Iterative solution methods for mesh variational inequalities

A. Lapin, Kazan State University, Russia
(Oulu University, Finland, 2009)

The lecture course is devoted to the iterative solution methods for finite-dimensional variational inequalities, which arise when approximating differential variational inequalities of mechanics and physics by finite element or finite difference methods. Such finite-dimensional inequalities are united by the notation "mesh variational inequalities".

The iterative methods for two classes of finite-dimensional variational inequalities are studied: inequalities with positive definite matrices and inequalities with "saddle" matrices. By "saddle" we call the symmetric matrices with both positive and negative eigenvalues.

Existence theorems and convergence results for the iterative methods are cited without proofs, which can be found in [11]. Main attention is paid to

applications of the general results to the mesh approximations of the differential variational inequalities,

discussing the implementation details of the iterative algorithms.

The lecture is organised as follows.

The first section contains simple examples of the mesh variational inequalities, elements of the convex functions theory and the equivalent formulations of the variational inequalities.

In the second section basic iterative methods for the variational inequalities with positive definite matrices are considered. These methods are: one-step stationary methods, relaxation methods for the potential problems and splitting iterative methods. General convergence results are applied to the mesh variational inequalities with simple constraints, when all of the considered iterative methods can be easily implemented.

The iterative methods for the variational inequalities approximating the differential problems with constraints on the gradient of a solution are the topic of the third section. Lagrange multipliers approach is used to transform these inequalities to ones containing simple constraints and saddle matrices instead of positive definite matrices. Uzawa and Arrow-Hurwicz methods and their generalisations, as well as splitting methods, are analysed for this class of variational inequalities.

Contents

§1 Extremal problems, variational inequalities and inclusions with multivalued operators **3**
1.1 Variational inequalities with sets of constraints 3
1.1.1 General remarks 3
1.1.2 Examples of mesh variational inequalities. 4

1.2 Variational inequalities with non-differentiable functions. 9
1.2.1 General remarks 9
1.2.2 Examples of mesh variational inequalities. 10
1.3 Variational inequalities and inclusions with multivalued operators. 11
1.3.1 Convex functions and subdifferentials 11
1.3.2 Equivalent formulations of the variational inequalities. . . 13

§2 Iterative methods for variational inequalities with positive definite matrices 14
2.1 One-step stationary method 14
2.1.1 General convergence result 14
2.1.2 Applications to the mesh variational inequalities 16
2.1.3 Numerical example 19
2.2 Preconditioned one-step stationary method 22
2.2.1 General convergence result 22
2.2.2 Applications to the mesh variational inequalities 23
2.2.3 Numerical example 27
2.3 Relaxation methods 28
2.3.1 General convergence result. 28
2.3.2 Jacobi, Gauss-Seidel and SOR-methods 29
2.3.3 Applications to the mesh variational inequalities. Numerical examples 31
2.4 Error control and stopping criteria 37
2.4.1 Generalities 37
2.4.2 Numerical example 38
2.5 Splitting iterative methods 40
2.5.1 General convergence theory 40
2.5.2 Applications to mesh variational inequalities 41
2.5.3 Numerical example 43

§3 Variational inequalities with saddle matrices 44
3.1 Problem formulation, generalities 44
3.2 Stationary one-step iterative methods. 47
3.2.1 Uzawa-type method 47
3.2.2 Arrow-Hurwicz-type methods 48
3.2.3 Applications to the mesh variational inequalities 48
3.2.4 Numerical example 53
3.3 Douglas-Rachford splitting method 54
3.3.1 General convergence result 54
3.3.2 Application to a mesh variational inequality 55

§4 Appendix 59
4.1 Some notations and results from the theory of matrices and functional spaces 59

§1 Extremal problems, variational inequalities and inclusions with multivalued operators

1.1 Variational inequalities with sets of constraints

1.1.1 General remarks

Consider the minimisation problem

$$t^* = \arg\min_{t\in[0,1]} F(t) \tag{1.1}$$

with a differentiable in the points of $[0,1]$ function F. If t^* is a solution of (1.1), then t^* satisfies the variational inequality

$$F'(t^*)(t-t^*) \geqslant 0 \ \ \forall t \in [0,1]. \tag{1.2}$$

In fact, when t^* is a solution of (1.1), then one of the following variants is true:

$$\begin{array}{l} t^* \in (0,1) \Rightarrow F'(t^*) = 0; \\ t^* = 0 \Rightarrow F'(t^*) \geqslant 0; \\ t^* = 1 \Rightarrow F'(t^*) \leqslant 0. \end{array}$$

All these variants can be combined in variational inequality (1.2).

Let now K be a closed and convex set in $\mathbb{R}^n$, function $F : \mathbb{R}^n \to \mathbb{R}$ be differentiable in K and $\nabla F(x)$ be its gradient in the point x.

Consider minimisation problem

$$x^* = \arg\min_{x\in K} F(x) \tag{1.3}$$

and variational inequality

$$x^* \in K : \ (\nabla F(x^*), x - x^*) \geqslant 0 \ \ \forall x \in K. \tag{1.4}$$

Lemma 1.1. *If x^* is a solution of* (1.3)*, then x^* satisfies variational inequality* (1.4)*. In case of convex function F both these problems,* (1.3) *and* (1.4)*, are equivalent.*

Proof. If x^* is a solution of (1.3), then for a fixed $x \in K$ function $\varphi(t) = F(x^* + t(x - x^*))$ attains its minimum over $[0,1]$ at the point $t = 0$, and from the previous result it follows

$$\varphi'(0) = (\nabla F(x^*), x - x^*) \geqslant 0.$$

Let now F be a convex function, then for any $x \in K$ and any $\lambda \in (0,1)$

$$F(x^* + \lambda(x - x^*)) - F(x^*) \leqslant \lambda\,(F(x) - F(x^*)),$$

whence $F(x) - F(x^*) \geqslant \lambda^{-1}\,(F(x^* + \lambda(x - x^*) - F(x^*))$. Passing to the limit for $\lambda \to 0$ one get

$$F(x) - F(x^*) \geqslant (\nabla F(x^*), x - x^*) \ \ \forall x \in K. \tag{1.5}$$

From inequality (1.5) obviously follows that a solution of variational inequality (1.4) is a minimum of $F(x)$ over K. □

A partial case of (1.3) is the problem to minimise a quadratical function

$$F(x) = \frac{1}{2}(Ax, x) - (f, x),$$

where $A \in \mathbb{R}^{n\times n}$ is a symmetric and positive definite matrix ($A = A^T > 0$) and $f \in \mathbb{R}^n$ is a given vector. In this case $\nabla F(x) = Ax - f$, and the variational inequality, which is equivalent to the minimisation problem, becomes

$$(Ax^*, x - x^*) \geqslant (f, x - x^*) \ \ \forall x \in K. \tag{1.6}$$

It is called as variational inequality with a linear main operator A.

1.1.2 Examples of mesh variational inequalities.

Example 1.1. Obstacle problem. Finite difference approximation.

Let $K = \{u \in H_0^1(\Omega) : u(t) \geqslant 0 \text{ in } \Omega\}$ be a convex set in Sobolev space $H_0^1(\Omega)$ and $f \in L_2(\Omega)$ be a given function. Obstacle problem is the following variational inequality: find $u \in K$, such that

$$\int\limits_\Omega \nabla u \cdot \nabla(v - u)dt \geqslant \int\limits_\Omega f(t)(v - u)dt \quad \forall v \in K. \tag{1.7}$$

Variational inequality (1.7) has a unique solution u, which is at the same time the unique solution of the minimisation problem

$$u = \arg\min_{v\in K}\{J(v) = \frac{1}{2}\int\limits_\Omega |\nabla v|^2 dt - \int\limits_\Omega fvdt\}. \tag{1.8}$$

Moreover, if $f(t)$ is continuous, then solution $u(t)$ is quite regular (belongs to Sobolev space $H^2(\Omega)$) and (1.7) can be written in the pointwise form

$$\begin{aligned} &(-\Delta u - f)(t) \geqslant 0,\ u(t) \geqslant 0,\ u(t)\,(\Delta u + f)(t) = 0 \text{ for } t \in \Omega, \\ &u(t) = 0 \text{ for } t \in \partial\Omega. \end{aligned} \tag{1.9}$$

First, we approximate the one-dimensional obstacle problem with $\Omega = (0, 1)$ by a finite-difference scheme on a uniform grid.

Supposing f to be continuous, we can use any of three formulation of the problem, namely, variational inequality (1.7), minimising problem (1.8) or pointwise form (1.9), to construct a finite-difference scheme. Now we choose (1.8), i.e. approximation of the functional

$$J(u) = \frac{1}{2}\int\limits_0^1 (u')^2 dt - \int\limits_\Omega fudt$$

over the set $K = \{u(t) \geqslant 0 \text{ for } t \in (0,1),\ u(t) = 0 \text{ for } t = 0 \text{ and for } t = 1\}$.

Let

$$\bar{\omega} = \{t_i = i\,h,\ i = 0, 1, \ldots, n+1\},\ (n+1)\,h = 1,$$

be a uniform mesh on the segment $[0,1]$ with meshsize $h > 0$, $u_i = u(t_i)(u_0 = u_{n+1} = 0)$ and $f_i = f(t_i)$. Functional J is approximated by a convex and differentiable (quadratical) function

$$J_h(u) = \frac{h}{2}\left(\frac{u_1^2}{h^2} + \sum_{i=1}^{n-1}\left(\frac{u_{i+1}-u_i}{h}\right)^2 + \frac{u_n^2}{h^2}\right) - h\sum_{i=1}^{n} f_i u_i.$$

When constructing this approximation, we use the difference quotients for the approximation of the first derivative

$$u'(t_i) \approx \frac{u_{i+1}-u_i}{h},\ \ u'(t_i) \approx \frac{u_i - u_{i-1}}{h}$$

and the quadrature rules for the approximation of the integrals

$$\int_0^1 F(t)dt \approx h\sum_{i=0}^{n} F(t_i),\ \ \int_0^1 F(t)dt \approx h\sum_{i=1}^{n+1} F(t_i) \text{ for any continuous function } F(t).$$

The gradient of J_h is $\nabla J_h(u) = h\,(Au - f)$, where

$$A = h^{-2}\begin{pmatrix} 2 & -1 & 0 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & 0 & -1 & 2 \end{pmatrix}. \tag{1.10}$$

Denote by $K = \{u \in \mathbb{R}^n : u_i \geqslant 0\ \forall i\}$ a convex set and by $(.,.)$ usual euclidian scalar product in $\mathbb{R}^n$. Owing to Lemma 1.1 the problem of the minimisation of J_h over K is equivalent to variational inequality

$$(Au, v-u) \geqslant (f, v-u)\ \ \forall v \in K. \tag{1.11}$$

The eigenvalues of the matrix A are known: $\lambda_k = 4h^{-2}\sin^2\dfrac{k\pi h}{2}$, $k = 1, 2, \ldots, n$. In particular, minimal and maximal eigenvalues are

$$\lambda_1 = 4h^{-2}\sin^2\frac{\pi h}{2} = O(1),\ \lambda_n = 4h^{-2}\cos^2\frac{\pi h}{2} = O(h^{-2}) \text{ as } h \to 0,$$

so, condition number of the matrix A is

$$\mathrm{cond}_2\, A = \frac{\lambda_n}{\lambda_1} = O(h^{-2}).$$

This simple example contains all **basic features** of the mesh problems, in particular, mesh variational inequalities which approximate the differential

problems: **high dimension** n, **big condition number, sparse matrix** (small number of nonzero entries in all rows and columns of the matrix). $\square$

Now, consider the two-dimensional obstacle problem and approximate it by using pointwise formulation (1.9).

Let $\Omega = (0,1) \times (0,1)$ be the unit square with the boundary $\partial\Omega$ and $\overline{\Omega} = \Omega \cap \partial\Omega$, let further $f(t) = f(t_1, t_2)$ be a continuous in Ω function. Denote by $\bar{\omega} = \{t = (ih, jh) : 0 \leqslant i, j \leqslant p+1,\ (p+1)h = 1\}$ a uniform grid on $\overline{\Omega}$ with meshsize $h > 0$, by $\gamma = \bar{\omega} \cap \partial\Omega$ the set of its boundary nodes and by $\omega = \bar{\omega} \setminus \gamma$ the set of its internal nodes.

Below u_h is a mesh function — function in p^2-dimensional space, — which is uniquely defined by its nodal values $u_{ij} = u_h(ih, jh)$ for $(ih, jh) \in \bar{\omega}$ and which is equal to zero in the nodes $(ih, jh) \in \gamma$. Let also f_h be the mesh function with nodal values $f_h(t) = f(t)$ for $t \in \omega$.

To approximate the derivatives of a smooth function $u(t)$ in the internal nodes $(ih, jh) \in \omega$ the following difference quotients are used:

$$\frac{\partial u}{\partial t_1}(ih, jh) \approx \frac{u_{ij} - u_{i-1j}}{h} \equiv \bar{\partial}_1 u_h, \quad \frac{\partial u}{\partial t_1}(ih, jh) \approx \frac{u_{ij+1} - u_{ij}}{h} \equiv \partial_1 u_h,$$

$$\frac{\partial^2 u}{\partial t_1^2}(ih, jh) \approx \frac{2u_{ij} - u_{i-1j} - u_{i+1j}}{h^2} \equiv \bar{\partial}_1 \partial_1 u_h = \partial_1 \bar{\partial}_1 u_h.$$

Further $\Delta_h u_h = \bar{\partial}_1 \partial_1 u_h + \bar{\partial}_2 \partial_2 u_h$ is mesh Laplace operator defined in the nodes of ω.

Finite-difference approximation of (1.9) reads as

$$\begin{aligned} &-\Delta_h u_h - f_h \geqslant 0,\ u_h \geqslant 0,\ u_h(\Delta_h u_h + f_h) = 0 \text{ in } \omega, \\ &u_h = 0 \text{ on } \gamma. \end{aligned} \tag{1.12}$$

Problem (1.12) has dimension $n = p^2$, and its unknowns are u_{ij}, $i, j = 1, 2, \ldots, p$. To write this problem in a matrix-vector form we need to present the set $\{u_{ij}\}$ in the form of a vector from $\mathbb{R}^n$. And this is equivalent to the choice of a enumeration (ordering) for the set of the mesh nodes ω. A traditional enumeration is so-called lexicographical one: "from left to right and from down to top". When using this enumeration one obtains a vector $y \in \mathbb{R}^n$ by the following rule:

$$y_1 = u_{1,1}, y_2 = u_{2,1}, \ldots, y_p = u_{p,1}, y_{p+1} = u_{1,2}, y_{p+2} = u_{2,2}, \ldots, y_{p^2} = u_{p,p}.$$

With lexicographical ordering for the set of the mesh nodes, the following symmetric and block-tridiagonal matrix $A \in \mathbb{R}^{n \times n}$ corresponds to the mesh Laplace operator $-\Delta_h$:

$$A = h^{-2} \begin{pmatrix} D & -E & 0 & \ldots & 0 \\ -E & D & -E & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & D & -E \\ 0 & 0 & \ldots & -E & D \end{pmatrix}, \ D = \begin{pmatrix} 4 & -1 & 0 & \ldots & 0 \\ -1 & 4 & -1 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & 4 & -1 \\ 0 & 0 & \ldots & -1 & 4 \end{pmatrix}, \tag{1.13}$$

where $D \in \mathbb{R}^{p\times p}$, E is unit $p \times p$ matrix.

Let vector $f \in \mathbb{R}^n$ corresponds to mesh function f_h, then system (1.12) becomes

$$(Ay - f)_i \geqslant 0, \;\; y_i \geqslant 0, \;\; (Ay - f)_i \, y_i = 0 \quad \forall i.$$

This is so-called complementarity problem, which is equivalent to variational inequality

$$y \in K : \; (Ay, z - y) \geqslant (f, z - y) \; \forall z \in K, \; K = \{z \in \mathbb{R}^n : \; z_i \geqslant 0 \; \forall i\}, \quad (1.14)$$

with symmetric and positive definite matrix A and closed convex set K. Also, because of the symmetry of A, variational inequality (1.14) is equivalent to the problem of minimisation of the function $\dfrac{1}{2}(Ay, y) - (f, y)$ over the set K.□

Example 1.2. Obstacle problem with diffusion-convection operator.

Consider variational inequality

$$\int_\Omega \nabla u \cdot \nabla (v - u) dt + \int_\Omega \bar{a} \cdot \nabla u (v - u) dt \geqslant \int_\Omega f(v - u) dt \quad \forall v \in K \qquad (1.15)$$

with a given constant vector $\bar{a}$ and set of constraints

$$K = \{u \in H_0^1(\Omega) : u(t) \geqslant 0 \text{ in } \Omega\}.$$

Variational inequality (1.15) has a unique solution $u(t)$ and if it is smooth enough, then (1.15) can be written in the pointwise form

$$-\Delta u + \bar{a} \cdot \nabla u - f \geqslant 0, \; u \geqslant 0, \; u \, (-\Delta u + \bar{a} \cdot \nabla u - f) = 0 \text{ in } \Omega \qquad (1.16)$$

with Dirichlet boundary condition $u = 0$ on $\partial\Omega$.

Let $\Omega = (0, 1) \times (0, 1)$ be the unit square. We approximate problem (1.16) by a finite-difference scheme on a uniform grid keeping the notations for the mesh sets difference quotients from Example 1.1. For definiteness we suppose the coordinates a_1 and a_2 of the vector $\bar{a}$ to be positive. Then finite-difference approximation of (1.16) reads as follows:

$$\begin{cases} -\Delta_h u_h + \bar{a} \cdot \overline{\nabla} u_h - f_h \geqslant 0, \; u_h \geqslant 0, \; u_h(-\Delta_h u_h + \bar{a} \cdot \overline{\nabla} u_h - f_h) = 0 \;\; \text{in } \omega, \\ u_h = 0 \text{ on } \gamma. \end{cases} \qquad (1.17)$$

Here $\bar{a} \cdot \overline{\nabla}_h u_h = a_1 \bar{\partial}_1 u_h + a_2 \bar{\partial}_2 u_h$ is the up-wind approximation of the convective term.

Let a vector $u \in \mathbb{R}^n$, $n = p^2$, corresponds to a mesh function u_h for lexicographical enumeration of the mesh nodes. Then u satisfies a variational inequality

$$u \in K : \;\; (\tilde{A} u, v - u) \geqslant (f, v - u) \;\; \forall v \in K, \; K = \{u \in \mathbb{R}^n : \; u_i \geqslant 0 \; \forall i\}$$

with a matrix $\tilde{A}$, which corresponds to mesh operator $-\Delta_h + \bar{a}\cdot\overline{\nabla}_h$. This matrix equals to the sum of the matrix A from (1.13) and the block-twodiagonal matrix

$$L = h^{-1}\begin{pmatrix} L_1 & 0 & \dots & 0 & 0 \\ -a_2E & L_1 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & L_1 & 0 \\ 0 & 0 & \dots & -a_2E & L_1 \end{pmatrix} \in \mathbb{R}^{n\times n},$$

where

$$L_1 = \begin{pmatrix} a_1+a_2 & 0 & \dots & 0 & 0 \\ -a_1 & a_1+a_2 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & a_1+a_2 & 0 \\ 0 & 0 & \dots & -a_1 & a_1+a_2 \end{pmatrix} \in \mathbb{R}^{p\times p}.$$

Below, in Example 2.2, we will prove that matrix $\tilde{A}$ is positive definite, so, once again we deal with a variational inequality with a positive definite matrix. □

Example 1.3. Obstacle problem. Finite element approximation.

Let $\Omega \in \mathbb{R}^2$ be a polygon and let $T_h = \{\delta_i\}_i$ be its conforming triangulation, i.e. decomposition into non-overlapping triangles, which can have only common sides or common vertices. Denote by h_i the diameter of a triangle δ_i, and by $h = \max\limits_i h_i$. Define the space of the linear functions $P_1 = \{p(t) = c_0 + c_1t_1 + c_2t_2,\ c_i \in \mathbb{R}\}$, the spaces of the mesh functions

$$V_h = \{u_h \in C(\overline{\Omega}) : u_h \in P_1\ \ \forall \delta \in T_h\}, \quad V_h^0 = \{u_h \in V_h : u_h(t) = 0\, \forall t \in \partial\Omega\}$$

and the set

$$K_h = \{u_h \in V_h^0 : u_h(t) \geqslant 0\ \ \forall t \in \Omega\}.$$

Approximation by finite element method of obstacle problem (1.7) is the following finite-dimensional variational inequality:

$$u_h \in K_h : \int_\Omega \nabla u_h \cdot \nabla(v_h - u_h)dt \geqslant \int_\Omega f(t)(v_h - u_h)dt\ \ \forall v_h \in K_h. \qquad (1.18)$$

Let $\omega_h = \{t_i\}_{i=1}^n$ be the set of the vertices in Ω of the triangles $\delta \in T_h$, $n = \operatorname{card}\omega_h$. Put in the correspondence to a function $v_h \in V_h^0$ the vector $v \in \mathbb{R}^n$ with the coordinates $v_i = v_h(t_i)$, $t_i \in \omega_h$, (using an enumeration of t_i). Further we will use the notation $v \Leftrightarrow v_h$ for this correspondence. Let $K = \{u \in R^n : u_i \geqslant 0\ \forall i\}$. Owing to the choice of the piecewise-linear functions in the construction of the space V_h, the constraints $u_h \in K_h$ are equivalent to the constraints $u \in K$ for the nodal values of u_h, i. e. $K_h \ni u_h \Leftrightarrow u \in K$.

A matrix A (called as stiffness matrix in finite element method) and a vector f (called as a load vector) are defined by

$$(Au, v) = \int\limits_{\Omega} \nabla u_h(t) \cdot \nabla v_h(t) dt, \;\; (f, v) = \int_{\Omega} f(t) v_h(t) dt, \;\; u \Leftrightarrow u_h, \; v \Leftrightarrow v_h.$$

Now variational inequality (1.18) can be written in the form

$$u \in K : \; (Au, v - u) \geqslant (f, v - u) \; \forall v \in K$$

with positive definite matrix A and closed convex set K. $\square$

1.2 Variational inequalities with non-differentiable functions.

1.2.1 General remarks

Let $F : \mathbb{R}^n \to \mathbb{R}$ be a differentiable function and $\varphi : \mathbb{R}^n \to \mathbb{R}$ be a convex function. Consider minimisation problem

$$x^* = \arg \min_{x \in \mathbb{R}^n} (F(x) + \varphi(x)) \tag{1.19}$$

and variational inequality

$$x^* \in \mathbb{R}^n : \; (\nabla F(x^*), x - x^*) + \varphi(x) - \varphi(x^*) \geqslant 0 \;\; \forall x \in \mathbb{R}^n. \tag{1.20}$$

Lemma 1.2. *If x^* is a solution of* (1.19)*, then x^* satisfies variational inequality* (1.20).

In case of convex function F problems (1.19) *and* (1.20) *are equivalent.*

Proof. If x^* is a solution of (1.19), then for any fixed $x \in \mathbb{R}^n$ and any $\lambda \in (0, 1)$

$$\begin{aligned} 0 &\leqslant (F(x^* + \lambda(x - x^*)) + \varphi(x^* + \lambda(x - x^*))) - (F(x^*) + \varphi(x^*)) \\ &\leqslant (F(x^* + \lambda(x - x^*)) - (F(x^*)) + \lambda(\varphi(x) - \varphi(x^*)). \end{aligned}$$

Dividing both parts of this inequality by λ and passing to the limit for $\lambda \to +0$, one immediately obtains variational inequality (1.20).

In case of a convex function F inequality (1.5) gives

$$F(x) - F(x^*) + \varphi(x) - \varphi(x^*) \geqslant (\nabla F(x^*), x - x^*) + \varphi(x) - \varphi(x^*) \; \forall x \in \mathbb{R}^n,$$

whence the second statement of the lemma. $\square$

In partial case

$$F(x) = \frac{1}{2}(Ax, x) - (f, x) + \varphi(x), \;\; A = A^T > 0,$$

with non-differentiable function φ minimisation problem is equivalent to variational inequality

$$(Ax^*, x - x^*) + \varphi(x) - \varphi(x^*) \geqslant (f, x - x^*) \;\; \forall x \in \mathbb{R}^n \tag{1.21}$$

with linear main operator A.

1.2.2 Examples of mesh variational inequalities.

Example 1.4. Consider a problem of the minimisation

$$J(u) = 1/2 \int_0^1 u'^2(t)dt - \int_0^1 f(t)\,u(t)dt + \int_0^1 |u(t)|dt. \tag{1.22}$$

over the space $H_0^1(0,1)$ and approximate it by a finite-difference scheme on a uniform grid.

Let $\bar{\omega} = \{t_i = i\,h,\ i = 0, \ldots, n+1;\ (n+1)\,h = 1\}$, $u_i = u(t_i)(u_0 = u_{n+1} = 0)$ and $f_i = f(t_i)$. Functional (1.22) is approximated by a convex function

$$J_h(u) = \frac{h}{2}\left(\frac{u_1^2}{h^2} + \sum_{i=1}^{n-1}\left(\frac{u_{i+1}-u_i}{h}\right)^2 + \frac{u_n^2}{h^2}\right) - h\sum_{i=1}^{n} f_i u_i + h\sum_{i=1}^{n}|u_i|.$$

Due to Lemma 1.2 the minimisation of this function over the vector space $\mathbb{R}^n$ is equivalent to solving variational inequality

$$u \in \mathbb{R}^n:\ (Au, v-u) + \varphi(v) - \varphi(u) \geqslant (f, v-u)\ \ \forall v \in \mathbb{R}^n$$

with positive definite and symmetric matrix A defined in (1.10), and with convex and continuous function $\varphi(u) = h\sum_{i=1}^{n}|u_i|$.□

Example 1.5. Model problem of the contact with friction.

Let Ω be a polygon in $\mathbb{R}^2$. The problem under consideration is to find a function $u \in H_0^1(\Omega)$, such that

$$\int_\Omega \nabla u \cdot \nabla(v-u)dt + \int_\Omega (|v| - |u|)dt \geqslant \int_\Omega f(t)(v-u)dt \quad \forall v \in H_0^1(\Omega). \tag{1.23}$$

We approximate variational inequality (1.23) by a finite element method. Let the triangulation T_h and the space V_h^0 be as in Example 1.3. To approximate non-differentiable functional the following quadrature formulae is used:

$$\int_\Omega |u_h(t)|dt \approx S_h(|u_h|) = \sum_{\delta \in T_h} S_\delta(|u_h|),\ S_\delta(|u_h|) = \frac{\mathrm{mes}\,\delta}{3}\sum_{i=1}^{3}|u_h(a_i)|,$$

where $\{a_i\}_{i=1}^3$ are the vertices of a finite element (triangle) $\delta \in T_h$.

Finite element scheme, approximating problem (1.23), is the following mesh variational inequality:

$$\begin{aligned} &\text{find } u_h \in V_h^0 \text{ such that for all } v_h \in V_h^0 \\ &\int_\Omega \nabla u_h \cdot \nabla(v_h - u_h)dt + S_h(|v_h|) - S_h(|u_h|) \geqslant \int_\Omega f(t)(v_h - u_h)dt. \end{aligned} \tag{1.24}$$

Let $\omega_h = \{t_i\}_{i=1}^n$ be the set of the vertices in Ω of the triangles $\delta \in T_h$, $n = \operatorname{card}\omega_h$. Define stiffness matrix A, load vector f and the function φ by the equalities

$$(Au, v) = \int\limits_{\Omega} \nabla u_h(t) \cdot \nabla v_h(t) dt,\ (f, v) = \int_{\Omega} f(t) v_h(t) dt,\ \varphi(u) = S_h(|u_h|)$$

for $u \Leftrightarrow u_h,\ v \Leftrightarrow v_h$. It is easy to see, that

$$\varphi(u) = \sum_{i=1}^{n} \alpha_i |u_i|\ \alpha_i > 0.$$

Now, mesh variational inequality (1.24) can be written in the form

$$u \in \mathbb{R}^n : \quad (Au, y-u) + \varphi(y) - \varphi(u) \geqslant (f, y-u) \quad \forall y \in \mathbb{R}^n$$

with positive definite and symmetric matrix A, and with convex and continuous function φ. $\square$

1.3 Variational inequalities and inclusions with multivalued operators.

1.3.1 Convex functions and subdifferentials

Function $F : \mathbb{R}^n \to \overline{\mathbb{R}}$ is called:

— convex, if $F(tx + (1-t)y) \leqslant tF(x) + (1-t)F(y)\ \ \forall x, y \in \mathbb{R}^n,\ \ \forall t \in (0,1)$;

— strictly convex, if $F(tx+(1-t)y) < tF(x)+(1-t)F(y)\ \forall x \neq y \in \mathbb{R}^n,\ \forall t \in (0,1)$;

— proper, if $F(x) > -\infty\ \forall x \in \mathbb{R}^n$ and its effective set is nonempty: $\operatorname{dom} F = \{x \in \mathbb{R}^n : F(x) < +\infty\} \neq \emptyset$;

— lower semicontinuous, if $x^k \to x \Rightarrow \liminf F(x^k) \geqslant F(x)$.

Let function $F : \mathbb{R}^n \to \overline{\mathbb{R}}$ be convex, proper and lower semicontinue. Then a vector $\mu \in \mathbb{R}^n$ is called **subgradient** of the function F at a point x, if

$$F(y) - F(x) \geqslant (\mu, y - x)\ \ \forall y \in \mathbb{R}^n.$$

The set of all subdgradients of F at a point x forms the subdifferential $\partial F(x)$ of F at a point x. Multivalued operator ∂F has a domain of definition $D(\partial F) \subseteq \operatorname{dom} F$ and a set of values in $\mathbb{R}^n$. To underline that the values of ∂F in general case are the sets in $\mathbb{R}^n$, they write $\partial F : \mathbb{R}^n \to 2^{\mathbb{R}^n}$.

Properties of subdifferentials

1) Subdifferential $\partial F(x)$ is a convex and closed set (possibly empty). If F is differentiable at a point $x \in \mathbb{R}^n$, then its subdifferential $\partial F(x)$ consists of only element – gradient of F: $\partial F(x) = \{\nabla F(x)\}$.

2) Operator ∂F is monotone:

$$(\mu^1 - \mu^2, x^1 - x^2) \geqslant 0 \ \ \forall x^1, x^2 \in D(\partial F), \ \forall \mu^i \in \partial F(x^i). \tag{1.25}$$

3) Let $F : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a convex, proper and lower semicontinuous function and B be $n \times m$ matrix. Then

$$\partial F(Bu) = (B^T \circ \partial F \circ B)(u),$$

Examples of subdifferentials

1) If $F(x) = x^+ = \max\{0, x\}$, then its subdifferential is so-called Heaviside function

$$H(x) = \begin{cases} 0 \text{ if } x < 0; \\ [0,1] \text{ f } x = 0; \\ 1 \text{ if } x > 0. \end{cases}$$

Effective set dom F and domain of definition $D(\partial F) = D(H)$ equal to the whole space $\mathbb{R}$ in this example.

2) Indicator function of a convex and closed set K, defined by

$$I_K(x) = \begin{cases} 0, \ x \in K \\ +\infty, \ x \notin K, \end{cases}$$

is convex, proper and lower semicontinuous. Its subdifferential

$$\partial I_K(x) = \{\mu \in \mathbb{R}^n : (\mu, y - x) \leqslant 0 \ \forall y \in K\}, \ \ D(\partial I_K) = \operatorname{dom} I_K = K.$$

3) Function $F : \mathbb{R}^n \to \overline{\mathbb{R}}$ is called separable, if $F(x) = \sum_{i=1}^{n} F_i(x_i), \ F_i : \mathbb{R} \to \overline{\mathbb{R}}$. Separable function F is convex, proper and lower semicontinuous if and only if the same properties have all functions F_i. Effective domain $\operatorname{dom} F = \operatorname{dom} F_1 \times \operatorname{dom} F_2 \times \cdots \times \operatorname{dom} F_n$.

Subdifferential of a separable function F is a diagonal operator $\partial F = \operatorname{diag}(\partial F_1, \partial F_2, \ldots, \partial F_n)$. In particular, if $K = \prod_{i=1}^{n} [a_i, b_i], \ -\infty \leqslant a_i < b_i \leqslant +\infty$, then $I_K = \sum_{i=1}^{n} I_{[a_i, b_i]}$ is a separable function and $\partial I_K = \operatorname{diag}(\partial I_{[a_1,b_1]}, \ldots, \partial I_{[a_n,b_n]})$, where

$$\partial I_{[a_i,b_i]}(x) = \begin{cases} (-\infty, 0] \text{ for } x = a_i \ \ (\text{ if } a_i > -\infty), \\ 0 \text{ for } a_i < x < b_i, \\ [0, +\infty) \text{ for } x = b_i \ \ (\text{ if } b_i < +\infty). \end{cases}$$

1.3.2 Equivalent formulations of the variational inequalities.

A variational inequality

$$(\nabla F(x^*), x - x^*) \geqslant 0 \ \ \forall x \in K, \ x^* \in K,$$

with convex closed set of constraints K can be written as

$$(\nabla F(x^*), x - x^*) + I_K(x) - I_K(x^*) \geqslant 0 \ \ \forall x \in \mathbb{R}^n, \ x^* \in \mathbb{R}^n,$$

where I_K is indicator function of K.

Thus,

$$(\nabla F(x^*), x - x^*) + \varphi(x) - \varphi(x^*) \geqslant 0 \ \ \forall x \in \mathbb{R}^n, \ x^* \in \mathbb{R}^n,$$

with a convex, proper ad lower semicontinuous function φ is a general form of writing both classes of the variational inequalities: with a set of constraints and with a convex non-differentiable function.

Lemma 1.3. *Let $F : \mathbb{R}^n \to \mathbb{R}$ be a convex and differentiable function, while $\varphi : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a convex, proper and lower semicontinuous function. Then three following problems are equivalent:*

$$x^* = \arg \min_{x \in \mathbb{R}^n} \{F(x) + \varphi(x)\},$$

$$x^* \in \mathbb{R}^n : \ (\nabla F(x^*), x - x^*) + \varphi(x) - \varphi(x^*) \geqslant 0 \quad \forall x \in \mathbb{R}^n,$$

$$\nabla F(x^*) + \partial \varphi(x^*) \ni 0.$$

Proof. The proof of the equivalence for a minimisation problem and corresponding variational inequality is given in Lemma 1.2. Equivalence of the variational inequality and the inclusion foolows from the definition of the subdifferential. □

Corollary 1.1. *If $A = A^T \geqslant 0$, then three following problems are equivalent:*

$$x^* = \arg \min_{x \in \mathbb{R}^n} \{\frac{1}{2}(Ax, x,) + \varphi(x)\},$$

$$(Ax^*, x - x^*) + \varphi(x) - \varphi(x^*) \geqslant 0 \quad \forall x \in \mathbb{R}^n,$$

$$Ax^* + \partial \varphi(x^*) \ni 0.$$

In case $A \neq A^T$ a variational inequality is also equivalent to corresponding inclusion, while there is now an equivalent to it minimisation problem.

§2 Iterative methods for variational inequalities with positive definite matrices

In this section, we consider a variational inequality

$$(Au, v-u) + \varphi(v) - \varphi(u) \geqslant (f, v-u) \; \forall v \in \mathbb{R}^n. \tag{2.1}$$

with a positive definite matrix $A \in \mathbb{R}^{n\times n}$.

Theorem 2.1. *Let $\varphi : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a convex, proper and lower semicontinuous function, and $A \in \mathbb{R}^{n\times n}$ be a positive definite matrix:*

$$(Ax, x) \geqslant m\|x\|^2 \; \forall x \in \mathbb{R}^n, \; m > 0.$$

Then

1) there exists a unique solution of variational inequality (2.1)*;*

2) if u_1, u_2 are solutions of the variational inequalities

$$(Au_i, v-u_i) + \varphi(v) - \varphi(u_i) \geqslant (f_i, v-u_i) \;\; \forall v \in \mathbb{R}^n,$$

then

$$\|u_1 - u_2\| \leqslant \frac{1}{m}\,\|f_1 - f_2\|.$$

2.1 One-step stationary method

2.1.1 General convergence result

Below we consider the equivalent to variational inequality (2.1) inclusion

$$Au + \partial\varphi(u) \ni f, \tag{2.2}$$

vector u^* being its unique solution.

One-step stationary iterative method for solving (2.2) reads as

$$\frac{u^{k+1} - u^k}{\tau} + Au^k + \partial\varphi(u^{k+1}) \ni f, \tag{2.3}$$

where $\tau > 0$ is an iterative parameter.

The iteration method is **correctly defined** in the sense, that for any k there exists a unique solution of (2.3). It follows from Theorem 2.1, applied in the case $A = E$ — identity matrix.

Note, that an iterative method

$$\frac{u^{k+1} - u^k}{\tau} + Au^k + \partial\varphi(u^k) \ni f$$

is not correctly defined, because the operator $\partial\varphi$ is multivalued and $\partial\varphi(u^k)$ is generally a set of values.

A calculated iteration u^{k+1} **have to be an argument of the multivalued operator** $\partial\varphi$ (at least, in combination with a known u^k). For the case of an indicator function $\varphi = I_K$ this condition can be interpreted, also, as the request for u^{k+1} to satisfy the constraints $u^{k+1} \in K$.

Remark 2.1. If $A = A^T > 0$ and $\varphi = I_K$ is the indicator function of a convex and closed set K, then variational inequality (2.1) is equivalent to the problem of minimisation of quadratical function $J(u) = \frac{1}{2}(Au, u) - (f, u)$ over the set K, while iterative method (2.3) is the gradient method with projection applied to this problem:

$$u^{k+1} = \mathrm{Pr}_K \left(u^k - \tau \nabla J(u^k)\right) \equiv \mathrm{Pr}_K \left(u^k - \tau(Au^k - f)\right). \tag{2.4}$$

Theorem 2.2. *Let $m\,E \leqslant A = A^T \leqslant M\,E$, $m > 0$. Then iterative method (2.3) converges if $\tau \in \left(0, \frac{2}{M}\right)$ and for any initial guess $u^0 \in \mathbb{R}^n$. Optimal iterative parameter is $\tau_0 = \frac{2}{M+m}$ and for $\tau = \tau_0$ the following estimate for rate of convergence holds:*

$$\|u^{k+1} - u^*\| \leqslant \frac{M-m}{M+m} \|u^k - u^*\| \quad \forall k. \tag{2.5}$$

If A is non-symmetric and

$$(Au, u) \geqslant m\|u\|^2, \quad (Au, v) \leqslant M^{1/2}(Au, u)^{1/2}\|v\|,$$

then iterative method (2.3) converges if $\tau \in \left(0, \frac{2}{M}\right)$ and for any initial guess $u^0 \in \mathbb{R}^n$. Optimal iterative parameter is $\tau_0 = \frac{1}{M}$ and for $\tau = \tau_0$

$$\|u^{k+1} - u^*\| \leqslant \left(1 - \frac{m}{M}\right)^{1/2} \|u^k - u^*\| \quad \forall k. \tag{2.6}$$

The implementation of method (2.3) is very easy if $\partial\varphi$ is a diagonal operator:

$$\partial\varphi = \mathrm{diag}(\partial\varphi_1, \partial\varphi_2, \ldots, \partial\varphi_n),$$

i. e. if it is the subdifferential of a separable function $\varphi(u) = \sum_{i=1}^{n} \varphi_i(u_i)$.

In fact, the implementation of one step of iterative method (2.3) consists of the multiplication of A by a known vector u^k, and of solving the inclusion

$$(E + \tau\partial\varphi)y \ni g^k = u^k + \tau(f - Au^k).$$

In the case of a diagonal operator $\partial\varphi$ this inclusion is decomposed into n scalar (one-dimensional) inclusions

$$u_i + \tau\partial\varphi_i(u_i) \ni g_i^k, \tag{2.7}$$

or, into n problems to minimise strictly convex, proper and lower semicontinuous functions

$$\frac{1}{2}u_i^2 + \tau\varphi_i(u_i) - g_i^k u_i.$$

Such problems can be easily solved by the methods of one-dimensional optimisation. Moreover, if the graph of $\partial\varphi_i$ is a piecewise-linear curve, then problem (2.7) can be solved directly. Such kind of the variational inequalities we consider below (cf. Examples 2.1— 2.3).

In the next section the iterative methods for a class of variational inequalities with non-diagonal operator $\partial\varphi$ will be studied.

2.1.2 Applications to the mesh variational inequalities

Example 2.1. In this example we apply stationary one-step iterative method (2.3) to the solution of the finite difference scheme for the two-dimensional obstacle problem (1.14)

$$y \in K:\ (Ay, z-y) \geqslant (f, z-y)\ \forall z \in K,\ K = \{z \in \mathbb{R}^n :\ z_i \geqslant 0\ \forall i\},$$

where matrix A corresponds to the mesh Laplace operator with homogeneous Dirichlet boundary conditions and is given by formula (1.13). As the matrix A is symmetric, then method (2.3) can be written in the form (2.4):

$$u^{k+1} = \mathrm{Pr}_K\left(u^k - \tau(Au^k - f)\right).$$

Owing to the structure of K, $K = \mathbb{R}^+ \times \mathbb{R}^+ \times \cdots \times \mathbb{R}^+$, the projection of a given vector g reduces to projection of every coordinates of g on $\mathbb{R}^+$, thus,

$$(Pr_K g)_i = y_i^+ \equiv \max\{y_i, 0\} \text{ for all coordinates } y_i.$$

Thus, the implementation of this method is very easy.

Now, let us estimate the rate of convergence of the method. Spectrum of the mesh operator $-\Delta_h$ with homogeneous Dirichlet boundary conditions is well-known, namely,

$\phi_{kl}(t) = \sin k\pi t_1 \sin l\pi t_2,\ t \in \overline{\omega},\ k, l = 1, 2, \ldots, p$, are eigenfunctions, and

$\lambda_{kl} = 4h^{-2}\left(\sin^2\dfrac{k\pi h}{2} + \sin^2\dfrac{l\pi h}{2}\right)$ are corresponding eigenvalues.

So, $m = 8h^{-2}\sin^2\dfrac{\pi h}{2}$ and $M = 8h^{-2}\cos^2\dfrac{\pi h}{2}$ are, respectively, minimal and maximal eigenvalues of the matrix A, and its condition number equals to $\mathrm{cond}_2(A) = \dfrac{M}{m} = O(h^{-2})$.

From Theorem 2.2 the optimal iterative parameter is

$$\tau_0 = \frac{2}{m+M} = \frac{h^2}{4}$$

and rate of convergence of method (2.3) is characterized by the factor

$$q = \frac{M-m}{M+m} = 1 - O(h^2)$$

in the estimate $\|u^{k+1} - u^*\| \leqslant q\|u^k - u^*\|$. It means that one needs

$$n(\varepsilon) = O(h^{-2} \ln \frac{1}{\varepsilon})$$

iterations to get the estimate $\|u^k - u\| \leqslant \varepsilon\|u^0 - u\|$.□

Example 2.2. Consider the finite difference approximation of the obstacle problem with diffusion-convection operator:

$$u \in K: \ \ (\tilde{A}u, v - u) \geqslant (f, v - u) \ \ \forall v \in K = \{u \in \mathbb{R}^n : \ u_i \geqslant 0 \ \forall i\} \tag{2.8}$$

with matrix $\tilde{A}$, corresponding to the mesh operator $-\Delta_h + \bar{a} \cdot \overline{\nabla}_h$. Matrix $\tilde{A}$ equals to the sum of the matrix A, corresponding to $-\Delta_h$, and the matrix $L \in \mathbb{R}^{n \times n}$, corresponding to $\bar{a} \cdot \overline{\nabla}_h$ (see details in Example 1.2).

Obviously, the implementation of (2.3) is similar to the implementation of the mesh obstacle problem with Laplace operator from previous example. In fact, the solution of the inclusion

$$(E + \tau\partial\varphi)y \ni g$$

with the diagonal operator

$$\partial\varphi = \mathrm{diag}(p, p, \ldots, p), \ \text{where } p(t) = \{(-\infty, 0] \text{ for } t \leqslant 0, \ 0 \text{ for } t > 0\}$$

is equivalent to the projection of the vector g on the set $K = \mathbb{R}^+ \times \mathbb{R}^+ \times \cdots \times \mathbb{R}^+$.

Let us estimate the rate of convergence. Symmetric part $\frac{1}{2}(L + L^T)$ of the matrix L corresponds to the mesh operator $-\frac{h}{2}(a_1\bar{\partial}_1\partial_1 + a_2\bar{\partial}_2\partial_2)$. Eigenfunctions of this operator are the same as eigenfunctions of $-\Delta_h$: $\phi_{kl}(t) = \sin k\pi t_1 \sin l\pi t_2, \ t \in \overline{\omega}$, and eigenvalues equal to

$$\lambda_{kl} = 4h^{-1}a_1 \sin^2 \frac{k\pi h}{2} + 4h^{-1}a_2 \sin^2 \frac{l\pi h}{2}, \ \ k, l = 1, 2, \ldots, p.$$

Because of this

$$\begin{gathered} (\tilde{A}u, u) \geqslant m\|u\|^2, \\ m = \lambda_{\min}(A) + \lambda_{\min}(\frac{1}{2}(L + L^T)) = \\ = 8h^{-2} \sin^2 \frac{\pi h}{2} + 4h^{-1}(a_1 + a_2) \sin^2 \frac{\pi h}{2} = O(1). \end{gathered} \tag{2.9}$$

Further, for $u \Leftrightarrow u_h, v \Leftrightarrow v_h$

$$(Lu, v) = h^2 \sum_{t \in \omega_h} \bar{a} \cdot \overline{\nabla}_h u_h(t) v_h(t) \leqslant \sqrt{2} \max\{a_1, a_2\} \left(h^2 \sum_{t \in \omega_h} (\partial_1 u_h(t))^2 + \right.$$

$$+(\partial_2 u_h(t))^2\Big)^{1/2}\left(h^2\sum_{t\in\omega_h}(v_h(t))^2\right)^{1/2} = \sqrt{2}\max\{a_1,a_2\}(Au,u)^{1/2}\|v\|.$$

As, in addition,

$$(Au,v)\leqslant\lambda_{\max}^{1/2}(A)(Au,u)^{1/2}\|v\|\leqslant 2\sqrt{2}h^{-1}(Au,u)^{1/2}\|v\|,$$

then

$$(\tilde{A}u,v)\leqslant(2\sqrt{2}h^{-1}+\sqrt{2}\max\{a_1,a_2\})(Au,u)^{1/2}\|v\|.$$

Finally,

$$\|\tilde{A}u\|^2\leqslant M(\tilde{A}u,u),\ M=(2\sqrt{2}h^{-1}+\sqrt{2}\max\{a_1,a_2\})^2. \tag{2.10}$$

Estimates (2.9) and (2.10) allow to choose optimal iterative parameter, and with this parameter

$$q=(1-m/M)^{1/2}=1-O(h^2),\ \ n(\varepsilon)=O(h^{-2}\ln\frac{1}{\varepsilon}),$$

as in the previous example.□

Example 2.3. Consider mesh variational inequality from Example 1.5:

$$u\in\mathbb{R}^n:\ (Au,y-u)+\varphi(y)-\varphi(u)\geqslant(f,y-u)\ \forall y\in\mathbb{R}^n,$$

where $\varphi(u)=\sum\limits_{i=1}^{n}\alpha_i|u_i|,\ \alpha_i>0$, and stiffness matrix A and vector F are defined by

$$(Au,v)=\int\limits_{\Omega}\nabla u_h(t)\cdot\nabla v_h(t)dt,\ (f,v)=\int_{\Omega}f(t)v_h(t)dt,\ \ u\Leftrightarrow u_h,\ v\Leftrightarrow v_h.$$

Define a symmetric and positive definite matrix D (called as mass matrix in finite element theory):

$$(Du,v)=\int\limits_{\Omega}u_h(t)v_h(t)dt,\quad u\Leftrightarrow u_h,\ v\Leftrightarrow v_h.$$

It is easy to prove the estimate

$$(Au,u)=\int\limits_{\Omega}|\nabla u_h|^2dt\leqslant c_1h_{\min}^{-2}\int\limits_{\Omega}u_h^2dt=c_1h_{\min}^{-2}(Du,u)\ \forall u, \tag{2.11}$$

where $h_{\min}$ is a minimal diameter of all finite elements $\delta\in T_h$ and constant c_1 does not depend on meshsize.

From well-known inequality

$$\int\limits_{\Omega}u^2dt\leqslant c_0\int\limits_{\Omega}|\nabla u|^2dt\ \ \forall u\in H_0^1(\Omega),c_0>0,$$

one get the estimate

$$(Au, u) \geqslant c_0^{-1} \int\limits_{\Omega} u_h^2 dt = c_0^{-1}(Du, u) \quad \forall u \tag{2.12}$$

with constant c_0, independent on meshsize.

Suppose, that the triangulation T_h is **quasiuniform:** there exists a constant $\alpha > 0$, such that $\alpha h \leqslant h_i \leqslant h$ for all i. Then

$$d_0 h^2 E \leqslant D \leqslant d_1 h^2 E, \tag{2.13}$$

where constants d_0 and d_1 don't depend on meshsize.

Estimates (2.11) - (2.13) yield the estimates for the minimal and maximal eigenvalues of the matrix A:

$$\lambda_{\min} = O(h^2),\ \lambda_{\max} = O(1),$$

and condition number of A appears as $O(h^{-2})$. Due to this, rate of convergence of method (2.3) with an optimal iterative parameter is asymptotically in h the same as in previous examples, namely,

$$q = 1 - O(h^2),\ \ n(\varepsilon) = O(h^{-2} \ln \frac{1}{\varepsilon}).$$

To implement the method we have to solve inclusion $(E + \tau\partial\varphi)y \ni g$ with diagonal operator $\partial\varphi = \mathrm{diag}(p_1, p_2, \ldots, p_n)$, where

$$p_i(t_i) = \{-\alpha_i \text{ for } t_i < 0;\ [-\alpha_i, \alpha_i] \text{ for } t_i = 0;\ \alpha_i \text{ for } t_i > 0\}.$$

It is decomposed into n one-dimensional inclusions $t_i + p_i(t_i) \ni g_i$, the solution of i-th inclusion is

$$t_i = \begin{cases} g_i + \tau\alpha_i \text{ for } g_i < -\tau\alpha_i, \\ 0 \text{ for } -\tau\alpha_i \leqslant g_i \leqslant \tau\alpha_i, \\ g_i - \tau\alpha_i \text{ for } g_i > \tau\alpha_i. \end{cases}$$

□

2.1.3 Numerical example

Let $\Omega = (0,1) \times (0,1)$ be the unit square, and ω_h be a unform mesh on $\bar{\Omega}$ with meshsize $h = 1/N$. So, ω_h contains $(N+1) \times (N+1)$ grid points. By $\partial\omega$ we denote the set of the boundary grid points, i. e. (ih, jh) for $i = 1$ or $i = N+1$ or $j = 1$ or $j = N+1$. We will write $(i,j) \in \partial\omega$.

Hereafter in numerical examples we use double numeration for the components of the mesh functions, namely, u_{ij} and f_{ij} with $i, j = 1, 2, \ldots, N+1$ for the mesh functions u_h and f_h.

Mesh Laplace operator $-\Delta_h$ for the internal points of ω_h is defined by

$$(-\Delta_h u)_{ij} = h^{-2}(-u_{i-1,j} - u_{i+1,j} + 4u_{ij} - u_{i,j-1} - u_{i,j+1}),\ 2 \leqslant i, j \leqslant N.$$

We consider the obstacle problem:

$$(-\Delta_h u)_{ij} + p(u_{ij}) \ni f_{ij} \text{ for } 2 \leqslant i,j \leqslant N,$$

$$u_{ij} = 0 \text{ for } (i,j) \in \partial\omega,$$

where $p(t) = \{(-\infty, 0] \text{ for } t \leqslant 0,\ 0 \text{ for } t > 0\}$.

One step iterative method (2.3) becomes

$$\begin{cases} \dfrac{u_{ij}^{k+1} - u_{ij}^k}{\tau} - \Delta_h u_{ij}^k + p(u_{ij}^{k+1}) \ni f_{ij} \text{ for all } 2 \leqslant i,j \leqslant N, \\ \\ u_{ij}^{k+1} = 0 \text{ for } (i,j) \in \partial\omega. \end{cases} \tag{2.14}$$

Recall, that minimal and maximal eigenvalues of this mesh Laplace operator are $m = 8h^{-2}\sin^2\dfrac{\pi h}{2}$ and $M = 8h^{-2}\cos^2\dfrac{\pi h}{2}$, so, the optimal iterative parameter

$$\tau_0 = \frac{2}{m+M} = \frac{h^2}{4}.$$

For $\tau = \tau_0$ formulas in (2.14) transform into

$$u_{ij}^{k+1} + \frac{h^2}{4} p(u_{ij}^{k+1}) \ni \frac{1}{4}\left(u_{i-1,j}^k + u_{i+1,j}^k + u_{i,j-1}^k + u_{i,j+1}^k\right) + \frac{h^2}{4} f_{ij},$$

whence

$$u_{ij}^{k+1} = \max\left\{0,\ \frac{1}{4}\left(u_{i-1,j}^k + u_{i+1,j}^k + u_{i,j-1}^k + u_{i,j+1}^k\right) + \frac{h^2}{4} f_{ij}\right\}, \quad 2 \leqslant i,j \leqslant N.$$

Algorithm

1. Take an initial guess $u_{ij}, 1 \leqslant i,j \leqslant N+1$, such that $u_{ij} = 0$ for $(i,j) \in \partial\omega$.
2. For $i = 2$ to N
   For $j = 2$ to N do
   $v = \dfrac{1}{4}\left(u_{i-1,j} + u_{i+1,j} + u_{i,j-1} + u_{i,j+1}\right) + \dfrac{h^2}{4} f_{i,j}$
   if $v < 0$, then $\hat{u}_{ij} = 0$, else $\hat{u}_{ij} = v$.
3. For $i = 2$ to N
   For $j = 2$ to N do
   $u_{ij} = \hat{u}_{ij}$.
4. If a stopping criterion is fulfilled, then Stop,
   otherwise go to n. 2.

Let exact solution u_h of the mesh problem is given by the values in the grid nodes of the function

$$u(x,y) = \begin{cases} 100\,y(y-1)(x-0.5)(x-1), & (x,y) \in (0.5,1] \times [0,1], \\ 0, & (x,y) \in [0,0.5] \times [0,1], \end{cases}$$

or, $u(x,y) = (100\,y(y-1)(x-0.5)(x-1))^+$ (coefficient 100 is taken to make $\max u \simeq 1$).

Using the form of writing the variational inequality

$$(-\Delta u_h)_{ij} + \gamma_{ij} = f_{ij},\ \gamma_{ij} \in p(u_{ij}),$$

we can easily construct a right-hand side. Namely, let $\gamma_{ij} = 0$ in the grid nodes $(ih, jh) \in (0.5,1) \times (0,1)$, where u_h is positive, while γ_{ij} be any non-positive number in the nodes $(ih, jh) \in (0,0.5] \times (0,1)$, where $u_h = 0$. Then $\gamma_{ij} \in p(u_{ij})$ and it remains to put

$$f_{ij} = (-\Delta_h u_h)_{ij} + \gamma_{ij} \text{ for all } 2 \leqslant i,j \leqslant N.$$

Numerical experiments were made for

$$f_{ij} = (-\Delta_h u_h)_{ij} \text{ in all grid points,}$$

and for

$$f_{i,j} = \begin{cases} (-\Delta_h u_h)_{ij}, & x > 0.5 \\ -u_{i+1,j}/h^2 & x = 0.5, \\ -1, & x < 0.5. \end{cases}$$

Initial guess was $u = 0$, stopping criterion $\|u - u^*\|_{L2} < \varepsilon = 0.001$, where u^* is the exact solution and

$$\|u\|_{L2} = \Big(\sum_{i,j=1}^{N} h^2\, u_{ij}^2 \Big)^{1/2}.$$

For both cases the results were the same and they are included in Table 1.

N	21	51	101	301
$n(\varepsilon)$	205	1290	5167	46522
$n(\varepsilon)h^2$	0.5125	0.516	0.5167	0.5169

Table 1: Number of iterations $n(\varepsilon)$ to achieve $\|u - u^*\|_{L2} < \varepsilon = 0.001$

We see that number of iterations is of order $N^2 = h^{-2}$, namely $n(\varepsilon) \simeq 0.5N^2$.

2.2 Preconditioned one-step stationary method

2.2.1 General convergence result

As we saw, one-step iterative method for all considered examples was very easy to implement but it is very slow convergent.

How to increase the rate of convergence?

From the theory for systems of linear equations it is known that **preconditioning** is a good approach to do this.

Below we study the following preconditioned one-step stationary method for variational inequality (2.2):

$$B\frac{u^{k+1}-u^k}{\tau}+Au^k+\partial\varphi(u^{k+1})\ni f. \tag{2.15}$$

Here preconditioner B is a symmetric and positive definite matrix.

Theorem 2.3. *1) Let $A=A^T>0$, $B=B^T>0$ and*

$$\alpha\, B\leqslant A\leqslant \beta\, B,\ \ \alpha>0. \tag{2.16}$$

Then method (2.15) *converges for any $\tau\in\left(0,\dfrac{2}{\beta}\right)$ and any $u^0\in\mathbb{R}^n$. Optimal iterative parameter is given by the equality $\tau_0=\dfrac{2}{\alpha+\beta}$ and for $\tau=\tau_0$ the following estimate holds:*

$$\|u^{k+1}-u^*\|_B\leqslant\frac{\beta-\alpha}{\beta+\alpha}\,\|u^k-u^*\|_B\quad\forall k.$$

Hereafter $\|u\|_B=(Bu,u)^{1/2}$ means so-called energetic norm of the symmetric and positive definite matrix B.

2) If matrix A is not symmetric and satisfied the following assumptions

$$(Au,u)\geqslant\alpha\|u\|_B^2,\quad (Au,v)\leqslant\beta^{1/2}(Au,u)^{1/2}\|v\|_B,\quad \alpha>0,$$

for all $u,v\in\mathbb{R}^n$, then iterative method (2.15) *converges for any $\tau\in\left(0,\dfrac{2}{\beta}\right)$. Optimal iterative parameter is $\tau_0=\dfrac{1}{\beta}$ and for $\tau=\tau_0$ the following estimate holds:*

$$\|u^{k+1}-u^*\|_B\leqslant(1-\alpha/\beta)^{1/2}\|u^k-u^*\|_B\quad\forall k.$$

When constructing the preconditioned iterative methods for variational inequalities one has to pay attention to the implementation problems. For example, in method (2.15) matrix B must be chosen close to A to ensure a good rate of convergence, but also in a such manner that the operator $B+\tau\partial\varphi$ is easy to

invert. The implementation of every iteration of the method (2.15) consists of the solution of the inclusion

$$Bu^{k+1} + \tau\partial\varphi(u^{k+1}) \ni Bu^k + \tau(f - Au^k) \equiv g^k,$$

which is equivalent to the variational inequality

$$(Bu^{k+1}, v - u^{k+1}) + \tau\varphi(v) - \tau\varphi(u^{k+1}) \geqslant (g^k, v - u^{k+1}) \; \forall v \in \mathbb{R}^n.$$

Its solving in case of a matrix B, which is close to A, can be of the same complexity as solving the initial variational inequality

$$(Au, v - u) + \varphi(v) - \varphi(u) \geqslant (f, v - u) \; \forall v \in \mathbb{R}^n.$$

In all previous examples for implementation of the one-step iterative method it needs the solution of the inclusion $(E + \tau\partial\varphi)(u) \ni g$. This solution reduces to solving n one-dimensional problems which can be solved directly because both operators, $\partial\varphi$ and unit matrix E, have diagonal form. The same will be if we change unit matrix by a diagonal matrix B. From the theory for systems of linear equations with a matrix A it is well-known, that the best diagonal preconditioner is the diagonal part of A. If A is a finite difference approximation of Laplace operator on the uniform grid, then the diagonal part of matrix A is a scalar matrix ($B = \dfrac{4}{h^2}E$ for two-dimensional case). Thus, iterative method (2.15) with such preconditioner is the same as non-preconditioned method (2.3) with a scaled iterative parameter and it has the same rate of convergence.

In case of finite element approximation of Laplace equation or in case of approximation of a differential operator with variable coefficients the choice B equals to the diagonal part of A is reasonable. But such choice does not improve the asymptotic in meshsize h rate of convergence.

Situation is much better for the problems in which the number of constraints is essentially less than the dimension of the discrete problem. Corresponding example is considered below (Example 2.5).

2.2.2 Applications to the mesh variational inequalities

Example 2.4. Obstacle problem. Finite element approximation.

Consider finite element approximation of the obstacle problem from Example 1.3:

$$u \in K: \quad (Au, v - u) \geqslant (f, v - u) \quad \forall v \in K, \tag{2.17}$$

where $K = \{u \in R^n : u_i \geqslant 0 \; \forall i\}$ and

$$(Au, v) = \int_\Omega \nabla u_h(t) \cdot \nabla v_h(t) dt, \;\; (f, v) = \int_\Omega f(t) v_h(t) dt, \;\; u \Leftrightarrow u_h, \; v \Leftrightarrow v_h.$$

Let us solve (2.17) by preconditioned iterative method (2.15) with a diagonal preconditioner B. B may be the diagonal part of A or a matrix, constructed

by application of a simplest quadrature formulae to the integral $\int\limits_\Omega u_h v_h dx$. We consider the second variant.

Let for any finite element $\delta \in T_h$ the following quadrature formulae is used:

$$\int\limits_\delta \phi(t)dt \approx S_\delta = \frac{\operatorname{mes}\delta}{3}\sum_{i=1}^{3}\phi(a_i), \quad \text{where } a_i \text{ are vertices of } \delta,$$

and let matrix $\tilde{D}$ is given by

$$(\tilde{D}u, v) = \sum_{\delta\in T_h} S_\delta(u_h v_h), \quad u \Leftrightarrow u_h, v \Leftrightarrow v_h.$$

It is easy to check, that $\tilde{D}$ is a diagonal matrix. Moreover, if D is mass matrix defined by the equality

$$(Du, v) = \int\limits_\Omega u_h(t)v_h(t)dt, \quad u \Leftrightarrow u_h, \ v \Leftrightarrow v_h,$$

then

$$d_0\tilde{D} \leqslant D \leqslant d_1\tilde{D}$$

with constants d_0 and d_1, independent on meshsize. Last inequalities and estimates (2.11),(2.12) yield

$$\alpha\,\tilde{D} \leqslant A \leqslant \beta\,\tilde{D}, \quad \alpha = O(1), \ \beta = O(h_{\min}^{-2}).$$

Thus, the rate of convergence of method (2.15) with the preconditioner $B = \tilde{D}$ is characterized by the factor $q = 1 - h_{\min}^2$.□

Example 2.5. Signorini problem.

Let the boundary of a domain $\Omega \in \mathbb{R}^2$ consists of two parts: $\partial\Omega = \Gamma_D \cup \Gamma_C$. Define the space $V = \{u \in H^1(\Omega) : u(t) = 0 \text{ on } \Gamma_D\}$ and the convex set

$$K = \{u \in V : u(t) \geqslant 0 \text{ on } \Gamma_C\}.$$

Signorini problem is the following variational inequality: for a given $f \in L_2(\Omega)$ find $u \in K$, such that

$$\int\limits_\Omega \nabla u \cdot \nabla(v-u)dt \geqslant \int\limits_\Omega f(t)(v-u)dt \quad \forall v \in K. \tag{2.18}$$

Variational inequality (2.18) has a unique solution which can be characterized as

$$-\Delta u = f \text{ in } \Omega, \quad u = 0 \text{ on } \Gamma_D,$$
$$u \geqslant 0, \ \frac{\partial u}{\partial n} \geqslant 0, \ u\frac{\partial u}{\partial n} = 0 \text{ on } \Gamma_C,$$

where n is the unit vector of external normal.

Let $\Omega = (0,1) \times (0,1)$, $f(t)$ is continuous in $\overline{\Omega}$ and $\Gamma_C = \{t \in \partial\Omega : t_1 = 0\}$. Approximate (2.18) by a finite-difference scheme. Let

$$\bar{\omega} = \{t = (ih, jh) : 0 \leqslant i, j \leqslant p+1,\ (p+1)h = 1\}$$

γ is the set of the boundary nodes of $\bar{\omega}$, $\omega = \bar{\omega} \setminus \gamma$ and $\gamma_C = \{t \in \gamma : t_1 = 0,\ 0 < t_2 < 1\}$, $\gamma_D = \gamma \setminus \gamma_C$. Define by V_h the space of mesh functions, which vanish in the nodes of γ_D. Let also f_h be a mesh function such that $f_h(t) = f(t)$ for $t \in \omega$.

Finite-difference scheme for (2.18) is

$$\begin{aligned} &-\Delta_h u_h = f_h \ \text{ in } \omega, \\ &u_h = 0 \text{ on } \gamma_D, \\ &u_h \geqslant 0,\ -\frac{1}{h}\partial_2 u_h - \bar{\partial}_1 \partial_1 u_h \geqslant f_h,\ u_h\,(\frac{1}{h}\partial_2 u_h + \bar{\partial}_1 \partial_1 u_h - f_h) = 0 \text{ on } \gamma_C, \end{aligned} \tag{2.19}$$

where $\Delta_h u_h = \bar{\partial}_1 \partial_1 u_h + \bar{\partial}_2 \partial_2 u_h$ for nodes in ω.

Define the bilinear form

$$a(u_h, v_h) = \sum_{t \in \omega \cup \gamma_C} (\partial_1 u_h(t)\, \partial_1 v_h(t) + \partial_2 u_h(t)\, \partial_2 v_h(t))$$

on $V_h \times V_h$ and the linear form $f_h(v_h) = \sum\limits_{t \in \omega_h \cup \gamma_C} f_h(t) v_h(t)$ and the set

$$K_h = \{u_h \in V_h :\ u_h(t) \geqslant 0\ \forall t \in \gamma_C\}$$

in the space V_h. Then finite-difference scheme (2.19) can be written as the following mesh variational inequality:

$$u_h \in K_h :\ a_h(u_h, v_h - u_h) \geqslant f_h(v_h - u_h)\ \forall v_h \in K_h.$$

Let now matrix $A \in \mathbb{R}^{n \times n}$, $n = p(p+1)$, and vector $f \in \mathbb{R}^n$ is defined by

$$(Au, v) = a(u_h, v_h),\ (f, v) = f_h(v_h) \quad \text{for } u \Leftrightarrow u_h, v \Leftrightarrow v_h,$$

with the lexicographical enumeration of the mesh nodes, while $K = \{u \in \mathbb{R}^n :\ u_i \geqslant 0 \text{ for } i = 1, 2, \ldots, p\}$. Them he mesh variational inequality becomes

$$u \in K :\ (Au, v - u) \geqslant (f, v - u)\ \forall v \in K.$$

Matrix A is symmetric and positive definite, its minimal and maximal eigenvalues are $m = O(1)$ and $M = O(h^{-2})$. Thus, all "traditional" for the stationary one-step iterative method convergence results hold, namely,

$$q = 1 - O(h^2),\ \ n(\varepsilon) = O(h^{-2} \ln \frac{1}{\varepsilon}).$$

The main feature of Signorini mesh problem :

the number of the constraints p is much less than the number of unknowns n.

This gives the possibility to use iterative methods with block-diagonal preconditioners. One such method is constructed below.

Let $u = (y, z)^T$ with $y = (u_1, u_2, \ldots, u_p)^T$, $z = (u_{p+1}, \ldots, u_n)^T$, i. e. y contains the coordinates of the vector u, corresponding to mesh points in γ_C. Corresponding to this decomposition of a vector $u \in \mathbb{R}^n$ are the following block representations of the matrix $A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$, vector $f = \begin{pmatrix} f^1 \\ f^2 \end{pmatrix}$ and operator $\partial I_k(u) = \begin{pmatrix} P(y) \\ 0 \end{pmatrix}$, where

$$P(y) = \operatorname{diag}(p(y_1), \ldots, p(y_p)),\ p(t) = \{(-\infty, 0] \text{ for } t = 0,\ 0 \text{ for } t > 0\}.$$

Using these notations the mesh variational inequality can be written as the following inclusion:

$$\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \begin{pmatrix} y \\ z \end{pmatrix} + \begin{pmatrix} P(y) \\ 0 \end{pmatrix} \ni \begin{pmatrix} f^1 \\ f^2 \end{pmatrix}. \tag{2.20}$$

Let us take the preconditioner

$$B = \begin{pmatrix} D_{11} & 0 \\ 0 & A_{22} \end{pmatrix},\ D_{11} = \operatorname{diag} A_{11} \in \mathbb{R}^p \text{ is the diagonal of } A_{11} \tag{2.21}$$

in iterative method (2.15):

$$D_{11} \frac{y^{k+1} - y^k}{\tau} + A_{11} y^k + A_{12} z^k + P(y^{k+1}) \ni f^1, \tag{2.22}$$

$$A_{22} \frac{z^{k+1} - z^k}{\tau} + A_{21} y^k + A_{22} z^k = f^2. \tag{2.23}$$

Implementation of (2.22) consists of the projection procedure for each coordinate:

$$y_i^{k+1} = a_{ii}^{-1} \big(y_i^k + h\tau (f_i^1 - (A_{11} y^k + A_{12} z^k)_i) \big)^+,$$

while (2.23) is a system of linear equations with symmetric and positive definite matrix A_{22}. There are a lot of effective methods for solving this system.

Let us estimate the constants of the spectral equivalence of the matrices A and B. First, note that $D_{11} = \dfrac{3}{h} E$ is a scalar matrix with the entries $\dfrac{3}{h}$ on the diagonal, and matrix A_{22} corresponds to the mesh Laplace operator $-\Delta_h$ defined in the space V_h^0 of the mesh functions with homogeneous Dirichlet conditions on the boundary γ:

$$(A_{22} u, v) = \sum_{t \in \omega} (\partial_1 u_h(t)\, \partial_1 v_h(t) + \partial_2 u_h(t)\, \partial_2 v_h(t))$$

for $u \Leftrightarrow u_h, v \Leftrightarrow v_h,\ u_h, v_h \in V_h^0$. To the decomposition of a vector $u \in \mathbb{R}^n$: $u = (y,0)^T + (0,z)^T$, $y \in \mathbb{R}^p, z \in \mathbb{R}^{p^2}$, corresponds decomposition of a mesh function $V_h \ni u_h \Leftrightarrow u$: $u_h = u_h^{(1)} + u_h^{(2)}$, where $u_h^{(2)} \in V_h^0$, and $u_h^{(1)}$ differs of zero only in the points of γ_C. Using this fact, for any $u \in \mathbb{R}^n$, $u \Leftrightarrow u_h$, one has

$$(Au,u) = a_h(u_h,u_h) \leqslant 2a_h(u_h^{(1)},u_h^{(1)}) + 2a_h(u_h^{(2)},u_h^{(2)}) =$$

$$= 2\sum_{t\in\gamma_C}\left((\partial_1 u_h^{(1)}(t))^2 + (\frac{1}{h}u_h^{(1)}(t))^2\right) + 2\sum_{t\in\omega}\left((\partial_1 u_h^{(2)}(t))^2 + (\partial_2 u_h^{(2)}(t))^2\right) \leqslant$$

$$\leqslant \frac{10}{h^2}\sum_{t\in\gamma_C}\left(u_h^{(1)}(t)\right)^2 + 2\sum_{t\in\omega}\left((\partial_1 u_h^{(2)}(t))^2 + (\partial_2 u_h^{(2)}(t))^2\right) =$$

$$= \frac{10}{3h}(D_{11}y,y) + 2(A_{22}z,z) \leqslant \frac{10}{3h}(Bu,u).$$

On the other hand, for any function $u_h \in V_h$ equality $-u_h(t_1,0) = h\sum\limits_{t_2=0}^{1}\partial_2 u_h(t)$ holds. So,

$$\sum_{t\in\gamma_C} u_h^2(t) \leqslant h\sum_{t\in\omega\cup\gamma_C}(\partial_2 u_h^{(2)}(t))^2,$$

whence $(Bu,u) \leqslant 3(Au,u)\ \forall u \in \mathbb{R}^n$. Finally, we get the following estimates of the spectral equivalence for matrices A and B:

$$\frac{1}{3}B \leqslant A \leqslant \frac{10}{3h}B. \tag{2.24}$$

It means, that the rate of convergence of the stationary one-step method with preconditioner B (2.21) is characterized by factor $q = 1 - O(h)$ and number of iterations $O(h^{-1}\ln\frac{1}{\varepsilon})$ to achieve accuracy ε.

This asymptotically in h one order better estimate than estimate for non-preconditioned method.□

2.2.3 Numerical example

We solved Signorini problem, approximating by a finite difference scheme on a uniform grin in the unit square. The one-sided (Signorini) condition was taken for $y = 0$. Preconditioned one-step method was used (as described in Example 2.5).

The exact solution u^* of the mesh variational inequality was the mesh function, corresponding to

$$u(x,y) = \begin{cases} 100x(1-y)(x-0.5)y, & 0 \leqslant x \leqslant 0.5, 0 \leqslant y \leqslant 1 \\ 100(1-x)(x-0.5)(1-y^2), & 0.5 < x \leqslant 1, 0 \leqslant y \leqslant 1, \end{cases}$$

N	21	51	101	301
$n(\varepsilon)$	57	156	321	984

Table 2: Number of iterations in preconditioned one-step iterative method for Signorini problem. Initial guess is $u^0 = 0$.

N	21	51	101	301
$n(\varepsilon)$	58	158	325	995

Table 3: Number of iterations in preconditioned one-step iterative method for Signorini problem. Initial guess is $u^0 = -1$.

and the right-hand side

$$f_{ij} = \begin{cases} -(u^*_{i+1,j} + u^*_{i-1,j} - 4u^*_{ij} + 2u^*_{i,j+1})/h^2, & j = 0,\ (y = 0), \\ -(u^*_{i-1,j} + u^*_{i+1,j} - 4u^*_{ij} + u^*_{i,j-1} + u^*_{i,j+1})/h^2, & \text{otherwise.} \end{cases}$$

Stopping criterion was $\|u^n - u^*\|_{L2} < \varepsilon = 0.001$.

2.3 Relaxation methods

2.3.1 General convergence result.

let $A \in \mathbb{R}^{n\times n}$ be a symmetric and positive definite matrix, $f \in \mathbb{R}^n$ and $\varphi : \mathbb{R}^n \to \overline{\mathbb{R}}$ be a convex, proper and lower semicontinuous function. We solve inclusion $Au + \partial\varphi(u) \ni f$, which is equivalent to the minimization problem

$$\text{find } \min_{u\in\mathbb{R}^n}\{J(u) = \frac{1}{2}(Au, u) + \varphi(u) - (f, u)\}. \tag{2.25}$$

By u^* a unique solution of this problem is denoted.

Consider a preconditioned stationary one-step method with a variable preconditioner

$$B_k(u^{k+1} - u^k) + Au^k + \partial\varphi(u^{k+1}) \ni f. \tag{2.26}$$

Further we suppose that

$$\{B_k\} \text{ is a bounded sequence of the } n \times n \text{ matrices,} \tag{2.27}$$

$$B_k - \frac{1}{2}A \geqslant c_0\, E, \quad c_0 = \text{const} > 0. \tag{2.28}$$

Owing to (2.28), matrix B_k for any k is positive definite, because of this problem (2.26) has a unique solution for any k (Theorem 2.1).

Theorem 2.4. *Let $A = A^T > 0$ and assumptions* (2.27), (2.28) *are true. Then iterations* (2.26) *converge to u^* for any initial guess u^0.*

2.3.2 Jacobi, Gauss-Seidel and SOR-methods

Extrapolated Jacobi method

Let for $A = A^T > 0$ the matrix $D = \mathrm{diag}(a_{11}, a_{22}, \ldots, a_{nn}) > 0$ be its diagonal part. Then iterative method

$$\frac{1}{\sigma} D(u^{k+1} - u^k) + Au^k + \partial\varphi(u^{k+1}) \ni f, \;\; \sigma > 0,$$

is the extrapolated Jacobi method. It is a partial case of (2.26), as well as of the preconditioned method (2.15), with the preconditioner $B = \dfrac{1}{\sigma} D$. As

$$D^{-1}Au = \lambda u \Rightarrow D^{-1/2}AD^{-1/2}v = \lambda v \;\text{ for } v = D^{1/2}u,$$

then the eigenvalues of matrices $D^{-1}A$ and $D^{-1/2}AD^{-1/2}$ coincide, so do their spectral radiuses. The convergence condition (2.28) of Theorem 2.4 reads as $B > \dfrac{1}{2}A$, and it is equivalent to the inequality $D^{-1/2}AD^{-1/2} < \dfrac{2}{\sigma}E$, i. e.

$$\lambda_{\max}(D^{-1/2}AD^{-1/2}) < \frac{2}{\sigma} \Rightarrow \sigma < \frac{2}{\rho(D^{-1/2}AD^{-1/2})} = \frac{2}{\rho(D^{-1}A)}.$$

The same follows from Theorem 2.3. Moreover, the last theorem gives the theoretically optimal parameter

$$\sigma_0 = \frac{2}{\alpha + \beta}, \;\; \text{where } \; \alpha = \lambda_{\min}(D^{-1}\,A), \; \beta = \lambda_{\max}(D^{-1}\,A).$$

Similar results are true for a block Jacobi method, where

$$D = \mathrm{diag}(A_{11}, A_{22}, \ldots, A_{ss}), \; A_{ii} \in \mathbb{R}^{n_i \times n_i}, \; \sum_{i=1}^{s} n_i = n,$$

is a block diagonal part of $A = A^T > 0$.

Successive overrelaxation method (SOR-method)

Let the matrix A be decomposed as $A = A^T = D + L + L^T > 0$, where $D = \mathrm{diag}(a_{11}, a_{22}, \ldots, a_{nn}) > 0$ is its diagonal part or $D = \mathrm{diag}(A_{11}, A_{22}, \ldots, A_{ss}), \; A_{ii} \in \mathbb{R}^{n_i \times n_i}$ is its block diagonal part, while L is strongly lower triangle part of A.

Taking $B_k = \dfrac{1}{\sigma_k} D + L, \; \sigma_k \in [\varepsilon, 2-\varepsilon], \; \varepsilon > 0$ in (2.26), one get SOR-method, (**point variant** in case of diagonal D and **block variant** if at least for one i the dimension of i-th block $n_i > 1$):

$$\left(\frac{1}{\sigma_k} D + L\right) u^{k+1} + \partial\varphi(u^{k+1}) \ni \left(\frac{1}{\sigma_k} - 1\right) Du^k - L^T u^k + f. \qquad (2.29)$$

It is easy to see, that for $\sigma_k \in [\varepsilon, 2-\varepsilon],\ \varepsilon > 0$,

$$\left((B_k - \frac{1}{2}A)u, u\right) = \left(\frac{1}{\sigma_k} - \frac{1}{2}\right)(Du, u) \geqslant \frac{\varepsilon}{2\,(2-\varepsilon)}\,\lambda_{\min}(A_{ii})\|u\|^2,$$

where $\lambda_{\min}(A_{ii})$ is the minimal eigenvalue of $A_{ii} > 0$. Thus, if $\sigma_k \in [\varepsilon, 2-\varepsilon],\ \varepsilon > 0$, then convergence condition (2.28) is fulfilled, and SOR-method (2.29) converges.

If $\sigma_k = \sigma$ for all k, then (2.28) is true for $\sigma \in (0, 2)$, and SOR-method (2.29)

$$\left(\frac{1}{\sigma}D + L\right)u^{k+1} + \partial\varphi(u^{k+1}) \ni \left(\frac{1}{\sigma} - 1\right)Du^k - L^T u^k$$

converges for $\sigma \in (0, 2)$.

Partial case of SOR-method (2.29) is **Gauss-Seidel method**, corresponding to choice $\sigma = 1$.

How to implement SOR-method (2.29)**?**

In case of a diagonal operator $\partial\varphi$ the implementation of a point variant of this method is as simple as for non-preconditioned one-step iterative method. In fact, on every iterative step of (2.29) one has to solve the inclusion $(\frac{1}{\sigma_k}D + L)u + \partial\varphi(u) \ni g$ with a triangle matrix $\frac{1}{\sigma_k}D + L$. This inclusion is solved recurrently. Namely, as $l_{ij} = 0$ for $j \geqslant i$ and $l_{ij} = a_{ij}$ fr $j < i$, then the following one-dimensional problems

$$\frac{1}{\sigma_k}a_{ii}u_i + \partial\varphi_i(u_i) \ni g_i - \sum_{j<i} a_{ij}u_j$$

are solved sequently for $i = 1, 2, \ldots, n$.

Block variant of method is reasonable to use in case, when the constraints are imposed not at all coordinates of the vector u, and blocks correspond to the coordinates without constraints.

For example, let $u = (y, z)^T$, where $y = (u_1, u_2, \ldots, u_p)^T$, $z = (u_{p+1}, \ldots, u_n)^T$, and there are no constrains to u_i for $i = p+1, p+2, \ldots, n$. Let also the operator $\partial\varphi$ is a diagonal one:

$$\partial\varphi(u) = (\partial\varphi_1(u_1), \partial\varphi_2(u_2), \ldots, \partial\varphi_p(u_p), 0, \ldots, 0).$$

For the decomposition $u = (y, z)^T$ of $u \in \mathbb{R}^n$ the following bock representation of the matrix $A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$ corresponds. Choose

$$D = \begin{pmatrix} D_{11} & 0 \\ 0 & A_{22} \end{pmatrix},\ D_{11} = \operatorname{diag}(a_{11}, \ldots, a_{pp}).$$

Then implementation of this block SOR-method consists of sequential solution of he one-dimensional inclusions for $i = 1, 2, \ldots, p$ and a system of linear equations with matrix A_{22}.

The described block variant of SOR-method can be used, for example, for solving mesh Signorini problem (cf. Example 2.5).

Symmetric successive overrelaxation method (SSOR-method).

Let $A = D + L + L^T = A^T > 0$, where $D = \operatorname{diag}(A_{11}, A_{22}, \ldots, A_{ss})$, $A_{ii} \in \mathbb{R}^{n_i \times n_i}$, $\sum_{i=1}^{s} n_i = n$, is a block diagonal part (in particular, diagonal part) of A, and L is strongly lower triangle part of A. Choosing

$$B_k = \frac{1}{\sigma_k} D + L \text{ for odd } k,\ B_k = \frac{1}{\sigma_k} D + L^T \text{ for even } k,$$

one get (block) SSOR-method

$$\left(\frac{1}{\sigma_k} D + L\right) u^{k+1/2} + \partial\varphi(u^{k+1/2}) \ni \left(\frac{1}{\sigma_k} - 1\right) D u^k - L^T u^k + f,$$
$$\left(\frac{1}{\sigma_k} D + L^T\right) u^{k+1} + \partial\varphi(u^{k+1}) \ni \left(\frac{1}{\sigma_k} - 1\right) D u^{k+1/2} - L u^{k+1/2} + f.$$

This method converges if $\sigma_k \in [\varepsilon, 2-\varepsilon]$, $\varepsilon > 0$ and in case of constant relaxation parameter — if $\sigma \in (0, 2)$.

The implementation of SSOR-method is similar to the implementation of SOR-method.

2.3.3 Applications to the mesh variational inequalities. Numerical examples

Once again consider the obstacle problem as in 2.1.3. Let $\Omega = (0,1) \times (0,1)$, ω_h be a unform mesh on Ω with $h = 1/N$. The mesh obstacle problem is

$$(-\Delta u_h)_{ij} + \gamma_{i,j} = f_{ij},\ \gamma_{ij} \in p(u_{i,j}),\ \text{for } 2 \leqslant i, j \leqslant N,$$

$$u_{ij} = 0 \text{ for } (i,j) \in \partial\omega,$$

where $p(t) = \{(-\infty, 0] \text{ for } t \leqslant 0,\ 0 \text{ for } t > 0\}$.

Let us fix the **lexicographical** ordering of the grid points. Then the formulas for Gauss-Seidel method read as

$$u_{ij}^{k+1} - \frac{1}{4}\left(u_{i-1,j}^{k+1} + u_{i,j-1}^{k+1}\right) + \frac{h^2}{4} p(u_{ij}^{k+1}) \ni \frac{1}{4}\left(u_{i+1,j}^{k} + u_{i,j+1}^{k}\right) + \frac{h^2}{4} f_{ij}.$$

It means, that **sequently** for $i = 2, 3, \ldots, N$ and for $j = 2, 3, \ldots, N$ one can find

$$u_{ij}^{k+1} = \max\left\{0,\ \frac{1}{4}\left(u_{i-1,j}^{k+1} + u_{i,j-1}^{k+1} + + u_{i+1,j}^{k} + u_{i,j+1}^{k}\right) + \frac{h^2}{4} f_{ij}\right\}.$$

Algorithm for Gauss-Seidel method

1. Take an initial guess $u_{ij}, 1 \leqslant i, j \leqslant N+1$, such that $u_{ij} = 0$ for $(i, j) \in \partial\omega$.

2. For $i = 2$ to N
   For $j = 2$ to N do
   $u_{ij} = \dfrac{1}{4}\left(u_{i-1,j} + u_{i,j-1} + u_{i+1,j} + u_{i,j+1}\right) + \dfrac{h^2}{4} f_{ij},$
   if $u_{ij} < 0$, then $u_{ij} = 0$.

3. If a stopping criterion is fulfilled, then Stop,
   otherwise go to n. 2.

For the same lexicographical ordering of the grid points the formulas for SOR-method are

$$\frac{1}{\sigma}u_{ij}^{k+1} - \frac{1}{4}\left(u_{i-1,j}^{k+1} + u_{i,j-1}^{k+1}\right) + \frac{h^2}{4}p(u_{ij}^{k+1}) \ni \left(\frac{1}{\sigma} - 1\right)u_{ij}^{k} + \frac{1}{4}\left(u_{i+1,j}^{k} + u_{i,j+1}^{k}\right) + \frac{h^2}{4}f_{ij}.$$

It means, that **sequently** for $i = 2, 3, \ldots, N$ and for $j = 2, 3, \ldots, N$ one can find

$$u_{ij}^{k+1} = \max\left\{0,\ (1-\sigma)u_{ij}^{k} + \frac{\sigma}{4}\left(u_{i-1,j}^{k+1} + u_{i,j-1}^{k+1} + u_{i+1,j}^{k} + u_{i,j+1}^{k}\right) + \frac{\sigma h^2}{4}f_{ij}\right\}$$

sequently for all $i = 2, 3, \ldots, N$ and for $i = 2, 3, \ldots, N$.

Algorithm for SOR-method

1. Take an initial guess $u_{ij}, 1 \leqslant i,j \leqslant N+1$, such that $u_{ij} = 0$ for $(i,j) \in \partial\omega$.

2. For $i = 2$ to N
   For $j = 2$ to N do

   $$u_{ij} = (1-\sigma)u_{ij}^k + \frac{\sigma}{4}\left(u_{i-1,j}^{k+1} + u_{i,j-1}^{k+1} + u_{i+1,j}^k + u_{i,j+1}^k\right) + \frac{\sigma h^2}{4} f_{ij},$$

   if $u_{ij} < 0$, then $u_{ij} = 0$.

3. If a stopping criterion is fulfilled, then Stop,
   otherwise go to n. 2.

SSOR-metod differs from SOR-method by the presence of iterations with relaxation on the inverse order. Thus:

Algorithm for SSOR-method

1. Take an initial guess $u_{ij}, 1 \leqslant i,j \leqslant N+1$, such that $u_{ij} = 0$ for $(i,j) \in \partial\omega$.

2. For $i = 2$ to N
   For $j = 2$ to N do

   $$u_{ij} = (1-\sigma)u_{ij}^k + \frac{\sigma}{4}\left(u_{i-1,j}^{k+1} + u_{i,j-1}^{k+1} + u_{i+1,j}^k + u_{i,j+1}^k\right) + \frac{\sigma h^2}{4} f_{ij},$$

   if $u_{ij} < 0$, then $u_{ij} = 0$.

3. For $i = N$ down to 2
   For $j = N$ down to 2 do

   $$u_{ij} = (1-\sigma)u_{ij}^k + \frac{\sigma}{4}\left(u_{i-1,j}^{k+1} + u_{i,j-1}^{k+1} + u_{i+1,j}^k + u_{i,j+1}^k\right) + \frac{\sigma h^2}{4} f_{ij},$$

   if $u_{ij} < 0$, then $u_{ij} = 0$.

4. If a stopping criterion is fulfilled, then Stop,
   otherwise go to n. 2.

Let the exact solution u_h of the mesh problem be given as before by the values in the grid nodes of the function $u(x,y) = (100\,y(y-1)(x-0.5)(x-1))^+$. Numerical experiments were made for

$$f_{i,j} = \begin{cases} (-\Delta_h u_h)_{ij}, & x > 0.5 \\ -u_{i+1,j}/h^2 & x = 0.5, \\ -1, & x < 0.5. \end{cases}$$

Initial guess was $u = 0$, stopping criterion $\|u - u^*\|_{L2} < \varepsilon = 0.001$.

Table 4 contains the comparison results for Jacobi method, Gauss-Seidel method and SOR-method with (numerically found) optimal relaxation parameter σ_0.

N	21	51	101	301	501
$n(\varepsilon)$ for Jacobi method	205	1290	5167	46522	too many
$n(\varepsilon)$ for Gauss-Seidel method	103	645	2584	23261	64616
$n(\varepsilon)$ for SOR-method	20	60	111	270	521
optimal σ_0	1.7	1.8	1.9	1.97	1.98

Table 4: Number of iterations $n(\varepsilon)$ to achieve $\|u - u^*\|_{L2} < \varepsilon = 0.001$

Remark 2.2. For SOR-method applied to solving a system of linear equations $Au = f$ with a symmetric and positive definite matrix A there is a theory of choosing the optimal relaxation parameter σ. This theory uses the properties of the matrix A, and optimal parameter is defined by these properties. SOR-method with the optimal parameter (for mesh problems it is σ, close to 2,) has a rate of convergence, which is essntially better than, for example, a rate of convergence for Gauss-Seidel method ($\sigma = 1$). For example, for solving a system of equations with the matrix A, corresponding to mesh Laplace operator, the rate of convergence for SOR-method with the optimal parameter ($\sigma \simeq 2 - O(h)$ for $h \to 0$), is characterised by a factor $q = 1 - O(h)$ instead of $q = 1 - O(h^2)$ as for Gauss-Seidel method.

For the mesh variational inequalities there is no similar theory. □

Below we give the results of numerical experiments, which purpose was to find the optimal relaxation parameters for different meshes, and compare them with the known optimal relaxation parameters for linear case.

For a matrix, corresponding to the mesh Laplace operator on the uniform grid with meshsize h the theoretically optimal relaxation parameter in SOR-method applied to the **system of linear equations** is

$$\sigma^* = 2/(1 + \sin(\pi h)).$$

In the following tables we give the number of SOR-iterations $n(\varepsilon)$ for different grids and different relaxation parameters. Initial guess is $u^0_{ij} = 0$, and stopping criterion is $\|u - u^*\|_{L2} < \varepsilon = 0.001$.

$N = 21$

σ	1.5	$\sigma_0 = 1.6$	1.7	$\sigma*$=1.74	1.8	1.9
$n(\varepsilon)$	31	20	30	29	29	48
$n(\varepsilon)h$		1.0		1.45		

$N = 51$

σ	1.5	1.6	1.7	$\sigma_0 = 1.8$	$\sigma*$=1.88	1.9
$n(\varepsilon)$	212	157	108	60	75	73
$n(\varepsilon)h$				1.2	1.5	

$N = 101$

σ	1.7	1.8	$\sigma_0 = 1.9$	$\sigma*$=1.94	1.95
$n(\varepsilon)$	450	278	111	151	147
$n(\varepsilon)h$			1.11	1.51	

$N = 301$

σ	1.9	1.95	1.96	$\sigma_0 = 1.97$	$\sigma*$=1.97934	1.98
$n(\varepsilon)$	1205	555	418	270	454	451
$n(\varepsilon)h$				0.9	1.513	

$N = 501$

σ	1.97	$\sigma_0 = 1.98$	$\sigma*$=1.98754	1.99
$n(\varepsilon)$	914	521	757	737
$n(\varepsilon)h$		1.042	1.514	

Table 5: Number of iterations to achieve $\|u - u^*\|_{L2} < \varepsilon = 0.001$. Comparison results for different relaxation parameters.

In the next table we collect for all meshes the following results: optimal for linear case relaxation parameter σ^*, experimentally founded optimal parameter σ_0 for the obstacle problem, dependence of the number of iterations upon the meshsize h.

N	21	51	101	301	501
σ^*	1.74	1.88	1.94	1.97934	1.98754
$n(\varepsilon)h$	1.45	1.5	1.51	1.513	1.514
σ_0	1.6	1.8	1.98	1.97	1.98
$n(\varepsilon)h$	1.0	1.2	1.11	0.9	1.042

Table 6: Dependence of iterations number upon meshsize h.

One more numerical test was made for the obstacle problem, which free boundary consists of many lines and is not smooth.

Namely, let the exact solution of the obstacle problem be

$$u_{ij} = (\sin(6\pi ih)\sin(6\pi jh))^+$$

and the right-hand side be chosen as

$$f_{ij} = \begin{cases} -1, & \text{if } u_{ij} < 0; \\ -(\Delta_h u)_{ij},, & \text{otherwise .} \end{cases}$$

Table 7 contain the results for the initial guess $u_{ij}^0 = 1$ and optimal relaxation parameter σ_0, which was found experimentally for every mesh.

N	21	51	101	301	501
σ_0	1.7	1.9	1.94	1.98	1.99
$n(\varepsilon)$	37	84	147	361	588
$n(\varepsilon)h$	1.85	1.68	1.47	1.2	1.176

Table 7: Number of iterations to achieve $\|u - u^*\|_{L2} < \varepsilon = 0.001$. Dependence of iterations number upon meshsize h.

Conclusions:

1) The optimal parameter σ_0 in SOR-method, applied to the mesh obstacle problem $Au + \partial\varphi(u) \ni f$, was close to the optimal parameter σ^* for the system of linear equations $Au = f$ with the same matrix A. Moreover,

smaller meshsize h $\Rightarrow$ closer σ_0 to σ^*.

2) The number of SOR-iterations with both relaxation parameters σ_0 and σ^* was proportional to $N = 1/h$.

These conclusions, made on the basis of several numerical tests, are really true for a wide class of variational inequalities.

Remark 2.3. In general, it is not possible to compute a priori the optimal value of the relaxation parameter σ. Frequently, for the systems $Au = f$ of linear mesh equations the following heuristic estimate is used:

$$\sigma^* \simeq 2 - ch,\ c = const.$$

In our numerical examples the optimal relaxation parameter for linear case is known and it just satisfies this estimate:

$$\sigma^* = 2/(1 + \sin(\pi h)) \simeq 2 - 2\pi h.$$

□

Let us look what happens when solving a variational inequality. In the tables below the result for the theoretically optimal (for corresponding matrix A) parameter σ^* and for experimentally found optimal (for variational inequality) parameter σ_0 are collected. First table contains the results for the obstacle problem with exact solution $u(x, y) = (100\, y(y-1)(x-0.5)(x-1))^+$ and the second one — with exact solution $u(x, y) = (\sin(6\pi x)\sin(6\pi y))^+$.

N	21	51	101	301	501
σ^*	1.74	1.88	1.94	1.97934	1.98754
$(2-\sigma^*)/h$	5.2	6.0	6.0	6.198	6.23
σ_0	1.6	1.8	1.9	1.97	1.98
$(2-\sigma_0)/h$	8	10	10	9	10

Table 8: Optimal parameters and constant c in the formula $\sigma = 2 - ch$. Exact solution $u(x,y) = (100\, y(y-1)(x-0.5)(x-1))^+$

N	21	51	101	301	501
σ_0	1.7	1.9	1.94	1.98	1.99
$(2-\sigma_0)/h$	6	5	6	6	5

Table 9: Optimal parameter and constant c in the formula $\sigma = 2 - ch$. Exact solution $u(x,y) = (\sin(6\pi x)\sin(6\pi y))^+$

From calculating results one can see that for optimal parameter σ_0 the dependence $\sigma_0 = 2 - c_0 h$ is also true.

This fact motivates the using of the following **method to find a relaxation parameter**, which is close to the optimal σ_0:

1) Find an optimal parameter σ_0 by numerical tests on a coarse grid with meshsize h_0.

2) Take $c = \dfrac{2-\sigma_0}{h_0}$ for calculating the relaxation parameter for a fine grid with meshsize h: $\sigma = 2 - ch$.

Note, that this method can be applied in the case of the **uniform grids** (or, at least, quasiuniform grids).

2.4 Error control and stopping criteria

2.4.1 Generalities

Error control and stopping criteria are very important aspects of the implementation of numerical methods. To control exactness of an iterative method when solving a mesh variational inequality there a several approaches.

First, if a rate of convergence is known:

$$\|u^{k+1} - u^*\|_s \leqslant q\, \|u^k - u^*\|_s \ \ \forall k,\ q < 1,$$

for a vector norm $\|.\|_s$, then one can estimate the norm of error $\|u^k - u^*\|_s$ by the norm of the difference of two current iterations $\|u^{k+1} - u^k\|_s$. Namely,

$$\|u^{k+1} - u^k\|_s \geqslant \|u^k - u^*\|_s - \|u^{k+1} - u^*\|_s \geqslant (1-q)\|u^k - u^*\|_s,$$

whence

$$\|u^k - u^*\|_s \leqslant \frac{1}{1-q}\, \|u^{k+1} - u^k\|_s. \tag{2.30}$$

The most universal method for error controlling is the estimating of a norm of a **residual vector.**

What is the residual vector for an iterative method, applied to solution of a variational inequality?

Let us consider iterative method (2.15). When implementing this method, the following inclusion is solved:

$$Bu^{k+1} + \tau\partial\varphi(u^{k+1}) \ni Bu^k + \tau(f - Au^k) \equiv g^k.$$

From this inclusion one can find not only u^{k+1}, but also

$$\gamma^{k+1} = \frac{1}{\tau}B(u^k - u^{k+1}) + f - Au^k \in \partial\varphi(u^{k+1}).$$

By residual vector we call the vector $r^{k+1} = Au^{k+1} + \gamma^{k+1} - f$.

The inclusion $Au + \partial\varphi(u) \ni f$, which we solve by iterative method (2.15), can be written as

$$Au + \gamma = f,\ \gamma \in \partial\varphi(u).$$

If (u^*, γ^*) is its solution, then $r^{k+1} = A(u^{k+1} - u^*) + (\gamma^{k+1} - \gamma^*)$. As matrix A is positive definite: $(Au, u) \geqslant m\|u\|^2$, and operator $\partial\varphi$ is monotone, then

$$(r^{k+1}, u^{k+1} - u^*) \geqslant (A(u^{k+1} - u^*), u^{k+1} - u^*) = \|u^{k+1} - u^*\|_A^2 \geqslant m\|u^{k+1} - u^*\|^2,$$

and from the estimate for norm of residual $\|r^{k+1}\|$ we have the following error estimates:

$$\|u^{k+1} - u^*\| \leqslant m^{-1}\|r^{k+1}\|,$$

$$\|u^{k+1} - u^*\|_A \leqslant \|r^{k+1}\|_{A^{-1}} \leqslant m^{-1/2}\|r^{k+1}\|.$$

Let us underline, that no information about the rate of convergence for an iterative method is needed to get these error estimates.

2.4.2 Numerical example

First, two different stopping criteria were used, namely, $\|r\|_{L2} < \varepsilon = 0.001$ and $\|r\|_C < \varepsilon = 0.001$ for solving the obstacle problem by SOR-method. Here $\|r\|_C$ is the maximum norm for residual vector. Number of iterations are denoted, respectively, by $n_r(\varepsilon)(L_2)$ for the first criterion and by $n_r(\varepsilon)(C)$ for the second one. We includes in the table the number of iterations to achieve the accuracy $\|u - u^*\|_{L_2} < \varepsilon = 0.001$, where u^* is the exact solution. Initial guess was $u = 0$.

The results are included in Table 10.

$N = 51$

σ	1.5	1.6	1.7	$\sigma_0 = 1.8$	$\sigma*$=1.88	1.9
$n_r(\varepsilon)(L_2)$	343	255	175	97	110	117
$n_r(\varepsilon)(C)$	377	280	192	107	120	141
$n(\varepsilon)$	212	157	108	60	75	73

$N = 101$

σ	1.7	1.8	$\sigma_0 = 1.9$	$\sigma*$=1.94	1.95
$n_r(\varepsilon)(L_2)$	727	449	176	222	255
$n_r(\varepsilon)(C)$	800	494	195	243	294
$n(\varepsilon)$	450	278	111	151	147

$N = 301$

σ	1.9	1.95	1.96	$\sigma_0 = 1.97$	$\sigma*$=1.97934	1.98
$n_r(\varepsilon)(L_2)$	1945	896	672	599	671	668
$n_r(\varepsilon)(C)$	2142	988	742	670	762	838
$n(\varepsilon)$	1205	555	418	270	454	451

Table 10: Number of iterations, when stopping criteria are norms of the residual.

N	21	51	101	301
$n(\varepsilon)$	373	2220	8735	77747
$\|u^{k+1} - u^k\|_{L_2}/(1-q)$	0.00098	0.00099	0.00099	0.00099
$\|u^k - u^*\|_{L_2}$	0.00051	0.00046	0.00042	0.00041
$\|r\|_{L2}$	0.01005	0.01004	0.00995	0.00990
q	0.98769	0.99803	0.99951	0.99995

Table 11: Comparison results for different stopping criteria.

The same mesh obstacle problem was solved by Jacobi method, which coincides for this problem with preconditioned one-step stationary method, the preconditioner being the diagonal part of the matrix A, corresponding to mesh Laplace operator for uniform grid. For this method the following estimate is valid:

$$\|u^{k+1} - u^*\|_{L_2} \leqslant q\|u^k - u*\|_{L_2}$$

with

$$q = \frac{M-m}{M+m},\ m = \frac{8}{h^2}\sin^2\frac{\pi h}{2},\ M = \frac{8}{h^2}\cos^2\frac{\pi h}{2}.$$

Thus,

$$\|u^k - u^*\|_{L_2} \leqslant \frac{1}{1-q}\|u^{k+1} - u^k\|_{L_2}.$$

We used the criterion $\dfrac{\|u^{k+1} - u^k\|_{L_2}}{1-q} < \varepsilon = 0.001$. Initial guess was $u_0 = 1$. Table 11 contains the calculated results.

These results ensure that one can use to control the error of the iterations either the norm of the difference of two current iterations (if the rate of convergence is known) or a norm of the residual vector. The last estimate is the most universal one.

2.5 Splitting iterative methods

2.5.1 General convergence theory

We will solve inclusion (2.2)

$$Au + \partial\varphi(u) \ni f$$

with a positive definite matrix A by the following iterative method:

$$DB\frac{u^{k+1} - u^k}{\tau} + Au^k + \partial\varphi(B(u^{k+1} - u^k) + u^k) \ni f, \tag{2.31}$$

where B is a regular matrix (i. e. exists B^{-1}), D is a symmetric and positive definite matrix, and $\tau > 0$ is an iterative parameter.

Inclusion (2.31) is equivalent to the system

$$D\frac{u^{k+1/2} - u^k}{\tau} + Au^k + \partial\varphi(u^{k+1/2}) \ni f, \tag{2.32}$$

$$B(u^{k+1} - u^k) = u^{k+1/2} - u^k, \tag{2.33}$$

whence the name "splitting" for the method.

The partial cases of (2.31) are generalised Peacemen-Rachford and Douglas-Rachford methods:

Peacemen-Rachford method ($B = \frac{1}{2}(E + \tau A)$ in (2.31))

$$\begin{cases} D\dfrac{u^{k+1/2} - u^k}{\tau} + Au^k + \partial\varphi(u^{k+1/2}) \ni f, \\ \dfrac{u^{k+1} - 2u^{k+1/2} + u^k}{\tau} + A(u^{k+1} - u^k) = 0; \end{cases} \tag{2.34}$$

Douglas-Rachford method ($B = E + \tau A$ in (2.31))

$$\begin{cases} D\dfrac{u^{k+1/2} - u^k}{\tau} + Au^k + \partial\varphi(u^{k+1/2}) \ni f, \\ \dfrac{u^{k+1} - u^{k+1/2}}{\tau} + A(u^{k+1} - u^k) = 0. \end{cases} \tag{2.35}$$

As we see, the first step (2.32) of method (2.31) coincides with the preconditioned one-step method, while the second step (2.33) can be viewed as a refinement of the iteration $u^{k+1/2}$.

The implementation of method (2.31) consists of the implementation of step (2.32), i. e. the preconditioned one-step method, which have been discussed before, and of solving system of linear equations (2.33).

Below the convergence results for the generalised Peacemen-Rachford and Douglas-Rachford methods in cases of symmetric and nonsymmetric matrix A are cited.

Theorem 2.5. *Let*

$$mD \leqslant A = A^T \leqslant MD,\ m > 0. \tag{2.36}$$

Then iterative methods (2.34) *and* (2.35) *converge for any initial guess* u^0 *and any iterative parameter* $\tau > 0$.

Optimal iterative parameter is $\tau_0 = \dfrac{1}{\sqrt{mM}}$ *and for* $\tau = \tau_0$ *the following estimates for rate of convergence are valid:*

$$\|(E + \tau_0 D^{-1}A)z^{k+1}\|_D \leqslant \frac{\sqrt{M} - \sqrt{m}}{\sqrt{M} + \sqrt{m}}\|(E + \tau_0 D^{-1}A)z^k\|_D$$

for method (2.34) *and*

$$\|(E + \tau_0 D^{-1}A)z^{k+1}\|_D \leqslant \frac{\sqrt{M}}{\sqrt{M} + \sqrt{m}}\|(E + \tau_0 D^{-1}A)z^k\|_D$$

for method (2.35).

Theorem 2.6. *Let*

$$(Au, u) \geqslant \delta\|u\|_D^2,\ \delta > 0, \quad (D^{-1}Au, Au) \leqslant \Delta(Au, u) \quad \forall u \in \mathbb{R}^n.$$

Then iterative methods (2.34) *and* (2.35) *converge for any initial guess* u^0 *and any iterative parameter* $\tau > 0$.

Optimal iterative parameter is $\tau_0 = \dfrac{1}{\sqrt{\Delta\,\delta}}$ *and for* $\tau = \tau_0$ *the following estimates for rate of convergence are valid:*

$$\|(E + \tau_0 D^{-1}A)z^{k+1}\|_D \leqslant q^{1/2}\|(E + \tau_0 D^{-1}A)z^k\|_D, \tag{2.37}$$

where $q = q_0 = \dfrac{\sqrt{\Delta} - \sqrt{\delta}}{\sqrt{\Delta} + \sqrt{\delta}}$ *for method* (2.34) *and* $q = \dfrac{1}{2} + \dfrac{q_0}{2}$ *for method* (2.35).

2.5.2 Applications to mesh variational inequalities

Splitting iterative methods can be applied to all mesh variational inequalities, considered in Examples 2.1 — 2.5, because matrices of these variational inequalities are positive definite.

The implementation of the non-preconditioned Peacemen-Rachford and Douglas-Rachford methods are similar to the implementation of non-preconditioned one-step iterative methods.

Now, let us estimate the rate of convergence. As was proved, the condition number $\mathrm{cond}(A)$ of a matrix A in any of the cited examples is $O(h^{-2})$. So, for the optimal iterative parameter τ_0 the factor

$$q = \frac{\sqrt{M} - \sqrt{m}}{\sqrt{M} + \sqrt{m}} = \frac{\sqrt{\mathrm{cond}(A)} - 1}{\sqrt{\mathrm{cond}(A)} + 1} = 1 - O(h).$$

It means, that non-preconditioned Peacemen-Rachford and Douglas-Rachford methods, applied to the mesh variational inequalities of Examples 2.1 — 2.5 request

$$n(\varepsilon) = O(h^{-1} \ln \frac{1}{\varepsilon})$$

iterations to get the estimate $\|u^k - u\| \leqslant \varepsilon \|u^0 - u\|$. $\square$

For the Signorini problem is possible to use the preconditioned methods. Namely, let $u = (y, z)^T$ with $y = (u_1, u_2, \ldots, u_p)^T$, $z = (u_{p+1}, \ldots, u_n)^T$, i. e. y contains the coordinates of the vector u, corresponding to mesh points in γ_C. Corresponding to this decomposition of a vector $u \in \mathbb{R}^n$ are the following block representations of the matrix $A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$, vector $f = \begin{pmatrix} f^1 \\ f^2 \end{pmatrix}$ and operator $\partial I_k(u) = \begin{pmatrix} P(y) \\ 0 \end{pmatrix}$, where

$$P(y) = \operatorname{diag}(p(y_1), \ldots, p(y_p)),\ p(t) = \{(-\infty, 0] \text{ for } t = 0,\ 0 \text{ for } t > 0\}.$$

Let us take the preconditioner

$$B = \begin{pmatrix} D_{11} & 0 \\ 0 & A_{22} \end{pmatrix},\ D_{11} = \operatorname{diag} A_{11} \in \mathbb{R}^p \text{ is the diagonal of } A_{11}$$

in iterative method (2.35):

$$\begin{cases} D_{11} \dfrac{y^{k+1/2} - y^k}{\tau} + A_{11} y^k + A_{12} z^k + P(y^{k+1}) \ni f^1, \\ \\ A_{22} \dfrac{z^{k+1/2} - z^k}{\tau} + A_{21} y^k + A_{22} z^k = f^2. \\ \\ \dfrac{u^{k+1} - u^{k+1/2}}{\tau} + A(u^{k+1} - u^k) = 0. \end{cases} \tag{2.38}$$

Implementation of (2.38) consists of the projection procedure:

$$y_i^{k+1} = a_{ii}^{-1} \big(y_i^k + h\tau (f_i^1 - (A_{11} y^k + A_{12} z^k)_i) \big)^+,$$

and the solution of the systems of linear equations.

Further, in Example 2.5 we have estimated the constants of the spectral equivalence of the matrices A and D (here we use notation D for the preconditioner):

$$\frac{1}{3} D \leqslant A \leqslant \frac{10}{3h} D \Rightarrow M = \frac{10}{3h},\ m = \frac{1}{3}.$$

Thus, for the optimal iterative parameter τ_0 the factor

$$q = q_0 = \frac{\sqrt{M} - \sqrt{m}}{\sqrt{M} + \sqrt{m}} = 1 - O(h^{1/2})$$

and number of iterations to get the estimate $\|u^k - u\| \leqslant \varepsilon \|u^0 - u\|$ equals to

$$n(\varepsilon) = O(h^{-1/2} \ln \frac{1}{\varepsilon})$$

□

2.5.3 Numerical example

We consider once again the obstacle problem with exact solution $u(x, y) = (100\, y(y-1)(x-0.5)(x-1))^+$ and right-hand side

$$f_{i,j} = \begin{cases} -(\Delta u)_{ij}, & x > 0.5 \\ -u_{i+1,j}/h^2 & x = 0.5, \\ f = -1, & x < 0.5. \end{cases}$$

The mesh variational inequality is solved by Douglas-Rachford method with the optimal iterative parameter.

Initial guess $u = 0$, stopping criterion $\|u - u^*\|_{L_2} < \varepsilon = 0.001$.

N	21	51	101	301	501
$n(\varepsilon)$	20	46	94	283	472
$\|r\|_{L_2}$	0.0003	0.02	0.02	0.03	0.03
$n(\varepsilon)$ for SOR-method with optimal parameter	20	60	111	270	521

Table 12: Number of iterations to achieve $\|u - u^*\|_{L_2} < 0.001$ and norm of residual for Douglas-Rachford method. Comparison with SOR-method

We see, that for Douglas-Rachford method with optimal iterative parameter $n(\varepsilon) \simeq N$, i. e. is proportional to h^{-1} as it was proved theoretically. The number of iterations is almost the same as in SOR-method with experimentally defined optimal relaxation parameter.

Deficiency of splitting iterative methods is the necessity to solve a system of linear equations at every iteration. It leads to more time consuming in comparison with SOR-method.

Merits:

1) Splitting iterative methods **converge for any iterative parameter** $\tau > 0$ and for optimal iterative parameter have asymptotically the same rate of convergence as SOR-method with the optimal relaxation parameter. **An optimal parameter** $\tau > 0$ **can be defined a priori** by the matrix properties. Experimentally defined optimal parameter (so, really optimal) almost coincides with the theoretical one. The rate of convergence is not very sensible to the choice of an iterative parameter.

2) Splitting iterative methods can be applied to problems with **nonsymmetric** positive definite matrices A and they have in this case asymptotically the same rate of convergence as for the symmetric case.

3) Splitting iterative methods converges also in the case of a positive semidefinite matrix A, as well as in the case of a non-linear monotone operator A.

§3 Variational inequalities with saddle matrices

3.1 Problem formulation, generalities

Example 3.1. Let us consider the minimization problem for the functional

$$J(u) = \frac{1}{2}\int\limits_0^1 u'^2(t)dt - \int\limits_0^1 b(t)\,u(t)dt, \quad b(t) \in C[0,1],$$

on the set $\{u(t) \in H_0^1(0,1) : |u'(t)| \leqslant 1 \ \text{ for } t \in (0,1)\}$. Let $\{t_i = i\,h,\, i = 0,\ldots,n+1;\, (n+1)\,h = 1\}$ be a uniform grid with meshsize $h > 0$ on the segment $[0,1]$, $u_i = u(t_i)(u_0 = u_{n+1} = 0)$ and $b_i = b(t_i)$. Finite difference scheme for the problem under consideration is

$$u^* = \arg\min_{u\in K} F(u) = \frac{1}{2}\left(\frac{u_1^2}{h^2} + \sum_{i=1}^{n-1}\left(\frac{u_{i+1}-u_i}{h}\right)^2 + \frac{u_n^2}{h^2}\right) - \sum_{i=1}^{n} b_i u_i, \qquad (3.1)$$

where $K = \left\{\dfrac{|u_i - u_{i-1}|}{h} \leqslant 1 \ \forall i = 1,\ldots,n+1\right\}$ is a convex and closed set.

This problem is equivalent to the variational inequality

$$(Au, v-u) \geqslant (b, v-u) \ \forall v \in K \qquad (3.2)$$

with matrix

$$A = L^T L,\ L = h^{-1}\begin{pmatrix} 1 & 0 & 0 & \ldots & 0 & 0 \\ -1 & 1 & 0 & \ldots & 0 & 0 \\ 0 & -1 & 1 & \ldots & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & -1 & 1 \\ 0 & 0 & 0 & \ldots & 0 & -1 \end{pmatrix} \in \mathbb{R}^{(n+1)\times n}.$$

As we know from the previous examples, matrix A is symmetric and positive definite, its spectrum is also known. Thus, all results on the convergence and rate of convergence of the considered above iterative methods are still valid.

However, when implementing all these (non-preconditioned) iterative methods we have to solve an inclusion, which reduces to the projection on the set K. And the projection on the set $K = \left\{\dfrac{|u_i - u_{i-1}|}{h} \leqslant 1 \ \forall i = 1,\ldots,n+1\right\}$, in contrast to all previous examples, is the problem, which can not be solved directly.

Let us discuss this problem in more details. The projection of a given vector g on a closed convex set K is a minimum of $\|u-g\|^2$ over K, that is equivalent to solution of the variational inequality

$$(u, v-u) \geqslant (g, v-u) \ \forall v \in K,$$

or, to solution of the inclusion

$$u + \partial I_K(u) \ni g. \tag{3.3}$$

Let $\theta(p)$ be the indicator function of the set $\{p \in \mathbb{R}^{(n+1)} : |p_i| \leqslant 1 \ \forall i\}$, then

$$\partial I_K(u) = L^T \circ \partial\theta \circ L(u),$$

with diagonal operator $\partial\theta$, and inclusion (3.3) becomes

$$u + L^T \circ \partial\theta \circ L(u) \ni g.$$

The solution of this inclusion is a problem of almost the same complexity, as the solution of the initial variational inequality (3.2), which can be equivalently written in the form of the inclusion

$$Au + L^T \circ \partial\theta \circ L(u) \ni b.$$

□

How to solve problem (3.1)**?**

The most reasonable approach is to use Lagrange multipliers method.

Using the introduced notations we write problem (3.1) in the form: find minimum of the function

$$\frac{1}{2}\|Lu\|^2 - (b, u) + \theta(Lu), \ u \in \mathbb{R}^n. \tag{3.4}$$

Let us define a new vector $p = Lu$, then (3.4) is equivalent to

$$\min_{Lu=p} \left(\frac{1}{2}\|p\|^2 - (b, u) + \theta(p) \right).$$

We will solve this problem by Lagrange multipliers method. Namely, let Lagrange function be defined as

$$\mathcal{L}(u, p, \lambda) = \frac{1}{2}\|p\|^2 - (b, u) + \theta(p) + (Lu - p, \lambda).$$

Then saddle point of $\mathcal{L}$ is a triple (u, p, λ), satisfying the system

$$\begin{aligned} &\frac{\partial\mathcal{L}}{\partial u}(u, p, \lambda) = 0 \Leftrightarrow L^T\lambda = b, \\ &\partial_p\mathcal{L}(u, p, \lambda) \ni 0 \Leftrightarrow p + \partial\theta(p) - \lambda \ni 0, \\ &\frac{\partial\mathcal{L}}{\partial\lambda}(u, p, \lambda) = 0 \Leftrightarrow Lu = p. \end{aligned} \tag{3.5}$$

It is well-known, that if there exists saddle point (u, p, λ) of Lagrange function $\mathcal{L}$, when u is a solution of (3.4) and $p = Lu$.

System (3.5) can be written as

$$\begin{pmatrix} 0 & 0 & L^T \\ 0 & E & -E \\ L & -E & 0 \end{pmatrix} \begin{pmatrix} u \\ p \\ \lambda \end{pmatrix} + \begin{pmatrix} 0 \\ \partial\theta(p) \\ 0 \end{pmatrix} \ni \begin{pmatrix} b \\ 0 \\ 0 \end{pmatrix}, \tag{3.6}$$

where E is the unit $(n+1) \times (n+1)$ matrix. Using the notations

$$\tilde{A} = \begin{pmatrix} 0 & 0 & L^T \\ 0 & E & -E \\ L & -E & 0 \end{pmatrix}, \ \tilde{P} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & \partial\theta & 0 \\ 0 & 0 & 0 \end{pmatrix}, \ \tilde{u} = \begin{pmatrix} u \\ p \\ \lambda \end{pmatrix}, \ \tilde{b} = \begin{pmatrix} 0 \\ b \\ 0 \end{pmatrix},$$

we can write (3.6) in "traditional" form

$$\tilde{A}\tilde{u} + \tilde{P}(\tilde{u}) \ni \tilde{b}.$$

Now operator $\tilde{P}$ is diagonal, matrix $\tilde{A}$ is symmetric. But it is not definite positive, it has both positive and negative eigenvalues. Such kind of matrices we will call "saddle matrices".

If we change the sign in the last equation of system (3.6), then it becomes

$$\begin{pmatrix} 0 & 0 & L^T \\ 0 & E & -E \\ -L & E & 0 \end{pmatrix} \begin{pmatrix} u \\ p \\ \lambda \end{pmatrix} + \begin{pmatrix} 0 \\ \partial\theta(p) \\ 0 \end{pmatrix} \ni \begin{pmatrix} b \\ 0 \\ 0 \end{pmatrix}. \tag{3.7}$$

The matrix

$$\mathcal{A} = \begin{pmatrix} 0 & 0 & L^T \\ 0 & E & -E \\ L & -E & 0 \end{pmatrix}$$

of the system (3.7) is not symmetric, but it is positive semidefinite:

$$(\mathcal{A}\tilde{u}, \tilde{u}) = \|p\|^2 \geqslant 0.$$

Below we will use both equivalent writing of the problem, (3.6) and (3.7), when constructing the iterative solution methods.

Now, let

$$A = \begin{pmatrix} 0 & 0 \\ 0 & E \end{pmatrix}, \ B = \begin{pmatrix} -L & E \end{pmatrix}, \ \partial\varphi = \begin{pmatrix} 0 & 0 \\ 0 & \partial\theta \end{pmatrix}, f = \begin{pmatrix} b \\ 0 \end{pmatrix}.$$

Then system (3.6) reads as

$$\begin{pmatrix} A & -B^T \\ -B & 0 \end{pmatrix} \begin{pmatrix} u \\ \lambda \end{pmatrix} + \begin{pmatrix} \partial\varphi(u) \\ 0 \end{pmatrix} \ni \begin{pmatrix} f \\ 0 \end{pmatrix}.$$

Theorem 3.1. *Let φ be a convex, proper and lower semicontinuous function,*

$$B \in \mathbb{R}^{s \times n}, s \leqslant n, \ \textit{is a matrix of full rank:} \ \operatorname{rank} B = s, \tag{3.8}$$

$$\{u \in \mathbb{R}^n : Bu = g\} \cap \operatorname{int} \operatorname{dom} \varphi \neq \emptyset \tag{3.9}$$

and of the following assumptions holds:

$$(Au, u) \geqslant m\|u\|^2 \quad \forall u \in \mathbb{R}^n,\ m > 0,\ \textit{or}$$

$$A = A^T \geqslant 0 \textit{ and } (Au, u) \geqslant m\|u\|^2 \quad \forall u \in \operatorname{Ker} B,\ \ m > 0.$$

Then problem

$$\begin{pmatrix} A & -B^T \\ -B & 0 \end{pmatrix} \begin{pmatrix} u \\ \lambda \end{pmatrix} + \begin{pmatrix} \partial\varphi(u) \\ 0 \end{pmatrix} \ni \begin{pmatrix} f \\ -g \end{pmatrix} \tag{3.10}$$

has a solution (u^*, λ^*)*, and its first component* u^* *is unique.*

Further we denote by $X = \{(u, \lambda)\}$ the set of the solutions of (3.10).

3.2 Stationary one-step iterative methods.

3.2.1 Uzawa-type method

Let A be positive definite, then inclusion $Au + \partial\varphi(u) \ni f$ has a unique solution and from the first equation in (3.10) we find u:

$$u = (A + \partial\varphi)^{-1}(B^T\lambda + f),$$

therefore λ satisfies equation

$$B \circ (A + \partial\varphi)^{-1}(B^T\lambda + f) = g. \tag{3.11}$$

Consider iterative method for solving (3.11):

$$\frac{1}{\tau} D(\lambda^{k+1} - \lambda^k) + B \circ (A + \partial\varphi)^{-1}(B^T\lambda^k + f) = g, \tag{3.12}$$

where D is a symmetric and positive definite matrix.

Its implementation consists of the sequential solution of the following problems:

$$Au^k + \partial\varphi(u^k) \ni B^T\lambda^k + f, \tag{3.13}$$

$$D\lambda^{k+1} = D\lambda^k + \tau(g - Bu^k). \tag{3.14}$$

Theorem 3.2. *Let assumptions* (3.8), (3.9) *be fulfilled and matrix* A *is positive definite:* $(Au, u) \geqslant m\|u\|^2$*. Then for*

$$D = D^T > \frac{\tau}{2m} BB^T \tag{3.15}$$

iterative method (3.12) *converges in the sense that* $(u^k, \lambda^k) \to (u^*, \lambda^*) \in X$.

Remark 3.1. *In case $A = A^T$ and for $D = E$ method* (3.13), (3.14) *coincides with Uzawa method for finding a saddle point of Lagrange function, corresponding to problem* (3.10)*:*

$$\mathcal{L}(u, \lambda) = \frac{1}{2}(Au, u) + \varphi(u) - (Bu, \lambda) - \psi(\lambda).$$

Uzawa method is the gradient method for finding $\max_{\lambda} \min_{u} \mathcal{L}(u, \lambda)$*, and it is as follows. For a known* λ^k *find* u^k *as a minimum of* $\mathcal{L}(u, \lambda^k)$ *(coincides with* (3.13)*), then execute one step of gradient method for finding* $\max_{\lambda} \mathcal{L}(u^k, \lambda)$ *(coincides with* (3.14)*).*

The main deficiency of method (3.12) is that at every iteration one needs to solve inclusion (3.13) when implementing. This inclusion (or variational inequality) with matrix A can be of great complexity in solution.

To avoid this deficiency we consider so-called Arrow-Hurwicz-type methods.

3.2.2 Arrow-Hurwicz-type methods

Consider the following iterative method for solving problem (3.10):

$$\begin{gathered} \frac{1}{\tau} D_u(u^{k+1} - u^k) + Au^k - B^T\lambda^k + \partial\varphi(u^{k+1}) \ni f, \\ \frac{1}{\tau} D_\lambda(\lambda^{k+1} - \lambda^k) + Bu^{k+1} = g \end{gathered} \tag{3.16}$$

with positive definite and symmetric matrices D_u and D_λ.

Theorem 3.3. *Let assumptions* (3.8), (3.9) *be fulfilled and matrix A be positive definite:* $(Au, u) \geqslant m\|u\|^2$*. Let further D_u and D_λ be symmetric and positive definite matrices, and*

$$(Au, v) \leqslant M^{1/2}(Av, v)^{1/2}\|u\|_{D_u} \ \forall u, v \in \mathbb{R}^n. \tag{3.17}$$

then for

$$\tau < \frac{2m}{mM + \|B\|_*^2}, \quad \|B\|_* = \sup_{u \neq 0, \lambda \neq 0} \frac{(Bu, \lambda)}{\|u\|_{D_u}\|\lambda\|_{D_\lambda}} \tag{3.18}$$

iterations (u^k, λ^k) *of method* (3.16) *converge to a solution* (u^*, λ^*) *of problem* (3.10)*.*

3.2.3 Applications to the mesh variational inequalities

Example 3.2. Go back to problem (3.4) from Example 3.1. Recall, it is the problem to minimize the function

$$\frac{1}{2}\|Lu\|^2 - (b, u) + \theta(Lu), \ u \in \mathbb{R}^n.$$

After introducing an artificial constraint $Lu = p$ it transforms to minimization problem

$$\min_{Lu=p} \left(\frac{1}{2}\|p\|^2 - (b,u) + \theta(p) \right),$$

which by using Lagrange multipliers method becomes the problem for finding saddle-point of the Lagrange function

$$\mathcal{L}(u,p,\lambda) = \frac{1}{2}\|p\|^2 - (b,u) + \theta(p) + (Lu - p, \lambda).$$

Its saddle-point is a triple $(u,p,\lambda$, satisfying the system

$$\begin{pmatrix} 0 & 0 & L^T \\ 0 & E & -E \\ L & -E & 0 \end{pmatrix} \begin{pmatrix} u \\ p \\ \lambda \end{pmatrix} + \begin{pmatrix} 0 \\ \partial\theta(p) \\ 0 \end{pmatrix} \ni \begin{pmatrix} b \\ 0 \\ 0 \end{pmatrix}. \tag{3.19}$$

So, it is a partial case of problem (3.10) with

$$A = \begin{pmatrix} 0 & 0 \\ 0 & E \end{pmatrix},\ B = \begin{pmatrix} -L & E \end{pmatrix},\ \partial\varphi = \begin{pmatrix} 0 & 0 \\ 0 & \partial\theta \end{pmatrix}, f = \begin{pmatrix} b \\ 0 \end{pmatrix}.$$

In our case matrix A is degenerate. Because of this we can not use Uzawa-type method (3.12) for its solution.

To avoid this deficiency of matrix A, we will do several identical transformations of (3.19). Namely, let us add to the first equation the third one, multiplying by rL^T with a positive constant r, and add to the second inclusion the third one, multiplying by r. As a result we get the system

$$\begin{pmatrix} rL^TL & -rL^T & L^T \\ -rL & (1+r)E & -E \\ L & -E & 0 \end{pmatrix} \begin{pmatrix} u \\ p \\ \lambda \end{pmatrix} + \begin{pmatrix} 0 \\ \partial\theta(p) \\ 0 \end{pmatrix} \ni \begin{pmatrix} b \\ 0 \\ 0 \end{pmatrix}. \tag{3.20}$$

Now matrix

$$A = \begin{pmatrix} rL^TL & -rL^T \\ -rL & (1+r)E \end{pmatrix}$$

is positive definite, because

$$(Ax,x) \geqslant m(r)(\|Lu\|^2 + \|p\|^2) \geqslant m(r)(\|u\|^2 + \|p\|^2),$$

where

$$m(r) = \frac{2r}{2r+1+\sqrt{4r^2+1}} > \frac{r}{2r+1}$$

is the minimal eigenvalue of the matrix $\begin{pmatrix} r & -r \\ -r & 1+r \end{pmatrix}$.

Above we used the fact, that $\|Lu\|^2 = (L^TLu, u)$, matrix L^TL corresponds to the mesh operator $-\partial\overline{\partial}$ with zero boundary conditions and its minimal eigenvalue equals to

$$\lambda_{min} = \frac{4}{h^2}\sin^2\frac{\pi h}{2} > 1.$$

Remark 3.2. *System* (3.20) *characterises saddle point of so-called augmented Lagrange function*

$$\mathcal{L}_r(u,p,\lambda) = \frac{1}{2}\|p\|^2 - (b,u) + \theta(p) + (Lu-p,\lambda) + \frac{r}{2}\|Lu-p\|^2, \;\; r>0.$$

Now, we can use Uzawa-type method (3.12) for solving (3.20). However, the implementation of this method requires the solution at each iteration the problem

$$\begin{pmatrix} rL^TL & -rL^T \\ -rL & (1+r)E \end{pmatrix}\begin{pmatrix} u \\ p \end{pmatrix} + \begin{pmatrix} 0 \\ \partial\theta(p) \end{pmatrix} \ni \begin{pmatrix} b \\ 0 \end{pmatrix},$$

which is an inclusion (or a variational inequality) with the positive definite matrix and a diagonal multivalued operator. This problem can not be solved directly, for its solution we have to use one of the iterative method (so-called internal iterations) from the previous section. Obviously, this makes the method (3.12) "weakly effective".

Let us apply to problem (3.20) Arrow-Hurwicz-type methods.

First, we consider non-preconditioned method (3.16), i. e. with $D_u = E$ and $D_\lambda = E$.

It reads as

$$\begin{aligned} &\frac{1}{\tau}(u^{k+1}-u^k) + rL^TLu^k - rL^Tp^k + L^T\lambda^k = b, \\ &\frac{1}{\tau}(p^{k+1}-p^k) - rLu^k + (1+r)p^k + \partial\theta(p^{k+1}) - \lambda^k \ni 0, \\ &\frac{1}{\tau}(\lambda^{k+1}-\lambda^k) - Lu^{k+1} + p^k = 0. \end{aligned} \tag{3.21}$$

The implementation of this method reduces to the multiplication of the matrices by the given vectors and solving the inclusion

$$p^{k+1} + \tau\partial\theta(p^{k+1}) \ni p^k + \tau(rLu^k - (1+r)p^k + \lambda^k) \equiv g^k,$$

which solution is

$$p^{k+1} = \begin{cases} -1 \text{ if } g^k \leqslant -1, \\ g^k \text{ if } -1 < g^k < 1, \\ 1 \text{ if } g^k \geqslant 1. \end{cases}$$

The convergence condition (3.18) for the iterative parameter becomes $\tau < \dfrac{2m}{mM + \|B\|^2}$. The norm

$$\|B\|^2 = \|B^T\|^2 = \lambda_{\max}(BB^T) = \lambda_{\max}(LL^T + E) = \frac{4}{h^2} + 1.$$

Here we use the fact, that the matrix LL^T corresponds to the mesh operator $-\partial\overline{\partial}$ with Neuman boundary conditions and its maximal eigenvalue equals to $\dfrac{4}{h^2}$.

Further, for all $x = (u, p)$ and $y = (v, q)$

$$(Ax, y) \leqslant r(\|Lu\| + \|p\|)(\|Lv\| + \|q\|) + \|p\|\|q\| \leqslant \frac{\text{const}}{m(r)} \left(1 + \frac{r}{h}\right) (Ax, x)^{1/2} \|y\|.$$

It means that $M \simeq \dfrac{1}{m(r)} \left(1 + \dfrac{r}{h}\right)^2$ and the convergence condition is

$$\tau < c(r, h) = \text{const}\, \frac{m(r) h^2}{r^2 + 1}.$$

□

Now, let us apply preconditioned method (3.16) with

$$D_u = \begin{pmatrix} rL^T L & 0 \\ 0 & (1 + r)E \end{pmatrix} \quad \text{and} \quad D_\lambda = E.$$

It reads as

$$\begin{aligned} &\frac{1}{\tau} rL^T L(u^{k+1} - u^k) + rL^T Lu^k - rL^T p^k + L^T \lambda^k = b, \\ &\frac{1}{\tau}(1 + r)(p^{k+1} - p^k) - rLu^k + (1 + r)p^k + \partial\theta(p^{k+1}) - \lambda^k \ni 0, \\ &\frac{1}{\tau}(\lambda^{k+1} - \lambda^k) - Lu^{k+1} + p^k = 0. \end{aligned} \tag{3.22}$$

Then implementing method (3.22) we have to solve at any iteration an equation with the matrix $L^T L$. This is a tridiagonal matrix, so, corresponding system of the linear equations can be solved by the direct methods very effectively.

Let us obtain an estimate for the iterative parameter providing the convergence.

From the inequality

$$(Ax, x) = r\|Lu - p\|^2 + \|p\|^2 \leqslant r(\|Lu\|^2 + \|p\|^2) + \|p\|^2 = (D_x x, x)$$

it follows $(Ax, y) \leqslant M^{1/2}(Ay, y)^{1/2} \|x\|_{D_x}$, so $M = 1$.

Further,

$$(Bx, \lambda) = (p - Lu, \lambda) \leqslant (\|p\| + \|Lu\|)\|\lambda\| \leqslant \sqrt{\frac{2}{r}} \|x\|_{D_u} \|\lambda\|,$$

therefore $\|B\|_*^2 \leqslant 2/r$ and the convergence condition reads as

$$\tau < \left(\frac{r}{r + 1}\right)^2.$$

Method (3.22) is much faster convergent than its non-preconditioned counterpart (3.21).□

Example 3.3. Here we describe one more possibility of the reformulation of problem (3.4), which allows to use for its solving preconditioned Uzawa-type methods.

Let us introduce once again the artificial constraint $Lu = p$ and rewrite (3.4) in the following equivalent form:

$$\min_{Lu=p} \left(\frac{1}{4}\|Lu\|^2 + \frac{1}{4}\|p\|^2 - (b,u) + \theta(p) \right). \tag{3.23}$$

The corresponding Lagrange function is

$$\mathcal{L}(u,p,\lambda) = \frac{1}{4}\|Lu\|^2 + \frac{1}{4}\|p\|^2 - (b,u) + \theta(p) + (Lu - p, \lambda)$$

and its saddle-point satisfies the system

$$\begin{pmatrix} \frac{1}{2}L^TL & 0 & L^T \\ 0 & \frac{1}{2}E & -E \\ L & -E & 0 \end{pmatrix} \begin{pmatrix} u \\ p \\ \lambda \end{pmatrix} + \begin{pmatrix} 0 \\ \partial\theta(p) \\ 0 \end{pmatrix} \ni \begin{pmatrix} b \\ 0 \\ 0 \end{pmatrix}. \tag{3.24}$$

Once again we get a partial case of problem (3.10), now with positive definite matrix

$$A = \begin{pmatrix} \frac{1}{2}L^TL & 0 \\ 0 & \frac{1}{2}E \end{pmatrix},$$

and the same, as before, function θ and matrix $B = \begin{pmatrix} -L & E \end{pmatrix}$.

This fact allows to use successfully both Uzawa and Arrow-Hurwicz methods. Let us consider a preconditioned Uzawa method:

$$\begin{aligned} &\frac{1}{2}L^TLu^{k+1} = b - L^T\lambda^k, \\ &\frac{1}{2}p^{k+1} + \partial\theta(p^{k+1}) \ni \lambda^k, \\ &\frac{1}{\tau}D(\lambda^{k+1} - \lambda^k) - Lu^{k+1} + p^{k+1} = 0 \end{aligned} \tag{3.25}$$

with matrix $D = BB^T = LL^T + E$.

As we remark above, the matrix LL^T corresponds to the mesh operator $-\partial\overline{\partial}$ with Neuman boundary conditions, so, implementation of method (3.25) includes solving at any iteration one mesh problem with Dirichlet boundary conditions (corresponds to solution a system with the matrix $\frac{1}{2}L^TL$) and one mesh problem with Neuman boundary conditions. Both are easy to solve by the direct methods.

Now,

$$m = \lambda_{min}(A) = \frac{2}{h^2}\sin^2\frac{\pi h}{2} + \frac{1}{2} > 1,$$

and from (3.15) the sufficient convergence condition for iterative parameter becomes

$$\tau \leqslant 1.$$

3.2.4 Numerical example

Let us consider the problem form Example 3.1, which is to minimise the functional

$$J(u) = \frac{1}{2}\int_0^1 u'^2(t)dt - \int_0^1 b(t)\,u(t)dt, \quad b(t) \in C[0,1],$$

over the set $\{u(t) \in H_0^1(0,1) : |u'(t)| \leqslant 1 \ \text{ for } t \in (0,1)\}$.

The finite difference approximation of this problem leads to the minimisation of the function

$$\frac{1}{2}\|Lu\|^2 - (b,u) + \theta(Lu),\ u \in \mathbb{R}^n$$

with matrix

$$A = L^T L,\ L = h^{-1}\begin{pmatrix} 1 & 0 & 0 & \dots & 0 & 0 \\ -1 & 1 & 0 & \dots & 0 & 0 \\ 0 & -1 & 1 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & -1 & 1 \\ 0 & 0 & 0 & \dots & 0 & -1 \end{pmatrix} \in \mathbb{R}^{(n+1)\times n}.$$

and $\theta(p)$ being the indicator function of the set $\{p \in \mathbb{R}^{(n+1)} : |p_i| \leqslant 1\ \forall i\}$.

We solved this problem by applying Arrow-Hurwicz method for finding a saddle point of augmented Lagrangian

$$\mathcal{L}_r(u,p,\lambda) = \frac{1}{2}\|p\|^2 - (b,u) + \theta(p) + (Lu - p, \lambda) + \frac{r}{2}\|Lu - p\|^2,\ \ r > 0,$$

and Usawa method for finding a saddle point of Lagrange function

$$\mathcal{L}(u,p,\lambda) = \frac{1}{4}\|Lu\|^2 + \frac{1}{4}\|p\|^2 - (b,u) + \theta(p) + (Lu - p, \lambda).$$

Exact solution is taken as

$$u^* = \begin{cases} x, & x < 0.25, \\ -16(x-0.25)^4 + 16(x-0.25)^3 - 6(x-0.25)^2 + x, & 0.25 < x < 0.75, \\ 1 - x, & x > 0.75 \end{cases}$$

and corresponding right-hand side

$$b = \begin{cases} 100, & x < 0.25, \\ (-u_{i-1} + 2u_i - u_{i+1})/h^2, & 0.25 < x < 0.75, \\ 100, & x > 0.75. \end{cases}$$

Stopping criterion was: $\|u - u^*\| < \varepsilon = 10^{-4}$.

Arrow-Hurwicz method for augmented Lagrangian
$r = 1$

N	51	51	51	51	101	101	101	101	501	1001
τ	0.8	0.7	0.6	0.5	0.8	0.7	0.6	0.5	0.5	0.5
$n(\varepsilon)$	76	66	58	59	76	66	58	59	59	59
$\|r_1\|_{L_2}$	0.08	0.03	0.02	0.01	0.08	0.3	0.3	0.01	0.01	0.01

$r = 10$

N	51	101	501	1001	1001	1001	1001	1001	1001	1001
τ	1	1	1	0.5	0.4	0.6	0.7	0.8	1	1.1
$n(\varepsilon)$	121	121	118	239	300	199	170	148	118	-
$\|r_1\|_{L_2}$	0.01	0.02	0.04	0.05	0.06	0.06	0.06	0.06		

$r = 0.1$

N	51	51	51	51	51	51	101	501	1001
τ	1	0.5	0.1	0.05	0.03	0.02	0.05	0.05	0.05
$n(\varepsilon)$	-	-	-	652	692	847	652	652	652
$\|r_1\|_{L_2}$	-	-	-	0.01	0.04		0.02	0.02	0.02

Uzawa method (3.25) **with preconditioner** $D = LL^T + E$**.**

N	51	101	101	101	501
τ	10	10	11	9	10
$n(\varepsilon)$	2443	8452	-	9392	greater than 100000
$\|r_1\|_{L_2}$	0.001	0.006	-	0.006	

Non-preconditioned Uzawa method (3.25)**:** $D = E$**.**

N	51	51	51	101	501	1001
τ	0.1	0.5	0.4	0.4	0.4	0.4
$n(\varepsilon)$	49	-	11	11	11	11
$\|r_1\|_{L_2}$	0.0005	-	0.05	0.04	0.05	0.05

3.3 Douglas-Rachford splitting method

3.3.1 General convergence result

Let us write problem (3.10) in the following form:

$$\begin{pmatrix} A & -B^T \\ B & 0 \end{pmatrix} \begin{pmatrix} u \\ \lambda \end{pmatrix} + \begin{pmatrix} \partial\varphi(u) \\ 0 \end{pmatrix} \ni \begin{pmatrix} f \\ g \end{pmatrix}. \tag{3.26}$$

We will solve problem (3.26) by Douglas-Rachford method:

$$\frac{1}{\tau} \begin{pmatrix} u^{k+1/2} - u^k \\ \lambda^{k+1/2} - \lambda^k \end{pmatrix} + \begin{pmatrix} A & -B^T \\ B & 0 \end{pmatrix} \begin{pmatrix} u^k \\ \lambda^k \end{pmatrix} + \begin{pmatrix} \partial\varphi(u^{k+1/2}) \\ 0 \end{pmatrix} \ni \begin{pmatrix} f \\ g \end{pmatrix}, \tag{3.27}$$

$$\frac{1}{\tau}\begin{pmatrix} u^{k+1}-u^{k+1/2} \\ \lambda^{k+1}-\lambda^{k+1/2} \end{pmatrix} + \begin{pmatrix} A & -B^T \\ B & 0 \end{pmatrix}\begin{pmatrix} u^{k+1}-u^{k} \\ \lambda^{k+1}-\lambda^{k} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}. \tag{3.28}$$

Theorem 3.4. *Let the assumptions* (3.8) *and* (3.9) *be valid, and matrix A be positive semidefinite. Then iterative method* (3.27), (3.28) *converges for any iterative parameter $\tau > 0$.*

The implementation of the first step (3.27) of the method consists of the solving an inclusion with of the operator $E + \tau\partial\varphi$. It is easy to do in case of a diagonal $\partial\varphi$.

On the second step (3.28) one needs to solve a system of the linear algebraic equations with a positive definite matrix $\begin{pmatrix} E+\tau A & -\tau B^T \\ \tau B & E \end{pmatrix}$.

3.3.2 Application to a mesh variational inequality

Example 3.4. A non-linear filtration problem

A mathematical model for a process of filtration of non-compressible liquid in a porous medium can pe formulated as the following variational inequality: find $u \in H_0^1(\Omega)$, suah that for all $v \in H_0^1(\Omega)$

$$\int\limits_\Omega \nabla u \cdot \nabla(v-u)dt + \int\limits_\Omega (|\nabla v| - |\nabla u|)dt \geqslant \int\limits_\Omega f(v-u)dt. \tag{3.29}$$

Variational inequality (3.29) is equivalent to the minimization over the space $H_0^1(\Omega)$ of the functional

$$J(u) = \frac{1}{2}\int\limits_\Omega |\nabla u|^2 dt + \int\limits_\Omega |\nabla u| dt - \int\limits_\Omega f u dt.$$

If u is a solution of the problem, then the domain Ω is divided into two subdomains:

$$\Omega_+ = \{t \in \Omega : |\nabla u(t)| > 0\} \text{ and } \Omega_0 = \{t \in \Omega : |\nabla u(t)| = 0\}.$$

In the points of Ω_+ a solution of (3.29) (in supposition that it is smooth enough) satisfies the equation

$$\operatorname{div}\left(-\nabla u(t) + \frac{\nabla u(t)}{|\nabla u(t)|}\right) = f(t).$$

Function $u(t)$ has a sense of a liquid pressure, while $\mathbf{v}(t) = -\nabla u(t) + \nabla u(t)/|\nabla u(t)|$ — a filtration velocity, which is a discontinuous function of the gradient of pressure: $|\mathbf{v}| = |\nabla u| + 1$, if $|\nabla u| > 0$, and $\mathbf{v} = 0$ for $|\nabla u| = 0$.

We approximate (3.29) by using finite element method. Let Ω be a polygon, $T_h = \{\delta_i\}_i$ be its conforming decomposition into triangles δ_i with the diameter h_i of a δ_i and $h = \max\limits_i h_i$. We suppose that the angles of all triangles δ_i are

bounded from below by a constant, independent on i. Define the space of continuous and piecewise linear functions

$$V_h^0 = \{u_h \in C(\overline{\Omega}) : u_h \in P_1 \ \forall \delta \in T_h, \ u_h(t) = 0 \, \forall t \in \partial\Omega\}$$

and the space of piecewise constant functions $W_h = \{u_h \in P_0 \ \forall \delta \in T_h\}$. The approximate problem is to find a minimum $u_h \in V_h^0$ of the function

$$J(u_h) = \frac{1}{2} \int\limits_\Omega |\nabla u_h|^2 dt + \int\limits_\Omega |\nabla u_h| dt - \int\limits_\Omega f u_h dt.$$

Let $\overline{p}_h = \nabla u_h$ for $t \in \delta_k$ and for all triangles $\delta_k \in T_h$, i. e. $p_{ih} = \partial u_h / \partial t_i \in W_h, \ i = 1, 2$. Then the equivalent formulation of the approximate problem is

$$\begin{cases} \min\limits_{(u_h, \overline{p}_h) \in K_h} \left\{ J(u_h, \overline{p}_h) = \dfrac{1}{2} \displaystyle\int\limits_\Omega |\overline{p}_h|^2 dt + \int\limits_\Omega |\overline{p}_h| dt - \int\limits_\Omega f u_h dt \right\}, \\ \\ K_h = \{(u_h, \overline{p}_h) \in V_h^0 \times (W_h)^2 : \overline{p}_h = \nabla u_h \text{ for } t \in \delta_k, \ \forall \delta_k \in T_h\}. \end{cases} \tag{3.30}$$

Let $\omega_h = \{t_i\}_{i=1}^m$ be the set of all vertices of the triangles $\delta \in T_h$, lying in Ω, $m = \operatorname{card} \omega_h$, and $\xi_h = \{t_i\}_{i=1}^s$ is the set of barycenters of $\delta \in T_h$. To a function $v_h \in V_h^0$ the vector $v \in R^m$ corresponds, which coordinates are $v_i = v_h(t_i)$, $t_i \in \omega_h$. Similarly, to a function $q_h \in W_h$ vector $q \in R^s$ with coordinates $q_i = q_h(t_i)$, $t_i \in \xi_h$ corresponds. As above, we use notations $v \Leftrightarrow v_h$, $q \Leftrightarrow q_h$ for this corresponding.

Let matrices D, L_1, L_2 and vector f be defined by:

$$(Dp, q) = \int\limits_\Omega p_h(t) q_h(t) dt, \ \ (\tilde{L}_i u, q) = \int\limits_\Omega \frac{\partial u_h}{\partial t_i}(t) q_h(t) dt, \ \ (f, v) = \int_\Omega f(t) v_h(t) dt$$

for $\mathbb{R}^m \ni u, v \Leftrightarrow u_h, v_h \in V_h^0, \ \mathbb{R}^s \ni p, q \Leftrightarrow p_h, q_h \in W_h$.

The equality $p_{ih} = \dfrac{\partial u_h}{\partial t_i}$ for all triangles $\delta_k \in T_h$ mean that $\displaystyle\int\limits_\Omega \left(\frac{\partial u_h}{\partial t_i} - p_{ih} \right) q_h dt =$ 0 for all $q_h \in W_h$, or,

$$(L_i u - D p_i, q) = 0 \ \forall q \Leftrightarrow L_i u = D p_i.$$

Let

$$L_i = D^{-1} \tilde{L}_i, \ L = (L_1, L_2)^T, \ \overline{D} = \operatorname{diag}(D, D).$$

The statement $(u_h, \overline{p}_h) \in K_h$ is equivalent to $L_i u = p_i, i = 1, 2$, or, $Lu = \overline{p}$. Further,

$$\int\limits_\Omega |\overline{p}_h| dt = \sum_{\delta_k \in T_h} d_{kk} |\overline{p}|_i = (D|\overline{p}|, 1),$$

where $D = \operatorname{diag}(d_{11}, d_{22}, \ldots, d_{ss})$. Below we denote $\varphi(\overline{p}) = (D|\overline{p}|, 1)$.

After introducing all these notations, we can write problem (3.30) as

$$\min_{Lu=\overline{p}} \left\{ f(u,\overline{p}) = \frac{1}{2}(\overline{D}\overline{p},\overline{p}) + \varphi(\overline{p}) - (f,u) \right\}. \tag{3.31}$$

Corresponding Lagrange function is

$$\mathcal{L}_r(u,\overline{p},\overline{\lambda}) = \frac{1}{2}(\overline{D}\overline{p},\overline{p}) + \varphi(\overline{p}) - (f,u) + (Lu-\overline{p},\overline{\lambda}),$$

and its saddle point satisfies the following system:

$$\begin{pmatrix} 0 & 0 & L^T \\ 0 & \overline{D} & -E \\ L & -E & 0 \end{pmatrix} \begin{pmatrix} u \\ \overline{p} \\ \overline{\lambda} \end{pmatrix} + \begin{pmatrix} 0 \\ \partial\varphi(\overline{p}) \\ 0 \end{pmatrix} \ni \begin{pmatrix} f \\ 0 \\ 0 \end{pmatrix}, \tag{3.32}$$

where E is the unit $s^2 \times s^2$ matrix.

Thus, we have a problem, which is similar to the problem of the previous example. For its solving we can use all iterative methods, described above: Uzawa and Arrow-Hurwicz methods, splitting methods.

The only difference in the implementation of these methods is that now instead of solving a system of scalar inclusions we have solve a system of two-dimensional inclusions

$$\overline{p}_i + \tau d_{ii}\partial\varphi_i(\overline{p}_i) \ni \overline{g}_i, \tag{3.33}$$

where $\overline{p}_i = (p_{1i}, p_{2i})$, $\varphi_i(\overline{p}_i) = \sqrt{p_{1i}^2 + p_{2i}^2}$ and g_i is a given vector. By definition of the subdifferential

$$\partial\varphi_i(\overline{p}_i) = \begin{cases} \overline{p}_i|\overline{p}_i|^{-1} \text{ if } \overline{p}_i \neq 0, \\ \text{closed unit ball } |\overline{p}_i| \leqslant 1 \text{ if } \overline{p}_i = 0. \end{cases}$$

Thus, the unique solution of (3.33) is given by

$$\overline{p}_i = 0, \text{ if } |\overline{g}_i| \leqslant \tau d_{ii},$$

$$|\overline{p}_i| = |\overline{g}_i| - \tau d_{ii}\ \overline{p}_i = \overline{g}_i \left(1 + \frac{\tau d_{ii}}{|\overline{p}_i|}\right)^{-1}, \text{ if } |\overline{g}_i| > \tau d_{ii}.$$

References

[1] Duvaut G., Lions J.-L., *Les inequations en mechanique et en physique,* Paris: Dunod, 1972.

[2] Kinderlehrer D., Stampacchia G, *An itroduction to variational inequalities and their applications,* N.Y.: Academic Press, 1980.

[3] Baiocchi C., Capelo A., *Variational and quasivariational inequalities. Applications to free boundary problems,* Chichester etc.: John Wiley & Sons, 1984.

[4] Friedman A., *Variational principles and free-boundary problems,* N.Y: John Wiley & Sons, 1982.

[5] Panagiotopoulos P., *Inequality problems in mechanics and applications,* Boston etc.: Birkhauser, 1985.

[6] Ekeland I., Temam R., *Convex analysis and variational problems,* North Holland, Amsterdam, 1976.

[7] Glowinski R., Lions J.-L., Tremolier R., *Analyse numerique des inequations variationnelles,* Paris, Dunod, 1976.

[8] Fortin M., Glowinski R. *Augmented Lagrangan methods,* Amsterdam, N.Y.: North- Holland, 1983.

[9] Glowinski R., LeTallec P. *Augmented Lagrangan and operator-splitting methods in nonlinear mechanics//* SIAM studies in applied mathematics, Philadelphia, 1989.

[10] Haslinger J., Hlavacek I., Necas J., *Numerical methods for unilateral problems in solids mechanics//* in: Handbook of numerical analysis, V. IV, Part 2 (edited by P.G. Ciarlet, J.L. Lions): Elsevier Science B. V., 1996.

[11] Lapin A., *Iterative solution methods for mesh variational inequalities (in Russian),* Kazan: Kazan State University, 2008.

[12] Hackbush W. *Iterative solution of large sparse systems of equations,* N.Y.: Springer Verlag, 1994.

[13] Axelsson O. *Iterative solution methods,* N.Y.: Cambridge University Press, 1996.

[14] Saad Y. *Iterative methods for sparse linear systems. Second edition,* Philadelphia, PA: SIAM, 2003.

§4 Appendix

4.1 Some notations and results from the theory of matrices and functional spaces

$\mathbb{R}$ is the space of real numbers, $\overline{\mathbb{R}} = \mathbb{R} \cup +\infty$, $x \in \mathbb{R}^n$ is an n-dimensional vector with real coordinates; $(x, y) = \sum_{i=1}^{n} x_i y_i$ is euclidian scalar product in vector space $\mathbb{R}^n$ and $\|x\| = (x, x)^{1/2}$ is the corresponding norm.

$A \in \mathbb{R}^{n \times m}$ is a rectangular $n \times m$ matrix (with n rows and m columns), $A^T \in \mathbb{R}^{m \times n}$ is its transpose matrix.

$\|A\| = \sup_{x \neq 0} \frac{\|Ax\|}{\|x\|}$ is the norm of the matrix A, subordinate to euclidian norm of the vectors;

in case of a symmetric matrix $A \in \mathbb{R}^{n \times n}$

$$\|A\| = \sup_{x \neq 0} \frac{(Ax, x)}{\|x\|^2} = \max_{1 \leqslant i \leqslant n} |\lambda_i(A)|,$$

where $\lambda_i(A)$ are eigenvalues of A.

In general case of a rectangular $n \times m$ matrix A, for its norm subordinated to euclidian norm of the vectors, the following equalities are true:

$$\|A\| = \max_{1 \leqslant i \leqslant n} \sqrt{\lambda_i(AA^T)} = \max_{1 \leqslant i \leqslant m} \sqrt{\lambda_i(A^T A)} = \|A^T\|.$$

$\|x\|_A = (Ax, x)^{1/2}$ is energetic norm of a vector x in case when A is a symmetric and positive definite matrix.

$\rho(A) = \max_{1 \leqslant i \leqslant n} |\lambda_i(A)|$ is the spectral radius of a matrix $A \in \mathbb{R}^{n \times n}$.

For any norm of a matrix $\|A\|$, which is subordinate to a norm of the vectors,

$$\rho(A) = \lim_{k \to \infty} \|A^k\|^{1/k} \leqslant \|A\|.$$

For any $\varepsilon > 0$ there exists a norm $\|.\|_*$ in $\mathbb{R}^n$, such that for a corresponding subordinate norm of a matrix $A \in \mathbb{R}^{n \times n}$ the inequality $\|A\|_* \leqslant \rho(A) + \varepsilon$ is true.

For any matrix $A \in \mathbb{R}^{n \times n}$ its spectral radius $\rho(A) < 1$ if and only if $\lim_{k \to \infty} A^k = 0$.

Rank of a matrix $A \in \mathbb{R}^{n \times m}$, denoted by rank$A$, is the maximal degree of a nonzero minor of A; rank$A \leqslant \min(n, m)$.

If rank$A = \min(n, m)$, then matrix A is called as matrix of full rank.

Ω is a bounded domain in $\mathbb{R}^2$ with a piecewise smooth boundary $\partial\Omega$.

$\nabla u = (\dfrac{\partial u}{\partial x_1}, \dfrac{\partial u}{\partial x_2})^T$ is the gradient of a function u; $\Delta u = \dfrac{\partial^2 u}{\partial x_1^2} + \dfrac{\partial^2 u}{\partial x_2^2}$ is a value of Laplace operator Δ on a function u.

$L_2(\Omega)$ is Lesbegue space of the functions $u(x)$, $x \in \Omega$, such that $u^2(x)$ are integrable in Ω; $L_2(\Omega)$ is a Hilbert space with the scalar product $(u, v) = \int\limits_\Omega u(x)v(x)dx$ and the corresponding norm $\|u\| = \left(\int\limits_\Omega u^2(x)\,dx\right)^{1/2}$.

$H^1(\Omega)$ is Sobolev space of the functions $u \in L_2(\Omega)$, which have first order generalized (weak) derivatives $\dfrac{\partial u}{\partial x_i} \in L_2(\Omega)\ \forall i$. $H^1(\Omega)$ is a Hilbert space with the scalar product $(u, v) = \int\limits_\Omega (\nabla u(x) \cdot \nabla v(x) + u(x)v(x))\,dx$ and the corresponding norm $\|u\| = \sqrt{(u, u)} = \left(\int\limits_\Omega (|\nabla u(x)|^2 + u^2(x))\,dx\right)^{1/2}$.

$H_0^1(\Omega) \subset H^1(\Omega)$ is the subspace of $H^1(\Omega)$, such that functions from $H_0^1(\Omega)$ vanish on the boundary $\partial\Omega$ (their traces on the boundary equal to zero). The H^1-norm of the space $H_0^1(\Omega)$ is equivalent to the norm

$$\|u\|_0 = \left(\int\limits_\Omega |\nabla u(x)|^2\,dx\right)^{1/2},$$

i. e. there exists a constant c (depending only on the domain Ω), such that for all $u \in H_0^1(\Omega)$

$$\left(\int\limits_\Omega |\nabla u(x)|^2\,dx\right)^{1/2} \leqslant \left(\int\limits_\Omega (|\nabla u(x)|^2 + u^2(x))\,dx\right)^{1/2} \leqslant c\left(\int\limits_\Omega |\nabla u(x)|^2\,dx\right)^{1/2}.$$