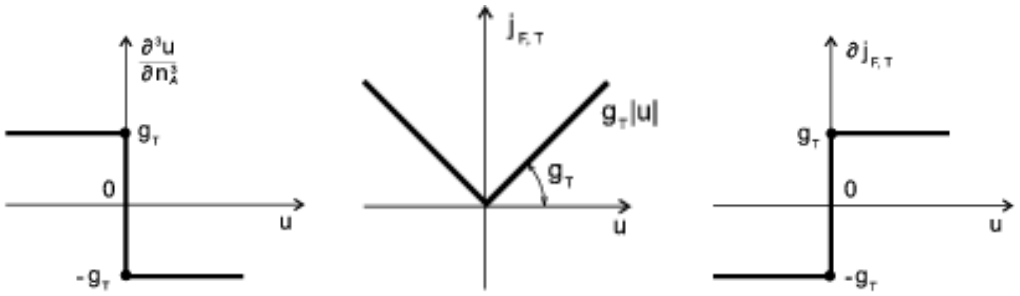
V případě daného tuhého tření omezujeme průběh posunutí či natočení velikostí reakcí a momentů pomocí následujících předpisů

$$\begin{gathered} |T(0)| \leq g_{T,0}, \quad |T(L)| \leq g_{T,L}, \quad |M(0)| \leq g_{M,0}, \quad |M(L)| \leq g_{M,L}, \\ g_{M,0}|Du(0)| - M(0)Du(0) = 0, \quad g_{M,L}|Du(L)| - M(L)Du(L) = 0, \\ g_{T,0}|u(0)| + T(0)u(0) = 0, \quad g_{T,L}|u(L)| + T(L)u(L) = 0 \end{gathered}$$

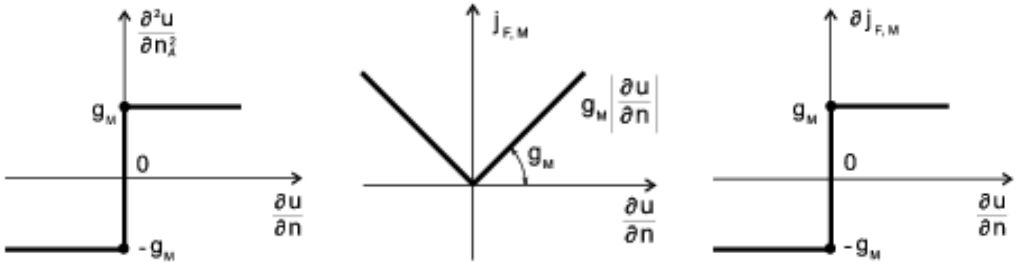
a formulace úlohy (variační nerovnice 2. druhu) s daným třením v podporách má tvar

$$a(u, v-u) + j(v) - j(u) \geq \langle f, v-u \rangle \quad \text{pro } \forall v \in \mathcal{V},$$

když nyní předpokládáme, že funkcionály $j_{1,x}^+ : H^2(\Omega) \to \mathbf{R}^1$, $i = 1, 2$, jsou pro $x = 0, L$ identicky nulové a $\mathcal{V} = H^2(\Omega)$. Připomínáme, že k zajištění řešitelnosti výše formulované úlohy musíme splnit opět podmínky bilancující (prostřednictvím zatížení) hodnoty daného tření s reakcemi v podporách, když nyní má odpovídající faktorprostor tvar $H^2(\Omega)/\mathcal{R}$ a zřejmě platí $\mathcal{V} \cap \mathcal{R} = \mathcal{R}$, kde $\mathcal{R} = \mathbf{P}_1$.



Obr. 8.1: Okrajové podmínky pro posunutí s daným „tuhým“ a třením

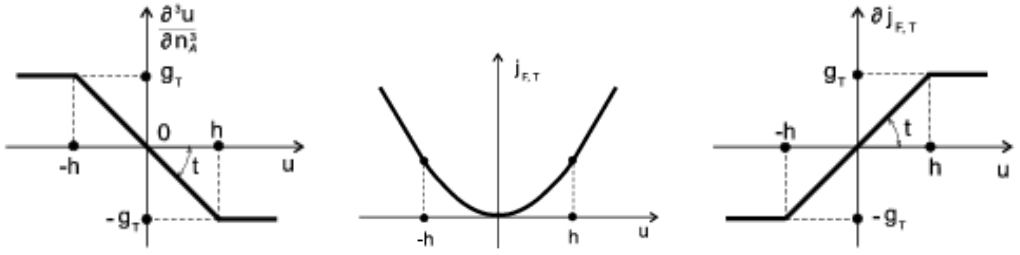


Obr. 8.2: Okrajové podmínky pro natočení s daným „tuhým“ třením

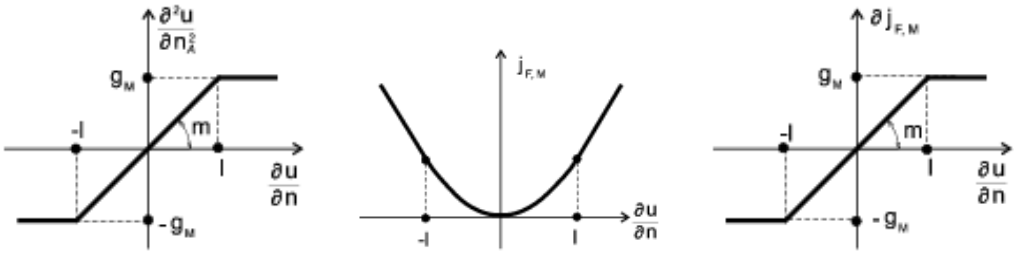
3.4.3 *Okrajové podmínky pro dané „pružné“ tření*

V tomto případě jde o typ okrajových podmínek kombinující podmínky pro dané tření s podmínkami Newtonova typu. Z pohledu mechaniky podmínky představují chování pružných podpor s předepsanými omezeními velikostí odpovídajících reakcí (analogie vetknutí s plastickým modelem).

Poznamenejme, že tuto úlohu je vhodné formulovat jako minimalizační úlohu, když práce tohoto modelu tření bude reprezentována potenciálem jehož charakter je z důvodů stručnosti schematicky znázorněn na obrázku 9.1 (pro posuvné tření) a 9.2 (pro natáčivé tření). Řešitelnost úlohy je v obecném případě třeba zajistit dalšími bilančními podmínkami.



Obr. 9.1: Okrajové podmínky pro posunutí s daným „pružným“ třením



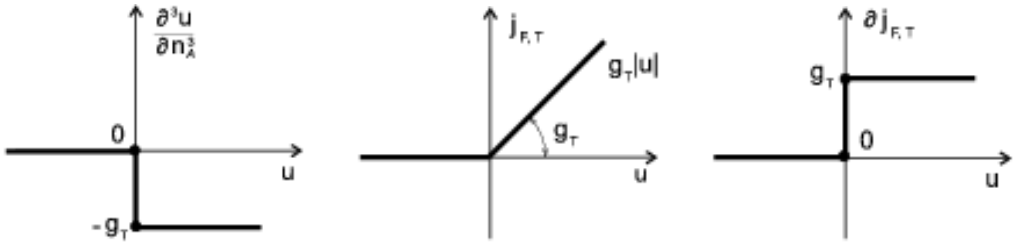
Obr. 9.2: Okrajové podmínky pro natočení s daným „pružným“ třením

3.5 Kombinované neklasické okrajové podmínky

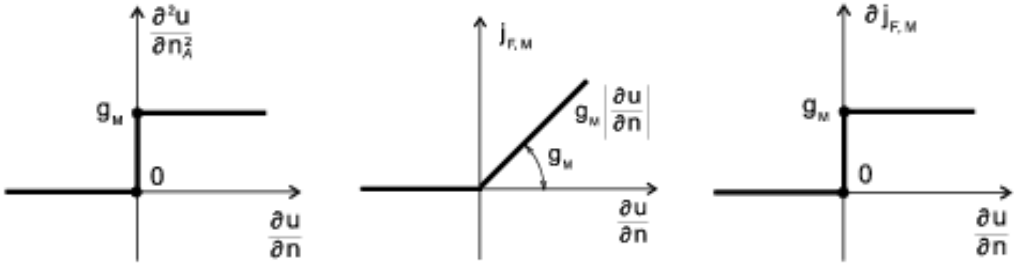
V této části uvedeme stručně (podrobnosti lze nalézt např. v [7] až [9]) charakter dalších možných typů okrajových podmínek, jež dostaneme kombinací předchozích typů.

3.5.1 *Jednostranné okrajové podmínky s daným „tuhým“ třením*

Je zřejmé, že doposud uvedené základní typy okrajových podmínek lze nejrůznějším způsobem kombinovat. Jedním z nejjednodušších způsobů zobecnění je následující postup kombinující bilaterální modely s jednostranným omezením. Pokud např. uvažujeme pouze jednostrannou podporu s modelem daného tuhého tření, je výsledný příspěvek takovéto okrajové podmínky k bilanci funkcionálu celkové potenciální energii možno charakterizovat potenciálem schematicky znázorněným na následujících obrázcích (pro posuvné i natáčivé podpory).



Obr. 10.1: Znázornění jednostranných podmínek pro posunutí se třením

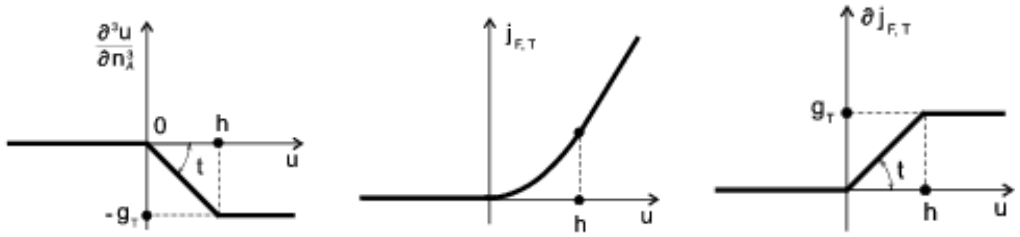


Obr. 10.2: Znázornění jednostranných podmínek pro natočení se třením

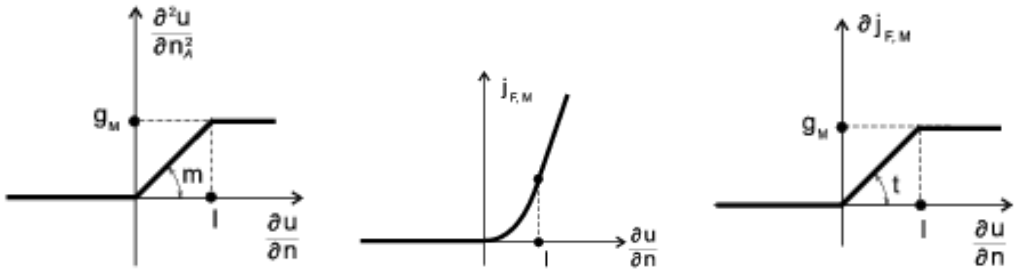
Další postup při matematické formulaci úlohy i analýze její řešitelnosti je analogický s předchozími, a nebudeme jej zde uvádět.

3.5.2 *Jednostranné okrajové podmínky s daným „pružným“ třením*

Analogicky jako v předchozím odstavci můžeme získat další model okrajové podmínky kombinací bilaterálního modelu pro pružné tření s jednostranným omezením. Z následujících obrázků jsou zřejmé jak ideje tohoto postupu tak i tvar příslušného potenciálu reprezentujícího práci tření. Všechny podrobnosti opět vynecháme.



Obr. 11.1: Znázornění jednostranných podmínek pro posunutí se třením



Obr. 11.2: Znázornění jednostranných podmínek pro natočení se třením

3.5.3 *Další kombinace: zobecnění*

Je zřejmé, že předložené typy a kombinace okrajových podmínek lze ještě dále kombinovat a zobecňovat. Např. na některé formulované okrajové podmínky lze přidat ještě další omezení, nebo naopak částečná uvolnění daných omezení. Zde jsme chtěli uvést, podle našeho názoru, typické podmínky ilustrující základní nebo významné odlišnosti ve vyšetřovaných modelových úlohách. Jednou z cest je například následující, pro potřeby praxe však velmi realistické zobecnění.

3.5.4 *Okrajové podmínky Signoriniho typu kombinované s Newtonovými a omezením*

Jestliže přidáme k podmínce reprezentující jednostrannou pružnou natáčivou nebo posuvnou podporu další omezení natočení či průhybu (např. z konstrukčních důvodů), dostaneme podmínku znázorněnou schematicky na obr. 12.

Obr. 12: Znázornění podmínek pro jednostrannou pružnou podporu s omezením

Tento typ omezení lze samozřejmě přidat prakticky ke všem zde uváděným okrajovým podmínkám. Takto získáme další zobecnění podmínek jež lépe vyjadřuje potřeby a realitu technické praxe.

Poděkování

Autor příspěvku považuje za svou milou povinnost poděkovat na tomto místě Grantové agentuře České republiky za finanční podporu jeho výzkumné činnosti. Předložená práce byla realizována v rámci grantu GA ČR 105/99/1651.

Reference

[1] Aubin, J. P.: *Approximation of Elliptic Boundary Value Problems.* Wiley-Interscience, London, 1972.

[2] Kufner, A., John, O., Fučík, S.: *Function Spaces.* Academia, Praha, 1977.

[3] Haslinger, J., a kol.: *Variační nerovnice v mechanice.* ALFA, Bratislava, 1979.

[4] Glowinski, R.: *Numerical Methods for Nonlinear Variationals Problems.* Springer Verlag, New York, 1984.

[5] Horák, J.: *On solvability of one special problem of coupled thermoelasticity, Part I.* Acta Universitatis Palackianae Olomucensis, Facultas Rerum Naturalium, Mathematica **34** (1995), 39–58.

[6] Haslinger, J., Miettinen, M., Panagiotopoulos, P. D.: *Finite Element Method for Hemivariational Inequalities. Theory, Methods, Applications.* Kluwer Academic Press, London, 1999.

[7] Horák, J. V.: *O okrajových podmínkách a řešitelnosti úlohy ohybu nosníku.* In: Sborník 8. semináře „Moderní matematické metody v inženýrství — 3μ“, Dolní Lomná, 9.–11. 6. 1999, ed. J. Doležalová, VŠB–TU Ostrava, 38–44.

[8] Horák, J. V.: *Poznámka k řešitelnosti semikoercivních případů úlohy ohybu nosníku na jednostranném podkladě a pro různé typy okrajových podmínek.* rukopis — nepublikováno, KMAaAM, PřF UP Olomouc, 1999.

[9] Horák, J. V.: *On One Class of Non-classical Boundary Conditions Splitting 1D Coupled Thermoelasticity Problems.* zasláno a přijato na IMSE 2000, Universita of Alberta, Kanada; Preprint, K MAaAM, PřF UP, Olomouc, 2000.

[10] Horák, J. V.: *Modelování ohybu kruhových desek s neklasickými okrajovými podmínkami.* rukopis — nepublikováno, KMAaAM, PřF UP, Olomouc, 2000