ESAIM: COCV 15 (2009) 626–652
DOI: 10.1051/cocv:2008042

ESAIM: Control, Optimisation and Calculus of Variations
www.esaim-cocv.org

MESH-INDEPENDENCE AND PRECONDITIONING FOR SOLVING PARABOLIC CONTROL PROBLEMS WITH MIXED CONTROL-STATE CONSTRAINTS

Michael Hintermüller[1], Ian Kopacka[2] and Stefan Volkwein[2]

Abstract. Optimal control problems for the heat equation with pointwise bilateral control-state constraints are considered. A locally superlinearly convergent numerical solution algorithm is proposed and its mesh independence is established. Further, for the efficient numerical solution reduced space and Schur complement based preconditioners are proposed which take into account the active and inactive set structure of the problem. The paper ends by numerical tests illustrating our theoretical findings and comparing the efficiency of the proposed preconditioners.

Mathematics Subject Classification. 49K20, 65K05.

Received December 13, 2006. Revised November 7, 2007 and March 31, 2008.
Published online July 19, 2008.

1. Introduction

In this work we study a locally superlinearly convergent algorithm for computing the solution of constrained distributed optimal control problems for processes governed by the linear heat equation

$$y_t - \alpha\Delta y = f + u \quad \text{in } Q, \quad y(0) = y_\circ \quad \text{on } \Omega, \tag{1.1}$$

where $\Omega \subset \mathbb{R}^d$ represents the domain of interest and Q is the space-time cylinder. Furthermore, f is a given source, $y_\circ$ denotes the initial temperature, and $\alpha > 0$ is a given constant reflecting heat conduction properties. We consider (1.1) together with homogeneous Dirichlet boundary conditions. In what follows we call y the state and u the control (variable), respectively.

Recently, there has been significant interest in optimally controlling (1.1) subject to pointwise mixed control-state constraints of the type

$$a \le y + cu \le b \quad \text{almost everywhere (a.e.) in } Q, \tag{1.2}$$

Keywords and phrases. Bilateral control-state constraints, heat equation, mesh independence, optimal control, PDE-constrained optimization, semismooth Newton method.

[1] University of Sussex Department of Mathematics Mantell Building Falmer, Brighton BN1 9RF, UK.
michael.hintermueller@uni-graz.at

[2] Karl-Franzens-University of Graz Department of Mathematics and Scientific Computing Heinrichstrasse 36, 8010 Graz, Austria.
ian.kopacka@uni-graz.at; stefan.volkwein@uni-graz.at

Article published by EDP Sciences

where $a, b \in L^q(Q)$, for some $q > 2$ and with $a < b$, and $c \in L^\infty(Q)$, with $c \geq \varepsilon_c > 0$ a.e. in Q or $c \leq \varepsilon_c < 0$ a.e. in Q. Mixed control-state constraints are of interest in several respects: (i) They occur in Lavrentiev-type regularized state constrained optimal control problems, where typically $c \equiv \epsilon > 0$. In contrast to the measure-valuedness of the Lagrange multiplier associated with pure state constraints (*i.e.*, $c \equiv 0$ in (1.2)), the multiplier pertinent to the mixed control-state constraints enjoys $L^2(Q)$-regularity; see, *e.g.*, [20]. This makes analytical investigations as well as the development of fast numerical solution methods amenable. (ii) On the other hand, mixed control-state constraints may appear in their own right. For instance, for $c < 0$ (1.2) can be interpreted as to restrict the control by some multiple of the state. In thermal processes this might be intended to avoid material tensions due to a significant difference between the state (temperature) and the control (heating or cooling) action.

Based on earlier experience [10,12,13], here we propose a primal-dual active set or, equivalently, semismooth Newton method for the numerical solution of the underlying constrained optimal control problems. It turns out that the method converges locally at a superlinear rate in function space as well as in finite dimensions after discretization. In addition we prove that the convergence is mesh-independent. In both cases we extend currently available work for the control of elliptic partial differential equations [9,11] to the parabolic case. This is of particular interest with respect to the mesh-independence, since there are no such results available even in the case where the mixed control-state constraints are replaced by the more accessible pointwise control constraints $a \leq u \leq b$ a.e. in Q. We emphasize that our findings for the mixed control-state constraints readily carry over to the case of pure control constraints.

As the discretization of time-dependent PDE-constrained optimization problems naturally results in an extremely large scale problem, preconditioned iterative solvers for the resulting subsystems have to be employed. For research papers on reliable preconditioning in (unconstrained) optimal control of PDEs we refer to, *e.g.*, [2–4]. The literature on preconditioning techniques in the case of additional inequality constraints and, in particular, in connection with active set solvers is relatively scarce. Therefore, another goal of the present work is to introduce a preconditioning technique which is tailored to the active respectively inactive set structure of our solver and, hence, is able to handle additional pointwise inequality constraints efficiently.

The subsequent sections are organized as follows. In Section 2 we introduce the optimal control problem under consideration and present first-order necessary and sufficient optimality conditions. Section 3 is devoted to the development and convergence analysis of our solution algorithms. Then, in Section 4, we prove mesh independence of our method. Finally, numerical examples illustrating the efficient performance of our solver are discussed in Section 5. This section also contains an investigation of appropriate preconditioners for the iterative solvers considered.

2. The optimal control problem

In this section we formulate the optimal control problem with mixed pointwise control-state constraints and review first-order necessary optimality conditions.

2.1. The constrained optimal control problem

Suppose that Ω is an open and bounded subset of $\mathbb{R}^d$, $d \in \{2, 3\}$, with Lipschitz boundary $\Gamma = \partial\Omega$. For $T > 0$ we set $Q = (0, T) \times \Omega$ and $\Sigma = (0, T) \times \Gamma$. Moreover, by $L^2(0, T; H_0^1(\Omega))$ we denote the space of (equivalence classes) of measurable abstract functions $\varphi : [0, T] \to H_0^1(\Omega)$, which are square integrable, *i.e.*,

$$\int_0^T \|\varphi(t)\|^2_{H_0^1(\Omega)} \, \mathrm{d}t < \infty.$$

For the definition of Sobolev spaces we refer the reader, *e.g.*, to [1,6]. In particular, $H^{-1}(\Omega)$ stands for the dual space of $H_0^1(\Omega)$. Note that the space $H_0^1(\Omega)$ is continuously embedded in $L^6(\Omega)$ for spatial dimension $d \leq 3$. When t is fixed, the expression $\varphi(t)$ stands for the function $\varphi(t, \cdot)$ considered as a function in Ω only.

Recall that

$$W(0,T) = \left\{\varphi \in L^2(0,T;H_0^1(\Omega)) : \varphi_t \in L^2(0,T;H^{-1}(\Omega))\right\}$$

is a Hilbert space supplied with its common inner product; see [5], p. 473, for instance. Since $H_0^1(\Omega)$ is continuously embedded in $L^{r(d)}(\Omega)$ with $r(1) \in [2,+\infty]$, $r(2) \in [2,+\infty)$ and $r(d) \in [2, 2d/(d-2)]$ for all $d \geq 3$, we have that $W(0,T)$ is continuously embedded in $L^2(0,T;L^{r(d)}(\Omega))$. Further, it is well-known that $W(0,T)$ is continuously embedded in $C([0,T];L^2(\Omega))$; see [21], Theorem 3.10. Notice that $L^2(0,T;L^2(\Omega))$ can be identified with $L^2(Q)$. By Aubin's lemma [18], $W(0,T)$ is compactly embedded into $L^2(Q)$. Moreover, it follows from

$$\begin{aligned}\int_0^T \int_\Omega |\varphi(t,\mathbf{x})|^3 \,\mathrm{d}\mathbf{x}\mathrm{d}t &\leq \int_0^T \left(\int_\Omega |\varphi(t,\mathbf{x})|^2 \,\mathrm{d}\mathbf{x}\right)^{1/2} \left(\int_\Omega |\varphi(t,\mathbf{x})|^4 \,\mathrm{d}\mathbf{x}\right)^{1/2} \mathrm{d}t \\ &= \int_0^T \|\varphi(t)\|_{L^2(\Omega)} \|\varphi(t)\|_{L^4(\Omega)}^2 \,\mathrm{d}t \\ &\leq \|\varphi\|_{C([0,T];L^2(\Omega))} \|\varphi\|_{L^2(0,T;L^4(\Omega))}^2\end{aligned}$$

for any $\varphi \in W(0,T)$ that $W(0,T)$ is continuously embedded into $L^3(Q)$ for spatial dimension $d \leq 3$.

We consider a distributed optimal control problem for the heat equation with mixed pointwise control-state constraints. The goal is to minimize the cost function $J : W(0,T) \times L^2(Q) \to [0,\infty)$ given by

$$\begin{aligned}J(y,u) \quad &= \frac{1}{2}\int_0^T \int_\Omega \alpha_Q \,|y - z_Q|^2 \,\mathrm{d}\mathbf{x}\mathrm{d}t + \frac{1}{2}\int_\Omega \alpha_\Omega \,|y(T) - z_\Omega|^2 \,\mathrm{d}\mathbf{x} \\ &\quad + \frac{\kappa}{2}\int_0^T \int_\Omega |u - u_d|^2 \,\mathrm{d}\mathbf{x}\mathrm{d}t,\end{aligned} \tag{2.1}$$

where the state y and the control u are coupled by the linear boundary value problem

$$y_t(t,\mathbf{x}) - \alpha\Delta y(t,\mathbf{x}) = f(t,\mathbf{x}) + u(t,\mathbf{x}) \qquad \text{for almost all } (t,\mathbf{x}) \in Q, \tag{2.2a}$$

$$y(t,\mathbf{s}) = 0 \qquad \text{for almost all } (t,\mathbf{s}) \in \Sigma, \tag{2.2b}$$

$$y(0,\mathbf{x}) = y_\circ(\mathbf{x}) \qquad \text{for almost all } \mathbf{x} \in \Omega. \tag{2.2c}$$

In (2.1) we assume that α_Q and α_Ω are non-negative weights satisfying $\alpha_Q \in L^\infty(Q)$ and $\alpha_\Omega \in L^\infty(\Omega)$, respectively. The desired states $z_Q \in L^2(Q)$, $z_\Omega \in L^2(\Omega)$ and the nominal control $u_d \in L^3(Q)$ are given, and $\kappa > 0$ denotes a regularization parameter. For the data in (2.2) we suppose that the inhomogeneity f belongs to $L^2(0,T;H^{-1}(\Omega))$, $\alpha > 0$ holds true, and the initial state satisfies $y_\circ \in L^2(\Omega)$. It is well-known that for any $u \in L^2(Q)$ there exists a unique solution $y \in W(0,T)$ of the state equation (2.2). Moreover, the mapping $u \mapsto y(u)$ is continuous from $L^2(Q)$ to $W(0,T)$; see [14,21]. If, in addition, $y_\circ \in H_0^1(\Omega)$, $f \in L^2(Q)$ hold and Ω is sufficiently smooth (*e.g.*, Ω is convex with Lipschitz-continuous boundary), then the state y belongs even to $L^2(0,T;H^2(\Omega) \cap H_0^1(\Omega)) \cap H^1(0,T;L^2(\Omega))$.

We also impose bilateral pointwise control-state constraints. For that purpose let $a, b \in L^3(Q)$ be given lower and upper bounds, respectively. Moreover, let $c \in L^\infty(Q)$ satisfy $c \geq \varepsilon_c > 0$ (or $c \leq \varepsilon_c < 0$) for almost all (f.a.a.) $(t,\mathbf{x}) \in Q$. We define the two Banach spaces

$$X = W(0,T) \times L^2(Q), \quad Y = L^2(0,T;H_0^1(\Omega)),$$

and denote the common compact embedding operators by $\imath : W(0,T) \to L^2(Q)$ and $\jmath : L^2(Q) \to Y'$. Then admissible state-control pairs (y,u) are required to belong to the closed convex set

$$X_{\mathsf{ad}} = \left\{(y,u) \in X \,\middle|\, a \leq \imath y + cu \leq b \text{ a.e. in } Q\right\}. \tag{2.3}$$

For a compact formulation of the optimal control problem we introduce the affine linear mapping $e : X \to Y'$ by

$$\langle e(y,u),\varphi\rangle_{Y',Y} = \int_0^T \langle y_t(t) - f(t), \varphi(t)\rangle_{H^{-1}(\Omega),H_0^1(\Omega)}\mathrm{d}t + \int_\Omega \alpha\nabla y\cdot\nabla\varphi - u\varphi\,\mathrm{d}\mathbf{x}\mathrm{d}t$$

for $(y,u) \in X$ and $\varphi \in Y$, where the dual Y' of Y is identified with $L^2(0,T;H^{-1}(\Omega))$, and $\langle\cdot,\cdot\rangle_{H^{-1}(\Omega),H_0^1(\Omega)}$ denotes the duality pairing between $H_0^1(\Omega)$ and its dual $H^{-1}(\Omega)$. The feasible set is given by

$$\Phi(\mathbf{P}) = \big\{x = (y,u) \in X_{\mathsf{ad}} \,\big|\, e(x) = 0 \text{ in } Y' \text{ and } y(0) = y_\circ \text{ in } L^2(\Omega)\big\}.$$

Throughout the paper we assume that $\Phi(\mathbf{P}) \neq \emptyset$.

Our infinite dimensional optimal control problem now reads

$$\min J(x) \quad \text{subject to (s.t.)} \quad x \in \Phi(\mathbf{P}). \tag{\mathbf{P}}$$

Since $\Phi(\mathbf{P}) \neq \emptyset$ by assumption, there exists a unique solution $x^* = (y^*,u^*)$ of ($\mathbf{P}$). The uniqueness follows from the strict convexity properties of the objective functional. If $y_\circ \in H_0^1(\Omega)$ and $f \in L^2(Q)$ hold, then $y^* \in L^2(0,T;H^2(\Omega)\cap H_0^1(\Omega))\cap H^1(0,T;L^2(\Omega))$.

The first-order optimality conditions of ($\mathbf{P}$) are stated in the next theorem.

Theorem 2.1. *Suppose that $\Phi(\mathbf{P}) \neq \emptyset$ and that $x^* = (y^*,u^*) \in \Phi(\mathbf{P})$ is the solution of ($\mathbf{P}$). Then there exists a unique Lagrange multiplier pair $(p^*,\lambda^*) \in W(0,T)\times L^2(Q)$ satisfying, together with (y^*,u^*), the dual system (here written in its strong form)*

$$-p_t^* - \alpha\Delta p^* + \lambda^* = -\alpha_Q(y^* - z_Q) \quad \text{in } Q, \tag{2.4a}$$
$$p^* = 0 \quad \text{on } \Sigma, \tag{2.4b}$$
$$p^*(T) = -\alpha_\Omega(y^*(T) - z_\Omega) \quad \text{in } \Omega, \tag{2.4c}$$
$$\kappa(u^* - u_d) - p^* + c\lambda^* = 0 \quad \text{in } Q, \tag{2.4d}$$
$$\lambda^* = \max(0,\lambda^* + \sigma(y^* + cu^* - b)) + \min(0,\lambda^* + \sigma(y^* + cu^* - a)) \quad \text{on } Q, \tag{2.4e}$$

where σ is an arbitrary function in $L^\infty(Q)$ with $\sigma(t,\mathbf{x}) \geq \underline{\sigma} > 0$ f.a.a. $(t,\mathbf{x}) \in Q$. In (2.4e) the min- *and* max-*operations are interpreted in the pointwise almost everywhere sense.*

Proof. The proof follows from arguments analogous to those given in [9], Section 2. □

Note that (2.4e) is a nonlinear complementarity problem (NCP) function based reformulation of the complementarity systems

$$\lambda_a^* \geq 0, \quad a - y^* - cu^* \leq 0, \quad \lambda_a^*(a - y^* - cu^*) = 0, \quad \text{a.e. in } Q, \tag{2.5}$$
$$\lambda_b^* \geq 0, \quad y^* + cu^* - b \leq 0, \quad \lambda_b^*(y^* + cu^* - b) = 0, \quad \text{a.e. in } Q, \tag{2.6}$$

with $\lambda^* = \lambda_b^* - \lambda_a^*$.

2.2. The reduced optimal control problem

Since, for any $u \in L^2(Q)$, there exists a unique solution $y = y(u) \in W(0,T)$ of the state equation (2.2), we can define the bounded and affine linear solution operator

$$\mathcal{S} : L^2(Q) \to W(0,T), \quad u \mapsto \mathcal{A}^{-1}(f + \jmath u),$$

where $\mathcal{A} = (\frac{\partial}{\partial t} - \alpha\Delta) : W(0,T) \to Y'$ is linear and bounded. For later use, let $\mathcal{A}^*$ denote the adjoint operator of $\mathcal{A}$. Next we introduce the so-called reduced cost functional

$$\hat{J}(u) = J(\mathcal{S}(u), u).$$

Further, we define the set of admissible controls

$$U_{\mathsf{ad}} = \big\{u \in L^2(Q) \,\big|\, a \le \imath\mathcal{S}(u) + cu \le b \text{ a.e. in } Q\big\},$$

which has the following properties.

Proposition 2.2. *The set U_{ad} is convex, closed and bounded in $L^2(Q)$.*

Proof. Closedness and convexity follow immediately. Therefore, we only focus on the boundedness of U_{ad}. Since $y = S(u)$, *i.e.*, $\mathcal{A}y = f + \jmath u$, we have $a \le \imath y + c\,u \le b$ a.e. in Q. Setting $v := \imath y + c\,u$ we obtain

$$\big(\mathcal{A} + \jmath c^{-1}\imath\,\mathrm{id}\big)\, y = f + \jmath c^{-1} v \quad \text{and} \quad a \le v \le b \text{ a.e. in } Q.$$

From this we infer

$$\frac{\|y\|^2_{L^2(Q)}}{\|c\|_{L^\infty(Q)}} \le \frac{1}{2}\|y_\circ\|^2_{L^2(\Omega)} + \big(\|f\|_{L^2(Q)} + |\epsilon_c|^{-1}\max(\|a\|_{L^2(Q)}, \|b\|_{L^2(Q)})\big)\,\|y\|_{L^2(Q)}.$$

Hence, y is bounded in $L^2(Q)$. Since $c \in L^\infty(Q)$ with $|c| \ge \epsilon_c > 0$ a.e. in Q, we infer that u is bounded in $L^2(Q)$. This proves the assertion. □

Remark 2.3. In the context of Lavrentiev-type regularization of state constrained optimal control problems one is interested in setting $c \equiv \epsilon_n > 0$ and studying $\epsilon_n \downarrow 0$. In this case, our arguments in the proof of Proposition 2.2 yield the boundedness of $y = y_{\epsilon_n}$ in $L^2(Q)$ uniformly with respect to ϵ_n.

Problem (**P**) can be equivalently expressed as

$$\min \hat{J}(u) \quad \text{s.t.} \quad u \in U_{\mathsf{ad}}. \tag{$\hat{\mathbf{P}}$}$$

For accessing the gradient of $\hat{J}$ we have to guarantee differentiability of $\mathcal{S}$. As in [21], one argues that $\mathcal{S}$ is continuously differentiable as a mapping from $L^2(Q)$ to $W(0,T)$. The action of the derivative $\mathcal{S}'(u)$ (we also write $y'(u)$) on some $v \in L^2(Q)$, *i.e.*, $\mathcal{S}'(u)v = w$, is characterized by the solution w of the initial-boundary value problem

$$\begin{aligned} w_t - \alpha\Delta w &= \jmath v && \text{in } Q,\\ w &= 0 && \text{on } \Sigma,\\ w(0) &= 0 && \text{in } \Omega. \end{aligned} \tag{2.7}$$

Considering the adjoint equations (2.4a)–(2.4c) with $y = y(u)$, we find that the adjoint state depends on u and λ, *i.e.*, $p = p(u,\lambda)$. Similarly, we obtain the differentiability of the adjoint state $p(u,\lambda)$ considered as a function of u and λ.

The derivative of $\hat{J}$ at a point $u \in L^2(Q)$ is represented by

$$\hat{J}'(u) = \mathcal{S}'(u)^* \frac{\partial J(\mathcal{S}(u),u)}{\partial y} + \frac{\partial J(\mathcal{S}(u),u)}{\partial u} = -\mathfrak{p}(u) + \kappa(u - u_d) \quad \text{in } Q,$$

where $\mathfrak{p}(u)$ solves the equation

$$\begin{aligned} -p_t - \alpha\Delta p &= \alpha_Q(z_Q - y(u)) && \text{in } Q,\\ p &= 0 && \text{on } \Sigma,\\ p(T) &= \alpha_\Omega(z_\Omega - y(u)(T)) && \text{in } \Omega. \end{aligned}$$

The first-order necessary optimality condition of ($\hat{\mathbf{P}}$) is given by the variational inequality

$$\langle \hat{J}'(u^*), u - u^* \rangle_{L^2(Q)} \geq 0 \quad \text{for all } u \in U_{\mathsf{ad}}.$$

This is equivalent to

$$\kappa(u^* - u_d) - p^* + c\lambda^* = 0, \tag{2.8a}$$

$$\lambda^* = \max(0, \lambda^* + \sigma(y(u^*) + cu^* - b)) + \min(0, \lambda^* + \sigma(y(u^*) + cu^* - a)) \tag{2.8b}$$

in Q for some arbitrarily fixed, positive $\sigma \in L^\infty(Q)$, and with $p^* = p(u^*)$ solving (2.4a)–(2.4c). Thus, the first-order necessary optimality conditions for ($\hat{\mathbf{P}}$) are given by (2.8) together with (2.4a)–(2.4c).

To express ($\hat{\mathbf{P}}$) as a bilateral control constrained problem we set $\tilde{a} = a - \imath_3\mathcal{A}^{-1}f$, $\tilde{b} = b - \imath_3\mathcal{A}^{-1}f$, with $\imath_3$ denoting the continuous embedding operator from $W(0,T)$ into $L^3(Q)$ for $d \leq 3$. Moreover, we define the linear and bounded operators $\mathcal{T} = \imath\mathcal{A}^{-1}\jmath : L^2(Q) \to L^2(Q)$ and $\mathcal{F} = \mathcal{T} + c\,\mathrm{id} : L^2(Q) \to L^2(Q)$. By assumption $c \neq 0$ is satisfied. Since $\imath$ is compact and $\jmath$ as well as $\mathcal{A}^{-1}$ are continuous, the operator $\mathcal{T}$ is compact. If

$$-c \text{ is no eigenvalue of } \mathcal{T}, \tag{2.9}$$

we infer from the Fredholm theory that the linear operator $\mathcal{F}$ admits a (unique) inverse. Thus, ($\hat{\mathbf{P}}$) can be expressed equivalently as a bilaterally control constrained problem for the new control variable $v := \mathcal{F}u$

$$\min \tilde{J}(v) \quad \text{s.t.} \quad v \in V_{\mathsf{ad}} = \{v \in L^2(Q) \,|\, \tilde{a} \leq v \leq \tilde{b} \text{ a.e. in } Q\}, \tag{$\tilde{\mathbf{P}}$}$$

with $\tilde{J} = \hat{J} \circ \mathcal{F}^{-1}$. Notice that ($\tilde{\mathbf{P}}$) is a minimization problem with bilateral control constraints, but with no equality constraints. Of course, $v^* = \mathcal{F}u^*$ is the solution of ($\tilde{\mathbf{P}}$). We will make use of ($\tilde{\mathbf{P}}$) when establishing a mesh-independence principle of our algorithm in Section 4.

Remark 2.4. Note that the smoothness of the bounds $\tilde{a}$ and $\tilde{b}$ depends on the smoothness of a, b and $\mathcal{A}^{-1}f$. In particular, if a and b are constant and $f \equiv 0$ holds, ($\tilde{\mathbf{P}}$) is an optimal control problem with constant box constraints. On the other hand, higher regularity properties can be ensured by proper assumptions on a, b and f. In our numerical test examples carried out in Section 5 we have $\tilde{a}$, $\tilde{b} \in C(\overline{\Omega}) \cap C^\infty(\Omega)$.

In order to ease the notation, in what follows we frequently neglect the embedding operators.

3. The semismooth Newton method

In this section, to solve ($\mathbf{P}$) numerically a Newton-type algorithm is applied to the first-order necessary optimality conditions of ($\hat{\mathbf{P}}$). For the generalized (Newton) differentiation of the min- and max-operators in function space we rely on the following definition which is due to [12].

Definition 3.1. Let V, W be two Banach spaces, $S \subset V$ a non-empty open set, $F : S \to W$ a given mapping, and $v^* \in S$. If there exists a neighborhood $N(v^*) \subset S$ and a family of mappings $G : N(v^*) \to L(V,W)$ such that

$$\lim_{\|v\|_V \downarrow 0} \frac{1}{\|v\|_V} \|F(v^* + v) - F(v^*) - G(v^* + v)v\|_W = 0, \tag{3.1}$$

then F is called *Newton-differentiable at* v^*, and $G(v^*)$ is said to be a *generalized derivative (or Newton map) for* F *at* v^*.

Here, $L(V,W)$ denotes the Banach space of all bounded and linear operators from V to W endowed with the common norm. Moreover, we write $L(V) = L(V,V)$.

Remark 3.2. The function $\max : L^p(Q) \to L^s(Q)$ is Newton differentiable for $1 \le s < p \le \infty$ (see [12]). If $F : L^r(Q) \to L^p(Q)$ is Fréchet differentiable for some $1 \le r \le \infty$, then the function

$$(t,\mathbf{x}) \mapsto \chi_A(t,\mathbf{x}) \cdot \nabla F(u(t,\mathbf{x})), \quad (t,\mathbf{x}) \in Q, \tag{3.2}$$

is a generalized derivative of $\max(0, F(\cdot)) : L^r(Q) \to L^s(Q)$. Here, χ_A denotes the characteristic function of the set $A \subset Q$, where $F(u(\cdot))$ is positive, *i.e.*, $\chi_A(t,\mathbf{x}) = 1$ if $F(u(t,\mathbf{x})) > 0$ and $\chi_A(t,\mathbf{x}) = 0$ otherwise. From $\min(0, F(\cdot)) = -\max(0, -F(\cdot))$, we see that an analogous differentiation formula holds true for the min-function.

Next observe that (2.8a) in (2.4a)–(2.4c) with $y^* = y(u^*)$ yields

$$(\mathcal{A}^* + c^{-1}\,\mathrm{id})p^* = \alpha_Q(z_Q - y(u^*)) + \kappa c^{-1}(u^* - u_d). \tag{3.3}$$

Consequently, assuming that

$$-c^{-1} \text{ is no eigenvalue of } \mathcal{A}^*, \tag{3.4}$$

we have $p^* = p(u^*) \in Y$ uniquely. Parabolic regularity results yield $p(u^*) \in W(0,T)$. This relation holds true whenever $y = y(u)$ in the right hand side of the adjoint system and $\kappa(u - u_d) - p + c\lambda = 0$ on Σ. Further we conclude $\lambda = \lambda(u)$.

Choosing $\sigma = \kappa/c^2 \in L^\infty(Q)$ with $\sigma \ge \kappa/\varepsilon_c^2 > 0$ in (2.8b) and taking into account (2.8a), we obtain

$$\begin{aligned} c^{-1}\left(\kappa(u_d - u^*) + p^*(u^*)\right) - \max\left(0, c^{-1}(\kappa u_d + p^*(u^*)) + \kappa c^{-2}(y(u^*) - b)\right) \\ - \min\left(0, c^{-1}(\kappa u_d + p^*(u^*)) + \kappa c^{-2}(y(u^*) - a)\right) = 0. \end{aligned} \tag{3.5}$$

Thus, we introduce the mapping $F : L^2(Q) \to L^2(Q)$ by

$$\begin{aligned} F(u) = c^{-1}\left(\kappa(u_d - u) + p(u)\right) - \max\left(0, c^{-1}(\kappa u_d + p(u)) + \kappa c^{-2}(y(u) - b)\right) \\ - \min\left(0, c^{-1}(\kappa u_d + p(u)) + \kappa c^{-2}(y(u) - a)\right). \end{aligned} \tag{3.6}$$

Then, (3.5) becomes the nonsmooth operator equation

$$F(u^*) = 0 \quad \text{in } L^2(Q). \tag{3.7}$$

Now suppose $\bar{u} \in L^2(Q)$ is some given approximation of u^*. Then, since $y(\bar{u})$ as well as $p(\bar{u})$ are continuously differentiable from $L^2(Q)$ to $W(0,T) \hookrightarrow L^3(Q)$ and $u_d, a, b \in L^3(Q)$ by assumption, Remark 3.2 provides Newton differentiability of the min- and max-terms in (3.5), respectively. A particular Newton map of F at $\bar{u}$ in direction $u \in L^2(Q)$ is given by

$$\begin{aligned} G(\bar{u})u = \frac{1}{c}\left(p'(\bar{u})u - \kappa u\right) - \frac{1}{c^2}\chi_{\left\{\lambda(\bar{u}) + \frac{\kappa}{c^2}(y(\bar{u}) + c\bar{u} - b) > 0\right\}}\left(cp'(\bar{u})u + \kappa y'(\bar{u})u\right) \\ - \frac{1}{c^2}\chi_{\left\{\lambda(\bar{u}) + \frac{\kappa}{c^2}(y(\bar{u}) + c\bar{u} - a) < 0\right\}}\left(cp'(\bar{u})u + \kappa y'(\bar{u})u\right), \end{aligned} \tag{3.8}$$

where $\delta y = y'(\bar{u})u \in W(0,T)$ solves the linearized state equations

$$\begin{aligned} \delta y_t - \alpha\Delta\delta y &= u && \text{in } Q, && (3.9a)\\ \delta y &= 0 && \text{on } \Sigma, && (3.9b)\\ \delta y(0) &= 0 && \text{in } \Omega, && (3.9c) \end{aligned}$$

and $\delta p = p'(\bar{u})u \in W(0,T)$ solves the linearized adjoint system

$$\begin{aligned} -\delta p_t - \alpha\Delta\delta p + c^{-1}\delta p &= -\alpha_Q \delta y + \kappa c^{-1} u && \text{in } Q, && (3.10a)\\ \delta p &= 0 && \text{on } \Sigma, && (3.10b)\\ \delta p(T) &= -\alpha_\Omega \delta y(T) && \text{in } \Omega. && (3.10c) \end{aligned}$$

It follows by standard arguments that for every $u \in L^2(Q)$ there exist uniquely determined $\delta y \in W(0,T)$ and $\delta p \in W(0,T)$ solving (3.9) and (3.10), respectively.

In Algorithm 1 we formulate the corresponding generalized Newton method for finding $u^* \in L^2(Q)$ such that (3.7) holds true. Due to the additional regularity of F implied by (3.1) we call it a semismooth Newton method.

Algorithm 1 (semismooth Newton method).

1: Choose $u^0 \in L^2(Q)$, and set $k = 0$.
2: **repeat**
3: Compute $G(u^k)$ according to (3.8) and solve for δu^k:

$$G(u^k)\delta u^k = -F(u^k)$$

with F given by (3.6).
4: Set $u^{k+1} = u^k + \delta u^k$, and $k = k + 1$.
5: **until** some stopping rule is satisfied.

We have the following convergence result; see [12].

Theorem 3.3. *Let $\{u^k\}_{k\in\mathbb{N}}$ be a sequence generated by Algorithm 1. Then, $\{u^k\}_{k\in\mathbb{N}}$ converges to the solution $u^* \in U_{\mathsf{ad}}$ of (**P**) at a q-superlinear rate provided that $u^0 \in L^2(Q)$ is sufficiently close to u^*.*

Algorithm 1 can be expressed equivalently as a primal-dual active-set strategy. In fact, using (3.2) and defining $\lambda^k = c^{-1}\left(\kappa(u_d - u^k) + p(u^k)\right)$ and

$$\begin{aligned} A_b^k &= \left\{(t,\mathbf{x}) \in Q \,\middle|\, \lambda^k + \frac{\kappa}{c^2}\left(y^k(u^k) + cu^k - b\right) > 0 \text{ a.e.}\right\},\\ A_a^k &= \left\{(t,\mathbf{x}) \in Q \,\middle|\, \lambda^k + \frac{\kappa}{c^2}\left(y^k(u^k) + cu^k - a\right) < 0 \text{ a.e.}\right\},\\ \mathcal{I}^k &= Q \setminus A^k, \quad \text{with } A^k = A_b^k \cup A_a^k, \end{aligned} \tag{3.11}$$

we obtain the following linearization of (3.5) at $(u, y(u), p(u))$ with respect to the independent variable u:

$$\begin{aligned} &\frac{1}{c}\left(\kappa(u_d - u^k - \delta u^k) + p(u^k) + p'(u^k)\delta u^k\right)\\ &-\frac{1}{c^2}\chi_{A_b^k}\Big(c(\kappa u_d + p(u^k) + p'(u^k)\delta u^k) + \kappa(y(u^k) + y'(u^k)\delta u^k - b)\Big)\\ &-\frac{1}{c^2}\chi_{A_a^k}\Big(c(\kappa u_d + p(u^k) + p'(u^k)\delta u^k) + \kappa(y(u^k) + y'(u^k)\delta u^k - a)\Big) = 0. \end{aligned} \tag{3.12}$$

Here $\delta u \in L^2(Q)$ represents the increment. A closer look reveals:

$$y(u^k) + y'(u^k)\delta u^k + cu^{k+1} = a \text{ on } A_a^k, \tag{3.13a}$$
$$y(u^k) + y'(u^k)\delta u^k + cu^{k+1} = b \text{ on } A_b^k, \tag{3.13b}$$
$$\kappa(u^{k+1} - u_d) - p(u^k) - p'(u^k)\delta u^k = 0 \text{ on } \mathcal{I}^k, \tag{3.13c}$$

where $u^{k+1} = u^k + \delta u^k$. Note that (3.13c) can be viewed as

$$\lambda^k + \delta\lambda^k = 0 \text{ on } \mathcal{I}^k, \tag{3.13d}$$

with $\delta\lambda^k = -\kappa\delta u^k + p'(u^k)\delta u^k$. The active respectively inactive set behavior of the variables in (3.13) motivates Algorithm 2.

Algorithm 2 (primal-dual active set strategy).

1: Choose starting values λ^0 and u^0, compute $y^0 = \mathcal{S}(u^0)$ and p^0 satisfying (2.4a)–(2.4d). Set $k = 0$.
2: **repeat**
3: Determine the active sets A_a^k, A_b^k and the inactive set I^k according to (3.11).
4: Compute the (unique) solution $x^k = (u^k, y^k)$ with pertinent multiplier λ^k and adjoint state p^k of

$$\begin{cases} \min \quad J(x) \quad \text{over} \quad x = (y,u) \in X \\ \text{s.t.} \quad e(x) = 0,\ y(0) = y_\circ, \\ \qquad y + cu = a \text{ on } A_a^k \text{ and } y + cu = b \text{ on } A_b^k. \end{cases} \tag{3.14}$$

5: Set $k = k + 1$.
6: **until** some stopping rule is satisfied.

Utilizing the setting on A_a^k and A_b^k, note that the feasible set of the minimization problem (3.14) in step 4 of Algorithm 2 becomes

$$y_t - \alpha\Delta y + c^{-1}\chi_{A_a^k \cup A_b^k} y = f + \chi_{I^k} u + c^{-1}(\chi_{A_a^k} a + \chi_{A_b^k} b),$$
$$y = 0 \quad \text{on } \Sigma, \quad y(0) = y_\circ \quad \text{in } \Omega.$$

Given u, this system admits a unique solution. The radial unboundedness of J with respect to u on the feasible set now guarantees the existence of a unique solution to (3.14).

Let $\delta y^k = y'(u^k)\delta u^k$ and $\delta p^k = p'(u^k)\delta u^k$. Then $y^{k+1} = y^k + \delta y^k$ solves

$$y_t^{k+1} - \alpha\Delta y^{k+1} = f + u^{k+1} \quad \text{in } Q, \tag{3.15a}$$
$$y^{k+1} = 0 \quad \text{on } \Sigma, \tag{3.15b}$$
$$y^{k+1}(0) = y_\circ \quad \text{in } \Omega. \tag{3.15c}$$

Furthermore, $p^{k+1} = p^k + \delta p^k$ satisfies

$$-p_t^{k+1} - \alpha\Delta p^{k+1} + \lambda^{k+1} = -\alpha_Q(y^{k+1} - z_Q) \quad \text{in } Q, \tag{3.16a}$$
$$p^{k+1} = 0 \quad \text{on } \Sigma, \tag{3.16b}$$
$$p^{k+1}(T) = -\alpha_\Omega(y^{k+1}(T) - z_\Omega) \quad \text{in } \Omega. \tag{3.16c}$$

Utilizing (3.13) we derive from (3.15a)

$$y_t^{k+1} - \alpha\Delta y^{k+1} + \frac{1}{c}\chi_{A^k} y^{k+1} - \frac{1}{\kappa}\chi_{I^k} p^{k+1} = f + \frac{1}{c}\left(\chi_{A_a^k} a + \chi_{A_b^k} b\right) + \chi_{I^k} u_d \tag{3.17}$$

with $A^k = A_a^k \cup A_b^k$. Since

$$\lambda^{k+1} = \lambda^k + \delta\lambda^k = \frac{1}{c}\left(p^{k+1} + \kappa(u_d - u^{k+1})\right)$$

(see (2.4d)), we conclude from (3.13d) that

$$-p_t^{k+1} + \left(-\alpha\Delta + \frac{1}{c}\chi_{A^k}\right)p^{k+1} + \left(\alpha_Q + \frac{\kappa}{c^2}\chi_{A^k}\right)y^{k+1} = \alpha_Q z_Q - \frac{\kappa}{c}\left(\chi_{A^k} u_d - \frac{1}{c}\left(\chi_{A_a^k} a + \chi_{A_b^k} b\right)\right). \tag{3.18}$$

Summarizing, (3.17) and (3.18) can be written compactly as

$$\left(\begin{array}{c|c} \alpha_Q + \frac{\kappa}{c^2}\chi_{A^k} & -\frac{\partial}{\partial t} - \alpha\Delta + \frac{1}{c}\chi_{A^k} \\ \hline \frac{\partial}{\partial t} - \alpha\Delta + \frac{1}{c}\chi_{A^k} & -\frac{1}{\kappa}\chi_{I^k} \end{array}\right)\left(\begin{array}{c} y^{k+1} \\ p^{k+1} \end{array}\right) = \left(\begin{array}{c} \alpha_Q z_Q - \frac{\kappa}{c}\left(\chi_{A^k} u_d - \frac{1}{c}(\chi_{A_a^k} a + \chi_{A_b^k} b)\right) \\ f + \frac{1}{c}\left(\chi_{A_a^k} a + \chi_{A_b^k} b\right) + \chi_{I^k} u_d \end{array}\right). \tag{3.19}$$

This is the reduced form of the Newton system which is used in our numerics; compare (5.1) in Section 5.

4. Mesh-independence

In this section we give sufficient conditions for the mesh-independent convergence of Algorithm 2. The proof technique is based on a combination of arguments in [15] and [9]. Throughout we assume $y_\circ$, z_Q, z_Ω, α_Q, and α_Ω are sufficiently regular; see [15], Table 1.

We proceed as in [15], Section 3, and [8]. Let Ω_h be a family of grids depending on the parameter $h > 0$. On these grids, P_h^0 and P_h^1 are the spaces of all piecewise constant respectively piecewise linear finite elements. Furthermore, we define the operators of orthogonal projections by $\mathcal{R}_h^i : L^2(\Omega) \to P_h^i$ for $i \in \{0, 1\}$. We suppose that the finite element grids are chosen in such a way that

$$\|\varphi - \mathcal{R}_h^i\varphi\|_{H^r(\Omega)} \le c_\Omega h^{s-r} \|\varphi\|_{H^s(\Omega)} \quad \text{for all } \varphi \in H^s(\Omega),$$

where $r \in [0, i]$, $s \in [r, i+1]$, $c_\Omega > 0$, and $i \in \{0, 1\}$. Next we introduce approximations for functions defined on Q. Let $t_j = jh_t$, $0 \le j \le n_t$, be a chosen grid in $[0, T]$ with step size $h_t = T/n_t$. To simplify the presentation, *i.e.*, to avoid terms of the type $\mathcal{O}(h_t + h^2)$ in our error analysis, we couple the time and spatial discretization in the following manner: we suppose that there are constants $0 < c_1 \le c_2$ such that

$$c_1 h^2 \le h_t \le c_2 h^2; \tag{4.1}$$

see, *e.g.*, [8].

Next we define the finite dimensional spaces

$$\begin{aligned} S_h^0 &= \left\{\varphi_h : Q \to \mathbb{R} \,\middle|\, \varphi_h(t) = \varphi_h(t_j) \in P_h^0,\ t \in [t_{j-1}, t_j),\ 1 \le j \le n_t\right\}, \\ S_h^1 &= \left\{\varphi_h : Q \to \mathbb{R} \,\middle|\, \varphi_h(t) = P_j^{j-1}(\varphi_h)(t),\ t \in [t_{j-1}, t_j],\ \varphi_h(t_j) \in P_h^1,\ 1 \le j \le n_t\right\}, \end{aligned}$$

where

$$P_j^{j-1}(\varphi_h)(t) = \frac{t_j - t}{h_t}\varphi_h(t_{j-1}) + \frac{t - t_{j-1}}{h_t}\varphi_h(t_j) \quad \text{for } t \in [t_{j-1}, t_j].$$

The corresponding restriction operators are $\mathcal{G}_h^i : L^2(Q) \to S_h^i$ with $i \in \{0, 1\}$. Note that the elements of S_h^0 are piecewise constant in space and time (on intervals $[t_{j-1}, t_j)$), and the elements of S_h^1 are piecewise linear in space and time.

The operator $\mathcal{A}^{-1}$ is approximated by $\mathcal{A}_h^{-1} : Y' \to S_h^1$ as follows: for every $g \in Y'$ the element $y_h = \mathcal{A}_h^{-1} g \in S_h^1$ solves

$$\int_\Omega \frac{y_h(t_j) - y_h(t_{j-1})}{h_t} \varphi_h + \alpha \nabla y_h(t_j) \cdot \nabla \varphi_h \, \mathrm{d}\mathbf{x} = \left\langle \frac{1}{h_t} \int_{t_{j-1}}^{t_j} g(t) \, \mathrm{d}t, \varphi_h \right\rangle_{H^{-1}(\Omega), H_0^1(\Omega)} \tag{4.2a}$$

for all $1 \le j \le n_t$ and $\varphi_h \in P_h^1$ together with the boundary and initial conditions

$$y_h(t_j, \cdot) = 0 \text{ on } \Gamma, \ 1 \le j \le n_t, \quad \text{and} \quad y_h(0) = \mathcal{R}_h^1 y_\circ \text{ in } \Omega. \tag{4.2b}$$

Hence, we have applied the implicit Euler method for the time integration and the piecewise linear finite element method for the spatial discretization.

By $\mathcal{S}_h : L^2(Q) \to S_h^1 \subset W(0,T)$, $u \mapsto \mathcal{A}_h^{-1}(f + u)$ we introduce an approximation of the affine linear operator $\mathcal{S}$ introduced in Section 2.2. The discretized set of admissible controls is defined as

$$U_{\mathsf{ad}}^h = \left\{ u_h \in S_h^0 \,\middle|\, a \le \imath_h \mathcal{S}_h(u_h) + c u_h \le b \text{ a.e. in } Q \right\},$$

where $\imath_h$ denotes the (orthogonal) projection of Y onto S_h^0. We may consider $\jmath|_{S_h^0} = \imath_h^*$, where $\imath_h^*$ denotes the adjoint of $\imath_h$. Further, for simplicity we assume $a, b, c, u_d, \alpha_Q \in S_h^0$; otherwise a corresponding discretization has to be considered.

Then, ($\hat{\mathbf{P}}$) is discretized by the family of finite-dimensional problems

$$\min \hat{J}_h(u_h) \quad \text{s.t.} \quad u_h \in U_{\mathsf{ad}}^h \tag{$\hat{\mathbf{P}}_h$}$$

with $\hat{J}_h(u_h) = J(\mathcal{S}_h(u_h), u_h)$ for $u_h \in S_h^0$.

Proceeding as in Section 3 we formulate a semismooth Newton method for ($\hat{\mathbf{P}}_h$). For that purpose we introduce the operator $F_h : S_h^0 \to S_h^0$,

$$\begin{aligned} F_h(u_h) = {} & c^{-1} \left(\kappa(u_d - u_h) + \imath_h p_h(u_h) \right) \\ & - \max\left(0, c^{-1}(\kappa u_d + \imath_h p_h(u_h)) + \kappa c^{-2}(\imath_h y_h(u_h) - b)\right) \\ & - \min\left(0, c^{-1}(\kappa u_d + \imath_h p_h(u_h)) + \kappa c^{-2}(\imath_h y_h(u_h) - a)\right), \end{aligned} \tag{4.3}$$

where $y_h(u_h) = \mathcal{A}_h^{-1}(f + \imath_h^* u_h) \in S_h^1$ and $p_h(u_h) \in S_h^1$ solves the adjoint equation of (4.2):

$$\int_\Omega \frac{p_h(t_j) - p_h(t_{j-1})}{h_t} \varphi_h - \alpha \nabla p_h(t_{j-1}) \cdot \nabla \varphi_h \, \mathrm{d}\mathbf{x} = \frac{1}{h_t} \int_{t_{j-1}}^{t_j} \left(\alpha_Q(t)(y_h(t) - z_Q(t)), \varphi_h\right)_{L^2(\Omega)} \mathrm{d}t \tag{4.4a}$$

for all $1 \le j \le n_t$ and $\varphi_h \in P_h^1$ together with the boundary and initial conditions

$$\begin{aligned} p_h(t_j, \cdot) &= 0 && \text{on } \Gamma, \ 0 \le j \le n_t - 1, \\ p_h(0) &= \mathcal{R}_h^1\big(\alpha_\Omega(z_\Omega - y_h(T))\big) && \text{in } \Omega. \end{aligned} \tag{4.4b}$$

As a mapping between finite dimensional spaces, F_h is Newton-differentiable. The generalized derivative of F_h at a point $\bar{u}_h \in S_h^0$ in direction u_h is given by

$$\begin{aligned} G_h(\bar{u}_h)u_h = {} & c^{-1} \left(\imath_h p_h'(\bar{u}_h)u_h - \kappa u_h \right) \\ & - c^{-2} \chi_{\left\{\lambda_h(\bar{u}_h) + \frac{\kappa}{c^2}(\imath_h y_h(\bar{u}_h) + c\bar{u}_h - b) > 0\right\}} \left(c\imath_h p_h'(\bar{u}_h)u_h + \kappa \imath_h y_h'(\bar{u}_h)u_h \right) \\ & - c^{-2} \chi_{\left\{\lambda_h(\bar{u}_h) + \frac{\kappa}{c^2}(\imath_h y_h(\bar{u}_h) + c\bar{u}_h - a) < 0\right\}} \left(c\imath_h p_h'(\bar{u}_h)u_h + \kappa \imath_h y_h'(\bar{u}_h)u_h \right), \end{aligned} \tag{4.5}$$

where $y_h = y_h'(\bar{u}_h)u_h \in S_h^1$ solves (compactly written)

$$\mathcal{A}_h y_h = \imath_h^* u_h \tag{4.6a}$$

together with the boundary and initial conditions

$$y_h(t_j, \cdot) = 0 \text{ on } \Gamma,\ 1 \le j \le n_t, \quad \text{and} \quad y_h(0) = 0 \text{ in } \Omega \tag{4.6b}$$

for all $1 \le j \le n_t$, and $p_h = p_h'(\bar{u}_h)u_h \in S_h^1$ is the solution to

$$\int_\Omega \frac{p_h(t_j) - p_h(t_{j-1})}{h_t}\varphi_h - \alpha\nabla p_h(t_{j-1}) \cdot \nabla\varphi_h - c(t_{j-1})^{-1}p_h(t_{j-1})\varphi_h \,\mathrm{d}\mathbf{x} = \frac{1}{h_t}\int_{t_{j-1}}^{t_j} (\alpha_Q(t)y_h(t) - \kappa c(t)^{-1}u_h(t), \varphi_h)_{L^2(\Omega)} \,\mathrm{d}t \tag{4.7a}$$

for all $1 \le j \le n_t$ and $\varphi_h \in P_h^1$ together with the boundary and initial conditions

$$p_h(t_j, \cdot) = 0 \text{ on } \Gamma,\ 0 \le j \le n_t - 1, \quad \text{and} \quad p_h(0) = -\mathcal{R}_h^1\big(\alpha_\Omega y_h(T)\big) \text{ in } \Omega. \tag{4.7b}$$

The discretized version of Algorithm 1 is given by Algorithm 3.

Algorithm 3 (discretized semismooth Newton method).

1: Choose $u_h^0 \in S_h^0$, and set $k = 0$.
2: **repeat**
3: Compute $G_h(u_h^k)$ according to (4.5) and solve for δu_h^k:

$$G_h(u_h^k)\delta u_h^k = -F_h(u_h^k),$$

with F_h given by (4.3).
4: Set $u_h^{k+1} = u_h^k + \delta u_h^k$, and $k = k + 1$.
5: **until** some stopping rule is satisfied.

We have the following convergence theorem; see [7,12].

Theorem 4.1. *Let $\{u_h^k\}_{k\in\mathbb{N}}$ be a sequence in S_h^0 generated by Algorithm 3. Then, $\{u_h^k\}_{k\in\mathbb{N}}$ converges to the solution $u_h^* \in U_{\mathsf{ad}}^h$ of $(\hat{\mathbf{P}}_h)$ at a superlinear rate provided that $u_h^0 \in L^2(Q)$ is sufficiently close to u_h^*.*

To apply the results in [15] we have to introduce an approximation for the bilaterally control constrained problem $(\tilde{\mathbf{P}})$. For that purpose we set $\mathcal{T}_h = \imath_h\mathcal{A}_h^{-1}\imath_h^* : S_h^0 \to S_h^0$ and $\mathcal{F}_h = (\mathcal{T}_h + c\,\mathrm{id})$.

Note that due to the existence of a unique solution of (4.2) and [8], Corollary 3.1, we have

$$\|\mathcal{T}_h\|_{L(S_h^0, L^2(Q))} = \sup_{\|u_h\|_{S_h^0}=1} \|\imath_h\mathcal{A}_h^{-1}\imath_h^* u_h\|_{L^2(Q)} \le c_{\mathcal{A}}$$

and further

$$\|\mathcal{F}_h\|_{L(S_h^0, L^2(Q))} \le c_{\mathcal{A}} + \|c\|_{L^\infty(Q)} =: c_{\mathcal{F}}.$$

Let $u \in L^2(Q)$ be given. For $y = \mathcal{T}u_h$ and $y_h = \mathcal{T}_h u_h$ we have the estimate [8], Corollary 3.1,

$$\|y - y_h\|_{L^2(Q)} \le c_{\mathcal{T}} h^2 \|y\|_{L^2(0,T;H^2(\Omega))\cap H^1(0,T;L^2(\Omega))} \tag{4.8}$$

with a constant $c_{\mathcal{T}} > 0$ provided that Ω is convex with Lipschitz-continuous boundary and $f \in L^2(Q)$. We infer from (4.8) that

$$\|\mathcal{F} - \mathcal{F}_h\|_{L(S_h^0, L^2(Q))} = \sup_{\|u_h\|_{S_h^0}=1} \big(\|\imath(\mathcal{A}^{-1} - \mathcal{A}_h^{-1})\imath_h^* u_h\|_{L^2(Q)} + \|(\imath - \imath_h)\mathcal{A}_h^{-1}\imath_h^* u_h\|_{L^2(Q)} \big) \to 0 \quad \text{for } h \to 0. \tag{4.9}$$

Then we have the existence of $\mathcal{F}_h^{-1}$ for h sufficiently small.

Recall that (2.9) ensures the invertibility of $\mathcal{F}$, and define the restriction operator $\hat{\imath}_h : L^2(Q) \to S_h^0$ with $\hat{\imath}_h|_{S_h^0} = \mathrm{id}_{S_h^0}$. Then, for all sufficiently small h,

$$\|\,\mathrm{id} - \hat{\imath}_h \mathcal{F}^{-1}_{|_{S_h^0}} \mathcal{F}_h\|_{L(S_h^0)} \le \frac{1}{2}.$$

Hence, by the Neumann lemma, $\hat{\imath}_h \mathcal{F}^{-1}_{|_{S_h^0}}$ is invertible for all sufficiently small h.

From the perturbation lemma [16], p. 45, we then deduce the uniform (w.r.t. h) boundedness of $\|\mathcal{F}_h^{-1}\|_{L(S_h^0)}$ for all sufficiently small h.

In the following we assume that $\tilde{a}, \tilde{b} \in L^\infty(Q)$, see ($\tilde{\mathbf{P}}$), and that the set of admissible controls is defined by

$$V_{\mathsf{ad}}^h = \left\{ v_h \in S_h^0 \,\middle|\, \tilde{a} \le v_h \le \tilde{b} \text{ in } Q \right\} = V_{\mathsf{ad}} \cap S_h^0 = \mathcal{R}_h^0 V_{\mathsf{ad}}.$$

For $\tilde{a}, \tilde{b} \in L^\infty(Q)$ with $\tilde{b} - \tilde{a} \ge \epsilon > 0$, this can always be achieved after the transformation

$$w := \frac{2}{\tilde{b} - \tilde{a}} v - \frac{\tilde{a} + \tilde{b}}{\tilde{b} - \tilde{a}}$$

since then $-1 \le w \le 1$ a.e. in Q.

We approximate ($\tilde{\mathbf{P}}$) by the family of problems

$$\min \tilde{J}_h(v_h) \quad \text{s.t.} \quad v_h \in V_{\mathsf{ad}}^h, \tag{$\tilde{\mathbf{P}}_h$}$$

where $\tilde{J}_h = \tilde{J} \circ \mathcal{F}_h^{-1}$. It follows that ($\tilde{\mathbf{P}}_h$) has a unique optimal solution $v_h^* \in V_{\mathsf{ad}}^h$ for every $h > 0$. Moreover, we have the following result [15], Theorem 3.2.

Theorem 4.2. *Suppose that Ω is convex with Lipschitz-continuous boundary, $\tilde{a}$, $\tilde{b} \in L^\infty(Q)$ and the inhomogeneity f is sufficiently smooth. Then there exists a constant $c_* > 0$ satisfying*

$$\|v^* - v_h^*\|_{L^2(Q)} \le c_* h \quad \textit{for every } h > 0, \tag{4.10}$$

where $v_h^* \in V_{\mathsf{ad}}^h$ and $v^* \in V_{\mathsf{ad}}$ denote the unique optimal solutions of ($\tilde{\mathbf{P}}_h$) and ($\tilde{\mathbf{P}}$), respectively.

From Theorem 4.2 we immediately derive the following corollary.

Corollary 4.3. *Let u^* and u_h^* denote the unique solutions to ($\hat{\mathbf{P}}$) and ($\hat{\mathbf{P}}_h$), respectively. Under the assumptions of the previous theorem we have*

$$\lim_{h \to 0} \|u^* - u_h^*\|_{L^2(Q)} = 0.$$

Proof. Note that $u^* = \mathcal{F}^{-1} v^*$ and $u_h^* = \mathcal{F}_h^{-1} v_h^*$. Using (4.9)–(4.10) it follows that

$$\|u^* - u_h^*\|_{L^2(Q)} \le \|\mathcal{F}^{-1} v^* - \mathcal{F}^{-1} v_h^*\|_{L^2(Q)} + \|\mathcal{F}^{-1} v_h^* - \mathcal{F}_h^{-1} v_h^*\|_{L^2(Q)} \le O(h) + \|\mathcal{F}^{-1}\|_{L(L^2(Q))} \|\mathcal{F}_h - \mathcal{F}\|_{L(S_h^0, L^2(Q))} \|\mathcal{F}_h^{-1}\|_{L(S_h^0)} \|v_h^*\|_{S_h^0} \to 0$$

for $h \to 0$, which proves the assertion. □

Remark 4.4. We note that

$$\lim_{h\to 0} \|(\imath - \imath_h)\mathcal{A}_h^{-1} g\|_{L^2(Q)} = O(h) \text{ for } g \in Y'$$

yields

$$\|\mathcal{F} - \mathcal{F}_h\|_{L(S_h^0, L^2(Q))} = O(h)$$

and further in Corollary 4.3

$$\|u^* - u_h^*\|_{L^2(Q)} = O(h).$$

We end this section by establishing the mesh independence result. For this purpose we first recall that

$$W(0,T) \hookrightarrow L^3(Q) \tag{4.11}$$

for $d \le 3$, which is of importance with respect to the Newton-differentiability of the max- and min-operators; compare Remark 3.2. Further note that F, *cf.* (3.6), can be written as

$$\begin{aligned} F(u) = c^{-1}(\kappa(u_d - u) + L_p u + f_p) &- \max(0, L_\ell u + f_\ell - \kappa c^{-1}(c^{-1}b - u_d)) \\ &- \min(0, L_\ell u + f_\ell - \kappa c^{-1}(c^{-1}a - u_d)) \end{aligned}$$

with

$$L_p = -(\mathcal{A}^* + c^{-1}\,\mathrm{id})^{-1}(\alpha_Q \mathcal{A}^{-1} + c^{-1}\kappa\,\mathrm{id}) \in L(Y', W(0,T)), \tag{4.12a}$$

$$L_\ell = c^{-1}L_p + \kappa c^{-2}\mathcal{A}^{-1} \in L(Y', L^3(Q)), \tag{4.12b}$$

$$f_p = (\mathcal{A}^* + c^{-1}\,\mathrm{id})^{-1}(\alpha_Q(z_Q - \mathcal{A}^{-1}f) + c^{-1}\kappa u_d) \in L^3(Q), \tag{4.12c}$$

$$f_\ell = c^{-1}f_p + \kappa c^{-2}\mathcal{A}^{-1}f \in L^3(Q). \tag{4.12d}$$

In the discrete setting we analogously obtain

$$\begin{aligned} F_h(u_h) = c^{-1}(\kappa(u_d - u_h) &+ \imath_h(L_{p,h}u_h + f_{p,h})) \\ &- \max(0, \imath_h(L_{\ell,h}u + f_{\ell,h}) - \kappa c^{-1}(c^{-1}b - u_d)) \\ &- \min(0, \imath_h(L_{\ell,h}u + f_{\ell,h}) - \kappa c^{-1}(c^{-1}a - u_d)). \end{aligned}$$

Hence, we are in a framework similar to the one considered in [9] for elliptic equations. For establishing the mesh independence result, it therefore remains to verify Assumption 4.1 of [9] on Q, *i.e.*, we have to show that

$$\lim_{h\to 0^+} \max(\|f_p - f_{p,h}\|_{L^3(Q)}, \|f_\ell - f_{\ell,h}\|_{L^3(Q)}) = 0, \tag{4.13}$$

$$\lim_{h\to 0^+} \|u^* - u_h^*\|_{L^2(Q)} = 0, \tag{4.14}$$

$$\lim_{h\to 0^+} \max(\|L_p u^* - L_{p,h}u_h^*\|_{L^3(Q)}, \|L_\ell u^* - L_{\ell,h}u_h^*\|_{L^3(Q)}) = 0, \tag{4.15}$$

$$\max(\|L_{p,h}\|_{L(Y',L^3(Q))}, \|L_{\ell,h}\|_{L(Y',L^3(Q))}) \le K \tag{4.16}$$

for some constant $K > 0$. For this purpose note first that (4.14) immediately follows from Corollary 4.3. Moreover, the estimate

$$\|y - y_h\|_{L^3(Q)} \le \tilde{c}_{\mathcal{T}} h \, \|y\|_{L^2(0,T;H^2(\Omega))\cap H^1(0,T;L^2(\Omega))} \tag{4.17}$$

is satisfied with a constant $\tilde{c}_{\mathcal{T}} > 0$ independent of h. Indeed, [8], Corollary 3.1, yields

$$\|y - y_h\|_{L^2(0,T;H^1(\Omega))\cap H^{1/2}(0,T;L^2(\Omega))} \le \hat{c}_{\mathcal{T}} h \, \|y\|_{L^2(0,T;H^2(\Omega))\cap H^1(0,T;L^2(\Omega))},$$

and Theorem A.1 in Appendix A implies the estimate (4.17). For the remaining conditions observe that

$$\lim_{h\to 0^+} \max(\|\mathcal{A}^{-1} - \mathcal{A}_h^{-1}\|_{L(Y',L^3(Q))}, \|\mathcal{A}^{-*} - \mathcal{A}_h^{-*}\|_{L(Y',L^3(Q))}) = 0 \tag{4.18}$$

and consequently

$$\lim_{h\to 0^+} \|(\mathcal{A}^* + c\,\mathrm{id})^{-1} - (\mathcal{A}^* + c\,\mathrm{id})_h^{-1}\|_{L(Y',L^3(Q))} = 0. \tag{4.19}$$

Indeed, (4.18) follows from the decomposition $\|y - y_h\|_{L^3(Q)} \le \|y - y^{\mathrm{LM}}\|_{L^3(Q)} + \|y^{\mathrm{LM}} - y_h\|_{L^3(Q)}$, where y^{LM} denotes the approximation of y obtained by the implicit line method with piecewise linear and continuous interpolation in time (identical to the interpolation in time for y_h). Note that

$$\|y - y^{\mathrm{LM}}\|_{L^3(Q)} \le C(T)\|y - y^{\mathrm{LM}}\|^{3/2}_{C(0,T;L^2(\Omega))}\|y - y^{\mathrm{LM}}\|^{3/2}_{L^2(0,T;L^6(\Omega))},$$

where $C(T) > 0$ denotes some constant depending on T. Then, according to [17], Theorem 11.1, we have $\|y - y^{\mathrm{LM}}\|_{L^3(Q)} \to 0$ as $h \to 0$. The convergence of $\|y^{\mathrm{LM}} - y_h\|_{L^3(Q)}$ follows from standard finite element estimates. From (4.18) and (4.19) we infer that there exist $h_p > 0$ and $K_p > 0$ (independently of h) such that

$$\lim_{h\to 0^+} \|L_p u^* - L_{p,h} u_h^*\|_{L^3(Q)} = 0 \quad \text{and} \tag{4.20a}$$

$$\|L_{p,h}\|_{L(Y',L^3(Q))} \le K_p \quad \text{for all } 0 < h \le h_p. \tag{4.20b}$$

Similarly one shows that there exist h_ℓ and $K_\ell > 0$ (independently of h) such that

$$\lim_{h\to 0^+} \|L_\ell u^* - L_{\ell,h} u_h^*\|_{L^3(Q)} = 0 \quad \text{and} \tag{4.21a}$$

$$\|L_{\ell,h}\|_{L(Y',L^3(Q))} \le K_\ell \quad \text{for all } 0 < h \le h_\ell. \tag{4.21b}$$

This verifies (4.15) and (4.16). Now, condition (4.13) immediately follows. Under these conditions which establish a $L^3(Q)$-$L^2(Q)$ norm gap in order to obtain Newton differentiability of the max- and min-terms in F, the following mesh independence result holds.

Theorem 4.5. *Let (4.13)–(4.16) be satisfied. Further assume that*

$$\mathrm{meas}(\{|L_\ell u^* + f_\ell - \kappa c^{-2} b| = 0\}) = 0, \tag{4.22a}$$

$$\mathrm{meas}(\{|L_\ell u^* + f_\ell - \kappa c^{-2} a| = 0\}) = 0 \tag{4.22b}$$

holds true. Then, for arbitrarily fixed $\theta \in (0,1)$, there exist $\delta^* > 0$ and $h^* > 0$ such that for all $h \le h^*$ and $k \in \mathbb{N}_0$

$$\begin{aligned} \|u^{k+1} - u^*\|_{L^2(Q)} &\le \theta\|u^k - u^*\|_{L^2(Q)}, \\ \|u_h^{k+1} - u_h^*\|_{S_h^0} &\le \theta\|u_h^k - u_h^*\|_{S_h^0} \end{aligned}$$

provided that $\max(\|u^0 - u^*\|_{L^2(Q)}, \|u_h^0 - u_h^*\|_{S_h^0}) \le \delta^*$.

Proof. The proof lies in the verification of (4.13)–(4.16) before the theorem and in [9] Lemma 4.1 and Theorem 4.1. □

Note that assumption (4.22) guarantees that the non-differentiability of the max- and min-operators in F is concentrated only on a set of measure zero. It corresponds to a strict complementarity assumption in connection with the pointwise inequality constraints. The assertion of the theorem states that, given a linear rate of convergence, for sufficiently small mesh sizes h and sufficiently good initial guesses u^0 and u_h^0 the continuous as well as the discrete Newton process converge at this specified linear rate.

Finally we remark that if one also has to consider discretizations a_h, b_h, $u_{d,h}$, c_h, f_h of the data a, b, u_d, c, f then, under our regularity assumptions, the corresponding interpolation errors tend to zero as $h \to 0^+$ and our results remain true.

5. Preconditioning and numerical experiments

In this section we validate our theoretical findings by numerical tests. Further, since the discretization yields large scale finite dimensional problems, we have to resort to iterative solvers for the subsystems occurring in our primal-dual active-set method. This requires suitable preconditioning of the system matrices. Here we propose preconditioning techniques taking into account the active/inactive set structure.

All test problems considered in this section are two-dimensional with $\Omega = (0,1) \times (0,1)$. Further, in all test cases the cost parameters had the values $\alpha_Q \equiv 1$, $\alpha_\Omega \equiv 10$, and $\kappa = 0.1$. We present three examples including bilateral control-state constraints, degenerate solutions or lack of strict complementarity. While Example 5.1.1 satisfies $c > 0$, in Examples 5.1.2–5.1.3 we consider situations where $c < 0$, *i.e.*, where a strict positivity assumption on c is violated. The latter assumption is standard in Lavrentiev-type regularization. However, as long as our assumptions (2.9) and (3.4) are satisfied our theoretical results remain true.

All coding is done in Matlab, and the computations are performed on a standard 1.7 GHz desktop PC.

5.1. Presentation of the examples

In this subsection we specify the numerical examples. We also highlight some properties of the respective solution such as degeneracy and lack of strict complementarity. By degeneracy we refer to situations where the primal quantity $c\,u^* + y^*$ and/or the corresponding Lagrange multiplier λ^* exhibit a very flat transition into the active set and/or zero. The problem is said to satisfy strict complementarity, if $(\{|c\,u^* + y^* - b| = 0\} \cup \{|c\,u^* + y^* - a| = 0\}) \cap \{|\lambda^*| = 0\}$ has zero measure. Lack of strict complementarity as well as degeneracy may complicate the numerical active set detection and, thus, may slow down a solution algorithm. It will turn out, however, that our semismooth Newton, or equivalently primal-dual active set, solver is not affected by these adverse situations.

5.1.1. *Example: Bilateral constraints*

We consider problem (**P**) with final time $T = 1$, heat conductivity $\alpha = 0.1$, weighting functions $\alpha_Q \equiv 1$ and $\alpha_\Omega \equiv 10$ in the cost functional, lower bound $a \equiv 0.02$ and upper bound $b \equiv 0.1$ for the inequality constraints, and $c \equiv 0.1$. Moreover, we have

$$\begin{aligned} y_\circ(\mathbf{x}) &= (1-x_1)(1-x_2)\sin(7\pi x_1 x_2)\cos(2\pi x_1), \\ z_Q(t,\mathbf{x}) &= 10 y_\circ(\mathbf{x})\cos(2\pi t x_1), \\ z_\Omega(\mathbf{x}) &= \frac{1}{10}\, y_\circ \cos(2.4\pi T x_1), \\ u_d(t,\mathbf{x}) &= (1-x_1)(1-x_2)\sin(2\pi x_1 x_2)\mathrm{e}^t, \\ f(t,\mathbf{x}) &= 20 x_1 x_2 (1-x_1)(1-x_2)\sin(\pi t x_1) \end{aligned}$$

for $(t,\mathbf{x}) \in \overline{Q}$ and $\mathbf{x} = (x_1, x_2) \in \overline{\Omega}$.

Figures 1–4 show the active/inactive set structure at times $t = 0.025$, $t = 0.05$, $t = 0.65$ and $t = T = 1$. The inactive set is displayed in white, the a-active set in gray, and in black the b-active set is shown. The solution is active from below throughout the whole time interval $(0,T)$. With respect to the upper bound b, we observe active and inactive zones over the entire time horizon.

5.1.2. *Example: Degenerate solution*

In contrast to the previous example, we consider unilateral constraints of the type $y + cu \le b$.

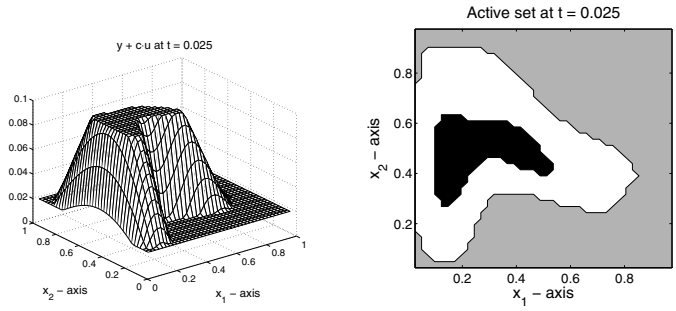

FIGURE 1. Example 5.1.1: Inequality and active sets at time step $t = 0.025$.

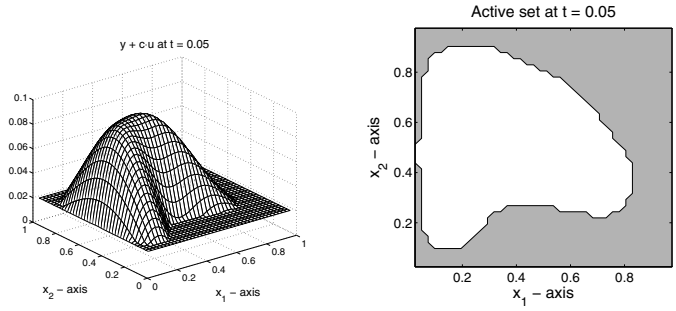

FIGURE 2. Example 5.1.1: Inequality and active sets at time step $t = 0.05$.

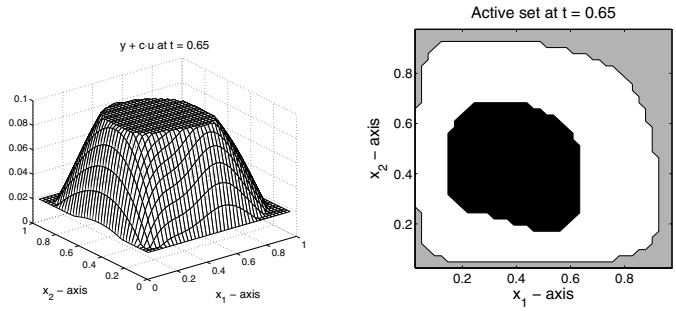

FIGURE 3. Example 5.1.1: Inequality and active sets at time step $t = 0.65$.

Further data values are $T = 1$, $\alpha \equiv 0.1$, $\alpha_Q \equiv 1$ and $\alpha_\Omega \equiv 1$. The upper bound is $b \equiv 1$, and $c \equiv -0.1$. Moreover, let

$$\begin{aligned}
y_\circ(\mathbf{x}) &= (1-x_1)(1-x_2)\sin(7\pi x_1 x_2)\cos(2\pi x_1),\\
z_Q(t,\mathbf{x}) &= \frac{1}{\alpha_Q}\Big(-\frac{\partial p^\dagger}{\partial t}(t,\mathbf{x}) - \Delta p^\dagger(t,\mathbf{x}) + \lambda^\dagger(t,\mathbf{x})\Big) + y^\dagger(t,\mathbf{x}),\\
z_\Omega(\mathbf{x}) &= y_\circ(\mathbf{x})(1+T)\sin(3.6\pi T x_2),\\
u_d(t,\mathbf{x}) &= u^\dagger(t,\mathbf{x}) - \frac{1}{\alpha}\big(p^\dagger(t,\mathbf{x}) + c\lambda^\dagger(t,\mathbf{x})\big),\\
f(t,\mathbf{x}) &= \frac{\partial y^\dagger}{\partial t}(t,\mathbf{x}) - \Delta y^\dagger(t,\mathbf{x}) - u^\dagger(t,\mathbf{x}),
\end{aligned}$$

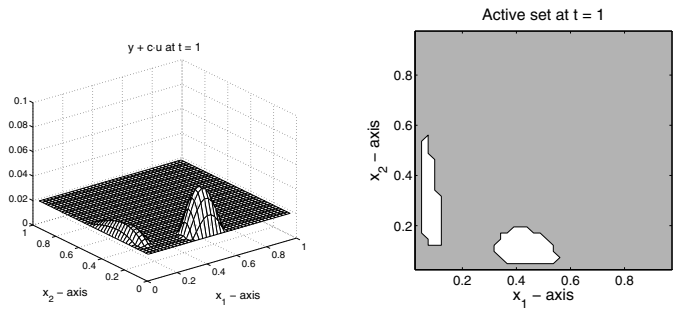

FIGURE 4. Example 5.1.1: Inequality and active sets at time step $t = T = 1$.

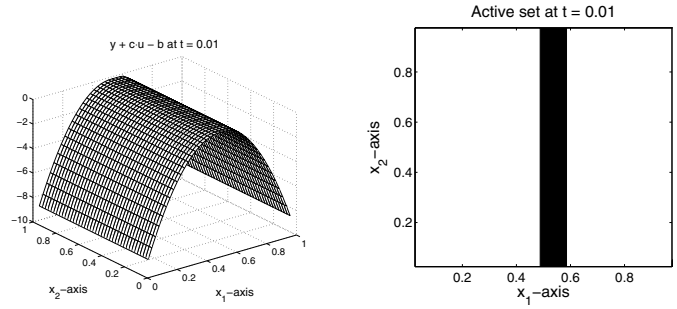

FIGURE 5. Example 5.1.2: Inequality and active sets at time step $t = 0.01$.

where

$$\begin{aligned} y^\dagger(t,\mathbf{x}) &= y_\circ \sin(3\pi t x_2)(1+t), \\ p^\dagger(t,\mathbf{x}) &= 10\sin(3\pi x_1)\cos(3\pi x_2 + 1.5\pi)\mathrm{e}^t, \\ u^\dagger(t,\mathbf{x}) &= \frac{1}{c}\left(y^\dagger(t,\mathbf{x}) - \vartheta(t)\left(\cos(3\pi t x_2)+1\right)\eta(\mathbf{x}) - d\right), \\ \lambda^\dagger(t,\mathbf{x}) &= \begin{cases} 10 y_\circ(\mathbf{x})^2 \mathrm{e}^t & \text{in } Q_1 \\ 0 & \text{else,} \end{cases} \end{aligned}$$

and

$$\eta(\mathbf{x}) = \begin{cases} 1 - 4x_1 + 4x_1^2 & \text{for } 0 < x_1 \le 0.5, \\ 0.9744 - 4x_1 + 4x_1^2 & \text{for } 0.58 \le x_1 < 1, \\ 0 & \text{else,} \end{cases}$$

$$\vartheta(t) = -5 + \frac{1}{9}\left(184.6875\, t - 241.375\, t^2 + 97.1875\, t^3\right),$$

$$Q_1 = \left\{(t,x_1,x_2) \in Q : \vartheta(t)\left(\cos(3\pi t x_2)+1\right)\eta(\mathbf{x}) > 0\right\}.$$

Figures 5–8 show the active/inactive sets at times $t = 0.01$, $t = 0.59$, $t = 0.60$ and $t = 0.61$. As in Example 5.1.1, the inactive set is displayed in white, the active set in black. Figures 6–8 clearly show degeneracy, *i.e.*, a very flat transition of $y + cu$ into the active set, with respect to space and, when considering the slowly (in x_2-direction) moving active area in the center of the figure, to a certain extent also in time.

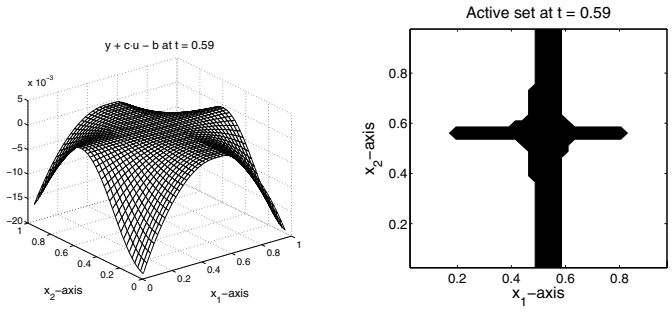

FIGURE 6. Example 5.1.2: Inequality and active sets at time step $t = 0.59$.

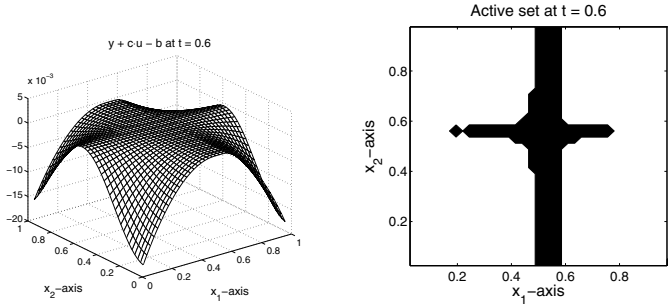

FIGURE 7. Example 5.1.2: Inequality and active sets at time step $t = 0.6$.

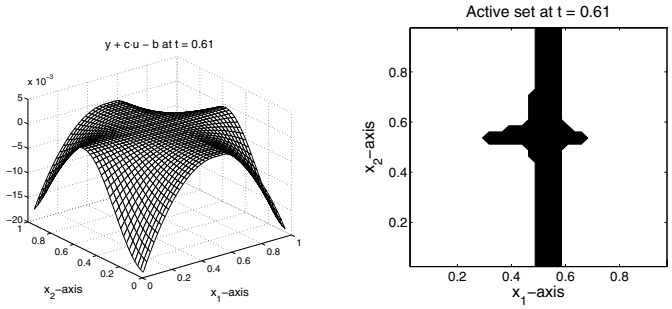

FIGURE 8. Example 5.1.2: Inequality and active sets at time step $t = 0.61$.

5.1.3. *Example: Lack of strict complementarity*

This unilaterally constrained example is constructed such that the solution has areas lacking strict complementarity.

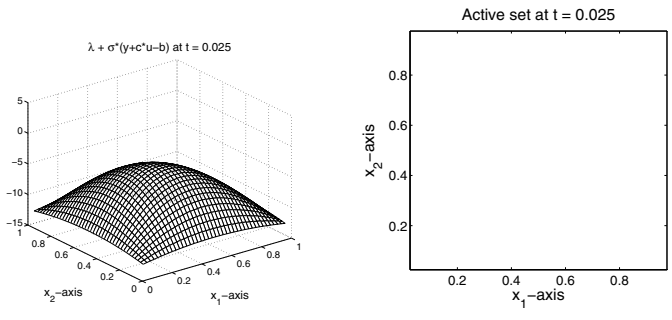

FIGURE 9. Example 5.1.3: Inequality and active sets at time step $t = 0.025$.

We have $T = 1$, $\alpha \equiv 0.1$, $\alpha_Q \equiv 1$ and $\alpha_\Omega \equiv 1$. The upper bound is $b \equiv 1$, and $c \equiv -0.1$. Moreover, we have

$$\begin{aligned}
y_\circ(\mathbf{x}) &= (1-x_1)(1-x_2)\sin(7\pi x_1 x_2)\cos(2\pi x_1),\\
z_Q(t,\mathbf{x}) &= \frac{1}{\alpha_Q}\left(-\frac{\partial p^\dagger}{\partial t}(t,\mathbf{x}) - \Delta p^\dagger(t,\mathbf{x}) + \lambda^\dagger(t,\mathbf{x})\right) + y^\dagger(t,\mathbf{x}),\\
z_\Omega(\mathbf{x}) &= y_\circ(\mathbf{x})(1+T)\sin(3.6\pi T x_2),\\
u_t(t,\mathbf{x}) &= u^\dagger(t,\mathbf{x}) - \frac{1}{\alpha}\left(p^\dagger(t,\mathbf{x}) + c\lambda^\dagger(t,\mathbf{x})\right),\\
f(t,\mathbf{x}) &= \frac{\partial y^\dagger}{\partial t}(t,\mathbf{x}) - \Delta y^\dagger(t,\mathbf{x}) - u^\dagger(t,\mathbf{x}),
\end{aligned}$$

where

$$\begin{aligned}
y^\dagger(t,\mathbf{x}) &= y_\circ(\mathbf{x})\left(1+\sin(\pi t^2)\right),\\
p^\dagger(t,\mathbf{x}) &= 10\sin(3\pi x_1)\cos(3\pi x_2 + 1.5\pi)\mathrm{e}^t,\\
u^\dagger(t,\mathbf{x}) &= -\frac{1}{c}\left(y^\dagger(t,\mathbf{x}) - \min\{\eta - t, 0\} + b\right),\\
\lambda^\dagger(t,\mathbf{x}) &= \begin{cases} 10 y_\circ(\mathbf{x})^2\mathrm{e}^t & \text{on } Q_1\\ 0 & \text{else,}\end{cases}
\end{aligned}$$

and

$$\begin{aligned}
\eta(\mathbf{x}) &= 20(1-x_1)(1-x_2)x_1 x_2\cos(\pi x_1 x_2) - 1.3,\\
Q_1 &= \{(t,\mathbf{x}) \in Q : \eta(\mathbf{x}) + \min\{t, 0.6\} > 0\}.
\end{aligned}$$

Figures 9–12 show the value for $\lambda^* + \sigma(y^* - c\cdot u^* - b)$ and the active sets at times $t = 0.025$, $t = 0.5$, $t = 0.7$ and $t = 1$. The inactive set is displayed in white, the strongly active set, *i.e.* the set where strict complementarity holds true, in black. The weakly active set, which is the set where $y^* + cu^* - b = 0$, and the dual variable $\lambda^* = 0$, are displayed in gray.

Figure 10 shows the active set at the center of the domain Ω. In Figures 11 and 12 an annulus area with lack of strict complementarity surrounding the active set can be observed.

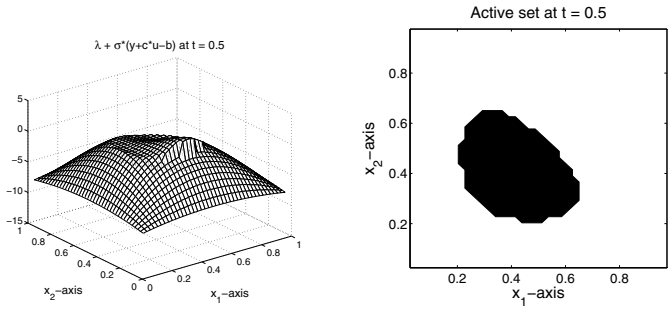

FIGURE 10. Example 5.1.3: Inequality and active sets at time step $t = 0.5$.

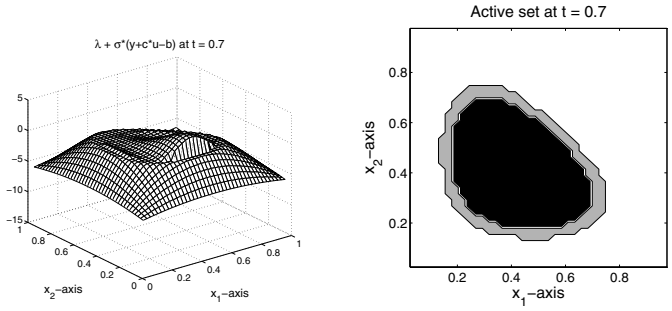

FIGURE 11. Example 5.1.3: Inequality and active sets at time step $t = 0.7$.

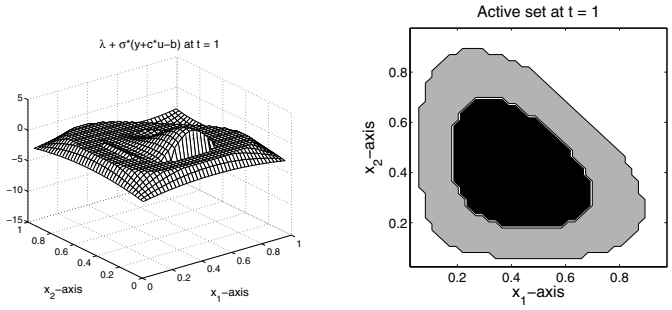

FIGURE 12. Example 5.1.3: Inequality and active sets at time step $t = 1$.

5.2. **Solving the linear system**

In every step of the discretized primal-dual active set method (resp. semismooth Newton method) a linear system of the form

$$\begin{pmatrix} M_1^k & M_2^k \\ (M_2^k)^T & M_4^k \end{pmatrix} \begin{pmatrix} y^{k+1} \\ p^{k+1} \end{pmatrix} = \begin{pmatrix} r_1^k \\ r_2^k \end{pmatrix} \tag{5.1}$$

has to be solved. Above, $M_1^k \in \mathbb{R}^{n^2m \times n^2m}$ is an invertible matrix, $M_2^k \in \mathbb{R}^{n^2m \times n^2m}$ is an upper block triangular matrix, and $M_4^k \in \mathbb{R}^{n^2m \times n^2m}$ is not invertible, in general. In our numerical tests we study several preconditioned iterative solvers.

5.2.1. *Preconditioned reduced system*

The system (5.1) can be solved by a GMRES iteration. In order to speed up the process we employ a problem related preconditioner which is based on the active and inactive set structure. In fact, we use an incomplete LU-factorization of an approximation of the M_2^k-block in (5.1). Recall that M_2^k is the discretization of the operator $-\frac{\partial}{\partial t} - \Delta - \frac{1}{c}\chi_{A^k}$. If we define $M_2 \in \mathbb{R}^{n^2 m}$ as the discretization of $-\frac{\partial}{\partial t} - \Delta$, then M_2^k can be written as the sum of a constant part and a part depending on the active set of the previous iteration; *i.e.*, $M_2^k = M_2 - D^k$, where D^k is the discretization of $\frac{1}{c}\chi_{A^k}$. Let $\tilde{L}$ and $\tilde{U}$ be the lower and upper triangular, respectively, incomplete LU-factors of the constant part M_2, *i.e.*, $\tilde{L}\tilde{U} \approx \tilde{P}M_2$, where $\tilde{P}$ is a permutation matrix. Then, instead of (5.1), we solve the equivalent system

$$\begin{pmatrix} \tilde{P}M_1^k\tilde{P}^T & \tilde{P}M_2^k \\ \left(M_2^k\right)^T\tilde{P}^T & M_4^k \end{pmatrix} \begin{pmatrix} \tilde{y}^{k+1} \\ p^{k+1} \end{pmatrix} = \begin{pmatrix} \tilde{P}r_1^k \\ r_2^k \end{pmatrix} \tag{5.2}$$

$$y^{k+1} = \tilde{P}^T\tilde{y}^{k+1}. \tag{5.3}$$

The preconditioner is realized by solving

$$\begin{pmatrix} 0 & \tilde{L}\tilde{U} \\ \tilde{U}^T\tilde{L}^T & 0 \end{pmatrix} \begin{pmatrix} \tilde{y}^{k+1} \\ p^{k+1} \end{pmatrix} = r \tag{5.4}$$

for some given right hand side $r \in \mathbb{R}^{2n^2 m}$. In our numerical tests the following correction term in (5.4) improved the effect of the preconditioner:

$$\begin{aligned} \tilde{U}^T\tilde{L}^T y^{k+1} &= \tilde{r}_2 \\ \tilde{L}\tilde{U}p^{k+1} &= \tilde{r}_1 - \tilde{P}M_1^k\tilde{P}^T y^{k+1}, \end{aligned} \tag{5.5}$$

where $\tilde{r}_1$ denotes the vector containing the first $n^2 m$ components of a given right hand side r and $\tilde{r}_2$ denotes the vector containing the second $n^2 m$ components of r. Note that the correction term on the right hand side of the second equation provides information contained in M_1^k.

5.2.2. *Preconditioned Schur complement*

As an alternative to the preceding approach, for solving the linear system (5.1) a reduction to a symmetric positive definite system and a subsequent application of the CG-method can be employed.

Computing the Schur complement, the linear system (5.1) is equivalent to

$$\left(\left(M_2^k\right)^T\left(M_1^k\right)^{-1}M_2^k - M_4^k\right)p^{k+1} = \left(M_2^k\right)^T\left(M_1^k\right)^{-1}r_1^k - r_2^k \tag{5.6}$$

$$y^{k+1} = \left(M_1^k\right)^{-1}\left(r_1^k - M_2^k p^{k+1}\right). \tag{5.7}$$

If M_2^k is invertible, then the system matrix in (5.6) is positive definite. In this case (5.6) can be solved by using a preconditioned conjugate gradient method. In this work, we tested several preconditioning techniques. The following scheme produced the best results.

Let the matrices L, U, P be the outcome of an incomplete LU-factorization of the constant part M_2^T, *i.e.* $LU \approx PM_2^T$. The idea behind the preconditioning of the equation (5.6) is to neglect the matrix M_4^k and to solve the equation $M_2^T\left(M_1^k\right)^{-1}M_2 p = r$ for a given right hand side r. We approximate the latter equation by using the incomplete factorization. This leads to the following system

$$\begin{aligned} LU\left(M_1^k\right)^{-1}U^T L^T\tilde{p} &= Pr \\ p &= P^T\tilde{p}. \end{aligned} \tag{5.8}$$

TABLE 1. Comparison of the total number of inner iterations and CPU-time for Example 5.1.1 solving the full system.

		Precond. (5.4)		Precond. (5.5)	
h^{-1}	h_t^{-1}	#it	CPU(s)	#it	CPU(s)
30	20	41	47.01	25	33.82
30	30	43	84.15	25	58.49
40	30	49	190.33	30	143.31
40	40	49	312.31	30	219.44
50	40	58	766.31	35	574.41

TABLE 2. Comparison of the total number of inner iterations and CPU-time for Example 5.1.1 solving the reduced system.

		Precond. (5.8)		SSOR-CG	
h^{-1}	h_t^{-1}	#it	CPU(s)	#it	CPU(s)
30	20	58	41.65	2392	266.77
30	30	61	70.64	2543	436.39
40	30	95	191.89	4414	1500.32
40	40	100	297.56	4716	2249.85
50	40	164	800.62	7162	5610.81

In our numerics, in order to further reduce the computational cost, we replace M_1^k by a diagonal matrix $\tilde{M}_1^k$, which arises from a mass lumping technique.

5.3. **Discussion**

Next we discuss the effect of the preconditioning techniques. Further we verify our theoretical mesh independence and fast local convergence results.

5.3.1. *Effect of preconditioning*

For Example 5.1.1, Table 1 compares the effect of the preconditioners (5.4) and (5.5), respectively, when using the GMRES-method for solving the reduced system (5.2). For the same test example, in Table 2 we compare the performance of the CG method for solving the reduced system (5.6) and when using the preconditioning scheme (5.8) and a standard SSOR-CG preconditioner, respectively.

Based on our experience resulting also from further test runs (including Examples 5.1.2 and 5.1.3) and based on the results displayed in Tables 1 and 2 we draw the following conclusions:

- The correction step added in (5.5) improves the performance of the preconditioner significantly.
- The preconditioning scheme (5.8) is clearly more effective than a standard SSOR-CG preconditioner.
- Compared to the preconditioners for the Schur complement, the preconditioners used for the reduced system show a remarkable stability of the number of iterations (#it) with respect to varying mesh size.
- The preconditioned GMRES solvers for the reduced system require more memory than the preconditioned Schur complement solvers.

5.3.2. *Dependence on the mesh-size*

In Table 3 we document the results for Examples 5.1.1–5.1.3 for various mesh sizes. In order to further reduce the computational burden, we employ an inexact semismooth Newton (respectively primal-dual active set) method. In fact, the Newton system in every iteration of the method is solved to a relative residual (res_{rel})

TABLE 3. Number of SSN iterations for Examples 5.1.1–5.1.3.

		Ex. 5.1.1	Ex. 5.1.2	Ex. 5.1.3
h^{-1}	h_t^{-1}	#it	#it	#it
30	20	4	6	6
30	30	4	6	8
40	30	4	5	7
40	40	4	6	6
50	40	4	5	6
50	50	4	6	6
60	60	5	5	6
64	64	5	5	6
64	96	5	5	6

norm accuracy smaller than a tolerance ε^k. The quantity ε^k is initialized by $\varepsilon^1 := 0.1$. Then, in each semismooth Newton iteration the tolerance is computed as

$$\varepsilon^{k+1} = \max\left\{\varepsilon_{min}, \min\left\{0.8\,\varepsilon^k, 10^{-3}\,\|res_{rel}\|\right\}\right\}$$

with $\varepsilon_{min} = 10^{-9}$. The semismooth Newton algorithm is terminated if either the active sets of two consecutive iterations do not change, or as soon as the norm of the NCP-function based reformulation of the complementarity system is smaller than ε_s and $\|res_{rel}\| < \varepsilon_s$. The stopping tolerance is chosen as $\varepsilon_s = \varepsilon_{min}/h^2$.

The results displayed in Table 3 clearly verify the mesh independent behavior for all three examples.

5.3.3. *Superlinear convergence*

In this section the convergence behaviour of the algorithm is investigated. Our theoretical results in Section 4 predict a locally superlinear rate.

For Example 5.1.1 with a mesh size $h^{-1} = 64$ and a time step size $h_t^{-1} = 96$, the quotients $\|u_h^{k+1} - u_h^*\|_{S_h^0}/\|u_h^k - u_h^*\|_{S_h^0}$ and $\|y_h^{k+1} - y_h^*\|_{S_h^1}/\|y_h^k - y_h^*\|_{S_h^1}$ are plotted in Figure 13. As before, we employ an inexact semismooth Newton method, but now with $\varepsilon_{min} = 10^{-10}$. The reference (exact) solution was computed by using smaller stopping tolerances, *i.e.*, $\varepsilon = 10^{-12}$ for the inner iterations and ε/h^2 for the complementarity condition and the norm of the relative residual for the outer iterations.

In all cases shown in Figures 13 and 14 we find that the respective quotient is decreasing to zero as the iteration number k is increased. This behavior indicates a superlinear convergence rate of our method. Moreover, combining these results with the ones of the preceding section the superlinear rate appears to be even mesh independent.

A. EMBEDDING RESULT

We define the Sobolev space

$$H^{1,1/2}(Q) = H^{1/2}(0,T;L^2(\Omega)) \cap L^2(0,T;H^1(\Omega)).$$

Theorem A.1. *The following continuous embedding holds true:*

$$H^{1,1/2}(Q) \subset L^q(Q) \text{ for } q \in \left(2, \frac{10}{3}\right).$$

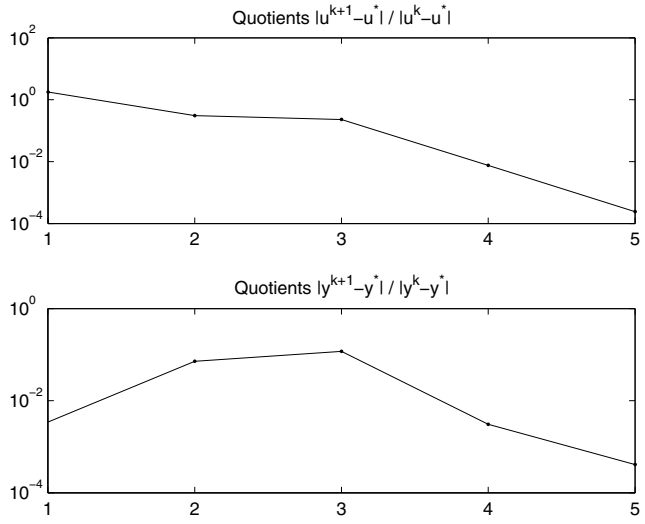

FIGURE 13. Quotients for convergence rate for Example 5.1.1.

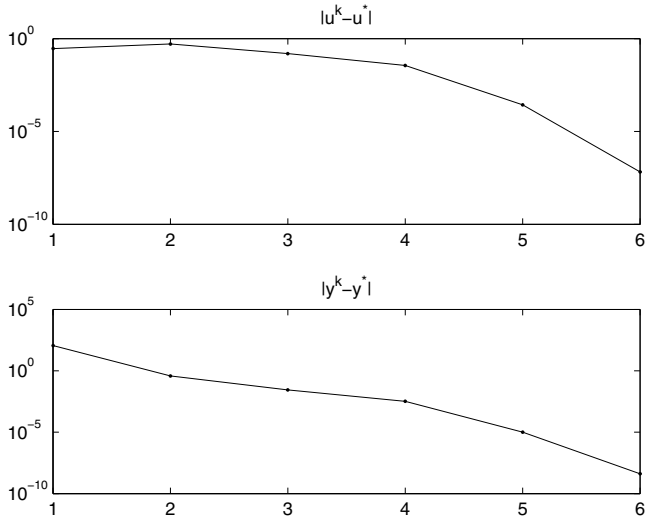

FIGURE 14. Errors for convergence rate for Example 5.1.1.

Proof. First recall that $H^1(\Omega) \subset L^6(\Omega)$ for $\Omega \subset \mathbb{R}^d$, $d \leq 3$. Moreover, we have

$$H^{1/2}(0,T;L^2(\Omega)) \subset L^{\hat{q}}(0,T;L^2(\Omega)) \text{ for } \hat{q} \in [2,+\infty),$$

which follows from [19], Remark 1, p. 328. Hence, we conclude

$$H^{1,1/2}(Q) \subset L^2(0,T;L^6(\Omega)) \cap L^{\hat{q}}(0,T;L^2(\Omega)) \text{ for } \hat{q} \in [2,+\infty). \tag{1.9}$$

Subsequently, let $2 < q = (q-r) + r$, $0 \le r \le q$, $p_1, p_2, s_1, s_2 \in [1, +\infty)$ with

$$p_1^{-1} + p_2^{-1} = 1 \text{ and } s_1^{-1} + s_2^{-1} = 1. \tag{1.10}$$

Further, the following conditions have to hold true:

$$p_1(q-r) = 6, \quad rp_2 = 2, \quad s_1(q-r) = 2, \quad rs_2 = \kappa, \tag{1.11}$$

with $\kappa > 0$. Then, we have

$$\begin{aligned} \int_0^T \int_\Omega |y(t,\mathbf{x})|^q \mathrm{d}\mathbf{x}\,\mathrm{d}t &\le \int_0^T \|y\|_{L^6(\Omega)}^{q-r} \|y\|_{L^2(\Omega)}^r \mathrm{d}t \\ &\le \|y\|_{L^2(0,T;L^6(\Omega))}^{q-r} \|y\|_{L^\kappa(0,T;L^2(\Omega))}^r, \end{aligned} \tag{1.12}$$

where we used $1/s_1 = 2/(q-r)$ and $1/s_2 = r/\kappa$.

The conditions in (1.10) yield

$$q = 6 - 2r \quad \text{and} \quad q = 2 + \frac{r(\kappa-2)}{\kappa}. \tag{1.13}$$

The requirement $q > 2$ yields $\kappa > 2$ and $r < 2$. The second relation in (1.13) implies $2 < q < 2 + r < 4$. Let $q = 4 - \epsilon$ with $\epsilon \in (0,2)$. Then,

$$p_1 = \frac{4}{2-\epsilon}, \quad p_2 = \frac{4}{2+\epsilon}, \quad s_1 = \frac{4}{3(2-\epsilon)}, \quad s_2 = \frac{2\kappa}{2+\epsilon}.$$

Now, (1.10) and $\kappa > 2$ imply $\epsilon \in (2/3, 2)$. Hence,

$$q \in \left(2, \frac{10}{3}\right).$$

The assertion follows from (1.9), which implies that the right hand side in (1.12) is bounded. □

Acknowledgements. The authors are indebted to the referees for their comments which helped to improve the presentation. The authors gratefully acknowledge support by the Austrian Science Fund FWF under grant No. P16760-N12. M.H. and I.K. also acknowledge support by the Austrian Federal Ministry of Education, Science and Culture (bm:bwk) and the FWF under START-program Y305 "Interfaces and Free Boundaries".

References

[1] R.A. Adams, *Sobolev Spaces, Pure and Applied Mathematics* **65**. Academic Press, New York-London (1975).
[2] A. Battermann and M. Heinkenschloss, Preconditioners for Karush-Kuhn-Tucker matrices arising in the optimal control of distributed systems, in *Control and estimation of distributed parameter systems (Vorau, 1996)*, *Internat. Ser. Numer. Math.* **126** (1998) 15–32.
[3] A. Battermann and E.W. Sachs, Block preconditioners for KKT systems in PDE-governed optimal control problems, in *Fast solution of discretized optimization problems (Berlin, 2000)*, *Internat. Ser. Numer. Math.* **138** (2001) 1–18.
[4] G. Biros and O. Ghattas, Parallel Lagrange-Newton-Krylov-Schur methods for PDE-constrained optimization. I. The Krylov-Schur solver. *SIAM J. Sci. Comput.* **27** (2005) 687–713.
[5] R. Dautray and J.-L. Lions, *Evolution Problems I, Mathematical Analysis and Numerical Methods for Science and Technology* **5**. Springer-Verlag, Berlin (1992).
[6] L.C. Evans, *Partial Differential Equations, Graduate Studies in Mathematics* **19**. American Mathematical Society, Providence, Rhode Island (1998).
[7] C. Geiger and C. Kanzow, *Theorie und Numerik restringierter Optimierungsaufgaben*. Springer-Verlag, Berlin (2002).
[8] W. Hackbusch, Optimal $H^{p,p/2}$ error estimates for a parabolic Galerkin method. *SIAM J. Numer. Anal.* **18** (1981) 681–692.

[9] M. Hintermüller, Mesh-independence and fast local convergence of a primal-dual active-set method for mixed control-state constrained elliptic control problems. *ANZIAM Journal* **49** (2007) 1–38.
[10] M. Hintermüller and M. Hinze, A SQP-semismooth Newton-type algorithm applied to control of the instationary Navier-Stokes system subject to control constraints. *SIAM J. Opt.* **16** (2006) 1177–1200.
[11] M. Hintermüller and M. Ulbrich, A mesh-independence result for semismooth Newton methods. *Math. Program. Ser. B* **101** (2004) 151–184.
[12] M. Hintermüller, K. Ito and K. Kunisch, The primal-dual active set strategy as a semi-smooth Newton method. *SIAM J. Opt.* **13** (2003) 865–888.
[13] M. Hintermüller, S. Volkwein and F. Diwoky, Fast solution techniques in constrained optimal boundary control of the semilinear heat equation. *Internat. Ser. Numer. Math.* **155** (2007) 119–147.
[14] J.-L. Lions, *Optimal control of systems governed by partial differential equations*. Springer-Verlag, Berlin (1971).
[15] K. Malanowski, Convergence of approximations versus regularity of solutions for convex, control-constrained optimal control problems. *Appl. Math. Optim.* **8** (1981) 69–95.
[16] J.M. Ortega and W.C. Rheinboldt, *Iterative Solution of Nonlinear Equations in several Variables, Computer Science and Applied Mathematics*. Academic Press, New York (1970).
[17] K. Rektorys, *The Method of Discretization in Time and Partial Differential Equations, Mathematics and Applications* **4**. D. Reichel Publishing Company, Boston-Dordrecht-London (1982).
[18] R. Temam, *Navier-Stokes Equations, Studies in Mathematics and its Applications*. North-Holland, Amsterdam (1979).
[19] H. Triebel, *Interpolation Theory, Function Spaces, Differential Operators*. North-Holland Publishing Company, Amsterdam (1978).
[20] F. Tröltzsch, Regular Lagrange multipliers for control problems with mixed pointwise control-state constraints. *SIAM J. Opt.* **15** (2005) 616–634.
[21] F. Tröltzsch, *Optimale Steuerung partieller Differentialgleichungen*. Vieweg Verlag, Wiesbaden (2005).