COMPUTATIONAL METHODS IN APPLIED MATHEMATICS, Vol. 10 (2010), No. 3, pp. 283–301

ON THE ITERATIVE SOLUTION METHODS FOR FINITE-DIMENSIONAL INCLUSIONS WITH APPLICATIONS TO OPTIMAL CONTROL PROBLEMS

E. LAITINEN[1], A. LAPIN[2], AND S. LAPIN[3]

Abstract — Iterative methods for finite-dimensional inclusions which arise in applying a finite-element or a finite-difference method to approximate state-constrained optimal control problems have been investigated. Specifically, problems of control on the right-hand side of linear elliptic boundary value problems and observation in the entire domain have been considered. The convergence and the rate of convergence for the iterative algorithms based on the finding of the control function or Lagrange multipliers are proved.

2000 Mathematics Subject Classification: 65K15, 65N30, 49M29, 49M30.

Keywords: constrained saddle-point problem, inclusion, state-constrained optimal control problem, finite element approximation, iterative methods.

Introduction

Large-scale finite dimensional inclusions with so-called saddle matrices and constrained saddle-point problems arise from the approximation of different applied problems. While the solution methods for unconstrained saddle-point problems have been thoroughly investigated (see, e.g., the survey paper [1] containing an exhaustive list of references on this subject), the development of efficient numerical methods for solving large-scale constrained saddle-point problems is still far from completed. For instance, the convergence of the Uzawa, Arrow-Hurwitz, and operator-splitting iterative methods for constrained saddle-point problems arising from an augmented Lagrangian approach to solving variational inequalities was investigated in [2] (see also the bibliography therein). Some iterative methods with the estimation of the rate of convergence for constrained saddle-point problems arising from a mixed hybrid finite element approximation of variational inequalities were proposed in [3].

State-constrained optimal control of systems governed by partial differential equations give rise to a class of constrained saddle-point problems, which causes problems to the the optimization methods (see, e.g., [4, 5, 6]). A common way to solve them consists of the approximation of the indicator function of the set of state constraints with further application of a gradient-type or Newton-type method [6]– [9].

In this paper, we develop iterative solution methods for constrained saddle-point problems and pay attention to the obtaining of estimates for the iterative parameters and the rate

[1] *Oulu University,* Oulu, Finland. E-mail: ejl@sun3.oulu.fi

[2] *Kazan State University,* Kazan, Russia. E-mail: alapin@ksu.ru

[3] *Washington State University,* Pullman, WA, USA. E-mail: slapin@wsu.edu

of convergence. As an illustrative example, we consider the finite element approximation of the state- and control-constrained optimal control problem governed by a linear elliptic partial differential equation. In this problem, the control is on the right-hand side of the equation or the Neumann boundary condition, and the observation - in the entire domain. Regardless of the fact that this problem seems to be very particular specific, it is of practical importance, especially since it can serve as an auxiliary problem in sequential quadratic programming (SQP)-methods (see [10]) and other preconditioning procedures to solve more general problems.

1. Solution of the finite dimensional inclusion

1.1. Existence results

Consider the minimization problem

$$\text{find } \min_{Ly=Su} \left\{ J(y,u) = \Theta(y) + \Upsilon(u) \right\}, \tag{1.1}$$

where

$$\begin{gathered} \Theta : \mathbb{R}^{N_y} \to \bar{\mathbb{R}} \text{ and } \Upsilon : \mathbb{R}^{N_u} \to \bar{\mathbb{R}} \quad (\bar{\mathbb{R}} = \mathbb{R} \cup \{+\infty\}) \text{ are proper,} \\ \text{convex and lower semicontinuous functions with closed domains} \\ \operatorname{dom}\Theta = \{y : \Theta(y) < +\infty\},\ \operatorname{dom}\Upsilon = \{u : \Upsilon(u) < +\infty\}, \end{gathered} \tag{1.2}$$

$$L \in \mathbb{R}^{N_y \times N_y},\ S \in \mathbb{R}^{N_y \times N_u} \text{ are matrices and } L \text{ is regular.} \tag{1.3}$$

Theorem 1.1. *Let* (1.2), (1.3) *be satisfied. Additionally, if*

$$\textit{there exists a pair } (y_0, u_0) \in \operatorname{dom}\Theta \times \operatorname{dom}\Upsilon : \quad Ly_0 = Su_0. \tag{1.4}$$

and one of the following assumptions holds:

$$\operatorname{dom}\Upsilon \textit{ is bounded,} \tag{5a}$$

$$\Upsilon \textit{ is coercive and } \Theta \textit{ is bounded below,} \tag{5b}$$

then problem (1.1) *has a solution.*

If moreover one of the following assumptions is satisfied:

$$\Upsilon \textit{ is strictly convex,} \tag{6a}$$

$$\Theta \textit{ is strictly convex and } \operatorname{Ker}S = \{0\}, \tag{6b}$$

then the solution is unique.

Proof. Owing to (1.2) – (1.4) the set $K = \{(y,u) : y \in \operatorname{dom}\Theta,\ u \in \operatorname{dom}\Upsilon,\ Ly = Su\}$ is closed, convex, and nonempty, while the function J is proper, convex, and lower semicontinuous. If domΥ is bounded, then K is bounded because of the inequality $\|y\| \leqslant \|L^{-1}\|\|S\|\|u\|$, and the function J attains its minimum on K. Let now Υ be coercive: $\lim \Upsilon(u) = +\infty$ as $u \in$ domΥ, $\|u\| \to \infty$, and Θ be bounded below: $\Theta(y) \geqslant \theta_0 = \text{const}$ for all y. Then J is coercive on K:

$$J(y_n, u_n) \to +\infty \text{ for } \{(y_n, u_n)\} \in K, \|y_n\| + \|u_n\| \to \infty.$$

In fact, if $\|y_n\| + \|u_n\| \to \infty$, then necessarily $\|u_n\| \to \infty$, and $J(y_n, u_n) \geqslant \theta_0 + \Upsilon(u_n) \to +\infty$. The proven properties of K and J ensure the existence of a solution (y, u) of problem (1.1).

To ascertain the uniqueness of the solution, we need only to prove the strict convexity of the function J on K.

Let Υ be strictly convex (assumption (6a)). If $(y_1, u_1) \in K$, $(y_2, u_2) \in K$ and $(y_1, u_1) \neq (y_2, u_2)$, then $u_1 \neq u_2$, because otherwise $y_1 - y_2 = L^{-1}S(u_1 - u_2) = 0$ and we get a contradiction. So, J is strictly convex on K as a sum of strictly convex Υ and convex Θ.

Let Θ be strictly convex and Ker$S = \{0\}$ (assumption (6b)). If $(y_1, u_1) \neq (y_2, u_2)$, then $y_1 \neq y_2$. Indeed, the equality $y_1 = y_2$ implies $S(u_1 - u_2) = L(y_1 - y_2) = 0$, whence $u_1 = u_2$ since Ker$S = \{0\}$. Again, J is strictly convex on K as a sum of strictly convex Θ and convex Υ. □

Now, define the Lagrange function for problem (1.1)

$$\mathcal{L}(y, u, \lambda) = \Theta(y) + \Upsilon(u) - (\lambda, Ly - Su). \tag{1.7}$$

The saddle-point of $\mathcal{L}$ is a triple $(y, u, \lambda) \in \mathbb{R}^{N_y} \times \mathbb{R}^{N_u} \times \mathbb{R}^{N_y}$ such that

$$\inf_{y,u} \sup_{\lambda} \mathcal{L}(y, u, \lambda) = \sup_{\lambda} \inf_{y,u} \mathcal{L}(y, u, \lambda).$$

It is known (cf. [11]) that the first two components $(y,\, u)$ of the saddle-point coincide with the solution of (1.1) and that (y, u, λ) is a saddle-point of Lagrangian (1.7) if and only if it is a solution of the system

$$\begin{aligned} \partial_y \mathcal{L}(y, u, \lambda) &= \partial\Theta(y) - L^T\lambda \ni 0, \\ \partial_u \mathcal{L}(y, u, \lambda) &= \partial\Upsilon(u) + S^T\lambda \ni 0, \\ \nabla_\lambda \mathcal{L}(y, u, \lambda) &= -Ly + Su = 0. \end{aligned} \tag{1.8}$$

Theorem 1.2. *Let* (1.2), (1.3) *and (5a-5b) be satisfied. Let also one of the following assumptions hold:*

$$\textit{there exists a pair } (y_0, u_0) \in \operatorname{int}\operatorname{dom}\Theta \times \operatorname{dom}\Upsilon : \quad Ly_0 = Su_0, \tag{1.9}$$

$$\begin{aligned} &\textit{there exists a pair } (y_0, u_0) \in \operatorname{dom}\Theta \times \operatorname{int}\operatorname{dom}\Upsilon : \quad Ly_0 = Su_0, \\ &\qquad \textit{and there exists } u_1 \in \mathbb{R}^{N_u} : y_1 = L^{-1}Su_1 \in \operatorname{int}\operatorname{dom}\Theta. \end{aligned} \tag{1.10}$$

Then there exists a saddle-point (y, u, λ) *of Lagrangian* (1.7). *The components* (y, u) *are defined uniquely if (6a-6b) holds. If, moreover,* Θ *is differentiable at the point* y *or if* Υ *is differentiable at the point* u *and there exists an inverse matrix* S^{-T}*, then* λ *is defined uniquely.*

Proof. Due to Theorem 1.1 the minimization problem (1.1) has a solution u, which is unique if (6a-6b) holds. Problem (1.1) can be written in the form

$$\text{find} \quad \min_{u \in \mathbb{R}^{N_u}} \left[\Theta(L^{-1}Su) + \Upsilon(u)\right],$$

which is equivalent to finding a solution of the inclusion $0 \in \partial\left[\Theta(L^{-1}Su) + \Upsilon(u)\right]$. Because of (1.9) or (1.10) (see the properties of subdifferentials in [11])

$$\begin{aligned} \partial\left[\Theta(L^{-1}Su) + \Upsilon(u)\right] &= \partial\left[\Theta(L^{-1}Su)\right] + \partial\Upsilon(u) \\ &= \big(S^T L^{-T} \circ \partial\Theta \circ L^{-1} S\big)(u) + \partial\Upsilon(u). \end{aligned}$$

Thus, u is the solution of the inclusion

$$\big(S^T L^{-T} \circ \partial\Theta \circ L^{-1} S\big)(u) + \partial\Upsilon(u) \ni 0. \tag{1.11}$$

In other words,

$$\exists \mu \in \partial\Upsilon(u),\ \exists \eta \in \partial\Theta(y) \text{ with } y = L^{-1}Su : S^T L^{-T}\eta + \mu = 0. \tag{1.12}$$

Denoting $\lambda = L^{-T}\eta$, we find that the triple $(y, u, \lambda) \in \mathbb{R}^{N_y} \times \mathbb{R}^{N_u} \times \mathbb{R}^{N_y}$ satisfies system (1.8). So, the existence of a saddle-point of Lagrangian (1.7) is proved.

Now, if the triple (y, u, λ) is a solution of (1.8), then it satisfies system (1.12). As a consequence, u is the solution of (1.11) and the pair (u, y), $y = L^{-1}Su$ is a solution to problem (1.1). So, with assumptions (6a-6b) (u, y) is defined uniquely owing to Theorem 1.1.

It remains to prove the uniqueness of λ. But if Θ is differentiable at the point y, then λ is defined uniquely from the first equation of (1.8): $\lambda = L^{-T}\nabla\Theta(y)$. Similarly, if there exist $\nabla\Upsilon(u)$ and S^{-T}, then $\lambda = -S^{-T}\nabla\Upsilon(u)$ from the second equation of (1.8). □

1.2. Iterative solution of the constrained saddle-point problem

Let system (1.8) have a solution (y, u, λ). We will consider the iterative methods for the inclusions constructed via transformations of system (1.8).

Case of the single-valued operator $\partial\Theta$

Let $\partial\Theta = \nabla\Theta$ be a single-valued operator, then from system (1.8) we can obtain the inclusion with respect to the vector u

$$\big(S^T L^{-T} \circ \nabla\Theta \circ L^{-1} S\big)(u) + \partial\Upsilon(u) \ni 0. \tag{1.13}$$

To solve it, we apply the stationary one-step iterative method

$$\frac{1}{\tau} B(u^{k+1} - u^k) + \big(S^T L^{-T} \circ \nabla\Theta \circ L^{-1} S\big)(u^k) + \partial\Upsilon(u^{k+1}) \ni 0, \tag{1.14}$$

where $B \in \mathbb{R}^{N_u \times N_u}$, $B = B^T > 0$ and $\tau > 0$. The iterative method (1.14) can be viewed as a preconditioned gradient-type method for finding the minimum of the function

$$\Theta(L^{-1}Su) + \Upsilon(u)$$

with the differentiable function Θ and nondifferentiable Υ. Its implementation consists of the following steps: for known u^k
1) find $y^k = L^{-1}Su^k$;
2) find $\lambda^k = -L^{-T}\nabla\Theta(y^k)$;
3) solve the inclusion

$$B\frac{u^{k+1} - u^k}{\tau} + \partial\Upsilon(u^{k+1}) \ni S^T\lambda^k. \tag{1.15}$$

Note that the choice of the preconditioner B is limited to the possibility to solve efficiently inclusion (1.15).

In the case $\partial\Upsilon = \partial\psi + M_u$ with a single-valued operator M_u and a convex, proper, and lower semicontinuous function ψ, we can consider the variant of method (1.14)

$$\frac{1}{\tau} B(u^{k+1} - u^k) + \big(S^T L^{-T} \circ \nabla\Theta \circ L^{-1} S\big)(u^k) + M_u(u^k) + \partial\psi(u^{k+1}) \ni 0.$$

In implementing this iterative method, we need to have an efficient solver for an inclusion with the operator $B + \tau\psi$.

Case of the single-valued operators $(\partial\Theta)^{-1}$ and $(\partial\Upsilon)^{-1}$

Suppose there exist single-valued operators $(\partial\Theta)^{-1}$ and $(\partial\Upsilon)^{-1}$. Then from system (1.8) we get the equation for λ

$$L\,(\partial\Theta)^{-1}(L^T\lambda) - S\,(\partial\Upsilon)^{-1}(-S^T\lambda) = 0 \tag{1.16}$$

with a single-valued operator $P = L \circ (\partial\Theta)^{-1} \circ L^T + (-S) \circ (\partial\Upsilon)^{-1} \circ (-S^T)$. To solve (1.16), we use the iterative method

$$B\frac{\lambda^{k+1} - \lambda^k}{\tau} + L\,(\partial\Theta)^{-1}(L^T\lambda^k) - S\,(\partial\Upsilon)^{-1}(-S^T\lambda^k) = 0 \tag{1.17}$$

with a preconditioner $B = B^T > 0$. Method (1.17) can be viewed as a preconditioned Uzawa method for finding the saddle point of Lagrangian (1.7). Its implementation consists of the following two steps: for known λ^k
1) find

$$y^k = (\partial\Theta)^{-1}(L^T\lambda^k) \text{ and } u^k = (\partial\Upsilon)^{-1}(-S^T\lambda^k);$$

2) solve the equation

$$B\frac{\lambda^{k+1} - \lambda^k}{\tau} + Ly^k - Su^k = 0.$$

Obviously, method (1.17) is of practical importance if inclusions with the operators $\partial\Theta$ and $\partial\Upsilon$ can be solved efficiently (first step of the algorithm). On the other hand, at the second step of the algorithm we solve the equation with a matrix B, so we can use a variety of preconditioners B.

Case of the single-valued operator $\partial\Upsilon$ and the regular matrix S

Let $\partial\Upsilon = \nabla\Upsilon$ and the matrix S be regular. Then system (1.8) can be transformed to the inclusion with respect to y

$$L^T\,S^{-T}\nabla\Upsilon\big(S^{-1}Ly\big) + \partial\Theta(y) \ni 0. \tag{1.18}$$

The stationary one-step iterative method for (1.18) reads as

$$B\frac{y^{k+1} - y^k}{\tau} + L^T\,S^{-T}\nabla\Upsilon\big(S^{-1}Ly^k\big) + \partial\Theta(y^{k+1}) \ni 0. \tag{1.19}$$

The iterative method (1.19) can be viewed as a preconditioned gradient-type method for finding the minimum of the function

$$\Theta(y) + \Upsilon(S^{-1}Ly)$$

with a differentiable function Υ and a nondifferentiable function Θ. Its implementation consists of the following steps: for known y^k
1) find $u^k = S^{-1}Ly^k$;
2) find $\lambda^k = -S^{-T}\nabla\Upsilon(u^k)$;
3) solve the inclusion

$$B\frac{y^{k+1} - y^k}{\tau} + \partial\Theta(y^{k+1}) \ni L^Ty^k.$$

If $\partial\Theta = \partial\theta + M_y$ with a single-valued operator M_y and a convex, proper and lower semicontinuous function θ, then we can consider the following variant of method (1.19):

$$B\frac{y^{k+1} - y^k}{\tau} + L^T\, S^{-T}\nabla\Upsilon\big(S^{-1}Ly^k\big) + A_y(y^k) + \partial\theta(y^{k+1}) \ni 0.$$

In implementing these iterative methods we need to have an efficient solver for the inclusion with the operator $B + \tau\Theta$ or with the operator $B + \tau\theta$, respectively.

1.3. Iterative methods for the general inclusion

The inclusions constructed in Section 1.2 for the vectors u, λ and y are particular cases of the general inclusion which we will consider in this section.

Consider the problem in $\mathbb{R}^n$

$$P(u) + Q(u) \ni 0 \tag{1.20}$$

with a (generally) multivalued maximal monotone operator Q and a continuous operator P. Further we suppose that inclusion (1.20) has a solution and apply for its solution the preconditioned one-step stationary iterative method

$$\frac{1}{\tau}B(u^{k+1} - u^k) + P(u^k) + Q(u^{k+1}) \ni 0, \tag{1.21}$$

with the matrix $B = B^T > 0$ and the iterative parameter $\tau > 0$.

Theorem 1.3. *Let $Q : \mathbb{R}^n \to 2^{\mathbb{R}^n}$ be a maximal monotone operator and $P = C^T \circ A \circ C$, where $C \in \mathbb{R}^{m\times n}$ and the operator $A : \mathbb{R}^m \to \mathbb{R}^m$ is uniformly inverse monotone (co-coercive)*[4]

$$(A(u) - A(v), u - v) \geqslant p_0\|A(u) - A(v)\|^2,\ p_0 > 0. \tag{1.22}$$

Then for

$$B = B^T > \frac{\tau}{2p_0}C^TC \tag{1.23}$$

the iterative method (1.21) *converges for any initial guess $u^0 \in \mathbb{R}^n$.*

Proof. Let u be a solution of (1.20), $z^k = u^k - u$. Multiplying the inclusion

$$\frac{1}{\tau}B(z^{k+1} - z^k) + P(u^k) - P(u) + Q(u^{k+1}) - Q(u) \ni 0$$

by $2\tau z^{k+1}$ and using the monotonicity of Q, we get

$$\|z^{k+1}\|_B^2 - \|z^k\|_B^2 + \|z^{k+1} - z^k\|_B^2 + 2\tau(A(Cu^k) - A(Cu), Cz^{k+1}) \leqslant 0. \tag{1.24}$$

Due to (1.22)

$$\begin{gathered}(A(Cu^k) - A(Cu), Cz^{k+1}) = (A(Cu^k) - A(Cu), Cz^k)\\ +(A(Cu^k) - A(Cu), C(u^{k+1} - u^k)) \geqslant p_0\|A(Cu^k) - A(Cu)\|^2\\ -\|A(Cu^k) - A(Cu)\|\|C(u^{k+1} - u^k)\| \geqslant -\frac{1}{4p_0}\|C(z^{k+1} - z^k)\|^2.\end{gathered} \tag{1.25}$$

[4] Hereafter we use the same notations $(.,.)$ and $\|.\|$ for Euclidian scalar products and norms in vector spaces of different dimensions.

Inequalities (1.24) and (1.25) yield

$$\|z^{k+1}\|_B^2 - \|z^k\|_B^2 + ((B - \frac{\tau}{2p_0}C^TC)(z^{k+1} - z^k), z^{k+1} - z^k) \leqslant 0.$$

Since there exists $\varepsilon > 0$ such that $B - \dfrac{\tau}{2p_0}C^TC \geqslant \varepsilon B$, then

$$\|z^{k+1}\|_B^2 + \varepsilon\|z^{k+1} - z^k\|_B^2 \leqslant \|z^k\|_B^2 \text{ for all } k. \tag{1.26}$$

Inequality (1.26) brings about the following statements:

(a) the sequence $\{\|z^k\|_B\}$ monotonically decreases and converges to a finite number, i.e, the sequence $\{u^k\}$ is bounded;

(b) $\|u^{k+1} - u^k\|_B = \|z^{k+1} - z^k\|_B \to 0$ as $k \to \infty$.

Let $u^{k_i} \to u^*$ for $k_i \to \infty$ be a convergent subsequence of the bounded sequence $\{u^k\}$. As $\|u^{k_i+1} - u^{k_i}\|_B \to 0$, then also $u^{k_i+1} \to u^*$ for $k_i \to \infty$.

Let us prove that u^* is a solution of (1.20). Recall that the maximal monotone operator Q is closed: $u^k \to u^*$ and $\gamma^k \to \gamma^*$, $\gamma^k \in Q(u^k)$ imply $\gamma^* \in Q(u^*)$.

Because of this property and the continuity of P, passing to the limit in the inclusion $Q(u^{k_i+1}) \ni -P(u^{k_i+1}) - \dfrac{1}{\tau}B(u^{k_i+1} - u^{k_i})$, we obtain $Q(u^*) \ni P(u^*)$. It means that u^* is a solution of (1.20).

Now let $u = u^*$ in all aforementioned arguments. Since the sequence $\{\|u^k - u^*\|_B\}$ monotonically decreases and its subsequence $\{\|u^{k_i} - u^*\|_B\}$ tends to zero, $\{\|u^k - u^*\|_B\}$ also tends to zero. □

Theorem 1.4. *Let $B = B^T > 0$, Q be the maximal monotone operator while P be a uniformly monotone and Lipshitz-continuous operator*

$$(P(u) - P(v), u - v) \geqslant \alpha\|u - v\|_B^2, \tag{1.27}$$

$$(P(u) - P(v), w) \leqslant \beta^{1/2}\,(P(u) - P(v), u - v)^{1/2}\|w\|_B. \tag{1.28}$$

Inclusion (1.20) *has a unique solution u, for $\tau \in (0, \dfrac{2}{\beta})$ the iterative method* (1.21) *converges starting from any initial guess u^0, and for the optimal parameter*

$$\tau = \tau_0 = \frac{1}{\beta}$$

the following estimate for the rate of convergence is valid:

$$\|u^{k+1} - u\|_B \leqslant \rho\|u^k - u\|_B,\ \rho = \big(1 - \frac{\alpha}{\beta}\big)^{1/2}. \tag{1.29}$$

Proof. Because of the uniform monotonicity and the Lipshitz continuity of the operator P, the operator $P + Q$ is maximally monotone and uniformly monotone. Thus, inclusion (1.20) has a unique solution u.

Let $z^k = u^k - u$. Multiplying the inclusion

$$\frac{1}{\tau}B(z^{k+1} - z^k) + P(u^k) - P(u) + Q(u^{k+1}) - Q(u) \ni 0$$

by $2\tau z^{k+1}$, we get

$$\|z^{k+1}\|_B^2 - \|z^k\|_B^2 + \|z^{k+1} - z^k\|_B^2 + 2\tau(P(u^k) - P(u), z^{k+1}) \leqslant 0.$$

Due to (1.27) and (1.28)

$$\begin{aligned} 2\tau(P(u^k) - P(u), z^{k+1}) &= 2\tau(P(u^k) - P(u), z^k) + 2\tau(P(u^k) - P(u), (u^{k+1} - u^k)) \\ &\geqslant (2\tau - \tau^2\beta)(P(u^k) - P(u), z^k) - \|z^{k+1} - z^k\|_B^2. \end{aligned}$$

Substituting this estimate into the previous inequality we get

$$\|z^{k+1}\|_B^2 \leqslant \big(1 - \tau\alpha(2 - \tau\beta)\big)\|z^k\|_B^2,$$

whence all results about the convergence and the rate of convergence for the optimal parameter τ_0 follow. □

2. Iterative solution of the state-constrained optimal control problem

2.1. Formulation of the problem and its approximation

Let $\Omega \subset \mathbb{R}^2$ be a bounded domain with the boundary $\partial\Omega = \Gamma_D \cup \Gamma_N$, $\operatorname{meas}\Gamma_D > 0$, and $V = \{u \in H^1(\Omega) : u(x) = 0 \text{ on } \Gamma_D\}$ be Sobolev space with an inner product $(u, v) = \int\limits_\Omega \nabla u \cdot \nabla v \, dx$ and norm $\|u\| = (u, u)^{1/2}$. Consider the weak formulation of the mixed boundary-value problem for the second order elliptic equation:

$$y \in V: \int\limits_\Omega \sum_{i,j=1}^2 (a_{ij} \frac{\partial y}{\partial x_j} \frac{\partial z}{\partial x_i} + a_0 y z) dx = \int\limits_\Omega f z dx + \int\limits_{\Gamma_N} q z dx \ \ \forall z \in V. \tag{2.1}$$

Suppose that the coefficients $a_{ij}(x)$ and $a_0(x)$ are continuous in the closed domain $\overline{\Omega}$ and

$$\sum_{i,j=1}^2 a_{ij}(x)\xi_j\xi_i \geqslant c_0 \sum_{i=1}^2 \xi_i^2, \ a_0(x) \geqslant 0 \ \forall x \in \overline{\Omega}, \ c_0 = \text{const} > 0.$$

Then the bilinear form $a(y, z)$ defined by the left-hand side of (2.1) is coercive and bounded

$$a(y, y) \geqslant c_0\|y\|^2, \ \forall y \in V; \ \ a(y, z) \leqslant c_1\|y\| \, \|z\|, \ \forall y, z \in V, \ c_1 = \text{const}.$$

Further, for any $f \in L_2(\Omega)$ and any $q \in L_2(\Gamma_N)$ the right-hand side of (2.1) defines a bounded linear functional in V. Therefore, owing to the Lax-Milgram theorem, problem (2.1) has a unique solution $y \in V$, and

$$\|y\|_V \leqslant k(\|f\|_{L_2(\Omega)} + \|q\|_{L_2(\Gamma_N)}), \ k = \text{const}. \tag{2.2}$$

Define the goal functional

$$J(y, f, q) = \frac{1}{2} \int\limits_\Omega (y - y_d)^2 dx + \frac{r_1}{2} \int\limits_{\Omega_1} f^2 dx + \frac{r_2}{2} \int\limits_{\Gamma_N} q^2 d\Gamma, \ r_i = \text{const} > 0,$$

with a given function $y_d(x) \in L_2(\Omega)$ and a subdomain $\Omega_1 \subseteqq \Omega$. Let the sets of constraints be

$$Y_{ad} = \{y \in V : y(x) \geqslant 0 \ \forall x \in \Omega\}, \ F_{ad} = \{f \in L_2(\Omega) : |f(x)| \leqslant f_d \ \forall x \in \Omega_1\},$$

$$Q_{ad} = \{q \in L_2(\Gamma_N) : |q(x)| \leqslant q_d \ \forall x \in \Gamma_N\}, \text{ where } f_d > 0, \, q_d > 0.$$

We consider the optimal control problem

$$\begin{gathered} \text{find} \min_{(y,f,q)\in K} J(y,f,q), \\ K = \{(y,f,q) : y \in Y_{ad}, f \in F_{ad}, \ q \in Q_{ad}, \ \text{Eq. (2.1) holds}\}. \end{gathered} \tag{2.3}$$

Lemma 2.1. *Optimal control problem* (2.3) *has a unique solution.*

Proof. The statement results from the following properties of K and J:
the set K is nonempty, closed and convex in $V \times L_2(\Omega) \times L_2(\Gamma_N)$, bounded because of the boundedness of F_{ad} and Q_{ad} and estimate (2.2);
functional J is continuous and strictly convex in $V \times L_2(\Omega) \times L_2(\Gamma_N)$. □

Suppose that the domains Ω and Ω_1 have polygonal boundaries and construct a finite element approximation of problem (2.3). Let $\overline{\Omega} = \bigcup_{e\in T_h} e$ be a conforming triangulation of $\overline{\Omega}$ ([12]), where T_h is a family of nonoverlapping closed triangles e (finite elements) and h is the maximal diameter of all $e \in T_h$. Let T_h generate triangulations T_h^1 on $\overline{\Omega}_1$ and ∂T_h on $\overline{\Gamma}_N$, namely, $\overline{\Omega}_1$ consists of an integer number of $e \in T_h$ and $\overline{\Gamma}_N$ consists of an integer number of sides ∂e of elements $e \in T_h$. Define the finite element space $V_h \subset V$ of the continuous and piecewise linear functions (linear on each e) which vanish on the boundary Γ_D and the finite element space $U_h \in L_2(\Gamma_N)$ of the piecewise linear functions on Γ_N (linear on each $\partial e \in \Gamma_N$), which are traces on Γ_N of the functions from V_h.

Let, for simplicity, the functions f, y_d and q be continuous and $f(x) = 0$ in $\Omega \setminus \overline{\Omega}_1$. Define $f_h \in V_h$ such that $f_h(x_i) = f(x_i)$ for all nodes x_i of triangulation T_h, and similar $q_h \in U_h$ and $y_{dh} \in V_h$. To approximate the integrals of the continuous function $g(x)$ over the finite element $e \in T_h$ or its side ∂e, we use the quadrature formulas

$$\int_e g(x)dx \approx S_e(g) = \frac{1}{3}\,\text{meas}\,(e) \sum_{\alpha=1}^{3} g(x_\alpha), \ x_\alpha \ \text{are the vertices of } e,$$

$$\int_{\partial e} g(x)d\Gamma \approx S_{\partial e}(g) = \frac{1}{2}\,\text{meas}\,(\partial e) \sum_{\alpha=1}^{2} g(x_\alpha), \ x_\alpha \ \text{are the vertices of } \partial e.$$

The corresponding composite quadrature formulas are

$$S_\Omega(g) = \sum_{e\in T_h} S_e(g), \quad S_{\Omega_1}(g) = \sum_{e\in T_h^1} S_e(g), \quad S_\Gamma(g) = \sum_{\partial e\in \partial T_h} S_{\partial e}(g).$$

Now we can define the discrete optimal control problem, namely, the state equation: find $y_h \in V_h$ such that

$$S_\Omega\left(\sum_{i,j=1}^{2} a_{ij}\,\frac{\partial y_h}{\partial x_j}\,\frac{\partial z_h}{\partial x_i} + a_0 y_h z_h\right) = S_{\Omega_1}(f_h\, z_h) + S_\Gamma(q_h\, z_h) \ \forall z_h \in V_h; \tag{2.4}$$

the goal function

$$J_h(y_h, f_h, q_h) = \frac{1}{2}\, S_\Omega((y_h - y_{dh})^2) + \frac{r_1}{2}\, S_{\Omega_1}(f_h^2) + \frac{r_2}{2}\, S_\Gamma(q_h^2);$$

the sets of constraints

$$Y_{ad}^h = \{y_h \in V_h :\ y_h(x) \geqslant 0 \text{ for } x \in \Omega\},\ Q_{ad}^h = \{q_h \in U_h :\ |q_h(x)| \leqslant q_d\ \forall x \in \Gamma_N\},$$

$$F_{ad}^h = \{f_h \in V_h :\ |f_h(x)| \leqslant f_d\ \forall x \in \overline{\Omega}_1,\ f_h(x) = 0\ \forall x \in \Omega \setminus \overline{\Omega}_1\},$$

$$\text{and } K_h = \{(y_h, f_h, q_h) :\ y_h \in Y_{ad}^h, f_h \in F_{ad}^h,\ q_h \in Q_{ad}^h,\ \text{ Eq. (2.4) holds}\};$$

the resulting discrete optimal control problem

$$\text{find} \min_{(y_h, f_h, q_h) \in K_h} J_h(y_h, f_h, q_h). \tag{2.5}$$

The bilinear form $a_h(y_h, z_h)$, defined by the left-hand side of Eq. (2.4), is uniformly in h coercive and bounded:

$$a_h(y_h, y_h) \geqslant \tilde{c}_0 \|y_h\|^2,\ \forall y_h \in V_h;\ \ a_h(y_h, z_h) \leqslant \tilde{c}_1 \|y_h\|\, \|z_h\|,\ \forall y_h, z_h \in V_h.$$

Because of this, Eq. (2.4) has a unique solution y_h for any $f_h \in V_h$, $q_h \in U_h$ and the following stability inequalities hold:

$$S_\Omega^{1/2}(|y_h|^2) \leqslant k_1\, S_\Omega^{1/2}(|\nabla y_h|^2) \leqslant k_f\big(S_{\Omega_1}^{1/2}(f_h^2) + S_\Gamma^{1/2}(q_h^2)\big) \tag{2.6}$$

with constants k_1 and k_f independent of h.

Lemma 2.2. *The discrete optimal control problem* (2.5) *has a unique solution* (y_h, f_h, q_h).

Proof. The proof immediately follows from the fact that the set K_h is nonempty, closed, convex, and bounded, while the function J_h is continuous and strictly convex. □

Below we formulate problem (2.5) in a "vector-matrix" form. Denoting by $y \in \mathbb{R}^{N_y}$ the vector of the nodal values of the function $y_h \in V_h$ ($N_y = \dim V_h$), we get the "onto" correspondence $y \Leftrightarrow y_h$. Similarly, we define $u \in \mathbb{R}^{N_u}$, $u \Leftrightarrow u_h$, $u_h \in U_h$, and $f \in \mathbb{R}^{N_f}$ as the vector of the nodal values for the restriction of the function $f_h(x)$ on the subdomain $\overline{\Omega}_1$.[5]

Define the stiffness matrix $L_y \in \mathbb{R}^{N_y \times N_y}$, the diagonal mass matrices $M_y \in \mathbb{R}^{N_y \times N_y}$, $M_f \in \mathbb{R}^{N_f \times N_f}$ and $M_q \in \mathbb{R}^{N_u \times N_u}$, and the rectangular matrices $S_q \in \mathbb{R}^{N_y \times N_u}$, $S_f \in \mathbb{R}^{N_y \times N_f}$ by the following equalities:

$$\begin{aligned}
&(L_y y, z) = S_\Omega\left(\sum_{i,j=1}^{2} a_{ij}\,\frac{\partial y_h}{\partial x_j}\,\frac{\partial z_h}{\partial x_i} + a_0 y_h z_h\right), && (M_y y, z) = S_\Omega(y_h z_h),\\
&(M_f f, g) = S_{\Omega_1}(f_h g_h), && (S_f f, z)\ = S_{\Omega_1}(f_h z_h),\\
&(M_q u, v) = S_\Gamma(u_h v_h), && (S_q u, z)\ = S_\Gamma(u_h z_h).
\end{aligned} \tag{2.7}$$

Above $y \Leftrightarrow y_h \in V_h$, $z \Leftrightarrow z_h \in V_h$, $u \Leftrightarrow u_h \in U_h$, $v \Leftrightarrow v_h \in U_h$, $f \Leftrightarrow f_h \in F_h$, $g \Leftrightarrow g_h \in F_h$.

[5] Since hereafter we consider only finite dimensional problems, we use the same notations for the vectors as for the functions.

With these notations the discrete state equation (2.4) can be written as a system of linear algebraic equations

$$L_y y = S_f f + S_q q$$

with a regular matrix L_y. The constraint sets become

$$F_{ad} = \{f \in \mathbb{R}^{N_f} : |f_i| \leqslant f_d \ \ \forall i\}, \ Q_{ad} = \{q \in \mathbb{R}^{N_u} : |q_i| \leqslant q_d \ \ \forall i\},$$
$$Y_{ad} = \{y \in \mathbb{R}^{N_y} : y_i \geqslant 0 \ \forall i\}.$$

Let further $\theta(y) = I_{Y_{ad}}(y)$, $\psi(f) = I_{F_{ad}}(f)$, and $\varphi(q) = I_{Q_{ad}}(q)$ be the indicator functions of the sets Y_{ad}, F_{ad}, and Q_{ad}, respectively.

The optimal control problem for the vectors of the nodal values of the grid functions is

$$\begin{aligned}
&\text{find} \min_{L_y y = S_f f + S_q q} \{J(y,f,q) = \Theta(y) + \Psi(f) + \Phi(q)\}, \\
&\text{where} \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (2.8)\\
&\Theta(y) = \frac{1}{2}(M_y y, y) - (g, y) + \theta(y), \ g = M_y y_d, \\
&\Psi(f) = \frac{r_1}{2}(M_f f, f) + \psi(f), \ \ \Phi(q) = \frac{r_2}{2}(M_q q, q) + \varphi(q).
\end{aligned}$$

The Lagrange function for (2.8) has the form

$$\begin{aligned}
\mathcal{L}(y,f,q,\lambda) = \frac{1}{2}(M_y y, y) - (g, y) + \theta(y) + \frac{r_1}{2}(M_f f, f) + \psi(f) \\
+ \frac{r_2}{2}(M_q q, q) + \varphi(q) - (L_y y - S_f f - S_q q, \lambda)
\end{aligned}$$

and its saddle point (y,f,q,λ) satisfies the system

$$\begin{pmatrix} M_y & 0 & 0 & -L_y^T \\ 0 & r_1 M_f & 0 & S_f^T \\ 0 & 0 & r_2 M_q & S_q^T \\ -L_y & S_f & S_q & 0 \end{pmatrix} \begin{pmatrix} y \\ f \\ q \\ \lambda \end{pmatrix} + \begin{pmatrix} \partial\theta(y) \\ \partial\psi(f) \\ \partial\varphi(q) \\ 0 \end{pmatrix} \ni \begin{pmatrix} g \\ 0 \\ 0 \\ 0 \end{pmatrix}. \tag{2.9}$$

Lemma 2.3. *Let the assumption*

$$\exists f_0 \in F_{ad}, \ \exists q_0 \in Q_{ad} : \ y_0 = L_y^{-1}(S_f f_0 + S_q q_0) \gg 0 \tag{2.10}$$

hold, where $v \gg 0$ means that $v_i > 0$ for all coordinates i of the vector v. Then system (2.9) *has a solution (y,f,q,λ) with unique (y,f,q).*

Proof. First, we list some properties of the matrices and the functions in problem (2.9):

$$\begin{array}{c} L_y \text{ is a positive definite matrix;} \\ M_y, M_f \text{ and } M_q \text{ are diagonal matrices with positive diagonals;} \\ S_f \text{ and } S_q \ \text{ are rectangular matrices with nonnegative entries;} \end{array} \tag{2.11}$$

$$\begin{array}{c} \Theta, \Psi \text{ and } \Phi \text{ are convex, lower semi-continuous functions} \\ \text{with domains} \quad \operatorname{dom}\Theta = Y_{ad}, \ \ \operatorname{dom}\Psi = F_{ad}, \ \ \operatorname{dom}\Psi = Q_{ad}. \end{array} \tag{2.12}$$

Now, we denote $u = (f,q)^T$, $L = L_y$, $S = (S_f, S_q)$, $\Upsilon(u) = \Psi(f) + \Phi(q)$ and use Theorem 1.2 to prove the solvability of (2.9). Properties (2.11) and (2.12) ensure the validity of assumptions (1.2), (1.3), (5a), and (6a) of Theorem 1.2. Assumption (2.10) corresponds to (1.9). Thus, all assumptions of Theorem 1.2 are fulfilled , whence the result. □

Below we give couple of examples when assumption (2.10) is fulfilled.

1) Suppose that the approximated equation has no mixed derivatives ($a_{ij} = 0$ for $i \neq j$) and the triangulation of the domain is of the acute type: all angles of $e \in T_h$ are less than or equal to $\pi/2$. In this case, the matrix L_y is irreducible diagonally dominant M-matrix and all entries of L_y^{-1} are strictly positive (cf. [13]). Thus, $y_0 = L_y^{-1}(S_f f_0 + S_q q_0) \gg 0$ for all nonnegative vectors $f_0 \in F_{ad}$ and $q_0 \in Q_{ad}$ such that at least one coordinate of f_0 or q_0 is positive.

2) Let the state problem be a Dirichlet problem and the control is in the entire domain. Then the discrete state equation becomes $L_y y = M_f f$, where M_f is a diagonal matrix with positive entries. Take any $\tilde{y} \gg 0$ and put $\tilde{f} = M_f^{-1} L_y \tilde{y}$. Then the pair (y_0, f_0) with $y_0 = \alpha\tilde{y}$, $f_0 = \alpha\tilde{f}$ and a positive $\alpha \leqslant f_d\,(\max_i |\tilde{f}_i|)^{-1}$ satisfies (2.10).

2.2. Iterative methods based on the finding of control variables

First, we study the convergence and the rate of convergence of method (1.14) for a particular case of problem (2.9) corresponding to the control on the right-hand side of the equation. For simplicity, we fix $q = 0$.

To apply method (1.14), we approximate $\theta(y) = I_{Y_{ad}}(y) = \{0 \text{ if } y_i \geqslant 0\ \forall i; +\infty \text{ otherwise}\}$ by

$$\theta_\varepsilon(y) = \frac{1}{2\,\varepsilon}(M_y y^-, y^-) \text{ with gradient } \nabla\theta_\varepsilon(y) = -\frac{1}{\varepsilon} M_y y^-, \tag{2.13}$$

where y^- is a vector with coordinates $y_i^- = 0.5(|y_i| - y_i)$ and $\varepsilon > 0$ is a small parameter. With these assumptions problem (2.9) becomes

$$\begin{pmatrix} M_y & 0 & -L_y^T \\ 0 & r_1 M_f & S_f^T \\ -L_y & S_f & 0 \end{pmatrix} \begin{pmatrix} y \\ f \\ \lambda \end{pmatrix} + \begin{pmatrix} \nabla\theta_\varepsilon(y) \\ \partial\psi(f) \\ 0 \end{pmatrix} \ni \begin{pmatrix} g \\ 0 \\ 0 \end{pmatrix}. \tag{2.14}$$

System (2.14) has a unique solution (y, f, λ). Eliminating the vectors y and λ from (2.14) gives the inclusion

$$Pf + \partial\psi(f) \ni L_y^{-T} g \tag{2.15}$$

with

$$Pf = \big(S_f^T\, L_y^{-T}\, M_y\, L_y^{-1}\, S_f + r_1 M_f\big) f + S_f^T\, L_y^{-T}\, \nabla\theta_\varepsilon(L_y^{-1} S_f f). \tag{2.16}$$

To solve (2.15), we use the iterative method

$$M_f \frac{f^{k+1} - f^k}{\tau} + Pf^k + \partial\psi(f^{k+1}) \ni L_y^{-T} g. \tag{2.17}$$

Note that the choice of the diagonal preconditioner M_f is very reasonable because the implementation of an inclusion with a diagonal operator $M_f + \tau\,\partial\psi$ reduces to the solution of a system of one-dimensional problems.

Theorem 2.1. *The iterative method* (2.17) *converges if* $\tau \in \big(0, \dfrac{2\varepsilon}{k_f^2(1+\varepsilon) + r_1\,\varepsilon}\big)$, *where* k_f *is a constant from inequality* (2.6).

For $\tau = \tau_0 = \dfrac{\varepsilon}{k_f^2(1+\varepsilon) + r_1\,\varepsilon}$ *the rate of convergence is characterized by the inequality*

$$\|f^{k+1} - f\|_{M_f} \leqslant \rho \|f^k - f\|_{M_f}, \quad \rho = 1 - \frac{r_1\,\varepsilon}{k_f^2(1+\varepsilon) + r_1\,\varepsilon}.$$

Proof. We apply Theorem 1.4 with $Q = \partial\psi$, $B = M_f$ and P given by (2.16). It is easy to check that

$$(\nabla\theta_\varepsilon(y) - \nabla\theta_\varepsilon(z), y - z) \geqslant 0,$$

$$(\nabla\theta_\varepsilon(y) - \nabla\theta_\varepsilon(z), x) \leqslant \frac{1}{\sqrt{\varepsilon}}\,(\nabla\theta_\varepsilon(y) - \nabla\theta_\varepsilon(z), y - z)^{1/2}\|x\|_{M_y}.$$

These inequalities imply the following estimates for the operator P:

$$\begin{aligned}(P(u) - P(v), u - v) = \big\|L_y^{-1}S_f(u-v)\big\|_{M_y}^2 + r_1\,\big\|u - v\big\|_{M_f}^2 + +\big(\nabla\theta_\varepsilon(L_y^{-1}S_f u)\\ - \nabla\theta_\varepsilon(L_y^{-1}S_f v), L_y^{-1}S_f(u-v)\big) \geqslant r_1\big\|u-v\big\|_{M_f}^2,\end{aligned} \tag{2.18}$$

and

$$\begin{aligned}(P(u) - P(v), w) &\leqslant \big\|L_y^{-1}S_f(u-v)\big\|_{M_y}\,\big\|L_y^{-1}S_f w\big\|_{M_y} + r_1\,\big\|u-v\big\|_{M_f}\big\|w\big\|_{M_f}\\ &+ \frac{1}{\sqrt{\varepsilon}}\big(\nabla\theta_\varepsilon(L_y^{-1}S_f u) - \nabla\theta_\varepsilon(L_y^{-1}S_f v), L_y^{-1}S_f(u-v)\big)^{1/2}\big\|L_y^{-1}S_f w\big\|_{M_y}\\ &\leqslant (P(u)-P(v), u-v)^{1/2}\big((1+\frac{1}{\varepsilon})\,\big\|L_y^{-1}S_f w\big\|_{M_y}^2 + r_1\,\big\|w\big\|_{M_f}^2\big)^{1/2}.\end{aligned}$$

Let us prove the inequality

$$(M_y L_y^{-1}S_f f, L_y^{-1}S_f f) \leqslant k_f^2(M_f f, f)\ \ \forall f. \tag{2.19}$$

Define y as a solution of the equation $L_y y = S_f f$ and let $y \leftrightarrow y_h \in V_h$, $f \leftrightarrow f_h \in F_h$. Then from (2.6) we get

$$(M_y y, y) = S_\Omega(|y_h|^2) \leqslant k_f^2 S_{\Omega_1}(f_h^2) = k_f^2(M_f\, f, f),$$

which is essentially (2.19). As a result,

$$(P(u) - P(v), w) \leqslant \big((1+\varepsilon^{-1})k_f^2 + r_1\big)^{1/2}(P(u)-P(v), u-v)^{1/2}\big\|w\big\|_{M_f}. \tag{2.20}$$

So, the constants in inequalities (1.27), (1.28) of the uniform monotonicity and Lipsitz continuity of the operator P can be taken equal to

$$\alpha = r_1,\quad \beta = (1+\varepsilon^{-1})k_f^2 + r_1,$$

and all statements of the theorem follow from Theorem 1.4. □

As follows from Theorem 2.1, the optimal iterative parameter τ_0 and the factor ρ are

$$\tau_0 = O(\varepsilon),\quad \rho = 1 - O(r_1\,\varepsilon).$$

This means that the number of iterations in method (2.17) for achieving the desired accuracy does not depend on the mesh size h, but depends linearly on $\dfrac{1}{r_1\varepsilon}$. Since the parameter $\varepsilon > 0$ is usually taken as $\varepsilon = \varepsilon(h)$, then the rate of convergence of method (2.17) can strictly depend on h.

Remark 2.1. If there are no state constraints, i. e., $\theta = 0$, then the constants in the inequalities of the uniform monotonicity and Lipsitz continuity of the operator P are equal to

$$\alpha = r_1, \quad \beta = k_f^2 + r_1,$$

thus, the optimal iterative parameter τ_0 and the factor ρ in method (2.17) are

$$\tau_0 = \frac{1}{k_f^2 + r_1} = O(1), \quad \rho = 1 - \frac{r_1}{k^2 + r_1} = 1 - O(r_1).$$

Now, we will briefly examine the problem with the control function on the right-hand side of the Neumann boundary condition with fixed $f = 0$ and θ_ε given by (2.13). In this case, problem (2.9) becomes

$$\begin{pmatrix} M_y & 0 & -L_y^T \\ 0 & r_2\, M_q & S_q^T \\ -L_y & S_q & 0 \end{pmatrix} \begin{pmatrix} y \\ q \\ \lambda \end{pmatrix} + \begin{pmatrix} \nabla\theta_\varepsilon(y) \\ \partial\varphi(q) \\ 0 \end{pmatrix} \ni \begin{pmatrix} g \\ 0 \\ 0 \end{pmatrix}. \tag{2.21}$$

We transform system (2.21) into the inclusion for finding the vector q

$$Pq + \partial\varphi(q) \ni L_y^{-T} g,$$

$$Pq = \big(S_q^T\, L_y^{-T}\, M_y\, L_y^{-1}\, S_q + r_2 M_q\big)q + S_q^T\, L_y^{-T}\, \nabla\theta_\varepsilon(L_y^{-1} S_q q),$$

and solve it by the one-step stationary iterative method with the diagonal preconditioner M_q

$$\frac{1}{\tau} M_q(q^{k+1} - q^k) + Pq^k + \partial\varphi(q^{k+1}) \ni L_y^{-T} g. \tag{2.22}$$

Theorem 2.2. *Method* (2.22) *converges if* $\tau \in (0, \dfrac{2\varepsilon}{k_f^2(1+\varepsilon) + r_2\,\varepsilon})$, *and for* $\tau = \tau_0 = \dfrac{\varepsilon}{\tilde{k}^2(1+\varepsilon) + r_2\,\varepsilon} = O(\varepsilon)$ *the following estimate for the rate of convergence holds:*

$$\|q^{k+1} - q\|_{M_q} \leqslant \rho \|q^k - q\|_{M_q}, \quad \rho = 1 - \frac{r_2\,\varepsilon}{k_f^2(1+\varepsilon) + r_2\,\varepsilon} = 1 - O(r_2\,\varepsilon).$$

Proof. Similarly to (2.18) and (2.20) we can prove the estimates

$$(P(u) - P(v), u - v) \geqslant r_2 \big\|u - v\big\|_{M_q}^2,$$

$$(P(u) - P(v), w) \leqslant \big((1 + \varepsilon^{-1})k_f^2 + r_2\big)^{1/2} (P(u) - P(v), u - v)^{1/2} \big\|w\big\|_{M_q},$$

i. e., the operator P is uniformly monotone and Lipshitz-continuous with constants $\alpha = r_2$ and $\beta = (1 + \varepsilon^{-1})k_f^2 + r_2$. To prove the Lipshitz continuity we use the inequality

$$(M_y\, L_y^{-1} S_q q, L_y^{-1} S_q q) \leqslant k_f^2 (M_q\, q, q) \quad \forall q, \tag{2.23}$$

which is a consequence of estimate (2.6). All formulations follow now from Theorem 1.4. □

2.3. Iterative methods based on finding Lagrange multipliers

Consider (2.9) with the control function on the right-hand side of the equation (fix $q = 0$) without supposition about the differentiability of θ

$$\begin{pmatrix} M_y & 0 & -L_y^T \\ 0 & r_1\, M_f & S_f^T \\ -L_y & S_f & 0 \end{pmatrix} \begin{pmatrix} y \\ f \\ \lambda \end{pmatrix} + \begin{pmatrix} \partial\theta(y) \\ \partial\psi(f) \\ 0 \end{pmatrix} \ni \begin{pmatrix} g \\ 0 \\ 0 \end{pmatrix}. \tag{2.24}$$

System (2.24) has a solution (y, f, λ) with unique (y, f) and a generally not unique Lagrange multiplier λ. Excluding y and f from this system we obtain the equation for the vector λ

$$P(\lambda) \equiv L_y\,(M_y + \partial\theta)^{-1}(L_y^T\lambda + g) - S_f\,(r_1\, M_f + \partial\psi)^{-1}(-S_f^T\lambda) = 0. \tag{2.25}$$

Let us apply for solving it the stationary one-step iterative method

$$L_y\, M_y^{-1}\, L_y^T \frac{\lambda^{k+1} - \lambda^k}{\tau} + P(\lambda^k) = 0 \tag{2.26}$$

and prove the convergence of this method as well as the rate of convergence in the case where θ is changed by the regularized function.

Theorem 2.3. *Let*

$$0 < \tau < \frac{2r_1}{k_f^2 + r_1}, \tag{2.27}$$

where k_f is a constant from inequality (2.6). *Then iterations of method* (2.26) *converge to the solution of* (2.25).

Proof. The operator P can be written in the form

$$P(\lambda) = L_y\, M_y^{-1/2} A_1(M_y^{-1/2}(L_y^T\lambda + g)) - S_f\,(r_1\, M_f)^{-1/2} A_2(-(r_1\, M_f)^{-1/2} S_f^T\lambda)$$

with

$$A_1 = M_y^{1/2} \circ (M_y + \partial\theta)^{-1} \circ M_y^{1/2} \text{ and } A_2 = (r_1\, M_f)^{1/2} \circ (r_1\, M_f + \partial\psi)^{-1} \circ (r_1\, M_f)^{1/2}.$$

To investigate the convergence of the iterative method (2.26), we apply Theorem 1.3 with

$$Q = 0,\ A = \begin{pmatrix} A_1 & 0 \\ 0 & A_2 \end{pmatrix},\ C = \begin{pmatrix} M_y^{-1/2}\, L_y^T \\ -(r_1\, M_f)^{-1/2} S_f^T \end{pmatrix}.$$

Using the notations $u_i = (M_y + \partial\theta)^{-1}(M_y^{1/2} y_i)$ we obtain

$$\begin{aligned}(A_1(y_1) - A_1(y_2), y_1 - y_2) &= ((M_y + \partial\theta)^{-1}(M_y^{1/2} y_1) \\ &\quad - (M_y + \partial\theta)^{-1}(M_y^{1/2} y_2), M_y^{1/2} y_1 - M_y^{1/2} y_2) \\ &\geqslant (M_y(u_1 - u_2), u_1 - u_2) = \|A_1(y_1) - A_1(y_2)\|^2. \end{aligned} \tag{2.28}$$

Similarly,

$$(A_2(f_1) - A_2(f_2), f_1 - f_2) \geqslant \|A_2(f_1) - A_2(f_2)\|^2. \tag{2.29}$$

Owing to (2.28) and (2.29), the assumptions of Theorem 1.3 are satisfied and the convergence condition (1.23) becomes

$$L_y\, M_y^{-1}\, L_y^T > \frac{\tau}{2}\big(L_y\, M_y^{-1}\, L_y^T + r_1^{-1}\, S_f\, M_f^{-1}\, S_f^T\big). \tag{2.30}$$

Let us prove the inequality

$$(S_f\, M_f^{-1}\, S_f^T\lambda, \lambda) \leqslant k_f^2(L_y\, M_y^{-1}\, L_y^T\lambda, \lambda). \tag{2.31}$$

We have

$$\|M_y^{-1/2}L_y^T\lambda\| = \sup_v \frac{(M_y^{-1/2}L_y^T\lambda, v)}{\|v\|} = \sup_y \frac{(\lambda, L_y y)}{\|M_y^{1/2}y\|} \geqslant \sup_f \frac{(\lambda, S_f f)}{\|M_y^{1/2}L_y^{-1}S_f f\|}.$$

Using inequality (2.19) gives

$$\|M_y^{-1/2}L_y^T\lambda\| \geqslant \frac{1}{k_f}\sup_f \frac{(S_f^T\lambda, f)}{\|M_f^{1/2}f\|} = \frac{1}{k_f}\|M_f^{-1/2}S_f^T\lambda\|,$$

i. e., inequality (2.31). Due to (2.31) the convergence condition (2.30) is fulfilled if

$$1 > \frac{\tau}{2}(1 + k_f^2 r_1^{-1}),$$

which is essentially (2.27). □

Theorem 2.4. *Let the function θ be changed by the regularized function $\theta_\varepsilon(y) = \frac{1}{2\,\varepsilon}(M_y y^-, y^-)$ in problem* (2.25)*. Then this problem has a unique solution, method* (2.26) *converges if*

$$0 < \tau < \frac{2r_1}{k_f^2 + r_1},$$

and for

$$\tau_0 = \frac{r_1}{k_f^2 + r_1}$$

the following estimate for the rate of convergence holds:

$$\|\lambda^{k+1} - \lambda\|_B \leqslant \big(1 - \frac{\varepsilon r_1}{(1+\varepsilon)(k_f^2 + r_1)}\big)^{1/2}\|\lambda^k - \lambda\|_B,\ B = L_y\, M_y^{-1}\, L_y^T.$$

Proof. In the case under consideration $P(\lambda) = P_1(\lambda) + P_2(\lambda)$, where

$$P_1(\lambda) = L_y\,(M_y + \nabla\theta_\varepsilon)^{-1}\, L_y^T, \quad P_2(\lambda) = -S_f\,(r_1\, M_f + \partial\psi)^{-1}(-S_f^T\lambda).$$

The following inequalities can be proved by direct calculations taking into account that the matrix M_y and the operator $\nabla\theta_\varepsilon$ are diagonal:

$$\big((M_y + \nabla\theta_\varepsilon)^{-1}(y_1) - (M_y + \nabla\theta_\varepsilon)^{-1}(y_2), y_1 - y_2\big) \geqslant \frac{\varepsilon}{1+\varepsilon}\,(M_y^{-1}(y_1 - y_2), y_1 - y_2),$$

$$\big((M_y + \nabla\theta_\varepsilon)^{-1}(y_1) - (M_y + \nabla\theta_\varepsilon)^{-1}(y_2), z\big)$$

$$\leqslant \big((M_y+\nabla\theta_\varepsilon)^{-1}(y_1)-(M_y+\nabla\theta_\varepsilon)^{-1}(y_2),y_1-y_2\big)^{1/2}\,(M_y^{-1}z,z)^{1/2}.$$

From these inequalities immediately follow

$$(P_1(\lambda_1)-P_1(\lambda_2),\lambda_1-\lambda_2)\geqslant\frac{\varepsilon}{1+\varepsilon}\,\|\lambda_1-\lambda_2\|_B^2,$$

$$(P_1(\lambda_1)-P_1(\lambda_2),\mu)\leqslant(P_1(\lambda_1)-P_1(\lambda_2),\lambda_1-\lambda_2)^{1/2}\,\|\mu\|_B.$$

Let us now use the notations $f_i=-(r_1\,M_f)^{-1/2}S_f^T\lambda_i$, $i=1,2$, and $A_2=(r_1\,M_f)^{1/2}\circ(r_1\,M_f+\partial\psi)^{-1}\circ(r_1\,M_f)^{1/2}$. Then

$$(P_2(\lambda_1)-P_2(\lambda_2),\lambda_1-\lambda_2)=(A_2(f_1)-A_2(f_2),f_1-f_2)\geqslant 0,$$

and in virtue of (2.29)

$$(P_2(\lambda_1)-P_2(\lambda_2),\mu)\leqslant(A_2(f_1)-A_2(f_2),f_1-f_2)^{1/2}\|(r_1\,M_f)^{-1/2}S_f^T\mu\|.$$

Using inequality (2.31) gives

$$(P_2(\lambda_1)-P_2(\lambda_2),\mu)\leqslant(P_2(\lambda_1)-P_2(\lambda_2),\lambda_1-\lambda_2)^{1/2}(k_f^2\,r_1^{-1})^{1/2}\|\mu\|_B.$$

Combining the estimates for the operators P_1 and P_2 yields

$$(P(\lambda_1)-P(\lambda_2),\lambda_1-\lambda_2)\geqslant\frac{\varepsilon}{1+\varepsilon}\,\|\lambda_1-\lambda_2\|_B^2,$$

$$(P(\lambda_1)-P(\lambda_2),\mu)\leqslant(1+r_1^{-1}k_f^2)^{1/2}\,(P(\lambda_1)-P(\lambda_2),\lambda_1-\lambda_2)^{1/2}\,\|\mu\|_B.$$

Thus, inequalities (1.27) and (1.28) are true with $\alpha=\dfrac{\varepsilon}{1+\varepsilon}$, $\beta=1+r_1^{-1}k_f^2$, and the statement of the theorem follows from Theorem 1.4. □

Remark 2.2. If $\theta=0$, then the optimal iterative parameter τ_0 and the factor ρ in method (2.26) are the same as in method (2.17) (cf. Remark 2.1):

$$\tau_0=\frac{1}{k_f^2+r_1},\quad \rho=1-\frac{r_1}{k_f^2+r_1}.$$

Consider now a problem with the control function on the right-hand side of the Neumann boundary condition with fixed $f=0$

$$\begin{pmatrix} M_y & 0 & -L_y^T\\ 0 & r_2\,M_q & S_q^T\\ -L_y & S_q & 0\end{pmatrix}\begin{pmatrix}y\\ q\\ \lambda\end{pmatrix}+\begin{pmatrix}\partial\theta(y)\\ \partial\varphi(q)\\ 0\end{pmatrix}\ni\begin{pmatrix}g\\ 0\\ 0\end{pmatrix}. \tag{2.32}$$

We transform system (2.32) into the equation for finding the vector λ

$$P(\lambda)\equiv L_y\,(M_y+\partial\theta)^{-1}(L_y^T\lambda+g)-S_q\,(r_2\,M_q+\partial\varphi)^{-1}(-S_q^T\lambda)=0. \tag{2.33}$$

Let us apply for solving it the stationary one-step iterative method

$$L_y\,M_y^{-1}\,L_y^T\frac{\lambda^{k+1}-\lambda^k}{\tau}+P(\lambda^k)=0. \tag{2.34}$$

Theorem 2.5. *Let*

$$0 < \tau < \frac{2r_2}{k_f^2 + r_2}, \tag{2.35}$$

where k_f is a constant from inequality (2.6)*. Then iterations of method* (2.34) *converge to the solution of* (2.33)*.*

If the function θ is changed by the regularized function $\theta_\varepsilon(y) = \frac{1}{2\,\varepsilon}(M_y y^-, y^-)$, then problem (2.33) *has a unique solution, method* (2.34) *converges if τ satisfied* (2.35)*, and for*

$$\tau_0 = \frac{r_2}{k_f^2 + r_2}$$

the following estimate for the rate of convergence holds:

$$\|\lambda^{k+1} - \lambda\|_B \leqslant \big(1 - \frac{\varepsilon r_2}{(1+\varepsilon)(k_f^2 + r_2)}\big)^{1/2} \|\lambda^k - \lambda\|_B,\ B = L_y\, M_y^{-1}\, L_y^T.$$

Proof. Similarly to the proof of Theorem 2.3 we can check that the assumptions of Theorem 1.3 are satisfied and the convergence condition (1.23) has the form

$$L_y\, M_y^{-1}\, L_y^T > \frac{\tau}{2}\big(L_y\, M_y^{-1}\, L_y^T + r_2^{-1}\, S_q\, M_q^{-1}\, S_q^T\big). \tag{2.36}$$

Further we use the inequality

$$(S_q\, M_q^{-1}\, S_q^T \lambda, \lambda) \leqslant k_f^2 (L_y\, M_y^{-1}\, L_y^T \lambda, \lambda), \tag{2.37}$$

whose proof is the same as inequality (2.31). After that it is easy to show that (2.36) is true when

$$1 > \frac{\tau}{2}(1 + k_f^2 r_2^{-1}).$$

The proof of the statements in the case $\theta = \theta_\varepsilon$ is just the same as in Theorem 2.4. Namely, we use, among other things, inequality (2.37) to prove the estimates

$$(P(\lambda_1) - P(\lambda_2), \lambda_1 - \lambda_2) \geqslant \frac{\varepsilon}{1+\varepsilon}\, \|\lambda_1 - \lambda_2\|_B^2,$$

$$(P(\lambda_1) - P(\lambda_2), \mu) \leqslant (1 + r_2^{-1} k_f^2)^{1/2}\, (P(\lambda_1) - P(\lambda_2), \lambda_1 - \lambda_2)^{1/2}\, \|\mu\|_B.$$

After that all formulated statements follow from Theorem 1.4.

□

Conclusions

On the basis of the proven convergence results for the iterative solution methods for State-constrained optimal control problem (2.9) we can draw the following conclusions.

In the case of $\theta = 0$ (problem without state constraints) or $\theta = \theta_\varepsilon$ (regularized indicator function of the set of state constraints), the theoretical estimates for the rates of convergence of methods (2.26), (2.34) and methods (2.17), (2.22) are asymptotically the same. The complexity is also the same – at each iteration in these methods we need to inverse L_y and L_y^T and solve an inclusion with a diagonal operator. On the other hand, methods (2.26), (2.34) have the following advantages:

- they can be applied to problems with a non-differentiable function θ without its regularization;
- in the case of a regularized function $\theta = \theta_\varepsilon$, the bounds for the iterative parameter τ which ensure the convergence and the optimal parameter τ_0 do not depend on ε;
- it is possible to use a preconditioner $B_0 = L_0\,M_y^{-1}\,L_0$ instead of $B = L_y\,M_y^{-1}\,L_y$ with a matrix B_0 which is spectrally equivalent to B; for example, L_0 may be a matrix corresponding to inexact inversion of L_y by an iterative method.

Acknowledgments. The work of A. Lapin was supported by the RFBR, project 10-01-0062977.

References

1. M. Benzi, G. Golub, and J. Liesen, *Numerical solution of saddle point problems*, Acta Numerica, **14** (2005), pp. 1–137.
2. R. Glowinski and P. LeTallec, *Augmented Lagrangan and operator-splitting methods in nonlinear mechanics*, SIAM studies in applied mathematics, Philadelphia, PA, 1989.
3. M. Ignatieva and A. Lapin, *Iterative solution of mixed hybrid finite element scheme for Signorini problem*, Comp. Methods in Appl. Math.,**4** (2004), no. 2, pp. 180–191.
4. Ph. Gill, W. Murray, and M. Wright, *Practical optimization*, London etc.: Academic Press, 1981.
5. D.P. Bertsekas, *Constrained optimization and Lagrange multiplier methods*, Computer Science and Applied Mathematics, N. Y., London etc.: Academic Press, 1982.
6. L.T. Biegler, O. Ghattas, M. Heinkenschloss, and B. van Bloemen Waanders (eds.), *Large-scale PDE-constrained optimization*, Lecture Notes in Computational Science and Engineering, 30, Berlin: Springer, 2003.
7. M. Bergounioux, *Augmented Lagrangian method for distributed optimal control problems with state constraints*, Optimization Theory Appl., **78** (1993), no. 3, pp. 493–521.
8. M. Bergounioux, V. Haddou, M. Hintermuller, and K. Kunisch,*A comparison of a Moreau-Yosida-based active set strategy and interior point methods for constrained optimal control problems*, SIAM J. Optim., **11** (2000), no. 2, pp. 495–521.
9. M. Bergounioux and K. Kunisch, *Primal-dual strategy for state-constrained optimal control problems*, Comput. Optim. Appl., **22** (2002), no. 2, pp. 193–224.
10. P.T. Boggs and J.W. Tolle, *Sequential quadratic programming*, Acta Numerica, 1995.
11. I. Ekeland and R. Temam, *Convex analysis and variational problems*, North Holland, Amsterdam, 1976.
12. Ph.G. Ciarlet, *The finite element method for elliptic problems*, North Holland, Amsterdam, 1978.
13. O. Axelsson, *Iterative solution methods*, Cambridge: Cambridge Univ. Press, 1996.